User Centred Evaluation of A Recommendation Based Image Browsing System

Teerapong Leelanupab, Frank Hopfgartner, Joemon M. Jose

University of Glasgow, Glasgow, G12 8RZ, United Kingdom,
{kimm,hopfgarf,jj}@dcs.gla.ac.uk,

Abstract. In this paper, we introduce a novel approach to recommend images by mining user interactions based on implicit feedback of user browsing. The underlying hypothesis is that the interaction implicitly indicates the interests of the users for meeting practical image retrieval tasks. The algorithm mines interaction data and also low-level content of the clicked images to choose diverse images by clustering heterogeneous features. A user-centred, task-oriented, comparative evaluation was undertaken to verify the validity of our approach where two versions of systems – one set up to enable diverse image recommendation – the other allowing browsing only – were compared. Use was made of the two systems by users in simulated work task situations and quantitative and qualitative data collected as indicators of recommendation results and the levels of user's satisfaction. The responses from the users indicate that they find the more diverse recommendation highly useful.

Key words: Image Browsing, Recommendation, Experiment Design

1 Introduction

Despite technological advances in multimedia information retrieval, image search still remains a challenging problem due to the Semantic Gap [1] between low-level features and high-level semantics. Low-level features of images alone are incapable of providing effective image retrieval. Furthermore, information seeking is a complicated task involving changes in user information needs and lack of knowledge of data collections. Users require a better environment that allows them to explore and browse their searching materials. An ideal image retrieval system would therefore support users in their complex information seeking task. The system should support independent search sessions, adapt retrieval results to the user's current information need, and provide additional recommendations. In this paper we introduce a new image retrieval system which provides users with additional image recommendations by exploiting their interactions with the system. The system supports adaptive browsing, multiple individual search sessions, and diverse recommendations based on user interactions. In order to evaluate our recommendation approach, we performed a user-centred, comparative evaluation.

The rest of this paper is organised as follows. In Section 2, we survey related work in this area. Section 3 introduces our approach of image browsing through recommendation and research questions. Section 4 introduces the system architecture of our recommendation system. The experimental methodology is detailed in Section 5, followed by a review of the results in Section 6. The main findings of the user study are discussed in Section 7. Finally, we conclude our work in Section 8.

2 Related Work

2.1 Image Browsing

One of the main challenges in information retrieval is to retrieve documents that match the user's interest. Salton and Buckley [2] argued that this challenge is further exacerbated by the problem of formulating *good* search queries. Users lack a good knowledge about the data collections and hence do not know which search queries to use. This problem is even more urgent in content based image retrieval, where efficient search queries might consist of a set of low-level features. Users face here the Semantic Gap [1], the difference between a user's understanding of an image and its representation by low-level features. Frames sharing similar low-level features do not necessarily represent similar concepts.

One approach towards bridging the Semantic Gap is to provide users with better searching environments which allow them to explore and browse their searching materials. A smart interface design can provide the user with an easy access to an image corpus.

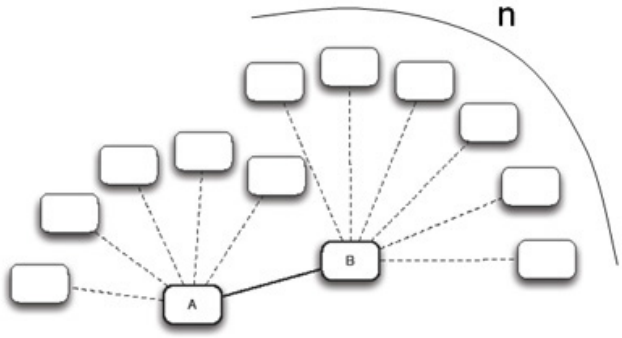

Fig. 1: Graph Based Image Browsing

A well studied approach (i.e. [3, 4]) to assist users in accessing their image collection is to offer a graph-based representation of retrieved images. Figure 1 illustrates an example of this approach. Given an image A as a node in a graph, similar images are shown as leaves of this node. Selecting on one of these leaves (i.e. image B) will implicitly provide relevance feedback, which the system adaptively retrieve other similar images related to that leave. Here, both node

A and B are considered as a query. This approach constructs an image graph by taking into account the structural relationship amongst images based on users' feedback. Urban and Jose [5] showed that using the images of a user's browsing path can be considered as relevance feedback and a search query can hence be expanded using this feedback. Besides, they illustrated that the ostensive model of developing information need [6], which was used to their study, can be applied to better adapt the search query to the user's current information need. This model considers the iteration when feedback was provided. Lower weighting was given to earlier iterations since they found that the user has most likely narrowed down his search interest in later few iterations. As a result, the subsequent information in path is assumed to be more relevant to the user.

2.2 Diversity Ranking

Traditional ranking models such as TF×IDF or BM 25 have shown their effectiveness in presenting retrieval results that best match the users' search query. Depending on the size of the indexed collection; however, the potentially large amount of returned documents can overwhelm the user. Hence, it is useful to present retrieval results not only based on their ranking with respect to the initial search query, but also based on a diverse set of documents capable of covering all possible aspects of search topics without harming precision.

Zhang et al. [7] carried out some initial work on diversification in the text retrieval domain. They introduce a novel ranking methodology, called Affinity Ranking, which ranks retrieval results based on their diversity and information richness. Similar work has been performed in the multimedia domain. Song et al. [8] argued for re-ranking image retrieval results based on topic richness, while keeping an acceptable retrieval performance. Moreover, van Zwol et al. [9] introduced a diversification model that is based on image annotations. They balance precision and diversity by estimating the query model from the distribution of tags which favours the dominant sense of the query. A problem within this approach, however, is the often missing textual annotation of multimedia documents. Therefore, Karypis [10] compared different diversity ranking approaches based on various low-level features extracted from images. Based on his experiments, he identifies a number of best dissimilarity measures.

2.3 Recommendation Approaches

Understanding the user's interest is an important factor towards satisfying the user and his information need. A common approach to identify this need is to exploit explicit or implicit user feedback. Traditionally, explicit relevance feedback has been used; however, there are a number of problems with this approach. Providing explicit feedback can be a cognitively taxing process. Users are forced to update their need constantly and this can be a difficult process when their information need is vague [11] or when they are unfamiliar with the document collection [2]. Also previous evaluations have found that users of explicit feedback

systems often do not provide sufficient levels of feedback for adaptive retrieval algorithms to work [12].

Implicit feedback has been shown to be a good indicator of interest in a number of areas in IR [13]. White et al. [14] used the concept of "search trails", meaning the search queries and document interactions sequences performed by the users during a search session, to enhance web search. Craswell and Szummer [15] applied a random walk on a graph of user click data, to help retrieve relevant images for user searches. Liu et al. [16] used a graph representation based on the textual features associated with a video to improve result list ranking.

3 Image Browsing through Recommendation

The above studies indicate the progress that has been made in image browsing and retrieval. At the same time, the solutions provided so far are not effective enough from a user's point of view. For example, the following problems arise: firstly, the difficulty in formulating and communicating user needs [17]; secondly, the inconsistency of the high-level semantics users have in mind and the low-level features used for matching; and the deficiency in supporting explorative image browsing. To address these problems, we introduce a new recommendation algorithm that takes advantage of interactions from users' browsing and diversifies results based on different low-level features. For instance, a user is searching for red car images by using a blue car image he receive from an initial search. Recommending images using a colour feature alone might not obtain red car images, but by using other features, such as edge or texture feature, a red car image may appear in recommendations. The system recommends images that can be used to find more relevant images and explore image collections. Our motivation is to confirm the benefit of these recommendations. We explore two main research questions:

1. Exploring user interactions can lead to effective image recommendations that assist the users to find new aspects of a search task.
2. Diversity in different low-level features of image recommendations helps the user identify different aspects of a task.

In order to investigate our research questions, we have created two different image retrieval interfaces implemented based on Leelanupab and Jose [18]. A baseline system allowing browsing only was compared with our system providing additional image recommendations. On the basis of the diversity of selection, the recommendations result from visually clustering three different low-level features (e.g. Colour Layout, Edge Histogram, and Homogeneous Texture). The systems and their respective performances were evaluated both qualitatively and quantitatively.

4 System Architecture

In this section, we first introduce the data set and according pre-processing steps in Section 4.1. Then, we introduce our recommendation approach in Section 4.2 and introduce the interfaces of our system in Section 4.3.

4.1 Data Pre-Processing

For the purpose of this evaluation we employ the photographic collection created from the CoPhIR[1] collection. The current collection contains 54 million images uploaded to Flickr[2] by real users. In our study, we select a subset of approximately 20000 images taken by unique users for 6 months between 1 October 2005 and 31 March 2006. We selected this time period because it covers the highest density of images from unique users. Images are enriched with textual annotations used for keyword search. The text is derived from titles, descriptions, and tags given by Flickr users. We use the open source retrieval engine Terrier [19] to remove stop words, stem the terms and index the collection. Okapi BM 25 is used to rank retrieval results. Moreover, three MPEG7 image features as mentioned in previous section have been extracted for each image.

4.2 Recommendation Module

As argued in Section 2.1, a graph-based representation of image retrieval results provide a user with an easy access to their image collections. However, Figure 1 illustrates a drawback of this presentation technique. Assuming that a search returns m relevant images, only a small set n of these results can be displayed to maintain the usability of the interface. This results in $(m - n)$ potentially relevant images which are not inspected by the user. We hypothesise that these ignored images can be used as a source to recommend further images. To evaluate this assumption, we opted for an existing browsing system based on Urban et al. [5]. This system has applied the ostensive model of developing information need [6] to trailer search queries to meet the user's current information need. We developed our recommendation algorithm on top of this browsing system.

To create a set of potentially relevant images, let Q_n be a set of n ostensive queries used in a browsing session or, in other words, is a set of all images selected by users during browsing, whereas q_i is an ostensive query at $i - th$ composed of images in a path that a user selects. $ORel(q_i)$ is a ostensive retrieval function that retrieve the top m ranked in the result lists where n is the number of images presented to the user and $(m - n)$ is the number of potentially relevant images collected for clustering. Since we want to provide the recommendation from this $(m - n)$ images, let us define $ORel_{(m-n)}(q_i)$ as the function that return only the $(m - n)$ images. A_i is a set of accumulated images to be clustered at the $i - th$ query within a search session. The algorithm that has been used for creating a set of accumulated images for clustering is outlined in Algorithm 1.

[1] `http://cophir.isti.cnr.it/`
[2] `http://www.flickr.com/`

Algorithm 1 Selecting a set of potentially relevant images to be clustered

Require: $Q_n = \{q_1, q_2, q_3, ..., q_n\}$, a set of ostensive queries q
Require: $q_i = \{img_1, img_2, img_3, ...\}$, a set of image documents in a selected path treated as a query.
$A_0 = \{\}$
for each $q_i \in Q_n$ **do**
$A_i = A_{i-1} \cup ORel_{(m-n)}(q_i)$
end for
return $A_n = \{img_x, img_y, img_z, ...\}$, a set of accumulated images to be clustered at the $n - th$ query

With the aim of assisting a user to explore an image collection, our underlying hypothesis is that a diverse representation of these images could identify more aspects for browsing with maintaining precision. We hence provide the users with additional recommendations diversely selected from all images in A_n accumulated in *all* iterations. In a first step of a diverse representation, we perform a hierarchical agglomerative clustering with the single linkage method to create groups of similar visual content. The Euclidean distance is employed as distance metric for three different features as mentioned. The algorithm generates three dendrograms, which are built by progressively merging the closest cluster until k clusters remain. We assume that each cluster has the potential to reflect different aspects of the user's information need. Recommending representative images from each cluster can hence provide users with a variety of distinct aspects. Thus, we select the medoid[3] as representative of that cluster since it is assumed that it could be the best representative of the cluster. To avoid overwhelming the user as suggested by Miller [20], we set $k = 5$ as a maximum of five selected images from each feature. A recommendation list can contain a minimum of five and a maximum of 15 images due to possible intersections amongst these images from different dendrograms. These images are then arranged in a random order to the recommendation list. The diversity of this recommendation list is two-fold: First of all, using images from each cluster results in a more diverse image selection. This diversity is further extended since the clusters are based on different low-level features. Finally, the random order of the images in the list guarantees that all clusters are treated fairly.

4.3 Interface Design

In this section, we illustrate graphical interfaces of our recommendation system. The system is composed of two main components: Browsing Interface (2) and Slideshow Window (3). The Browsing Interface (2) can be divided into two main panels. The left panel consists of: Full View tab (A), showing a full size visualisation of the image, accompanied with its textual metadata; Presentation tab (B), containing list of relevant lists and Recommendation tab (C).

[3] The image closest to the centroid of the cluster

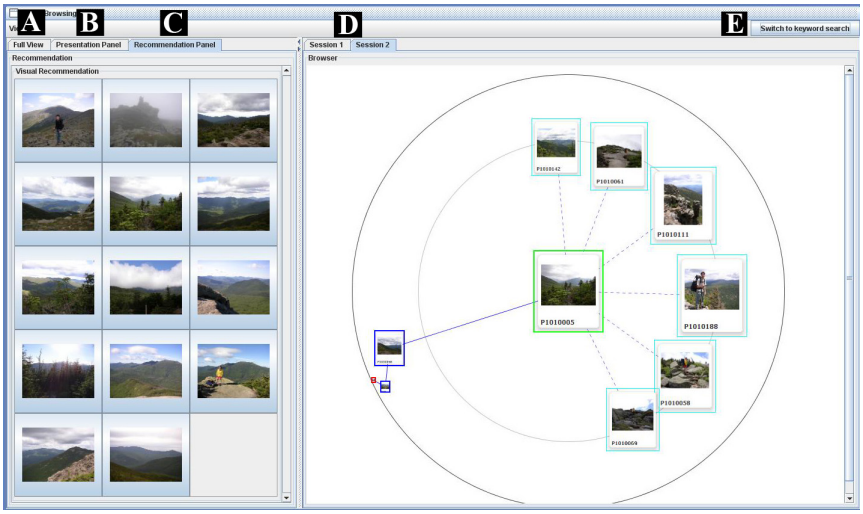

Fig. 2: Browsing Interface

The Recommendation tab (C) *only* appears in the recommendation system. Here, recommended images are presented. Users can click on these images to either start a new browsing session in the right hand side of the interface or to add the images to the presentation tab (B) as relevant images. Moreover, browsing sessions can be initiated by selecting images from a keyword search or other browsing sessions. These independent sessions are visualised as tabs in the Browsing Panel (D). It is hypothesised that each session is created for different aspects based on the user's viewpoint as suggested by Hopfgartner et al. [21]. At the top right of the frame, a Switching Mode button (E) is provided in order to offer the users the option to change search methods between traditional keyword search and adaptive browsing.

In each browsing panel, images are visualised following the approach introduced by Urban et al. [5]. When selecting an image, the system computes a set of six related images which are visualised as leaves in a graph, where the initial image is the node of these leaves. The more relevant an image is to a given query, the larger the image is in the graph. When a user clicks on any of these candidate images, the system will centre on the according image and return a new set of similar images. This action will result in an image path that depicts the way that users have browsed through the collection. In each step, the nodes of this path will be used as a search query to return new query candidates. Following Campbell [6], we consider later steps in the path as being closer to the users' current information need. Hence, the adapted search query will give a higher weighting to more recent images.

In Figure 3, a screenshot of an active Presentation Tab (B) is shown. In our evaluation, users are expected to select relevant images for a given search scenario and store these images in this presentation panel. The interface allows users to simply drag according images from the browsing panel and drop it to

Fig. 3: Slideshow Window

this presentation panel, or to select images from the recommendation panel. The users can modify the presentation by inserting, updating and deleting images in the presentation panel. A click on the play button (1) will start an animated slideshow in a slideshow window (2). In this window, the users can move forward or backward through an size-increased presentation of each image. Moreover, they can trigger an automated slideshow, where each image will be displayed for one second, followed by the successive image.

5 Experimental Methodology

In order to address the introduced hypotheses, we performed a user study. The settings of this user-oriented evaluation scheme is described in this section.

5.1 Experimental Design

We adopted a variance of the Graeco-Latin Square design where participants of our user study were asked to carry out four different search tasks using our two interfaces. Both the order of the tasks and the order of the systems were varied to avoid learning effects which could effect the outcome of the study. Each participant was given ten minutes of training on each system with a different training task for each system. For each task, the participants had a maximum of twenty minutes. With the aim of evaluation, we investigated the nature of task exploration using six measures: (1) user perception of search experience; (2) the number of clicks performed for browsing; (3) the number of textual queries executed; (4) the number of sessions created; (5) the number of image results found; and (6) the distribution of recommendation sources from different low-level features.

The users' interactions with the system were logged and they were asked to fill out a number of questionnaires. The experiment started with an entry questionnaire, where users were asked to provide personal details and to rate their experience of image retrieval. After each search task, we asked them to fill out a post-task questionnaire, aimed at understanding their opinion about the task and the system used for that task. Finally, an exit questionnaire was provided where the participants were asked to compare the two interfaces.

5.2 Participants

24 participants took part in the user study. The group consisted of 16 males and 8 females with an average age of 29 years and different professions. Before being introduced to the experimental tasks and systems, the participants were asked to fill out an entry questionnaire to provide a background in multimedia search. The questionnaire revealed that they have a high experience in dealing with multimedia. All of the subjects create images in digital format and own their own private photo collections. The most common approach to organise these images is to store them under a hierarchical folder structure, e.g. images taken in 2009 are stored in sub directories of a folder called "2009". Sub directory are self-explanatory, e.g. "Holiday in Greece" or "Bob's Birthday Party'.

Most participants stated that they rely on search engines such as Google or Yahoo to search for images online. The photo sharing portal Flickr was named often as well. Using these text query based retrieval services was generally considered to be easy and satisfactory. One hence noticed a different interaction behaviour for different kind of images. While the participants preferred to browse their own images, they feel confident searching for other people's pictures by providing search queries. Asked for the features of an "ideal photo management system", the most desired features were to sort pictures based on the date or location they were taken, or based on contextual information such as events. Moreover, the participants stated that they would like to have a feature which retrieves images with a similar visual appearance.

5.3 Search Tasks

We created four simulated work task situations as suggested by Borlund [22] to define a context for participants to help them better understand the task and arouse their information needs. All tasks asked for different aspects of a search topic and provided some examples. In all tasks, the participants were asked to collect images for creating a slideshow presentation. For example, in Task 1 participants were asked to find different aspects of wild living creatures. The simulated situation was "Imagine you are a graphic designer whose task is to prepare a graphic leaflet with images relating to various subjects of the Wildlife Conservation (WLC). You will give a short presentation of these materials to an educated audience on this subject. The presentation is aimed at raising general awareness about endangered species and preservation of their habitats. You want to create a short presentation about a variety of wild living creatures."

The task were...?	
clear	unclear
easy	difficult
simple	complex
familiar	unfamiliar

The search was...?	
relaxing	stressful
interesting	boring
restful	tiring
easy	difficult

While using a system, you felt...?	
in control	not in control
comfortable	uncomfortable
confident	unconfident

The retrieved image set was...?	
relevant	not relevant
appropriate	inappropriate
complete	incomplete
expected	surprising

The system was...?	
wonderful	terrible
satisfying	frustrating
easy	difficult
effective	ineffective
flexible	rigid
reliable	unreliable

Table 1: 21 Semantic Differentials

The remainder of tasks from 2 to 4, used in the evaluation, entitled "Find different aspects of vehicles", "Find different aspects of natural water", and "Find different aspects of open scenery" respectively.

6 Results

6.1 User Perception

On completion of each task provided, participants were asked to describe various aspects of their experience of using each system in post-search questionnaires, by rating the performance of the system on a set of 21 semantic differentials on Five-Point Likert scales. 4 of these differentials focused on the task they had just performed; 4 focused on the search they had just carried out; 3 focused on their feeling in interaction with the system during the search; 4 focused on the set of images retrieved; and 6 focused on the system itself (See Table 1).

In this evaluation, we were interested in feedback on the user satisfaction with the system's features and responses, and the quality of images retrieved from searches and recommendations. For the semantic differentials related to the task performed, the participants states that the tasks provided were clear, roughly simple, and familiar. However, having analysed the questionnaires by one-way ANOVA, we found that there are significant differences ($p < 0.05$) between the level of task difficulty. It disclosed that Task 2 was the most difficult task followed by Task 3 and 4 whereas Task 1 was the easiest. For other differentials related to our systems, we did not find any difference between the baseline and recommendation systems since the participants felt that they both were effective for solving the task, as they helped them to explore the collection, to find relevant images, and to focus their search. Some questions indicated that the selected images matched what they had in mind before starting the search task and that browsing the collection made it easy to find these images. They stated, however, that the idea of the type of images they were searching for changed

Which system...	*A*	*B*	=
did you find best overall?	3	13	8
did you find easier to learn to use?	3	9	12
did you find easier to use?	6	7	11
did you prefer?	3	15	6
changed your perception of the task?	1	14	9
did you find more effective?	3	13	8
Percentage	13.2%	49.3%	37.5%

Table 2: User Perception of both systems

while performing the tasks. Comments were: "I almost never changed my query word and yet reached many different pictures. So I think the browsing system works well.", "I found browsing quite efficient, as new aspects or ideas came up in terms of different images." or "I preferred browsing a lot rather than keyword searching since browsing helped me in finding more images without posing new queries.'

In addition to the semantic differentials, the post-recommendation-system-questionnaire contained questions aimed at evaluating the quality of the provided recommendations. The averaged answers indicate that they found the recommendations very useful, since the recommendations effectively supported them in their search task. Besides, they asserted that the recommendations gave some more new ideas about how to formulate search queries and that the recommendations helped them to find more relevant images. Some quotations: "Recommendations [...] were quite related to images I searched for", it "revealed images that otherwise would not appear" and "the recommendations were easy to manage, they appeared automatically". Other participants, however, said that "sometimes recommendations drew my attention from browsing".

At the end of the user study, we asked all 24 users to evaluate both systems based on various questions. Table 2 shows the users' preferences for each of the questions. A represents the baseline system and B stands for the recommendation system. The last column represents a neutral perception about the systems of the users. Nearly 50% of all participants selected the recommendation system B as the best performing system, since it was considered being more effective and supportive to find new aspects of the task. Even though it provided an additional feature, the participants did not find it more difficult to use the system.

Our analysis of the questionnaires suggest that the participants had more positive perceptions on the recommendation system, which indicates the success of our recommendation approach. In a next step, we analysed the resulting log files of their interactions with the interfaces in order to compare the performance of the two interfaces.

6.2 Logfile Analysis

Agichtein et al. [23] argue that analysing the users' behaviour while using the system can be a valuable source for improving retrieval results. Hence, we assume

	# browses		*# queries*		*# sessions*		*# results*	
Task	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*
T1	11.4(4.3)	19.1(8.3)	16.2(4.4)	12.6(9.8)	14.1(3.4)	13.4(2.8)	19.0(12.3)	19.6(21.1)
T2	11.5(2.1)	13.6(4.3)	22.4(17.6)	26.3(9.2)	18.5(4.2)	17.7(5.4)	13.5(14.8)	10.0(6.9)
T3	22.3(8.0)	13.9(5.1)	15.8(4.6)	15.0(12.4)	13.8(3.1)	15.3(4.6)	11.9(7.5)	13.9(8.9)
T4	14.3(6.7)	19.7(6.4)	10.9(8.0)	9.4(5.4)	12.8(3.9)	14.1(5.1)	18.8(8.4)	18.5(6.6)
Avg	14.9(5.8)	16.6(4.8)	16.3(10.1)	15.8(9.4)	14.8(3.7)	15.1(4.5)	15.8(11.5)	15.5(12.8)

Table 3: User Interaction Statistics (Mean and SD)

that the users' behaviour patterns, captured in the log files, can be a strong indicator of the efficiency of the two interface approaches. Assuming that behaviour patterns are directly influenced by the features of the graphical interface, we expect to identify different patterns for our two interfaces. In the baseline system, users enter search queries and need to perform similar actions on retrieved relevant and non-relevant results; Users will click on the result, browse through the image collection and/or drag and drop the result. The recommendation system, however, automatically updates recommendations. Assuming that these recommendations are relevant to the users' information need, they will adopt their interaction strategy accordingly, resulting in a different behaviour pattern with respect to the results.

Table 3 shows a mean of four other measures of exploration, illustrating the user's interaction with the baseline system (A) and the recommendation system (B) over all four tasks T1 – T4. The first column denoted "# browses" lists the number of clicks user performed for browsing using the different interface. The second column denoted "# queries" shows the number of textual queries executed on search. The next column denoted "# sessions" shows the number of sessions created for different aspects. The last column denoted "# results" depicts the total number of images added to the presentation panel.

	Task	*Total*	*Colour*	*Edge*	*Texture*
Sessions	*T1*	18.1%	6.8%	9.3%	6.2%
	T2	14.6%	3.7%	5.7%	13.3%
	T3	18.6%	7.1%	6.6%	8.7%
	T4	15.7%	7.6%	7.6%	5.9%
	Avg	16.8%	6.3%	7.3%	8.5%
Results	*T1*	33.2%	15.8%	12.8%	14.0%
	T2	6.7%	1.6%	3.3%	2.5%
	T3	43.7%	16.2%	17.4%	19.8%
	T4	23.4%	8.6%	10.8%	7.7%
	Avg	26.8%	10.6%	11.1%	11.0%

Table 4: Number of recommended images (in percentage) exploited in a recommendation system

In Table 4, we show the number of recommended images used to create sessions and selected to presentations (in percentage) in the recommendation

system (B) as our last measure. Moreover, the table shows which low-level feature was used to retrieve the recommended image. The abbreviations stand for Colour Layout (CL), Edge Histogram (EH) and Homogeneous Texture (HT), respectively. Recommended images can accrue from the union of different low-level features. The total number of recommendations in the table is hence smaller than the sum of the presented features.

Moreover, we were interested in analysing how the participants interacted with both systems of various time points during their search sessions. Figures 4 and 5 show the numbers of created sessions and the number of images that were dropped to the presentation panel using both systems, respectively. The two figures reveal an interesting search pattern. In the first ten minutes of the search session, the participants create more sessions in the recommendation system B, but at the same time dropped less images to the presentation panel. After 15 minutes, however, this pattern changes towards creating more sessions using the baseline system A and dropping more relevant images using the recommendation system B. At the end of the search session, the pattern reverses again.

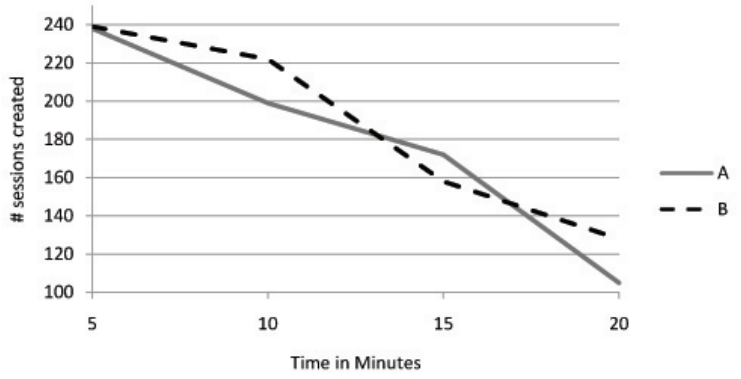

Fig. 4: Sessions per minute

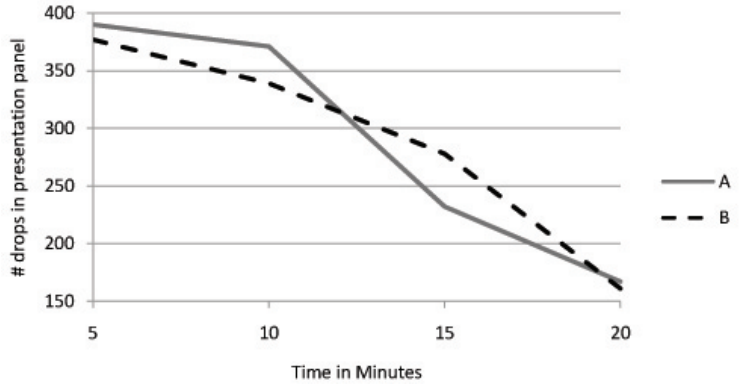

Fig. 5: Results per minute

7 Discussion

The aim of our first research question was analysing whether interpreting user interactions can lead to an effective image recommendation which help users in finding new aspects of a search task. The analysis of the questionnaires revealed that the recommendations were well accepted by our participants. This indicates that the recommendations contained images that were relevant for the given task and hence were of good quality.

An ANOVA analysis of the results did not reveal any significant differences between the number of browses, queries, sessions, or results for the two systems. Nevertheless, the results suggested that diverse recommendations can improve an effectiveness of image browsing systems. As Table 3 shows, the recommendation system B outperforms the baseline system A in terms of average number of clicks for browsing and the number of textual queries executed. This suggests that System B assists the users to rely more on browsing and less on search queries. The results suggest that Task 1 appears to match our recommendation approach the most since users found more results with the less amount of effort in formulating queries. Furthermore, the result shows that the number of new sessions created is higher in two out of four cases, Task 3 and 4. The questionnaires reveal that Task 2 was perceived as the most difficult task followed by Task 3, 4, and 1. The questionnaires correspond to the total number of results for Task 2 that is lower than for the other tasks. One of the main problems in Task 2 was that the participants found it difficult to formulate an initial search query that would retrieve useful results which they could interact with. Another reason supporting this issue is the level of specification for given topics due to the nature of the collection. Task 2 may be the "narrowest" in comparison to Task 1, 3, and 4. To explain this, we analysed the level of agreement between the set of results, assuming that there will be less agreement amongst users for broader tasks, which require a greater extent of interpretation. For Task 2, 38.6% of unique results were selected by two or more users. For Task 1, 3, and 4, two or more users selected 29.0%, 29.9%, and 27.6% respectively. The greater number of agreement amongst users in Task 2 is consistent with Task 2 being the most specific task. Importantly, the average number of created sessions supports the benefit of our recommendation approach. The users created more sessions, suggesting they found more new aspects related to search tasks.

Table 4 illustrates the number of recommended images to create new sessions and be added to a result list. As can be seen, roughly every sixth created aspect was based on an recommended image. Moreover, it shows that almost one quarter of all images that were added to the presentation panel came from the recommendation panel. The participants often relied on the provided recommendations. Figures 4 and 5 show different interaction patterns between system A and B. In contrast to the user interaction with System A, the participants created new search sessions using System B and then selected relevant results after 15 minutes. A possible explanation for this behaviour is that the provided recommendations in System B provided them with image examples that they then used as a starting point to create new sessions. This would again support

our hypothesis that the provided recommendations were useful for identifying new aspects of the given search topic. We hence conclude that the recommendations, which were created by exploiting user interactions, were useful to identify new aspects.

The second research question was that the diverse representation of image recommendations helps the users in identifying new aspects of a topic. Table 4 shows that users did not prefer any specific result lists, since they relied equally on recommendations coming from different low-level retrieval lists. This suggests that the diverse presentation of the recommended images relieved the participants from relying on the results from one low-level feature only. We therefore conclude that a diversity is a useful means to present recommendations to the users.

8 Conclusion

In this paper, we have presented a new image recommendation approach based on exploiting users interactions with a graph-based image retrieval system. This approach has the potential to allow users to explore a data collection to a greater extent. We employed a hierarchical clustering technique to identify various recommendations and proposed them in a diverse order to the users to assist them in their information seeking task.

We evaluated two research questions using a user-centred evaluation methodology. A user study was performed using a subset of a large scale real user image collection. Our research interest was two fold: Firstly, we wanted to evaluate whether exploring user interactions with a retrieval system can be used to provide effective image recommendations. Secondly, we were interested in seeing if a diverse visualisation of the images assists users in finding new aspects of a search task. Both questions were evaluated on both the user perception and an analysis of the log files of the performed user study.

The introduced approach has the potential to assist users in exploring large scale image collections.

9 Acknowledgements

This research was supported by the Royal Thai Government and the European Commission under contract FP6-027122-SALERO. It is the view of the authors but not necessarily the view of the community.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12) (2000) 1349–1380
2. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Readings in information retrieval (1997) 355–364

3. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization. IEEE TVCG. **6**(1) (2000) 24–43
4. Viaud, M.L., Thièvre, J., Goëau, H., Saulnier, A., Buisson, O.: Interactive components for visual exploration of multimedia archives. In: Proc. CIVR '08, New York, NY, USA (2008) 609–616
5. Urban, J., Jose, J.M., van Rijsbergen, C.J.: An adaptive technique for content-based image retrieval. MTAP **31**(1) (2006) 1–28
6. Campbell, I.: Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. Inf. Retr. **2**(1) (2000) 89–114
7. Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W.Y.: Improving web search results using affinity graph. In: Proc. SIGIR '05. (2005) 504–511
8. Song, K., Tian, Y., Gao, W., Huang, T.: Diversifying the image retrieval results. In: Proc. ACM MM '06. (2006) 707–710
9. van Zwol, R., Murdock, V., Garcia Pueyo, L., Ramirez, G.: Diversifying image search with user generated content. In: Proc. MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval, New York, NY, USA, ACM (2008) 67–74
10. Karypis, G.: Evaluation of item-based top-n recommendation algorithms. In: Proc. CIKM '01, New York, NY, USA, ACM (2001) 247–254
11. Spink, A., Greisdorf, H., Bateman, J.: From highly relevant to not relevant: examining different regions of relevance. Inf. Process. Manage. **34**(5) (1998) 599–621
12. Hancock-Beaulieu, M., Walker, S.: An evaluation of automatic query expansion in an online library catalogue. J. Doc. **48**(4) (1992) 406–421
13. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum **37**(2) (2003) 18–28
14. White, R.W., Bilenko, M., Cucerzan, S.: Studying the use of popular destinations to enhance web search interaction. In: Proc. SIGIR '07, New York, NY, USA, ACM (2007) 159–166
15. Craswell, N., Szummer, M.: Random walks on the click graph. In: Proc. SIGIR '07, New York, NY, USA, ACM (2007) 239–246
16. Liu, J., Lai, W., Hua, X.S., Huang, Y., Li, S.: Video search re-ranking via multi-graph propagation. In: Proc. MM '07, New York, NY, USA, ACM (2007) 208–217
17. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool for interactive content-based image retrieval. Circuits and Systems for Video Technology, IEEE Trans. on **8**(5) (1998) 644–655
18. Leelanupab, T., Jose, J.M.: An adaptive browsing-based approach for creating a photographic story. In: Proc. SAMT '08. (2008) 196–197
19. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper (2007)
20. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review **63** (1956) 81–97
21. Hopfgartner, F., Urruty, T., Villa, R., Jose, J.M.: Facet-based Browsing in Video Retrieval: A Simulation-based Evaluation. In: Proc. MMM'09. (01 2009)
22. Borlund, P.: The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research **8**(3) (2003)
23. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proc. SIGIR '06, New York, NY, USA, ACM Press (2006) 19–26