An information foraging theory based user study of an adaptive user interaction framework for content-based image retrieval

Conference or Workshop Item

An Information Foraging Theory Based User Study of an Adaptive User Interaction Framework for Content-based Image Retrieval

Haiming Liu[1], Paul Mulholland[1], Dawei Song[2], Victoria Uren[3], Stefan Rüger[1]

[1]Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK
[2]School of Computing, The Robert Gordon University, Aberdeen, AB25 1HG, UK
[3]Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
{h.liu,p.mulholland,s.rueger}@open.ac.uk
d.song@rgu.ac.uk;v.uren@dcs.shef.ac.uk

Abstract. This paper presents the design and results of a task-based user study, based on Information Foraging Theory, on a novel user interaction framework - uInteract - for content-based image retrieval (CBIR). The framework includes a four-factor user interaction model and an interactive interface. The user study involves three focused evaluations, 12 simulated real life search tasks with different complexity levels, 12 comparative systems and 50 subjects. Information Foraging Theory is applied to the user study design and the quantitative data analysis. The systematic findings have not only shown how effective and easy to use the uInteract framework is, but also illustrate the value of Information Foraging Theory for interpreting user interaction with CBIR.

Key words: Information Foraging Theory, User interaction, Four-factor user interaction model, uInteract, content-based image retrieval

1 Introduction

In an effort to improve the interaction between users and search systems, some researchers have focused on developing user interaction models and/or interactive interfaces.

Spink et al. (1998) proposed a three-dimensional spatial model to support user interactive search [8]. Campbell (2000) proposed the Ostensive Model (OM), which indicates the degree of relevance relative to when a user selected the evidence from the results set [1]. Ruthven et al. (2003) adapted two dimensions from Spink et al.'s model combined with OM [7]. Liu et al. (2009) proposed an adaptive four-factor user interaction model (FFUIM) based on above models for content-based image retrieval (CBIR) [3].

The interaction models need to be delivered by visual interactive interfaces for further improving the user interaction. For instance, Urban et al. (2006) developed a visual image search system based on the OM [10]. Liu et al. (2009) proposed an interactive CBIR interface that successfully delivered the FFUIM and allowed users to manipulate the model effectively [4].

To date, most of the evaluations of interactive search systems are still system-oriented. For instance, the search results of an automatic pseudo or simulated user evaluation are measured by precision and recall. However, users in real-life seek to optimize the entire search process, not just results accuracy. Evaluation of output alone is not enough to explain the effectiveness of the systems or users' search experience [2].

Pirolli (2007) stated in Information Foraging Theory [6] that the two inter-related environments, namely task environment and information environment, will affect the information search process. 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.*" [5]. We consider that a clear task environment and a rich information environment determine a forager's effective and enjoyable search experience.

With respect to Information Foraging Theory, our task-based user study applies simulated searching tasks with different complexity levels, and employed users with different age and image search experience. This way, we can investigate how the different task environments and the users' different information environments affect evaluation results.

2 uInteract Framework

The uInteract Framework aims to improve user interaction and users' overall search experience. The framework includes a four-factor user interaction model (FFUIM) and an interactive interface. HSV colour feature, City Block dissimilarity measure and ImageCLEF2007 collection are employed by the framework.

2.1 Four-factor User Interaction Model

The four factors taken into account in the model are relevance region, relevance level, time and frequency [3].

The relevance region comprises two sub-regions: relevant (positive) evidence and non-relevant (negative) evidence. The two sub-regions contain a range of relevance levels. The relevance level is a quantitative level, which indicates how relevant/non-relevant the evidence is. The time factor adapts OM [1], which indicates the degree of relevance/non-relevance relative to when the evidence was selected. The frequency factor captures the number of appearances of an image in the user selected evidence both for positive and negative evidence separately.

The FFUIM works together with two fusion approaches, namely Vector Space Model (VSM) for the positive query only scenario and K-Nearest Neighbours (KNN) for the both positive and negative queries scenario.

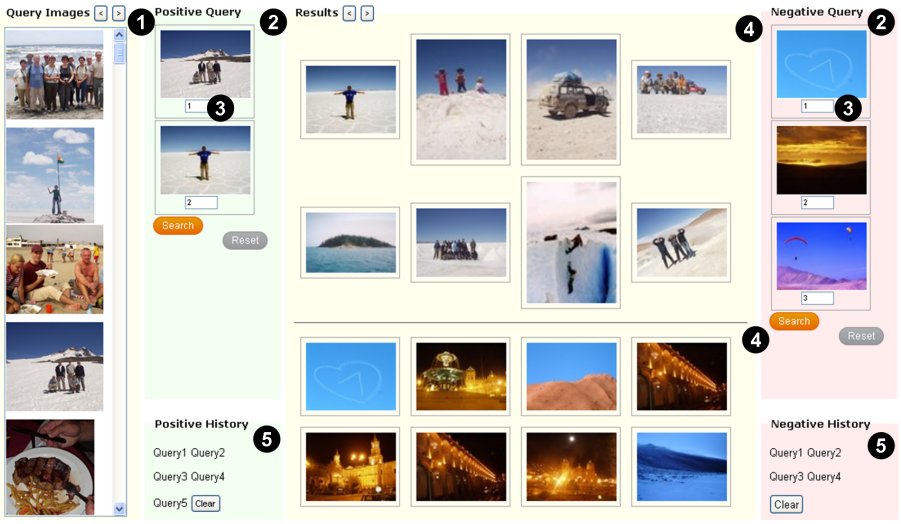


Fig. 1. The uInteract interface (the keys are explained in the main text.)

2.2 uInteract Interface

The key features of the uInteract interface (Figure 1) [4] are as follows: (1) **Query images** panel provides a browsing functionality that facilitates the selection of the initial query images. (2) Users can provide both positive and negative examples to a search query in the **positive and negative query** panel, and further expand or remove images from that query. (3) By allowing the user to override the system-generated **scores** (integer 1-20) of positive and negative query images, users can directly influence the relevance level of the feedback. (4) The displaying of the results in **results** shows not only the best matches but also the worst matches. This functionality can enable users to gain a better understanding of the data set they are searching. (5) Combining both **positive and negative query history** functionality has not previously been undertaken in CBIR. 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.

3 User Study Methodology

Our user study contained three focused evaluations: evaluation1 (E1) was to evaluate the ease of use and usefulness of the uInteract interface; evaluation2 (E2) was to evaluate the performance of the four profiles of the OM; evaluation3 (E3) was to evaluate the effectiveness of the four different settings of the FFUIM.

Fifty subjects were employed for the user study. They were 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. The subjects were classified into two categories - inexperienced or experienced - based on their image search experience. We considered that people were experienced subjects if they searched images at least once a week, and otherwise they were inexperienced subjects. The 50 subjects were divided into three groups, 17, 16 and 17 subjects assigned to E1, E2 and E3 respectively. In each evaluation, the subjects were asked to complete four search tasks with different complexity level on four systems randomly (limited to five minutes for each task).

The complexity level of each task in E1 was reflected by the task description. Task1 (E1T1) provided both search topic and example images, so we considered it the easiest task in term of the "easiness" of formulating the query and identifying the information need. Task2 (E1T2) gave example images without a topic description, so we considered it harder than E1T1. Task3 (E1T3) had only a topic but no image examples, which was even harder than E1T2. Task4 (E1T4) described a broad search scenario without any specific topic and image examples, so it was the hardest task in our view.

The four testing systems of E1 were: system1 (I1) had a baseline interface, where users were allowed to give positive feedback from search results through a simplified interface; system2 (I2) - an interface based on Urban et al.'s [10] model, provided positive query history functionality which was an addition to I1; system3 (I3) - an interface based on Ruthven et al.'s [7] model, enhanced I2 by allowing users to assign a relevance value to the query images; system4 (I4) was the uInteract interface [4], which added negative query, negative result and negative query history functionalities based on I3.

The four tasks in E2 and E3 used the same description structure, which had both specific search topic and three example images. The complexity level of each task was based on the search accuracy of the query images of the tasks from our earlier lab-based simulated experiments results. The mean average precision (MAP) of task1 (T1), task2 (T2), task3 (T3) and task4 (T4) was 0.2420, 0.0872, 0.0294, 0.0098 respectively. We considered T1 was the easiest task with the highest precision, and then it was followed by T2 and T3. T4 had lowest precision thus we took it as the hardest task.

The four testing systems of E2 were: system1 (OM1) applied the increasing profile of the OM; system2 (OM2) applied the decreasing profile of the OM; system3 (OM3) applied the flat profile of the OM; system4 (OM4) applied the current profile of the OM.

The four testing systems of E3 were: system1 (FFUIM1) delivered the relevance region factor and time factor of the FFUIM and here we apply the increasing profile of the OM to both positive and negative queries; system2 (FFUIM2) delivered the relevance region factor, time factor and relevance level factor of the FFUIM, and here we combined the increasing profile of the OM with the relevance scores provided by the users for both positive and negative queries; system3 (FFUIM3) delivered the relevance region factor and time factor and frequency factor of the FFUIM, and here we combined 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) delivered the relevance region factor, time factor, relevance level factor and frequency factor of the FFUIM, and here we combined 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.

The data was collected by questionnaires and actual search results. The questionnaires used five point Likert scales, and included entry questionnaire, post-search questionnaire, and exit questionnaire.

3.1 Main Performance Indicators and Nine Hypothesis of Quantitative Analysis

The main performance indicators (PIs) of the qualitative data are generated from the questionnaires and actual search results. The main PIs of E1, E2 and E3 are listed in Figure 1.

The following nine evaluation hypotheses aims for investigating not only the effectiveness and ease of use of the uInteract framework, but also how the different task environments and the users' information environments will affect the performance indicators.

- **Hypothesis1**: Task Order (PI5) and System Order (PI7) will affect the PI8-33 provided by subjects because of familiarity or fatigue;
- **Hypothesis2**: System (PI6) will affect the PI8-33;
- **Hypothesis3**: Task (PI4) will affect the PI8-33 provided by subjects because of different complexity level;
- **Hypothesis4**: The interaction between Task (PI4) and System (PI6) will influence the scores of the PI8-33;
- **Hypothesis5**: Person (PI1) will affect the PI8-33, based on individual differences;
- **Hypothesis6**: The subjects' Age (PI2) and prior Image Search Experience (PI3) of the subjects will affect subjects' opinion on overall search experience (PI8-21);
- **Hypothesis7**: The subjects' Age (PI2) and prior Image Search Experience (PI3) of the subjects will affect subjects' opinion on the functionalities of the interfaces (PI22-33);
- **Hypothesis8**: System (PI6) and Task (PI4) will have impact on Precision (PI34) of the search results;
- **Hypothesis9**: System (PI6) and Task (PI4) will have impact on Recall (PI35) of the search results.

3.2 Quantitative Data Analysis Procedure

Quantitative data analysis is supported by the use of statistical software - SPSS. The analysis procedure is as follows:

	Performance Indicator	Description	From
1	Person	Subject (User) ID	Entry questionnaires of E1, E2 and E3
2	Age	Age of subjects	
3	ImageSearchExperience	Image search experience of subjects	
4	Task	Task ID	Post-search questionnaires of E1, E2 and E3
5	TaskOrder	Performing order of tasks	
6	System	System ID	
7	SystemOrder	Performing order of systems	
8	TaskGeneralFeeling	Subjects' general feeling to tasks	
9	TaskGeneralPerformance	The general performance of tasks	
10	EnoughTime	Subjects feel they have enough time to complete takes	
11	NextAction	Subjects know what to do next	
12	ResultSatisfaction	Subjects satisfy with search results	
13	HaveInitialIdea	Subjects have initial idea on what they are looking for	
14	MatchedInitialIdea	Subjects think the search result matches their initial idea	
15	SystemGeneralFeeling	Subjects' general feeling to systems	
16	SystemNovelty	How subjects think the novelty of systems	
17	FeelInControl	Subjects feel in control when they perform the tasks	
18	FeelComfortable	Subjects feel comfortable on using systems	
19	SystemSatisfaction	Subjects satisfied with systems	
20	KnowCollection	Interface help subjects to understand the quality of the collection where they are searching from	
21	SearchInNaturalWay	Systems support subjects natural search strategy	
22	QueryHistoryEasyToUse	Subjects think query history is easy to use	Post-search questionnaires of E1, E2 and E3 Exit questionnaires of E2 and E3
23	QueryHistoryUseful	Subjects think query history can be useful	
24	QueryHistoryUsefulHere	Subjects think query history is useful for this task	
25	PQScoringEasyToUse	Subjects think scoring positive query images is easy to use	
26	PQScoringUseful	Subjects think scoring positive query images can be useful	
27	PQScoringUsefulHere	Subjects think scoring positive query images is useful for this task	
28	NQueryEasyToUse	Subjects think negative query is easy to use	
29	NQueryUseful	Subjects think negative query can be useful	
30	NQueryUsefulHere	Subjects think negative query is useful for this task	
31	NResultUseful	Subjects think negative result can be useful	
32	NResultUsefulHere	Subjects think negative result is useful for this task	
33	NScoringAsUsefulAsPScoring	Subjects think scoring negative query images is as useful as scoring positive query images	
34	Precision	Search precision based 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 1. The main performance indicators from the three evaluations for qualitative data analysis

1. Identify so-called 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 asked 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 to every image in a union result set of a task, and the raters do the same to the result images of the other 11 tasks. We test the reliability of the raters' scores of all the images for the 12 tasks by Cronbach's Alpha statistics 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. 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 complete the task;

2. Obtain the figures for the performance indicators listed in Figure 1 from the questionnaires and the actual search results for the three focused evaluations, and test the nine hypotheses that we intend to investigate in Section 3.1 by factorial ANOVA statistical tests;
3. Analyze the testing results that we have obtained from the ANOVA test based on Information Foraging Theory.

4 Evaluation Results and Analysis

Table 2 shows the results of the three evaluations that are obtained by ANOVA analysis (with $\alpha = 0.05$) of the main PIs based on the nine hypotheses.

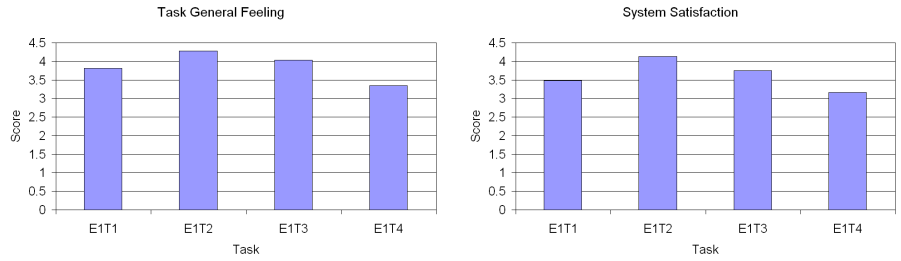
From Table 2 we can see that (1) the different complexity level of the tasks and the different age and image experience of the users have very strong effect on the PIs, which confirms the importance of the task and information environment stated in Information Foraging Theory; (2) the performing order of the tasks and systems does not affect the PIs, which implies the familiarity or fatigue with the task and the system does not make a difference to the subjects' scores on the indicators; (3) there is no significant difference between the testing systems from three evaluations. This may be because that Task and Person indicator strongly impinge on the PIs. The following sections will report how the different tasks (task environment) and different users - different age and image search experience (information environment) affect the scores of the PIs.

4.1 Effects of the task environment

Task (PI4) strongly influences most PIs in three evaluations.

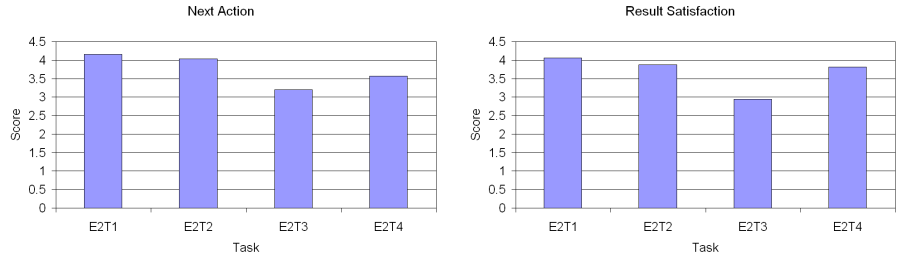
Hypotheses	E1	E2	E3
Hypothesis1: Task Order (PI5) and System Order (PI7) will affect the PI8-33 provided by subjects because of familiarity or fatigue	Not supported	Not supported	Not supported
Hypothesis2: System (PI6) will affect the PI8-33	Not supported	Not supported	Not supported
Hypothesis3: Task (PI4) will affect the PI8-33 provided by subjects because of different complexity level	Partially supported	Partially supported	Partially supported
Hypothesis4: The interaction between Task (PI4) and System (PI6) will influence the scores of the PI8-33	Partially supported	Partially supported	Partially supported
Hypothesis5: Person (PI1) will affect the PI8-33, based on individual differences	Partially supported	Partially supported	Partially supported
Hypothesis6: The subjects' Age (PI2) and prior Image Search Experience (PI3) of the subjects will affect subjects' opinion on overall search experience (PI8-21)	Partially supported	Partially supported	Partially supported
Hypothesis7: The subjects' Age (PI2) and prior Image Search Experience (PI3) of the subjects will affect subjects' opinion on the functionalities of the interfaces (PI22-33)	Partially supported	Partially supported	Partially supported
Hypothesis8: System (PI6) and Task (PI4) will have impact on Precision (PI34) of the search results	Partially supported	Not supported	Partially supported
Hypothesis9: System (PI6) and Task (PI4) will have impact on Recall (PI35) of the search results	Partially supported	Partially supported	Partially supported

Table 2. How the nine hypotheses have been supported or rejected in E1, E2 and E3 (partially = part of the PIs have significantly supported the hypotheses)



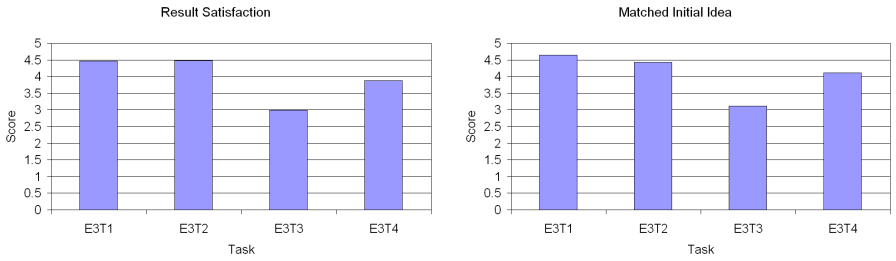
(a) E1_TaskGeneralFeeling (b) E1_SystemSatisfaction

Fig. 2. E1: examples of effects of Task on performance indicators (8-33)



(a) E2_NextAction (b) E2_ResultSatisfaction

Fig. 3. E2: examples of effects of Task on performance indicators (8-33)



(a) E3_ResultSatisfaction (b) E3_MatchedInitialIdea

Fig. 4. E3: examples of effects of Task on performance indicators (8-33)

For E1T2, E1T3 and E1T4 in E1, subjective feelings decrease as task difficulty increases. This is the case for a number of PIs, e.g. (Figure 2), TaskGeneralFeeling (PI8), SystemSatisfaction (PI19), etc. However, E1T1 subjective feelings are relatively low even though the task is easier. This may be have been because the image examples used in the task being difficult to interpret, therefore making the task more difficult than intended.

For T1, T2 and T3 in E2, there is a decrease in subjective feelings as task difficulty increases, e.g. (Figure 3), NextAction (PI11), ResultSatisfaction (PI12), etc. However, subjective feelings were relatively high for T4 even though it was the hardest task. This may be because subjects tended to give an over-generous definition of what images were relevant to the solution, therefore making the task easier for themselves. This was reflected in the low precision scores for this task.

One interesting observation from the analysis is that the trend in subjective feelings for T1, T2 and T3 in E3 become more negative as the task becomes harder, e.g. (Figure 4), ResultSatisfaction (PI12), MatchedInitialIdea (PI14), etc. As in E2, the subjects are relatively more positive about T4 because they had a generous definition of what images were relevant to the solution.

4.2 Effects of the information environment

Table 3 shows how the different users' information environments - different age and image search experience - relate to the scores of the PIs.

For E1 and E2 we can see that:

- Some PIs tend to correlate with age - i.e. more positive feelings toward the task or system with age.
- Some PIs tend to correlate with age only for experienced users and inversely correlate for inexperienced users - i.e. increasingly positive feelings for experienced users as they get older, decreasingly positive feelings for inexperience users as they get older.
- Some PIs tend to be higher for experienced users.

Relationship	E1	E2	E3
Correlate with age	PI117, PI18, PI19, PI20	PI10, PI11, PI31, PI32, PI29, PI30	PI10, PI16, PI18, PI19, PI20, PI24, PI33, PI8
Inversely correlate with age			
Correlate with age for experienced users	PI12, PI14	PI18, PI19, PI21, PI23, PI33	PI21
Inversely correlate with age for experienced users			PI28, PI29, PI30, PI31
Correlate with age for inexperienced users			PI28, PI29, PI30, PI31
Inversely correlate with age for inexperienced users	PI12, PI14	PI18, PI19, PI21, PI23, PI33	PI21
Higher for experienced users		PI22, PI25, PI28, PI31	PI23, PI24, PI26
Higher for inexperienced users			PI19, PI20

Table 3. The relationship between the scores of the main performance indicators and the information environment in E1, E2 and E3

For E3 there are some exceptions, in which some factors (PI28, PI29, PI30 and PI31) correlate with age for inexperienced users, with inexperienced users also having higher scores for PI19 and PI20. This can be inferred from PI20, PI28, PI29, PI30 and PI31 all being related to user perception of negative functions - i.e. inexperienced users can adapt the new negative functions easier than experienced users, and inexperienced users have increasingly positive perception to the new functions as they get older.

In summary, the lesson learnt from the data can be mapped back to task and information environment, along the lines of the following:

- Task difficulty can effect a range of measures and the difficulty of the task might differ from expectations depending on how users interpret the materials and instructions.
- Age can affect perception of the task and and system, with older subjects perhaps more likely to have a positive perception.
- Experience can effect perception of the task and and system, with experienced subjects more likely to have positive feelings for the specific functionalities of the system, and with inexperienced subjects likely to have positive feelings for the entire system.
- Age may interact with experience in certain ways depending on the subjects' perception of the functionalities and the search process in general.

These findings have implications for how CBIR evaluations are designed and analysed: choice of PIs, selection of tasks and selection of subjects, etc.

5 Conclusions and Future Work

The quantitative data analysis results of E1, E2 and E3 show that the different tasks and different users have stronger effects on the performance indicators than the different systems. This finding reinforces the importance of the task and information environment concept of Information Foraging Theory for interactive

CBIR study. A clear trend is found from the influence of the Task (PI4) indicator: the subjects tend to give higher scores to the performance indicators when they perform an easier task although there are exceptions (1) when the image examples are not intuitive, and (2) how the subjects perform the tasks. However, the results of the three evaluations do not show a clear trend on how the Person (PI1) indicator affected the performance indicators. We have tested the effects of the Age (PI2) and Image Search Experience (PI3) of the subjects, but found varied results across the three evaluations. Therefore, we realize that the simple user classification based on Age (PI2) and Image Search Experiences (PI3) is not sufficient, so that we will need to investigate in-depth how to better classify user types and how the different user types affect the users' search preferences.

6 Acknowledgments

This work was partially supported by AutoAdapt project funded by the UK's Engineering and Physical Sciences Research Council, grant number EP/F035705/1.

References

1. I. Campbell. Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. *Journal of Information Retrieval*, 2(1), 2000.
2. K. Järvelin. 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, pages 289–296, 2009.
3. H. Liu, V. Uren, D. Song, and S. Rüger. 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)*, 2009.
4. H. Liu, S. Zagorac, V. Uren, D. Song, and S. Rüger. Enabling effective user interactions in content-based image retrieval. In *Proceedings of the Fifth Asia Information Retrieval Symposium (AIRS)*, 2009.
5. P. Pirolli. *Information Foraging Theory Adaptive Interaction with Information*. Oxford University Press, Inc, 2007.
6. P. Pirolli and S. K. Card. Information foraging. *Psychological Review*, 106:643–675, 1999.
7. I. Ruthven, M. Lalmas, and K. van Rijsbergen. Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology*, 54(6):528–548, 2003.
8. A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing Management*, 34(5):599–621, 1998.
9. J. Urban and J. M. Jose. Evaluating a workspace's usefulness for image retrieval. *Multimedia Systems Journal*, 12(4-5):355–373, 2006.
10. J. Urban, J. M. Jose, and K. van Rijsbergen. An adaptive technique for content-based image retrieval. *Multimedia Tools and Applications*, 31:1–28, July 2006.