USER QUERY BEHAVIOR: IMPLICATIONS FOR THE DESIGN OF IMAGE RETRIEVAL SYSTEMS

Ya-Xi Chen
Media Informatics, University of Munich
Amalienstr. 17, 80333 Munich, Germany

ABSTRACT

In the last decade, Image Retrieval (IR) has received substantial attention and experienced a rapid growth. While most research has investigated the actual retrieval algorithms, much less is known about the UI design of IR systems. In this paper we present the results of an observational user study, in which we observed the users' natural behavior in image searching and tagging. Based on both quantitative and qualitative analysis, we derive implications for the design of IR systems: Concerning the evaluation of IR systems, besides traditional quantitative evaluation parameters such as recall and precision, other qualitative evaluation strategies focusing on the user's perspective should be integrated to provide a more user-centered evaluation. IR systems should offer more advanced options such as for sorting, filtering or grouping. Beyond simple text extraction, an adaptive understanding of the semantics of the query is also required. Because of the instability, Content-Based Image Retrieval (CBIR) was mainly perceived useful as an inspiration or for discovering unexpected new results. Tags are a strong alternative to content analysis, but in hybrid IR systems both of them could be combined to achieve a better performance. There might be a remarkable tagging gap between the indexer and searcher, and we propose some tagging principles which might help to narrow this gap.

KEYWORDS

User study, interface design, image retrieval, image tagging.

1. INTRODUCTION

In the field of Image Retrieval (IR), most recent work is focusing on the technical problems of Content-Based Image Retrieval (CBIR), such as low-level feature extraction and image understanding, automatic annotation, relevance feedback and machine learning. Much less is known about how searchers actually define their image needs. The users' requirements and retrieval strategies, as well as their variation across different contexts have attracted little attention, which should be well explored to enable the successful design of image retrieval systems and image tagging UIs. While algorithmic performance can be easily characterized and measured, there are few evaluation studies on the perceived performance of existing retrieval systems for the end user. In this paper, we explore several IR systems from the perspective of their end users. Based on an analysis of the searchers' practical behavior, we investigate their relevant query behavior and derive implications for the design of future IR systems.

2. RELATED WORK

In the last decade, IR has received great interest and achieved a rapid growth. A large number of IR systems and technologies have been proposed, some of which are particularly dedicated to CBIR. [Datta et al., 2005] conducted a brief survey on relevant work in CBIR and proposed some guidelines for the design of IR systems. Based on an analysis of publication trends in the CBIR field, they claimed that although CBIR systems receive a high degree of attention, application-oriented issues such as the UI design, visualization of retrieval results and end user evaluation have received less consideration. [Kherfi and Ziou, 2004] addressed IR from a World Wide Web perspective. They surveyed the main features of the most often cited systems and discussed the main issues related to the design and implementation of a Web IR engine. Recently some researchers have realized the value of integrating users more tightly into the IR process. In the work of [Cui and Zhang, 2007], users can choose an area of interest by drawing strokes to provide better relevance feedback. [Kaester et al., 2003] extended IR systems by multiple modalities, such as speech and haptics in order to provide an easier and more intuitive user interface.

2.1 Searching Behavior with Personal Collections

Although personal collections are quite different from public large image sources, we might still be able to transfer some of the insights from studies about personal collections. [Kirk et al., 2006] studied the behavior of searching, browsing and selecting with home collections. Their observations suggested possible methods to support search: content analysis, for example, could help users to cluster and view large collections; users tend to search for particular categories of objects; filtering and grouping along different criteria allows users to see their collections in new ways. [Rodden and Wood, 2003] conducted an evaluation with a photograph management tool named Shoebox, which offers query facilities. They found that text-based indexing does not motivate users efficiently to tag their photos, because they seldom conduct an explicit keyword search within their personal collections. Even within the annotation trial, they found problems, such as name inconsistencies for the same person.

2.2 Image Tagging

One well-known topic in the IR field is automatic image annotation, whereas the natural tagging behavior of users in practical situations is a much less studied field. [Enser, 1993] conducted the first analysis based on real requests submitted to a picture archive. The study revealed that most of the requests were querying about a person, object or event and implied the requirement for a browsing functionality in some cases. From the perspective of an experienced art librarian, [Layne, 1994] stated that the four most important types of image attributes for indexing and retrieval are biographical, subject, exemplification and relationship attributes.

Some studies also showed that there is a relation between retrieval behavior and the retrieval task. [Efthimiadis and Fidel, 2000] found that the nature of the retrieval task may affect searching behavior. They divided these tasks into three groups: Data Pole in which images are used as a source of information, Object Pole in which images are needed as objects, and Pole In-Between. The authors then summarized the characteristics of searching behavior in different task in detail. There was no evidence that color, shape and texture are important for Object Pole. Furthermore, they state that for the performance evaluation of IR systems, precision and recall (which are generally used as major evaluation criteria) have many shortcomings and might not be adequate.

There are abundant literatures in the information science community on tag disagreement between searchers and authors and most of them studied the annotations of text objects. In the specific application domain of newspaper image archives, [Markkula and Sormunen, 2000] found a big difference between the archivists' tags and the keywords used by searchers. Most of the indexes are related to the photo's context while 69% of the searcher requests are of a more visual nature. There are more recent works on analyses of collaborative tags from online communities [Marlow et al., 2006] [Scott et al., 2006].

3. MOTIVATION

In contrast to the mostly task-based existing surveys, we wanted to conduct a behavior-driven study, which we felt was more adequate. In the field of Human-Computer Interaction (HCI), it has become standard procedure to conduct interviews with intended end users beforehand and observe their natural behaviors in order to establish meaningful and well-founded design principles. In this paper, we follow the same approach and use face-to-face interviews as well as user observation to study the users' behavior. We believe that this method has advantages over task-driven approaches in CBIR, because it puts the focus on the performance and benefits perceived by the user, rather than technical performance measures.

Concerning different user types, for example normal users and photographic specialist, there might be a noticeable difference in searching behavior and requirements for the UI, which are differentiated by their experience with images. Members of the first group normally use general online search engines, while the other group mostly uses commercial databases or archives. The press and broadcasting business form a specific domain, in which images originate from commercial databases and archivists are in charge of image maintenance and indexing. Existing studies [Markkula and Sormunen, 2000] and [Ornager, 1995] describe the distinct attributes of indexing and searching behavior in this field. We have confirmed these insights in a preliminary study with Signum (http://www.signumbt.com/), one of our research partners, about their commercial IR system, which targets broadcasters in Germany. Our discussions confirmed that the situation in broadcasting is quite different from the general public. The archivist, more commonly named the editor, receives images with internationally standardized IPTC (International Press Telecommunications Council) data. The archivist can correct and modify all metadata. The searcher thereby benefits from a high reliability of the metadata, but has relatively low flexibility for searching. Although searchers are the end-users of the IR system, they have no right to make any changes of the metadata. Content-based search is not yet considered, because there is a strict and well-established definition of indexes and the potential and capabilities of content-based search are still largely unknown.

Since broadcasting and press is such a special application field and differs drastically from other online search engines users, we decided to exclude the corresponding user type and mainly focus on the normal users. In this paper, we explore their behavior regarding searching and tagging with some representative online IR systems. Furthermore, we wish to study the problem what we call the tagging gap between the image indexer and the searcher and specifically investigate from the aspect of the searcher's visual perception of unfamiliar images. Based on these observations, we identify possible problems and solutions to enhance image searching, which leads to implications for the design of IR systems. The following critical issues need to be explored:

Can fully automatic technologies, such as color-, sketch- or example- based search be helpful?

How do users assess the relevance of images, and which attributes are used in the final decision process?

Is there any general tagging principle? How can the tagging gap between indexer and searcher be bridged, in order to improve the searching efficiency?

4. USER STUDY

In order to explore these issues and thereby lay the foundation for understanding and enhancing IR systems, we conducted an exploratory study with user behavior of image searching and tagging. We recruited 14 participants from University of Munich. The subjects were all right handed, 6 female and 8 male with a mean age of 26.3 years. All participants were regular users of computers and web search engines. According to their experience with photography, they were divided into two trial groups. Group 1 (7 participants) stated that they were advanced hobby photographers and therefore had a relatively high (yet no professional) experience with photos. Group 2 (also 7 participants) had no special experience with photography.

4.1 Settings and Procedure

During the interview, the participants were equipped with a PC, keyboard and mouse. During the user study they could freely browse the designate websites.

The study was consisted of a pre-questionnaire, semi-controlled interviews and a post-questionnaire. First all participants filled out a questionnaire about their experience and background. The following interviews were consisted of two sessions concerning image searching and tagging behavior respectively. At the beginning of each session, each participant was given an explanation of the tasks they would be involved in. In the interview, we observed the participants' behavior which was also video recorded for later analysis. The Think-Aloud protocol was applied, which helped the participants to express their strategies in a more natural and flexible way.

Although there are already abundant IR researches, most of them still stay in the lab status and not many IR systems aim at public usage. In order to understand the impact of existing IR systems, in the first session of interview, we chose four representative IR systems as the experimental platforms concerning their popularity and different functionalities. This allowed us to examine some aspects of current technologies and make suggestions for future development. Engine 1 is Google images (http://images.google.com/), which came out as the dominating one from our pre-questionnaire. Since it is one of the most popular engines, we do not introduce its functionalities here. Engine 2 is stockxpert (http://www.stockxpert.com/, see Figure 1(a)), which offers more search options, such as filtering by category, color and resolution. Engine 3 is ALIPR (http:// www.alipr.com, see Figure 1(b)), where users can conduct an example-based search and get more images by clicking a 'related' or 'similar' button. Engine 4 is Retrievr (http://labs.systemone.at/retrievr/, see Figure 1(c)), which is a sketch-based searching engine.

In the first session, all subjects were introduced to the main functionalities of each system, and then they were required to come up one common query and conduct it in all the engines, which could reflect their most natural requirement and behavior. Since subjects conducted different tasks, the time elapsed was not recorded in this section. After trials of four systems, they completed a post-questionnaire exploring functionalities preferred in each system.

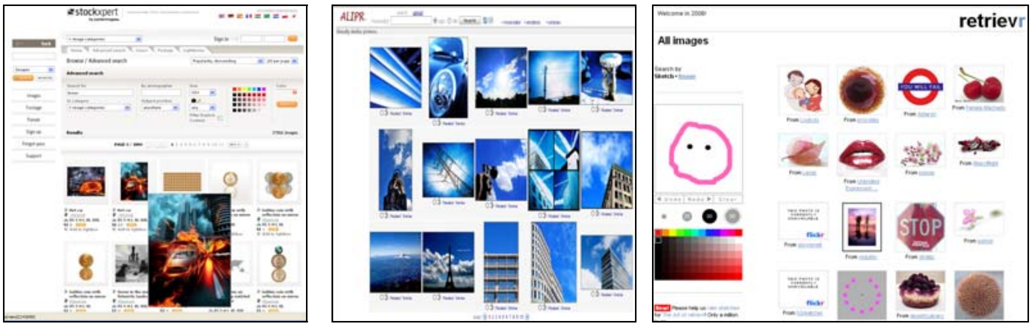

Figure 1. Interfaces of three selected search engines. (a) stockxpert. (b) ALIPR. (c) Retrievr.

The second session of interview focuses on the tagging behavior. The subjects were asked to tag the first 23 photos returned by Flickr (http://www.flickr.com.) with the keyword 'trip'. They tagged these photos freely with keyboard and there was no restriction of tag quantity or length.

4.2 Results

We analyzed the questionnaire, the answers and the video recorded during the interview. The following results were discovered:

4.2.1 General Experience

According to the questionnaire, users in group 1 more frequently manipulated their own photo collections. In particular, 63.2% of group 1 owns a collection containing more than 5000 photos while only 22% of group 2 does. Group 1 add new photos to their own collections much more often than group 2 (4.8 and 1.71 times per week respectively). When asking about the use of online search engines, we found that Google images dominates in this domain because of its large data source, its general popularity and its ease of use. For the relevant usage frequency per week, the average answer was higher in group 1 (4.2) than in group 2 (1.6). Regarding the motivation for using an online search engine, there was no big difference and the top three motivations were illustration or inspiration for work, amusement and obtaining knowledge. Participants were required to evaluate their most often used search engine. Google images received the same score of 3.82 (5 for perfect) in both groups, which means that it is acceptable but might benefit from improvement, such as more search options, semantic analysis, content analysis for untagged photos or an adjustable interactive UI with a good overview of retrieval results. In order to observe the actual searching behavior, each participant was asked to conduct a trial search and record all the operations. All participants claimed that they changed keywords when nothing was found in the first few pages. Interestingly, the time to complete the search varied from a few seconds to 2 hours without any direct correlation between the time elapsed and the user's consent degree.

4.2.2 Image Searching Behavior

In combination with the video analysis, we compare the user searching behavior with the four selected engines. Table 1 shows the average scores of the main functionalities for each engine (5 = perfect). Engine 1 (Google images) performed remarkably better regarding the perceived abundant sources. Subjects scored keyword-based search as the most important functionality and also expressed a strong requirement for advanced search. Although engine 1 offers some of this functionality, in the practical trials no one actually used it. Concerning the advanced search, engine 2 was scored better than engine 1, because it offers more options, such as search by category, predominant color, resolution and type. With a decent filtering functionality, the user can express this requirement better and it also helps narrowing down the search area.

Table 1. Average scores of main functionalities for each engine

Functionality \ Engine		1	2	3	4
Source abundance		4.6	2.4	1.4	1.8
Search option	Keyword-based	4.9	4.5	3.7	\
	Advanced search	3.6	4.3	\	\
Retrieval result	Relevance	3.6	4.6	3.5	1.2
	Visualization	3.5	4.8	3.2	2.6
Average score		4.8	4.2	2.9	1.9

Regarding the relevance of retrieval results, engine 2 and 3 performed better or equal to engine 1. One reason might be that all the images were uploaded by users who also gave a meaningful title or tags to them. Therefore they worked better with keyword-based search. The ‘related’ and ‘similar’ functionality of engine 3 was appreciated. One interesting fact about engine 2 is that, although it offers different sorting criteria for the retrieval results, no one actually used other settings, as long as the default setting (by relevance) was satisfied. For the visualization of results, engine 2 was rated better because it offers additional functionality for browsing the results, such as a preview on mouse-over and an adjustable number of images per page. Engine 4 was rated lowest almost in each aspect because of retrieval errors, unexpected results from the sketch-based search, and a generally suboptimal interpretation of the sketch for the proper search results.

4.2.3 Image Tagging Behavior

In the second session of interview, the participants were asked to tag some unfamiliar photos. To measure the agreement degree, tags of each subject were compared with the popular tags among all the subjects, then compared with those given by the original indexer. The two distinct values (see table 2) implicate that there is big understanding difference between the indexer and the searcher, while among searchers the agreement is relatively higher. The subjects who are more experienced with photos tend to spend more time with tagging and got better tag quality. For example, experienced user 9 and user 10 spent much more time on tagging and consequently performed best. Based on the tags and the video analysis we also found out some interesting details about our subjects’ tagging behavior:

Searchers always lack the context and therefore mostly just enumerate all visible objects in a photo. The visual importance of objects is proportional to their size, shape or color, i.e. their visual weight. With traceable hints, subjects may guess the event, feeling, status, location, and time of a photo.

For photos with a special effect or specific object, tag quality is determined by pre-knowledge. For example, a photo with multiple exposures was only recognized by one subject. Another subject noticed the special effect but did not know how to express it.

Some linguistic problems with free tagging were discovered, such as synonymy (inconsistencies within the same object) and ambiguity (one term with several meanings). For example, all subjects noticed the same object but came up with different tags, such as (girl / woman / female) or (collage / collection / composed image).

Subjects were influenced by the order of the pictures. For half of the subjects, all photos were displayed randomly and they did not notice any similar photos. For the other half, similar photos were ordered near to each other. Most of the subjects noticed this and thought about how to distinguish the similar photos from each other. Tagging behavior is also affected by the user's daily manipulation of photos. When user 9, who achieved the best tag quality, was asked why it took him so long to finish this tagging task, his answer was quite illustrative: Tagging is a serious job for me. When I was tagging, I not only focused on what I thought about this photo, but also how other people interpret it.

Table 2. Quantitative analysis of the tagging task

Participant	Time elapsed (minute)	Tag quantity	Average tags/photo	Tag agreement among participants	Tag agreement with indexer
1	7:00	75	3.3	0.55	0.18
2	4:58	55	2.4	0.40	0.11
3	8:30	64	2.8	0.29	0.15
4	7:30	93	4.0	0.60	0.19
5	12:47	112	4.9	0.48	0.16
6	6:45	80	3.5	0.55	0.19
7	9:00	71	3.1	0.55	0.18
8	4:46	61	2.7	0.48	0.12
9	18:41	159	6.9	0.62	0.22
10	17:20	132	5.7	0.58	0.20
11	8:24	56	2.4	0.42	0.10
12	11:01	92	4.0	0.56	0.18
13	5:53	65	2.8	0.32	0.15
14	4:38	48	2.1	0.40	0.10

5. IMPLICATIONS

During the descriptive video analysis and statistical analysis of pre- and post- questionnaires, as well as the two semi-controlled interviews, we derive a number of observations, which might help IR system design in the aspects of developing novel UI concepts, improving the efficiency of keyword-based searching and image tagging.

5.1 UI Design Concepts

Based on the analysis of computer assisted searching, some design concepts were discovered, which might improve searching.

Search options play a key role in IR systems. Since the images originate from diverse sources with quite different resolution, context, content and license, the system should offer functionalities such as sorting and filtering, which help narrowing down the query and steering the direction of searching. Users are used to keyword-based search which is convenient and easy to use.

For the retrieval results, systems should offer multiple sorting criteria but keep relevance as the default setting. Users should also have the options to further filter or re-group the retrieval results based on certain attributes. Users stick to the standard grid-based display of results. To optimize the visualization of results, systems should facilitate browsing within abundant result sets. For instance, the user may have the option to adjust the quantity of images per page. Since users always make their final selection from pre-selected candidates, a container for possible candidates would be appreciated to relieve temporal memory and avoid a continuous switch in the UI.

Beyond simple text extraction, as provided by Google images, users have a stronger requirement for an adaptive understanding of the semantics of the query. Concerning the content analysis, although [McDonald et al., 2001] claimed that color-, sketch- or example- based search could help users to express their needs in a visual form, in our study there were only two content-related queries. Because of its instability, users would not apply content-based search when they have serious or concrete needs. CBIR was mainly perceived useful as an inspiration or for discovering unexpected new results. User would appreciate if content-based searching could be combined with keyword-based, in order to discover something new while still staying within the expected realm.

Source abundance and response speed are the key issues for any IR system. For the evaluation of an IR system, besides traditional quantitative evaluation parameters such as recall and precision, other qualitative evaluation strategies focusing on the user's perspective should be integrated to provide a more user-centered evaluation.

5.2 Bridge the Tagging Gap

In the second interview, we conducted a photo tagging experiment. Our subjects are usually in the role of the searcher, therefore, the way in which they formulate their queries with textual descriptions, what general tagging principles they follow, and how big the difference of image understanding between indexer and searcher is, will offer implications for both indexer and searcher and thereby enable more successful search queries.

Based on the quantitative and qualitative analysis, a big tagging gap between the indexer and searcher was revealed. The indexer focuses more on the contextual aspect, while for the searcher, the visual objects in the photo are the only available hints. If both sides think more about the other side, more agreements would be achieved for the benefit of both sides. With proper keywords, the searcher can conduct more successful queries and get more relevant results. On the other hand, for the indexer, his/her work piece will become more popular and appear more frequently in the retrieval results, which is one of the intuitive motivations for making contributions by tagging. In order to improve the quality of tags, indexers should take tagging more seriously, take the searcher's thoughts into consideration, and think about possible principles the searcher will apply during querying.

From the system design aspect, Taggers could think about more aspects which are difficult to detect for computer-driven techniques, such as feeling, status or semantic aspects. Tags are a strong alternative to content analysis, but in hybrid IR systems both of them could be combined to achieve a better performance. Since the tagger's behavior will also be influenced by the order and similarity of photos, pre-grouping similar photos could facilitate more efficient tagging. Tag recommendations based on Collaborative Filtering (CF) might also help users to produce more reasonable tags while reducing redundancy and errors.

6. CONCLUSION AND FUTURE WORK

In this paper we conducted behavior-driven user study, in which users' searching behavior and tagging strategy were explored. We derived no big behavior difference between the normal user and the experienced photo user, and the user's behavior is more determined by the practical IR tasks. Based on an analysis of relevant user behavior, we also obtained implications for the IR systems concerning the UI design and improvement of searching and tagging. Although they are preliminary results, we believe they may bring insights for Image Retrieval from the end user's aspect.

In order to obtain more detailed and well-founded guidelines for the system design, the end-users' needs should be further extensively studied. In our future work, we will design formal user study based on the implications we gained from the current work and conduct a large-scale exploration to derive more reliable and in-depth implications for the design and evaluation of IR systems.