A Framework for Understanding User Interaction with Content-based Image Retrieval: Model, Interface and Users

Thesis submitted for the degree of

Doctor of Philosophy

at The Open University.

by

Haiming Liu

Knowledge Media Institute

The Open University.

July 2010

Abstract

User interaction is essential to the communication between users and content-based image retrieval (CBIR) systems. User interaction covers three key elements: an interaction model, an interactive interface and users. The three key elements combine to enable effective interaction to happen. Many studies have investigated different aspects of user interaction. However, there is lack of research in combining all three elements in an integrated manner, especially through well-principled data analysis based on a systematic user study. In this thesis, we investigate the combination of all three elements for interactive CBIR.

We first propose uInteract - a framework including a novel four-factor user interaction model (FFUIM) and an interactive interface. The FFUIM aims to improve interaction and search accuracy of the relevance feedback mechanism for CBIR. The interface delivers the FFUIM visually, aiming to support users in grasping how the interaction model functions and how best to manipulate it. The framework is tested in three task-based and user-oriented comparative evaluations, which involves 12 comparative systems, 12 real life scenario tasks and 50 subjects. The quantitative data analysis shows encouraging observations on ease of use and usefulness of the proposed framework, and also reveals a large variance of the results depending on different user types.

Accordingly, based on Information Foraging Theory, we further propose a user classification model along three user interaction dimensions: information goals (I), search strategies (S) and evaluation thresholds (E) of users. To our best knowledge, this is the first principled user classification model in CBIR. The model is operated and verified by a systematic qualitative data analysis based on multi linear regression on the real user interaction data from comparative user evaluations. From final quantitative and qualitative data analysis based on the ISE model, we have established what different types of users like about the framework and their preferences for interactive CBIR systems. Our findings offer useful guidelines for interactive search system design, evaluation and analysis.

Declaration

The studies outlined in this thesis were undertaken in the Knowledge Media Institute (KMi), The Open University and supervised by Dr. Paul Mulholland, Prof. Stefan Rüger, Dr. Victoria Uren and Prof. Dawei Song. I declare that all of the work was performed by the author except where otherwise indicated, and this thesis has not been submitted at any other university.

Research work presented in some sections has been previously published, in particular: the dissimilarity measures comparison study (Section 1.1.2) has been published in (Liu et al. 2008), the proposed four-factor user interaction model (Chapter 2) has been published in (Liu et al. 2009), the design of the uInteract interface (Chapter 2) has been published in (Liu et al. 2009), the definition and verification of ISE model (Chapter 7) will be published in the proceeding of Information Interaction in Context Conference (IIiX2010) in August 2010. There are some further publications of the author which are referenced in the appropriate chapters.

Haiming Liu
Milton Keynes, UK, July 2010

Acknowledgements

I would like to express my profound gratitude to my supervisors: Dr. Paul Mulholland, Prof. Stefan Rüger, Dr. Victoria Uren and Prof. Dawei Song, for their valuable suggestions, encouragement, and supervision throughout my PhD research.

I am grateful to The Open University for granting me full studentship, and to Dr. Zdenek Zdrahal and Dr. Peter Scott for providing additional funding, so that I may complete my research.

Special thanks to: Prof. Ian Ruthven for his reading list which got my PhD started, and Prof. Paul Garthwaite for his statistics advisory service during my data analysis, and Dr. Jana Urban for the useful discussion and templates, and Srđan Zagorac for his help on developing the evaluation systems. Thanks are also made to all my colleagues in the Knowledge Media Institute (KMi), The Open University and in the Multimedia and Information Systems group, Department of Computing, Imperial College London for their support and help during my study.

Thanks are also made to many others, my friends and relatives and people I met at different occasions, such as, CRC PG forum, summer schools, conferences, workshops, meetings, talks, coffee and lunch breaks, etc., for sharing your valuable knowledge and experiences with me and contributing to my study in different ways.

I am as ever, especially indebted to my parents Mr Enping Liu and Ms Xiuchun Ren for their love and support throughout every aspects of my life.

Finally, and most importantly, thanks to my husband Paul Sellers for his everlasting love, support, encouragement, patience, understanding and constant motivation to keep me focused throughout my studies.

Thank you!

Haiming

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The overall aim of the thesis is to systematically explore the key elements for supporting and understanding user interactions in exploratory content-based image retrieval systems. The research will be built on the background of content-based image retrieval, the challenge of the semantic gap and the limitations of the current relevance feedback techniques. Particularly, this research is motivated by the importance of user interaction during the search process, especially the exploratory search process. In our opinion, user interaction involves three key elements: an interaction model, an interactive interface and users. Apart from investigating the user interactive model and the interactive interface, we also systematically analyze users based on Information Foraging Theory.

In this Chapter, we will firstly explain the background of our research from five aspects: content-based image retrieval (CBIR) (Section 1.1), the semantic gap (Section 1.2), relevance feedback (Section 1.3), user interaction (Section 1.4), exploratory search (Section 1.5) and Information Foraging Theory(Section 1.6). We will then highlight our research aim, objectives and questions generated from the background literature in Section 1.7. To address the research questions, our methodology and major contributions will be stated in Section 1.8. Finally, Section 1.9 outlines how the thesis is organized.

1.1 Content-based Image Retrieval

Current commercial image search engines retrieve images mainly based on their keyword annotations. This approach has two limitations. First, the manual annotation of images requires significant effort and thus may not be practical for large image collections. Second, as the complexity of the images increases, capturing image content by text alone becomes increasingly more difficult.

In seeking to overcome these limitations, content-based image retrieval (CBIR) was proposed in the early 1990's (Rui et al. 1998; Marques and Furht 2002). CBIR systems have since been primarily used for image searches on collections with limited annotations, or for image searches where annotation is not required, such as trademark searches (Eakins et al. 2003). To date, Google has launched a new application, called "Google Goggles" for Google Android mobile phones[1], which is a content-based search application and allows people to search for more information about a famous landmark or work of art simply by taking a photo of that object (Jamaal 2010).

A basic CBIR system should be able to interpret the content of the images in a query and a collection, match the similarity between the images in the query and the object images in the collection, rank the object images in the collection according to their degree of relevance to the user's query (Marques and Furht 2002). These key technical components of the CBIR system will be introduced in the following sections.

1.1.1 Visual Feature Extraction

CBIR systems index images by reference to the low-level features of the image itself, such as colour, texture and structure features (Pickering and Rüger 2003; Howarth and Rüger 2005b). The visual content of an image is then represented as a feature vector of floating numbers.

[1] $http://www.google.com/mobile/goggles/\#landmark$

RGB is a simple colour feature used to represent digital images. The colour of each pixel contains a different proportion of red, green and blue light. *HSV* is another colour feature. The hue coordinate H is angular and represents the colour, the saturation S represents the pureness of the colour and is the radial distance, finally the brightness V is the vertical distance.

One of the most popular signal processing based approaches for texture feature extraction has been the use of Gabor filters, a bank of filters at different scales and orientation information. This can be used to decompose the image into texture features.

Convolution (or Konvolution) was designed to find horizontal, vertical, and diagonal edges at good levels of filtering and different arrangements of these edges. This feature generation process is not only determined by the structure of the image, but also concerns colour and texture.

Findings from our pilot study show that the HSV perform the best among these features (Liu et al. 2008).

1.1.2 Dissimilarity Measures

After the low-level features are extracted, CBIR computes the dissimilarity between query images and object images in the collection based on their low-level content feature vectors. Dissimilarity is used to describe the how unlike each other two images or more images are in the high dimensional feature vectors, such as Euclidean distance, city-block metric, etc (Marques and Furht 2002).

Although there have been some attempts in theoretically summarizing existing dissimilarity measures (Chen and Chu 2005; Howarth and Rüger 2005a; Kokare et al. 2003; Noreault et al. 1980; Ojala et al. 1996; Puzicha 2001; Puzicha et al. 1997; Puzicha et al. 1999; Rubner et al. 2004; Zhang and Lu 2003), there is still a lack of systematic investigation into the applicability and performance of different dissimilarity measures in the CBIR field and the investigation into various dissimilarity measures on different features for large-scale CBIR.

In our pilot study (Liu et al. 2008) and (Hu et al. 2008), we have reviewed fourteen dissimilarity measures, and divided them into three categories: geometry, information theory and statistics, in terms of their theoretical characteristics and functionality. In addition, these dissimilarity measures have been empirically compared on six typical content-based image features, and their combinations on the standard Corel and ImageCLEF image collections. From the experimental results, we found the city-block dissimilarity measure is one of the best performing measures. Due to its computational simplicity, we recommend it for use in CBIR systems.

1.1.3 Fusion Approaches

For a single image query, the dissimilarity measure automatically produces a ranked list of results. When the query consists of several example images, however, we need to use a fusion approach to aggregate the results with respect to different example images to produce a ranked list. There are two well known adaptive approaches namely Vector Space Model (VSM) and K-nearest neighbours (KNN).

- The VSM has been widely used in text retrieval (Salton 1989). For an object image in the collection, the adapted VSM sums up all its dissimilarity scores to the images in a query. Following the recommendation from the literature (Pickering and Rüger 2003), in this thesis, we apply the adapted VSM to fuse the positive query images only. The retrieved similar images are ranked by the final dissimilarity scores:

  $$D_{VSM} = \sum_{i} (d_{ij}), \tag{1.1}$$

  where the D_{VSM} is the sum of the dissimilarity value d_{ij} between a query image i $1 \leq i \leq P$ and an object image j $1 \leq j \leq M$ (the query contains P positive examples and the collection contains M images).

- The original KNN approach was proposed by (Mitchell 1997). Our adapted approach is based on the idea that, given both positive and negative query images, the object images can be classified according to their proximity to these examples. To rank an object image in the collection we need to identify how the image is dissimilar with positive query images or negative query images. If the dissimilarity is lower with positive examples than negative examples, the image should be ranked higher, although it will depend on the true value (Howarth and Rüger 2005a). The dissimilarity measure is given by

  $$D_{KNN} = \frac{\sum_{i \in N} (d_{ij} + \varepsilon)^{-1}}{\sum_{i \in P} (d_{ij} + \varepsilon)^{-1} + \varepsilon}, \tag{1.2}$$

  where D_{KNN} is the dissimilarity value between an object image j with all the example images (positive and negative) in the query. d_{ij} is a dissimilarity value between a query i and an object image j. ε is a small positive number (e.g. 0.00001) to avoid division by zero. P and N denote the sets of positive and negative images in the query.

1.2 Semantic Gap

CBIR techniques are important when text annotations are nonexistent or incomplete (Smeulders et al. 2000; Eakins et al. 2003). However, CBIR systems index images by their low-level features such as colour, texture and structure, and such features do not necessarily mean anything to users. Further, the low-level feature based dissimilarity search algorithms are not as intuitive nor as user-friendly as they are expected to be. Thus there is a “gap” between the users and the systems (Lew et al. 2006; Zhou and Huang 2003; Crucianu et al. 2004). The problem is a well known challenge in the field of CBIR, called the “semantic gap”, which is a gap between the low-level features (RGB, for example) of an image and its high-level meaning (semantic) to a user (Marques and Furht 2002). Very often, when a user thinks some images in the search result are not relevant at all based on their semantic meaning, the computer may consider the images as relevant to what the user is looking for based on the dissimilarity values computed from the images' low-level features. A question then arises here, how can we close the gap and make the CBIR systems more effective?

1.3 Relevance Feedback

One way to bridge the semantic gap is to introduce relevance feedback to CBIR. Relevance feedback (RF) brings users into the search loop, so that users have the opportunity to provide feedback to help refine the query based on previous search results. The systems can then learn users' preferences from their feedback to improve the search performance (Crucianu et al. 2004; Zhou and Huang 2003; Marques and Furht 2002; Ruthven and Lalmas 2003; Rui et al. 1998).

There are two types of interactive feedback for learning users' preferences: explicit feedback and implicit feedback. The explicit feedback is given actively and consciously by the user to instruct the system what to do. Whereas, implicit feedback is inferred by the system from the way the user has interacted with the system. In other words, explicit feedback means the user is actively controlling the search process whilst implicit feedback means the system is controlling the search process. Many researchers suggest to use explicit or implicit or both to enhance search performance (Hopfgartner et al. 2007; White et al. 2006).

The relevance information learnt from users can be utilized by using query point shifting, query expansion or feature re-weighting, etc (Heesch and Rüger ; Heesch 2005). Query point shifting aims to move the query point towards positive examples by changing the query image. Query expansion is when additional example images are added to the query to better describe what the users are looking for. Feature re-weighting aims to give higher weights to the specific features of images that users prefer.

Currently most of the RF techniques only allow users to provide positive feedback. However, studies have shown that allowing users to provide both negative and positive feedback can improve the search performance (Müller et al. 2000; Heesch 2005). For example, Müller et al. (2000) compared a variety of strategies for positive and negative feedback. They employed an automated query expansion scheme to obtain negative judgements when users were not able to provide sufficient feedback. The experiment showed that negative feedback images improve search performance significantly. Specifically, a query from a user who initially uses positive feedback can only be improved by automatically supplying non-selected images as negative feedback. Heesch (2005) proposed a framework of relevance feedback for the K-nearest neighbours approach. They used query shifting and feature re-weighting techniques to recompute both relevant and irrelevant images by user's feedback. The framework does not only improve the retrieval performance, but also demonstrates a better user interaction.

These studies show the interactive CBIR systems with RF techniques can improve not only the search performance, but also the communication between the users and the system. To be able to learn more useful information from users and better engage the users during the search process, the interaction between the users and the system is vital (Zhou and Huang 2003; Urban et al. 2003; Urban 2007; Lew et al. 2006; Wegner 1997).

1.4 User Interaction

In our opinion, user interaction involves three key elements: the user interaction model, the interactive interface for delivering the user interaction model, and the users. The three elements combine to enable effective interaction to happen.

1.4.1 User Interaction Models

In an effort to make the RF mechanisms more interactive, some researchers have focused on developing user interaction models to formalize different factors for improving the interaction.

For example, Spink et al. (1998) proposed a three-dimensional spatial model to support user interactive search for text retrieval. The three dimensions are regions of relevance, levels of relevance and time.

Other studies have focused more on some individual dimensions of Spink et al.'s model, such as levels of relevance. Taylor et al. (2007) further showed the importance of the levels of relevance for information searching process.

Brini and Boughanem (2003) adapted the regions of relevance to their text retrieval system. Wu et al. (2004) and Cheng et al. (2008) applied the regions of relevance to image retrieval (Cheng et al. 2008; Wu et al. 2004). Their results showed the effectiveness of the method.

Campbell (2000) has focused on the time dimension and proposed an Ostensive Model (OM) that incorporates the degree of relevance relative to when a user selected the evidence from the results set. Later, Browne and Smeaton (2004) and Urban et al. (2006) applied the so called increasing profile of the OM to video IR and CBIR respectively (Urban et al. 2006; Browne and Smeaton 2004). Their studies showed that a system based on the OM outperforms, and is preferred by users, over traditional RF techniques in CBIR. Further, Fuhr (2008) suggested that the OM supports the dynamic nature of information needs.

Ruthven et al. (2003) adapted two dimensions from Spink et al.'s model combined with OM in their study. Their experimental results showed that combining partial relevance and time relevance did help the interaction between the user and the system.

More details on the user interaction models will be provided in Section 2.1.1.

1.4.2 User Interactive Interfaces

The user interaction models aim to provide an enhanced search experience in terms of the level of interaction between the system and users and search accuracy. However, without a visual search interface, we are not able to facilitate and test the interaction aspects.

When providing new search functionality, we should decide how the new functionality should be delivered to users (White and Ruthven 2006; Bates 1990). Some studies have focused on improving the interaction of the relevance feedback mechanisms on the visual interface. For instance, Flexible Image Retrieval Engine (FIRE) (Deselaers et al. 2005) is a tool that allows users to provide non-relevant feedback from the result set. Indeed, the research in (Heesch and Rüger 2003; Pickering and Rüger 2003; Müller et al. 2000) also suggested the importance of providing both negative and positive examples as feedback. Urban et al. (2006) developed an image search system based on the Ostensive Model. Later, Urban and Jose (2006a) presented another system - Effective Group Organization (EGO), which is a personalized image search and management tool that allows the user to search and group the results. Hopfgartner et al. (2007) investigated a video search system with explicit and implicit feedback.

More details on interactive interface design will be provided in Section 2.2.1.

1.4.3 User Evaluation

The performance of interactive CBIR systems is influenced by the users, tasks and systems (**?**; Järvelin 2009). Taking users into account, the system should be evaluated by users, not just by lab-based precision and recall measurement (Rijsbergen 1979; Baeza-Yates and Ribeiro-Neto 1999). Some researchers have applied different types of user-oriented evaluation design to their user studies as in (Ingwersen 1992; Borlund and Ingwersen 1997; Jose et al. 1998; White et al. 2005; Urban and Jose 2006b; Käki and Aula 2008).

Our user-oriented evaluation will apply simulated real world searching tasks, which allow the users to develop their own interpretation of the task, and use their own judgement for choosing relevant images as feedback and result, and discover different functionalities of the interface to support their search. Further, the evaluation data will also allow a systematic analysis of users, for identifying different user types and their search preferences and behaviours.

1.5 Exploratory Search

At this point, we need to introduce exploratory search, because we consider information seeking tasks, in particular interactive CBIR, will involve different levels of exploration depending on the users.

Exploratory search is recently emerging to support more user-centric information seeking and interactive search. It aims to shift the research focus from getting the highest precision (query-document matching) toward finding guidance at all stages of the information-seeking process to support a broader set of users' searching and interaction behaviours (White et al. 2007; White and Roth 2009).

Exploratory search is hard to define exactly, as almost all searches are somehow exploratory. However, one definition is that exploratory search is any search with a combination of a querying and a browsing strategy to enable learning and investigation (Marchionini 2006; White et al. 2006; White et al. 2007; Mulholland et al. 2008; Marchionini and White 2009). Marchionini (2006) emphasizes that the traditional “lookup” search on its own is not exploratory search; however, exploratory search contains the lookup stage.

Further, White et al. (2006) presented a few more definitions of exploratory search from the users’ perspective:

- An exploratory search can happen when the presence of the search technology and information objects is meaningful to users.
- An exploratory search is motivated by a complex information problem that users are not looking for a single answer to.
- An exploratory search can happen when the users are unfamiliar with the domain of the task.
- An exploratory search is required when the users are uncertain of the ways to achieve their goal by the technology.
- An exploratory search is needed when the users are unsure of their goals in the first place.
- An exploratory search will happen if the users have a lack of knowledge of the data they are searching from.

To support exploratory search, we need to consider how the information is found; how the information is presented; how the information needs are described; how the information is used by users; how the information seeking behaviour is defined by analyzing the exploratory data, and how the exploratory search is evaluated.

White et al. (2006) suggested that exploratory search is related to Information Foraging Theory (Pirolli and Card 1995; Pirolli and Card 1999) in the aspect of finding an optimal path to reach users' information goal during search. For instance, how users search for information based on their information goal, how users apply their searching strategy, and how users decide what information to use, etc. Indeed, Mulholland et al. (2008) have shown that Information Foraging theory can interpret the effects of the exploratory search technologies. They identified two distinct strategies of exploratory search, namely risky search strategy and cautious search strategy. Their findings can be considered as a step forward in supporting exploratory search.

1.6 Information Foraging Theory

User contexts can be very different when different users use search systems. Some people know what they want, and some people only know when they find it (ter Hofstede et al. 1996). Some people are patient, but some are not. Some people frequently change their mind on what they are looking for, but some do not. Some people like to use both query by example search model and browsing search model to retrieve their idea. Some people are satisfied with the result they get after a few rounds, but some are not (Urban et al. 2003).

Information Foraging Theory suggests that the way humans seek information is not unlike the way of wild animals gather food (Pirolli and Card 1995; Pirolli and Card 1999; Pirolli 2007). First they will find a path to food resource (scents); next they will select what to eat (diet); and then they have to decide when to hunt elsewhere (patch) (Nielsen 2003; Stephens and Krebs 1986).

To adapt the food hunting behaviour to humans online information seeking, the interpretation will be: foragers will find an information patch that they think would bring the outcome they desire based on their information scents; the foragers then will decide which information resource they will select based on their information diet; the foragers also need to decide how long they will stay with this information patch and when to go to a different patch of information. To decide which information resource is the start point and when to move elsewhere, the foragers need to consider the cost and benefit trade-offs. Different foragers will make different decisions on these stages based on the different search preferences and behaviours. Thus, we are motivated to identify different types of foragers and find out different types of preferences and behaviours they have during the information seeking process.

In this thesis, we will adapt Information Foraging Theory to the CBIR scenario: the information patch in our case will be a set of result images from the initial search; the information scent will be the clues that users get from task descriptions, query images, result images and past search experience to formulate their information goal and navigate their search process; the information diet will be the way that users select the feedback and result images.

More details on Information Foraging Theory will be provided in Section 7.1.

1.7 Research Questions

Motivated and inspired by the literature, the main aim of the thesis is to systematically explore the three key elements of user interactions in exploratory CBIR systems, including the user interaction model, the interactive interface and users. The three elements combine to enable effective interaction to happen. However, there is lack of research in understanding them in an integrated manner. In this thesis, we investigate the combination of all three elements for interactive CBIR.

Specifically, the objectives of the research are:

- to propose a novel user interaction model for content-based image retrieval;
- to deliver the model by a visual interactive interface that allows users to effectively manipulate the model;

- to evaluate the effects of the interactive model through both simulated experiments and user evaluation;
- to evaluate the effects of the visual interface through user evaluation;
- to propose a well-principled user classification model based on Information Foraging Theory to identify different user types;
- to verify and operationalize the user classification model based on extensive real user interaction data collected from the user evaluations;
- to use the user classification model to better understand user interactions in CBIR and find user preferences and behaviours on interactive CBIR system design based on different user types.

The objectives lead to the following research questions:

- Q1: What factors in the user interaction model are important to the interaction between the users and interactive CBIR systems? (addressed in Chapter 2)
  Q1.1: What effects do the four profiles of the Ostensive Model have on the users? (addressed in Chapter 5)
  Q1.2: What effects do the factors of the four-factor user interaction model have on the users? (addressed in Chapter 6)
  Q1.3: What effects do the users have on the preferences for the user interaction models? (addressed in Chapter 5 and Chapter 6)
- Q2: How can the users best interact with the system through a visual interactive interface? (addressed in Chapter 2)
  Q2.1: What effects do the visual interfaces have on the users? (addressed in Chapter 4, Chapter 5 and Chapter 6)
  Q2.2: What effects do the users have on the preferences for the interfaces? (addressed in Chapter 4, Chapter 5 and Chapter 6)

- Q3: What are the preferences for the different user types on the interactive CBIR framework design? (addressed in Chapter 8)

  Q3.1: How many user types are there? (addressed in Chapter 7)

  Q3.2: Will the search precision be different based on the different user types? (addressed in Chapter 8)

  Q3.3: What preferences do the different user types have on the interactive interfaces? (addressed in Chapter 8)

  Q3.4: What preferences do the different user types have on the user interaction models? (addressed in Chapter 8)

  Q3.5: What are the comments of different user types to the uInteract framework? (addressed in Chapter 8)

1.8 Contributions

The research begins with a comprehensive literature review. The related work motivates and inspires us to propose a new interactive CBIR framework - uInteract. The framework includes a four-factor user interaction model and a visual interactive interface to deliver the model and allow users to manipulate the model. On top of lab-based simulated experiments, a series of task-based user evaluations with real image search scenarios are carried out to evaluate the effects of the user interaction model and the visual interface. The user evaluations also provide a large amount of quantitative and qualitative data including the search results, user comments and interactions with the systems. An extensive quantitative result analysis shows the effects of the user interaction model and the interactive interface to the users as well as the users' impact on the preferences to the uInteract framework. A novel user classification model, called ISE (I: information goals; S: search strategies; E: evaluation thresholds) is proposed based on Information Foraging Theory for better understanding the user interactions in CBIR. The model is further verified by an in-depth analysis of the user interaction data gathered from our user evaluations.

Applying the ISE model to the quantitative and qualitative data analysis from our user study allows us to explore the users' preferences for the uInteract framework in relation to user types. The final findings provide valuable insights on the interactive CBIR framework design, evaluation and analysis based on different user types.

To the best of our knowledge, this is the first systematic and principled investigation in exploring the three key elements of user interactions, i.e., interaction model, interface and users, in a integrated manner.

1.9 Organization of the Thesis

The remainder of the thesis is organized as follows:

- In Chapter 2, we will present a new exploratory CBIR framework - uInteract. We will first review the related work in user interaction models, and then propose a new four-factor user interaction model. Results from a lab-based simulated evaluation will show the effectiveness of the proposed interaction model with both multi-image positive and negative queries for CBIR. Secondly, we will review the existing interactive interfaces for CBIR and then present our visual interactive interface for uInteract for delivering the four-factor user interaction model.

- In Chapter 3, we will describe in general our task-based and user-oriented evaluation methodology. Particularly, three user evaluations are carried out to evaluate the uInteract interface (detailed experimental setup and results are reported in Chapter 4), the Ostensive Model (detailed experimental setup and results are reported in Chapter 5) and the four-factor user interaction model (detailed experimental setup and results are reported in Chapter 6) respectively.

- In Chapter 7, we will review Information Foraging Theory that motivates us to propose an ISE user classification model. The ISE model will be verified by qualitative data analysis.

- In Chapter 8, we will apply the ISE user classification model to the quantitative and qualitative data obtained in our user evaluations. The quantitative and qualitative data analysis results will show users' preferences on interactive CBIR framework design based on different user types.
- In Chapter 9, we will conclude the thesis by reviewing the contributions and suggesting future work;
- In Appendix A, we will report the task descriptions and questionnaires used for evaluating the uInteract interface (E1).
- In Appendix B, we will report the task descriptions and questionnaires used for evaluating the Ostensive Model (E2) and the different settings of the four-factor user interaction model (E3).

Chapter 2

uInteract Framework: Four-factor User Interaction Model and Interactive Interface

In Chapter 1 we reviewed the research background, addressed the research questions and stated the main contributions of the thesis. From the literature we have learnt that user interaction is essential to the communication between users and content-based image retrieval (CBIR) systems. In our opinion, user interaction covers three key elements: an interaction model, an interactive interface and users. The three elements combine to enable effective interaction to happen. Many studies have investigated different aspects of user interaction. However, there is lack of research in combining all three elements in an integrated manner.

In this Chapter, we will introduce a novel interactive CBIR framework - uInteract, which aims to tackle the first two elements[1]: the user interaction model and the interactive interface. In Section 2.1.2, we will first propose an adaptive four-factor user interaction model (FFUIM) based on the literature review, and then we will investigate the performance of the FFUIM through simulated evaluations on a large

[1] The last element of the user interaction - users - will be tackled in Chapter 7 and Chapter 8.

image collection. In section 2.2.2, a visual interactive interface is designed based on existing design guidelines to deliver the FFUIM visually and aims to provide a user-oriented search platform.

2.1 User Interaction Models

To improve user interaction between a CBIR system and users, we firstly need to have a good user interaction model, which should not only support the interaction but also enhance search performance.

2.1.1 Literature Review on User Interaction Models

In this section, we review a number of existing user interaction models and describe how our FFUIM harnesses their advantages, whilst addressing some of their limitations.

Three-dimensional Spatial Model

In order to improve the interaction between the users and the system, Spink et al. (1998) proposed a three-dimensional spatial model, consisting of levels of relevance, regions of relevance and time dimension of relevance, for text retrieval. They firstly applied Saracevic's five levels of relevance (Saracevic 1996) to indicate why the feedback is relevant, which includes system's or algorithmic relevance, topical or subject relevance, cognitive relevance or pertinence, situational relevance or utility, motivational or affective relevance. Second, the regions of relevance indicate the degree of users' relevance judgements to a feedback. The four regions are: relevant, partially relevant, partially not relevant and not relevant. Third, they proposed a time dimension in their framework, because they found that humans seek information on a particular information problem in stages over time. The time of relevance is measured in formats such as information seeking stage and successive searches. We consider the three-dimensional spatial model as a useful starting point to develop a more advanced user interaction model and techniques.

Other existing research has focused more on a single dimension, such as levels of relevance. Taylor et al. (2007) further showed the importance of the levels of relevance for the information searching process. Their results show that relevance is multi-context and dynamic. Moreover, they also suggested that non-binary relevance assessment is important within every context.

Brini and Boughanem (2003) adapted another dimension - regions of relevance from Spink et al. (1998)'s model to their text retrieval system. They considered that partial relevance is close to human reasoning. Their experimental result showed that the partial relevance feedback approach outperformed the binary relevance feedback approach. Wu et al. (2004) and Cheng et al. (2008) applied the regions of relevance to their relevance feedback mechanism for image retrieval. The multi-level relevance measurement was utilized by query expansion and feature re-weighting according to relevance level of query images indicated by the user (Cheng et al. 2008; Wu et al. 2004).

Ostensive Model

Campbell (2000) has focused on the time dimension. He proposed the Ostensive Model (OM) that indicates the degree of relevance relative to when a user selected the evidence from the results set. The OM includes four ostensive relevance profiles: decreasing, increasing, flat and current profiles, respectively. With the increasing profile the latest feedback is deemed the most important, whereas with the decreasing profile it is the earliest feedback that is regarded as the most important. With the flat profile all feedback is given equal importance, regardless of when the feedback was provided. Finally, the current profile gives the latest feedback the highest weight and earlier feedback is ignored. Campbell found that for text retrieval the increasing, flat and current profile showed overall better accuracy than the decreasing model, and the increasing profile was the most robust (Campbell 2000). Fuhr (2008) suggested that the OM supports the dynamic nature of information needs.

Browne and Smeaton (2004) and Urban et al. (2006) adapted the OM from text retrieval for image and video retrieval to help overcome interaction problems between users and multimedia search systems (Urban et al. 2006; Browne and Smeaton 2004). In their studies, only the increasing profile was applied. The results indicated that, whilst users found the OM easy to use, they found it difficult to control the RF process without greater interaction. Furthermore, the traditional OM accepted only positive feedback, whereas in reality users may wish to refine their searches by providing both negative and positive feedback. Indeed, some research (Dunlop 1997; Pickering and Rüger 2003; Müller et al. 2000) has shown that including negative examples into the RF can actually help improve the image retrieval accuracy. Therefore, we are motivated to test the performance of the four profiles of the OM on the multi-image query and both negative and positive feedback search scenarios.

Partial and Ostensive Evidence

Ruthven et al. adapted and combined two dimensions from Spink et al. (1998)'s three-dimensional spatial model, namely: regions of relevance and time, for ranking query expansion terms in text-based information retrieval (Ruthven et al. 2003; Ruthven et al. 2002). The region of relevance in their study is called partial evidence, which is a range of relevance level from one to ten, which is different from Spink et al. (1998)'s definition. In addition, they applied the OM to the time dimension, which is called ostensive evidence. The ostensive evidence is measured by iterations of feedback. Their study shows that combining RF techniques with the user interaction factors is preferred by users over RF techniques alone. However, to our knowledge, neither the combined model nor the three-dimensional spatial model have previously been applied to CBIR. It will be interesting to see how the combined model performs in our CBIR system.

2.1.2 A Four-factor User Interaction Model (FFUIM)

Based on these interesting studies, we propose a new model named 'four-factor user interaction model (FFUIM)', which combines the three-dimensional spatial model with the OM and, further, adds another factor - frequency. The FFUIM includes: relevance region, relevance level, time and frequency. We introduce the four factors in the following sections.

Relevance Region

Instead of Spink et al. (1998)'s four regions of relevance, the relevance region here comprises two parts: relevant (positive) evidence and non-relevant (negative) evidence. Both relevance regions contain a range of relevance levels.

Relevance Level

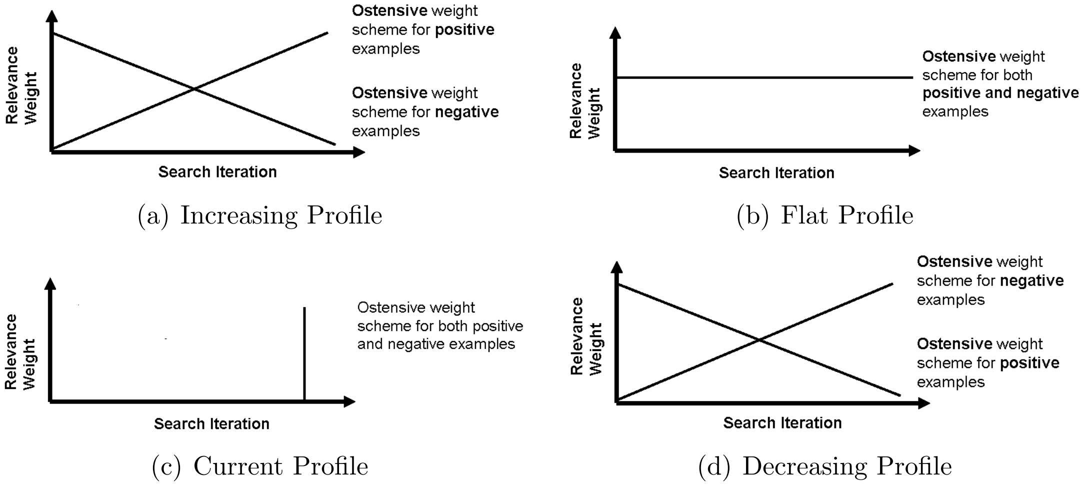
The relevance level here indicates how relevant/non-relevant the evidence is on the related relevance region, which implies a quantitative difference, and differs from Saracevic's definition used in Spink et al. (1998). This factor is measured by a range of relevance levels (integers 1-20) indicated by users. The distance function with the relevance level factor is given by

$$D_{ij} = d_{ij}/W_p, \tag{2.1}$$

where $D_{ij}(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ is the final distance between a query image i and an object image j; d_{ij} is the original distance between the query image i and an object image j; W_p is the partial weight, $W_p = r$ for the positive examples, and

$W_p = \frac{1}{r}$ for the negative examples (r is the level of the relevance provided by the user, an integer between 1 and 20)[2].

Time



(a) Increasing Profile

(b) Flat Profile

(c) Current Profile

(d) Decreasing Profile

Figure 2.1: Four Profiles of the Ostensive Model (time factor)

We adapted the OM to the time factor to indicate the degree of relevance relative to when the evidence was selected. In this study, we have taken the OM a step further. In addition to using the increasing profile, we have also tested the flat profile, current profile and the decreasing profile (Figure 2.1). For our study, the increasing / decreasing profile means ostensive relevance weights for positive / negative examples increase / decrease respectively with further search iterations. The fundamental difference between our studies and Urban et al. (2006)'s study is that we have applied these ostensive relevance weights to both the positive and negative feedback, and applied the weight to more than one image in every query. We propose the following distance function with the ostensive weight:

$$D_{ij} = d_{ij}/W_o, \tag{2.2}$$

[2] Note that we have tested a number of other weighting functions for W_y (y can be o,p,f), e.g., $W_y = x$, $W_y = 2^x$ and $W_y = \ln(x)$ (x can be r,s,t) for positive examples, but there was no significant difference in performance (MAP). Here we use the linear setting for simplicity.

where W_o, the ostensive weight, can be different depending on the profile. $W_o = s$ for the positive examples, and $W_o = \frac{1}{s}$ for the negative examples (for the increasing profile, s is iterations of feedback; for the decreasing profile, s the i-th iteration of feedback in the contrary order; for the flat profile, s is 1; for the current profile, s is 1 for the current iteration, but 0 for the previous iterations)[3].

Frequency

While we are investigating the combined models, we find that the same images can be used as positive/negative examples in different feedback iterations. Thus, we wonder: can the number of times (frequency) an image appears across all the iterations contribute to the model? To answer this question, we propose a new factor - frequency, which captures the number of appearances of an image in the user selected positive and negative evidence separately. The distance function with frequency is given by

$$D_{ij} = d_{ij}/W_f, \tag{2.3}$$

where W_f, the frequency weight, is how often an image has been chosen as a relevant or non-relevant example: $W_f = t$ for the positive examples, and $W_f = \frac{1}{t}$ for the negative examples (t is the number of times the image was chosen as feedback)[4].

2.1.3 Simulated Evaluation

Our empirical experiments aim to find possible interaction settings of the FFUIM that improve the search accuracy in comparison with a CBIR system without any interaction. The evaluation was a lab-based systematic comparison. We tested some individual and combined factors of the FFUIM. The performance indicator used was

[3] Please see more detail in footnote 2.

[4] Please see more detail in footnote 2.

Mean Average Precision (MAP), and we used the ranking of images in the entire data set to compute the MAP for each experiment.

Experimental Setup

The ImageCLEFphoto2007 collection (Grubinger et al. 2006) was used, which consists of 20,000 real life images and 60 query topics. We applied colour feature HSV to all of the images. The city-block distance (a special case of the Minkowski distance family) as suggested in our pilot study (Liu et al. 2008; Hu et al. 2008) was used to compute the distance between query images and object images.

Two Fusion Approaches

Heesch et al. (2003) investigated three fusion approaches, namely: Support Vector Machine (SVM), adapted Vector Space Model (VSM) and adapted K-nearest neighbours (KNN), with multi-image queries on interactive CBIR systems for the pseudo feedback setting. Their experimental result suggested that KNN fusion approach is the best for interactive CBIR systems with both positive and negative multi-image queries.

We used two fusion approaches to support two different feedback scenarios. Firstly, the Vector Space Model (VSM) (Pickering and Rüger 2003) is deployed for positive relevance feedback only. By adding the weighting scheme of the FFUIM into the VSM, the approach is represented by:

$$D_{VSM} = \sum_i (d_{ij}/W_z), \tag{2.4}$$

where the D_{VSM} is the sum of the distance value between a query (containing i positive examples) and an object image j. W_z can be one of the three factors' weight W_o,W_p,W_f, or any combination weight[5] of all three factors, depending upon

[5] Multiplication is employed to combine the weight together.

which factor or combined factors is/are being tested.

Secondly, because the VSM in (Pickering and Rüger 2003) only uses positive feedback, we apply k-nearest neighbours (k-NN) when both positive and negative feedback are used (Pickering and Rüger 2003). Here, by taking into account the weighting scheme, k-NN is given by:

$$D_{KNN} = \frac{\sum_{i\in N}(d_{ij}/W_z + \varepsilon)^{-1}}{\sum_{i\in P}(d_{ij}/W_z + \varepsilon)^{-1} + \varepsilon}, \tag{2.5}$$

where D_{KNN} is the distance value between an object image j with all the example images (positive and negative) in the query. ε is a small positive number (e.g. 0.00001) to avoid division by zero. N and P denote the sets of positive and negative images in the query.

Two Interaction Approaches

Our experiments used two interaction approaches: pseudo feedback and a method we call simulated user feedback.

Firstly, pseudo feedback was applied - a method widely used in information retrieval. Here there is no user interaction functionality with the feedback approach. The system automatically takes the top and bottom three images from the ranked last iteration search result of each query as positive and negative examples, respectively, to expand the current queries. The reason we take the bottom three images as negative feedback to expand the current queries is because, from our previous experiment, this approach outperforms the use of randomly chosen negative examples.

Secondly, a so-called simulated user feedback was used. This approach uses three truly relevant images from the top ranked results of each query and three truly non-relevant images from the bottom as tested against the official relevance judgments file. We derive this method to provide an automatic means of feedback which is closer to real user behaviour. The reason we limit feedback to three positive images and three negative ones is because we want to make the experimental results more comparable with equal numbers of image examples in the queries.

For consistency of approach, we used three image examples in each original query and each of the feedback iterations. Further, we limited the number of iterations to three, where iteration one is the search by original queries without feedback, and iterations two and three are with feedback. The time and relevance region factors are applied to all the queries on each iteration, whilst the relevance level and frequency factor is applied only to the latest iteration.

Experimental Results

Our experiment tested the performance of 16 interaction settings of the FFUIM, which includes four profiles of OM (time factor): flat profile, increasing profile, current profile, decreasing profile respectively, and the four profiles combined with the relevance level factor respectively, and the four profiles combined with the frequency factor respectively, and the four profiles combined with the relevance level and the frequency factor respectively. Each of the 16 settings were tested using positive feedback only as well as both positive and negative feedback (relevance region factor). The models have been tested against a large image collection and two interaction approaches as previously described. The following insights have been observed by doing statistical significance tests (the Wilcoxon signed ranks test with $\alpha = 0.05$):

Firstly, simulated user feedback has a better performance than pseudo feedback. Secondly, with the pseudo feedback approach, accuracy falls with increasing iterations. Thirdly, under simulated user feedback approach, the performance clearly improves with each search iteration for all the results.

Apart from these generic insights, other results vary depending on the different settings and iterations. Since iteration three is the last iteration in our experiment and the weights should show more effect on the results, and, in addition, the simulated

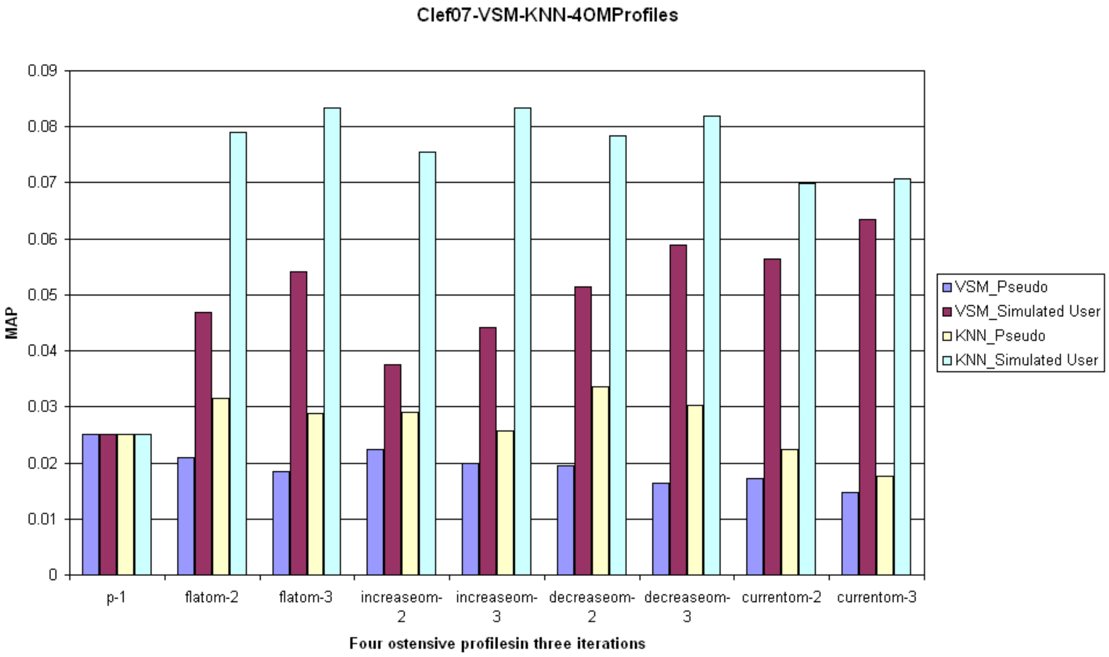


Figure 2.2: Effects of the relevance region and time factor

user feedback outperforms pseudo feedback and is closer to the real search scenario, we have undertaken further detailed analysis of the simulated feedback at iteration three based on different search settings as follows:

Effects of using the positive examples only and both positive and negative examples (relevance region factor). Figure 2.2 shows that the use of both positive and negative example feedback with the k-NN approach performs significantly better than the positive example only feedback with VSM approach. The promising result encourages us to include the negative functionalities to our future visual search system, and then we need to think about how to deliver these functionalities to users through the interface.

Effects of the four profiles of the Ostensive Model (time factor). Figure 2.2 shows that under the positive feedback only setting, the decreasing and current profiles show consistently good performance, and the flat profile outperforms the increasing profile in most tests. Under both the positive and negative feedback settings, the decreasing, flat and increasing profiles are not significantly different, but the current profile shows statistically worse performance than the other three profiles. The results do not show the same observation as previous OM studies, namely that the latest feedback expresses best the user's information needs. This may be because the relevance judgement file was developed against the original query that is the oldest feedback iteration. Thus the decreasing profile performs consistently well in different circumstances. These models need further testing in a real, as opposed to a simulated, CBIR search environment.

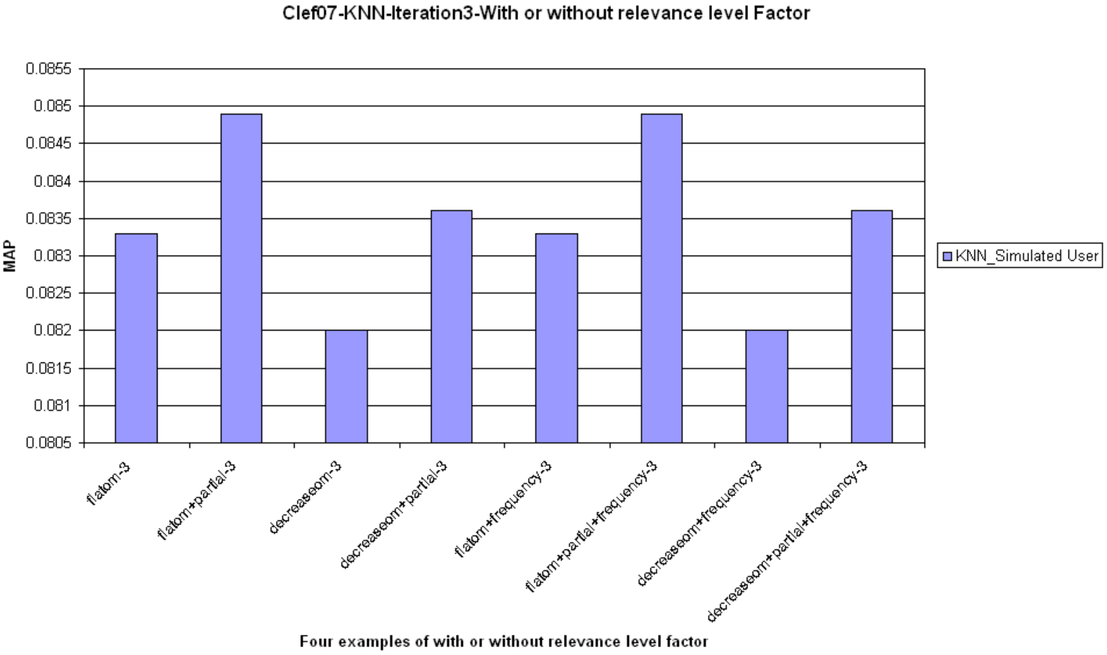


Figure 2.3: Effects of the relevance level factor

Effects of relevance level factor. Figure 2.3 shows that in all of the tests, the relevance level when combined with the OM is not significantly different to the OM alone. This factor also needs further testing under a real user as opposed to simulated user evaluation.

Effects of frequency factor. The frequency factor when combined with the other factors does not lead to significantly better performance than the factors without frequency factor. This may be because the limited number of search iterations means that the frequency weight has little impact. This result may be clearer when we run further iterations of the experiment, or even under a real as opposed to simulated user evaluation.

2.2 Interactive interface

After proposing the four-factor user interaction model, we need an interactive interface to deliver the model to users visually, so that the users will be able to manipulate the model through the interface.

2.2.1 Literature Review on Interactive Interfaces for CBIR

One reason that CBIR is not yet widely applied is that most existing CBIR systems are designed principally for evaluating search accuracy. Less attention has been paid to designing interactive visual systems that support users in grasping how feedback algorithms work and how they can be manipulated.

To improve the usability of CBIR systems and to make the CBIR system more human-centric, the system should deliver a user-oriented search making the user feel that they, rather than the system, are driving the search process. Bates (1990) addressed two issues for search system design: "*(1) the degree of user vs. system involvement in the search, and (2) the size, or chunking, of activities; that is, how much and what type of activity the user should be able to direct the system to do at once.*"

To investigate the first issue, we have developed an interactive relevance feedback (RF) mechanism named four-factor user interaction model in Section 2.1.2, which aims to improve the interaction between users and the content-based image retrieval (CBIR) system and in turn users' overall search experience. According to the results of our simulated experiments, the model can improve the search accuracy in some circumstances. However, we are not able to carry out user evaluation on the ease of use and usefulness of the interactive functionalities without an interactive visual search interface.

In terms of the second issue, White and Ruthven (2006) has also stated "*When pro-*

viding new search functionality, system designers must decide how the new functionality should be offered to users. One major choice is between (a) offering automatic features that require little human input but give little human control; or (b) interactive features which allow human control over how the feature is used, but often give little guidance over how the feature should be best used." One question arises here for our study: How should the functionalities be presented visually to the user by the interface to enable users to directly control the model in an effective way?

A user interface for an Information Retrieval (IR) system normally includes two parts: a query formulation part and a result presentation part (Marques and Furht 2002). Here we will review the related work and explain our motivation for resolving the interaction issue (the degree of search control deployed to users and system) and the design issue (the best way to deliver the framework functionalities to users through interface) for the two parts.

When providing new search functionality, we should decide how the new functionality should be delivered to users (White and Ruthven 2006; Bates 1990). In this section we investigate a number of search interfaces in order to explain why we have developed the search interface in the way we did.

Flexible Image Retrieval Engine (FIRE) (Deselaers et al. 2005) is one tool that allows users to provide non-relevant feedback from the result set. The research in (Heesch and Rüger 2003; Pickering and Rüger 2003; Müller et al. 2000) also usefully referred to the importance of providing both negative and positive examples as feedback. In addition, from the results of our simulated experiments, we found that limiting user's selection of non-relevant feedback to the poorest matches in the results list will improve search accuracy, but we realized this is not going to be intuitive to users. Therefore, we are encouraged to design the system to enable users to provide the negative examples from the worst matches in a natural way.

Urban et al. (2006) developed an image search system based on the Ostensive Model. Like FIRE, this is a browsing based search system, which uses a dynamic tree view to display the query path and results, thus enabling users to re-use their previous queries at a later stage. Whilst the query path functionality is useful, the user display becomes overly crowded even after a relatively small number of iterations. This limitation would become even more evident were the system to allow the user to provide negative as well as positive examples. Why not then harness the benefits of the query path functionality but in a search-based system, which separates query and results and applies the linear display to both queries and results?

Later, Urban and Jose (2006a) presented another system—Effective Group Organization (EGO), which is a personalized image search and management tool that allows the user to search and group the results. The user's groupings are then used to influence the outcome of the results of the next search iteration. This system supports long-term user and search activity by capturing the user's personalized grouping history, allowing users to break and re-commence later without the need to re-create their search groupings from scratch. From this study, we can see that providing a personalized user search history can improve the interaction between the system and users.

Hopfgartner et al. (2007) applied explicit and implicit feedback to a video retrieval system. Their simulated user study results showed that combining implicit RF with explicit RF may provide better search results than explicit RF by itself. We are then encouraged to combine the implicit and explicit RF in our system.

In the following sections, we will present our proposed uInteract interface, which will implement the ideas we have developed to overcome the shortcomings of the related work. Table 2.1 shows how the related work maps to the features of the uInteract interface (note that in this thesis we only compare the CBIR features and ignore the textual search features). Moreover, we will describe how we developed the interface to deliver our four-factor user interaction model.

Feature	Deselaers et al.	Urban et al.	Urban et al.(EGO)	Hopfgartner et al.	uInteract
Search-based system	No	No	Yes	No	Yes
Providing positive feedback	Yes	Yes	Yes	Yes	Yes
Providing negative feedback	Yes	No	No	Yes	Yes
Range of (non)relevant level	No	No	No	No	Yes
Query history functionality	No	Yes	No	No	Yes
Showing negative result	No	No	No	No	Yes

Table 2.1: How the related work maps to the features of uInteract

2.2.2 The uInteract Interface

In our view, an appropriate interface is vital to allow our new interaction CBIR framework to fully function because the interface is the communication platform between the system and user. We will outline our developed interface and describe how it underpins the four-factor interaction model.

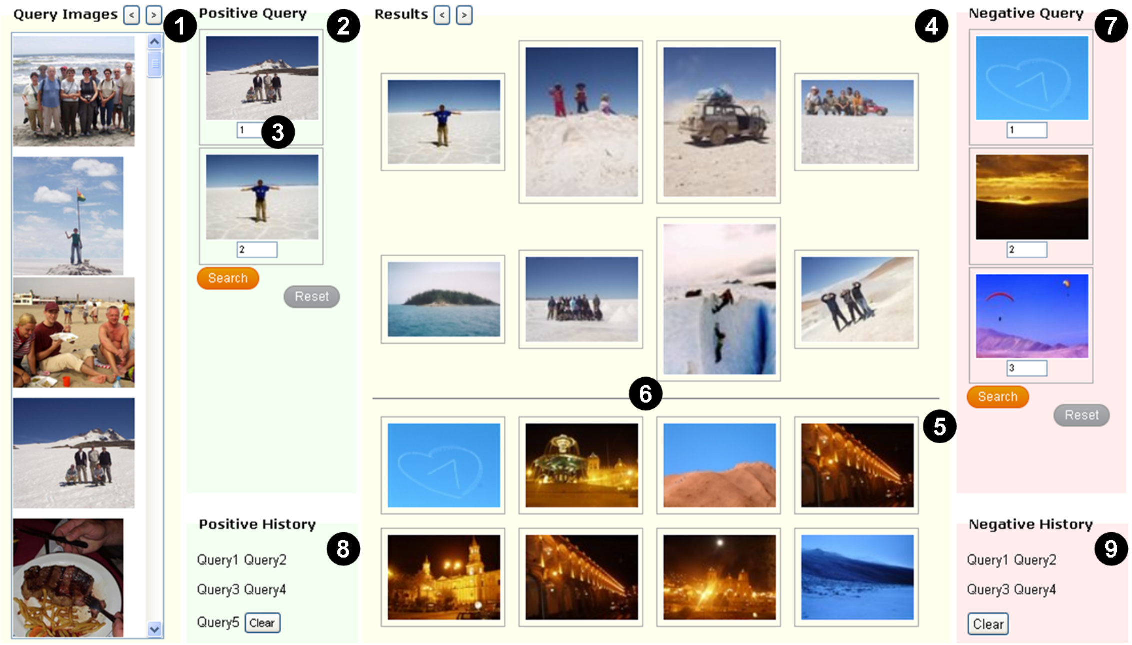


Figure 2.4: The uInteract interface. Key: [1] The browsing based query images where the initial query is selected; the initial query images go into [2] as a positive query to start the search; users can score (integer 1-20, bigger is better) the selected images in [3] with their preferences; [4] and [5] the search result shows the best matches and worst matches to the query respectively; [6] a horizontal line divides the two parts of the results visually; [7] negative query examples that users selected from previous results; [8] positive query history records the positive queries that were used previously; [9] negative query history records the negative queries from the previous search.

The search interface (Figure 2.4) takes on a simple search-based grid style so that the user does not need to learn the new visual layout before they start a search. Different colour backgrounds have been applied to the different panels which is aimed at supporting user navigation and appreciation of the differences between the panels. Each panel provides a different level of interaction to the user, where some of the four factors are controlled indirectly and others more directly. Table 2.2 shows how the interface supports each of the four factors (note: the numbers on the table indicate the functionalities on the screen shot). The rest of this section describes the features of those panels.

Factor	**Functionality**
Relevance region	Positive and negative feedback in [2] and [7]
Relevance level	score in [3]
Time	Positive and negative query history in [8] and [9]
Frequency	Positive and negative query history in [8] and [9]

Table 2.2: Which parts of the interface support the four-factor user interaction model

Query Image Browsing Panel (Region 1)

The query image panel is a browsing panel. The user browses the query panel and selects one or more images from the provided query images as an initial query image(s) prior to starting the search.

Positive Query Panel (Region 2)

The positive query panel contains images that the user considers are good positive examples of what they are searching for. Users can provide as many images as they want as positive queries. These images can be selected from the query images, the search results or a combination of both. Users are also able to eliminate positive examples by simply clicking on them.

After the user selects positive images, the system automatically gives their importance score by their display order. If the user is not happy with the default score, he can re-score the importance of the images by changing the number (integer 1-20, bigger is better) in the text box underneath each image. This functionality delivers the ‘relevance level’ factor. The intention of the design is to provide users an explicit control of the importance level of the query image examples.

Negative Query Panel (Region 7)

The negative query panel has similar functionality to the positive query panel but this time for negative queries. The only difference is that negative examples may only be selected from the previous search results. The score of these negative example images indicates the level of non-relevance (integer 1-20, bigger is worse).

In summary, both the positive and negative query panels deliver the ‘relevance region’ factor, such as relevant and non-relevant region. The score of image examples in both panels indicates the ‘relevance level’ factor—a scale of relevance and non-relevance. Combining the findings in (Spink et al. 1998; Ruthven et al. 2003) and (Pickering and Rüger 2003; Müller et al. 2000), our hypothesis is that blending the non-binary relevance level with both positive and negative regions will enhance user interaction on the one hand and increase search accuracy on the other.

Results Panel (Regions 4 and 5)

Whereas a common linear display search system may display only the best matching results, our system displays both the best and poorest matches. In our view, this added functionality allows users to gain a better understanding of the data set they are searching. By seeing both good and bad results, the user can gain a better understanding of the data they are searching. Additionally, for experienced users, the extreme results can aid their special search purposes, for instance, when a user searches for two extremely different colour images, say one pink and one blue.

Furthermore, users can indicate positive examples from the good matches and negative examples from the poorest matches by selecting them with a single mouse click. The selected images will appear automatically in either the positive or negative query panels. According to our simulated experimental results, taking the worst matches as negative query examples outperforms the query example from good matches. Therefore, we designed the interface to support the search mechanism by showing the poorest as well as the best matches. Users will need some training in the way that the interface works. We assume that the users will be able to search naturally after a couple of search iterations although this functionality is not intuitive to start with.

To aid navigation, we have inserted a horizontal line between the good and bad results to clearly divide the two.

Positive History Panel (Region 8)

This is an important feature of our search system. This panel records the user's earlier positive queries used during previous search iterations. This enables the user to go back and reuse a previous query if required. This might be needed, for instance, if the user got lost during the search process.

In addition, this panel delivers two important factors to our four-factor user interaction model: Firstly, the 'time' factor which is computed by the Ostensive Model and takes a search iteration as a time unit. Secondly, the 'frequency' factor that judges the importance of an image by reference to how many time the image was used as a query.

These two factors are fully controlled by the system, and all previous queries will be taken into account in the final weighting scheme.

Negative History Panel (Region 9)

This panel is similar to the positive history panel but instead records the negative queries selected from each search iteration. The negative query history is introduced together with the negative query as two of the new features of our search interface. The introduction of query history functionality has been encouraged (Campbell 2000; Urban et al. 2006; Chen et al. 2000) and we would like to investigate the effects on user interaction and search accuracy by adding the negative factor.

Summary of the uInteract interface

In summary, the key features of the proposed interface are:

(1) Users can provide both positive and negative examples to a search query, and further expand or reformulate the query. This is a way to deliver the 'relevance region' factor.

(2) By allowing the user to override the automatically generated score of positive and negative query images, we are enabling the user to directly influence the importance level of the feedback. The 'relevance level' factor is generated by the score functionality.

(3) The display of the results in the interface takes a search-based linear display format but with the addition of showing not only the best matches but also the worst matches. This functionality aims to enable users to control the model directly in a natural way.

(4) The query history not only provides users with the ability to reuse their previous queries, but also enables them to expand future search queries by taking previous queries into account. The positive and negative history panels together with the current query feed the 'time' and 'frequency' factor of our four-factor user interaction model.

2.3 Summary

In this chapter, we proposed a framework - uInteract, which includes a user interaction model and interactive interface.

In an effort to alleviate the limitations of current user interaction (UI) models and to find a UI model to deliver a better interaction and search accuracy for CBIR, we have proposed a new four-factor user interaction model (FFUIM) based on relevance region, relevance level, time and frequency. We have also empirically investigated different settings of the proposed model.

The following main observations have been made from the lab-based simulated experiment results: (1) bringing the user into the loop will enhance CBIR; (2) allowing both positive and negative feedback improves search performance; (3) combining the relevance level and frequency factor with other factors may make the user interaction model more usable and may improve the search accuracy.

We then developed an interactive visual interface. The interface is developed to achieve two objectives: (a) to deliver an effective interactive CBIR framework, in particular through a novel four-factor user interaction model, (b) to design the interaction activities of the interface to enable users to directly control the model in a natural way.

Overall, the development and investigation of the uInteract framework has answered the research questions Q1 and Q2 in Section 1.7. Whilst the framework is developed for our research purposes, we believe the factors in the model and functionalities on the interface could be adapted to any content-based search framework.

In the next Chapters, we will test the ease of use and usefulness of the new search functionalities through a user study.

Chapter 3

User Evaluation Methodology

In Chapter 2, we introduced the uInteract framework including a four-factor user interaction model (FFUIM) and a interactive interface for delivering the FFUIM visually. From this Chapter, we will start to evaluate the framework by a task-based user study. This Chapter will introduce our user evaluation methodology.

Section 3.1 reviews the background of user evaluation methodology for interactive search systems. Section 3.2 describes the evaluation setup for our user evaluations. The evaluation procedure on how we organize the user evaluations is introduced in Section 3.3. Section 3.4 and Section 3.5 state the main performance indicators, the hypothesis to test, and the procedures for the quantitative results analysis of the three evaluations. A summary of the Chapter is given in Section 3.6.

3.1 Background

To date, most of the evaluations of relevance feedback techniques for content-based image retrieval are still system-oriented. For instance, the automatic pseudo or simulated user evaluation are applied on a standard benchmark (e.g. the Benchathlon network (Ben), ImageCLEF (Ima) and TRECVID (Tre) are currently online) with fixed queries, testing data and relevance judgement file, and the search results will be measured by precision and recall (Rijsbergen 1979; Baeza-Yates and Ribeiro-Neto 1999) against the relevance judgement file.

However, searchers in real-life seek to optimize the entire search process, not just results accuracy, thus, evaluation of output alone is not enough to explain searchers' behaviour (Järvelin 2009; Lew et al. 2006). In particular, the provided relevance judgement file is no longer suitable for the interactive search when users start to indicate the more appropriate query examples from the result set by relevance feedback techniques. Users' relevance assessment changes with the actual search process, such as via learning from the results and reformulating their information needs. Therefore, user-oriented evaluation is needed for evaluating the interactive relevance feedback techniques by real, as opposed to simulated, users. Some researchers have applied different types of user-oriented design to their studies (Ingwersen 1992; Borlund and Ingwersen 1997; Jose et al. 1998; White et al. 2005; Urban and Jose 2006b).

3.2 User Evaluation Setup

We adapt the design of Urban and Jose (2006b) to our evaluation, which applies natural life scenarios to formulate the tasks. The natural search scenario is aimed at recreating tasks from an individual's real life searching. This allows the users to develop their own interpretation of the task and use their own judgement for choosing relevant images. This way, we can study how information needs evolve and what influence the interface has on their search and how users manage to adapt to the search strategy that the model requires.

From our early simulated user evaluation results, we have got positive findings on the effects of the relevance region factor, the effects of the four profiles of the Ostensive Model, the effects of the relevance level factor and the effects of the frequency factor. Therefore, we would like to find out how the effects of these factors and their combinations are going to be under real users' assessment.

In our user study, we have three focused evaluations: evaluation1 (E1) is to evaluate the ease of use and usefulness of the functionalities on the uInteract interface; evaluation2 (E2) is to evaluate the performance of the four profiles of the Ostensive Model; evaluation3 (E3) is to evaluate the effectiveness of the different settings of the four-factor user interaction model.

White and Morris (2007) find that users' behaviours are different for querying, result clicking and post-query navigation when comparing search experts to common users. To take their findings into account, we employ a total of 50 subjects[1] for the three focused evaluations. They are a mixture of males and females, undergraduate and postgraduate students and academic staff from a variety of departments with different ages and levels of image search experience. Subjects can be classified into two categories - inexperienced or experienced - based on their image search experience. We consider that people are experienced subjects if they search images at least once a week, and otherwise they are inexperienced subjects.

The 50 subjects were divided into three groups. 17, 16 and 17 subjects assigned to E1, E2 and E3 respectively based on the minimum sample size (16) suggested by the TREC interactive track (Dumais and Belkin 2005). In each evaluation, the subjects attempted four different complexity levels of search tasks on the four systems randomly in a random order (limited to five minutes for each task) and provided feedback on their search experiences through questionnaires and comments made during informal interviews. The detailed setups of the three evaluations will be described in the next chapters. The evaluation systems are different for different evaluations.

The data is collected by means of questionnaires, informal interviews, actual search results of every task and screen captures for the evaluation with video and audio

[1] We will call users "subjects" in Chapter3, Chapter4, Chapter5 and Chapter6.

input. The questionnaires use five point Likert scales, and include entry questionnaire, post-search questionnaire, and exit questionnaire. The entry questionnaire is used to find out the subjects' age, background, experience on searching images and expectations on image search tools. The information can be used to classify different subjects' profiles based on age or image search experience. The post-search questionnaire is to assess the task they have just performed, and that system and search results. The information will show subjects' opinion on an individual task, system and search experience. The exit questionnaire is to compare the four tasks, underlying systems and search results that the subjects have just processed. The information will show subjects' general opinion on the evaluation. The informal interview happens during the search process and after completing the evaluation, to get users' feedback on the tasks, systems and search experiences, which they have not be able to provide in the questionnaires. We will be able to extract the search accuracy from subjects' actual search results of the completed tasks. The screen capture with video and audio input will provide rich user interaction data for our qualitative data analysis on finding person profiles.

We will use the data from questionnaires and actual search results for quantitative analysis, and use the data from screen capture and information interview for qualitative analysis.

3.3 User Evaluation Procedure

The evaluation procedure for each subject is as follows:

- an introduction to the purpose of the evaluation;
- an entry questionnaire;
- a hand out of pre-ordered written instructions for four tasks and four pre-ordered post-search questionnaires (the order is random, so everybody might

get different combination of tasks and systems, and also test the systems in a different order);

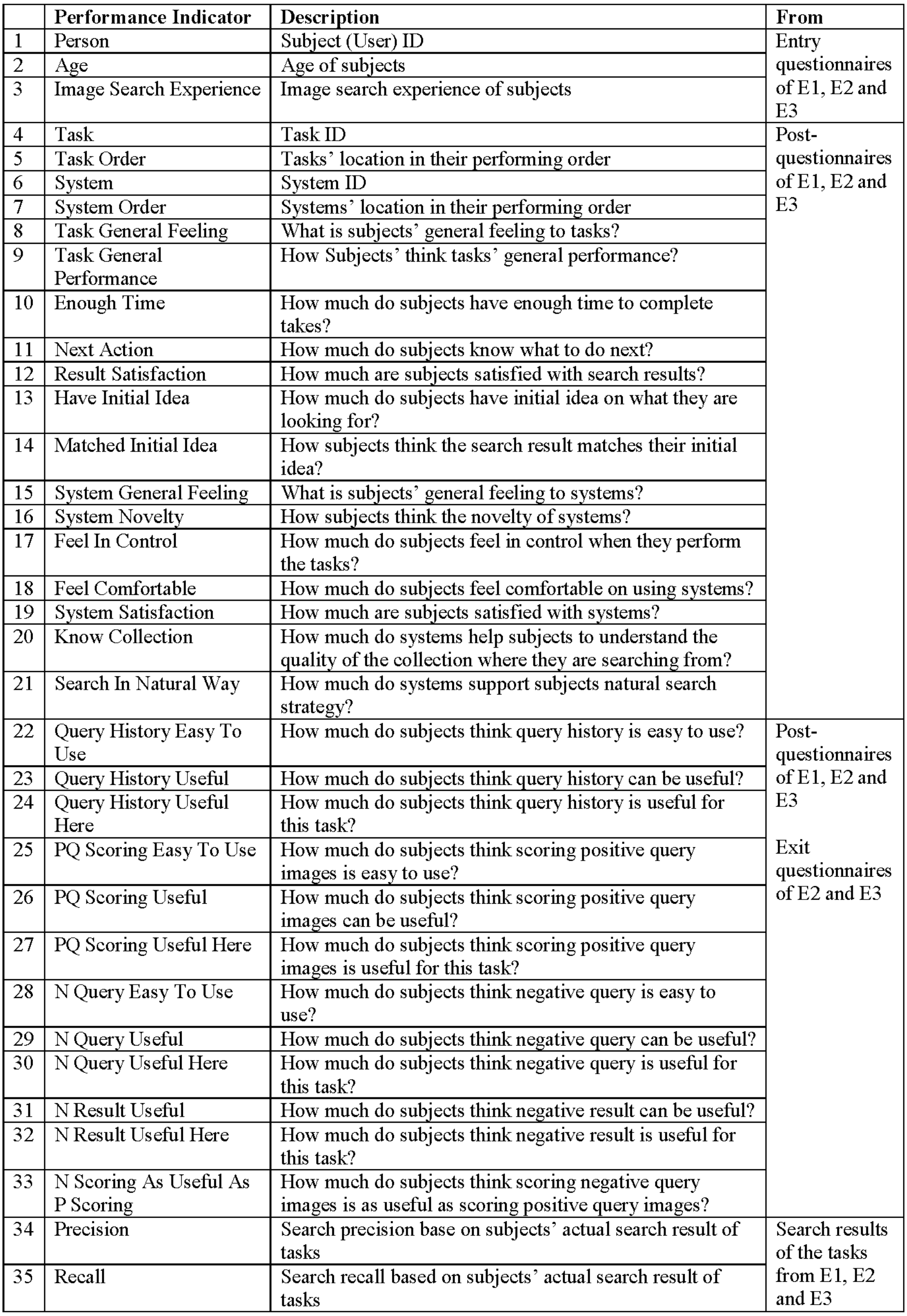
- a training session on the systems with which the subject were to test and how to read the task instructions and how to complete the questionnaires;
- the first search session in which the subject interacted with the first system in the order and its matched task;
- a post-search questionnaire;
- the second search session in which the subject interacted with the second system in the order and its matched task;
- a post-search questionnaire;
- the third search session in which the subject interacted with the third system in the order and its matched task;
- a post-search questionnaire;
- the fourth search session in which the subject interacted with the fourth system in the order and its combined task;
- a post-search questionnaire;
- an exit questionnaire;
- an informal interview;
- the whole process was recorded by screen capture with video and audio input.

3.4 Main Performance Indicators and Hypothesis of Quantitative Analysis

The main performance indicators of the qualitative data are generated from the questionnaires (please refer to Appendix A and Appendix B) and actual search results. The main indicators of E1, E2 and E3 are listed in Table 3.1.

In order to answer the research questions Q1.1, Q1.2, Q1.3, Q2.1 and Q2.2 addressed in Section 1.7, we propose nine hypotheses and test them by a quantitative data analysis. The nine hypotheses are:

- Hypothesis1: Task Order and System Order will affect the performance indicators (8-33) provided by subjects because of familiarity or fatigue;
- Hypothesis2: System will affect the performance indicators (8-33);
- Hypothesis3: Task will affect the performance indicators (8-33) provided by subjects because of different complexity levels;
- Hypothesis4: The interaction between Task and System will influence the scores of the performance indicators (8-33);
- Hypothesis5: Person will affect the performance indicators (8-33), based on individual differences;
- Hypothesis6: The subjects' Age and prior Image Search Experience of the subjects will affect subjects' opinion of the overall search experience (8-21);
- Hypothesis7: The subjects' Age and prior Image Search Experience of the subjects will have effects on the subjects' opinion on the functionalities of the interfaces (22-33);
- Hypothesis8: System and Task will have an impact on Precision of the search results (34);



	Performance Indicator	Description	From
1	Person	Subject (User) ID	Entry questionnaires of E1, E2 and E3
2	Age	Age of subjects	
3	Image Search Experience	Image search experience of subjects	
4	Task	Task ID	Post-questionnaires of E1, E2 and E3
5	Task Order	Tasks' location in their performing order	
6	System	System ID	
7	System Order	Systems' location in their performing order	
8	Task General Feeling	What is subjects' general feeling to tasks?	
9	Task General Performance	How Subjects' think tasks' general performance?	
10	Enough Time	How much do subjects have enough time to complete takes?	
11	Next Action	How much do subjects know what to do next?	
12	Result Satisfaction	How much are subjects satisfied with search results?	
13	Have Initial Idea	How much do subjects have initial idea on what they are looking for?	
14	Matched Initial Idea	How subjects think the search result matches their initial idea?	
15	System General Feeling	What is subjects' general feeling to systems?	
16	System Novelty	How subjects think the novelty of systems?	
17	Feel In Control	How much do subjects feel in control when they perform the tasks?	
18	Feel Comfortable	How much do subjects feel comfortable on using systems?	
19	System Satisfaction	How much are subjects satisfied with systems?	
20	Know Collection	How much do systems help subjects to understand the quality of the collection where they are searching from?	
21	Search In Natural Way	How much do systems support subjects natural search strategy?	
22	Query History Easy To Use	How much do subjects think query history is easy to use?	Post-questionnaires of E1, E2 and E3 Exit questionnaires of E2 and E3
23	Query History Useful	How much do subjects think query history can be useful?	
24	Query History Useful Here	How much do subjects think query history is useful for this task?	
25	PQ Scoring Easy To Use	How much do subjects think scoring positive query images is easy to use?	
26	PQ Scoring Useful	How much do subjects think scoring positive query images can be useful?	
27	PQ Scoring Useful Here	How much do subjects think scoring positive query images is useful for this task?	
28	N Query Easy To Use	How much do subjects think negative query is easy to use?	
29	N Query Useful	How much do subjects think negative query can be useful?	
30	N Query Useful Here	How much do subjects think negative query is useful for this task?	
31	N Result Useful	How much do subjects think negative result can be useful?	
32	N Result Useful Here	How much do subjects think negative result is useful for this task?	
33	N Scoring As Useful As P Scoring	How much do subjects think scoring negative query images is as useful as scoring positive query images?	
34	Precision	Search precision base on subjects' actual search result of tasks	Search results of the tasks from E1, E2 and E3
35	Recall	Search recall based on subjects' actual search result of tasks	

Table 3.1: The main performance indicators from the three evaluations for qualitative data analysis

- Hypothesis9: System and Task will have an impact on Recall of the search results (35).

3.5 Quantitative Data Analysis Procedure

The qualitative data analysis was supported by the use of statistical software, namely Statistical Package for Social Science (SPSS). The procedure adopted for the qualitative data analysis was as follows:

1. Identify precision value and recall for the 12 tasks preformed by 50 subjects;

   - Get result images:

     We firstly get the union ($\bigcup$) of result images of one task from all the result images selected by all of the subjects who did this task. Then we do the same to the other 11 tasks (4 tasks in each evaluation) to get 12 result images union sets;

   - Get independent raters to rate the result images:

     We ask 5 independent raters to rate all images in the 12 result union sets with 1 to 5 scales (5 is the most relevant). The raters give a relevance value (between 1 and 5) to every image in a union result set of a task, and the rater will do the same to the result images of the other 11 tasks. We test the reliability of the raters' rating value of all the images for the 12 tasks by Cronbach's Alpha statistical test according to a reliability of 0.70 or higher in SPSS, and find the reliability for all of the 12 tasks across the three evaluations;

   - Get the precision value:

     The precision value for each result image is the mean rating value provided by the five raters to the image with 1 to 5 scales. The precision value of a task is the mean precision value of all the result images of the task;

- Get the recall value:

  The recall of a task is the number of images selected by a subject to the result for completing the task;

2. Obtain the figures for the performance indicators listed in Table 3.1 from the questionnaires and the actual search results for the three focused evaluations, and test the nine hypothesis we intended to investigate in Section 3.4 by factorial ANOVA statistical tests;

3. Analyze the testing results we obtained from the ANOVA test.

3.6 Summary

This Chapter has described our user study methodology and quantitative data analysis methodology. The setup and results of the qualitative data analysis for three focused evaluations will be reported in Chapter 4, Chapter 5 and Chapter 6, respectively.

Chapter 4

Evaluation of the Effects of the uInteract Interface

In Chapter 3, we described our user evaluation methodology and quantitative data analysis methodology for the three focused evaluations. This Chapter will report the setup (Section 4.1) and results (Section 4.2) of evaluation 1 (E1). The goal of E1 is to test whether users find the uInteract interface is useful and easy to use. Section 4.3 summarizes the Chapter.

4.1 Evaluation Setup

Seventeen subjects participated in E1. They were asked to complete four search tasks on four interfaces in a random order, and provide feedback on their search experiences through questionnaires and comments made during informal interviews. The tasks were designed at different complexity levels. The task descriptions and questionnaires of this evaluation are provided in Appendix A.

The complexity level of each task in E1 is reflected by the task description. Task1 (T1) provides both search topic and example images, so we consider it the easiest

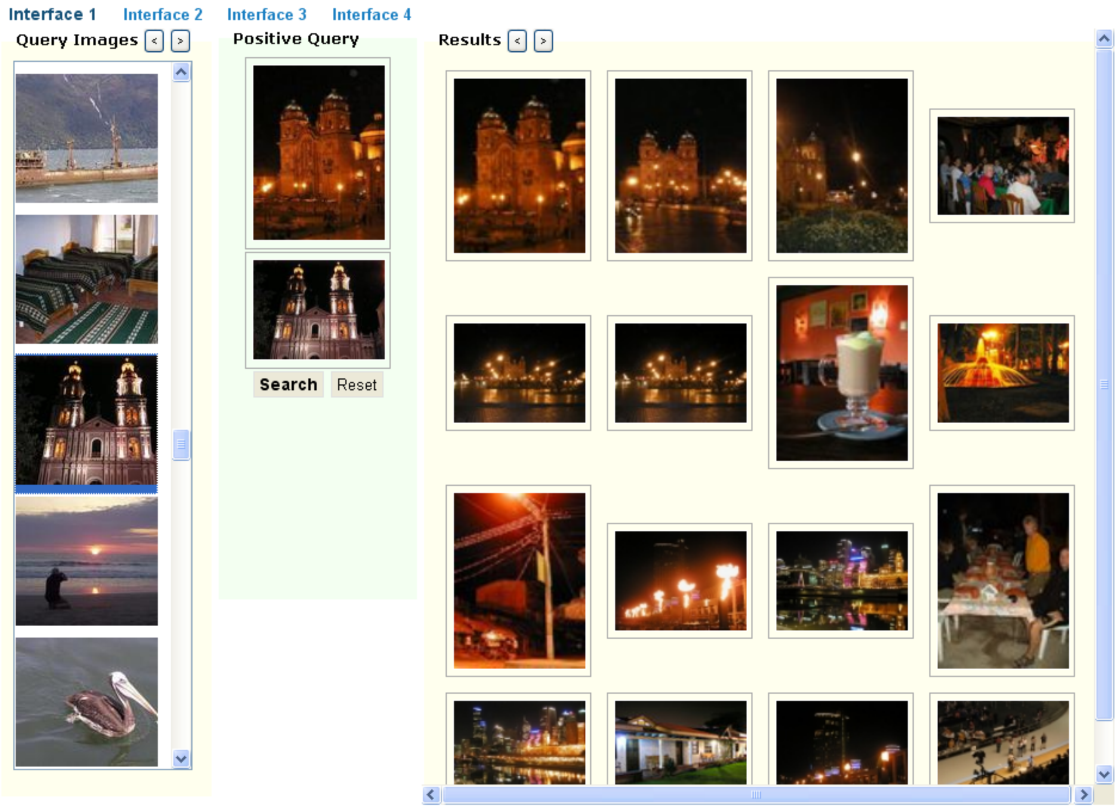


Figure 4.1: E1_Interface1 (I1)

task in term of the "easiness" of formulating the query and identifying the information need. Task2 (T2) gives example images without a topic description, so we consider it harder than T1. Task3 (T3) has only a topic but no image examples, which is even harder than T2. Task4 (T4) describes a broad search scenario without any specific topic and image examples, so it is the hardest task in our view.

We created four testing systems. System1 (I1) (Figure 4.1) has a typical Relevance Feedback (RF) interface, where users are allowed to give positive feedback from search results through a simplified interface. System2 (I2) (Figure 4.2) - an interface based on Urban et al. (2006) Ostensive Model, provides positive query history functionality which is an addition to I1. System3 (I3) (Figure 4.3) - an interface based on Ruthven et al. (2003) interaction model, enhances I2 by adding partial relevance (we call it **importance score** here) functionality on the interface. System4 (I4) (Figure 4.4) is the uInteract interface that we proposed in Section 2.2.2 based on our four-factor user interaction model.

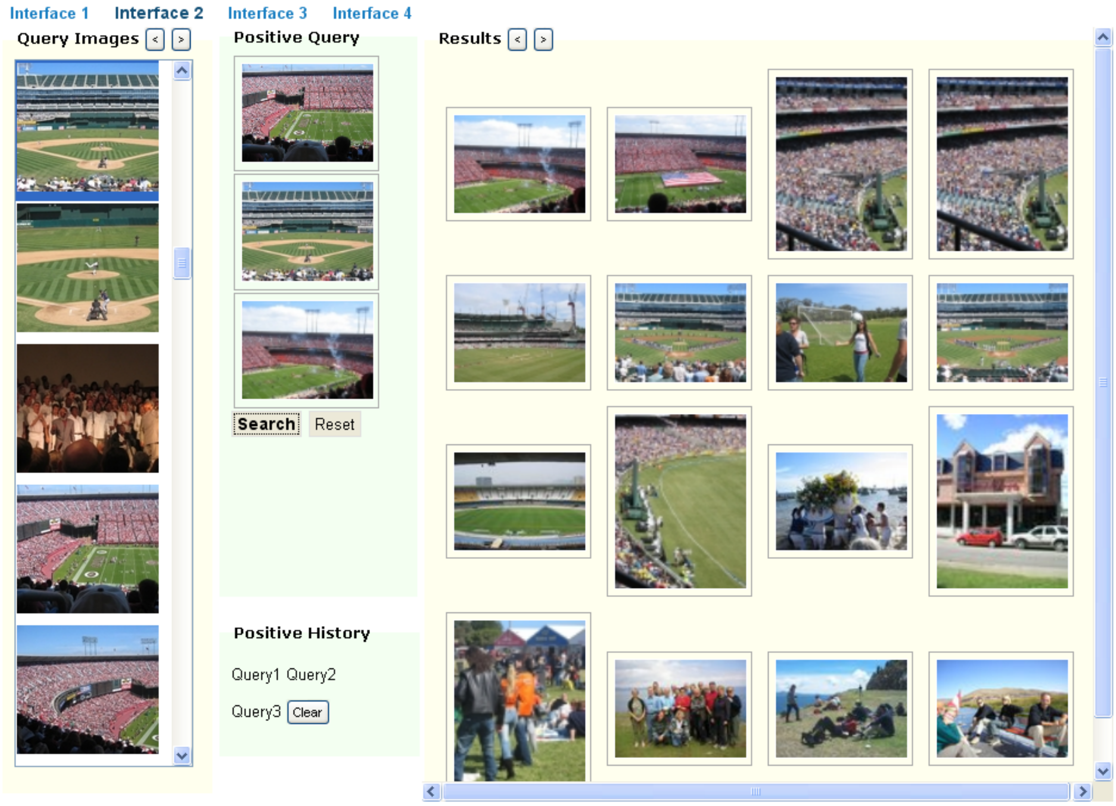


Figure 4.2: E1_Interface2 (I2)

4.2 Evaluation Results and Analysis

The following results were obtained by applying ANOVA analysis (with $\alpha = 0.05$) on the experimental results in terms of the main performance indicators (introduced in Table 3.1). We have broken down our main goal of this evaluation into nine hypotheses in Section 3.4. The analysis will focus on the individual hypotheses. Only statistically significant results will be listed.

Hypothesis1: Task Order and System Order will affect the performance indicators (8-33) provided by subjects because of familiarity or fatigue.

The factorial ANOVA results showed that Task Order and System Order did not significantly affect the scores provided by subjects on all the performance indicators. Nor was there a significant interaction between the effects of Task Order and System Order on any of the performance indicators neither.

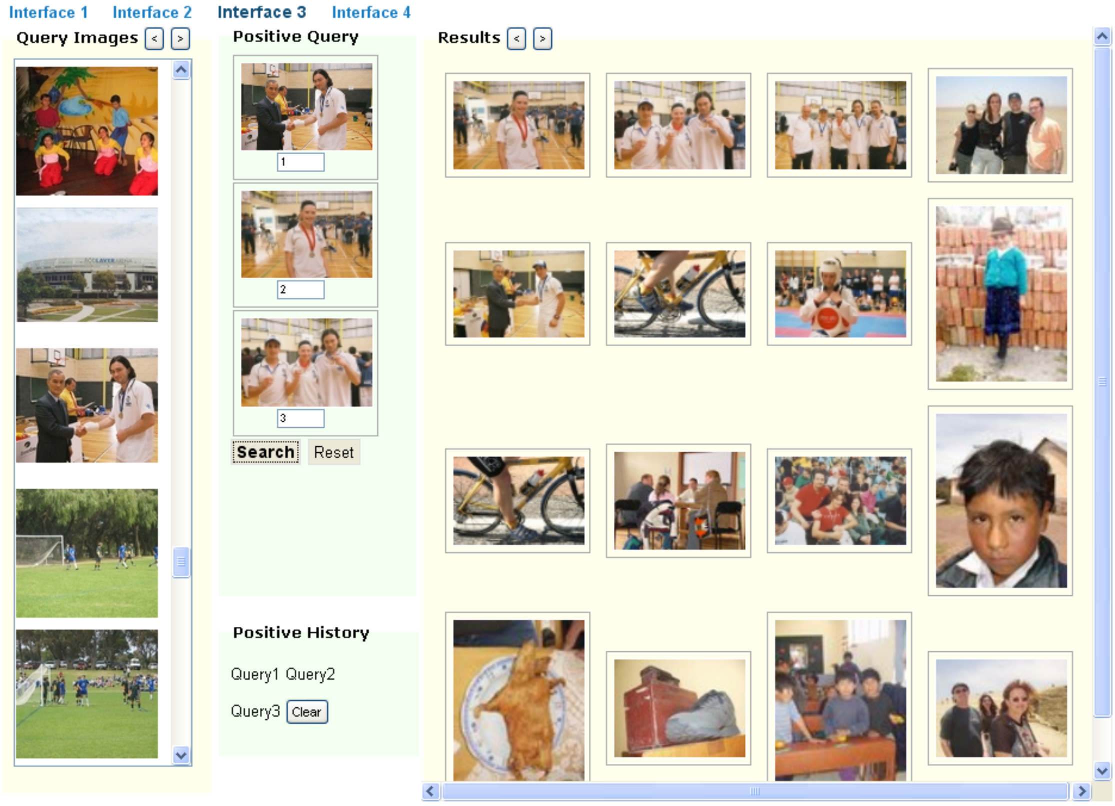


Figure 4.3: E1_Interface3 (I3)

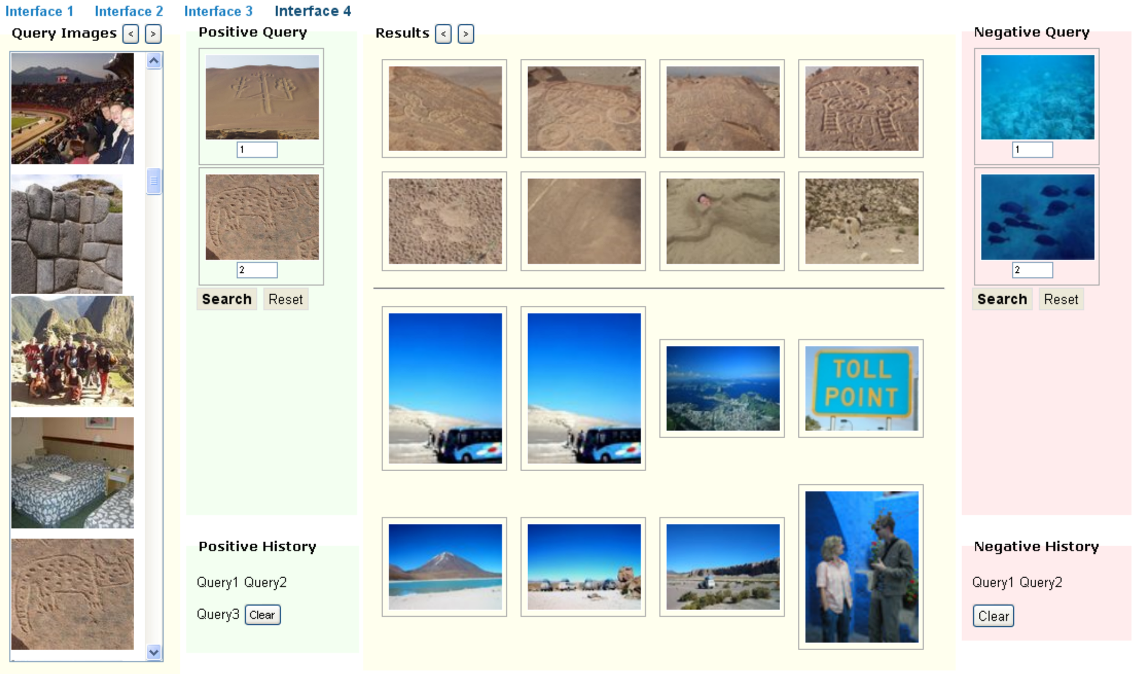


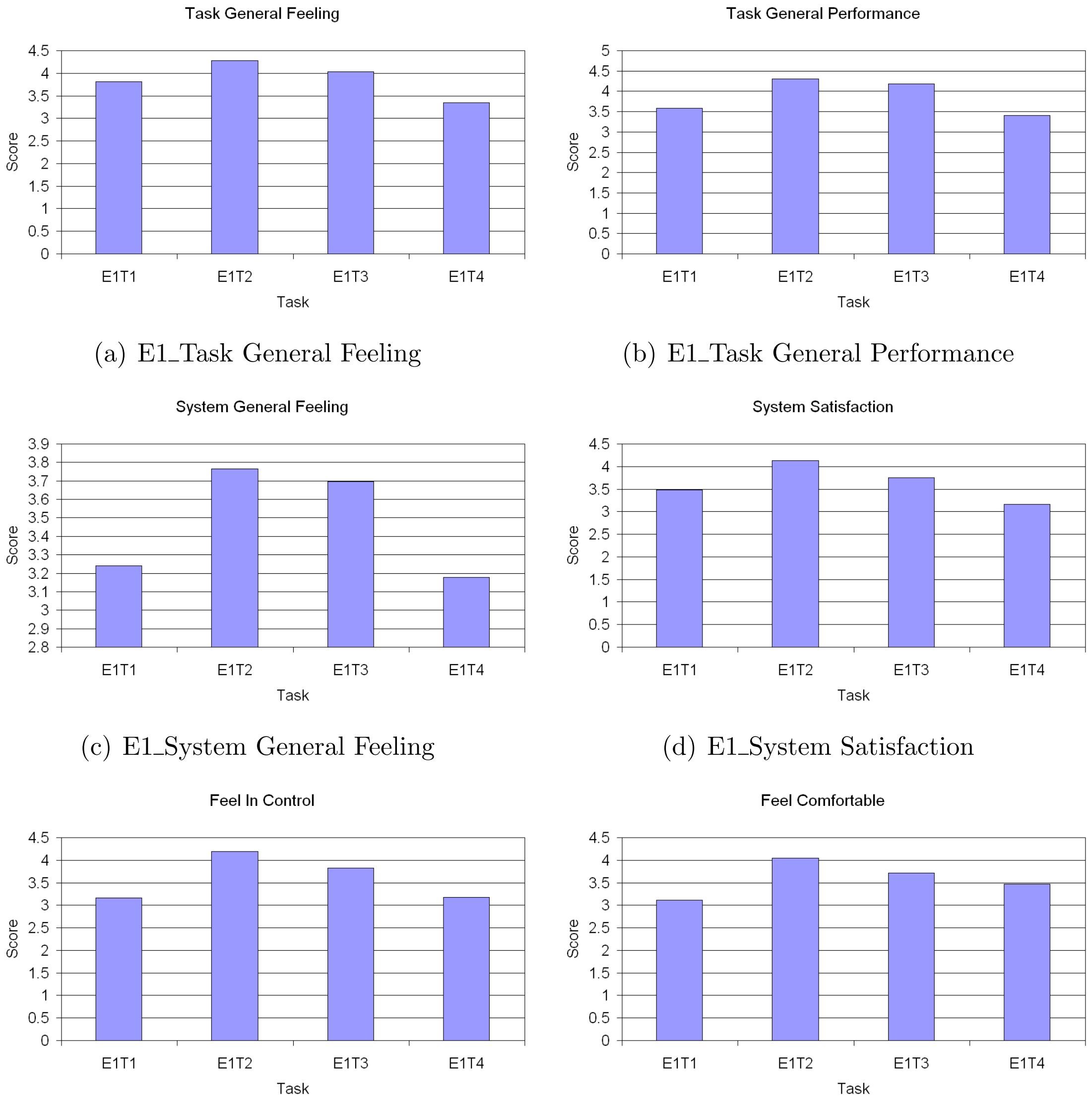
Figure 4.4: E1_Interface4 (I4)

Hypothesis2: System will affect the performance indicators (8-33).

The factorial ANOVA results showed that System had no significant effects on any of the performance indicators.

Hypothesis3: Task will affect the performance indicators (8-33) provided by subjects because of different complexity levels.

The factorial ANOVA results showed that Task significantly impinged on the following performance indicators (Figure 4.5):



(a) E1_Task General Feeling

(b) E1_Task General Performance

(c) E1_System General Feeling

(d) E1_System Satisfaction

(e) E1_Feel In Control

(f) E1_Feel Comfortable

Figure 4.5: E1: Effects of Task on performance indicators (8-33)

- **Task General Feeling [F(3,61)=2.63, p=0.013]** (Figure 4.5(a)). The pairwise comparison analysis revealed that the subjects performing T2 (p=0.002) and T3 (p=0.019) gave higher scores on the Task General Feeling than the hardest task (T4).

- **Task General Performance [F(3,61)=3.25, p=0.019]** (Figure 4.5(b)). The pairwise comparison analysis revealed that the subjects performing T2 (p=0.009) and T3 (p=0.023) gave higher scores on the Task General Performance than the hardest task (T4). However, the easiest task (T1) performed worse than a harder task (T2) (p=0.033).

- **System General Feeling [F(3,61)=4.88, p=0.004]** (Figure 4.5(c)). The pairwise comparison analysis revealed that when subjects performed T2 (p=0.004) and T3 (p=0.010), they tended to gave higher scores on System General Feeling than the hardest task (T4). However, when the subjects performed the easiest task (T1), they gave lower scores than when they performed two harder tasks T2 (p=0.009) and T3 (p=0.022).

- **System Satisfaction [F(3,61)=5.04, p=0.003]** (Figure 4.5(d)). The pairwise comparison analysis revealed that the subjects were more satisfied with the system when they performed T2 (p=0.000) and T3 (p=0.026) than the hardest task (T4). However, the subjects were more satisfied with the system when they performed a harder task (T2) than the easiest task (T1) (p=0.014).

- **Feel In Control [F(3,61)=4.56, p=0.006]** (Figure 4.5(e)). The pairwise comparison analysis revealed that the subjects felt more in control on completing the tasks when they performed easier task (T2) (p=0.003) than the hardest task (T4). However, the subjects felt more in control when they performed a harder task (T2) (p=0.014) than the easiest task (T1).

- **Feel Comfortable [F(3,61)=2.96, p=0.039]** (Figure 4.5(f)). The pairwise comparison analysis revealed that the subjects felt more comfortable using the systems when they performed a harder task (T2) than the easiest task (T1) (p=0.005).

In summary, Figure 4.5 shows that the subjects tend to give higher scores to the performance indicators when they perform easier tasks, such as T2 and T3. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree that T4 is the hardest task, and T3 is harder than T2. However, they think T2 and T3 are easier than T1, although based on the task description T1 is regarded as the easiest task because T1 has both text and image description. This may be because: first the colour of the three initial query image examples given in T1 is more complex than in T2; second the task description of T1 is actually more constraining while they can interpret T2 more freely.

Hypothesis4: The interaction between Task and System will influence the scores of the performance indicators (8-33).

System Novelty	**E1T1**	**E1T2**	**E1T3**	**E1T4**
E1I1	3.5	2.4	5	3.75
E1I2	4.25	4	3.75	3.2
E1I3	4.2	4	3.75	3.5
E1I4	3.75	4.25	4	4.25

Table 4.1: E1: Effects of the interaction between Task and System on performance indicators (8-33)

The factorial ANOVA results showed there was no significant interaction between the effects of Task and System on most performance indicators. However, there was a significant interaction between Task and System on the scores of System Novelty [F(9,52) = 3.49, p = 0.002], although the pairwise interaction comparison analysis did not reveal any significant difference. The interaction scores between Task and System are shown in Table 4.1.

Hypothesis5: Person will affect the performance indicators (8-33), based on individual differences.

The factorial ANOVA results showed that the differences between individual users (Person) significantly affected their scores on most performance indicators. The affected indicators were:

- Next Action, F(16,45)=4.33, p=0.000;
- Have Initial Idea, F(16,45)=2.40, p=0.011;
- System Novelty, F(16,45)=7.65, p=0.000;
- Feel In Control, F(16,45)=2.89, p=0.003;
- Feel Comfortable, F(16,45)=3.41, p=0.001;
- Know Collection, F(16,45)=2.45, p=0.009;
- Search In Natural Way, F(16,45)=10.34, p=0.000.

From the above results we can see that Person is another important factor which affects many performance indicators. However, the results do not show how Person affects these indicators. White and Morris (2007) find that the behaviour is different between search experts and common users. Thus, we wonder whether the subjects' image search experience is a key factor of the effects? Further, will the subjects' age be a key factor? In an effort to find how Person influences the performance indicators, we take the age and image search experience into account in the following investigation on the Person factor.

Hypothesis6: The subjects' Age and prior Image Search Experience of the subjects will affect subjects' opinion of the overall search experience (8-21).

The factorial ANOVA with covariate results showed that Age and Image Search Experience significantly affected following performance indicators:

- **Result Satisfaction.** Age affected Result Satisfaction [F(1,64) = 5.06, p = 0.028]. Image Search Experience also affected Result Satisfaction [F(1,64) = 5.93, p = 0.018], and so did their interaction [F(1,64) = 5.85, p = 0.018]. The resultant equations[1] (Dowdy et al. 2004; Calder 1996) were: for inexperienced people, Result Satisfaction[2] = 6.882 - 0.130*Age - 2.952*1 + 0.122*1*Age

[1] The regression equations is derived from the ANOVA results (Dowdy et al. 2004)

[2] In this regression equation, $ResultSatisfaction$ is the score that we want to predict, 6.882 is the intercept B value (regression coefficient) from ANOVA results which is the point at which the

=3.930 - 0.008*Age; for experienced people, Result Satisfaction[3] = 6.882 - 0.130*Age - 2.952*2 + 0.122*2*Age = 0.978 + 0.114*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.008 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.114 per year as the age increased.

- **Matched Initial Idea.** Age affected Matched Initial Idea [F(1,64) = 5.10, p = 0.027]. Image Search Experience also affected Matched Initial Idea [F(1,64) = 5.85, p = 0.018], and so did their interaction [F(1,64) = 5.52, p = 0.022]. The resultant equations were: for inexperienced people, Matched Initial Idea = 6.671 - 0.125*Age - 2.796*1 + 0.113*1*Age =3.875 - 0.012*Age; for experienced people, Matched Initial Idea = 6.671 - 0.125*Age - 2.796*2 + 0.113*2*Age = 1.079 + 0.101*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.012 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.101 per year as the age increased.

- **Feel In Control.** Age affected Feel In Control [F(1,64) = 4.13, p = 0.046]. Image Search Experience also affected Feel In Control [F(1,64) = 5.00, p = 0.029], and so did their interaction [F(1,64) = 6.59, p = 0.013]. The resultant equations were: for inexperienced people, Feel In Control = 6.241 - 0.136*Age - 3.122*1 + 0.149*1*Age = 3.119 + 0.013*Age; for experienced people, Feel In Control = 6.241 - 0.136*Age - 3.122*2 + 0.149*2*Age = -0.003 + 0.162*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.013 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.162 per year as the age increased.

regression line cuts the Y axis, -0.130 is the B value of Age from ANOVA results which is the slope or the gradient of the line of Age, -2.952 is the B value of Image Search Experience from ANOVA results which is the slope of the line of Image Search Experience, 1 is the real value to indicate that a user is inexperienced, 0.122 is the B value of interaction between Age and Image Search Experience (Age*Image Search Experience) which is the slope of the line of Age*Image Search Experience.

[3] This regression equation is similar with the equation above. The only one difference is that 2 is the real value to indicate that a user is experience.

- **Feel Comfortable.** The equation for predicting Feel Comfortable was: Feel Comfortable = 1.595 + 0.033*Age + 0.785*Image Search Experience. Age affected Feel Comfortable [F(1,65) = 4.04, p = 0.049]: the scores of Feel Comfortable increased by 0.033 per year increase in age. Image Search Experience also affected Feel Comfortable [F(1,65) = 10.16, p = 0.02]: if the subjects had higher level of image search experience, then their scores of Feel Comfortable were, on average, 0.785 higher.

- **System Satisfaction.** Age affected System Satisfaction [F(1,65) = 5.92, p = 0.018]: the System Satisfaction scores were estimated to increase by 0.035 per year increase in age.

- **Know Collection.** Age affected subjects' opinions on Know Collection [F(1,65) = 9.02, p = 0.004]: the scores increased by 0.050 per year increase in age.

Hypothesis7: The subjects' Age and prior Image Search Experience of the subjects will have effects on the subjects' opinion on the functionalities of the interfaces (22-33).

The one-way ANOVA results showed that the Image Search Experience of the subjects also significantly influenced the opinions on Query History Useful [F(1,15) = 7.67, p = 0.014], and N Query Easy To Use [F(1,15) = 8.06, p = 0.012]. The experienced subjects gave higher scores to these indicators than inexperienced subjects. Simple effects of different Age were not significant, indicating that the subjects had the same opinion about the new functionalities regardless of the difference in age.

Hypothesis8: System and Task will have an impact on Precision of the search results (34).

The factorial ANOVA results in Figure 4.6 showed that Task significantly impinged the Precision (the precision calculation procedure is in Section 3.5) of the actual image search results [F(3,61) = 2.94, p = 0.040]. The pairwise comparison analysis revealed that the Precisions of T2 (p=0.010) and T3 (p=0.025) were higher than

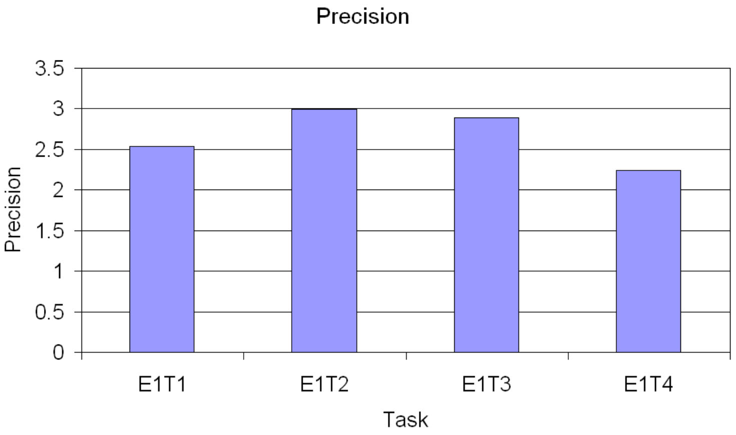


Figure 4.6: E1: Effects of the Task on Precision

the hardest task (T4). There were no significant differences between the Precision of the search results for different systems.

Hypothesis9: System and Task will have an impact on Recall of the search results (35).

Recall	**E1T1**	**E1T2**	**E1T3**	**E1T4**
E1I1	4.75	2	3	2.75
E1I2	3.75	2	3	4
E1I3	2	2	3	6
E1I4	3.75	2	3.4	3.5

Table 4.2: E1: Effects of the Task and System on Recall

The factorial ANOVA results showed that Task significantly impinged the Recall of the actual image search results [F(3,52) = 5.99, p = 0.001]. The pairwise comparison analysis revealed that the Recall of T1 (p=0.003), T3 (p=0.035) and T4 (p=0.000) were higher than the Recall of T2. This is because we limited the number of result images in each task description, and the number of result images in T1, T3 and T4 was higher than the number in T2. There were no significant differences between the Recall of the search results for different systems.

There was a significant interaction between Task and System on Recall [F(9,52) = 2.13, p = 0.043]. The interaction scores between Task and System are shown in

Table 4.2.

4.3 Summary

In this Chapter, we reported the evaluation setup for evaluating the ease of use and usefulness of the uInteract system (E1), and the results obtained from the ANOVA analysis (with $\alpha = 0.05$) on the main performance indicators based on the nine hypotheses, corresponding to the research questions Q2.1 and Q2.2 addressed in Section 1.7.

We summarize the quantitative analysis results of E1 based on the nine hypotheses as below (Table 6.4):

- Hypothesis1 is not supported because System Order and Task Order do not affect the performance indicators at all. This implies the familiarity or fatigue with the task and the system does not make a difference to the subjects' scores on the indicators.
- Hypothesis2 is not supported because System does not influence the performance indicators at all, meaning there are no significant differences between different systems.
- Hypothesis3 is partially supported because Task has a strong impact on most performance indicators, meaning the complexity level of a task does affect the subjects' opinions related to the performance indicates.
- Hypothesis4 is partially supported because the interaction between Task and System significantly affect the scores of System Novelty, although there is no significant impact from System only.
- Hypothesis5 is partially supported because Person affects most performance indicators, implying different individuals have very different preferences.

- Hypothesis6 is partially supported because Age and Image Search Experience and their interaction significantly affect the subjects' opinions on most performance indicators.

- Hypothesis7 is partially supported because the experienced subjects give higher scores on Query History Useful and N Query Easy To Use than inexperienced subjects.

- Hypothesis8 is partially supported because Task significantly affects the Precision of the search results, although there is no significant impact from System and the interaction between Task and System.

- Hypothesis9 is partially supported because Task and the interaction between Task and System significantly affect the Recall of the search results although there are no significant differences between systems.

All in all, we conclude the Chapter by summarizing key results for three main aspects: system, task and users. (1) It is difficult to identify the effects of the uInteract interface because Task and Person factors strongly impinge on the scores of the performance indicators (related to Q2.1). (2) Task strongly influences the performance indicators. One interesting observation from our analysis of the results is that the subjects tend to give higher scores to the performance indicators when they perform easier tasks. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree that T4 is the hardest task, and T3 is harder than T2. However, they think the T2 is easier than T1 although the description the T1 is more comprehensive. (3) Person is a very important factor which affects most performance indicators. There is clear evidence that subjects with different Age and Image Search Experience have different preferences on the interactive interfaces. The trend is that the subjects tend to be more satisfied with the system and understand the quality of the data collection better and feel more comfortable to use the system, with the increase in age. For the inexperienced subjects, their satisfaction with the search results and agreement on matching their initial idea tend to decrease as the age increases. However, for the experienced subjects, the scores tend to increase with the increase in age. It is also observed on Feeling In Control that for both experienced and inexperienced subjects, the scores tend to increase with the increase in age (related to Q2.2).

Chapter 5

Evaluation of the Effects of the Four Profiles of the OM

In Chapter 4, we reported the first focused evaluation (E1) setup and results on the ease of use and usefulness of the interactive uInteract interface. This Chapter will report the second focused evaluation (E2) on the effects of the four profiles of the Ostensive Model (OM). The goal of E2 is to test the performance of the four profiles of the Ostensive Model and whether users find the uInteract interface useful.

Section 5.1 reports the evaluation setup of E2. Section 5.2 reports the evaluation results of E2. The final findings will be stated in Section 5.3.

5.1 Evaluation Setup

Sixteen subjects participated in E2. They were asked to complete four search tasks on four systems in a random order, and provide feedback on their search experiences through questionnaires and comments made during informal interviews. The tasks were designed at different complexity levels. The task descriptions and questionnaires of this evaluation are provided in Appendix B.

The four tasks in E2 use the same description structure with both specific verbal search topic and three example images. The complexity level of each task is based on the search accuracy of the query images of the tasks from the lab-based simulated experimental results. The mean average precision (MAP) of task1 (T1), task2 (T2), task3 (T2) and task4 (T4) is 0.2420, 0.0872, 0.0294, 0.0098 respectively. We consider T1 is the easiest task with the highest precision, followed by T2, T3. T4 has lowest precision, thus we take it as the hardest task.

We created four testing systems[1]. System1 (OM1) applies the increasing profile of the Ostensive Model delivered by the uInteract interface. System2 (OM2) applies the decreasing profile of the Ostensive Model delivered by the uInteract interface. System3 (OM3) applies the flat profile of the Ostensive Model delivered by the uInteract interface. System4 (OM4) applies the current profile of the Ostensive Model delivered by the uInteract interface.

5.2 Evaluation Results and Analysis

The following results were obtained by applying ANOVA analysis (with $\alpha = 0.05$) on the experimental results in terms of the main performance indicators (introduced in Table 3.1). We have broken down our main goal of this evaluation into nine hypotheses in Section 3.4. The result analysis will focus on the individual hypotheses. Only statistically significant results will be listed.

Hypothesis1: Task Order and System Order will affect the performance indicators (8-33) provided by subjects because of familiarity or fatigue.

The factorial ANOVA results showed that Task Order and System Order did not significantly affect the scores provided by subjects on all the performance indicators. There was no significant interaction between the effects of Task Order and System Order on all the performance indicators either.

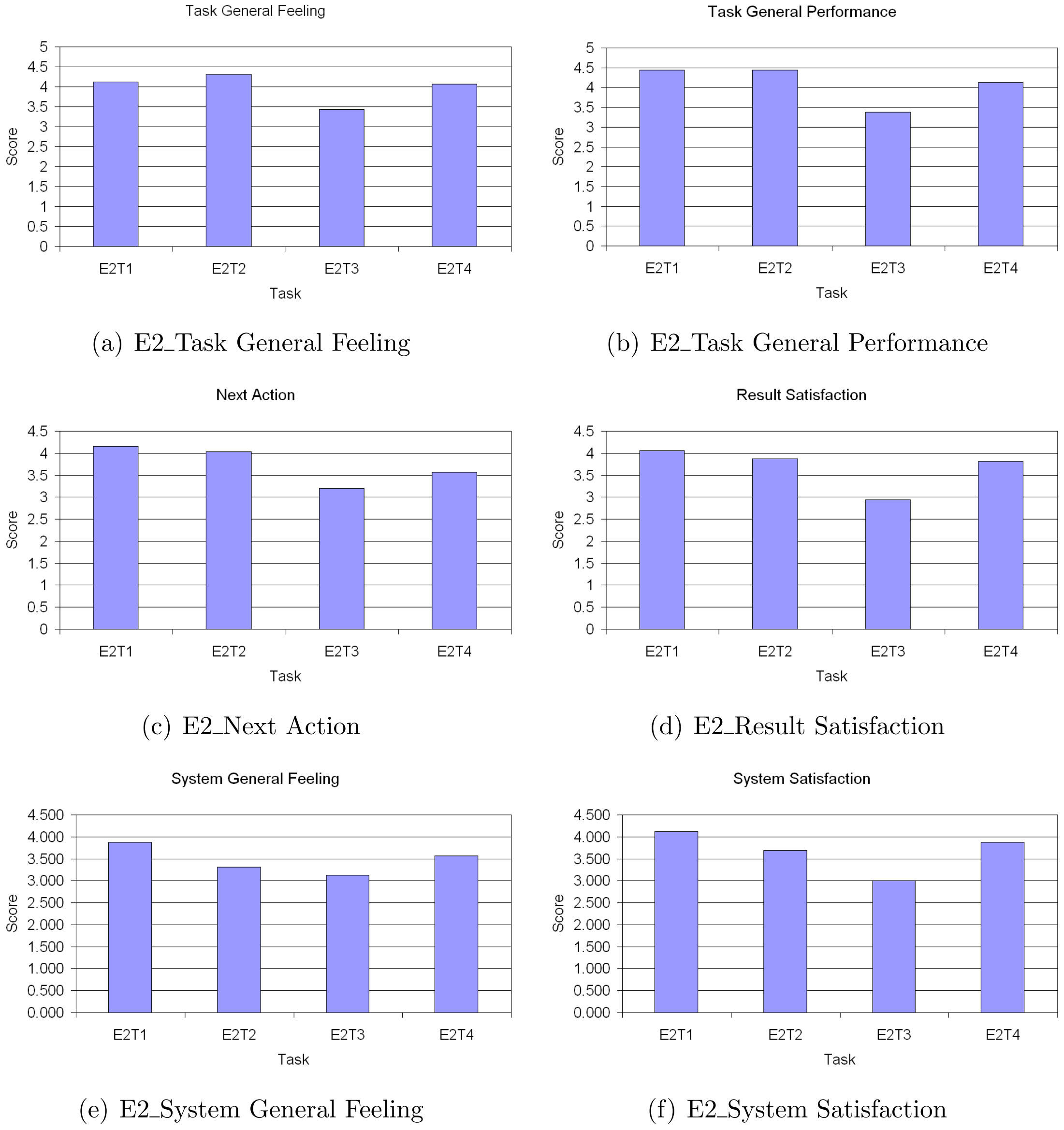
[1] The four systems are delivered by the uInteract interface (Figure 2.4).

Hypothesis2: System will affect the performance indicators (8-33).

The factorial ANOVA results showed that Systems had no significant effects on any of the performance indicators.

Hypothesis3: Task will affect the performance indicators (8-33) provided by subjects because of different complexity levels.

The factorial ANOVA results showed that the tasks with different complexity levels significantly impinged on the following performance indicators (Figure 5.1):



(a) E2_Task General Feeling

(b) E2_Task General Performance

(c) E2_Next Action

(d) E2_Result Satisfaction

(e) E2_System General Feeling

(f) E2_System Satisfaction

Figure 5.1: E2: Effects of Task on performance indicators (8-33)

- **Task General Feeling [F(3,57)=3.94, p=0.013]** (Figure 5.1(a). The pairwise comparison analysis revealed that the subjects performing the two easier tasks T1 (p=0.014) and T2 (p=0.002) gave higher scores on Task General Feeling than a harder task (T3). However, Task General Feeling scores of T3 were lower than the hardest task (T4) (p=0.025).

- **Task General Performance [F(3,57)=5.79, p=0.002]** (Figure 5.1(b)). The pairwise comparison analysis revealed that the subjects performing the two easier tasks T1 (p=0.001) and T2 (p=0.001) gave higher scores on Task General Performance than a harder task (T3). However, the T3 performed worse than the hardest task (T4) (p=0.014).

- **Next Action [F(3,26)=3.19, p=0.040]** (Figure 5.1(c)). The pairwise comparison analysis revealed that the subjects knew better what to do next when they performed the easiest task (T1) than a harder task (T3) (p=0.010).

- **Result Satisfaction [F(3,57)=5.48, p=0.002]** (Figure 5.1(d)). The pairwise comparison analysis revealed that the subjects felt more satisfied on the search results when they performed two easier tasks T1 (p=0.000) and T2 (p=0.003)than a harder task (T3). However, the subjects felt less satisfied with the search results when they perform T3 than the hardest task (T4) (p=0.005).

- **System General Feeling [F(3,57)=3.54, p=0.020]** (Figure 5.1(e)). The pairwise comparison analysis revealed that when subjects performed the easiest task T1, they tended to gave higher scores on System General Feeling than two harder tasks T2 (p=0.025) and T3 (p=0.003).

- **System Satisfaction [F(3,57)=4.66, p=0.006]** (Figure 5.1(f)). The pairwise comparison analysis revealed that the subjects were more satisfied with the system when they performed two easier tasks T1 (p=0.001) and T2 (p=0.034) than a harder task (T3). However, the subjects were more satisfied with the system when they performed the hardest task (T4) than T3 (p=0.008).

In summary, Figure 5.1 shows that the subjects tend to give higher scores to the performance indicators when they perform easier tasks, such as T1 and T2. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree T3 is harder than T1 and T2. However, they think that the hardest task (T4) is easier than T3 although the precision of the T3 is higher than the precision of T4 from our previous lab-based simulated experiment results. This may be because: first the colour of the three initial query image examples given in T3 is more complex than in T4, which makes the colour-based search difficult; second the given verbal search topic of T4 is easier to form a clear search goal than the topic of T3.

Hypothesis4: The interaction between Task and System will influence the scores of the performance indicators (8-33).

Search In Natural Way	**E2T1**	**E2T2**	**E2T3**	**E2T4**
E2OM1	2.5	3.25	3.25	3.5
E2OM2	3.75	3.25	3.75	3.25
E2OM3	4.25	4.5	3.25	2.75
E2OM4	4	2.75	3	3.5

Table 5.1: E2: Effects of the interaction between Task and System on performance indicators (8-33)

The factorial ANOVA results showed that there was no significant interaction between the effects of Task and System on most performance indicators. However, there was a significant interaction between Task and System on Search In Natural Way [F(9,48) = 2.34, p = 0.028], although the pairwise interaction comparison analysis did not reveal any significant difference. The interaction scores between Task and System are shown in Table 5.1.

Hypothesis5: Person will affect the performance indicators (8-33), based on individual differences.

The factorial ANOVA results showed that the differences between individual users (Person) significantly affected their scores on most performance indicators. The affected indicators were:

- Task General Feeling, F(15,42)=2.99, P=0.003;
- Enough Time, F(15,42)=2.85, P=0.004;
- Result Satisfaction, F(15,42)=2.54, P=0.009;
- Have Initial Idea, F(15,42)=2.16, p=0.025;
- Matched Initial Idea, F(15,42)=2.24, p=0.020;
- System General Feeling, F(15,42)=4.28, p=0.000;
- System Novelty, F(15,42)=7.79, p=0.000;
- Feel In Control, F(15,42)=2.54, p=0.009;
- Feel Comfortable, F(15,42)=5.41, p=0.000;
- Query History Easy To Use, F(15,42)=4.2, p=0.000;
- Query History Useful, F(15,42)=5.42, p=0.000;
- Query History Useful Here, F(15,42)=2.27, p=0.019;
- PQ Scoring Easy To Use, F(15,42)=3.63, p=0.000;
- PQ Scoring Useful, F(15,42)=4.65, p=0.000;
- PQ Scoring Useful Here, F(15,42)=3.07, p=0.002;
- N Query Easy To Use, F(15,42)=8.29, p=0.000;
- N Query Useful, F(15,42)=7.83, p=0.000;
- N Query Useful Here, F(15,42)=2.58, p=0.008;
- N Result Useful, F(15,42)=11.40, p=0.000;
- N Result Useful Here, F(15,42)=3.27, p=0.001;
- N Scoring As Useful As P Scoring, F(15,42)=2.43, p=0.012;
- System Satisfaction, F(15,42)=3.52, p=0.001;
- Know Collection, F(15,42)=7.03, p=0.000;
- Search In Natural Way, F(15,42)=3.92, p=0.000.

Thus, we could see that Person was another important factor which affected almost all of the performance indicators. Like we did in E1, we also took the subjects' age and image search experience into account in the following investigation on the

Person factor.

Hypothesis6: The subjects' Age and prior Image Search Experience of the subjects will affect subjects' opinion of the overall search experience (8-21).

The factorial ANOVA with covariate results showed that Age and Image Search Experience significantly affected the following performance indicators:

- **Task General Performance.** The interaction between Age and Image Search Experience affected the scores of Task General Performance [F(1,60) = 4.30, p = 0.042]. The resultant equations were: for inexperienced people, Task General Performance = 6.978 - 0.119*Age - 2.722*1 + 0.110*1*Age = 4.265 - 0.009*Age; for experienced people, Task General Performance = 6.978 - 0.119*Age - 2.722*2 + 0.110*2*Age = 1.543 + 0.101*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.009 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.101 per year as the age increased.

- **Enough Time.** Image Search Experience affected Enough Time [F(1,60) = 5.62, p = 0.021]. The interaction between Image Search Experience and Age also affected Enough Time [F(1,60) = 4.59, p = 0.036]. The resultant equations were: for inexperienced people, Enough Time = 7.682 - 0.118*Age - 3.497*1 + 0.120*1*Age = 4.185 + 0.002*Age; for experienced people, Enough Time = 7.682 - 0.118*Age - 3.497*2 + 0.120*2*Age = 0.688 + 0.122*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.002 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.122 per year as the age increased.

- **Next Action.** Image Search Experience affected Next Action [F(1,59) = 4.05, p = 0.049]. The interaction between Image Search Experience and Age also affected Next Action [F(1,59) = 4.19, p = 0.045]. The resultant equations were: for inexperienced people, Next Action = 6.956 - 0.124*Age - 2.750*1 + 0.106*1*Age = 4.206 - 0.018*Age; for experienced people, Next Action = 6.956 - 0.124*Age - 2.750*2 + 0.106*2*Age = 1.456 - 0.538*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.018 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.538 per year as the age increased.

- **Feel Comfortable.** Age affected Feel Comfortable [F(1,60) = 15.64, p = 0.000]. Image Search Experience also affected Feel Comfortable [F(1,60) = 16.11, p = 0.000], and so did their interaction (F(1,60)=18.07, p=0.000). The resultant equations were: for inexperienced people, Feel Comfortable = 8.806 - 0.209*Age - 4.217*1 + 0.170*1*Age = 4.589 - 0.039*Age; for experienced people, Feel Comfortable = 8.806 - 0.209*Age - 4.217*2 + 0.170*2*Age = 0.372 + 0.131*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.039 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.131 per year as the age increased.

- **Search In Natural Way.** Image Search Experience affected Search In Natural Way [F(1,60) = 5.18, p = 0.026]. The interaction between Image Search Experience and Age also affected Search In Natural Way [F(1,60) = 6.27, p = 0.015]. The resultant equations were: for inexperienced people, Search In Natural Way = 6.191 - 0.122*Age - 2.891*1 + 0.121*1*Age = 3.3 - 0.001*Age; for experienced people, Search In Natural Way = 6.191 - 0.122*Age - 2.891*2 + 0.121*2*Age = 0.409 + 0.12*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.001 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.12 per year as the age increased.

Hypothesis7: The subjects' Age and prior Image Search Experience of the subjects will have effects on the subjects' opinion on the functionalities

of the interfaces (22-33).

The factorial ANOVA with covariate results showed that Age and Image Search Experience significantly impinged on the following performance indicators:

- **Query History Easy To Use.** Image Search Experience affected Query History Easy To Use [F(1,61) = 9.64, p = 0.003]: if people had a higher level of image search experience then their scores on Query History Easy To Use were, on average, 0.565 higher.
- **Query History Useful.** Age affected Query History Useful [F(1,60) = 8.14, p = 0.006], Image Search Experience also affected Query History Useful [F(1,60) = 5.11, p = 0.027], and so did their interaction [F(1,60) = 8,21, p = 0.006]. The resultant equations were: for inexperienced people, Query History Useful = 7.406 - 0.148*Age - 3.705*1 + 0.141*1*Age = 3.701 - 0.007*Age; for experienced people, Query History Useful = 7.406 - 0.148*Age - 3.705*2 + 0.141*2*Age = -0.004 + 0.134*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.007 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.134 per year as the age increased.
- **PQ Scoring Easy To Use.** Image Search Experience affected PQ Scoring Easy To Use [F(1,61) = 5.40, p = 0.024]: if people had a higher level of image search experience then their scores on PQ Scoring Easy To Use were, on average, 0.473 higher.
- **N Query Easy To Use.** Image Search Experience affected N Query Easy To Use [F(1,61) = 5.93, p = 0.018]: if people had a higher level of image search experience then their scores on N Query Easy To Use were, on average, 0.597 higher.
- **N Query Useful.** Image Search Experience affected N Query Useful [F(1,60) = 7.29, p = 0.009]. The interaction between Image Search Experience and

Age also affected N Query Useful [F(1,60) = 8.40, P = 0.005]. The resultant equations were: for inexperienced people, N Query Useful = 5.890 - 0.118*Age - 3.614*1 + 0.147*1*Age = 2.276 + 0.029*Age; for experienced people, N Query Useful = 5.890 - 0.118*Age - 3.614*2 + 0.147*2*Age = -1.338 + 0.176*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.029 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.176 per year as the age increased.

- **N Query Useful Here.** Image Search Experience affected N Query Useful Here [F(1,60) = 7.94, p = 0.007]. The interaction between Image Search Experience and Age also affected N Query Useful Here [F(1,60) = 7.55, P = 0.008]. The resultant equations were: for inexperienced people, N Query Useful Here = 7.917 - 0.211*Age - 5.954*1 + 0.221*1*Age = 1.963 + 0.01*Age; for experienced people, N Query Useful Here = 7.917 - 0.211*Age - 5.954*2 + 0.221*2*Age = -3.991 + 0.231*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.01 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.231 per year as the age increased.

- **N Result Useful.** Age affected N Result Useful [F(1,61) = 11.51, p = 0.001]: the scores of N Result Useful increased by 0.080 per year increase in age. Image Search Experience affected N Result Useful [F(1,61) = 4.32, p = 0.042]: if people had a higher level of image search experience then their scores on N Result Useful were, on average, 0.490 higher.

- **N Result Useful Here.** Age affected N Result Useful Here [F(1,61) = 9.86, p = 0.003]: the scores of N Result Useful Here increased by 0.096 per year increase in age.

- **N Scoring As Useful As P Scoring.** Age affected N Scoring As Useful As P Scoring [F(1,60) = 6.60, p = 0.013]. Image Search Experience also affected N Scoring As Useful As P Scoring [F(1,60) = 9.01, p = 0.004], and so did their interaction [F(1,60) = 9.52, P = 0.003]. The resultant equations were: for inexperienced people, N Scoring As Useful As P Scoring = 6.992 - 0.181*Age - 4.204*1 + 0.164*1*Age = 2.788 - 0.017*Age; for experienced people, N Scoring As Useful As P Scoring = 6.992 - 0.181*Age - 4.204*2 + 0.164*2*Age = -1.488 + 0.147*Age. In other words, for the inexperienced people, the scores were estimated to decrease by 0.017 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.147 per year as the age increased.

Hypothesis8: System and Task will have an impact on Precision of the search results (34).

The factorial ANOVA results showed that there were no significant effects on the Precision (the precision calculation procedure is in Section 3.5) of the actual search results from System, Task or the interaction between System and Task.

Hypothesis9: System and Task will have an impact on Recall of the search results (35).

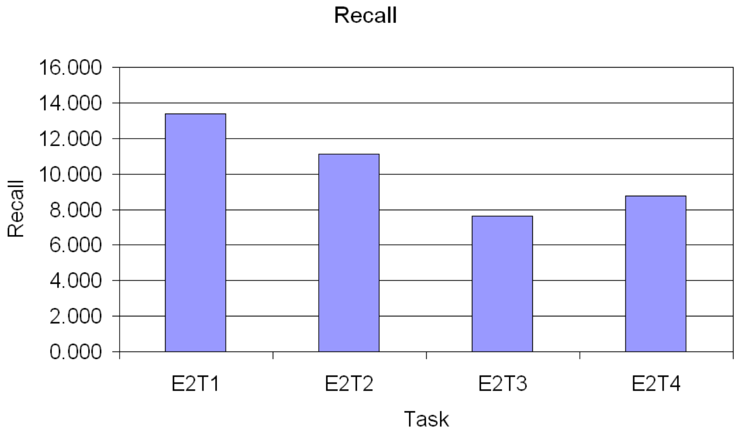


Figure 5.2: E2: Effects of the Task on Recall

The factorial ANOVA results in Figure 5.2 showed that Task significantly affected the Recall of the actual search results [F(3,57) = 3.29, p = 0.027]. The pairwise comparison analysis revealed that the Recall of T1 was higher than the Recall of

T3 (p=0.006) and T4 (p=0.024). There were no significant differences between the Recall of the search results for different systems.

5.3 Summary

In this Chapter, we reported the evaluation setup for evaluating the effects of the four profiles of the Ostensive Model and the usefulness of the uInteract interface (E2). We also reported the results obtained from the ANOVA analysis (with $\alpha = 0.05$) on the main performance indicators based on the nine hypotheses. The analysis results answered the research questions Q1.1, Q1.3, Q2.1 and Q2.2 addressed in Section 1.7.

We summarize the quantitative analysis results of E2 based on the nine hypothesis as below (Table 6.4):

- Hypothesis1 is not supported because System Order and Task Order do not affect the performance indicators at all. This implies that neither the familiarity or fatigue with the task nor the system make a difference to the subjects' scores on the indicators.

- Hypothesis2 is not supported because System does not influence the performance indicators at all, meaning there are no significant differences between the different systems.

- Hypothesis3 is partially supported because Task has a strong impact on most performance indicators, meaning the complexity level of a task does affect the subjects' opinions related to the performance indicators.

- Hypothesis4 is partially supported because the interaction between Task and System significantly affect the scores of Search In Natural Way, although there is no significant impact from System only.

- Hypothesis5 is partially supported because Person affects most performance indicators, implying different individuals have very different preferences.

- Hypothesis6 is partially supported because Age and Image Search Experience and their interaction significantly affect the subjects' opinions on the performance indicators of the search experience.

- Hypothesis7 is partially supported because Age and Image Search Experience and their interaction significantly affect the subjects' opinions on the functionalities of the interfaces.

- Hypothesis8 is not supported because Task or System or both do not affect the Precision of the search results.

- Hypothesis9 is partially supported because Task significantly affects the Recall of the search results, although there are no significant differences between the systems.

All in all, we conclude the Chapter by summarizing the key results for three main aspects: system, task and users. (1) It is difficult to identify the effects of the four profiles of the Ostensive Model because Task and Person factors strongly impinge on the scores of the performance indicators (Related to Q1.1 and Q2.1). (2) Task strongly influences the performance indicators. One interesting observation from our analysis of the results is that the subjects tend to give higher scores to the indicators when they perform easier tasks. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree that T3 is harder than T1 and T2. However, they think that T4 is easier than T3 although the precision of the T3 is higher than the precision of T4 from our previous lab-based simulated experiments. (3) Person is a very important factor which affects most performance indicators. There is clear evidence that users' age, image search experience and their interaction significantly affect the subjects' opinions on most performance indicators. The trend is for the inexperienced subjects, their satisfaction with the task performance, their opinion on feeling comfortable using the system and agreement on their natural search strategy supported by the systems, decrease as the age increases. However, for the experienced subjects, the scores on these indicators increase with the increase in age. In addition, for both the experienced and inexperienced subjects, the agreement on having enough time to complete the task increases with the increase in age, and the agreement on knowing next action decreases with the decrease in age. Further, the experienced subjects give higher scores on Query History Ease To Use, PQ Scoring Easy To Use, N Query Easy To Use and N Result Useful than inexperienced subjects. The scores on N Result Useful and N Result Useful Here increase with the increase in age. For the inexperienced subjects, the scores on Query History Useful and N Scoring As Useful As P Scoring decrease as the age increases; for the experienced subjects, the scores increase with the increase in age. The scores on N Query Useful and N Query Useful Here increase as the age increases for both experienced and inexperienced subjects (Related to Q1.3 and Q2.2).

Chapter 6

Evaluation of the Effects of the Four-factor User Interaction Model

In Chapter 5, we reported the second focused evaluation (E2) setup and results on the effects of the four profiles of the Ostensive Model (OM). This Chapter will report the third focused evaluation (E3) on the effects of the four-factor user interaction model (FFUIM). The goal of E3 is to test the effectiveness of the four settings of the FFUIM and whether users find the uInteract interface useful.

Section 6.1 reports the evaluation setup of E3. Section 6.2 reports the evaluation results of E3. The final finding will be stated in Section 6.3.

6.1 Evaluation Setup

The evaluation setup of E3 is similar with the setup of E2 (please refer to Section 5.1 and Appendix B). The only differences are that 17 subjects participated in E3 and the evaluation systems are different. The four testing systems[1] we used for E3 are:

[1] The four systems are delivered by the uInteract interface (Figure 2.4).

- System1 (FFUIM1) delivers the relevance region factor and time factor of the FFUIM [2], and here we apply the increasing profile [3] of the OM to both positive and negative queries;
- System2 (FFUIM2) delivers the relevance region factor, the time factor and relevance level factor of the FFUIM [4], and here we combine the increasing profile of the OM with the relevance scores provided by the users for both positive and negative queries;
- System3 (FFUIM3) delivers the relevance region factor and time factor and frequency factor of the FFUIM, and here we combine the increasing profile of the OM with the number of times (frequency) images appeared in the feedback for both positive and negative queries;
- system4 (FFUIM4) delivers the relevance region factor, time factor, relevance level factor and frequency factor of the FFUIM, and here we combine the increasing profile of the OM and the relevance scores provided by the users and the number of times (frequency) images appeared in the feedback for both positive and negative queries.

6.2 Evaluation Results and Analysis

The following results were obtained by applying ANOVA analysis (with $\alpha = 0.05$) on the experimental results in terms of the main performance indicators (introduced in Table 3.1). We have broken down our main goal of this evaluation into nine hypotheses in Section 3.4. The results analysis will focus on the individual hypotheses.

[2] This setting of the FFUIM is based on the Urban et al. (2006) Ostensive Model. The differences are that we use both positive and negative feedback and multi-image query here.

[3] We apply the increasing profile of the Ostensive Model here because (1) it is one of the best performing profile based on our experimental results; (2) it is the only one profile widely applied in related work.

[4] This setting of the FFUIM is based on the Ruthven et al. (2003) interaction model. The differences are that we apply the model to content-based image search and we allow both positive and negative feedback here.

Only statistically significant results will be listed.

Hypothesis1: Task Order and System Order will affect the performance indicators (8-33) provided by subjects because of familiarity or fatigue.

The factorial ANOVA results showed that the task position and system position did not significantly affect the scores provided by subjects on all the performance indicators. There was no significant interaction between the effects of task and system position on all the performance indicators.

Hypothesis2: System will affect the performance indicators (8-33).

The factorial ANOVA results showed that Systems had no significant effects on any of the performance indicators.

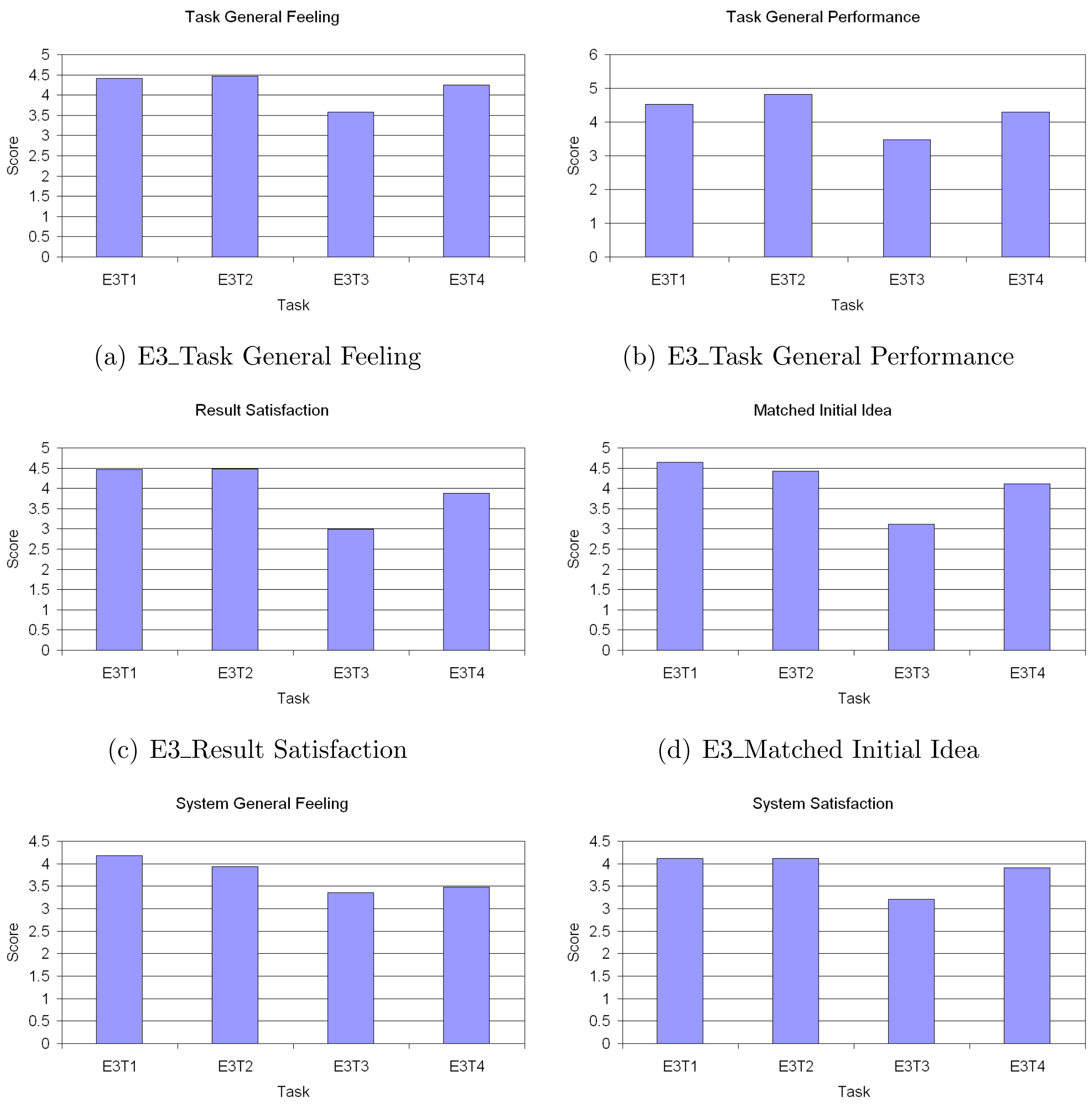
Hypothesis3: Task will affect the performance indicators (8-33) provided by subjects because of different complexity levels.

The factorial ANOVA results showed that the different complexity level of tasks significantly impinged on the following performance indicators (Figure 6.1):

- **Task General Feeling [F(3,61)=5.86, p=0.001]** (Figure 6.1(a). The pairwise comparison analysis revealed that the subjects performing two easier tasks T1 (p=0.001) and T2 (p=0.000) gave higher scores on Task General Feeling than a harder task (T3). However, the scores of T3 were worse than the hardest task (T4) (p=0.006).

- **Task General Performance [F(3,61)=9.58, p=0.000]** (Figure 6.1(b). The pairwise comparison analysis revealed that the subjects performing two easier tasks T1 (p=0.000) and T2 (p=0.000) gave higher scores on Task General Performance than a harder task (T3). However, T3 performed worse than the hardest task (T4) (p=0.003).

- **Result Satisfaction [F(3,61)=19.92, p=0.000]** (Figure 6.1(c). The pairwise comparison analysis revealed that the subjects felt more satisfied on the



(a) E3_Task General Feeling (b) E3_Task General Performance

(c) E3_Result Satisfaction (d) E3_Matched Initial Idea

(e) E3_System General Feeling (f) E3_System Satisfaction

Figure 6.1: E3: Effects of Task on performance indicators (8-33)

search results when they performed two easier tasks T1 (p=0.000, p=0.010) and T2 (p=0.000, p=0.008) than the two harder tasks (T3) and (T4). However, the subjects felt less satisfied with the search results of T3 than those for the hardest task (T4) (p=0.000).

- **Matched Initial Idea [F(3,61)=12.81, p=0.000]** (Figure 6.1(d). The pairwise comparison analysis revealed that the search results of T1 (p=0.000) and T2 (p=0.000) better matched the subjects' initial idea than a harder task (T3). However, the search results of the hardest task (T4) better matched the subjects initial idea than the search results of T3 (p=0.000).

- **System General Feeling [F(3,61)=5.34, p=0.002]** (Figure 6.1(e). The pairwise comparison analysis revealed that when subjects performed the easiest tasks (T1), they tended to gave higher scores on System General Feeling than when they performed two harder tasks T3 (p=0.001) and T4 (p=0.004).
- **System Satisfaction [F(3,61)=2.99, p=0.038]** (Figure 6.1(f). The pairwise comparison analysis revealed that the subjects were more satisfied with the system when they performed two easier tasks T1 (p=0.013) and T2 (p=0.012) rather than a harder task (T3).

In summary, Figure 6.1 shows that the subjects tend to give higher scores to the performance indicators when they perform easier tasks, such as T1 and T2. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree that T3 is harder than T1 and T2. However, they think the hardest task (T4) is easier than T3, although the precision of the T3 is higher than the precision of T4 from our previous lab-based simulated experiment results. This may be because: first the colour of the three initial query image examples given in T3 is more complex than in T4, which makes the colour-based search difficult; second the given verbal search topic of T4 is easier to form a clear search goal than the topic of T3.

Hypothesis4: The interaction between Task and System will influence the scores of the performance indicators (8-33).

Feel Comfortable	**E3T1**	**E3T2**	**E3T3**	**E3T4**
E3FFUIM1	4.4	4.75	3.5	4.25
E3FFUIM2	4.75	4.25	3.4	4.25
E3FFUIM3	5	3.5	2.75	4
E3FFUIM4	4.25	3.6	4.75	4.25

Table 6.1: E3: Effects of the interaction between Task and System on performance indicators (8-33)

The factorial ANOVA results showed that there was no significant interaction between the effects of Task and System on most performance indicators. However,

N Query Useful	**E3T1**	**E3T2**	**E3T3**	**E3T4**
E3FFUIM1	1.8	3.25	2.75	4.75
E3FFUIM2	2.5	2.25	4.4	3.5
E3FFUIM3	2.25	4.25	1.75	2.4
E3FFUIM4	4.5	3.6	2.75	2

Table 6.2: E3: Effects of the interaction between Task and System on performance indicators (8-33)

there was a significant interaction between Task and System on Feel Comfortable [F(9,52) = 2.58, p = 0.015], and on N Query Useful [F(9,52) = 3.23, p = 0.003], although the pairwise interaction comparison analysis did not reveal any significant difference. The interaction scores between Task and System are shown in Table 6.1 and Table 6.2 respectively.

Hypothesis5: Person will affect the performance indicators (8-33), based on individual differences.

The factorial ANOVA results showed that the differences between individual users (Person) significantly affected their scores on most performance indicators. The affected indicators were:

- Enough Time, F(16,45)=3.18, P=0.001;
- Next Action, F(16,45)=2.60, p=0.006;
- Have Initial Idea, F(16,45)=2.97, p=0.002;
- System General Feeling, F(16,45)=2.19, p=0.020;
- System Novelty, F(16,45)=18.56, p=0.000;
- Feel In Control, F(16,45)=1.99, p=0.035;
- Feel Comfortable, F(16,45)=3.09, p=0.001;
- Query History Easy To Use, F(16,45)=13.33, p=0.000;
- Query History Useful, F(16,45)=4.67, p=0.000;
- Query History Useful Here, F(16,45)=6.34, p=0.000;
- PQ Scoring Easy To Use, F(16,45)=4.54, p=0.000;
- PQ Scoring Useful, F(16,45)=3.83, p=0.000;

- PQ Scoring Useful Here, F(16,45)=1.94, p=0.042;
- N Query Easy To Use, F(16,45)=13.76, p=0.000;
- N Query Useful, F(16,45)=11.75, p=0.000;
- N Query Useful Here, F(16,45)=10.03, p=0.000;
- N Result Useful, F(16,45)=9.14, p=0.000;
- N Result Useful Here, F(16,45)=3.69, p=0.000;
- N Scoring As Useful As P Scoring, F(16,45)=10.36, p=0.000;
- System Satisfaction, F(16,45)=2.93, p=0.002;
- Know Collection, F(16,45)=3.99, p=0.000;
- Search In Natural Way, F(16,45)=8.84, p=0.000.

Thus, we could see that Person was another important factor which affected almost all of the performance indicators. As we did in E1 and E2, we took the subjects' age and image search experience into account in the following investigation on the Person factor.

Hypothesis6: The subjects' Age and prior Image Search Experience of the subjects will affect subjects' opinion of the overall search experience (8-21).

The factorial ANOVA with covariate results showed that Age and Image Search Experience significantly affected the following performance indicators:

- **Task General Performance.** Image Search Experience affected Task General Performance [F(1,65) = 4.52, p = 0.037]: if people had a higher level of image search experience then their scores on Task General Performance were, on average, 0.472 lower.
- **Enough Time.** Age affected Enough Time [F(1,65) = 8.87, p = 0.004]: the Enough Time scores increased by 0.038 per year increase in age.
- **System General Feeling.** Image Search Experience affected System General Feeling [F(1,64) = 5.34, p = 0.024], and the interaction between Image Search

Experience and Age [F(1,64) = 5.17, p = 0.026]. The resultant equations were: for inexperienced people, System General Feeling = 6.222 - 0.087*Age - 1.561*1 + 0.055*1*Age = 4.661 + 0.032*Age; for experienced people, System General Feeling = 6.222 - 0.087*Age - 1.561*2 + 0.055*2*Age = 3.1 + 0.023*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.032 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.023 per year as the age increased.

- **System Novelty.** Age affected System Novelty [F(1,65) = 45.40, p = 0.000]: the System Novelty scores increased by 0.063 per year with increase in age.
- **Feel Comfortable.** Age affected Feel Comfortable [F(1,65) = 8.82, p = 0.004]: the scores of Feel Comfortable increased by 0.030 per year increase in age.
- **System Satisfaction.** Age affected System Satisfaction [F(1,65) = 10.20, p = 0.002]: the scores of System Satisfaction increased by 0.039 per year increase in age.
- **Know Collection.** Age affected Know Collection [F(1,65) = 15.07, p = 0.000]: the scores of Know Collection increased by 0.048 per year increase in age. Image Search Experience also affected Know Collection [F(1,65) = 7.82, p = 0.007]: if people had higher level of image search experience then their scores of Know Collection were, on average, 0.714 lower.
- **Search In Natural Way.** Age affected Search In Natural Way [F(1,64) = 6.73, p = 0.012]. Image Search Experience also affected Search In Natural Way [F(1,64) = 10.83, p = 0.002], and so did their interaction [F(1,64) = 9.46, p = 0.003]. The resultant equations were: for inexperienced people, Search In Natural Way = 8.458 - 0.165*Age - 3.113*1 + 0.104*1*Age = 5.345 - 0.061*Age; for experienced people, Search In Natural Way = 8.458 - 0.165*Age - 3.113*2 + 0.104*2*Age = 2.232 + 0.043*Age. In other words, for the inexperienced

people, the scores were estimated to decrease by 0.061 per year as the age increased; for the experienced people, the scores were estimated to increase by 0.043 per year as the age increased.

Hypothesis7: The subjects' Age and prior Image Search Experience of the subjects will have effects on the subjects' opinion on the functionalities of the interfaces (22-33).

The factorial ANOVA with covariate results showed that Age and Image Search Experience significantly impinged on the following performance indicators:

- **Query History Useful.** Image Search Experience affected Query History Useful [F(1,65) = 13.60, p = 0.000]: if people had a higher level of image search experience then their scores on Query History Useful were, on average, 1.232 lower.

- **Query History Useful Here.** Age affected Query History Useful Here [F(1,65) = 16.70, p = 0.000]: the scores of Query History Useful Here increased by 0.080 per year increase in age. Image Search Experience also affected Query History Useful Here [F(1,65) = 41.29, p = 0.000]: if people had a higher level of image search experience then their scores on Query History Useful Here were, on average, 2.619 lower.

- **N Query Easy To Use.** Age affected N Query Easy To Use [F(1,64) = 3.99, p = 0.050], and the interaction between Age and Image Search Experience [F(1,64) = 5.33, p = 0.024]. The resultant equations were: for inexperienced people, N Query Easy To Use = 0.280 + 0.163*Age + 2.158*1 - 0.100*1*Age = 2.438 + 0.063*Age; for experienced people, N Query Easy To Use = 0.280 + 0.163*Age + 2.158*2 - 0.100*2*Age = 4.596 - 0.037*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.063 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.037 per year as the age increased.

- **N Query Useful.** Age affected N Query Useful [F(1,64) = 8.12, p = 0.006]. Image Search Experience also affected N Query Useful [F(1,64) = 5.36, p = 0.024], and so did their interaction [F(1,64) = 9.23, P = 0.003]. The resultant equations were: for inexperienced people, N Query Useful = -2.669 + 0.263*Age + 3.184*1 - 0.149*1*Age = 0.515 + 0.114*Age; for experienced people, N Query Useful = -2.669 + 0.263*Age + 3.184*2 - 0.149*2*Age = 3.699 - 0.035*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.114 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.035 per year as the age increased.

- **N Query Useful Here.** Age affected N Query Useful Here [F(1,64) = 8.31, p = 0.005]. Image Search Experience also affected N Query Useful Here [F(1,64) = 5.96, p = 0.017], and so did their interaction [F(1,64) = 9.48, P = 0.003]. The resultant equations were: for inexperienced people, N Query Useful Here = -4.977 + 0.334*Age + 4.215*1 - 0.190*1*Age = -0.762 + 0.144*Age; for experienced people, N Query Useful Here = -4.977 + 0.334*Age + 4.215*2 - 0.190*2*Age = 3.453 - 0.046*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.144 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.046 per year as the age increased.

- **N Result Useful.** Age affected N Result Useful [F(1,64) = 4.30, p = 0.042]. The interaction between Age and Image Search Experience also affected N Result Useful [F(1,64) = 5.34, p = 0.024]. The resultant equations were: for inexperienced people, N Result Useful = -0.594 + 0.163*Age + 1.888*1 - 0.097*1*Age = 1.294 + 0.066*Age; for experienced people, N Result Useful = -0.594 + 0.163*Age + 1.888*2 - 0.097*2*Age = 3.182 - 0.031*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.066 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.031 per year as the age increased.

- **N Result Useful Here.** Age affected N Result Useful Here [F(1,64) = 5.25, p = 0.025]. The interaction between Age and Image Search Experience also affected N Result Useful Here [F(1,64) = 5.58, p = 0.021]. The resultant equations were: for inexperienced people, N Result Useful Here = -2.045 + 0.202*Age + 2.424*1 - 0.111*1*Age = 0.379 + 0.091*Age; for experienced people, N Result Useful Here = -2.045 + 0.202*Age + 2.424*2 - 0.111*2*Age = 2.803 - 0.02*Age. In other words, for the inexperienced people, the scores were estimated to increase by 0.091 per year as the age increased; for the experienced people, the scores were estimated to decrease by 0.02 per year as the age increased.

- **PQ Scoring Useful.** Image Search Experience affected PQ Scoring Useful [F(1,65) = 3.99, p = 0.050]: if people had a higher level of image search experience then their scores on PQ Scoring Useful were, on average, 0.570 lower.

- **N Scoring As Useful As P Scoring.** Age affected N Scoring As Useful As P Scoring [F(1,65) = 6.43, p = 0.014]: the scores of N Scoring As Useful As P Scoring increased by 0.036 per year increase in age.

Hypothesis8: System and Task will have an impact on Precision of the search results (34).

Precision	**E3T1**	**E3T2**	**E3T3**	**E3T4**
E3FFUIM1	3.086	2.788	2.315	3.033
E3FFUIM2	3.555	3.175	2.946	2.685
E3FFUIM3	2.703	2.973	3.318	3.262
E3FFUIM4	3.07	3.488	3.28	3.12

Table 6.3: E3: Effects of the interaction between Task and System on Precision

The factorial ANOVA results in Figure 6.2 showed that there were significant differences between the Precision (the precision calculation procedure is in Section 3.5) of the search results for different systems [F(3,52) = 2.80, p = 0.049]. The pairwise

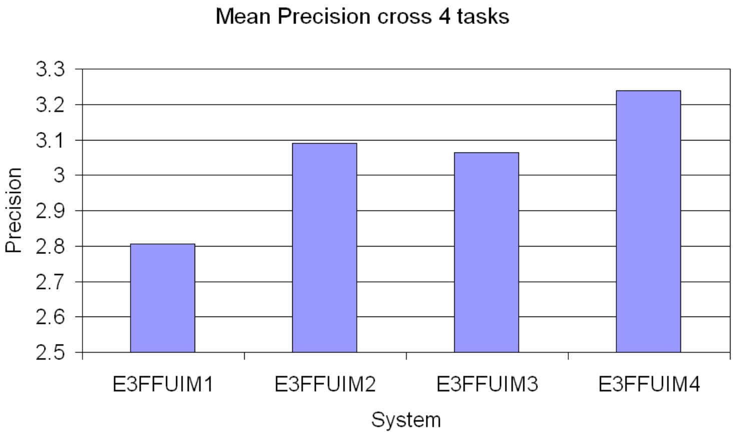


Figure 6.2: E3: Effects of the System on Precision

comparison analysis revealed that the Precision of the FFUIM4 (a combination of relevance region, relevance level, time and frequency factor) was significantly higher than the FFUIM1 (a combination of relevance region and time factor), p=0.006.

There was also significant interaction between System and Task on Precision (Table 6.3) [F(9,52) = 2.51, p = 0.018]. The pairwise comparison analysis revealed that the Precision of the FFUIM4 was the best, and then it was followed by FFUIM2 (a combination of relevance region, relevance level and time factor), FFUIM3 (a combination of relevance region, frequency and time factor), FFUIM1. The mean average precision of the FFUIM1, FFUIM2, FFUIM3, FFUIM4 across the four tasks was 2.81, 3.09, 3.06, 3.24.

Hypothesis9: System and Task will have an impact on Recall of the search results (35).

The factorial ANOVA results shows that Task significantly affected the Recall of the search results [F(3,61) = 3.26, p = 0.027]. The pairwise comparison analysis revealed that the Recall of T2 was higher than the Recall of T3 (p=0.008) and T4 (p=0.013). There were no significant differences between the Recall of the search results for the different systems.

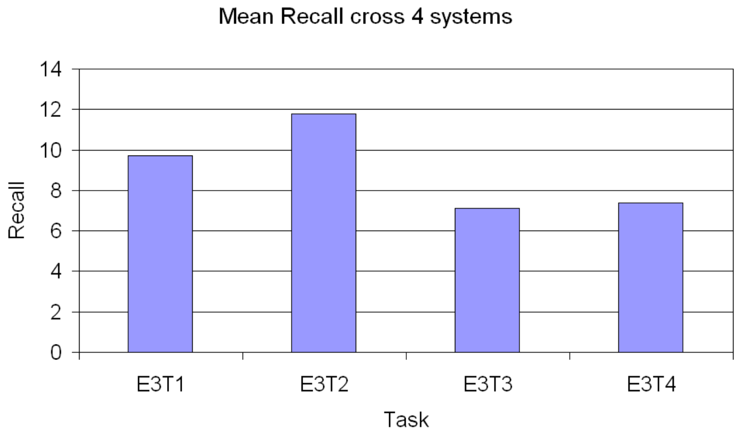


Figure 6.3: E3: Effects of the Task on Recall

6.3 Summary

Hypotheses	**E1**	**E2**	**E3**
Hypothesis1: Task Order and System Order will affect the performance indicators (8-33) provided by subjects because of familiarity or fatigue	Not supported	Not supported	Not supported
Hypothesis2: System will affect the performance indicators (8-33)	Not supported	Not supported	Not supported
Hypothesis3: Task will affect the performance indicators (8-33) provided by subjects because of different complexity levels	Partially supported	Partially supported	Partially supported
Hypothesis4: The interaction between Task and System will influence the scores of the performance indicators (8-33)	Partially supported	Partially supported	Partially supported
Hypothesis5: Person will affect the performance indicators (8-33), based on individual differences	Partially supported	Partially supported	Partially supported
Hypothesis6: The subjects' Age and prior Image Search Experience of the subjects will affect subjects' opinion of the overall search experience (8-21)	Partially supported	Partially supported	Partially supported
Hypothesis7: The subjects' Age and prior Image Search Experience of the subjects will have subjects' effects on the opinion on the functionalities of the interfaces (22-33)	Partially supported	Partially supported	Partially supported
Hypothesis8: System and Task will have an impact on Precision of the search results (34)	Partially supported	Not supported	Partially supported
Hypothesis9: System and Task will have an impact on Recall of the search results (35)	Partially supported	Partially supported	Partially supported

Table 6.4: How the nine hypotheses have been supported or rejected in E1, E2 and E3

In this Chapter, we reported the evaluation setup for evaluating the effects of the four settings of the four-factor user interaction model and the usefulness of the uInteract interface (E3). We also reported the results obtained from the ANOVA analysis (with $\alpha = 0.05$) on the main performance indicators based on the nine hypotheses. The analysis results answered the research questions Q1.2, Q1.3, Q2.1

and Q2.2 addressed in Section 1.7.

We summarize the quantitative analysis results of E3 based on the nine hypothesis as below (Table 6.4):

- Hypothesis1 is not supported because System Order and Task Order do not affect the performance indicators at all. This implies the familiarity or fatigue with the task and the system does not make a difference to the subjects' scores on the indicators.

- Hypothesis2 is not supported because System does not influence the performance indicators at all, meaning there is no significant differences between different systems.

- Hypothesis3 is partially supported because Task has a strong impact on most performance indicators, meaning the complexity level of a task does affect the subjects' opinions related to the performance indicators.

- Hypothesis4 is partially supported because the interaction between Task and System significantly affect the scores of Feel Comfortable and N Query Useful, although there is no significant impact from System only.

- Hypothesis5 is partially supported because Person affects most performance indicators, implying different individuals have very different preferences.

- Hypothesis6 is partially supported because Age and Image Search Experience and their interaction significantly affect the subjects' opinions on the performance indicators of the search experience.

- Hypothesis7 is partially supported because Age and Image Search Experience and their interaction significantly affect the subjects' opinions on the functionalities of the interfaces.

- Hypothesis8 is partially supported because System and the interaction between Task and System significantly affect the Precision of the search results.

- Hypothesis9 is partially supported because Task significantly affects the Recall of the search results, although there are no significant differences between systems.

All in all, we conclude the Chapter by summarizing key results from three main aspects: system, task and users. (1) It is difficult to identify the effects of the four settings of the four-factor user interaction model because Task and Person factors strongly impinge on the scores of the performance indicators. However, System significantly affects the Precision and the pairwise comparison results shows that FFUIM4 (combination of relevance region, relevance level, time and frequency) outperforms FFUIM1 (combination of relevance region and time) (Related to Q1.2 and Q2.1). (2) Task strongly influences the performance indicators. One interesting observation is that the subjects tend to give higher scores to the indicators when they perform easier tasks. In most cases, the subjects' perception of task difficulty is not the same as the difficulty level we had intended. They agree that T3 is harder than T1 and T2. However, they think that T4 is easier than T3, although the precision of the T3 is higher than the precision of T4 from our previous lab-based simulated experiments. (3) Person is a very important factor which affects most performance indicators. There is clear evidence that age, image search experience and their interaction significantly affect the subjects' opinions on most performance indicators. The trend is that the experienced subjects give lower scores on general task performance and knowing data quality supported by the system than the inexperienced subjects. The scores on enough time to complete the tasks, system novelty, feeling comfortable using the systems, satisfaction by the systems and knowing the data quality supported by the system increase with increase in age. Further, for the inexperienced subjects, the opinion on their natural search strategy being supported is decreased as the age increases; for the experienced subjects, the scores on this factor are increased with the increase in age. In addition, for both the experienced and inexperienced subjects, the scores on general feeling about the systems increase with the increase in age. Moreover, the experienced subjects give higher scores on

Query History Useful, Query History Useful Here and PQ Scoring Useful than inexperienced subjects. The scores on Query History Useful Here and N Scoring As Useful As P Scoring increase with the increase in age. Further, for the inexperienced subjects, the scores on N Query Easy To Use, N Query Useful, N Query Useful Here, N Result Useful and N Result Useful Here increase as the age increases; for the experienced subjects, the scores decrease with the increase in age (Related to Q1.3 and Q2.2).

Chapter 7

ISE: A User Classification Model based on Information Goals (I), Search Strategies (S) and Evaluation Thresholds (E)

Chapters 4, 5 and 6 suggested that there was no significant difference between systems, but there was a strong and significant impact of the Task and Person indicators. The quantitative data analysis results of E1, E2 and E3 showed the same trend on the influence from the Task indicator: the subjects tended to give higher scores to the performance indicators when they performed an easier task, and the subjects had different opinions on the complexity level of the tasks. However, the quantitative data analysis results did not show the trend on how Person indicator affected the scores of the performance indicators, although we further tested the effects of Age and Image Search Experience of the subjects. Therefore, we realize that the simple user classification based on Age and Image Search Experiences is not sufficient, and we need to investigate in-depth how to better classify user types and how the Person indicator impinges on the users' preferences and search behaviours.

This Chapter proposes an ISE (Information goal, Search strategy, Evaluation threshold) user classification model, based on Information Foraging Theory (Pirolli and Card 1995; Pirolli 2007), for understanding user interaction with content-based image retrieval (CBIR), to tackle the last key element "users" of the user interaction. The proposed ISE model is verified by a multiple linear regression analysis based on 50 users' qualitative data collected from the extensive task-based user evaluations reported in the previous chapters. To the best of our knowledge, this proposed model is the first principled user classification model in CBIR verified by a formal systematic data analysis based on extensive user interaction data from a real interactive image search scenario.

We will firstly introduce the background knowledge on Information Foraging Theory (Section 7.1), and then present the proposed ISE user classification model in Section 7.2. The ISE model will be operationalized and verified by statistical multiple linear regression analysis in Section 7.3. Section 7.4 will conclude the chapter.

7.1 Information Foraging Theory (IFT)

To provide adaptive strategies for information foraging in a complex information environment, Pirolli and Card (1995) proposed Information Foraging Theory, which aims "*to explain and predict how people will best shape themselves for their information environments and how information environments can best be shaped for people.*" (P.3) (Pirolli 2007). The methodology of the Information Foraging Theory is adapted from the framework of optimal foraging theory in biology (Stephens and Krebs 1986).

The optimal foraging theory was developed to explain food seeking and prey selection behaviours among animals. Consider a hypothetical predator, such as a bird of prey. The environment surrounding this bird will have a patchy structure, with different types of habitat and different kinds and amounts of prey. Thus, the bird needs to find the best solution to catch more food per unit energy cost within the environment constraints. Stephens and Krebs (1986) introduced two conventional models: (a) the patch model, which addresses decisions related to searching and exploiting an environment that has a patchy distribution of resources, and (b) the diet model, which addresses what kind of things to eat and what to ignore.

Pirolli and Card (1999) adapted two conventional models from the optimal foraging theory (Stephens and Krebs 1986), originally applied to the food hunting environment, to the information seeking environment. Further, they proposed three information models for IFT: information patch model, information diet model and information scent model, which will be explained in detail in the next subsections.

7.1.1 The Information Patch Model

The aim of the information patch model is to predict the amount of time a forager would forage within an information patch or searching for new patches when the information forager deals with information that is distributed in a patchy manner. For instance, there are a variety of information items on my work desk, such as books, printed papers, notes, electronic files in my computer and an internet connection. Some of these items are located in within arm's reach of my desk, and some items are stored on the bookcase or in the filing cabinet. The relevant information to my current task can be found on the desk and on the bookcase. If I identify the arms' reachable area is one patch and the bookcase as another patch, my information foraging process will be within-patch and between-patch activities. I will need to decide whether I stay longer in the arms' reachable patch to look for relevant information or I should go to dig the information from the books on the bookcase. The decision will be made depending on the prevalence and profitability of the patches. A higher prevalence of patches may contain many relevant items to the task, and a higher profitability patch may contain the most relevant information to the task. All in all, the decision to do within-patch or between-patch activity or a bit of both will be based on finding the most relevant information to complete the task in the shortest time.

Unlike the conventional patch model, the information patch model deals with a mouldable environment. The information forager can modify the environment to fit the available strategy. This process is called enrichment. The first kind of enrichment is to reduce the cost of between-patch activity. For example, I can reorganize the the desk, the book case and the filing cabinet to make the access easier for completing the task by moving them closer to each other. The second kind of enrichment is to improve the within-patch activity. For instance, within my arms' reachable patch, I can reorganize the items on the desk based on information category rather than based on item size or issue time to make the access easier for completing the task.

7.1.2 The Information Diet Model

The question that the conventional diet model deals with is: when a predator lives in an environment containing a number of potential kinds of food sources, what kinds of things should the predators prey on, and what kinds of things should they ignore? One way to answer this question is in terms of diet concept: a generalized diet includes a broad type of prey, but a specialized diet includes only a few types. "*If a predator is too specialized, it will do very narrow searching. If the predator is too generalized, then it will pursue too much unprofitable prey (p.39)* (Pirolli 2007)." Thus, the diet model in Information Foraging Theory can be explained in terms of the conventional diet model: if I have a generalized diet, I will complete the task with a wide range of relevant information with diverse dimensions; if I have a specialized diet, I will complete the task with only a few relevant information sources focusing on one dimension.

The Scatter/Gather browser is a cluster-based retrieval tool on large text collections, which was used for demonstrating the information patch and information diet models of IFT (Pirolli and Card 1995). Later Pirolli et al. (1996) found the tool was more useful in supporting exploratory search activities than searching with a specific goal. They suggested to use the theoretical models of IFT to evaluate information access and search behaviours. Further, Pirolli (1997) introduced the notion of information scent as "*Users must rely on such terse representations of content as a kind of information scent whose trail leads to information of interest (p.1).*", while analyzing the data from the user interacting with the Scatter/Gather browser.

7.1.3 The Information Scent Model

The information scent model is a psychological theory, which explains how people identify the value of the information based on cues such as result clusters on the interface in order to gain an overall sense of the contents of information collections. If the scent is strong, the forager will be able to move fairly directly. If there is no scent, the forager will perform a random walk (Pirolli and Card 1999; Pirolli et al. 2005).

As the most popular concept of IFT, the information scent model has been applied to investigate effective information scent cues in aiding navigation. For instance, Chi et al. (2001) proposed two computational methods for modeling users' information needs and actions on the web, based on the concept of information scent. The first situation is to predict users' surfing pattern given users' information needs. The second situation is to infer a user's information needs given a user's particular pattern of surfing. Their general finding is that the two models will help researchers better understand the usage of the Web, help the design of better web sites, and make users' information seeking activities more efficient. Pirolli et al. (2003) compared the performance of the Hyperbolic Tree Browser and the Microsoft Windows File Browser. Their finding suggested that a good cue was not always good. Whether a cue (navigation/presentation) is good depends on how the cue matches the information goal of users. Another general finding was that an interface with good information scent would improve usability.

7.1.4 Applications of IFT in Information Seeking

Information Foraging Theory (IFT) has also been suggested and applied to dealing with problems in human-information interaction for understanding of information-seeking behaviours and guiding new designs (White et al. 2007; Käki 2005; Marchionini and White 2009; Mulholland et al. 2008; Pirolli 2007).

White et al. (2007) proposed a new information seeking paradigm - exploratory search. They propose that exploratory search can be explained and supported by Information Foraging Theory in some respects, such as users searching for information to meet their information goals, the impact from users' searching behaviours and an optimal path/navigation leading to their goal to be achieved, and the knowledge gained during the search process. Marchionini and White (2009) identified exploratory search as an information seeking process that includes recognizing the need, accepting the problem, formulating the problem, expressing the need, examining results, reformulating the problem and transition to use. They recommended the use of Information Foraging Theory to model these sub-processes because they consider the theory is highly adaptive to the information environment. Mulholland et al. (2008) analyzed users' exploratory search behaviours in Semantic Web data based on Information Foraging Theory. They successfully identified two types of search strategy: a risky strategy and a cautious strategy, based on their qualitative data analysis. They also suggested that the findings will have implications for the intelligent scaffolding of exploratory search. Kules and Shneiderman (2008) also proposed a set of guidelines for the design of exploratory search interfaces drawn on their qualitative data analysis. They find that some participants explore categories instead of providing a new query. The finding is consistent with the concept of the information scent model of IFT.

Nielsen (2003) suggested that information foraging is the most important concept in the human-computer interaction field. He finds that Web users behave like wild animals in the jungle based on the three information models of the IFT. Ivory et al.

(2004) investigated sighted and blind users' decision-making behaviours and performance during the search process. Their finding is consistent with the basic concept of Information Foraging Theory in that foragers attempt to maximize the benefit with minimum costs. Berendt and Kralisch (2009) applied Information Foraging Theory to analyze users' behaviours and attitudes when using multilingual tools online. A finding also suggested that users' decision making depends heavily on whether the action is worth the effort. Käki (2005) proposed two enhanced result categorization algorithms for text search systems, effective interfaces for delivering these algorithms, and user-oriented evaluation methodologies (normalized search speed measure, qualified search speed measure and immediate accuracy measure), which were partially motivated by the information patch model and Scatter/Gather browser.

7.1.5 Task and Information Environment and Forager

In the studies mentioned above, the models and concepts of IFT have been applied to improve the design perspective for interactive search and in turn to improve users search experience. However, there is lack of research on applying IFT to understand user interaction based on the users' perspective.

Pirolli (2007)(p.20) stated that to understand information foraging requires analysis of the environment and analysis of the forager. The two interrelated environments during an information search process are the task environment and the information environment. The definition of the task environment "*refers to an environment coupled with a goal, problem or task - the one for which the motivation of the subject is assumed*". "*The information environment is a tributary of knowledge that permits people to more adaptively engage their task environments.*" In other words, "*What we know, or do not know, affects how well we function in the important task environments that we face in life.*". Our understanding of the task and information environments is that they should be part of the information scent concept from a forager's point of view. A clear task environment and a rich information environment will determine a forager's strong information scent (goal). A forager with strong information scent (goal) should find the right information resource quicker with the support of a well designed interface. Moreover, different forager types and the same type of forager within different environments will show very different search strategies and behaviours.

The above discussion corresponds with our findings from the quantitative data analysis results in Chapter 4, Chapter 5 and Chapter 6. Task complexity and user characteristics significantly influence the evaluation results. Based on the experience from our task-based user study and motivated by Mulholland et al. (2008)'s findings, we consider users can be classified into different user types based on their profile, and the users within the same user type have similar search preferences and search behaviours. We then decide to undertake an in depth investigation into how many different user types we can identify and what search preferences and behaviours each user type has, based on Information Foraging Theory and qualitative analysis of the real user interaction data in interactive CBIR.

7.2 Definition of the ISE Model

In this Section, and based on Information Foraging Theory (IFT), we propose a new user classification model, called ISE model. The model includes three criteria: information goals (I), search strategies (S) and evaluation thresholds (E). Each criterion categorizes users into two types based on two different user characters[1]: I - fixed information goal or evolving information goal; S - risky search strategy or cautious search strategy; E - weak evaluation threshold or precise evaluation threshold. We take our user study described in Chapter 3 as an example to explain how we map the concepts between Information Foraging Theory and the ISE model.

[1] There are in total six characters in the ISE model.

7.2.1 Information Goal

The information goal can be explained by the information scent model of IFT.

After reading the task description, the searchers may or may not have a clear information goal (i.e., an idea of what they are looking for) to start the search. In IFT terms, the searchers might or might not get a strong information scent from reading the task based on their information environment (knowledge). Thus, the searchers can be categorized into two types based on the information scent concepts: one type with fixed information goal and the other with evolving information goal. According to the information scent concept, if the searchers have a fixed information goal, they will focus on what they are looking for and likely make consistent decisions at every stage. On the other hand, if the searchers have an evolving information goal, their search will be more exploratory. They will randomly walk about and learn from the data before they make a decision although the decision might not be a correct one.

The assumptions of the fixed and evolving information goal based on interactive image search scenario are: (1) the searchers with evolving goals will be likely to perform trial and error types of search so that it will take them longer to find the best result image for completing search tasks. For example, they will reformulate queries with completely different image examples, and they are likely to go back to previous queries if the current query returns less relevant results; (2) the searchers with fixed goals are likely to have opposite behaviours to the searchers with evolving goals. For instance, they will refine queries with small changes to the image examples in the queries, and they are likely to get increasingly better results with every query refinement, so that they do not need to reuse previous queries and are likely to get satisfying result images quickly for completing the search tasks.

7.2.2 Search Strategy

The search strategy can be explained by the information patch model of IFT.

When the searchers start the search, they will submit the first query, which can be seen as an initial effort to find the first information patch, and then they might or might not walk around within the patch and evaluate what they have found before they provide feedback to refine or reformulate the query to start a new search (we can consider this as looking for a new patch). In IFT terms, the searchers can decide whether they would like to go between or within patch activities based on their search strategy. Thus, we can categorize the searchers into two types based on the information patch model: motivated by the findings of Mulholland et al. (2008), we suggest that one type of searchers will have a cautious search strategy and a second type of searchers have a risky search strategy. According to the information patch concept, the searchers with cautious search strategy will do more within-patch activities, which means they will carefully search through the current patch before they move to the next patch (e.g. refining the query to start a new search); the searchers with the risky search strategy, on the other hand, will be more adventurous and perform more between-patch activities, which means they will skip over the current patch and move quickly to the next patch.

The assumptions of the cautious and risky search strategy for interactive image search scenario are: (1) the searchers with a cautious search strategy will look through the search results carefully page by page, spend a long time to analyze the results before they refine the query to start a new search; They will not select the result images until they think no better images exist in the result set; (2) the searchers with a risky strategy will only look at the first few pages and select the result images from the pages while they are viewing, and then they will reformulate a new query to start another search.

7.2.3 Evaluation Threshold

The evaluation threshold can be explained by the information diet model of IFT.

When searchers select the result images for completing the tasks, they need to decide which images to choose from the results. In IFT terms, some foragers like easy-to-catch prey, but others like hard-to-catch prey. Thus, the searchers can be categorized into two types based on the information diet concepts: one type with a weak evaluation threshold and the other with a precise evaluation threshold. According to the information diet concepts, the searchers with a weak evaluation threshold will be likely to go for easy-to-catch information, although the information may be just slightly relevant to the their information goal; the searchers with a precise evaluation threshold will instead go for hard-to-catch information: for example, they will not select the information unless it is highly relevant to their information goal.

The assumptions of the weak and precise evaluation goal based on interactive image search scenario are: (1) the searchers with a weak evaluation threshold will select a large number of images based on diverse relevance to their search information goal. For example, if they are looking for a picture of an apple, they will be happy with any picture as long as there is an apple on the picture; (2) the searchers with precise evaluation threshold will only select very relevant images to their search information goal, for instance, if they are looking for a picture of apple, they will not select an image unless there is a red apple in the image, and they will refine the query carefully and try to achieve the precise results.

In summary, Table 7.1 shows the mapping between the IFT and the ISE models (including three categories and six characters), and Table 7.2 shows the definition of the six user characters of the ISE model.

Information Foraging Theory	**ISE Criteria**	**Character**
Information scent models	Information goal	fixed; evolving
Information patch models	Search strategy	cautious; risky
Information diet models	Evaluation threshold	weak; precise

Table 7.1: ISE user classification model based on the Information Foraging Theory

Character	**Definition**
fixed	Searchers with fixed information goal know what they are looking for.
evolving	Searchers with evolving information goal are not sure what they are looking for.
cautious	Searchers with cautious search strategy move slowly between patches.
risky	Searchers with risky search strategy move quickly between patches.
weak	Searchers with weak evaluation threshold are lenient on selecting the results.
precise	Searchers with precise evaluation threshold are strict on selecting the results.

Table 7.2: Definition of the six characters

7.3 Verification of the ISE model

The above definitions of the six user characters of the ISE model are based on the mapping between Information Foraging Theory and the interactive image search scenario. In order to verify the ISE model, we need to operationalize the definitions of the six user characters by mapping them to concrete user interaction features based on real user interaction data collected from an extensive user study that we have performed and described in Chapter 3, and then verify the model by a qualitative data analysis of the interaction data.

7.3.1 The Verification Procedure

The procedure of the ISE model verification is as follows:

1. extract user interaction features of the three evaluations;
2. produce an operational definition of the six characters based on the extracted interaction features;
3. apply the multiple linear regression test to the interaction features;
4. check whether the regression models match the assumptions;
5. describe the assumptions based on the regression model in Information Foraging Theory terms.

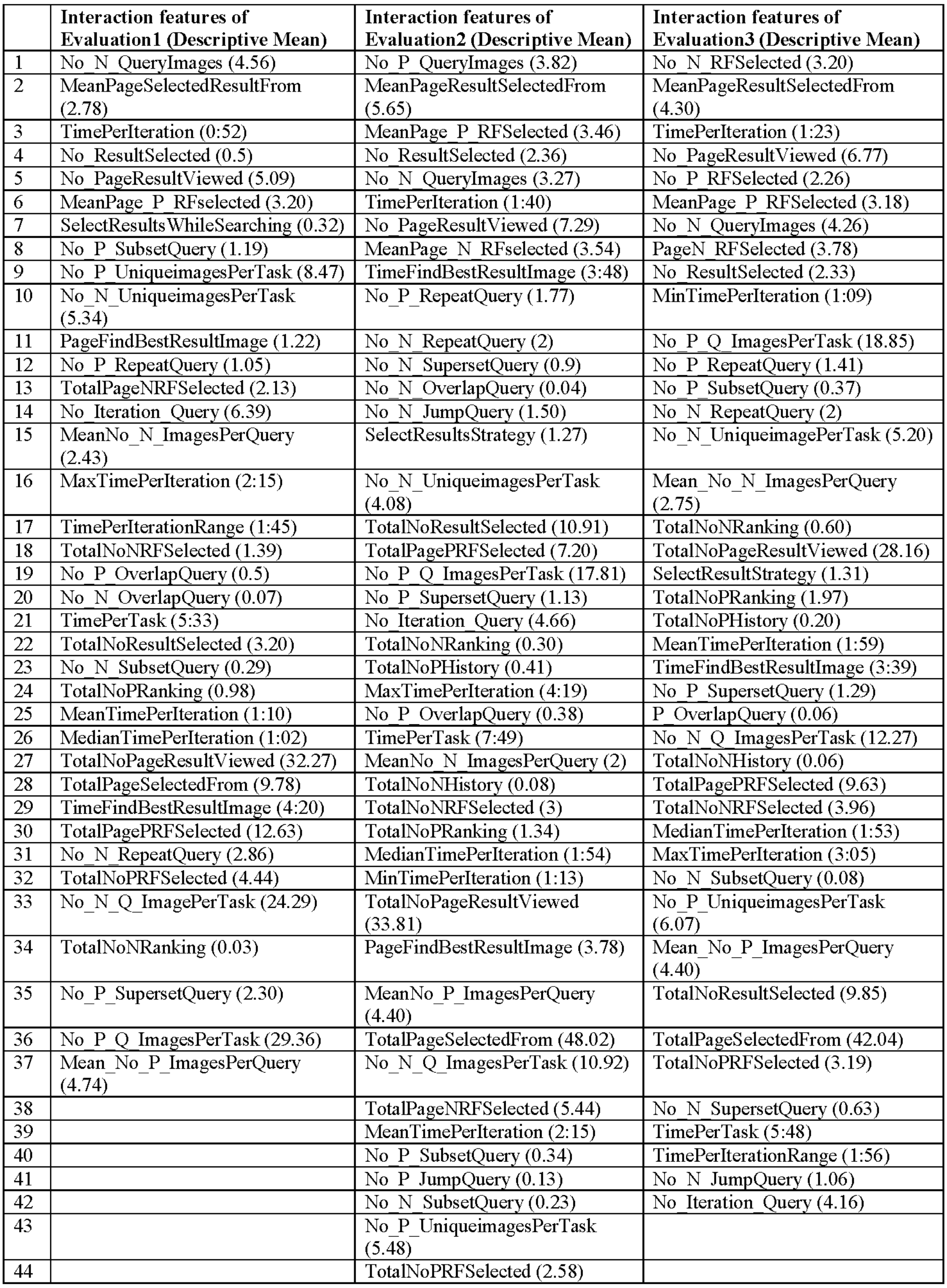
7.3.2 The Interaction Features

A substantial amount of qualitative user interaction features are extracted from the screen capture of the three evaluations in our real user study. There are in total 50 users' screen captures. Every screen capture is about two hours long with both audio and video input. We extract, in total, 123 interactive features from the screen captures (37 from evaluation 1, 44 from evaluation 2, 42 from evaluation 3). Table 7.3 shows the 123 interaction features and their descriptive mean values based on the three evaluations. Some interaction features apply to more than one evaluation, however, the values of the features are different due to different evaluation setups. There are 48 unique features within the total 123 interaction features. Table 7.4 shows the 48 interaction features and their descriptions.

We can basically categorize the 48 unique user interaction features into six groups:

- **time and iteration:** time to complete each iteration, time to complete task, time to find the best result, number of iterations/queries per task;
- **results page:** number of result pages viewed, page results selected from, page found the best result [2], page positive feedback selected from, page negative feedback selected from;
- **image:** number of images per query (positive and negative query), number of feedback images selected (positive and negative query), number of results selected;
- **functionality used:** number of times positive/negative ranking used, number of times positive/negative history used;
- **select results strategy:** some users select results while searching, whilst others select results at the end of the search;

[2] The best result here is judged based on the rating results of the five raters described in chapter 3.



	Interaction features of Evaluation1 (Descriptive Mean)	**Interaction features of Evaluation2 (Descriptive Mean)**	**Interaction features of Evaluation3 (Descriptive Mean)**
1	No_N_QueryImages (4.56)	No_P_QueryImages (3.82)	No_N_RFSelected (3.20)
2	MeanPageSelectedResultFrom (2.78)	MeanPageResultSelectedFrom (5.65)	MeanPageResultSelectedFrom (4.30)
3	TimePerIteration (0:52)	MeanPage_P_RFSelected (3.46)	TimePerIteration (1:23)
4	No_ResultSelected (0.5)	No_ResultSelected (2.36)	No_PageResultViewed (6.77)
5	No_PageResultViewed (5.09)	No_N_QueryImages (3.27)	No_P_RFSelected (2.26)
6	MeanPage_P_RFselected (3.20)	TimePerIteration (1:40)	MeanPage_P_RFSelected (3.18)
7	SelectResultsWhileSearching (0.32)	No_PageResultViewed (7.29)	No_N_QueryImages (4.26)
8	No_P_SubsetQuery (1.19)	MeanPage_N_RFselected (3.54)	PageN_RFSelected (3.78)
9	No_P_UniqueimagesPerTask (8.47)	TimeFindBestResultImage (3:48)	No_ResultSelected (2.33)
10	No_N_UniqueimagesPerTask (5.34)	No_P_RepeatQuery (1.77)	MinTimePerIteration (1:09)
11	PageFindBestResultImage (1.22)	No_N_RepeatQuery (2)	No_P_Q_ImagesPerTask (18.85)
12	No_P_RepeatQuery (1.05)	No_N_SupersetQuery (0.9)	No_P_RepeatQuery (1.41)
13	TotalPageNRFSelected (2.13)	No_N_OverlapQuery (0.04)	No_P_SubsetQuery (0.37)
14	No_Iteration_Query (6.39)	No_N_JumpQuery (1.50)	No_N_RepeatQuery (2)
15	MeanNo_N_ImagesPerQuery (2.43)	SelectResultsStrategy (1.27)	No_N_UniqueimagePerTask (5.20)
16	MaxTimePerIteration (2:15)	No_N_UniqueimagesPerTask (4.08)	Mean_No_N_ImagesPerQuery (2.75)
17	TimePerIterationRange (1:45)	TotalNoResultSelected (10.91)	TotalNoNRanking (0.60)
18	TotalNoNRFSelected (1.39)	TotalPagePRFSelected (7.20)	TotalNoPageResultViewed (28.16)
19	No_P_OverlapQuery (0.5)	No_P_Q_ImagesPerTask (17.81)	SelectResultStrategy (1.31)
20	No_N_OverlapQuery (0.07)	No_P_SupersetQuery (1.13)	TotalNoPRanking (1.97)
21	TimePerTask (5:33)	No_Iteration_Query (4.66)	TotalNoPHistory (0.20)
22	TotalNoResultSelected (3.20)	TotalNoNRanking (0.30)	MeanTimePerIteration (1:59)
23	No_N_SubsetQuery (0.29)	TotalNoPHistory (0.41)	TimeFindBestResultImage (3:39)
24	TotalNoPRanking (0.98)	MaxTimePerIteration (4:19)	No_P_SupersetQuery (1.29)
25	MeanTimePerIteration (1:10)	No_P_OverlapQuery (0.38)	P_OverlapQuery (0.06)
26	MedianTimePerIteration (1:02)	TimePerTask (7:49)	No_N_Q_ImagesPerTask (12.27)
27	TotalNoPageResultViewed (32.27)	MeanNo_N_ImagesPerQuery (2)	TotalNoNHistory (0.06)
28	TotalPageSelectedFrom (9.78)	TotalNoNHistory (0.08)	TotalPagePRFSelected (9.63)
29	TimeFindBestResultImage (4:20)	TotalNoNRFSelected (3)	TotalNoNRFSelected (3.96)
30	TotalPagePRFSelected (12.63)	TotalNoPRanking (1.34)	MedianTimePerIteration (1:53)
31	No_N_RepeatQuery (2.86)	MedianTimePerIteration (1:54)	MaxTimePerIteration (3:05)
32	TotalNoPRFSelected (4.44)	MinTimePerIteration (1:13)	No_N_SubsetQuery (0.08)
33	No_N_Q_ImagePerTask (24.29)	TotalNoPageResultViewed (33.81)	No_P_UniqueimagesPerTask (6.07)
34	TotalNoNRanking (0.03)	PageFindBestResultImage (3.78)	Mean_No_P_ImagesPerQuery (4.40)
35	No_P_SupersetQuery (2.30)	MeanNo_P_ImagesPerQuery (4.40)	TotalNoResultSelected (9.85)
36	No_P_Q_ImagesPerTask (29.36)	TotalPageSelectedFrom (48.02)	TotalPageSelectedFrom (42.04)
37	Mean_No_P_ImagesPerQuery (4.74)	No_N_Q_ImagesPerTask (10.92)	TotalNoPRFSelected (3.19)
38		TotalPageNRFSelected (5.44)	No_N_SupersetQuery (0.63)
39		MeanTimePerIteration (2:15)	TimePerTask (5:48)
40		No_P_SubsetQuery (0.34)	TimePerIterationRange (1:56)
41		No_P_JumpQuery (0.13)	No_N_JumpQuery (1.06)
42		No_N_SubsetQuery (0.23)	No_Iteration_Query (4.16)
43		No_P_UniqueimagesPerTask (5.48)	
44		TotalNoPRFSelected (2.58)	

Table 7.3: 123 Interaction features generated for Evaluation1, 2 and 3 and the features' descriptive means

- **query transitions:** we adapted the five query transitions from Mulholland et al. (2008) study to our analysis. The five transitions for both positive and

	Interaction feature	**Description**
1	No_N_QueryImages	Number of images in a negative query per iteration
2	MeanPageSelectedResultFrom	Mean page number the result images selected from per iteration
3	TimePerIteration	Time used per search iteration
4	No_ResultSelected	Number of result images selected per iteration
5	No_PageResultViewed	Number of result pages viewed per iteration
6	MeanPage_P_RFselected	Mean page number the positive feedback images selected from per iteration
7	SelectResultsStrategy	The way of users selecting the search result for completing tasks
8	No_P_SubsetQuery	Number of subset query transition used in positive queries per task
9	No_P_UniqueimagesPerTask	Number of unique images used in positive queries per task
10	No_N_UniqueimagesPerTask	Number of unique images used in negative queries per task
11	PageFindBestResultImage	Page to find the best result image against the ground truth
12	No_P_RepeatQuery	Number of repeat query transition used in positive queries per task
13	TotalPageNRFSelected	Total page number the negative feedback images selected from per task
14	No_Iteration_Query	Number of queries used per task
15	MeanNo_N_ImagesPerQuery	Mean number of image examples used in negative queries per task
16	MaxTimePerIteration	The longest search iteration to complete a task
17	TimePerIterationRange	Time range between the longest search iteration and the shortest search iteration to complete a task
18	TotalNoNRFSelected	Total number of negative feedback images selected per task
19	No_P_OverlapQuery	Number of overlap query transition used in positive queries per task
20	No_N_OverlapQuery	Number of overlap query transition used in negative queries per task
21	TimePerTask	Time used per task
22	TotalNoResultSelected	Total number of result images selected per task
23	No_N_SubsetQuery	Number of subset query transition used in negative queries per task
24	TotalNoPRanking	Total number of positive query image scoring functionality used per task
25	MeanTimePerIteration	Mean time of all search iterations used to complete a task
26	MedianTimePerIteration	Median time of all search iterations used to complete a task
27	TotalNoPageResultViewed	Total number of result pages viewed per task
28	TotalPageSelectedResultFrom	Total page number the result images selected from per task
29	TimeFindBestResultImage	Time when the best result image (against the ground truth) found
30	TotalPagePRFSelected	Total page number the positive feedback images selected from per task
31	No_N_RepeatQuery	Number of repeat query transition used in negative queries per task
32	TotalNoPRFSelected	Total number of positive feedback images selected per task
33	No_N_Q_ImagePerTask	Number of image examples used in negative queries per task
34	TotalNoNRanking	Total number of negative query image scoring functionality used per task
35	No_P_SupersetQuery	Number of superset query transition used in positive queries per task
36	No_P_Q_ImagesPerTask	Number of image examples used in positive queries per task
37	MeanNo_P_ImagesPerQuery	Mean number of image examples used in positive queries per task
38	No_P_QueryImages	Number of images in a positive query per iteration
39	MeanPage_N_RFselected	Mean page number of negative feedback images selected per iteration
40	No_N_SupersetQuery	Number of superset query transition used in negative queries per task
41	No_N_JumpQuery	Number of jump query transition used in negative queries per task
42	TotalNoPHistory	Total number of positive query history functionality used per task
43	TotalNoNHistory	Total number of negative query history functionality used per task
44	MinTimePerIteration	The shortest search iteration to complete a task
45	No_P_JumpQuery	Number of jump query transition used in positive queries per task
46	No_N_RFSelected	Number of negative feedback images selected per iteration
47	No_P_RFSelected	Number of positive feedback images selected per iteration
48	PageN_RFSelected	Total page number the negative feedback images selected from per iteration

Table 7.4: 48 unique interaction features generated from the screen capture of the 3 evaluations

negative queries are given in Table 7.5.

Query transition	**description**
Repeat	Consecutive positive or negative query contains identical images.
Subset	The next positive or negative query contains a subset of the query images.
Superset	The next positive or negative query contains all the previous images plus one or more additional images.
Overlap	The next positive or negative query contains some but not all of the previous images plus one or more additional images.
Jump	There is no intersection between the images used in consecutive positive or negative queries.

Table 7.5: The adapted five query transitions

7.3.3 The Analysis Assumptions: An Operational ISE Model based on the 48 unique User Interaction Features

Characters	**Operational definition**
Fixed	1. use small number of jump query transitions; 2. use small number of history functions; 3. find the best result image early.
Evolving	1. use large number of jump query transitions; 2. use large number of history functions; 3. find the best result image late.
Cautious	1. view large number of result pages; 2. spend a long time per search iteration; 3. select results at the end of the search.
Risky	1. view small number of result pages; 2. spend a short time per search iteration; 3. select results while searching.
Weak	1. select a large number of results; 2. select a large number of feedback; 3. use a small number of search iterations.
Precise	1. use lots of subset query transition; 2. use the query image scoring functionality many times; 3. use a large number of search iterations.

Table 7.6: Operational definition of the six characters

After defining the ISE model (including three criteria and six user characters) based on Information Foraging Theory, we examine the 48 unique interaction features from the collected qualitative data, and assign the six characters or their combinations to all the 48 interaction features based on the definition provided in Table 7.1[3]. Comments are provided on why the assignments were made (Table 7.7 gives an example of how we assign the characters or their combinations to the 48 unique interaction features).

Table 7.6 summarizes the operational definitions of the six characters based on the character allocation results of 48 unique interaction features. We can then verify the ISE model by a qualitative data analysis of all the 123 interaction features based on the operational definitions of the six user characters in the ISE model.

7.3.4 Multiple Linear Regression

Multiple linear regression is applied to our qualitative analysis, because we want to find out the correlations across the interaction features, and also want to generate models for predicting the interaction features. The 123 features are the input of the multiple linear regression. We carry out the regression test using SPSS, a statistical analysis tool.

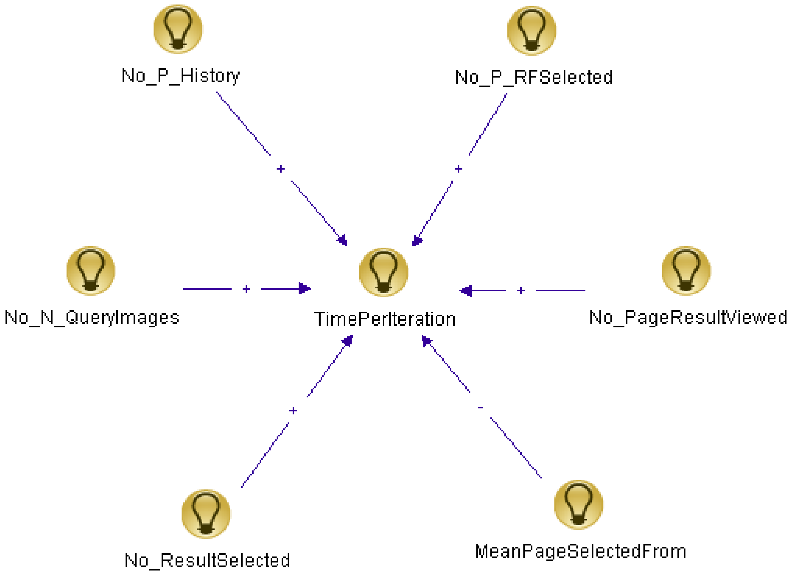
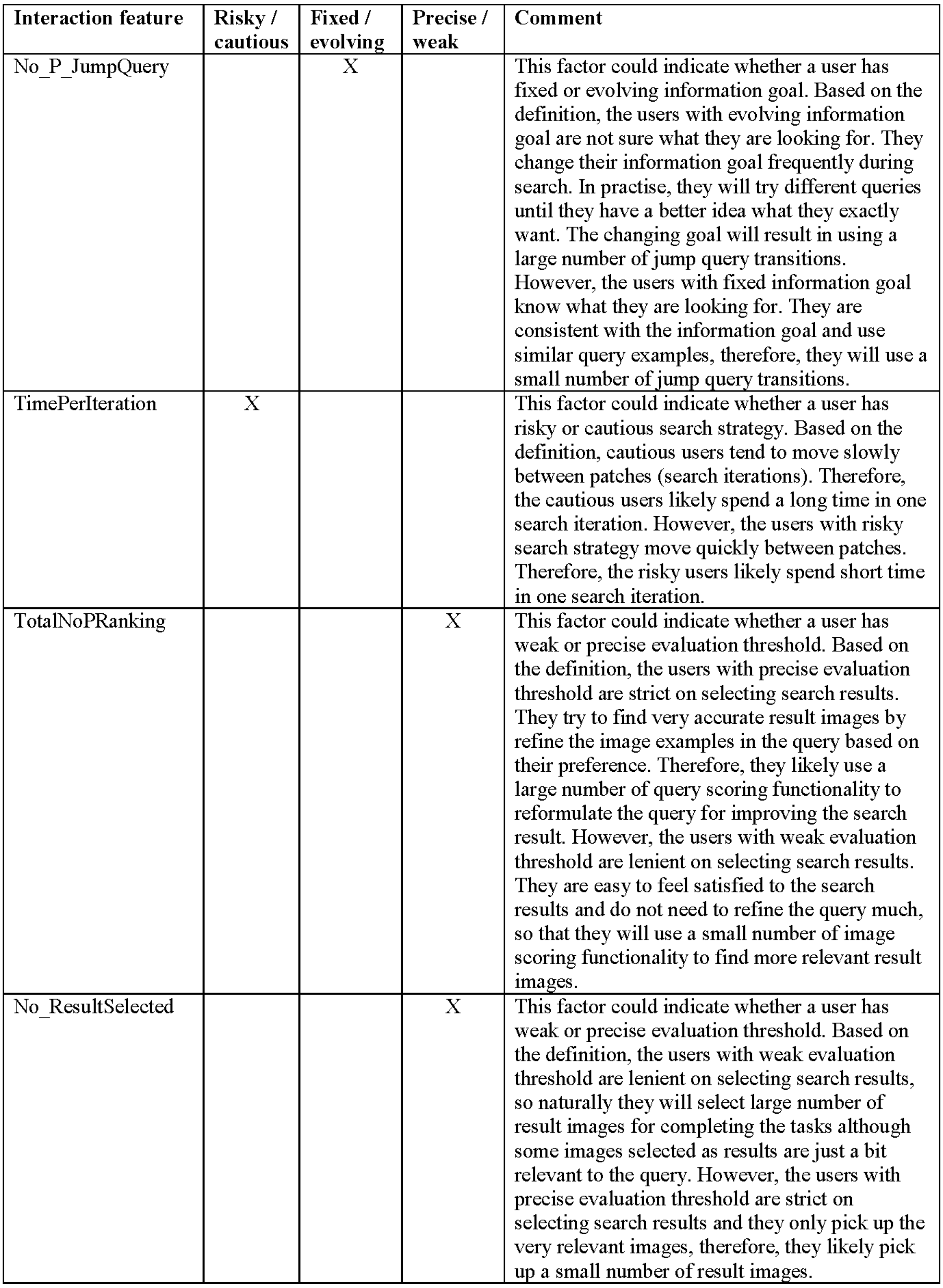


Figure 7.1: An example of visualized multiple linear regression model

[3] We looked at the 48 unique features because we wanted to let users decide which interaction features would be good to operationalize the ISE model.



Interaction feature	Risky / cautious	Fixed / evolving	Precise / weak	Comment
No_P_JumpQuery		X		This factor could indicate whether a user has fixed or evolving information goal. Based on the definition, the users with evolving information goal are not sure what they are looking for. They change their information goal frequently during search. In practise, they will try different queries until they have a better idea what they exactly want. The changing goal will result in using a large number of jump query transitions. However, the users with fixed information goal know what they are looking for. They are consistent with the information goal and use similar query examples, therefore, they will use a small number of jump query transitions.
TimePerIteration	X			This factor could indicate whether a user has risky or cautious search strategy. Based on the definition, cautious users tend to move slowly between patches (search iterations). Therefore, the cautious users likely spend a long time in one search iteration. However, the users with risky search strategy move quickly between patches. Therefore, the risky users likely spend short time in one search iteration.
TotalNoPRanking			X	This factor could indicate whether a user has weak or precise evaluation threshold. Based on the definition, the users with precise evaluation threshold are strict on selecting search results. They try to find very accurate result images by refine the image examples in the query based on their preference. Therefore, they likely use a large number of query scoring functionality to reformulate the query for improving the search result. However, the users with weak evaluation threshold are lenient on selecting search results. They are easy to feel satisfied to the search results and do not need to refine the query much, so that they will use a small number of image scoring functionality to find more relevant result images.
No_ResultSelected			X	This factor could indicate whether a user has weak or precise evaluation threshold. Based on the definition, the users with weak evaluation threshold are lenient on selecting search results, so naturally they will select large number of result images for completing the tasks although some images selected as results are just a bit relevant to the query. However, the users with precise evaluation threshold are strict on selecting search results and they only pick up the very relevant images, therefore, they likely pick up a small number of result images.

Table 7.7: Example of assigning the six characters to the interaction features

We first test the multiple linear regression on the interaction features of the three evaluations respectively. Then we get a model to predict each interaction feature (Figure 7.1). For example, in Figure 7.1, the interaction feature *TimePerIteration* is predicted by the other six interation features: *No_P_RFselected*, *No_PageResultViewed*, *No_N_QueryImages*, *No_ResultSelected*, *No_P_History* and *MeanPageSelectedResultFrom* interaction features. There are relation lines between the interaction features and predicted feature. The direction of each arrow is the predicting direction. The + and − on the line denotes whether the prediction is positive or negative. For instance, the *No_PageResultViewed* predicts *TimePerIteration* positively, i.e., if a user views more result pages, they are likely to spend a longer time per search iteration.

7.3.5 Regression model analysis

We have obtained the 123 regression models of the 123 interaction features involved in all the three user evaluations. We then need to investigate whether the operational definitions of the proposed six user characters are supported by the regression models.

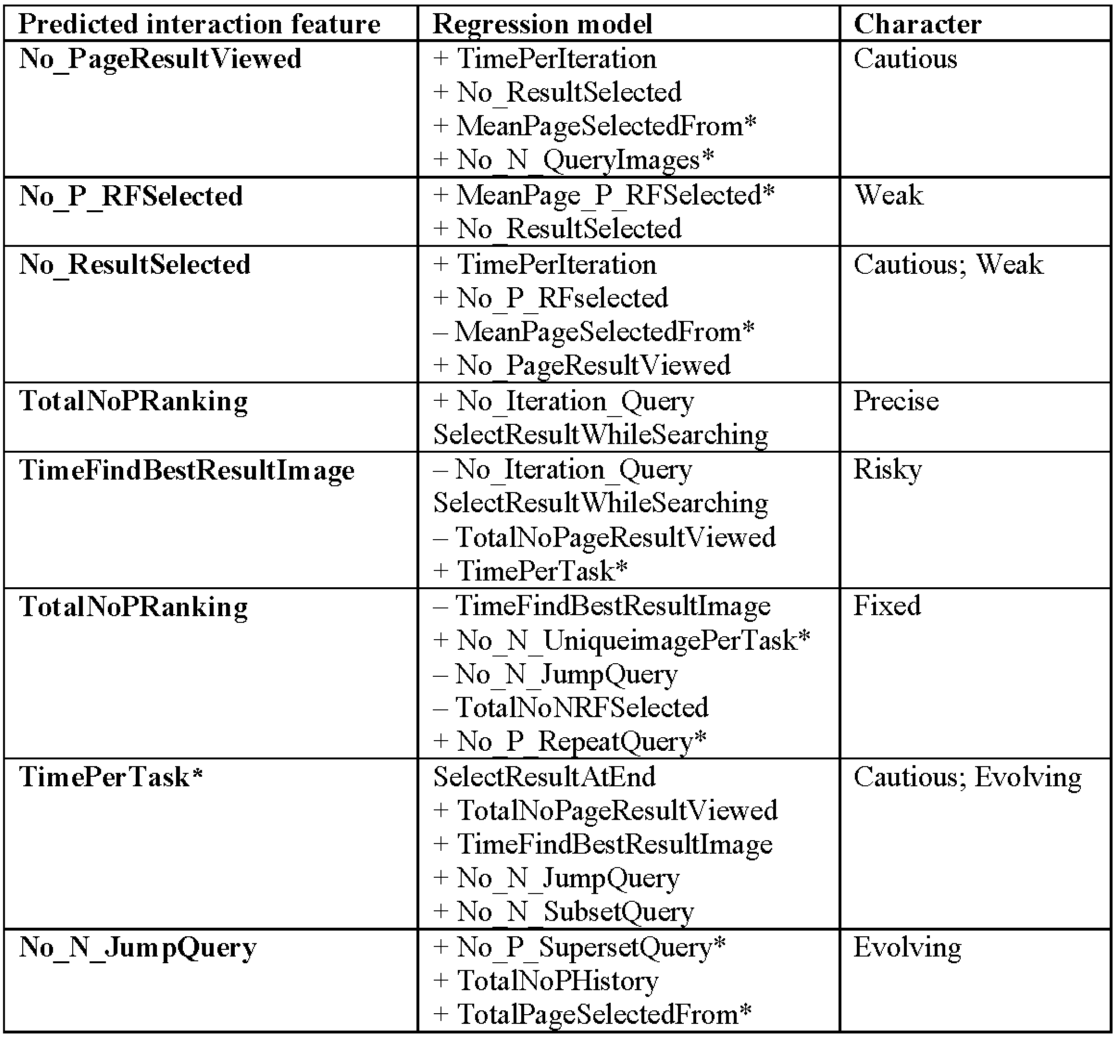
We assign the six characters in ISE model and their combinations to the 123 models[4]. **The justification method for assigning a character to a model confidently is that the model has to contain at least two interaction features that are relative to the character's operational definition of the ISE model.** Examples of the assignments of characters to the regression models are given in Table 7.8[5].

The results show that the models can be described by the the six characters or their combinations, and the descriptions fit[6] the operational definitions of the six characters. Take the regression model predicting *No_PageResultViewed* in Table 7.8 as an example. *No_PageResultViewed* is positively predicted by *TimePerIteration*,

[4] The reason of performing multiple linear regression on all the 123 interaction features is for the verification of the consistency of the manually-defined operational features with the correlated features revealed by the multiple linear regression models.

[5] In Table 7.8, the "+/−" shows how the features in the regression models predict the interaction features in the second column. "+" means the prediction is positive, and "−" means the prediction is negative. "∗" indicates that the interaction features are not mentioned in the operational definitions in Table 7.6.

[6] "Fit" means all the mentioned features in a regression model are correctly detected based on the operational definitions.



Predicted interaction feature	Regression model	Character
No_PageResultViewed	+ TimePerIteration + No_ResultSelected + MeanPageSelectedFrom* + No_N_QueryImages*	Cautious
No_P_RFSelected	+ MeanPage_P_RFSelected* + No_ResultSelected	Weak
No_ResultSelected	+ TimePerIteration + No_P_RFselected – MeanPageSelectedFrom* + No_PageResultViewed	Cautious; Weak
TotalNoPRanking	+ No_Iteration_Query SelectResultWhileSearching	Precise
TimeFindBestResultImage	– No_Iteration_Query SelectResultWhileSearching – TotalNoPageResultViewed + TimePerTask*	Risky
TotalNoPRanking	– TimeFindBestResultImage + No_N_UniqueimagePerTask* – No_N_JumpQuery – TotalNoNRFSelected + No_P_RepeatQuery*	Fixed
TimePerTask*	SelectResultAtEnd + TotalNoPageResultViewed + TimeFindBestResultImage + No_N_JumpQuery + No_N_SubsetQuery	Cautious; Evolving
No_N_JumpQuery	+ No_P_SupersetQuery* + TotalNoPHistory + TotalPageSelectedFrom*	Evolving

Table 7.8: Example of assigning the six characters to the regression models

No_ResultSelected, *MeanPageSelectedFrom* and *No_N_QueryImages*, which means that users will view lots of result pages if they spend a long time per search iteration, select a large number of result images, select result images from late pages, and use large number of negative query images. According to the operational definitions and our justification method, this model can be described by the cautious character because the model contains two interaction features that are related to the operational definition of the cautious character, and the description of the model fits the operational definition well, e.g., spends a long time per search iteration and views a large number of result pages.

From the 123 regression model analysis results we can see most of the characters or

Character	No. of models
Cautious	7
Risky	2
Evolving	7
Fixed	1
Weak	12
Precise	12
Cautious+Evolving	2
Cautious+Weak	2
N/A	78
Total	123

Table 7.9: Summary of characters and no. of supporting regression models

their combinations correspond to the regression models. Some regression models can be described by single characters but some models need to be described by different combinations of the six characters. Some characters correspond to a large number of regression models but some only correspond to a couple of regression models. The user characters and their combinations corresponded to at least one regression models are listed in Table 7.9[7].

Table 7.9 shows that the 45 regression models confidently identify 8 character groups based on our justification method[8]. Each character group[9] is identified by an average of 5.6 regression models. We suggest the four character groups identified by more than six regression models are well represented character groups in our user study, namely: *cautious*, *evolving*, *weak* and *precise*. Seventy eight regression models cannot be clearly described by any character groups. Within the 78 regression models, 8 models do not include any interaction features that are relative to the operational definitions of the six characters in the ISE model, and 70 models contain one/more single interaction feature that is/are relative to the operational definition of one/more characters. According to the judgement method for the ISE model

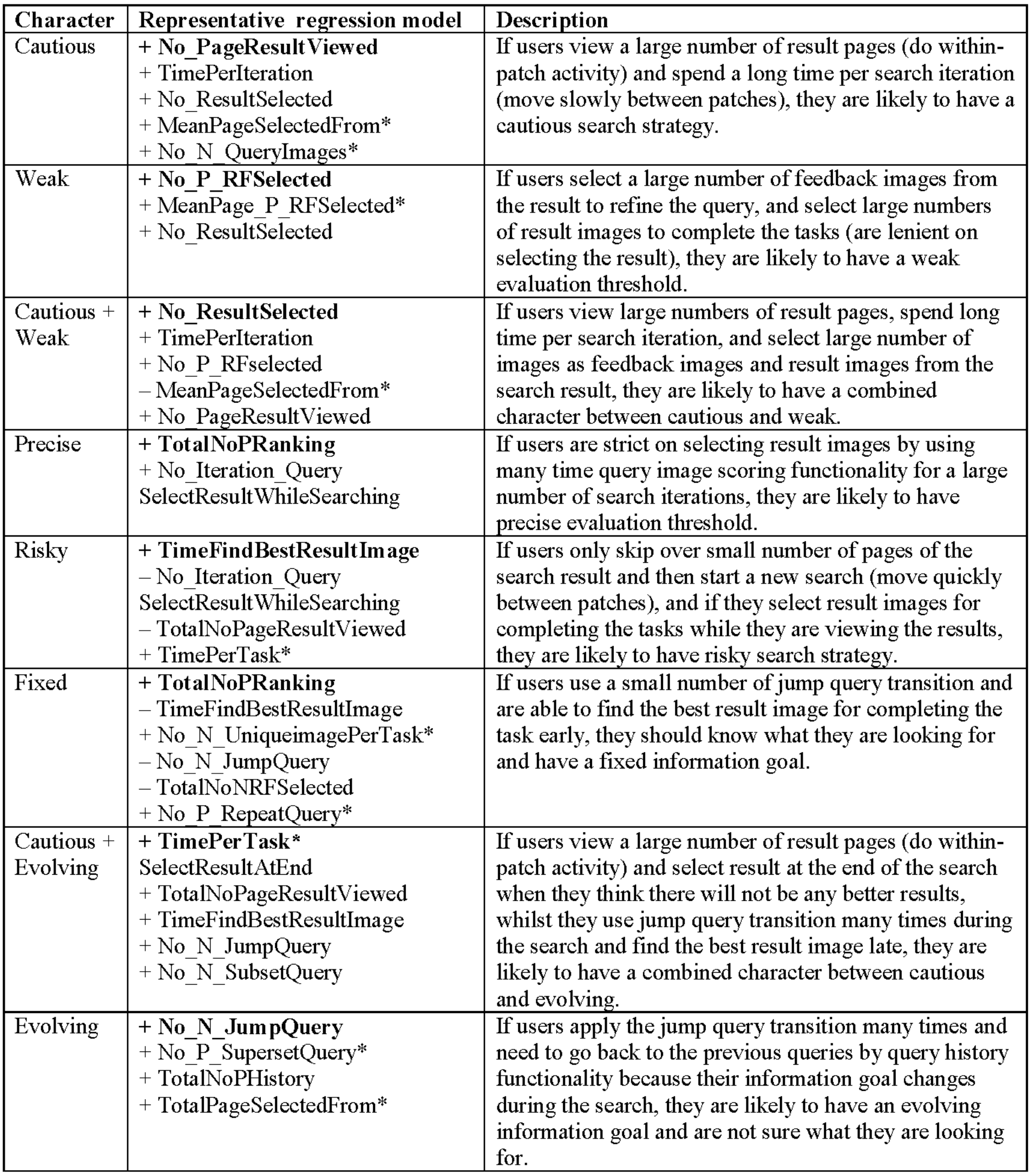
[7] In Table 7.9, N/A = there is no more than one interaction feature relative to the operational definition of any character in the regression model.

[8] The interaction features in the regression models show reasonable predictions. Further, there are at least two interaction features in each model that fit the operational definitions of the six characters in the ISE model.

[9] The character group can be any single character or a combination of characters. We consider a character group is a user type.

verification, the 70 regression models should not be used to verify the ISE model, although the models are all consistent with the operational definition. Therefore, we only take into consideration the 45 models that contain at least two interaction features that are relative to the characters' operational definition.

7.3.6 Description of 8 Character Groups



Character	Representative regression model	Description
Cautious	**+ No_PageResultViewed** + TimePerIteration + No_ResultSelected + MeanPageSelectedFrom* + No_N_QueryImages*	If users view a large number of result pages (do within-patch activity) and spend a long time per search iteration (move slowly between patches), they are likely to have a cautious search strategy.
Weak	**+ No_P_RFSelected** + MeanPage_P_RFSelected* + No_ResultSelected	If users select a large number of feedback images from the result to refine the query, and select large numbers of result images to complete the tasks (are lenient on selecting the result), they are likely to have a weak evaluation threshold.
Cautious + Weak	**+ No_ResultSelected** + TimePerIteration + No_P_RFselected – MeanPageSelectedFrom* + No_PageResultViewed	If users view large numbers of result pages, spend long time per search iteration, and select large number of images as feedback images and result images from the search result, they are likely to have a combined character between cautious and weak.
Precise	**+ TotalNoPRanking** + No_Iteration_Query SelectResultWhileSearching	If users are strict on selecting result images by using many time query image scoring functionality for a large number of search iterations, they are likely to have precise evaluation threshold.
Risky	**+ TimeFindBestResultImage** – No_Iteration_Query SelectResultWhileSearching – TotalNoPageResultViewed + TimePerTask*	If users only skip over small number of pages of the search result and then start a new search (move quickly between patches), and if they select result images for completing the tasks while they are viewing the results, they are likely to have risky search strategy.
Fixed	**+ TotalNoPRanking** – TimeFindBestResultImage + No_N_UniqueimagePerTask* – No_N_JumpQuery – TotalNoNRFSelected + No_P_RepeatQuery*	If users use a small number of jump query transition and are able to find the best result image for completing the task early, they should know what they are looking for and have a fixed information goal.
Cautious + Evolving	**+ TimePerTask*** SelectResultAtEnd + TotalNoPageResultViewed + TimeFindBestResultImage + No_N_JumpQuery + No_N_SubsetQuery	If users view a large number of result pages (do within-patch activity) and select result at the end of the search when they think there will not be any better results, whilst they use jump query transition many times during the search and find the best result image late, they are likely to have a combined character between cautious and evolving.
Evolving	**+ No_N_JumpQuery** + No_P_SupersetQuery* + TotalNoPHistory + TotalPageSelectedFrom*	If users apply the jump query transition many times and need to go back to the previous queries by query history functionality because their information goal changes during the search, they are likely to have an evolving information goal and are not sure what they are looking for.

Table 7.10: Regression model explanations for 8 characters

To further verify the ISE model, we will describe the 8 character groups that correspond to at least one regression model in Table 7.9. We first choose a representative regression model for each character group, and then describe the regression model based on the definition of the ISE model (Table 7.2) and Information Foraging Theory. The description of the 8 corresponding character groups is given in Table 7.10[10].

7.4 Summary

In an effort to understand the users' interaction preferences and behaviours based on different user types for CBIR, we have proposed a user classification model - ISE - based on Information Foraging Theory. The ISE model contains three criteria: information goal (I), search strategy (S) and evaluation threshold (E). There are different types of user characters in each criterion. They are fixed information goal and evolving information goal (I); risky search strategy and cautious search strategy (S); weak evaluation threshold and precise evaluation threshold (E).

In order to verify the ISE model, we have first operationalized the ISE model based on the 48 unique interaction features extracted from the screen capture of our user study. A multiple linear regression has then been performed on the total number of 123 interaction features involved in all the 3 evaluations in the user study, resulting in 123 regression models. Finally, we have investigated whether the operational definitions of the six user characters in the ISE model are consistent with the regression models based on a regression model analysis.

The ISE user classification model has been successfully verified by the qualitative data analysis. The findings show that all regression models are sensible and consistent with the operational definitions of the six characters in the ISE model. Eight user character groups (user types) are confidently identified by 45 regression models.

[10] "*" indicates that the interaction features are not mentioned in the operational definition in Table 7.6.

This practise has not only helped to find different user types for future user-focused design, study and analysis, but also reinforced the usefulness of Information Foraging Theory for exploratory search, especially for exploratory CBIR search.

Chapter 8

Quantitative and Qualitative Analysis of User Evaluation Results Based on the ISE Model

We proposed and verified an ISE user classification model in Chapter 7. We found that the different user types affect users' search behaviours and their search preferences. However, what are the search behaviours and search preferences based on different user types, and how they are reflected in the interactive content-based image retrieval (CBIR) framework design, evaluation and analysis? In this Chapter, we are going to perform further quantitative and qualitative analysis of our user evaluation results based on the ISE model, and further investigate the search behaviours and preferences of different user types, and their implications for future CBIR studies.

In Section 8.1 we will describe the methodology we applied to identify the characters of the 50 users in our evaluations. Section 8.2 will report the findings on search results with regard to different user types based on quantitative data, and different types of users' expectations of image search tools, their experience (satisfaction) of the system, and their suggestions on how to improve the evaluation systems, based on qualitative data obtained from our user study, e.g., users' comments on the questionnaires. The summary of the analysis results will be stated in Section 8.3.

8.1 Methodology

We apply the definitions and operational definitions of the six characters of the ISE model (Table 7.2 and Table 7.6) to the qualitative data extracted from the screen captures of the three evaluations in order to find the user type[1] of each individual subject in our user study. Further, we group the 50 users based on their user types to find the different search preferences and search behaviours based on user groups[2]. The concrete methodology is detailed as follows:

8.1.1 Identifying User Types

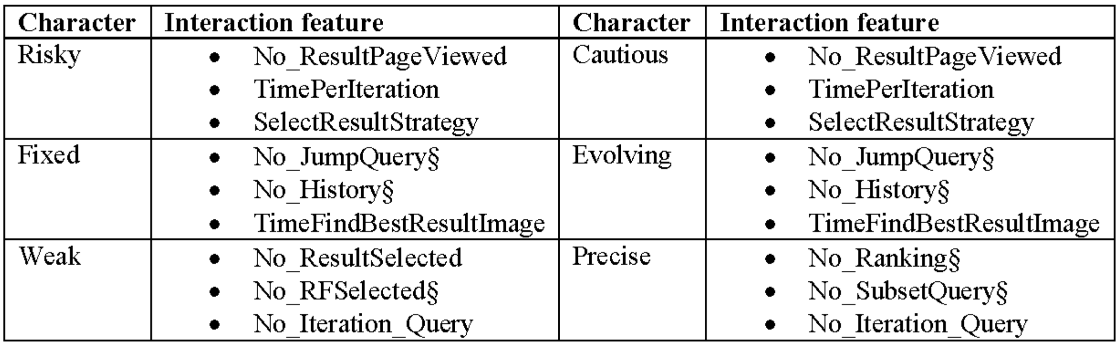
The following steps describe how we identify characters for each individual user.

Step 1 - Find 3 interaction features and their values for the 3 operational definitions of the 6 characters. We extract 3 interaction features that match the 3 operational definitions of each character, per task per evaluation, from the qualitative data. For example, 17 users completed 4 tasks in evaluation 3, and we extracted 43 interactive features from every user's screen capture (Table 7.3). We have identified 6 characters, and each character has 3 operational definitions (Table 7.6). Each operational definition will be supported by one of 43 interaction features. For instance, the operational definitions of *Risky* character are (1) view small number of result pages; (2) spend a short time per search iteration; (3) select results while searching. In this case, the best supportive interaction features from the 43 interaction feature in evaluation 3 are (1) $No_ResultPageViewed$; (2)

[1] A user type could include more than one character from different classification criteria of the ISE model.

[2] A user group contains users with the same user type.

TimePerIteration; (3) *SelectResultStrategy*. Accordingly, we find the 3 interaction features that match the 3 operational definitions of each character for evaluation 3. Table 8.1[3] shows the 3 interaction features we used for each of the 6 characters of the ISE model.



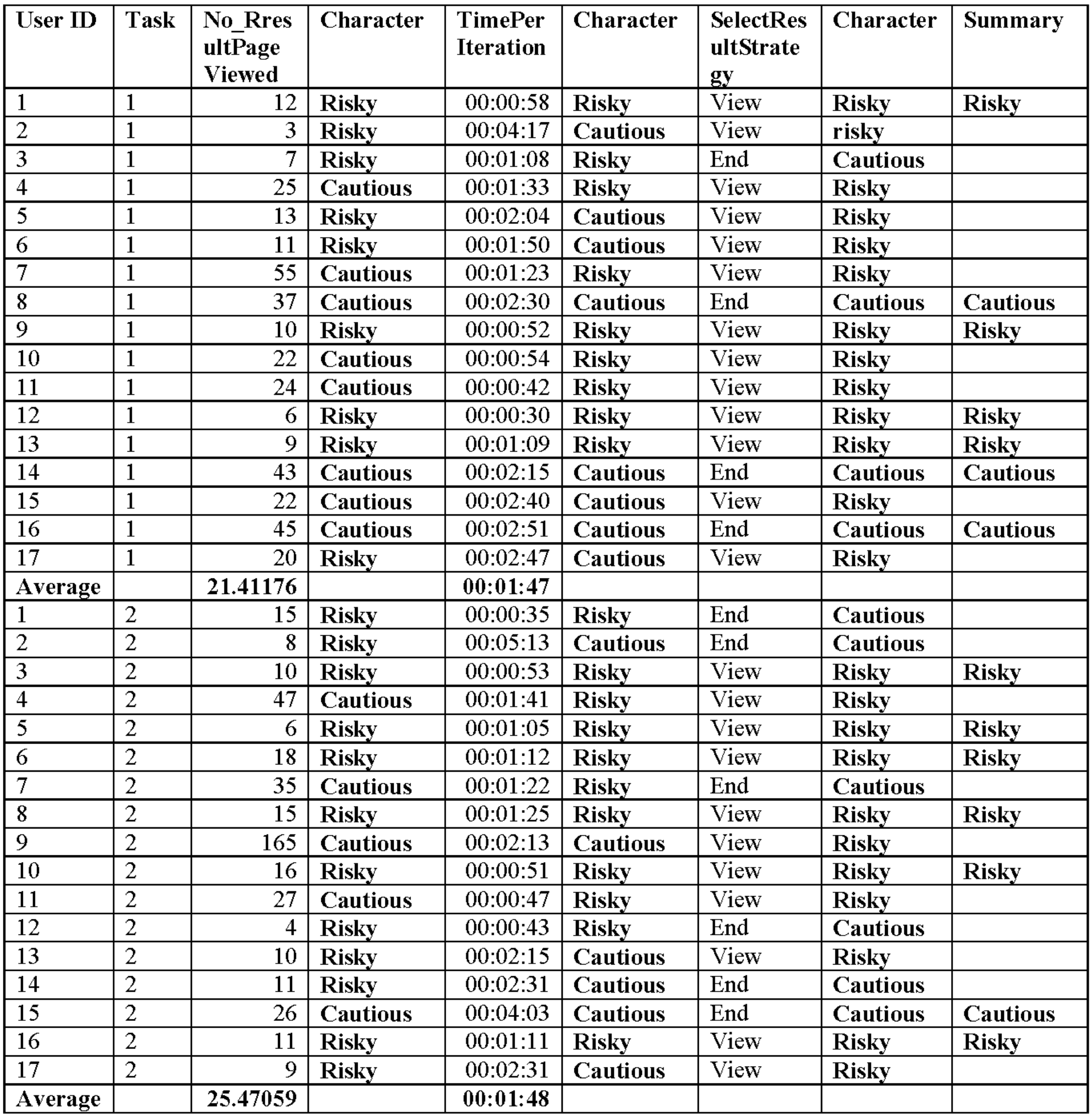
Character	Interaction feature	Character	Interaction feature
Risky	• No_ResultPageViewed • TimePerIteration • SelectResultStrategy	Cautious	• No_ResultPageViewed • TimePerIteration • SelectResultStrategy
Fixed	• No_JumpQuery§ • No_History§ • TimeFindBestResultImage	Evolving	• No_JumpQuery§ • No_History§ • TimeFindBestResultImage
Weak	• No_ResultSelected • No_RFSelected§ • No_Iteration_Query	Precise	• No_Ranking§ • No_SubsetQuery§ • No_Iteration_Query

Table 8.1: Interaction features that support the operational definitions of the six characters

Step 2 - Identify the characters of every user for each task. We calculate the mean value of the data for every interaction feature with regard to each task and each evaluation across all the users. We then judge the character of a user based on whether the value of an interaction feature for the user is larger than the mean value or not. For the interaction features with binary values such as *SelectResultStrategy*, we do not calculate the mean value and instead judge the character of the user based on the data itself. The final character of a user for that specific task is cautious when the cautious character emerges from all 3 interaction features. A user will be identified risky when he shows risky character with regard to all the 3 interaction features, otherwise the user will not be risky nor cautious. We applied the same methodology on checking fixed or evolving character and weak or precise character. Table 8.2 and Table 8.3 are examples to show how we identified whether the user is risky or cautious [4] for the 4 tasks of evaluation 3.

[3] "§" indicates the interaction feature may be relative to positive query or negative query. In total 16 interaction features from the qualitative data support the operational definitions in the ISE model.

[4] It happens here that the 3 interaction features for the operational definitions of "risky" and "cautious" are the same.

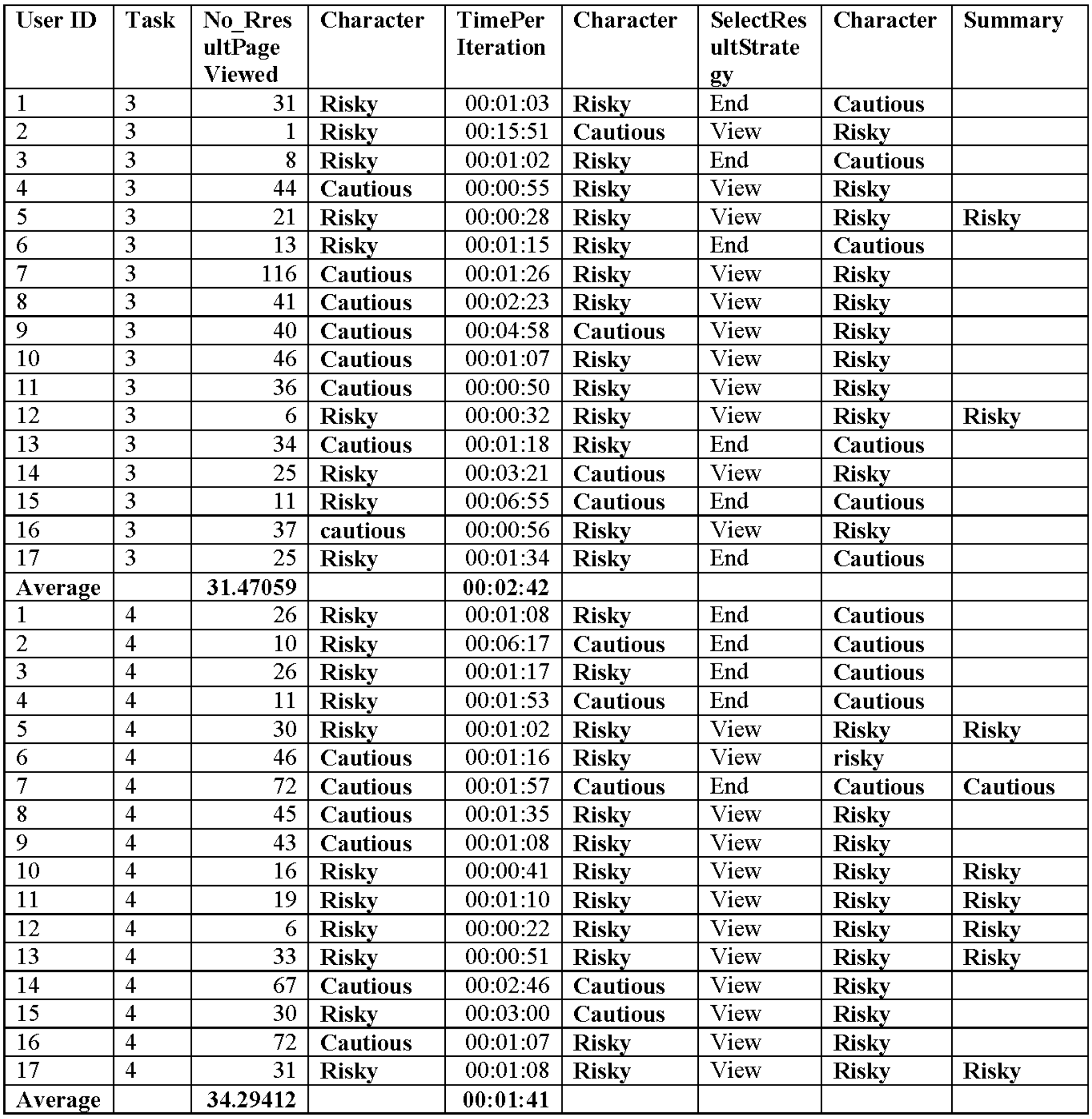


User ID	Task	No_RresultPage Viewed	Character	TimePer Iteration	Character	SelectResultStrategy	Character	Summary
1	1	12	**Risky**	00:00:58	**Risky**	View	**Risky**	**Risky**
2	1	3	**Risky**	00:04:17	**Cautious**	View	**risky**	
3	1	7	**Risky**	00:01:08	**Risky**	End	**Cautious**	
4	1	25	**Cautious**	00:01:33	**Risky**	View	**Risky**	
5	1	13	**Risky**	00:02:04	**Cautious**	View	**Risky**	
6	1	11	**Risky**	00:01:50	**Cautious**	View	**Risky**	
7	1	55	**Cautious**	00:01:23	**Risky**	View	**Risky**	
8	1	37	**Cautious**	00:02:30	**Cautious**	End	**Cautious**	**Cautious**
9	1	10	**Risky**	00:00:52	**Risky**	View	**Risky**	**Risky**
10	1	22	**Cautious**	00:00:54	**Risky**	View	**Risky**	
11	1	24	**Cautious**	00:00:42	**Risky**	View	**Risky**	
12	1	6	**Risky**	00:00:30	**Risky**	View	**Risky**	**Risky**
13	1	9	**Risky**	00:01:09	**Risky**	View	**Risky**	**Risky**
14	1	43	**Cautious**	00:02:15	**Cautious**	End	**Cautious**	**Cautious**
15	1	22	**Cautious**	00:02:40	**Cautious**	View	**Risky**	
16	1	45	**Cautious**	00:02:51	**Cautious**	End	**Cautious**	**Cautious**
17	1	20	**Risky**	00:02:47	**Cautious**	View	**Risky**	
Average		**21.41176**		**00:01:47**				
1	2	15	**Risky**	00:00:35	**Risky**	End	**Cautious**	
2	2	8	**Risky**	00:05:13	**Cautious**	End	**Cautious**	
3	2	10	**Risky**	00:00:53	**Risky**	View	**Risky**	**Risky**
4	2	47	**Cautious**	00:01:41	**Risky**	View	**Risky**	
5	2	6	**Risky**	00:01:05	**Risky**	View	**Risky**	**Risky**
6	2	18	**Risky**	00:01:12	**Risky**	View	**Risky**	**Risky**
7	2	35	**Cautious**	00:01:22	**Risky**	End	**Cautious**	
8	2	15	**Risky**	00:01:25	**Risky**	View	**Risky**	**Risky**
9	2	165	**Cautious**	00:02:13	**Cautious**	View	**Risky**	
10	2	16	**Risky**	00:00:51	**Risky**	View	**Risky**	**Risky**
11	2	27	**Cautious**	00:00:47	**Risky**	View	**Risky**	
12	2	4	**Risky**	00:00:43	**Risky**	End	**Cautious**	
13	2	10	**Risky**	00:02:15	**Cautious**	View	**Risky**	
14	2	11	**Risky**	00:02:31	**Cautious**	End	**Cautious**	
15	2	26	**Cautious**	00:04:03	**Cautious**	End	**Cautious**	**Cautious**
16	2	11	**Risky**	00:01:11	**Risky**	View	**Risky**	**Risky**
17	2	9	**Risky**	00:02:31	**Cautious**	View	**Risky**	
Average		**25.47059**		**00:01:48**				

Table 8.2: An example of how to identify risky or cautious (1)

Step 3 - Identify the characters for every user. After checking every user's character for each task (Step 2), we then need to summarize the user's overall characters for the 3 evaluations respectively. We decide the type of user based on the following criteria:

1. Risky (R) user: $\geq$ 2 tasks shown risky and 0 tasks shown cautious;
2. Cautious (C) user: $\geq$ 2 tasks shown cautious and 0 tasks shown risky;
3. MixRC user: $>$ 0 tasks shown risky and $>$ 0 tasks shown cautious;
4. NoneRC user: 0 tasks shown cautious and 0 tasks shown risky;



User ID	Task	No_RresultPage Viewed	Character	TimePer Iteration	Character	SelectResultStrategy	Character	Summary
1	3	31	**Risky**	00:01:03	**Risky**	End	**Cautious**	
2	3	1	**Risky**	00:15:51	**Cautious**	View	**Risky**	
3	3	8	**Risky**	00:01:02	**Risky**	End	**Cautious**	
4	3	44	**Cautious**	00:00:55	**Risky**	View	**Risky**	
5	3	21	**Risky**	00:00:28	**Risky**	View	**Risky**	**Risky**
6	3	13	**Risky**	00:01:15	**Risky**	End	**Cautious**	
7	3	116	**Cautious**	00:01:26	**Risky**	View	**Risky**	
8	3	41	**Cautious**	00:02:23	**Risky**	View	**Risky**	
9	3	40	**Cautious**	00:04:58	**Cautious**	View	**Risky**	
10	3	46	**Cautious**	00:01:07	**Risky**	View	**Risky**	
11	3	36	**Cautious**	00:00:50	**Risky**	View	**Risky**	
12	3	6	**Risky**	00:00:32	**Risky**	View	**Risky**	**Risky**
13	3	34	**Cautious**	00:01:18	**Risky**	End	**Cautious**	
14	3	25	**Risky**	00:03:21	**Cautious**	View	**Risky**	
15	3	11	**Risky**	00:06:55	**Cautious**	End	**Cautious**	
16	3	37	**cautious**	00:00:56	**Risky**	View	**Risky**	
17	3	25	**Risky**	00:01:34	**Risky**	End	**Cautious**	
Average		**31.47059**		**00:02:42**				
1	4	26	**Risky**	00:01:08	**Risky**	End	**Cautious**	
2	4	10	**Risky**	00:06:17	**Cautious**	End	**Cautious**	
3	4	26	**Risky**	00:01:17	**Risky**	End	**Cautious**	
4	4	11	**Risky**	00:01:53	**Cautious**	End	**Cautious**	
5	4	30	**Risky**	00:01:02	**Risky**	View	**Risky**	**Risky**
6	4	46	**Cautious**	00:01:16	**Risky**	View	**risky**	
7	4	72	**Cautious**	00:01:57	**Cautious**	End	**Cautious**	**Cautious**
8	4	45	**Cautious**	00:01:35	**Risky**	View	**Risky**	
9	4	43	**Cautious**	00:01:08	**Risky**	View	**Risky**	
10	4	16	**Risky**	00:00:41	**Risky**	View	**Risky**	**Risky**
11	4	19	**Risky**	00:01:10	**Risky**	View	**Risky**	**Risky**
12	4	6	**Risky**	00:00:22	**Risky**	View	**Risky**	**Risky**
13	4	33	**Risky**	00:00:51	**Risky**	View	**Risky**	**Risky**
14	4	67	**Cautious**	00:02:46	**Cautious**	View	**Risky**	
15	4	30	**Risky**	00:03:00	**Cautious**	View	**Risky**	
16	4	72	**Cautious**	00:01:07	**Risky**	View	**Risky**	
17	4	31	**Risky**	00:01:08	**Risky**	View	**Risky**	**Risky**
Average		**34.29412**		**00:01:41**				

Table 8.3: An example of how to identify risky or cautious (2)

5. UndefinedRC user: does not match 1 - 4;
6. Fixed (F) user: $\geq$ 2 tasks shown fixed and 0 tasks shown evolving;
7. Evolving (E) user: $\geq$ 2 tasks shown evolving and 0 tasks shown fixed;
8. MixFE user: $>$ 0 tasks shown fixed and $>$ 0 tasks shown evolving;
9. NoneFE user: 0 tasks shown fixed and 0 tasks shown evolving;
10. UndefinedFE user: does not match 6 - 9;
11. Weak (W) user: $\geq$ 2 tasks shown weak and 0 tasks shown precise;
12. Precise (P) user: $\geq$ 2 tasks shown precise and 0 tasks shown weak;

13. MixWP user: > 0 tasks shown weak and > 0 tasks shown precise;
14. NoneWP user: 0 tasks shown weak and 0 tasks shown precise;
15. UndefinedWP user: does not match 11 - 14;

Table 8.4 shows an example of how we identified the user types based on the above criteria for evaluation 3.

User ID	**User Type**	**Risky / Cautious**				**Fixed / Evolving**				**Weak / Precise**			
		Task1	Task2	Task3	Task4	Task1	Task2	Task3	Task4	Task1	Task2	Task3	Task4
1	**UndefinedRC; Undefined FE; Undefined WP**	Risky							Evolving	Precise			
2	**NoneRC; NoneFE; NoneWP**												
3	**UndefinedRC; Undefined FE; UndefinedWP**		Risky				Fixed						Precise
4	**NoneRC; NoneFE; NoneWP**												
5	**Riksy; MixedFE; NoneWP**		Risky	Risky	Risky	Evolving			Fixed				
6	**UndefinedRC; NoneFE; NoneWP**		Risky										
7	**UndefinedRC; NoneFE; NoneWP**				Cautious								
8	**MixedRC; NoneFE; NoneWP**	Cautious	Risky										
9	**UndefinedRC; MixedFE; UndefinedWP**	Risky				Fixed	Evolving						Weak
10	**Risky; NoneFE; NoneWP**		Risky		Risky								
11	**UndefinedRC; NoneFE; NoneWP**				Risky								
12	**Risky; NoneFE; UndefinedWP**	Risky		Risky	Risky						Precise		
13	**Risky; NoneFE; NoneWP**	Risky			Risky								
14	**UndefinedRC; Undefined FE; NoneWP**	Cautious							Fixed				
15	**UndefinedRC; UndefinedFE; UndefinedWP**		Cautious			Fixed							Weak
16	**MixedRC; MixedFE; NoneWP**	Cautious	Risky					Evolving	Fixed				
17	**UndefinedRC; UndefinedFE; UndefinedWP**				Risky		Fixed				Weak		

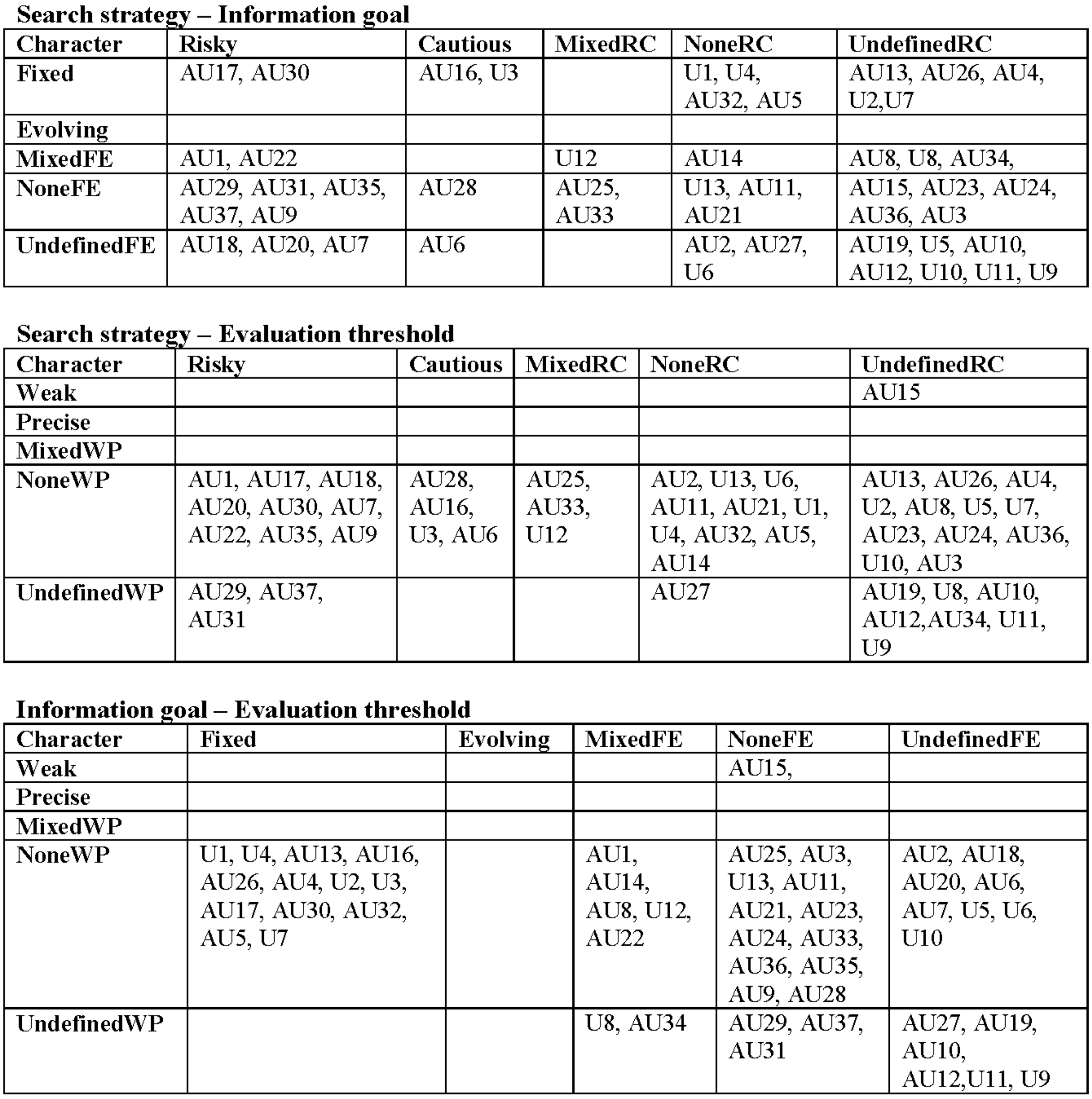
Table 8.4: An example of how to identify user types

8.1.2 Grouping Users Based on Their User Types

From the second column of Table 8.4 we can see that each user has more than one character. Can we independently group the users along each criterion? The following steps will describe how we categorize the 50 users into character groups and how we identify each users' character.

Step 1 - Put the users into character cross tables. As we can see in Table 7.2, there are three character criteria namely: search strategy, information goal and evaluation threshold. Each criterion contains two characters: risky and cautious, fixed and evolving, weak and precise respectively. However, in reality we find five characters in each criterion from our qualitative data, such as risky, cautious, mixedRC, noneRC and undefinedRC. Firstly, we make a cross table for each pair of character criteria: for instance, one cross table between search strategy and information goal; one cross table between search strategy and evaluation threshold; one cross table between information goal and evaluation threshold. Secondly, we assign the 50 users into every cross table. Each cross table has six rows and six columns. Each cell indicates the number of users identified as the crossed characters. Table 8.5 shows the three assigned cross tables. The Chi square test on the independence between the five row characters and the five column characters shows there is no significant relationship between the two categorical variables, which suggests that we can analyze the row or column characters independently.

Step 2 - Group the users. From Table 2 and 3 of Table 8.5 we can see there is insufficient variation among the characters in the evaluation threshold criterion. For instance, only one user shows weak character and the rest of the users carries noneWP and undefinedWP characters, so we decide we are not going to analyze the characters in this criterion any further. We then focus on the other two criteria: search strategy and information goal. The columns of the top table of the Table 8.5 show 12 risky users, 4 cautious users, 3 mixedRC users, 11 noneRC users and 20 undefinedRC users. The rows of of the table 1 shows 13 users with fixed goals, 7



Search strategy – Information goal

Character	Risky	Cautious	MixedRC	NoneRC	UndefinedRC
Fixed	AU17, AU30	AU16, U3		U1, U4, AU32, AU5	AU13, AU26, AU4, U2,U7
Evolving					
MixedFE	AU1, AU22		U12	AU14	AU8, U8, AU34,
NoneFE	AU29, AU31, AU35, AU37, AU9	AU28	AU25, AU33	U13, AU11, AU21	AU15, AU23, AU24, AU36, AU3
UndefinedFE	AU18, AU20, AU7	AU6		AU2, AU27, U6	AU19, U5, AU10, AU12, U10, U11, U9

Search strategy – Evaluation threshold

Character	Risky	Cautious	MixedRC	NoneRC	UndefinedRC
Weak					AU15
Precise					
MixedWP					
NoneWP	AU1, AU17, AU18, AU20, AU30, AU7, AU22, AU35, AU9	AU28, AU16, U3, AU6	AU25, AU33, U12	AU2, U13, U6, AU11, AU21, U1, U4, AU32, AU5, AU14	AU13, AU26, AU4, U2, AU8, U5, U7, AU23, AU24, AU36, U10, AU3
UndefinedWP	AU29, AU37, AU31			AU27	AU19, U8, AU10, AU12,AU34, U11, U9

Information goal – Evaluation threshold

Character	Fixed	Evolving	MixedFE	NoneFE	UndefinedFE
Weak				AU15,	
Precise					
MixedWP					
NoneWP	U1, U4, AU13, AU16, AU26, AU4, U2, U3, AU17, AU30, AU32, AU5, U7		AU1, AU14, AU8, U12, AU22	AU25, AU3, U13, AU11, AU21, AU23, AU24, AU33, AU36, AU35, AU9, AU28	AU2, AU18, AU20, AU6, AU7, U5, U6, U10
UndefinedWP			U8, AU34	AU29, AU37, AU31	AU27, AU19, AU10, AU12,U11, U9

Table 8.5: The 50 users assigned into three character cross tables

users with mixed fixed and evolving goals, 16 users with neither fixed nor evolving goals (based on the Step3 of Section 8.1.1), 14 users with undefined fixed or evolving goals (based on the Step3 of Section 8.1.1).

8.1.3 Linking User Evaluation Results to User Types

In this Section we will go back to the quantitative data (search results) and the users' comments from the questionnaires and informal interviews of the 50 users. We will look at each of 5 character groups respectively from the search strategy and information goal criteria in Table 1 of the Table 8.5. The analysis will be carried out from different angles based on different character groups.

Step 1 - Users' performance and opinions with respect to search strategy. We firstly group the 50 users into five character groups under the search strategy category: risky, cautious, mixedRC, noneRC and undefinedRC. Then we check the search performance in terms of the search precision and users' opinions on the best performing system from the quantitative data.

Step 2 - Users' performance and opinions with respect to information goal. We also group the 50 users into five character groups under the information goal category: fixed, evolving, mixedFE, noneFE and undefinedFE. Then we check the search performance in terms of the search precision and users' opinions on the best performing system from the quantitative data.

Step 3 - Users' comments with respect to search strategy. Apart from checking the quantitative data, we also analyze the 50 users comments under the 5 character groupings under the search strategy criterion. We firstly group the users with the same criteria as Step 1. The users' comments can be classified into three classes: expected image search tool, search experience and suggestions to the evaluation systems. We then analyze the users' comments in different classes based on different character groups.

Step 4 - Users' comments with respect to information goal. This analysis will follow the same approach as step 3, but use the same grouping criteria as Step 2.

After analyzing the users' performance and preferences based on the 5 character groups for each criterion, we find that some user characters correspond to similar preferences and performance. Thus, we decide to carry out more analysis of a coarser grouping, by merging the 5 characters into two groups: with style and no style.

Step 5 - Users' performance and opinions with respect to search strategy

(2 groups). We merge the five user characters into two groups under the search strategy criterion: with-RC-style (including risky, cautious, and mixedRC) and no-RC-style (including noneRC and undefinedRC). We perform the same analysis with respect to the two groups as Step 1.

Step 6 - Users' performance and opinions with respect to information goal (2 groups). We merge the five user characters into two groups under the information goal criterion as well: with-FE-style (including fixed, evolving, mixedFE), and no-FE-style (including noneFE and undefinedFE). We perform the same analysis with respect to the two groups as Step 2.

Step 7 - Users' comments with respect to search strategy (2 groups). We do the same analysis as Step 3, with respect to the 2 groups for the search strategy criterion.

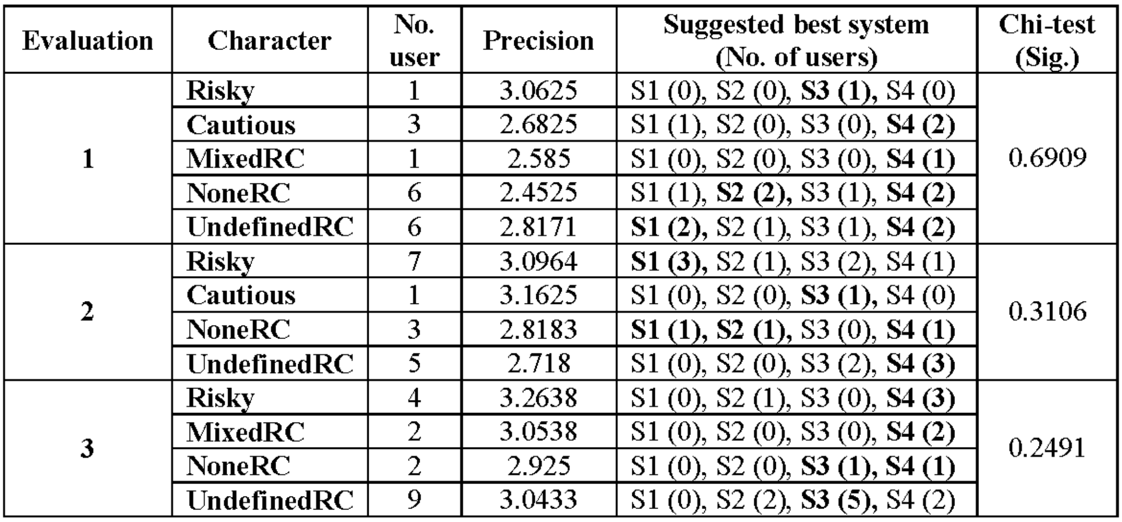
Step 8 - Users' comments with respect to information goal (2 groups). Again, this analysis will follow the same methodology as Step 3 with respect to the 2 groups for the information goal criterion.

8.2 Findings and Suggestions

In the following subsections, we will report the findings from the qualitative and quantitative data analysis based on the ISE Model. The findings and suggestions will be organized based on different test data and different group settings of each character criterion for the three evaluations. The t-test is used to test the statistical difference between different user groups on their average precision. We also use Chi-square(d) test to test the statistical interaction between different user characters based on the testing data.

8.2.1 Precision and Suggested Best System with Respect to Search Strategy (Five Groups)

The analysis results under this test setting are shown in Table 8.6.



Evaluation	Character	No. user	Precision	Suggested best system (No. of users)	Chi-test (Sig.)
1	**Risky**	1	3.0625	S1 (0), S2 (0), **S3 (1),** S4 (0)	0.6909
	Cautious	3	2.6825	S1 (1), S2 (0), S3 (0), **S4 (2)**	
	MixedRC	1	2.585	S1 (0), S2 (0), S3 (0), **S4 (1)**	
	NoneRC	6	2.4525	S1 (1), **S2 (2),** S3 (1), **S4 (2)**	
	UndefinedRC	6	2.8171	**S1 (2),** S2 (1), S3 (1), **S4 (2)**	
2	**Risky**	7	3.0964	**S1 (3),** S2 (1), S3 (2), S4 (1)	0.3106
	Cautious	1	3.1625	S1 (0), S2 (0), **S3 (1),** S4 (0)	
	NoneRC	3	2.8183	**S1 (1), S2 (1),** S3 (0), **S4 (1)**	
	UndefinedRC	5	2.718	S1 (0), S2 (0), S3 (2), **S4 (3)**	
3	**Risky**	4	3.2638	S1 (0), S2 (1), S3 (0), **S4 (3)**	0.2491
	MixedRC	2	3.0538	S1 (0), S2 (0), S3 (0), **S4 (2)**	
	NoneRC	2	2.925	S1 (0), S2 (0), **S3 (1), S4 (1)**	
	UndefinedRC	9	3.0433	S1 (0), S2 (2), **S3 (5),** S4 (2)	

Table 8.6: Analysis of precision and suggested best system with respect to the five user groups based on the search strategy

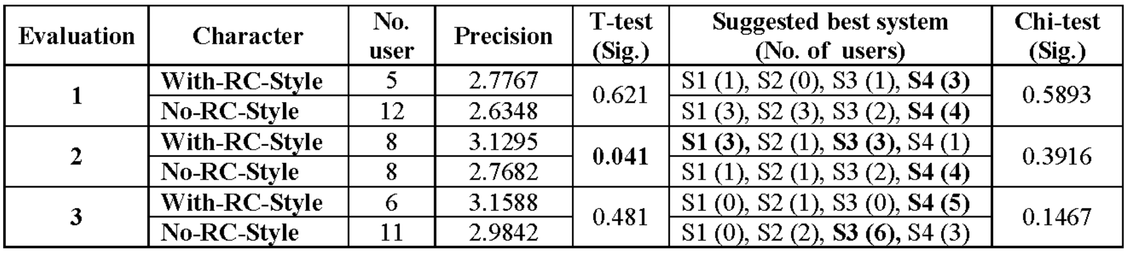
Evaluation1: In terms of "Precision", there is no significant difference among the five occurring characters in E1. As there is only one person showing risky and mixedRC character in E1, we cannot show the difference by a statistical test. From the column of "Suggested best system" we can see that users with a cautious character or any character with cautious elements prefer I4, which is the uInteract interface delivering our four-factor user interaction model. Risky people prefer I3, which is the interface delivering the history and ranking functionalities but no negative functionality. NoneRC and undefinedRC users can be divided into two groups. One group prefers simpler interfaces, and the other group prefers the interface with richer functionalities.

Evaluation2: Four characters occurred in E2: risky, cautious, noneRC and undefinedRC. Again the mean average precisions from different characters do not show any statistically significant. From the column of "Suggested best system" we can see risky users prefer OM1, which is the system delivering the increasing profile of the Ostensive Model. Cautious users prefer OM3, which delivers the flat profile of the Ostensive Model. NoneRC and undefinedRC users both prefer OM4 that delivers the current profile of the Ostensive Model, but noneRC users also like OM1 and OM2 that deliver the increasing and decreasing profile of the Ostensive Model.

Evaluation3: Risky, mixedRC, noneRC and undefinedRC occurred in E3. There is no statistical difference (p=0.180) among the average precisions for the four characters. The preferences on “Suggested best system” focus on FFUIM3 that delivers the interaction model with the relevance region + time + frequency and FFUIM4 that delivers the relevance region + time + relevance level + frequency. Risky and mixedRC users mostly like FFUIM4 only, and undefinedRC likes FFUIM3. Only noneRC likes both FFUIM3 and FFUIM4.

8.2.2 Precision and Suggested Best System with Respect to Search Strategy (Two Groups)

The analysis results under this test setting are shown in Table 8.7.



Evaluation	Character	No. user	Precision	T-test (Sig.)	Suggested best system (No. of users)	Chi-test (Sig.)
1	**With-RC-Style**	5	2.7767	0.621	S1 (1), S2 (0), S3 (1), **S4 (3)**	0.5893
	No-RC-Style	12	2.6348		S1 (3), S2 (3), S3 (2), **S4 (4)**	
2	**With-RC-Style**	8	3.1295	**0.041**	**S1 (3),** S2 (1), **S3 (3),** S4 (1)	0.3916
	No-RC-Style	8	2.7682		S1 (1), S2 (1), S3 (2), **S4 (4)**	
3	**With-RC-Style**	6	3.1588	0.481	S1 (0), S2 (1), S3 (0), **S4 (5)**	0.1467
	No-RC-Style	11	2.9842		S1 (0), S2 (2), **S3 (6),** S4 (3)	

Table 8.7: Analysis on precision and suggested best system with respect to the two user groups based on the search strategy

Evaluation1: From the column of “Precision”, we can see that there is no difference between with-RC-style character and no-RC-style character (p=0.621) in this evaluation. From the column of “Suggested best system”, we can see that users with-RC-style and no-RC-style both prefer I4, and users with-RC-style prefer I4 more than users with no-RC-style.

Evaluation2: There is a significant difference between with-RC-style users and no-RC-style users on their mean average precision (p=0.041). The users with-RC-style perform better than the users with no-RC-style. This suggests that users with a style engaged with the evaluation and understood the tasks and system better than the users with no style. From the column "Suggested best system", we can see users with-RC-style prefer OM1 and OM3, which delivers the increasing and flat profile of the Ostensive Model. On the other hand, users with no-RC-style prefer OM4 that delivers the current profile of the Ostensive Model.

Evaluation3: There is no significant difference between the users with-RC-style and the users with no-RC-style on the mean average precision (p=0.481). The users with style prefer OM4, and the users with no style prefer OM3. This may be because users with style like the rich functionality system more than users with no style.

8.2.3 Precision and Suggested Best System with Respect to Information Goal (Five Groups)

The analysis results under this test setting are shown in Table 8.8.

Evaluation	**Character**	**No. user**	**Precision**	**Suggested best system (No. of users)**	**Chi-test (Sig.)**
1	**Fixed**	8	2.6172	S1 (1), S2 (1), S3 (1), **S4 (5)**	0.5651
	MixedFE	2	2.8538	**S1 (1),** S2 (0), **S3 (1),** S4 (0)	
	NoneFE	5	2.7315	**S1 (2),** S2 (1), S3 (1), S4 (1)	
	UndefinedFE	2	2.505	S1 (0), **S2 (1),** S3 (0), **S4 (1)**	
2	**Fixed**	5	2.93	S1 (1), S2 (0), **S3 (2), S4 (2)**	0.4146
	MixedFE	2	2.7363	S1 (0), **S2 (1),** S3 (0), **S4 (1)**	
	NoneFE	2	3.0713	**S1 (1), S2 (1),** S3 (0), S4 (0)	
	UndefinedFE	7	2.9454	S1 (2), S2 (1), **S3 (3),** S4 (1)	
3	**MixedFE**	3	3.1617	S1 (0), S2 (0), S3 (0), **S4 (3)**	0.2710
	NoneFE	9	3.1586	S1 (0), S2 (2), S3 (3), **S4 (4)**	
	UndefinedFE	5	2.928	S1 (0), S2 (1), **S3 (3),** S4 (1)	

Table 8.8: Analysis on precision and suggested best system with respect to the five user groups based on the information goal

Evaluation1: From the column “Precision”, we can see that there is no significant difference among the five occurring characters in E1 (p=0.241). From the column “Suggested best system”, we can see that users with fixed information goal prefer I4, which is the uInteract interface that delivers our four-factor user interaction model. NoneFE users prefer I1, which is the baseline interface with relevance feedback only. MixedFE users are equally populated on supporting I1 and I3. Half of the undefinedFE users prefer I2 and the other half prefer I4.

Evaluation2: Four user characters occurred in E2: fixed, mixedFE, NoneFE and undefinedFE. Again the mean average precision of the four user characters do not show a significant difference (p=0.896). From the column of “Suggested best system”, we can see the users with fixed goals prefer OM3 and OM4, which are the systems deliver the flat and current profiles of the Ostensive Model. The users with mixedFE prefer OM2 and OM4, which delivers the decreasing and current profiles of the Ostensive Model. NoneRC prefers OM1 and OM2 that deliver the increasing and decreasing profiles of the Ostensive Model. Finally, the undefinedFE like OM3 that delivers the flat profile of the Ostensive Model.

Evaluation3: MixedFE, noneFE and undefinedFE occurred in E3. There is no statistical difference (p=0.368) among the mean average precision of the four user characters. The preferences of the characters on “Suggested best system” focus on FFUIM3 that delivers the interaction model with relevance region + time + frequency and FFUIM4 that delivers the relevance region + time + relevance level + frequency. MixedFE and noneFE users like FFUIM4, and undefinedFE users like FFUIM3.

8.2.4 Precision and Suggested Best System with Respect to Information Goal (Two Groups)

The analysis results under this test setting are shown in Table 8.9.

Evaluation	Character	No. user	Precision	T-test (Sig.)	Suggested best system (No. of users)	Chi-test (Sig.)
1	With-FE-Style	10	2.7355	0.846	S1 (2), S2 (1), S3 (2), **S4 (5)**	0.6895
	No-FE-Style	7	2.6181		**S1 (2), S2 (2),** S3 (1), **S4 (2)**	
2	With-FE-Style	7	2.8331	0.279	S1 (1), S2 (1), S3 (2), **S4 (3)**	0.5088
	No-FE-Style	9	3.0083		**S1 (3),** S2 (2), **S3 (3),** S4 (1)	
3	With-FE-Style	3	3.1617	0.090	S1 (0), S2 (0), S3 (0), **S4 (3)**	0.1409
	No-FE-Style	14	3.0433		S1 (0), S2 (3), **S3 (6),** S4 (5)	

Table 8.9: Analysis on precision and suggested best system with respect to the two user groups based on the information goal

Evaluation1: From the column of "Precision", we can see that there is no significant difference between with-FE-style users and no-FE-style users (p=0.846) in this evaluation. From the column of "Suggested best system", we can see that users with-FE-style prefer I4. Users with no-FE-style prefer I1, I2 and I4.

Evaluation2: There is no significant difference between with-FE-style users and no-FE-style users on their mean average precision (p=0.279). From the column of "Suggested best system", we can see users with-FE-style prefer OM4 that delivers the current profile of the Ostensive Model. On the other hand, the users with no-FE-style prefer OM3 that delivers the flat profile of the Ostensive Model.

Evaluation3: There is no significant difference between the users with-FE-style and the users with no-FE-style on the mean average precision (p=0.090). The users with-FE-style prefer FFUIM4, and the users with no-FE-style prefer FFUIM3. This may be because users with-FE-style prefer the rich functionality system.

8.2.5 Comments with Respect to the Five User Groups Based on Search Strategy

The analysis results under this test setting are shown in Table 8.10[5].

[5] Table 8.10 shows the users' response to the three categories: expected image search tool, search experience and suggestions to the evaluation systems, with regard to the five user groups based on users' search strategy. In each user group column, the percentage shows the number of users responding to a comment. $N = x$ means there are x users in that user group combined across three evaluations.

Category	Comment	Risky (N=12)	Cautious (N=4)	MixedRC (N=3)	NoneRC (N=11)	UndefinedRC (N=20)
Expected image search tool	Good and large data source	0	0.25	0.33	0.36	0.25
	Fast	0.25	0	0	**0.55**	0.05
	Accurate and diverse result	**0.75**	**0.5**	0.33	**0.64**	**0.55**
	Easy to use	0.25	0.25	**0.67**	0.18	0.25
	Rich functionalities for search	**1**	**1**	**1**	**1**	**1**
Search experiences	Satisfied with the result	0.25	**0.75**	1	0.27	0.25
	Not satisfied with the result	0.25	0	0	0.18	0.25
	Satisfied with the systems	0.42	**0.75**	**0.67**	0.45	0.25
	Not satisfied with the systems	0.08	0.25	0	0.18	0.1
	Tasks are interesting and clear	0.25	0.5	0.33	**0.64**	0.35
	Had goal in mind	0.08	0.25	0	0.18	0.15
	Relevance feedback is useful	0.33	**0.75**	0	0.45	0.25
	Negative query is useful	**0.67**	**0.75**	**0.67**	0.36	**0.6**
	Negative result is useful	0.17	0.25	0.33	0.09	0.35
	Ranking is useful	**0.58**	**0.5**	**0.67**	0.45	0.45
	History is useful	0.33	0.25	0.33	0	**0.5**
	No need to use functions although they could be useful	0.17	**0.75**	0	0.36	0.1
	System accuracy depending on the tasks	0.17	0	0	0.09	0.1
	Hard to decide the relevant result	0	0	0.33	0.36	0.2
	Search accuracy drop with more image examples in the query	0.08	0	0	0	0.1
	More image examples in the query improve the search result	0	0.25	0	0	0.15
	Functions are more useful when the tasks are more difficult.	0	0	0	0	0.15
	Initial idea changed and the system supported the change	0	0	0	0.18	0.15
	Content-based image search is better	0.33	0	**1**	0.27	0.1
	Keyword-based image search is better	0.08	0	0	0.18	0.2
Suggestions to the evaluation systems	Combine keyword-based and content-based search	0.17	0.5	0	0.36	0.1
	Make the negative functionalities optional	0.25	0	0.33	0.27	0.15
	Improve history functionality	0.25	0	0	0.45	0.25
	Improve ranking functionality	0.25	0	0	0.36	0.15
	Improve negative functionality	0.08	0	**1**	0	0.25
	General improvement of the interface	**1**	0.25	**0.67**	**0.73**	**0.5**

Table 8.10: Analysis on comments with respect to the five user groups based on the search strategy

Risky: The users with a risky search strategy prefer accurate and diverse results and care less about the data source quality and where they search from. They prefer rich functionalities to support different search aspects, so that they can find good results quickly and easily. They judge the effectiveness of the system depending on the tasks they perform. They tend to think the system is good when they perform an easy task using the system and get good results fairly quickly; otherwise, they think the system is poor. As our evaluation systems support multiple images for each query, the users with risky search strategy feel the search accuracy drops with more image examples in the query. This might be because they are likely to provide diverse images as query examples based on colour, shape and semantic relevance, which does not suit the nature of our colour, but only based image search evaluation systems. However, a risky user could perform quite well if s/he gets the supportive functions needed. This is why risky users provided many useful suggestions about improving the usability of the evaluation systems, such as adding an egg timer, image zooming, and incorporating drag and drop, etc.

Cautious: The users with cautious search strategy are another group that showed a clear pattern. Like risky users, they hope the search system is accurate and with rich functionalities to support different search aspects. They do not care much about the search speed. This might be because cautious people are usually patient. They are more satisfied with the search results and the evaluation systems than risky users. They did not need to use all of the provided functionalities on completing some tasks, although they are more likely to think the functions could be useful. The difference between risky users and cautious users is that the cautious users feel that using more image examples in a query improves the search results. This might be because cautious users are more likely to be careful with query refinement and they understand the nature of the colour-based evaluation systems and the tasks better than risky users. Half of the cautious user population think the tasks are interesting and clear, and thus they are satisfied with the search performance. They suggested only minor improvements to the evaluation systems, such as better graphic design for the interfaces. They strongly suggested combining keyword-based and content-based search. This might also because it is hard for cautious users to change their search strategy completely. They like the content-based search strategy, but they also want to keep their normal keyword-based search strategy.

MixedRC: Some comments provided by the users with mixedRC are similar to the comments from the risky and cautious users. For instance, they do not care about the speed of the system, they like rich functionalities on the image search system, they are satisfied with the search results and the evaluation systems, they think the negative query and query scoring functionalities are useful and have many suggestions on improving the evaluation systems. However, they commented on something that the risky and cautious users have not mentioned. For instance, they strongly believe the image search tool is easy to use, they think the content-based image search strategy is better than a keyword-based search strategy, and they would like to see the negative results become optional.

NoneRC: The users with noneRC like fast and accurate systems and prefer rich functionalities to support different search aspects. They think the tasks are interesting and clear. They do not think the history functionality is useful at all. They find that it is hard to decide the relevant results for the tasks. Their initial search idea changes during the searching but they think the system supports the change well. They strongly suggest improving the usability of the query history and query image scoring functionalities by showing thumbnail images in query history section, ranking the query images by a slide bar or dragging and dropping to a different position in a query. They do not have many comments on the negative functionalities.

UndefinedRC: The users with undefinedRC like an accurate and rich functionality search system. They are more likely to think the negative query functionality is useful in the evaluation systems. They also like the negative result functionality because they think they get to know the data collection quality better by seeing the negative results in the result panel. They are more likely to think the query history functionality is useful. Like risky users, they also feel the search accuracy drops with more image examples included in a query. They feel the functionalities are more useful when they perform more difficult tasks. Their initial search idea changes during the search, and the evaluation systems support the change well. They think the usability of the functionalities can be improved by showing query history automatically rather than having to press reset, showing diverse negative results rather than based on colour only, showing page number, etc.

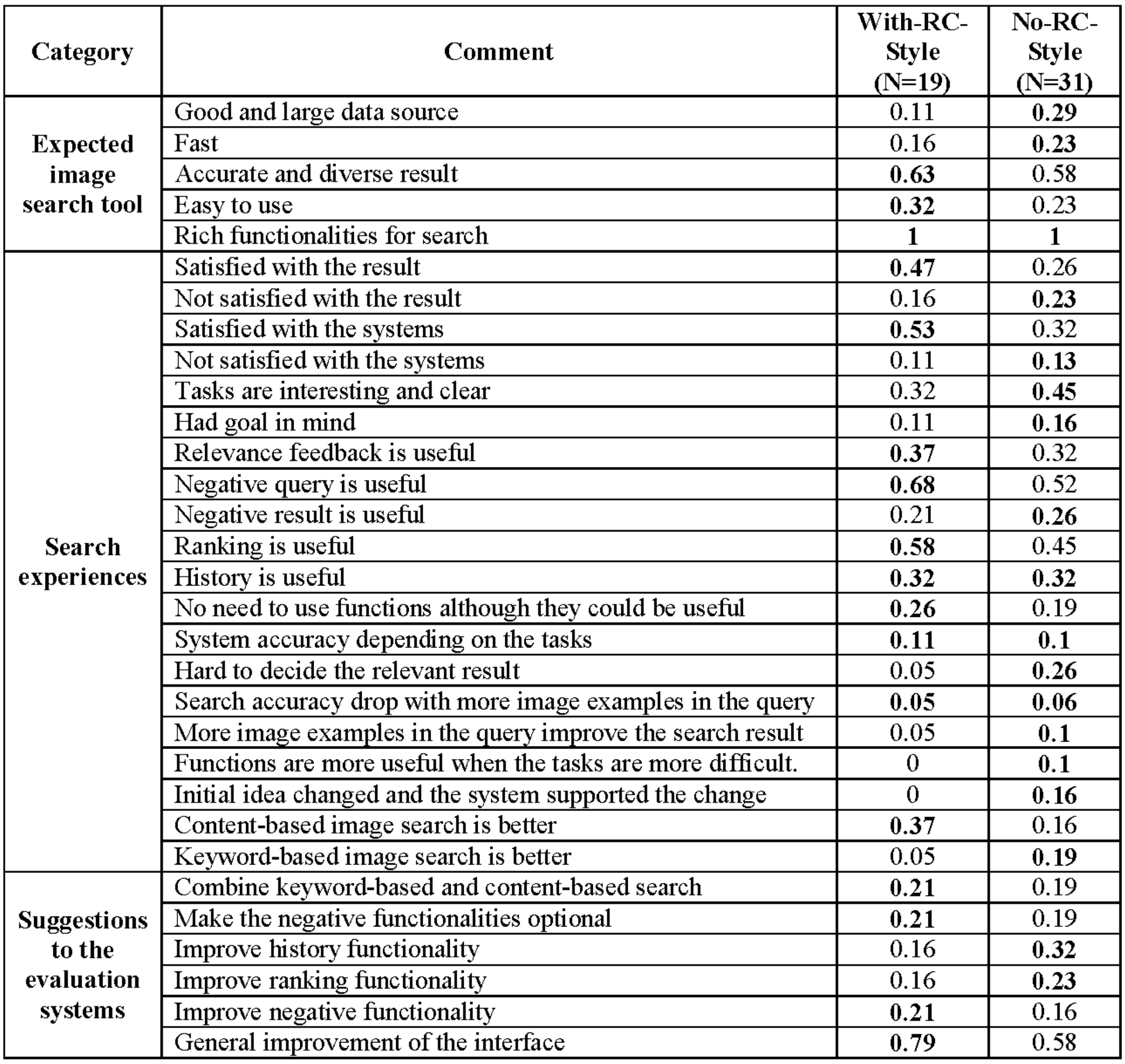
8.2.6 Comments with Respect to Two User Groups Based on Search Strategy

The analysis results under this test setting are shown in Table 8.11[6].

With-RC-style: The users with-RC-style like accurate and easy to use image search tools with rich functionalities. They are more likely to be satisfied with the search results and the evaluation systems. They are more likely to think the positive and negative feedback functions are useful. They are also likely to think the query image scoring and query history functions are useful. They prefer content-based or content-based related search strategy. Whilst they think all the provided functionalities are useful, they suggest to improve the general functions, such as making negative results optional, image zooming, and incorporating drag and drop, etc.

No-RC-style: The users with no-RC-style expect the image search tool to have a good quality and large data source as well as fast and rich functionalities. They are less satisfied with the search results and evaluation systems than the users with-RC-style. They think that the tasks are interesting and clear, and set a goal before

[6] Table 8.11 shows users' response to the three categories: expected image search tool, search experiences and suggestions to the evaluation systems, with regard to the two user groups based on users' search strategy. In each user group column, the percentage shows the amount of users respond to a comment. $N = x$ means there are x users in that user group combined across three evaluations.



Category	Comment	With-RC-Style (N=19)	No-RC-Style (N=31)
Expected image search tool	Good and large data source	0.11	**0.29**
	Fast	0.16	**0.23**
	Accurate and diverse result	**0.63**	0.58
	Easy to use	**0.32**	0.23
	Rich functionalities for search	**1**	**1**
Search experiences	Satisfied with the result	**0.47**	0.26
	Not satisfied with the result	0.16	**0.23**
	Satisfied with the systems	**0.53**	0.32
	Not satisfied with the systems	0.11	**0.13**
	Tasks are interesting and clear	0.32	**0.45**
	Had goal in mind	0.11	**0.16**
	Relevance feedback is useful	**0.37**	0.32
	Negative query is useful	**0.68**	0.52
	Negative result is useful	0.21	**0.26**
	Ranking is useful	**0.58**	0.45
	History is useful	**0.32**	**0.32**
	No need to use functions although they could be useful	**0.26**	0.19
	System accuracy depending on the tasks	**0.11**	**0.1**
	Hard to decide the relevant result	0.05	**0.26**
	Search accuracy drop with more image examples in the query	**0.05**	**0.06**
	More image examples in the query improve the search result	0.05	**0.1**
	Functions are more useful when the tasks are more difficult.	0	**0.1**
	Initial idea changed and the system supported the change	0	**0.16**
	Content-based image search is better	**0.37**	0.16
	Keyword-based image search is better	0.05	**0.19**
Suggestions to the evaluation systems	Combine keyword-based and content-based search	**0.21**	0.19
	Make the negative functionalities optional	**0.21**	0.19
	Improve history functionality	0.16	**0.32**
	Improve ranking functionality	0.16	**0.23**
	Improve negative functionality	**0.21**	0.16
	General improvement of the interface	**0.79**	0.58

Table 8.11: Analysis on comments with respect to the two user groups based on the search strategy

the search. They find showing negative results in the result panel is useful, which helps them to discover what is available in the collection. They find it is hard to decide the relevant results for the tasks as their initial goal keeps changing during the search. However, they agree that the systems supports the change well. They also find the functionalities are even more supportive when they perform difficult exploratory tasks. They especially suggest that improvements can be made to the query history, negative query and query image scoring functionalities.

8.2.7 Comments with Respect to Five User Groups Based on the Information Goal

The analysis results under this test setting are shown in Table 8.12[7].

[7] Table 8.12 shows users' response to the three categories: expected image search tool, search experiences and suggestions to the evaluation systems, with regard to the five user groups based on users' information goal. In each user group column, the percentage shows the number of users responding to a comment. $N = x$ means there are x users in that user group combined across three evaluations.

Category	Comment	Fixed (N=13)	MixedFE (N=7)	NoneFE (N=16)	UndefinedFE (N=14)
Expected image search tool	Good and large data source	0.23	0.29	0.31	0.07
	Fast	0.08	0.29	0.31	0.14
	Accurate and diverse result	**0.62**	0.43	**0.69**	**0.57**
	Easy to use	0	0.29	0.38	0.36
	Rich functionalities for search	**1**	**1**	**1**	**1**
Search experiences	Satisfied with the result	0.31	0.29	**0.56**	0.14
	Not satisfied with the result	0	0.29	0.31	0.21
	Satisfied with the systems	0.23	0	**0.88**	0.21
	Not satisfied with the systems	0.23	0.14	0.06	0.07
	Tasks are interesting and clear	0.15	**0.86**	0.44	0.36
	Had goal in mind	0.31	0.14	0	0.14
	Relevance feedback is useful	0.38	0.29	0.19	**0.5**
	Negative query is useful	0.38	0.29	**0.88**	**0.57**
	Negative result is useful	0.23	0.43	0.19	0.21
	Ranking is useful	0.23	**1**	**0.56**	0.43
	History is useful	0.23	**0.57**	0.31	0.29
	No need to use functions although they could be useful	0.15	0.43	0.13	0.29
	System accuracy depending on the tasks	0	0	0	0.36
	Hard to decide the relevant result	0	0.14	0.31	0.21
	Search accuracy drop with more image examples in the query	0	0.14	0.13	0
	More image examples in the query improve the search result	0.08	0	0.19	0
	Functions are more useful when the tasks are more difficult.	0.08	0.29	0	0
	Initial idea changed and the system supported the change	0	0	0.25	0.07
	Content-based image search is better	0.31	0.29	0.31	0.07
	Keyword-based image search is better	0.23	0.14	0.13	0.07
Suggestions to the evaluation systems	Combine keyword-based and content-based search	0.23	0.14	0.19	0.21
	Make the negative functionalities optional	0	0.14	0.38	0.21
	Improve history functionality	0.15	**0.57**	0.19	0.29
	Improve ranking functionality	0.23	0.29	0.19	0.14
	Improve negative functionality	0.15	0.43	0.13	0.14
	General improvement of the interface	**0.92**	0.29	**0.56**	**0.71**

Table 8.12: Analysis on comments with respect to the five user groups based on the information goal

Fixed: the users with a fixed information goal like accurate systems with rich functionalities. They are satisfied with the search results, and they are basically satisfied with the evaluation systems, though feel that the usability of the interface needs could be improved. They have a clear information goal in mind before starting the search, and the goal does not change during the search. They feel the search results get increasingly better with each query refinement. They find it easy to make decisions on results selection. Whilst they prefer the content-based search, they also like keyword-based search, thus they suggest combining the two. They like all the functionalities provided, but again they think the usability of some functionalities can be improved, for example, by ranking query images by a scale bar or dragging and dropping, showing image thumbnails in the query history section, starting with keyword-based search, etc.

MixedFE: The users with mixedFE have fewer expectations of the system accuracy than users with other characters, but they have the same expectation with the other types of users on rich functionalities to support different search aspects. They think the tasks are interesting and clear. They think all the provided functionalities are useful especially the query image scoring and query history functionality. They also find that the functionalities are more useful when they perform more difficult tasks. Whilst they prefer the content-based search, they also like keyword-based search and the combination of keyword-based and content-based search. As they tried many functionalities for completing the tasks, they provide numerous suggestions on improving the functions, such as ranking query images by a slide bar, showing image thumbnails in query history section, providing a colour histogram or pie chart for selecting negative colour examples, etc.

NoneFE: Compared to the users with other characters, users with noneFE like good a quality and large data source, a fast and easy to use system, accurate search results and rich functions to support different search aspects. These users are satisfied with both the search results and the evaluation systems. They think the tasks are fairly interesting and clear. They do not know what they are looking for before they start the search. They think the negative query is extremely useful. They also like the query image scoring function. They sometimes find it hard to decide the relevance of the results for the tasks. Their ideas change during the search and the systems support the changes well. They suggest making the negative results optional. As with other types of users, they think the usability of the interface can be improved by providing drag and drop and image zoom functionality.

UndefinedFE: The undefinedFE users expect the image search tool to be accurate and easy to use, and have rich functionalities. They are more satisfied with the search experience with the systems than with the search results because they judge the system accuracy based on the complexity of the tasks. They are satisfied with the search results when they perform easier tasks, and they are not satisfied with the search results when they perform harder tasks. They think the positive and negative feedback functions are useful. They suggested improvements to the interfaces of the evaluation systems, such as adding drag and drop, image zoom and providing diverse negative results rather than based solely on colour only, etc.

8.2.8 Comments with Respect to Two User Groups Based on Information Goal

The analysis results under this test setting are shown in Table 8.13[8].

With-FE-style: The users with-FE-style prefer the quality and size of the data source where they search, and they like accurate search results and rich functionalities to support different search aspects. They have a clear information goal in mind before they start the search. They feel the functionalities are more useful when they perform harder tasks. They are satisfied with the search results but not satisfied

[8] Table 8.13 shows users' responses to the three categories: expected image search tool, search experiences and suggestions to the evaluation systems, with regard to the two user groups based on the users' information goal. In each user group column, the percentage shows the number of users responding to a comment. $N = x$ means there are x users in that user group combined across three evaluations.

Category	Comment	With-FE-Style (N=20)	No-FE-Style (N=30)
Expected image search tool	Good and large data source	**0.25**	0.2
	Fast	0.15	**0.23**
	Accurate and diverse result	0.55	**0.63**
	Easy to use	0.1	**0.37**
	Rich functionalities for search	**1**	**1**
Search experiences	Satisfied with the result	0.3	**0.37**
	Not satisfied with the result	0.1	**0.27**
	Satisfied with the systems	0.15	**0.57**
	Not satisfied with the systems	**0.2**	0.07
	Tasks are interesting and clear	**0.4**	**0.4**
	Had goal in mind	**0.25**	0.07
	Relevance feedback is useful	**0.35**	0.33
	Negative query is useful	0.35	**0.73**
	Negative result is useful	**0.3**	0.2
	Ranking is useful	**0.5**	**0.5**
	History is useful	**0.35**	0.3
	No need to use functions although they could be useful	**0.25**	0.2
	System accuracy depending on the tasks	0	**0.17**
	Hard to decide the relevant result	0.05	**0.27**
	Search accuracy drop with more image examples in the query	0.05	**0.07**
	More image examples in the query improve the search result	0.05	**0.1**
	Functions are more useful when the tasks are more difficult.	**0.15**	0
	Initial idea changed and the system supported the change	0	**0.17**
	Content-based image search is better	**0.3**	0.2
	Keyword-based image search is better	**0.2**	0.1
Suggestions to the evaluation systems	Combine keyword-based and content-based search	**0.2**	**0.2**
	Make the negative functionalities optional	0.05	**0.3**
	Improve history functionality	**0.3**	0.23
	Improve ranking functionality	**0.25**	0.17
	Improve negative functionality	**0.25**	0.13
	General improvement of the interface	**0.7**	0.63

Table 8.13: Analysis on comments with respect to the two user groups based on the information goal

with the evaluation systems, although they have tried some of the functionalities and have agreed that the functionalities can be useful. They suggested improvements could be made to the ease of use of the systems. Users with-FE-style prefer content-based image search, although they also like keyword-based image search and the combined content-based and keyword-based image search.

No-FE-style: The users with no-FE-style like a fast, accurate and easy to use image search tool. Like users with-FE-Style, they prefer rich functionalities to support the search. They are satisfied with the search results as well as the evaluation systems. They think the negative query functionality is useful. They judge the effectiveness of the systems based on the tasks. They say the system is better when the task they are performing is easier. For some tasks they find it hard to decide which result is relevant. Their initial idea changes during the search process and they think the system supports the change well. They prefer to have the content-based search element in the search system over a purely keyword-based search systems. They strongly suggest making the negative results optional because they do not feel the negative results are needed for all tasks.

8.3 Summary

In this Chapter, we categorized the 50 users in our user study into different groups based on the ISE user classification model. We have found that only the users grouped based on the search strategy and information goal criteria are evenly spread to every character, so we decided to discuss the characters of these two criteria only. After grouping the 50 subjects into the characters, we have extracted some quantitative and qualitative data from our user study introduced in Chapter 3, such as search precision of the actual search results for the tasks, suggested best evaluation system from the exit questionnaire and users' comments on the entry, post-search and exit questionnaires. Through analyzing these quantitative and qualitative data based on different characters, we have found clear evidence concerning users' different search performances for the tasks, their different search preferences for the evaluation systems, their differing expectations of image search tools, their varying search experiences during the evaluation, and have made suggestions to improve the evaluated systems, for the different user types. The summary of the findings and suggestions (Table 8.14) will be stated in the following sections.

User type	Summary
Risky (R)	1. They prefer I3, increasing profile, and FFUIM4; 2. They like accurate and diverse result, rich functions; 3. They suggest multi-modal search system.
Cautious (C)	1. They prefer I4, and flat profile; 2. They are satisfied with the evaluation systems; 3. They suggest to combine keyword-based and content-based search.
MixedRC	1. They prefer I4, and FFUIM4; 2. They are satisfied with the evaluation systems, and like rich functions; 3. They suggest to make the negative results optional.
NoneRC	1. They do not have a clear preference of the evaluation systems; 2. Their information goal changes during search and the systems support the change; 3. They suggest a trial session before evaluation, and longer time needed.
UndefinedRC	1. They prefer flat profile and FFUIM3, but have no preference of the interfaces; 2. They like negative result function; 3. Their search is more exploratory.
Fixed (F)	1. They prefer I4 and the flat profile; 2. They have a clear information goal before search; 3. They like all the functions provided, but suggest to improve the ease of use.
MixedFE	1. They prefer FFUIM4, but have no clear preference of interfaces and OM profiles; 2. They like all the functions provided; 3. They prefer to combine keyword-based and content-based search.
NoneFE	1. They prefer I1 and FFUIM4, but have no clear preference of the OM profiles; 2. Their information goal changes during search, and the systems support the change; 3. They suggest to make the negative results optional.
UndefinedFE	1. They prefer the flat profile and FFUIM3, but have no clear preference of the interfaces; 2. Their favorite functions are the positive and negative feedback; 3. They are satisfied with the evaluation systems.
With-RC-style	1. They prefer I4, the flat profile and FFUIM3; 2. They are satisfied with the evaluation systems; 3. They suggest to combine the keyword-based and content-based search.
No-RC-style	1. They prefer I4, the current profile and FFUIM3; 2. Their information goal changes during search, and the systems support the change; 3. Their search is more exploratory.
With-FE-style	1. They prefer I4, the current profile and FFUIM4; 2. They have clear information goal before the search starts; 3. They think the easy of use of the systems can be improved.
No-FE-style	1. They prefer the flat profile and FFUIM3, but have no clear preference of the interfaces; 2. They have no clear information goal before the search; 3. They think the search system supported their exploratory search well.

Table 8.14: Summary of the final outcomes with regard to 13 user types

8.3.1 Risky (R)

Risky people prefer the interface with the query history and query image scoring functionalities. They think the increasing profile of the Ostensive Model performs better than the other profiles, and prefer the combination of the four factors of the four-factor user interaction model, namely: relevance region, relevance level, time and frequency.

Risky people focus more on the results and seek out accurate and diverse result images. They prefer rich functionalities to support their search process. They would like to judge the effectiveness of the system depending on the complexity level of the tasks they perform. They complain that the search results get worse when they reformulate the query with more image examples.

These preferences and behaviours match the risky character definition in terms of Information Foraging Theory. They like to move between patches frequently. They prefer that the system provides good results in early pages because they do not like to look through many pages of search results. They always reformulate the query to change the patch, so that they need rich functionality to support the reformulation. The risky people like to explore different aspects of the tasks, so that colour-based image search seems insufficient to support their diverse requirements. Therefore, we need to improve the interface and combine different multi-modal search systems for the risky people. As soon as their search behaviours are supported, they can perform quite well.

8.3.2 Cautious (C)

Cautious people prefer the uInteract interface over the other three interfaces. They like the flat profile of the Ostensive Model better than other profiles.

Cautious people do not require a fast search system. They find the search results improve with more image examples in a query. They also like rich functionalities although they do not need to use them for all tasks. They are satisfied with the evaluation systems and search results and so have few suggestions on system improvement. The cautious people largely agreed that keyword-based search and content-based search should be combined.

These preferences and behaviours match the cautious character definition in terms of Information Foraging Theory. They like to stay in one patch consistently. They do not mind spending time looking for good results. They are careful in selecting feedback to refine the queries and search result images. They can adapt to the new content-based search strategy well, but they still like their common search strategy, i.e., keyword-based search.

8.3.3 MixedRC

People with a mixed risky and cautious strategy prefer the uInteract interface. They also like the FFUIM4 system that delivers a combination of the four factors of the four-factor user interaction model, i.e., relevance region, relevance level, time and frequency.

MixedRC people do not mind the search speed. They also like rich search functionalities. They are satisfied with the evaluation systems and search results. They think the provided functionalities are useful, but the functionality can be improved to support ease of use. For example, they suggest making the negative results optional.

The preferences and behaviours of the MixedRC show a combination of risky and cautious characters' preferences and behaviours. This matches the nature of the MixedRC character. Of course, there are also unique comments from MixedRC people that risky and cautious people did not make. For instance, the people with MixedRC character think the content-based image search strategy is better than a keyword-based search strategy.

8.3.4 NoneRC

People with NoneRC (neither risky nor cautious) character can be divided into two groups. One group prefers simpler interfaces such as the basic interface with relevance feedback function only. The other group prefers the interface with richer functionalities such as the uInteract interface. They like the current profile of the Ostensive Model the best, and they also like the increasing and decreasing profiles. They prefer the combinations of the four factors of the four-factor user interaction model, as well as the combination of the three factors of the four-factors user interaction model: relevance region, time and frequency.

NoneRC people like a fast, accurate search system with rich functionalities. They like the negative functionality but do not like the query history functions. They find it is difficult to select the relevant result images for the tasks. Their initial idea changes during the search.

These preferences and search behaviours support the nature of the noneRC people, who do not have risky people's rushing around, nor cautious people's steady behaviours. These people have quite different opinions on the system preferences. Their search behaviours also show they cannot any the clear difference in the performance of the systems. This might be because they need a longer time to develop a new search strategy. It might be even better if we could give them a trial session before the evaluation.

8.3.5 UndefinedRC

People with undefinedRC (undefined risky and cautions) character, can be divided into two groups. One group prefers a simpler interface, such as the baseline interface with relevance feedback only. The other group prefers the interface with rich functionalities, such as the uInteract interface. However, they have consistent positive views on the flat profile of the Ostensive Model and the combination of the three factors of the four-factor user interaction model, namely: relevance region, time and frequency.

UndefinedRC people also like accurate systems with rich functionality. They like the negative and query history functionalities. They think displaying negative results helps them judge the data quality. Like risky users, they also feel the search accuracy drops by adding more image examples in a query. They feel the functionalities are more useful when they perform more difficult tasks. Their initial idea sometimes changes during the search.

From their search preferences and behaviours, we can see the people with undefinedRC have cross characters of risky, cautious, mixedRC and noneRC. Their search process should be more exploratory as they do not have a clear character. This is why they like to see the negative results from the result panel that is designed especially for supporting exploratory search.

8.3.6 Fixed (F)

People with fixed goals prefer the uInteract interface. They also prefer the flat and current profiles of the Ostensive Model.

People with fixed goals like accurate systems with rich functionalities. They had a clear information goal in mind before starting the search and the goal did not change during the search. They feel the search results increasingly improve with each query refinement. They find it easy to make decisions on results selection. They like all of the functionalities, although they think their ease of use could be improved.

These preferences and search behaviours match the fixed character definition very well in terms of Information Foraging Theory. People with fixed goals are clear about what they are looking for. Every movement they make to refine the query improves the results. They find it easy to make decisions on which images to select as feedback or results.

8.3.7 MixedFE

People with mixed fixed and evolving goals equally like the simplest interface with relevance feedback only and the interface with query history and query image scoring. They also prefer the decreasing and current profiles of the Ostensive Model. Their favorite setting is the combination of the four factors of the four-factor user interaction model.

MixedFE people do not care much about the search accuracy, but prefer the rich functions to support their search. They like all the provided functions, especially query history and query image scoring. They feel these functions are more useful when performing more difficult tasks. They would prefer to combine the content-based and keyword-based search together.

From the preferences and behaviours, we can see that the people with mixed fixed and evolving goals combine the nature of the fixed and evolving character in terms of Information Foraging Theory. They sometimes know what they are looking for but sometimes they do not. We can see that people with mixedFE have some similarity with fixed goal people, but it is hard to tell the similarity with evolving goal people because evolving character did not occur in our evaluation.

8.3.8 NoneFE

People with neither fixed nor evolving goals prefer the simplest interface with relevance feedback only. They like the increasing and decreasing profile of the Ostensive Model. Again, the system delivering the combination of the four factors of the four-factor user interaction model is their favorite setting.

NoneFE people like a good quality and large data source to search from. They also like accurate and fast systems. They are satisfied with the evaluation systems and the search results. They state they do not know what they are looking for before the search. Their idea changes during the search. They also find it difficult to select relevant results for the tasks. Although it is not easy to complete the task, they find the provided functions supports the search process well. They think the ease of use of the interface can be improved by, for example, by making the negative results optional.

From the preferences and behaviours of the NoneFE people, we can see these people showed completely different search behaviours from the people with a fixed information goal. For example, they are unsure what they are looking for, before the search starts and their ideas change during the search process also. People with noneFE character prefer the simplest interface. This should be different from what the evolving goal people would prefer. People with evolving information goal would like rich functionality to support their search, especially the query history functionality[9].

8.3.9 UndefinedFE

People with undefinedFE can be divided into two groups. One group prefers the system with relevance feedback and query history functions. The other group prefers the uInteract interface with relevance feedback, query history, query image scoring and negative functions. All of the undefinedFE people like the flat profile of the Ostensive Model and the combined three factors of the four-factor user interaction model, namely: relevance region, time and frequency.

UndefinedFE people expect an accurate and easy to use system with rich functionalities. They are satisfied with the evaluation systems. Their satisfaction with the search results depends on the performance of the tasks. Their prefered functions are the positive and negative feedback.

From their search preferences and behaviours, we can see the people with undefinedRC do not have specific preferences and unique search behaviours. They have cross characters of fixed, mixedFE and noneFE.

From the above analysis of the characters of the two criteria, we find that risky, cautious and fixed goal people have clear and unique preferences and search behaviours. People with mixedRC or mixedFE show the mixed preferences and behaviours. However, like noneFE people, users with undefined FE have an unstable and unclear search pattern. Further, there is no significant difference among the characters on the search precision of the actual search results. It will be interesting

[9] As the evolving character did not occur in our evaluation, we do not have evidence about what the evolving goal people prefer. However, we suppose that the people with evolving goal should like the interfaces with rich functions to support their goal changes during the search based on the definition of the evolving character.

to see the results of combining some of the characters together, and we expect to see a clear pattern through the new groups. The new groups are as follows:

- risky + cautious + mixedRC ⇒ with-RC-style;
- noneRC + undefinedRC ⇒ no-RC-style;
- fixed + mixedFE ⇒ with-FE-style;
- noneFE + undefinedFE ⇒ no-FE-style.

The analysis based on these new groups shows a clear pattern. The analysis summary is set out below.

8.3.10 With-RC-style

People with-RC-style prefer the uInteract interface. They like the increasing and flat profile of the Ostensive Model and the combined three factors of the four-factor user interaction model, namely: relevance region, time and frequency.

People with-RC-style like accurate and easy to use systems with rich functionalities. They are satisfied with the evaluation systems and search results. They like all the functionalities provided on the evaluation systems. They prefer the content-based and content-based related search strategy. They suggest improving the ease of use of the interfaces by combining the keyword-based and content-based searches.

These people have clear preferences. There is also a clear pattern to their search behaviours.

8.3.11 No-RC-style

People with no-RC-style also prefer uInteract interface. They like the current profile of the Ostensive Model. Like the people with-RC-style, they also prefer the setting with three factors of the four-factor user interaction model, namely: relevance region, time and frequency.

People with no-RC-style expect a good quality, large data source and a fast system with rich functionalities. They are not satisfied with the search results and the evaluation systems. They feel their information goal changes during the search. They agree that the system supports the changes well. They think all the functions are useful, especially showing the negative results that helps them judge data quality. They find the functions are more useful when they perform harder tasks.

These people have similar preferences to the people with-RC-style, but there is a big difference shown in their search behaviours and experience. People with no-RC-style carry out more exploratory search. From their comments, we can see their exploratory search is supported by the systems.

8.3.12 With-FE-style

People with-FE-style prefer the uInteract interface. They prefer the current profile of the Ostensive Model and the combination of the four factors of the four-factor user interaction model, namely: relevance region, relevance level, time and frequency.

People with-FE-style like a good quality and large data source to search from. They also like accurate systems with rich functionalities. They have a clear information goal before the search starts. They are satisfied with the search results, although they think the ease of use of the systems needs to be improved. They prefer the combination of the content-based and keyword-based search.

These preferences and behaviours match more with people with a fixed goal than people with an evolving goal. This may be because most of the people in our study have fixed goals and only a few people have mixedFE goals related to the evolving information goal.

8.3.13 No-FE-style

People with no-FE-style prefer three interfaces as follows: the interface with relevance feedback only, the interface with relevance feedback and query history, and the uInteract interface. These people prefer the flat profile of the Ostensive model and the combination of the three factors of the four-factor user interaction model: relevance region, time and frequency.

People with no-FE-style like fast, accurate and easy to use systems with rich functionalities. They are satisfied with the search results and the evaluation systems. Their initial idea changes during the search and they find it difficult to select the feedback and results for the tasks. They find all the functions useful and especially the negative functions, although they think the negative results should be optional. They feel the functions are most useful when they perform hard tasks. They think combining the keyword-based search with the content-based search would significantly enhance the system.

These people do not have clear views on system preferences, but their search behaviours and experiences show their search is rather exploratory, and the system supports the exploratory search well.

Overall, we have learnt how to apply the ISE model to analyze the quantitative and qualitative data from the user study based on our interactive CBIR systems.

Chapter 9

Conclusions and Future Work

The overall aim of our research was to systematically explore the three key elements of user interaction for content-based image retrieval (CBIR): the interaction model, interactive interface and users, in an integrated and principled manner. The objectives were (1) to develop a framework for interactive CBIR including a user interaction model and a visual interactive interface to deliver the model; (2) to evaluate the usefulness and effectiveness of the framework by user-oriented evaluations; (3) to demonstrate how interactive CBIR search can be analyzed in the context of tasks and user characters.

9.1 Research Contributions

In an effort to better understand and improve the interaction between users and CBIR systems, we have proposed a novel exploratory CBIR framework called uInteract. The framework contains a four-factor user interaction model and an interactive interface. The four-factor user interaction model was developed to overcome the limitations of related work. The four-factor user interaction model is delivered by an interactive interface visually to users. A lab-based simulated experiment was employed to test the effectiveness of the model, and a task-based user study was employed to evaluate the usefulness and effectiveness of the model and visual interface. From the quantitative data analysis results, we have not only demonstrated the usefulness and effectiveness of the framework, but have also observed that users have very different opinions on the usefulness and effectiveness of the different components and functionalities of the framework. In an effort to find different preferences and search behaviours based on different user types, we proposed a user classification model, called ISE (information goals, search strategies, evaluation thresholds), based on Information Foraging Theory. The ISE model was verified by an in-depth analysis of the real user interaction data collected from our user study. The verified ISE model was applied back to the quantitative and qualitative data in our user study and we have shown different user types have very different preferences and search behaviours. Therefore, we suggest that when designing, developing and evaluating a new search tool to make an effective user interaction happen, we need to consider the user interaction model, interactive interface and user types as a whole.

By doing this research, we have made the following contributions:

- proposed a four-factor user interaction model for interactive CBIR systems (Chapter 2);
- designed interactive uInteract interface to deliver the four-factor user interaction model (Chapter 2);
- evaluated the effects of the uInteract interface (Chapter 4);
- evaluated the effects of the four profiles of the Ostensive Model with multiple image query and both positive and negative feedback (Chapter 5);
- evaluated the effects of the different settings of the four-factor user interaction model (Chapter 6);
- proposed a principled ISE user classification model based on the Information Foraging Theory (Chapter 7);
- Findings for future interactive CBIR framework design from applying the ISE model (Chapter 8).

9.1.1 Four-factor User Interaction Model

In an effort to improve the interaction between the users and the CBIR system as well as search accuracy, we have proposed and investigated a four-factor user interaction model (FFUIM), which includes relevance region, relevance level, time and frequency. Development and testing of the FFUIM was motivated and inspired by a growing interest in making the search system more interactive, and by the ongoing research on user interaction models to support interactive search. Notably, with the recent research interest in exploratory search, supporting users' interaction and communication with the system becomes increasingly more important. Whilst the model was developed for our research purposes, we believe it could be adapted to any interactive search system. Therefore, the model is not only useful in this thesis, but also contributes to the further development of more advanced user interaction models.

9.1.2 uInteract Interface

The uInteract interface was developed mainly for delivering the FFUIM visually to users. However, we have gone beyond the existing design guidelines and considered how to support users' exploratory search and how to let users manipulate the search mechanism naturally. The uInteract interface combines interactive search features in a novel way. The key achievements are: (1) we allow users to provide negative examples from specific search results to refine the query; (2) we provide the negative as well as positive search results to users to support their understanding of the data quality; at the same time, the negative results play a crucial part in helping users manipulate the search algorithm; (3) we also provide the query history function that supports users exploratory search.

9.1.3 Evaluation

We have not only applied simulated precision-based experiments to investigate the effectiveness of the four-factor user interaction model, but also, motivated by the related literature on evaluating interactive search systems, conducted extensive user-oriented evaluations. The key contributions of our user study are: (1) we have evaluated the uInteract interface that delivers the FFUIM against the basic relevance feedback interface, the interface based on the Ostensive Model, and the interface based on the partial model. By testing the effect of the uInteract interface, we have demonstrated that different types of users have different preferences. (2) we have investigated the four profiles of the Ostensive Model for multi-image queries with both positive and negative feedback CBIR scenarios, which previously has not been extensively tested in content-based image retrieval. In general, our evaluation outcome on the effect of the four profile of the Ostensive Model is similar to the evaluation results of previous studies, i.e., users prefer the increasing and flat profiles to decreasing and current profiles; (3) we have evaluated the different combinations of the four factors of the FFUIM. Again the findings from the evaluation show different user types have different preferences.

The design of the evaluations follows the best practice in the literature. At the same time, the outcome of the evaluations contributes to the literature: firstly, by demonstrating the usefulness of the four-factor user interaction model and proving the usefulness of the existing user interaction models in CBIR; secondly, by finding the strong impact on the search results, preferences and behaviours from different user types.

9.1.4 ISE User Classification Model

In an effort to define and identify different user types, we have proposed an ISE user classification model based on Information Foraging Theory. The ISE model includes three criteria corresponding to three key aspects during the search process: users' information goal (I), users' search strategy (S) and users' evaluation threshold (E). The three aspects are generated based on the three models of Information Foraging Theory, namely: the information scent model, the information patch model and the information diet model. The Information Foraging Theory is a well known theory adapted from food foraging theory in biology to explain and predict the human information seeking behaviours. The information scent and information patch models have been applied to information seeking and browsing, and Information Foraging Theory has been suggested to analyze searchers' preferences and behaviours for exploratory search. However, the theory has not previously been applied to real search scenarios, especially to content-based image retrieval scenarios.

The proposed ISE model was verified by a multiple linear regression analysis of the qualitative user interaction data gathered from our real user evaluations. To the best of our knowledge, this model is the first principled user classification model in CBIR verified by a systematic data analysis based on large real interaction data. This practise has not only helped to identify different user types for future user-focused design, study and analysis, but also reinforces the usefulness of the Information Foraging Theory for information seeking, especially for interactive CBIR searches.

9.1.5 Findings and Suggestions with Regard to User Types

In order to find the users' preferences concerning the user interaction models and the interactive interfaces in our study, and further concerning CBIR search, we applied the ISE user classification model to the analysis of the quantitative and qualitative data we obtained from the user study. The analysis was done based on the five characters that occurred in two classification criteria. For instance, under the search strategy criterion, we have risky, cautious, mixedRC, noneRC and undefinedRC characters, under the information goal criterion, we have fixed, mixedFE, noneFE and undefinedFE characters. The quantitative and qualitative data analysis results based on the five characters have provided valuable observations on the preferences of different types of users. The observations match closely the characters' definitions in the ISE model. We further merged the five characters under each criterion into two groups: with-style or no-style. For example, for search strategy, we grouped risky, cautious and mixedRC together into with-RC-style, and noneRC and undefinedRC together into no-RC-style. Similarly for the information goal, we grouped fixed and mixedFE together into with-FE-style, and noneFE and undefinedFE together into no-FE-style. The analysis based on the two groups shows that the search process of the users with no style is rather exploratory and the systems, especially the uInteract interface, supported the search well. Further, these users like three of the four factors of the four-factor user interaction model: relevance region, time and frequency, and tend not to care much about the relevance level factor. However, the users with-FE-style like the relevance level factor, which may be because they have a fixed search goal so that they can make fully use of this function in the interface.

The suggestion we have made from the findings of the quantitative and qualitative data analysis based on the ISE user classification model is that the ISE model produces a usable way to identify user types precisely in ways that enable us to predict with some degree of accuracy, what different types of users want - at least in this study. The findings on the preferences and search behaviours based on the different user types provide valuable guidelines on designing, evaluating and analyzing information seeking systems, especially interactive content-based image retrieval systems.

9.2 Future Work

Whilst we have provided useful observations for future image search tools, especially interactive CBIR search system development to suit different user types, we have also discovered areas where improvement can be made through future work.

Our research has demonstrated that the four-factor user interaction model is useful and effective for CBIR search, and it could also be adapted to other interactive search systems. The model can still be expanded by adding fresh factors, such as users' social recommendations, users' preferences, etc. The uInteract interface has successfully delivered the four-factor user interaction model, enabling users to manipulate the model visually. However, users have suggested the need to improve the ease of use of the interface, e.g., by introducing drag and drop, enlarging the image, visualizing the query history, enabling to provide negative feedback by a colour pie chart rather than using negative search results, improving the graphic design of the buttons and interface layout, etc.

The uInteract framework including the four-factor user interaction model and the interactive interface has demonstrated that it is possible for users to effectively manipulate the relevance feedback mechanism through a visual interface without impacting on how the underlying search mechanism works. As Lew et al. (2006) suggested, a CBIR system should support searching by various media including text and content information. Other researchers have also suggested supporting varying user contexts, the search system needs to include different search models and should also be able to integrate these models with flexible user interfaces (Marques and Furht 2002). However, the integration of the different search models and the integration between relevance feedback mechanism and user interaction technologies is still a challenge in CBIR (Barecke et al. 2006). Our quantitative data analysis results based on users' comments also suggest some related challenges. For instance, users would like to see advanced visualization and interaction options by applying more user interface design techniques; they also expect a retrieval tool that integrates all user interactions in a single interface by making some functions optional. Therefore, developing a system with different interfaces to customize different search skill levels could improve the search activity. Further, users would like to see more diverse results rather than just colour-based results. This again emphasizes the importance of combining multi-modality into a single search system, so that the results can be provided based on diverse search results using different search models. This is also a key component of exploratory search.

We have evaluated the framework by a series of simulated and user-oriented evaluations. The evaluation provides valuable insights on future interactive CBIR system design, evaluation and analysis. However, there are several lessons learnt from the evaluations: (1) The task affected the performance indicators much more than we had expected. Thus we should always consider how the nature of the tasks will affect the study before carrying out evaluations. As such the effects of the different complexity levels of tasks will be ruled out; (2) We did not consider the user types when we recruited the users so that the number of users for each user type occurred in our user study was not evenly distributed. As a consequence, we could not make suggestions on some types of users due to the lack of data. To improve this situation, we should revise our entry questionnaire to ensure more user types are covered in our future evaluations, or even carry out a user type test before performing the formal user study; (3) The evaluations we performed were controlled, short-term user studies. The users only used the system once, and were restricted to 5 and 10 minutes to complete each task. The findings can be limited based on the data obtained from this kind of study. A long-term and less controlled evaluation may provide richer observations and insights on user types, search preferences and behaviours.

We have successfully applied Information Foraging Theory to understand user interaction based on the users perspective, and proposed the first principled user classification model - ISE (information goal (I), search strategy (S) and evaluation threshold (E)). The ISE model has been verified and tested based on the user interaction data of our user study. We suggest the ISE model needs to be further verified in an independent study on a broader range of interactive search systems.

Appendix A

Tasks and questionnaires of evaluation1 (E1)

A.1 Four tasks of E1

A.2 Questionnaires of E1

E1: TASK 1

Your task is to find **the best positive query** from the “positive query panel” and **the best negative query** from the “negative query panel” (if applicable) that you used to retrieve images for the topic of “**people observing football match**” (the query can contain one or more images). Please use the 3 images below as your initial positive query (you can find the 3 example images from the “query image panel”), you may then make any query that you wish.

Please record your results in the “**Answers**” box below.

TASK 1: ANSWERS

Positive query:

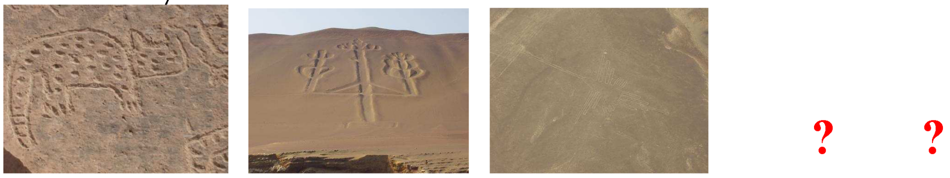
Negative query (if applicable):

E1: Task 2

Consider the three images provided below, they all share a common theme. Please use the 3 images that you can find from the "query image panel" as your initial positive query, you may then make any query that you wish. Your task is two fold:

1. Find **one** similar image (from the result) that complements the set, and
2. Find **one** image (from the result) that stands out as having a completely different-theme.

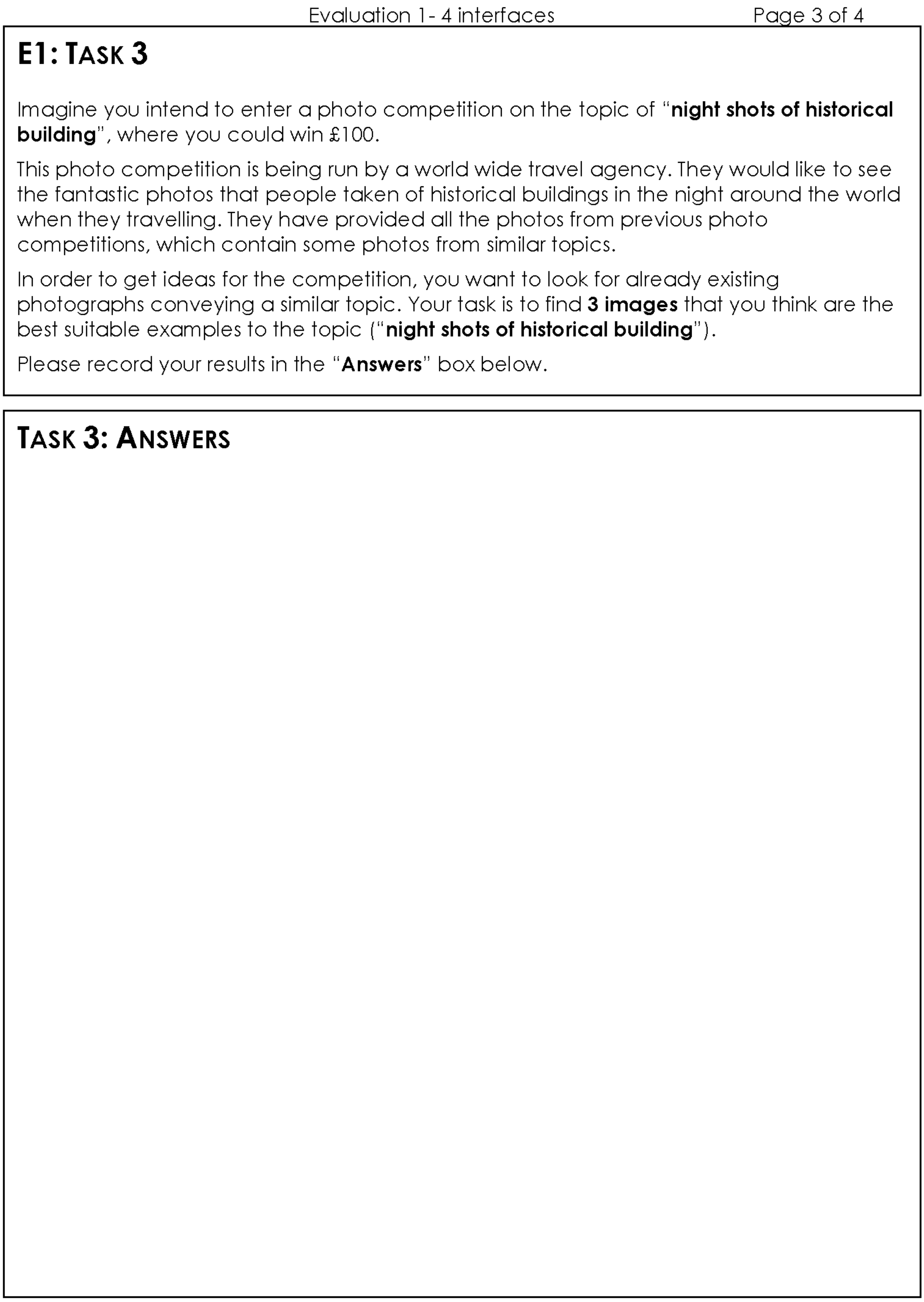
Please record your results in the "**Answers**" box below.



Task 2: Answers

One similar-theme image:

One different-theme image:



Evaluation 1- 4 interfaces Page 3 of 4

E1: Task 3

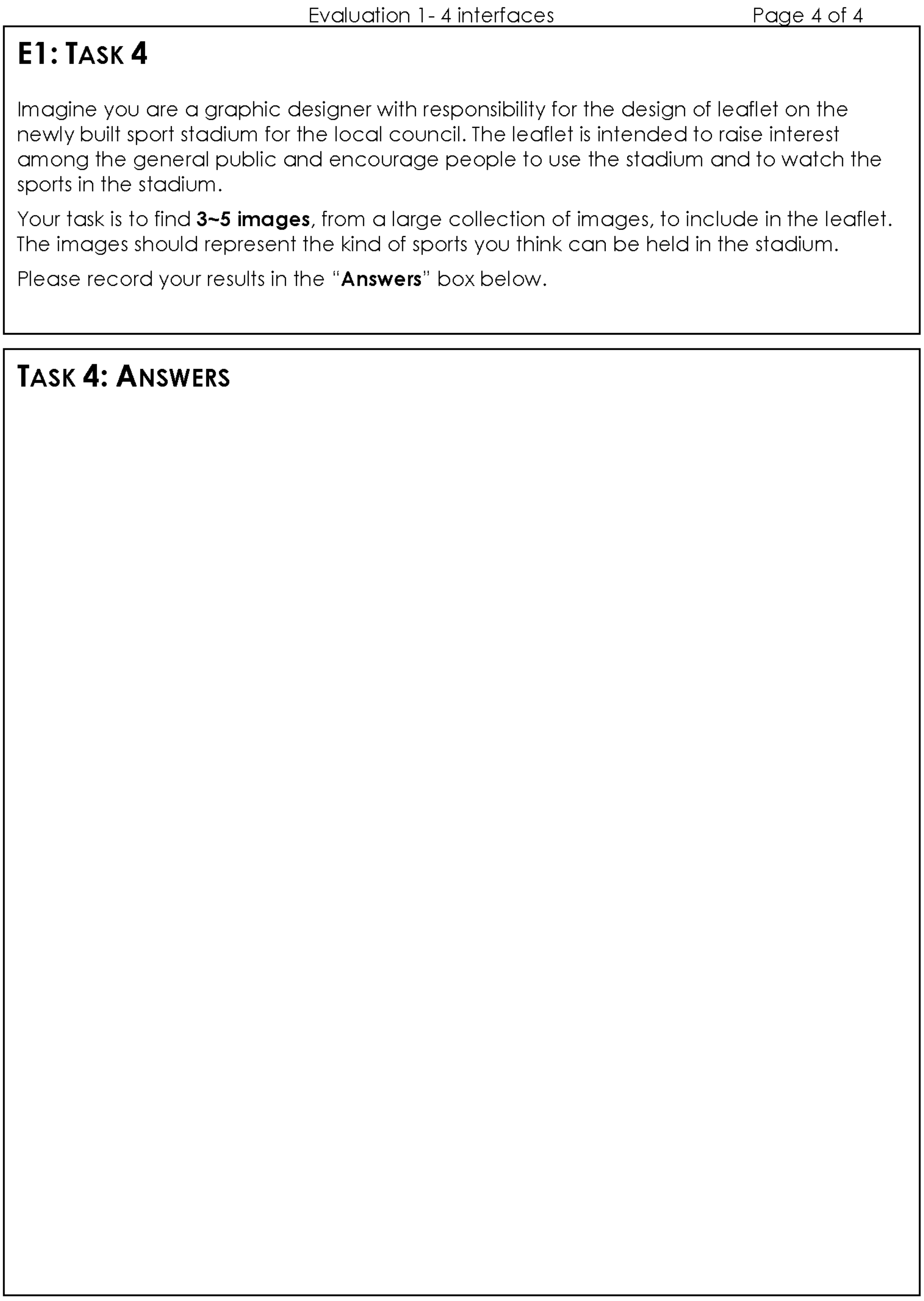
Imagine you intend to enter a photo competition on the topic of “**night shots of historical building**”, where you could win £100.

This photo competition is being run by a world wide travel agency. They would like to see the fantastic photos that people taken of historical buildings in the night around the world when they travelling. They have provided all the photos from previous photo competitions, which contain some photos from similar topics.

In order to get ideas for the competition, you want to look for already existing photographs conveying a similar topic. Your task is to find **3 images** that you think are the best suitable examples to the topic (“**night shots of historical building**”).

Please record your results in the “**Answers**” box below.

Task 3: Answers



Evaluation 1- 4 interfaces Page 4 of 4

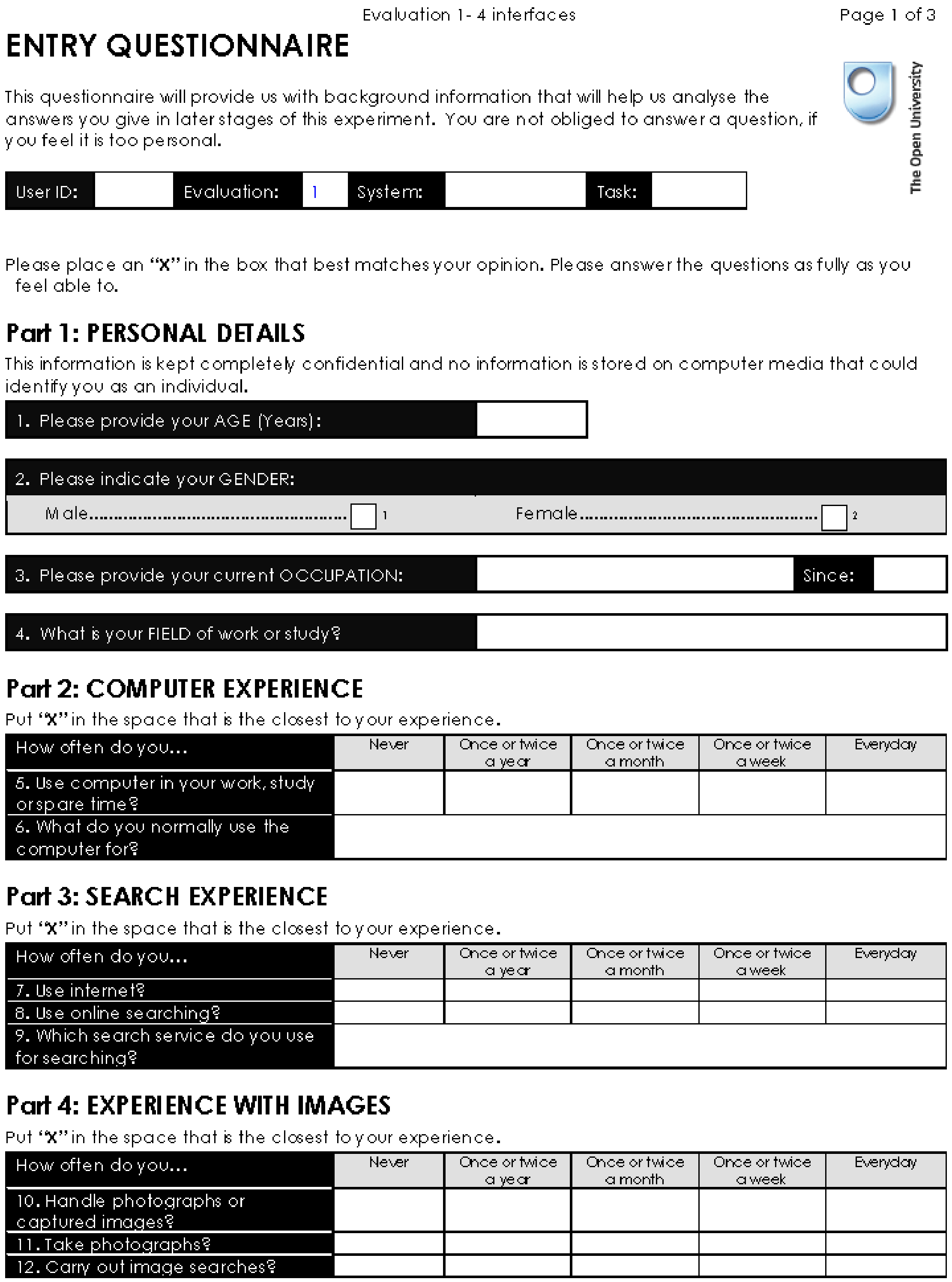
E1: TASK 4

Imagine you are a graphic designer with responsibility for the design of leaflet on the newly built sport stadium for the local council. The leaflet is intended to raise interest among the general public and encourage people to use the stadium and to watch the sports in the stadium.

Your task is to find **3~5 images**, from a large collection of images, to include in the leaflet. The images should represent the kind of sports you think can be held in the stadium.

Please record your results in the "**Answers**" box below.

TASK 4: ANSWERS



Evaluation 1- 4 interfaces

Page 1 of 3

ENTRY QUESTIONNAIRE

The Open University

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question, if you feel it is too personal.

User ID:		Evaluation:	1	System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer the questions as fully as you feel able to.

Part 1: PERSONAL DETAILS

This information is kept completely confidential and no information is stored on computer media that could identify you as an individual.

1. Please provide your AGE (Years):	

2. Please indicate your GENDER:

Male.. □ 1   Female.. □ 2

3. Please provide your current OCCUPATION:		Since:	

4. What is your FIELD of work or study?	

Part 2: COMPUTER EXPERIENCE

Put "X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
5. Use computer in your work, study or spare time?					
6. What do you normally use the computer for?					

Part 3: SEARCH EXPERIENCE

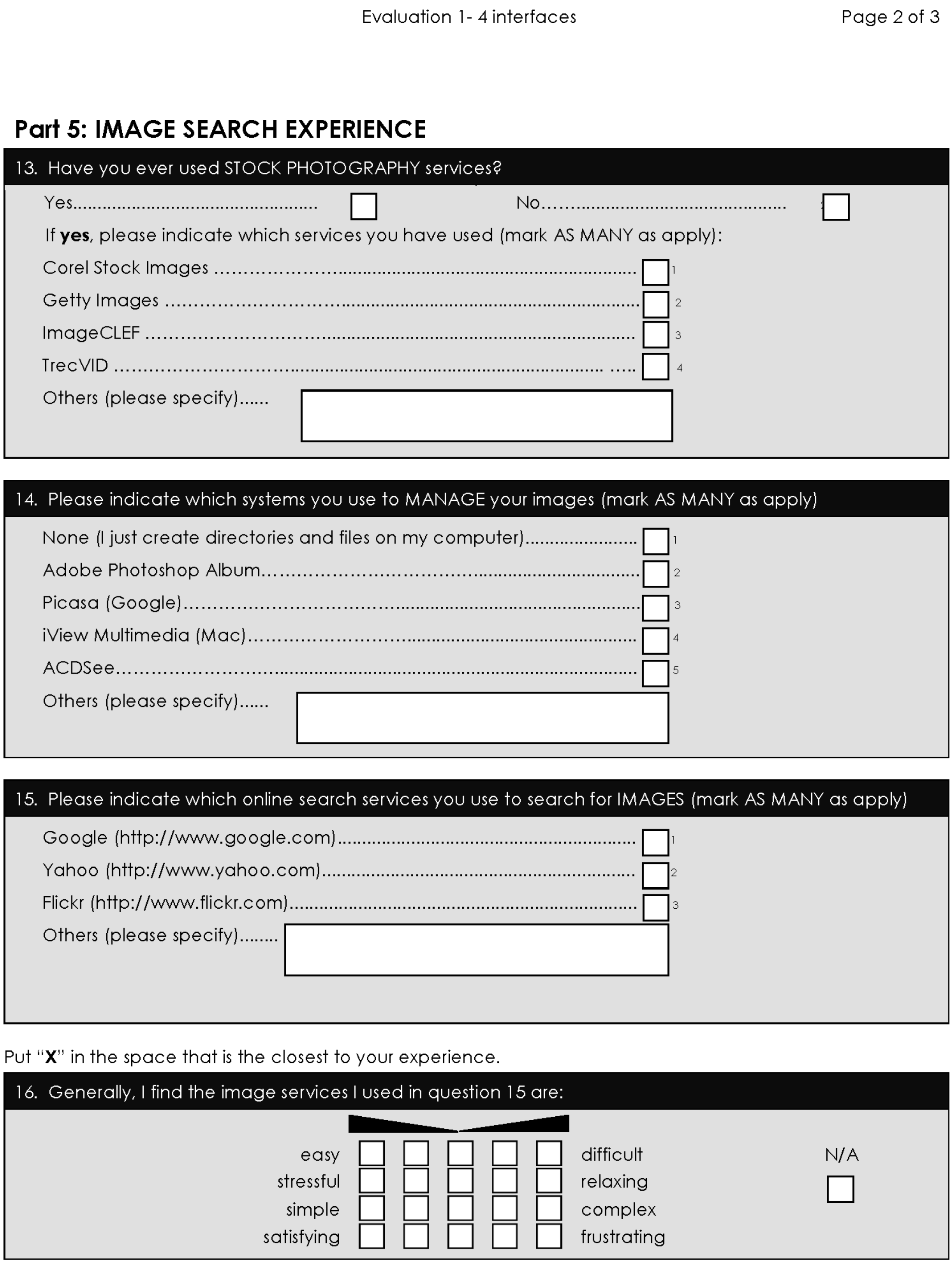
Put "X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
7. Use internet?					
8. Use online searching?					
9. Which search service do you use for searching?					

Part 4: EXPERIENCE WITH IMAGES

Put "X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
10. Handle photographs or captured images?					
11. Take photographs?					
12. Carry out image searches?					



Evaluation 1- 4 interfaces Page 2 of 3

Part 5: IMAGE SEARCH EXPERIENCE

13. Have you ever used STOCK PHOTOGRAPHY services?

Yes.. □ No.. □

If **yes**, please indicate which services you have used (mark AS MANY as apply):

Corel Stock Images .... □ 1
Getty Images .... □ 2
ImageCLEF .... □ 3
TrecVID .... □ 4
Others (please specify)...... □

14. Please indicate which systems you use to MANAGE your images (mark AS MANY as apply)

None (I just create directories and files on my computer)...................... □ 1
Adobe Photoshop Album... □ 2
Picasa (Google).. □ 3
iView Multimedia (Mac)... □ 4
ACDSee.. □ 5
Others (please specify)...... □

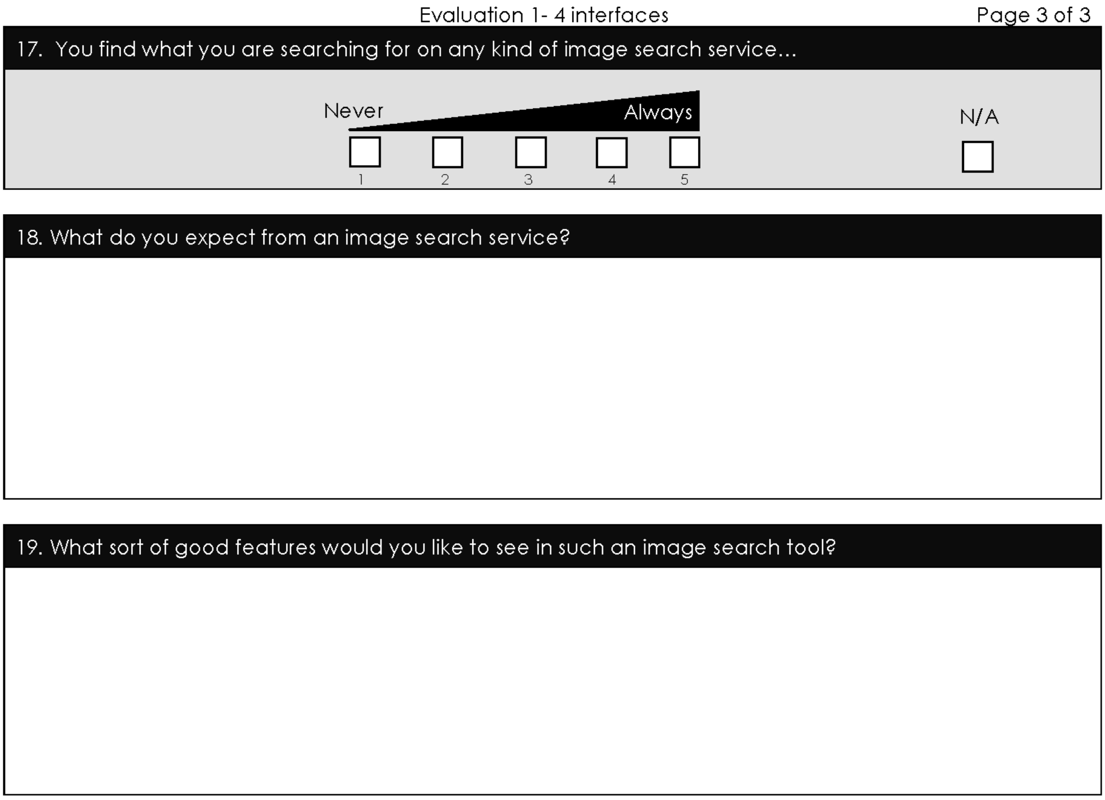
15. Please indicate which online search services you use to search for IMAGES (mark AS MANY as apply)

Google (http://www.google.com).. □ 1
Yahoo (http://www.yahoo.com)... □ 2
Flickr (http://www.flickr.com)... □ 3
Others (please specify)........ □

Put "**X**" in the space that is the closest to your experience.

16. Generally, I find the image services I used in question 15 are:

easy	□	□	□	□	□	difficult	N/A
stressful	□	□	□	□	□	relaxing	□
simple	□	□	□	□	□	complex	
satisfying	□	□	□	□	□	frustrating	



Evaluation 1- 4 interfaces

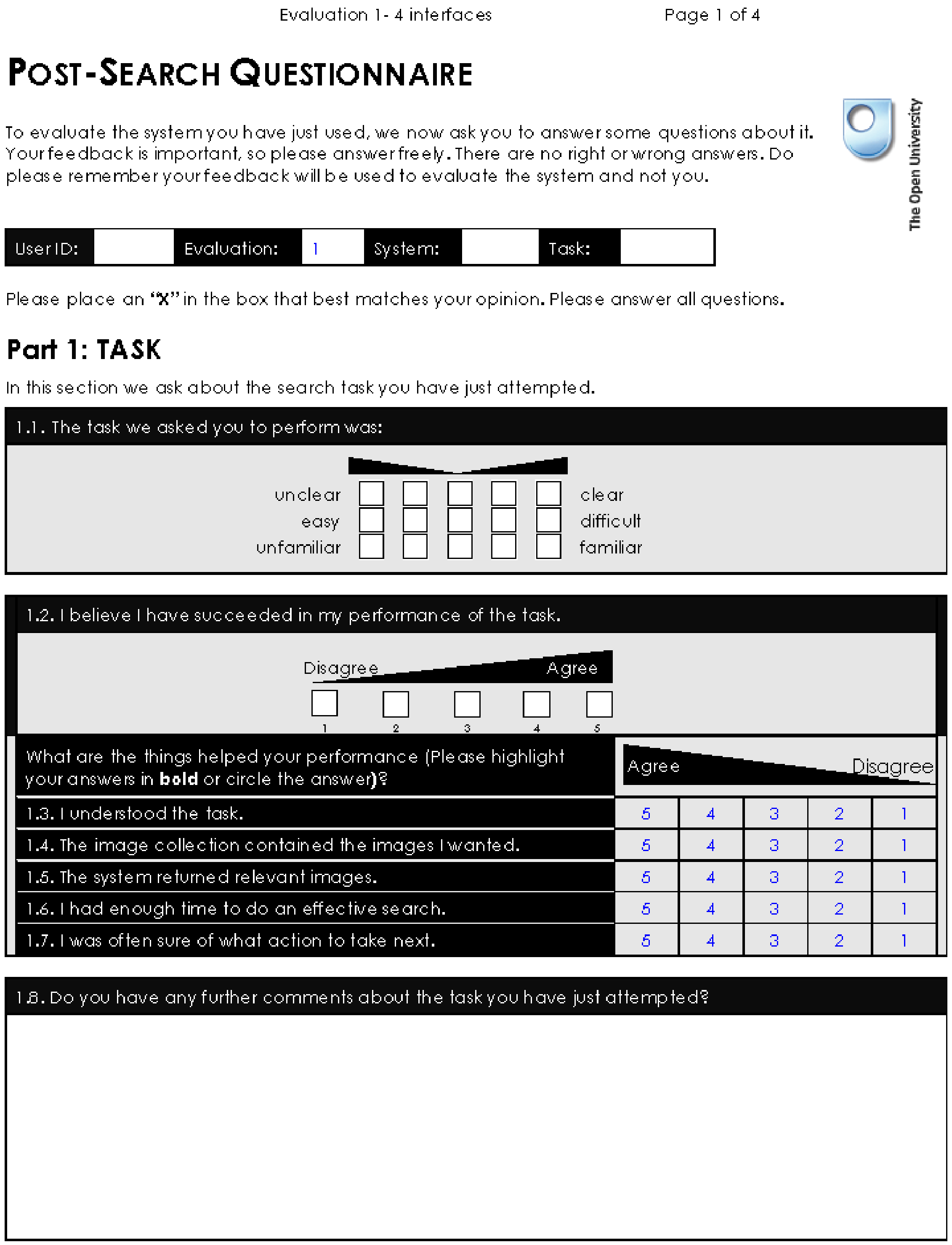
Page 3 of 3

17. You find what you are searching for on any kind of image search service...

Never				Always	N/A
□	□	□	□	□	□
1	2	3	4	5	

18. What do you expect from an image search service?

19. What sort of good features would you like to see in such an image search tool?



Evaluation 1- 4 interfaces Page 1 of 4

POST-SEARCH QUESTIONNAIRE

The Open University

To evaluate the system you have just used, we now ask you to answer some questions about it. Your feedback is important, so please answer freely. There are no right or wrong answers. Do please remember your feedback will be used to evaluate the system and not you.

User ID:		Evaluation:	1	System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer all questions.

Part 1: TASK

In this section we ask about the search task you have just attempted.

1.1. The task we asked you to perform was:

unclear	□	□	□	□	□	clear
easy	□	□	□	□	□	difficult
unfamiliar	□	□	□	□	□	familiar

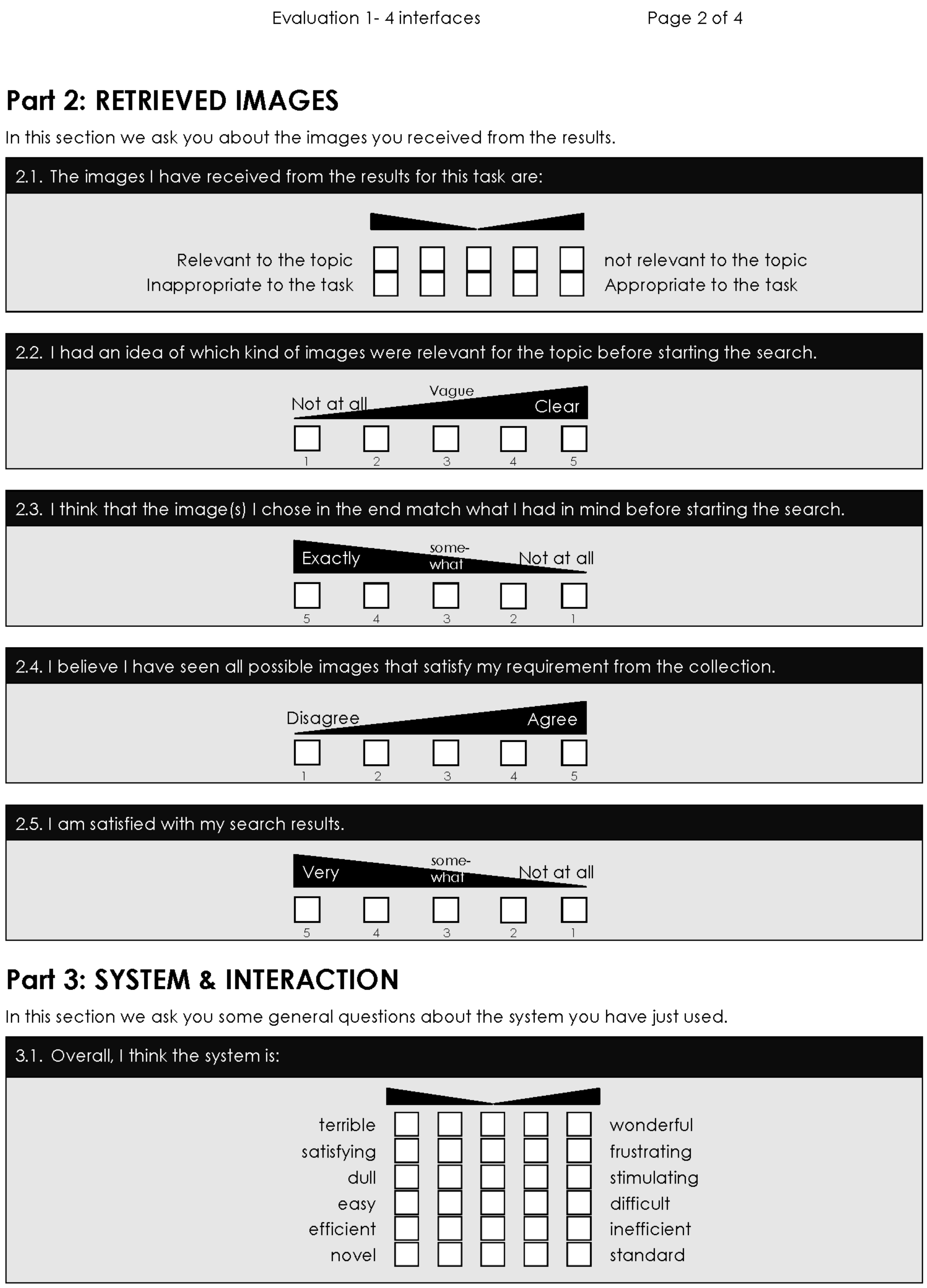
1.2. I believe I have succeeded in my performance of the task.

Disagree Agree

□	□	□	□	□
1	2	3	4	5

What are the things helped your performance (Please highlight your answers in **bold** or circle the answer)?	Agree				Disagree
1.3. I understood the task.	5	4	3	2	1
1.4. The image collection contained the images I wanted.	5	4	3	2	1
1.5. The system returned relevant images.	5	4	3	2	1
1.6. I had enough time to do an effective search.	5	4	3	2	1
1.7. I was often sure of what action to take next.	5	4	3	2	1

1.8. Do you have any further comments about the task you have just attempted?



Evaluation 1- 4 interfaces Page 2 of 4

Part 2: RETRIEVED IMAGES

In this section we ask you about the images you received from the results.

2.1. The images I have received from the results for this task are:

Relevant to the topic	□	□	□	□	□	not relevant to the topic
Inappropriate to the task	□	□	□	□	□	Appropriate to the task

2.2. I had an idea of which kind of images were relevant for the topic before starting the search.

Not at all Vague Clear

□ □ □ □ □

1 2 3 4 5

2.3. I think that the image(s) I chose in the end match what I had in mind before starting the search.

Exactly some-what Not at all

□ □ □ □ □

5 4 3 2 1

2.4. I believe I have seen all possible images that satisfy my requirement from the collection.

Disagree Agree

□ □ □ □ □

1 2 3 4 5

2.5. I am satisfied with my search results.

Very some-what Not at all

□ □ □ □ □

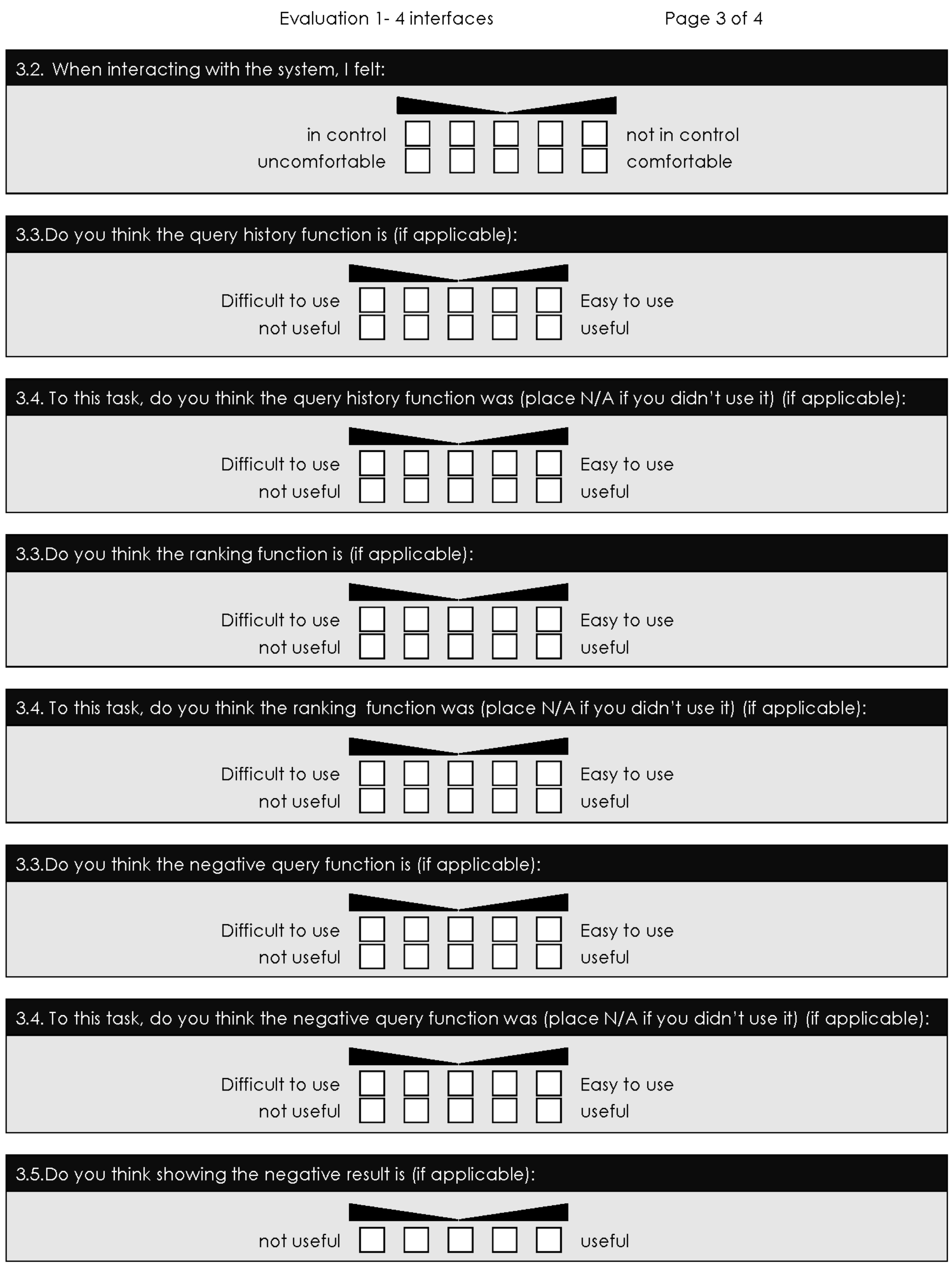
5 4 3 2 1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall, I think the system is:

terrible	□	□	□	□	□	wonderful
satisfying	□	□	□	□	□	frustrating
dull	□	□	□	□	□	stimulating
easy	□	□	□	□	□	difficult
efficient	□	□	□	□	□	inefficient
novel	□	□	□	□	□	standard



Evaluation 1- 4 interfaces Page 3 of 4

3.2. When interacting with the system, I felt:

in control □ □ □ □ □ not in control
uncomfortable □ □ □ □ □ comfortable

3.3.Do you think the query history function is (if applicable):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.4. To this task, do you think the query history function was (place N/A if you didn't use it) (if applicable):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.3.Do you think the ranking function is (if applicable):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.4. To this task, do you think the ranking function was (place N/A if you didn't use it) (if applicable):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.3.Do you think the negative query function is (if applicable):

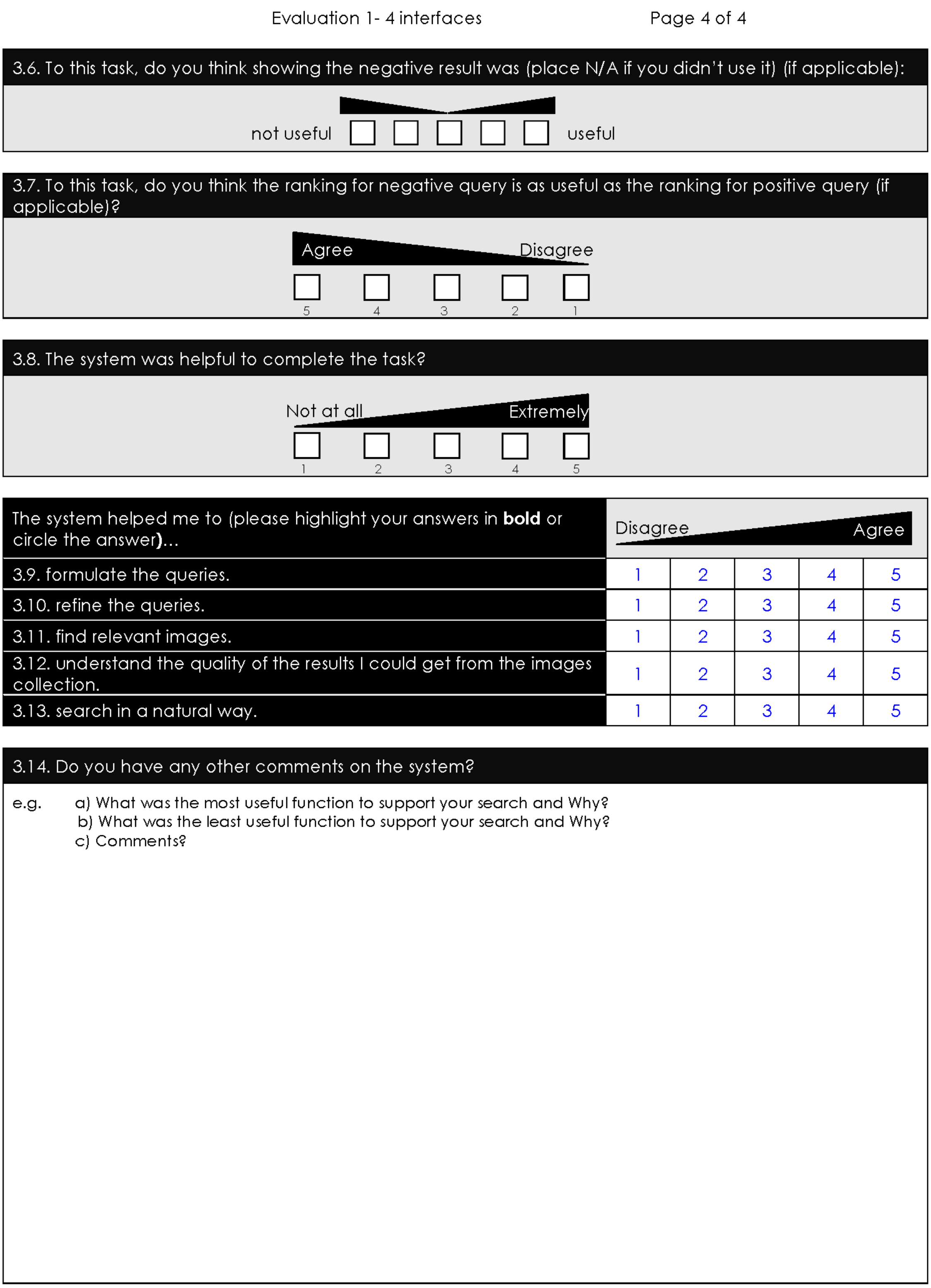
Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.4. To this task, do you think the negative query function was (place N/A if you didn't use it) (if applicable):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.5.Do you think showing the negative result is (if applicable):

not useful □ □ □ □ □ useful



Evaluation 1- 4 interfaces Page 4 of 4

3.6. To this task, do you think showing the negative result was (place N/A if you didn't use it) (if applicable):

not useful □ □ □ □ □ useful

3.7. To this task, do you think the ranking for negative query is as useful as the ranking for positive query (if applicable)?

Agree Disagree

□ □ □ □ □

5 4 3 2 1

3.8. The system was helpful to complete the task?

Not at all Extremely

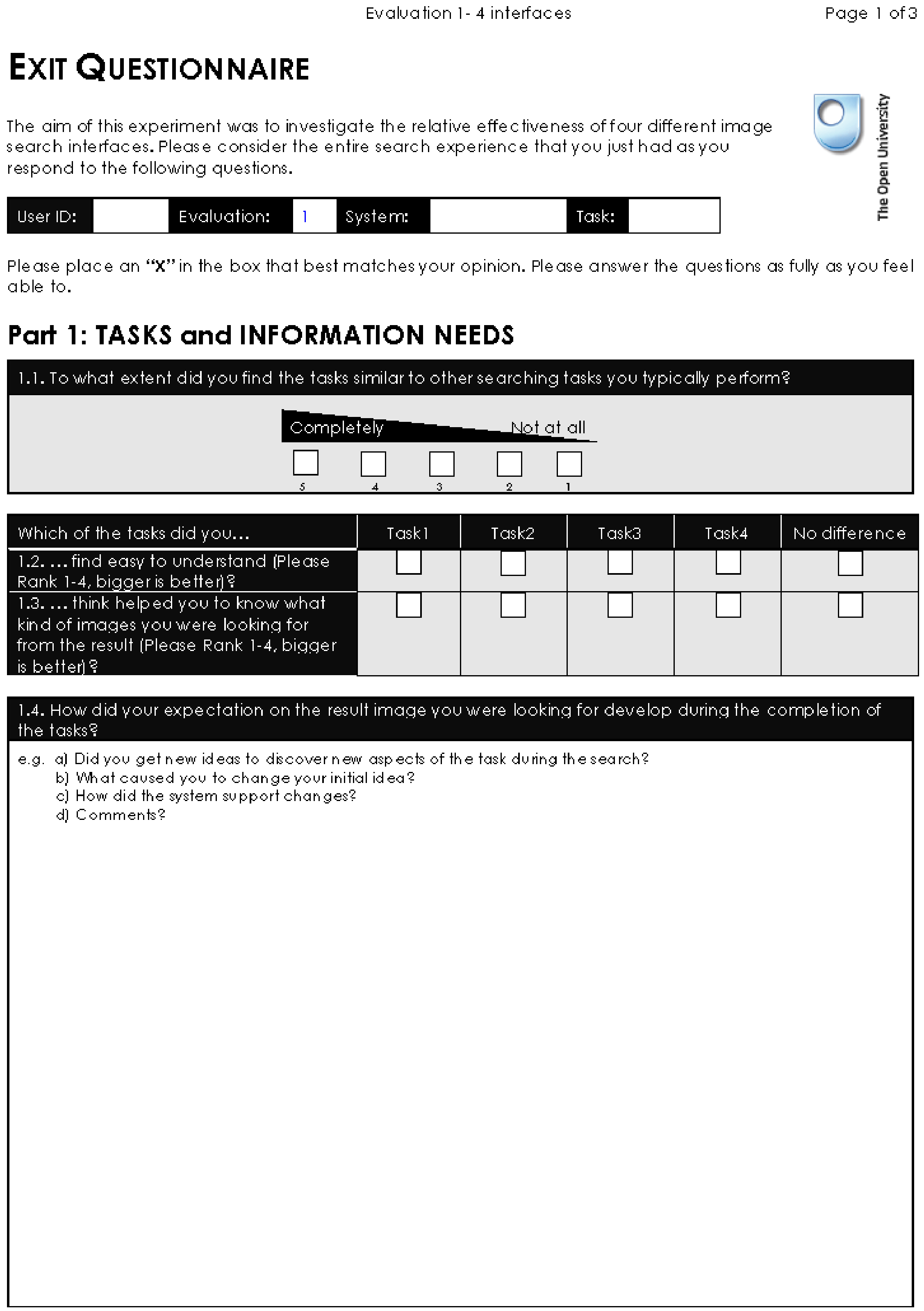
□ □ □ □ □

1 2 3 4 5

The system helped me to (please highlight your answers in **bold** or circle the answer**)**...	Disagree				Agree
3.9. formulate the queries.	1	2	3	4	5
3.10. refine the queries.	1	2	3	4	5
3.11. find relevant images.	1	2	3	4	5
3.12. understand the quality of the results I could get from the images collection.	1	2	3	4	5
3.13. search in a natural way.	1	2	3	4	5

3.14. Do you have any other comments on the system?

e.g. a) What was the most useful function to support your search and Why?
b) What was the least useful function to support your search and Why?
c) Comments?



Evaluation 1- 4 interfaces

Page 1 of 3

EXIT QUESTIONNAIRE

The aim of this experiment was to investigate the relative effectiveness of four different image search interfaces. Please consider the entire search experience that you just had as you respond to the following questions.

The Open University

User ID:		Evaluation:	1	System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer the questions as fully as you feel able to.

Part 1: TASKS and INFORMATION NEEDS

1.1. To what extent did you find the tasks similar to other searching tasks you typically perform?

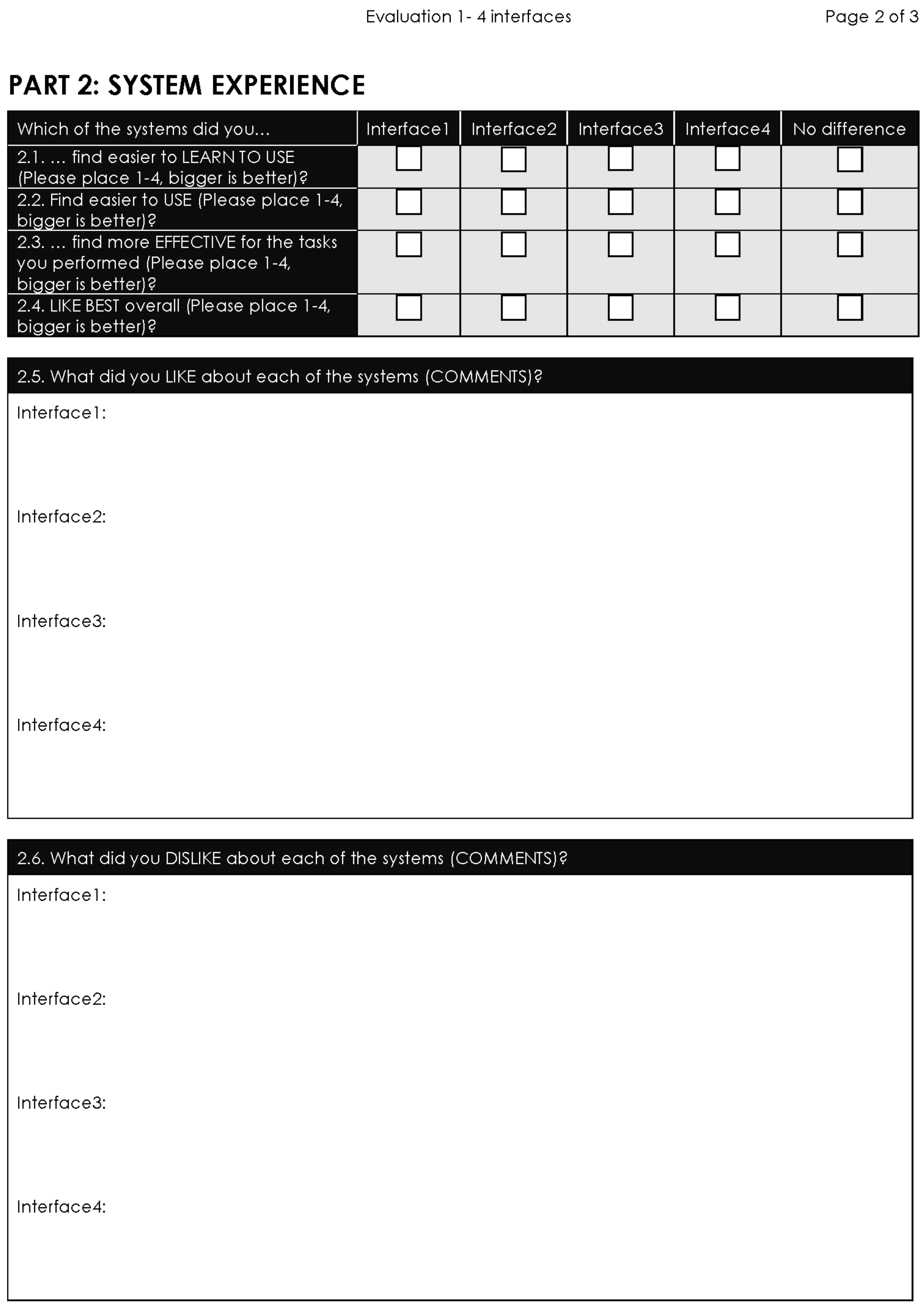
Completely — Not at all

□ 5 □ 4 □ 3 □ 2 □ 1

Which of the tasks did you...	Task1	Task2	Task3	Task4	No difference
1.2. ... find easy to understand (Please Rank 1-4, bigger is better)?	□	□	□	□	□
1.3. ... think helped you to know what kind of images you were looking for from the result (Please Rank 1-4, bigger is better)?	□	□	□	□	□

1.4. How did your expectation on the result image you were looking for develop during the completion of the tasks?

e.g. a) Did you get new ideas to discover new aspects of the task during the search?
b) What caused you to change your initial idea?
c) How did the system support changes?
d) Comments?



Evaluation 1- 4 interfaces Page 2 of 3

PART 2: SYSTEM EXPERIENCE

Which of the systems did you…	Interface1	Interface2	Interface3	Interface4	No difference
2.1. … find easier to LEARN TO USE (Please place 1-4, bigger is better)?	☐	☐	☐	☐	☐
2.2. Find easier to USE (Please place 1-4, bigger is better)?	☐	☐	☐	☐	☐
2.3. … find more EFFECTIVE for the tasks you performed (Please place 1-4, bigger is better)?	☐	☐	☐	☐	☐
2.4. LIKE BEST overall (Please place 1-4, bigger is better)?	☐	☐	☐	☐	☐

2.5. What did you LIKE about each of the systems (COMMENTS)?

Interface1:

Interface2:

Interface3:

Interface4:

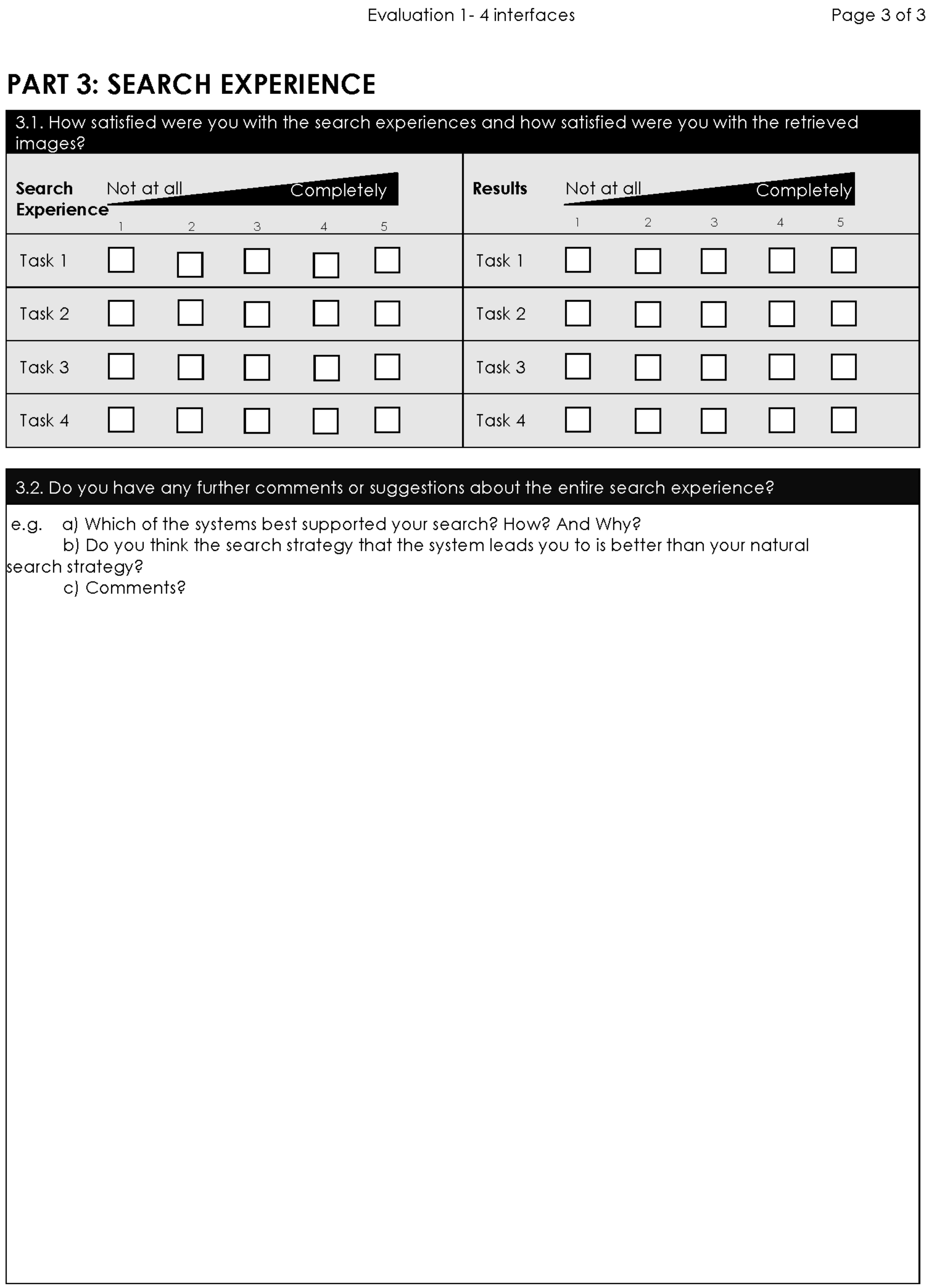
2.6. What did you DISLIKE about each of the systems (COMMENTS)?

Interface1:

Interface2:

Interface3:

Interface4:



Evaluation 1- 4 interfaces Page 3 of 3

PART 3: SEARCH EXPERIENCE

3.1. How satisfied were you with the search experiences and how satisfied were you with the retrieved images?

Search Experience	Not at all				Completely	Results	Not at all				Completely
	1	2	3	4	5		1	2	3	4	5
Task 1	□	□	□	□	□	Task 1	□	□	□	□	□
Task 2	□	□	□	□	□	Task 2	□	□	□	□	□
Task 3	□	□	□	□	□	Task 3	□	□	□	□	□
Task 4	□	□	□	□	□	Task 4	□	□	□	□	□

3.2. Do you have any further comments or suggestions about the entire search experience?

e.g. a) Which of the systems best supported your search? How? And Why?
b) Do you think the search strategy that the system leads you to is better than your natural search strategy?
c) Comments?

Appendix B

Tasks and questionnaires of evaluation2 (E2) and evaluation3 (E3)

B.1 Four tasks of E2 and E3

B.2 Questionnaires of E2 and E3

E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 1 of 4

E2/E3: Task 1

Your task is to find **as many images as possible** for the topic of "**drawings in deserts**" . Please use the 3 images below as your initial positive query (you can find the 3 example images from the "query image panel") and you can rank them however you like, you may then make any query that you wish.

Please record your results in the "**Answers**" box below.

Task 1: Answers

E2/E3: TASK 2

Your task is to find **as many images as possible** for the topic of "**scenes of footballers in action**" . Please use the 3 images below as your initial positive query (you can find the 3 example images from the "query image panel") and you can rank them however you like, you may then make any query that you wish.

Please record your results in the "**Answers**" box below.

TASK 2: ANSWERS

E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 3 of 4

E2/E3: Task 3

Your task is to find **as many images as possible** for the topic of "**sports people with prizes**". Please use the 3 images below as your initial positive query (you can find the 3 example images from the "query image panel") and you can rank them however you like, you may then make any query that you wish.

Please record your results in the "**Answers**" box below.

Task 3: Answers

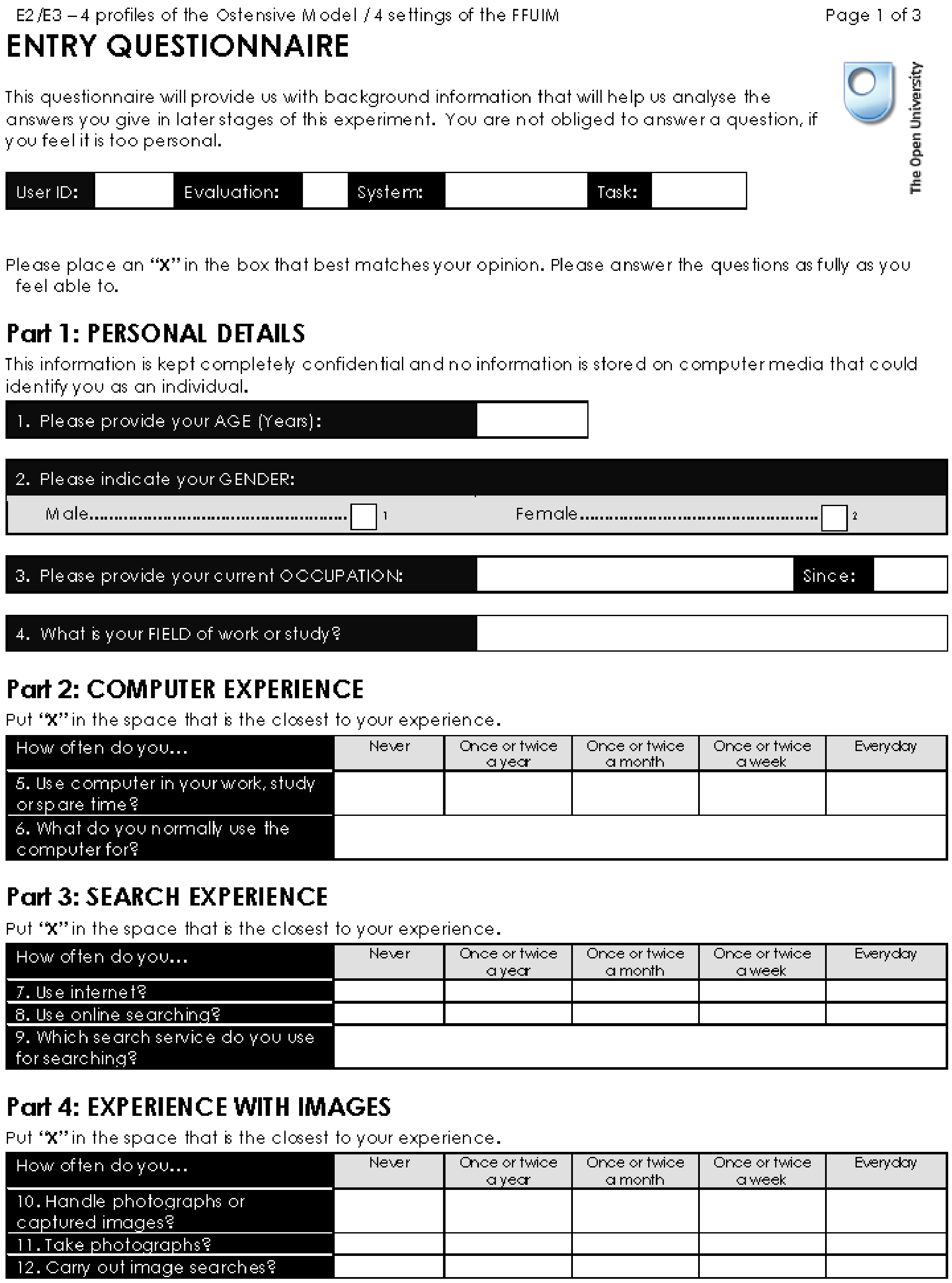
E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 4 of 4

E2/E3: TASK 4

Your task is to find **as many images as possible** for the topic of "**sunset over water**" . Please use the 3 images below as your initial positive query (you can find the 3 example images from the "query image panel") and you can rank them however you like, you may then make any query that you wish.

Please record your results in the "**Answers**" box below.

TASK 4: ANSWERS



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 1 of 3

ENTRY QUESTIONNAIRE

The Open University

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question, if you feel it is too personal.

User ID:		Evaluation:		System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer the questions as fully as you feel able to.

Part 1: PERSONAL DETAILS

This information is kept completely confidential and no information is stored on computer media that could identify you as an individual.

1. Please provide your AGE (Years):	

2. Please indicate your GENDER:

Male.. □ 1 Female.. □ 2

3. Please provide your current OCCUPATION:		Since:	

4. What is your FIELD of work or study?	

Part 2: COMPUTER EXPERIENCE

Put 'X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
5. Use computer in your work, study or spare time?					
6. What do you normally use the computer for?					

Part 3: SEARCH EXPERIENCE

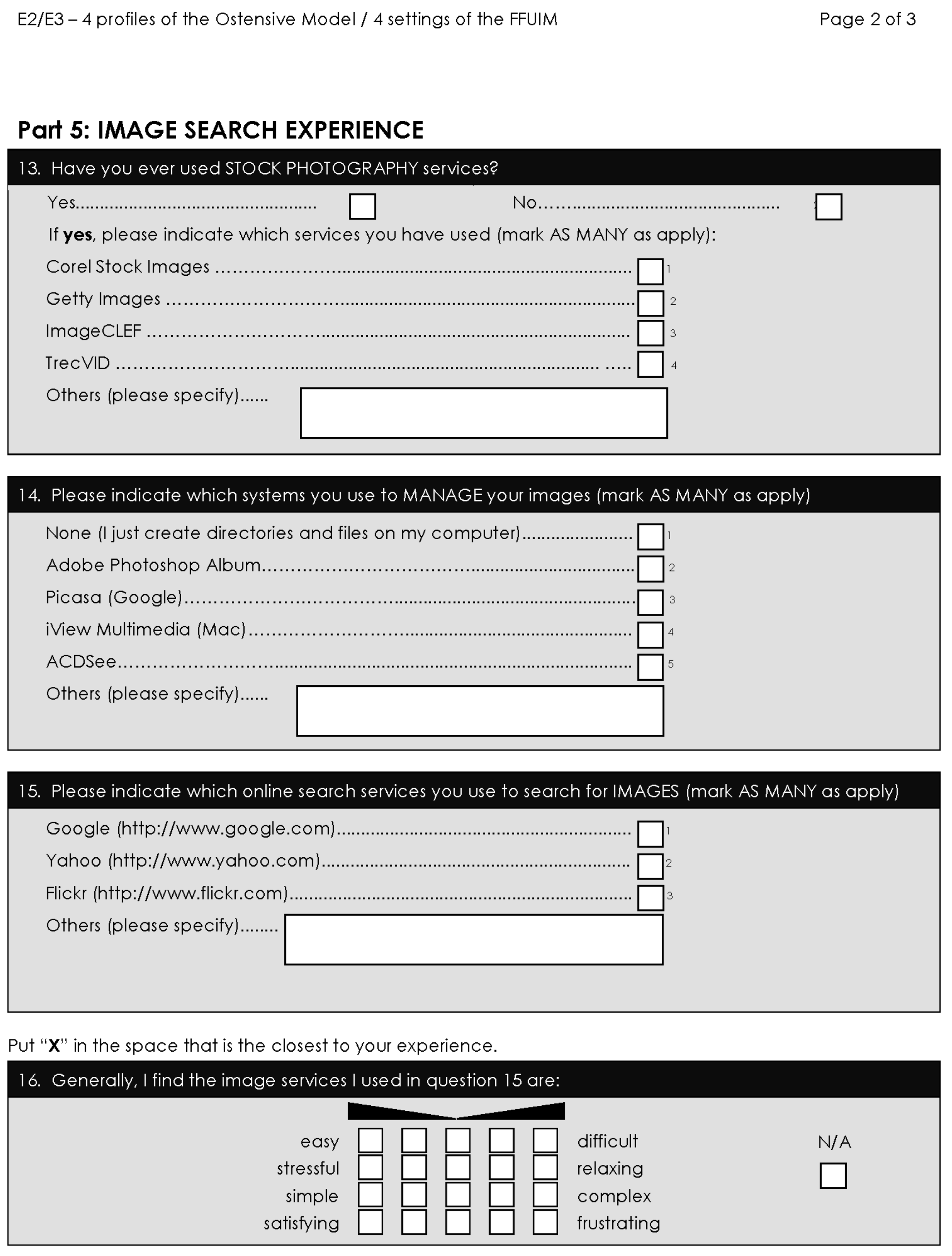
Put 'X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
7. Use internet?					
8. Use online searching?					
9. Which search service do you use for searching?					

Part 4: EXPERIENCE WITH IMAGES

Put 'X" in the space that is the closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Everyday
10. Handle photographs or captured images?					
11. Take photographs?					
12. Carry out image searches?					



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 2 of 3

Part 5: IMAGE SEARCH EXPERIENCE

13. Have you ever used STOCK PHOTOGRAPHY services?

Yes.. □ No.. □

If **yes**, please indicate which services you have used (mark AS MANY as apply):

Corel Stock Images .... □ 1

Getty Images .... □ 2

ImageCLEF .... □ 3

TrecVID .... □ 4

Others (please specify)...... □

14. Please indicate which systems you use to MANAGE your images (mark AS MANY as apply)

None (I just create directories and files on my computer)...................... □ 1

Adobe Photoshop Album.. □ 2

Picasa (Google)... □ 3

iView Multimedia (Mac)... □ 4

ACDSee... □ 5

Others (please specify)...... □

15. Please indicate which online search services you use to search for IMAGES (mark AS MANY as apply)

Google (http://www.google.com)... □ 1

Yahoo (http://www.yahoo.com).. □ 2

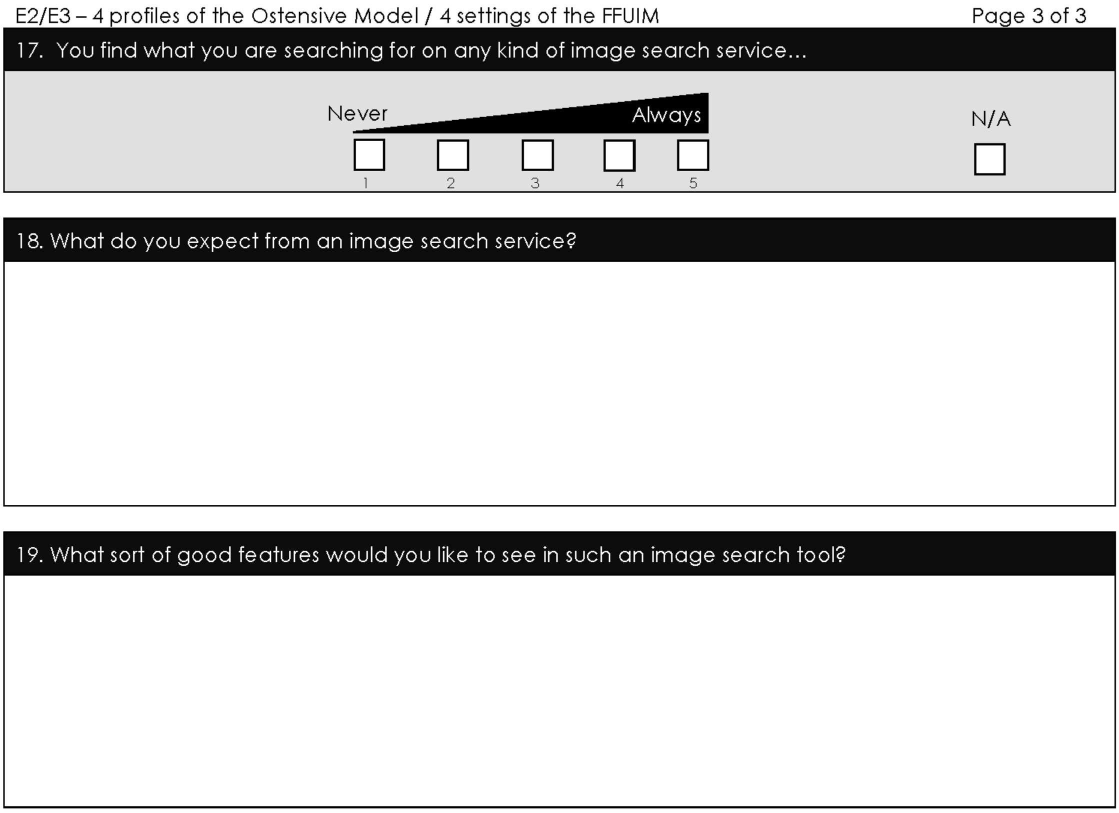
Flickr (http://www.flickr.com)... □ 3

Others (please specify)........ □

Put "**X**" in the space that is the closest to your experience.

16. Generally, I find the image services I used in question 15 are:

easy	□	□	□	□	□	difficult	N/A
stressful	□	□	□	□	□	relaxing	□
simple	□	□	□	□	□	complex	
satisfying	□	□	□	□	□	frustrating	



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 3 of 3

17. You find what you are searching for on any kind of image search service…

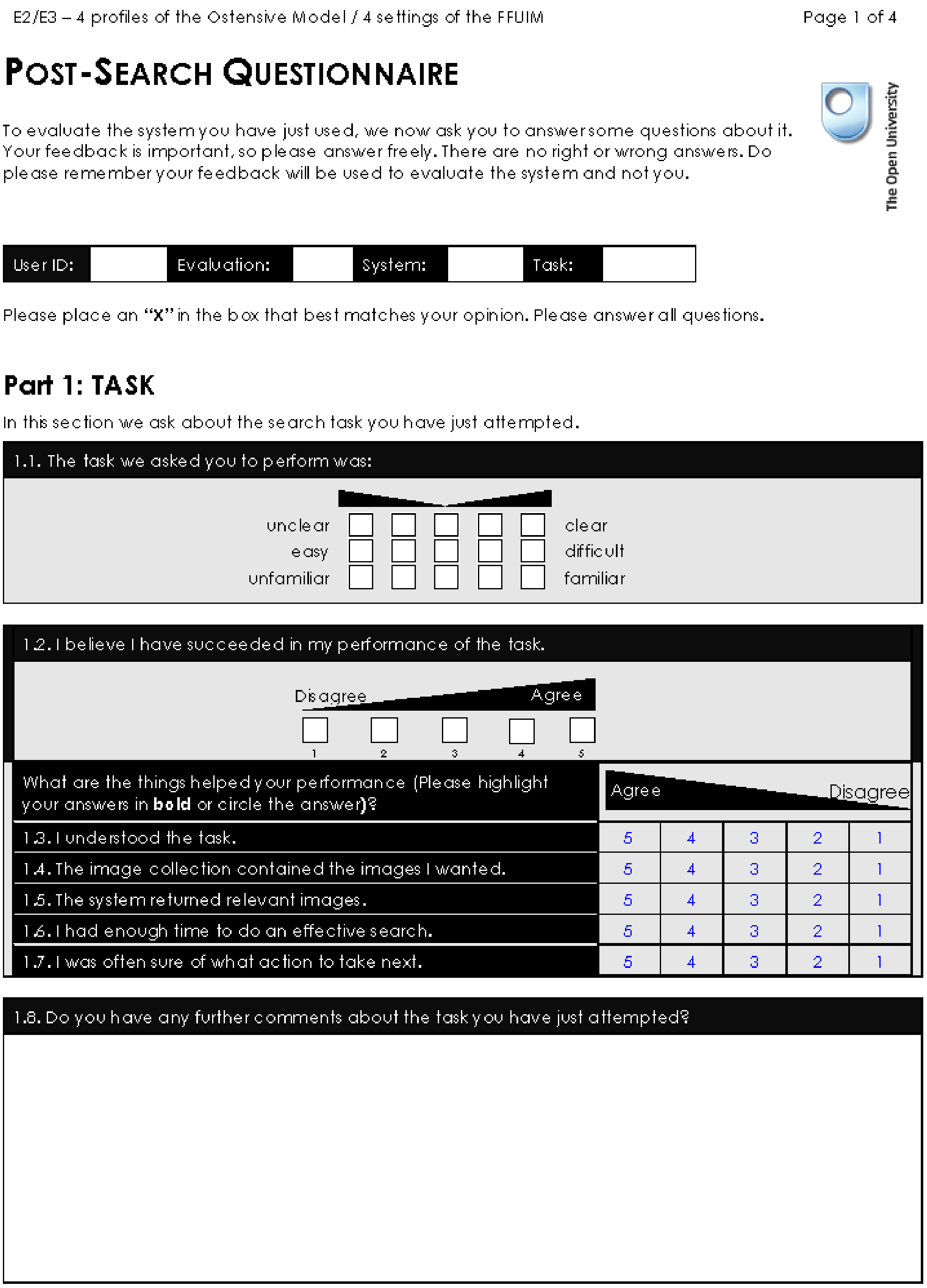
Never — Always

□ 1 □ 2 □ 3 □ 4 □ 5

N/A □

18. What do you expect from an image search service?

19. What sort of good features would you like to see in such an image search tool?



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 1 of 4

The Open University

POST-SEARCH QUESTIONNAIRE

To evaluate the system you have just used, we now ask you to answer some questions about it. Your feedback is important, so please answer freely. There are no right or wrong answers. Do please remember your feedback will be used to evaluate the system and not you.

User ID:		Evaluation:		System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer all questions.

Part 1: TASK

In this section we ask about the search task you have just attempted.

1.1. The task we asked you to perform was:

unclear	□	□	□	□	□	clear
easy	□	□	□	□	□	difficult
unfamiliar	□	□	□	□	□	familiar

1.2. I believe I have succeeded in my performance of the task.

Disagree Agree

□	□	□	□	□
1	2	3	4	5

What are the things helped your performance (Please highlight your answers in **bold** or circle the answer)?	Agree				Disagree
1.3. I understood the task.	5	4	3	2	1
1.4. The image collection contained the images I wanted.	5	4	3	2	1
1.5. The system returned relevant images.	5	4	3	2	1
1.6. I had enough time to do an effective search.	5	4	3	2	1
1.7. I was often sure of what action to take next.	5	4	3	2	1

1.8. Do you have any further comments about the task you have just attempted?

E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 2 of 4

Part 2: RETRIEVED IMAGES

In this section we ask you about the images you received from the results.

2.1. The images I have received from the results for this task are:

Relevant to the topic	□	□	□	□	□	not relevant to the topic
Inappropriate to the task	□	□	□	□	□	Appropriate to the task

2.2. I had an idea of which kind of images were relevant for the topic before starting the search.

Not at all		Vague		Clear
□	□	□	□	□
1	2	3	4	5

2.3. I think that the image(s) I chose in the end match what I had in mind before starting the search.

Exactly		some-what		Not at all
□	□	□	□	□
5	4	3	2	1

2.4. I believe I have seen all possible images that satisfy my requirement from the collection.

Disagree				Agree
□	□	□	□	□
1	2	3	4	5

2.5. I am satisfied with my search results.

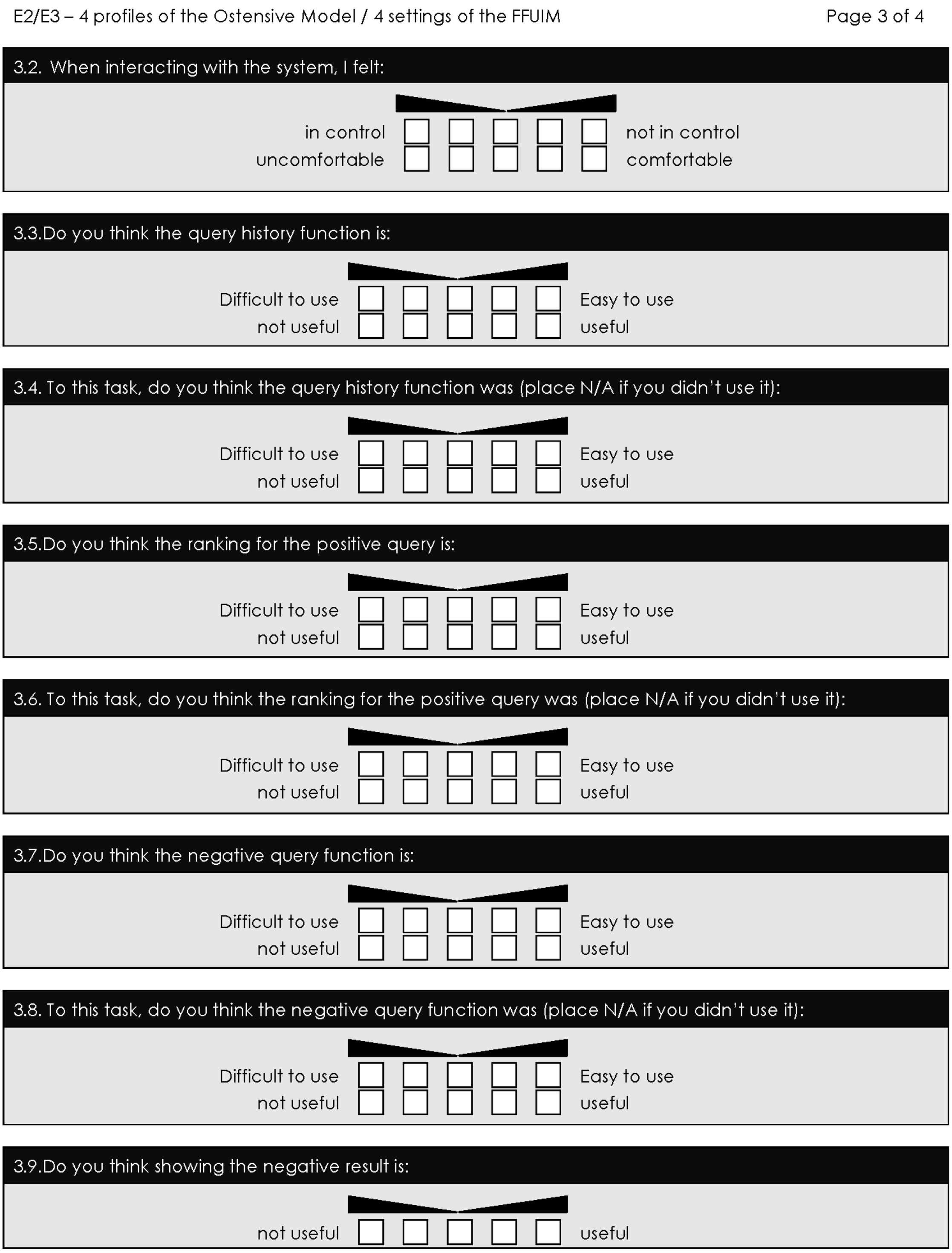
Very		some-what		Not at all
□	□	□	□	□
5	4	3	2	1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall, I think the system is:

terrible	□	□	□	□	□	wonderful
satisfying	□	□	□	□	□	frustrating
dull	□	□	□	□	□	stimulating
easy	□	□	□	□	□	difficult
efficient	□	□	□	□	□	inefficient
novel	□	□	□	□	□	standard



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 3 of 4

3.2. When interacting with the system, I felt:

in control □ □ □ □ □ not in control
uncomfortable □ □ □ □ □ comfortable

3.3.Do you think the query history function is:

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.4. To this task, do you think the query history function was (place N/A if you didn't use it):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.5.Do you think the ranking for the positive query is:

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.6. To this task, do you think the ranking for the positive query was (place N/A if you didn't use it):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.7.Do you think the negative query function is:

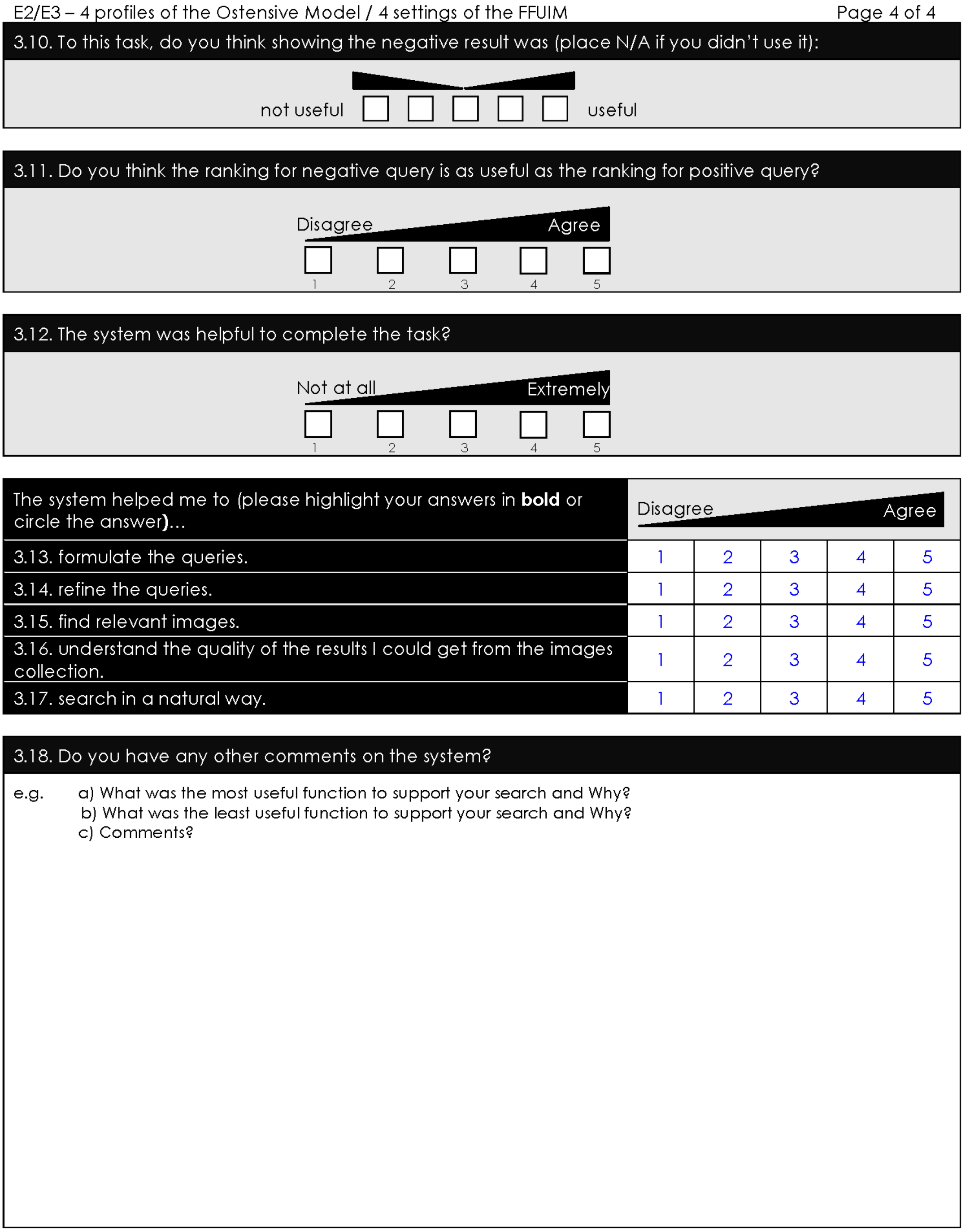
Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.8. To this task, do you think the negative query function was (place N/A if you didn't use it):

Difficult to use □ □ □ □ □ Easy to use
not useful □ □ □ □ □ useful

3.9.Do you think showing the negative result is:

not useful □ □ □ □ □ useful



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 4 of 4

3.10. To this task, do you think showing the negative result was (place N/A if you didn't use it):

not useful □ □ □ □ □ useful

3.11. Do you think the ranking for negative query is as useful as the ranking for positive query?

Disagree Agree

□ □ □ □ □

1 2 3 4 5

3.12. The system was helpful to complete the task?

Not at all Extremely

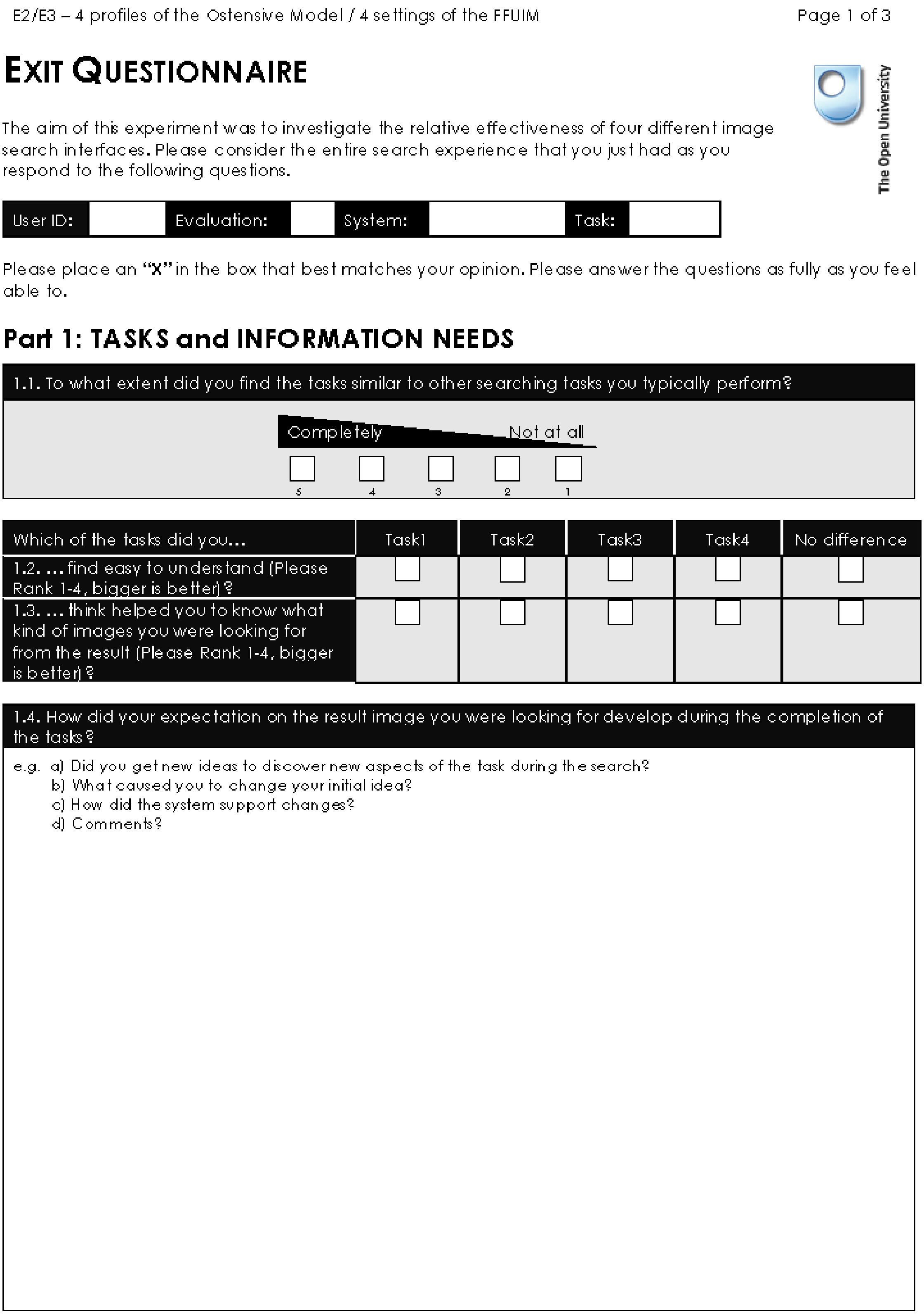
□ □ □ □ □

1 2 3 4 5

The system helped me to (please highlight your answers in **bold** or circle the answer)...	Disagree Agree				
3.13. formulate the queries.	1	2	3	4	5
3.14. refine the queries.	1	2	3	4	5
3.15. find relevant images.	1	2	3	4	5
3.16. understand the quality of the results I could get from the images collection.	1	2	3	4	5
3.17. search in a natural way.	1	2	3	4	5

3.18. Do you have any other comments on the system?

e.g. a) What was the most useful function to support your search and Why?
b) What was the least useful function to support your search and Why?
c) Comments?



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM

Page 1 of 3

EXIT QUESTIONNAIRE

The Open University

The aim of this experiment was to investigate the relative effectiveness of four different image search interfaces. Please consider the entire search experience that you just had as you respond to the following questions.

User ID:		Evaluation:		System:		Task:	

Please place an "X" in the box that best matches your opinion. Please answer the questions as fully as you feel able to.

Part 1: TASKS and INFORMATION NEEDS

1.1. To what extent did you find the tasks similar to other searching tasks you typically perform?

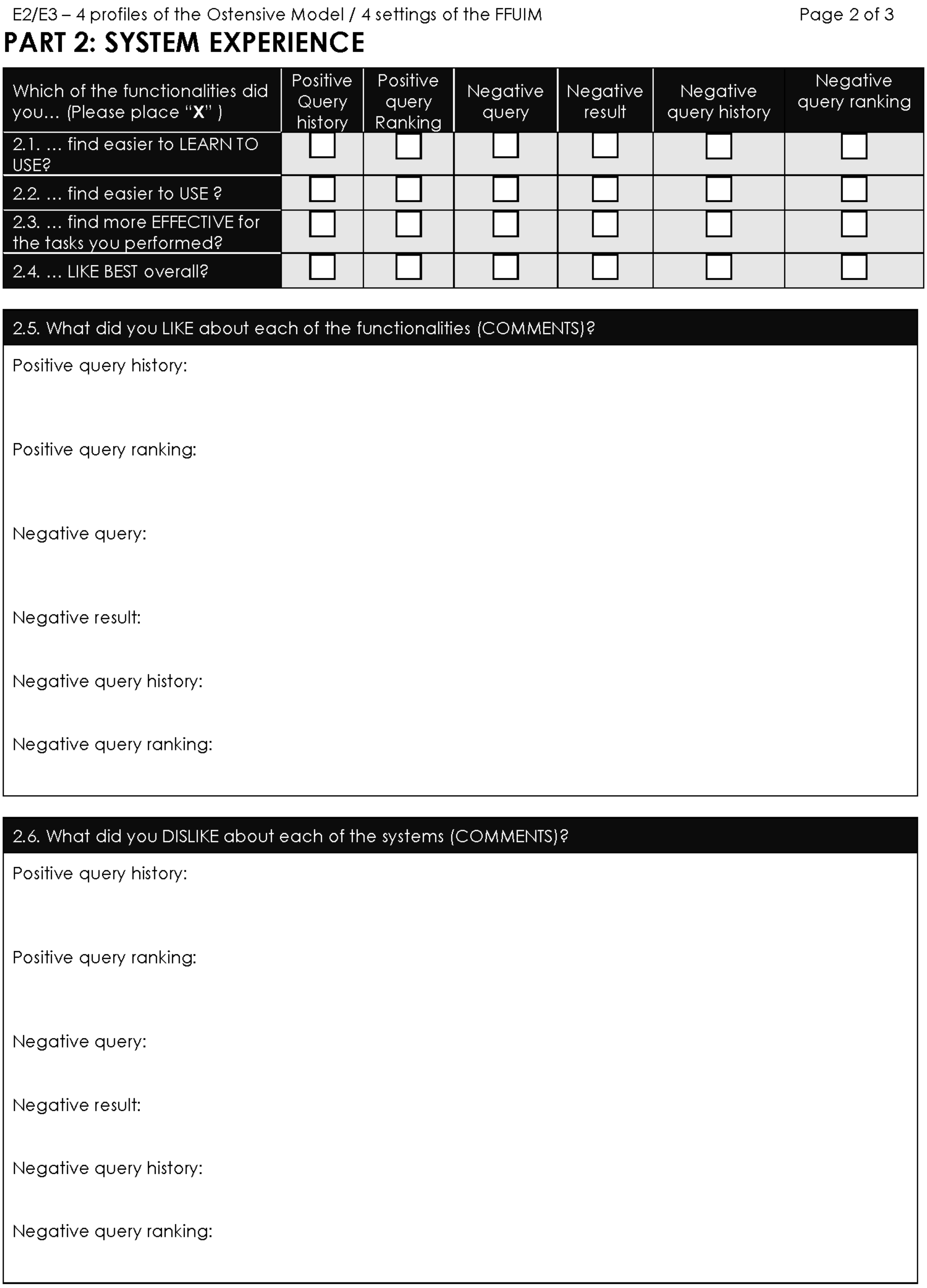
Completely — Not at all

□	□	□	□	□
5	4	3	2	1

Which of the tasks did you…	Task1	Task2	Task3	Task4	No difference
1.2. …find easy to understand (Please Rank 1-4, bigger is better)?	□	□	□	□	□
1.3. … think helped you to know what kind of images you were looking for from the result (Please Rank 1-4, bigger is better)?	□	□	□	□	□

1.4. How did your expectation on the result image you were looking for develop during the completion of the tasks?

e.g. a) Did you get new ideas to discover new aspects of the task during the search?
b) What caused you to change your initial idea?
c) How did the system support changes?
d) Comments?



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 2 of 3

PART 2: SYSTEM EXPERIENCE

Which of the functionalities did you… (Please place "**X**")	Positive Query history	Positive query Ranking	Negative query	Negative result	Negative query history	Negative query ranking
2.1. … find easier to LEARN TO USE?	□	□	□	□	□	□
2.2. … find easier to USE ?	□	□	□	□	□	□
2.3. … find more EFFECTIVE for the tasks you performed?	□	□	□	□	□	□
2.4. … LIKE BEST overall?	□	□	□	□	□	□

2.5. What did you LIKE about each of the functionalities (COMMENTS)?

Positive query history:

Positive query ranking:

Negative query:

Negative result:

Negative query history:

Negative query ranking:

2.6. What did you DISLIKE about each of the systems (COMMENTS)?

Positive query history:

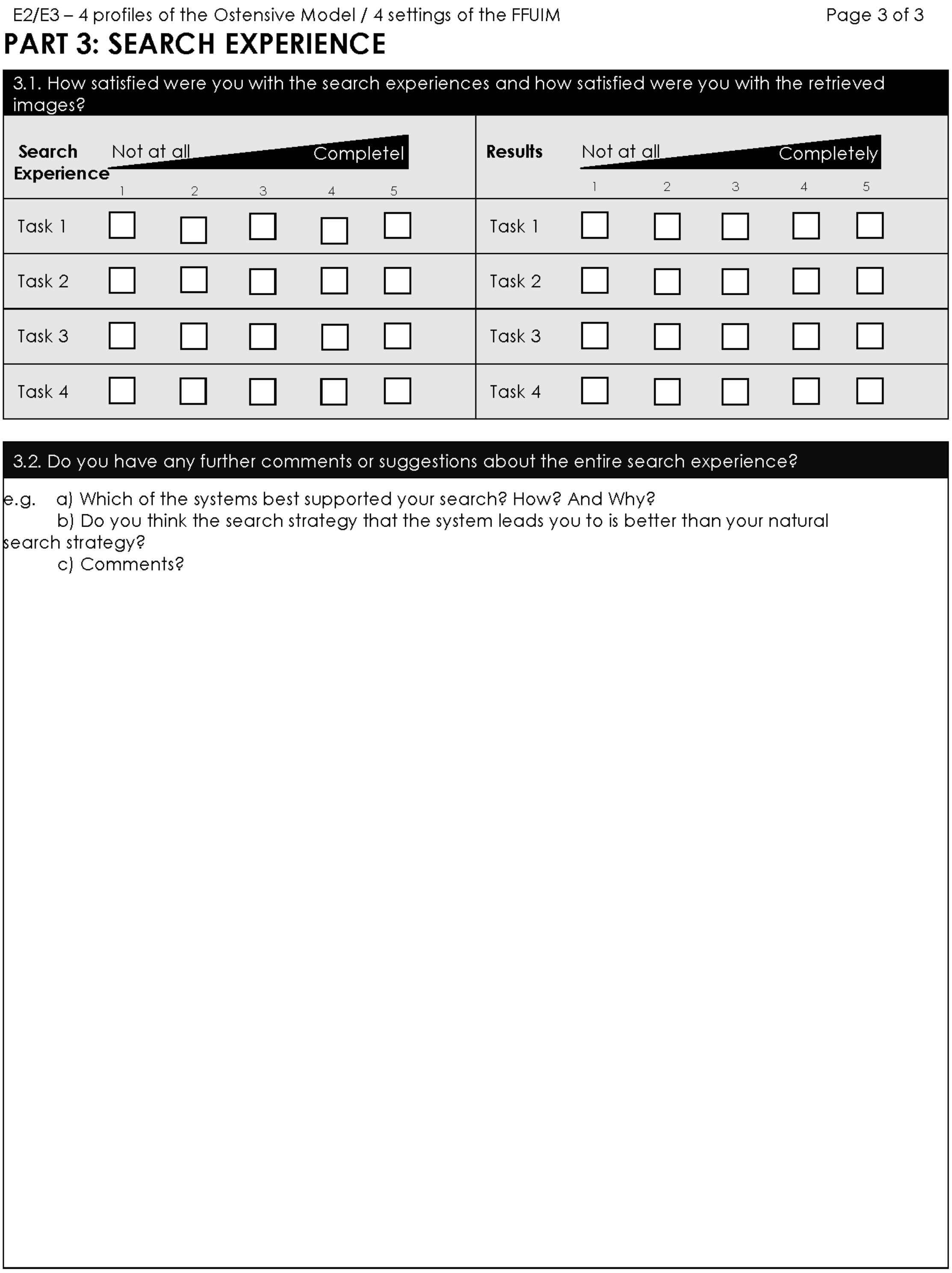
Positive query ranking:

Negative query:

Negative result:

Negative query history:

Negative query ranking:



E2/E3 – 4 profiles of the Ostensive Model / 4 settings of the FFUIM Page 3 of 3

PART 3: SEARCH EXPERIENCE

3.1. How satisfied were you with the search experiences and how satisfied were you with the retrieved images?

Search Experience	Not at all				Completel	**Results**	Not at all				Completely
	1	2	3	4	5		1	2	3	4	5
Task 1	□	□	□	□	□	Task 1	□	□	□	□	□
Task 2	□	□	□	□	□	Task 2	□	□	□	□	□
Task 3	□	□	□	□	□	Task 3	□	□	□	□	□
Task 4	□	□	□	□	□	Task 4	□	□	□	□	□

3.2. Do you have any further comments or suggestions about the entire search experience?

e.g. a) Which of the systems best supported your search? How? And Why?
b) Do you think the search strategy that the system leads you to is better than your natural search strategy?
c) Comments?

Bibliography

The benchathlon network, home of cbir benchmarking. Online. Available online at http://www.benchathlon.net/.

Imageclef - the clef cross language image retrieval track. Online. Available online at http://www.imageclef.org/.

Trec video retrieval evaluation. Online. Available online at http://www-nlpir.nist.gov/projects/trecvid/.

Baeza-Yates, Ricardo A. and Ribeiro-Neto, Berthier A. (1999). *Modern Information Retrieval.* ACM Press / Addison-Wesley.

Barecke, Thomas, Kijak, Eva, Nümberger, Andreas and Detyniecki, Marcin (2006, July). Summarizing video information using self-organizing maps. In *Proceeding of IEEE Internation Conference on Fuzzy Systems*, Canada, pp 540–546.

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

Berendt, Bettina and Kralisch, Anett (2009). A user-centric approach to identifying best deployment strategies for language tools: the impact of content and access language on web user behaviour and attitudes. *Information Retrieval 12*(3), 380–399.

Borlund, Pia and Ingwersen, Peter (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation 53*, 225–250.

Brini, Asma H. and Boughanem, Mohand (2003). Relevance feedback: introduction of partial assessments for query expansion. In *Proceeding of the 2nd EUSFLAT Conference*, pp 67–72.

Browne, Paul and Smeaton, Alan F. (2004). Video information retrieval using objects and ostensive relevance feedback. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pp 1084–1090.

Calder, Judith (1996). *Data Collection and Analysis*, Chapter Statistical Techniques, pp 225–261. The Open University.

Campbell, Iain (2000). Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *Journal of Information Retrieval 2*(1).

Chen, Chaomei, Gagaudakis, George and Rosin, Paul (2000). Similarity-based image browsing. In *Proceedings of the 16th IFIP World Computer Congress (International Conference on Intelligent Information Processing)*, Beijing, China, pp 206–213.

Chen, Chaur-Chin and Chu, Hsueh-Ting (2005). Similarity measurement between images. In *Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC'05)*. IEEE.

Cheng, Pei-Cheng, Chien, Been-Chian, Ke, Hao-Ren and Yang, Wei-Pang (2008). A two-level relevance feedback mechanism for image retrieval. *Expert Systems with Applications 34*(3), 2193–2200.

Chi, Ed H., Pirolli, Peter, Chen, Kim and Pitkow, James (2001). Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp 490–497. ACM.

Crucianu, Michel, Ferecatu, Marin and Boujemaa, Nozha (2004). Relevance feedback for image retrieval: a short survey. In state of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction, Including Datamodels and Languages, Report of the DELOS2 European Network of Excellence (6th Framework Programme).

Deselaers, Thomas, Keysers, Daniel and Ney, Hermann (2005). Fire – flexible image retrieval engine: Imageclef 2004 evaluation. In *Proceeding of Multilingual Information Access for Text, Speech and Images – Fifth Workshop of the Cross-Language Evaluation Forum (CLEF2004)*, Volume 3491, Bath, UK, pp 688–698. LNCS.

Dowdy, Shirley, Weardon, Stanley and Chilko, Daniel (2004). *Statistics for research* (Third ed). Wiley-Interscience.

Dumais, Susan T. and Belkin, Nicholas J. (2005). *The TREC Interactive Tracks: Putting the User into Search*, Chapter 6, pp 122–152. The MIT Press.

Dunlop, Mark D. (1997). The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems (TOIS) 15*(2), 137–153.

Eakins, John P., Riley, K. Jonathan and Edwards, Jonathan D. (2003). Shape feature matching for trademark image retrieval. In *Proceeding of International Conference on Image and Video Retrieval (CIVR)*, pp 28–38.

Fuhr, Norbert (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval 11*(3), 251–265.

Grubinger, Michael, Clough, Paul, Müller, Henning and Deselaers, Thomas (2006). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *In Proceedings of International Workshop OntoImageŠ2006 Language Resources for Content-Based Image Retrieval*, pp 13–23.

Heesch, Daniel (2005). *The NN^k Technique for image searching and browsing.* PhD thesis, Department of Electrical and Electronic Engineering, Imperial College London.

Heesch, Daniel and Rüger, Stefan. *Interaction models and relevance feedback in content-based image retrieval.*

Heesch, Daniel and Rüger, Stefan (2003). Performance boosting with three mouse clicks-relevance feedback for cbir. In *Proceeding of the European Conference on IR Research 2003*.

Heesch, Daniel, Yavlinsky, Alexei and Rüger, Stefan (2003). Performance comparison of different similarity models for cbir with relevance feedback. In et al., E.M.Bakker (Ed), *Proceeding of CIVR 2003*, pp 456–466.

ter Hofstede, A.H.M., Proper, H.A. and van der Weide, Th.P. (1996, September). Query formulation as an information retrieval problem. *The Computer Journal 39*(4), 255–274.

Hopfgartner, Frank, Urban, Jana, Villa, Robert and Jose, Joemon (2007). Simulated testing of an adaptive multimedia information retrieval system. In *In proceeding of Content-Based Multimedia Indexing (CBMI)*, pp 328–335.

Howarth, Peter and Rüger, Stefan (2005a, March). Fractional distance measures for content-based image retrieval. In *Proceedings of ECIR 2005 : European conference on IR research*, Santiago de Compostela , ESPAGNE. Springer, Berlin, ALLEMAGNE (2005) (Monographie).

Howarth, Peter and Rüger, Stefan (2005b). Robust texture features for still-image retrieval. *IEE Proc on Vision, Image and Signal 6*(152 (6)), 868–874.

Hu, Rui, Rüger, Stefan, Song, Dawei, Liu, Haiming and Huang, Zi (2008). Dissimilarity measures for content-based image retrieval. In *Proceeding of IEEE International Conference on Multimedia and Expo (ICME)*, pp 1365–1368.

Ingwersen, Peter (1992). *Information Retrieval Interaction.* Taylor Graham, London.

Ivory, Melody Y., Yu, Shiqing and Gronemyer, Kathryn (2004). Search result exploration: a preliminary study of blind and sighted users' decision making and performance. In *CHI '04: extended abstracts on Human factors in computing systems*, New York, NY, USA, pp 1453–1456. ACM.

Jamaal, Qudsia (2010, February). Google goggles - use pictures to search the web. Online.

Järvelin, Kalervo (2009). Explaining user performance in information retrieval: Challenges to ir evaluation. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR)*, Volume 5766, pp 289–296.

Jose, Joemon M., Furner, Jonathan and Harper, David J. (1998). Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp 232–240. ACM.

Käki, Mida (2005). *Enhancing Web Search Result Access with Automatic Categorization.* PhD thesis, Faculty of Information Sciences, Department of Computer Sciences, University of Tampere, Finland.

Käki, Mika and Aula, Anne (2008). Controlling the complexity in comparing search user interfaces via user studies. *Information Processiong and Management 44*(1), 82–91.

Kokare, Manesh, Chatterji, B.N. and Biswas, P.K. (2003). Comparison of similarity metrics for texture image retrieval. In *Proceeding of IEEE Conf. on Convergent Technologies for Asia-Pacific Region*, Volume 2, pp 571–575.

Kules, Bill and Shneiderman, Ben (2008). Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing and Management 44*(2), 463–484.

Lew, Michael S., Sebe, Nicu, Djeraba, Chabane and Jain, Ramesh (2006, February). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Cimmunications and Applications 2*(1), 1–19.

Liu, Haiming, Song, Dawei, Rüger, Stefan, Hu, Rui and Uren, Victoria (2008). Comparing dissimilarity measures for content-based image retrieval. In *AIRS2008*, pp 44–50.

Liu, Haiming, Uren, Victoria, Song, Dawei and Rüger, Stefan (2009). A four-factor user interaction model for content-based image retrieval. In *Proceeding of the 2nd international conference on the theory of information retrieval (ICTIR).*

Liu, Haiming, Zagorac, Srđan, Uren, Victoria, Song, Dawei and Rüger, Stefan (2009). Enabling effective user interactions in content-based image retrieval. In *Proceedings of the Fifth Asia Information Retrieval Symposium (AIRS).*

Marchionini, Gary (2006). Exploratory search: from finding to understanding. *Communications of the ACM 49*(4), 37–39.

Marchionini, Gary and White, Ryen W. (2009, March). Beyond search: Information seeking support systems. *IEEE Computer 42*(3), 30–32. Guest Editor's Introduction.

Marques, Oge and Furht, Borko (2002). *Content-Based Image and Video Retrieval.* Kluwer Academic Publishers.

Mitchell, Tom M. (1997). *Machine Learning.* New York: McGraw-Hill.

Mulholland, Paul, Zdrahal, Zdenek and Collins, Trevor (2008). Investigating the effects of exploratory semantic search on the use of a museum archive. In *Proceeding of the IEEE 2008 International Conference on Distributed Human-Machine Systems.*

Müller, Henning, Müller, Wolfgang, Marchand-Maillet, Stéphane and Pun, Thierry (2000, September). Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of the International Conference on Pattern Recognition (ICPR'2000)*, Volume 1, Barcelona, Spain, pp 1043–1046.

Nielsen, Jakob (2003, June). Information foraging: Why google makes people leave your site faster. Online. Available at: http://www.useit.com/alertbox/20030630.html.

Noreault, Terry, McGill, Michael and Koll, Matthew B. (1980). A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In *Proceeding of the 3rd annual ACM*

Conference on Research and development in inforamtion retreval, SIGIR'80, Kent, UK, pp 57–76. ACM: Butterworth Co.

Ojala, Timo, Pietikainen, Matti and Harwood, David (1996). Comparative study of texture measures with classification based on feature distributions. *Pattern Recognition 29*(1), 51–59.

Pickering, Marcus J. and Rüger, Stefan (2003, November). Evaluation of key frame-based retrieval techniques for video. *Computer Vision and Image Understanding 92*(2-3), 217–235.

Pirolli, Peter (1997). Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the Human Factors in Computing, CHI'97 Conference*, Atlanta GA, pp 3–10. ACM Press.

Pirolli, Peter (2007). *Information Foraging Theory Adaptive Interaction with Information.* Oxford University Press, Inc.

Pirolli, Peter and Card, Stuart (1995). Information foraging in information access environments. In *CHI '95: Mosaic of Creativity*, New York, NY, USA, pp 51–58. ACM.

Pirolli, Peter and Card, Stuart K. (1999). Information foraging. *Psychological Review 106*, 643–675.

Pirolli, Peter, Card, Stuart K. and Wege, Mija M. Van Der (2003). The effects of information scent on visual search in the hyperbolic tree browser. *ACM Trans. Comput.-Hum. Interact. 10*(1), 20–53.

Pirolli, Peter, tat Fu, Wai, Chi, Ed and Farahat, Ayman (2005, July). Information scent and web navigation: theory, models, and automated usability evaluation. In *Proceedings of Human Computer Interaction International 2005.*

Pirolli, Peter, Schank, Patricia, Hearst, Marti and Diehl, Christine (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp 213–220. ACM.

Puzicha, Jan (2001). *Distribution-Based Image Similarity*, Chapter 7, pp 143–164. Kluwer Academic Publishers.

Puzicha, Jan, Hofmann, Thomas and Buhmann, Joachim M. (1997). Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan.

Puzicha, Jan, Rubner, Yossi, Tomasi, Carlo and Buhmann, Joachim M. (1999, September). Empirical evaluation of dissimilarity measures for color and texture. In *Proceeding of the international conference on computer vision*, Volume 2, pp 1165–1172.

Rijsbergen, C.J.va (1979). *Information Retrieval* (Second ed), Volume 7. Butterworth Co(Publishers)Ltd.

Rubner, Yossi, Tomasi, Carlo and Guibas, Leonidas J. (2004, November). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision 40*(2), 99–121.

Rui, Yong, Huang, Thomas S., Ortega, Michael and Mehrotra, Sharad (1998). Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE transactions on circuits and video technology 8*(5), 644–655.

Ruthven, Ian and Lalmas, Mounia (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review 18*(2), 95–145.

Ruthven, Ian, Lalmas, Mounia and van Rijsbergen, Keith (2002). Ranking expansion terms with partial and ostensive evidence. In *Proceeding of the 4th International Conference on Conceptions of Library and Information Science (CoLIS4)*, pp 199–220.

Ruthven, Ian, Lalmas, Mounia and van Rijsbergen, Keith (2003). Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology 54*(6), 528–548.

Salton, Gerard (1989). *Automatic Text Processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley.

Saracevic, Tefko (1996, October). Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, Copenhagen,Denmark, pp 210–218.

Smeulders, Arnold W.M., Worring, Marcel, Santini, Simone, Gupta, Amarnath and Jain, Ramesh (2000, December). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(12), 1349–1380.

Spink, Amanda, Greisdorf, Howard and Bateman, Judy (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing Management 34*(5), 599–621.

Stephens, David W. and Krebs, John R. (1986). *Foraging Theory.* Princeton University Press.

Taylor, Arthur R., Cool, Colleen, Belkin, Nicholas J. and Amadio, William J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing and Management 43*(4), 1071–1084.

Urban, Jana (2007). *An Adaptive Approach for Image Organisation and Retrieval.* PhD thesis, Faculty of Information and Mathematical Sciences, the University of Glasgow.

Urban, Jana and Jose, Joemon M. (2006a). Ego: A personalized multimedia management and retrieval tool. *International Journal of Intelligent Systems 21*, 725 – 745.

Urban, Jana and Jose, Joemon M. (2006b). Evaluating a workspace's usefulness for image retrieval. *Multimedia Systems Journal 12*(4-5), 355–373.

Urban, Jana, Jose, Joemon M. and van Rijsbergen, Keith (2003). An adaptive approach towards content-based image retrieval. In *Proceeding of the Third*

International Workshop on Content-based Multimedia Indexing (CBMI'03), Rennes, France, pp 119–126.

Urban, Jana, Jose, Joemon M. and van Rijsbergen, Keith (2006, July). An adaptive technique for content-based image retrieval. *Multimedia Tools and Applications 31*, 1–28.

Wegner, Peter (1997). Why interaction is more powerful than algorithms. *Communications of the ACM 40*(5), 80–91.

White, R. W. and Ruthven, I. (2006). A study of interface support mechanisms for interactive information retrieval. *Journal of the American Society for Information Science and Technology*.

White, Ryen W., Drucker, Steven M., Marchionini, Gary, Hearst, Marti and m. c. schraefel (2007). Exploratory search and hci: designing and evaluating interfaces to support exploratory search interaction. In *CHI '07: extended abstracts on Human factors in computing systems*, New York, NY, USA, pp 2877–2880. ACM.

White, Ryen W., Jose, Joemon M. and Ruthven, Ian (2006, January). An implicit feedback approach for interactive information retrieval. *Information Processing and Management 42*(1), 166–190.

White, Ryen W., Kules, Bill, Drucker, Steven M. and Schraefel, M. C. (2006). Supporting exploratory search. *Communications of the ACM 49*(4), 37–39.

White, Ryen W. and Morris, Dan (2007, July). Investigating the querying and browsing behavior of advanced search engine users. In *Proceeding of The 30th Annual International ACM SIGIR Conference*, Amsterdam, pp 255–262.

White, Ryen W. and Roth, Resa A. (2009). *Exploratory Search: Beyond the Query - Response Paradigm*. Morgan and Claypool Publishers.

White, Ryen W., Ruthven, Ian and Jose, Joemon M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceeding of SIGIR 2005*.

Wu, Hong, Lu, Hanqing and Ma, Songde (2004). Willhunter: Interactive image retrieval with multilevel relevance measurement. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR).*

Zhang, Dengsheng and Lu, Guojun (2003, December). Evaluation of similarity measurement for image retrieval. In *Procedding of IEEE International Conference on Neural Networks Signal*, Nanjing, pp 928–931. IEEE.

Zhou, Xiang Sean and Huang, Thomas S. (2003, April). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems 8*(6), 536–544.