

Criteria that have an effect on users while making image relevance judgements

Rahayu A Hamid

School of Computer Science and IT
RMIT University
Victoria 3001 Australia
rahayu.ahamid@student.rmit.edu.au

James A. Thom

School of Computer Science and IT
RMIT University
Victoria 3001 Australia
james.thom@rmit.edu.au

Abstract *This paper reports the result of an exploratory user study investigating criteria that are important to users when judging relevance while performing an image search. Data was collected from 12 participants using questionnaires and screen capture recordings. Users were required to perform three image search tasks which are specific, general and abstract image search and judge relevance based on ten criteria identified from previous studies. Findings show that some criteria were important when making relevance judgements, with topicality, appeal of information and composition being the common criteria across the search tasks. However the order of importance of the criteria differ between the image search tasks.*

Keywords *Information retrieval, user studies involving documents, Web image search, Relevance criteria, Relevance judgement*

1 Introduction

In the last decade, a large number of digital images have been made available and accessible due to the prevalence of digital imaging technology as well as the growth of the Internet. This has contributed to the development of various image retrieval systems, which in turn has made the process of storing and retrieving images much easier. However, research studies that explore users' relevance judgements for image retrieval are not that common.

Although considerable work has been done in identifying criteria users employ when making text retrieval relevance judgements (for example [1, 7, 12, 14]), little is known about what criteria users employ when making image relevance judgements. Therefore, it is important to explore how users select images in order to develop better retrieval systems with more effective user interfaces.

Relevance is a fundamental notion in information retrieval. Over the years, the field of infor-

mation retrieval has gained knowledge about relevance, its factors and effects. However, it has mainly focused on traditional textual information retrieval. Relevance, especially in an image is difficult to define satisfactorily. A relevant image is one judged similar in the context of a query. But it depends on the person judging it and in what context is the image relevant. Furthermore, humans are seldom consistent when making judgements. For that matter, there is no way one can guarantee that a user will be consistent in making judgement, especially given the considerable amount of images presented to them. As Volkmer et al. [17] observe, it is difficult to determine whether an image should be judged as relevant or irrelevant, because with an image, there is always room for ambiguity.

The purpose of this exploratory research is to understand people's behaviour when performing image search. The goal is to identify criteria that might be important to a user when they perform image search. Findings from the study will be used to enhance the image search process in order to minimise the users' effort. The rest of the paper is organised as follows. In Section 2 we present some background on users' relevance criteria. In Section 3 we describe the approach and methods used in the study. Results and analysis of the study are discussed in Section 4. Finally we conclude in Section 5 and suggest future work.

2 Related Work

Relevance is an elusive concept that has long been discussed in information retrieval, yet it is still difficult to define clearly. We discuss relevance in Section 2.1 and previous research regarding relevance in the area of image retrieval in Section 2.2.

2.1 Relevance Judgements Criteria

According to Saracevic [13], relevance is not stated, but implied. Different users want different kinds of information. The same information means different things to different people. The same user wants different kinds of information at different times. The same information can mean different things

to the same people viewing it at different times. Nonetheless, according to Borlund [3], it can be agreed that relevance involves users' perception of information, at a certain point in time, based on their need situation.

Since the 1990s, there has been a surge of studies on relevance judgement made by real users when given real text retrieval tasks. These studies have been conducted to elicit user's relevance judgement criteria [1, 5, 7, 9, 12, 14, 15]. Saracevic [13] identified these studies as "clues to research". The clues represent artifacts of the search process and the criteria used by the subjects are the attributes which describe these clues. These studies investigated a wide range of criteria and came up with different lists and classifications. For example:

- *accuracy, depth and scope, clarity, recency* [1];
- *authority, accessibility, interesting, topicality, quality* [7];
- *presentation quality, currency, reliability, accuracy* [14].

Although each of the studies were widely varied, they made similar observations about the relevance criteria, which can be generalised as follows [13]:

- Searchers use the same criteria but assign different weights to these criteria.
- The importance of these criteria changes with task, progress in task over time, and varies by some categorisation or class of user.
- Criteria may interact with each other.

However, due to differences between text and image information, users' criteria for image relevance may be very different from textual document relevance judgements.

2.2 Studies on Image Relevance Judgements

Research studies that explore users' relevance judgement on image retrieval are not that common. These studies have explored user's relevance by applying specific information needs and then identifying relevance criteria utilised by the users while making relevance inference [5, 8, 9, 15]. The focus is on criteria users apply while thinking of what is or is not relevant and to what degree it may be relevant.

Choi and Rasmussen [5] conducted a study to observe users' relevance criteria and how these criteria change as expressed before and after the search. Thirty eight faculty and graduate students from the Department of History at Carnegie Mellon University, Duquesne University and the University of Pittsburgh that participated were interviewed. They were using the Library of Congress American Memory photo archives. The authors used

nine common criteria which include *topicality, accuracy, time frame, suggestiveness, novelty, completeness, accessibility, appeal of information and technical attributes of images*. These criteria were selected from those mentioned by end-users in previous studies. However, they noted that these were not the only criteria and expected users to mention other criteria as well. Users were interviewed to elicit their information need, and they were also asked to rate the importance of each relevance criteria. Information needs were then used by the researchers to perform searches and retrieve images. After providing the participants with the set of retrieved images for their information need, they were once again asked to rate the importance of each relevance criteria. From the results, they observed a significant change in the importance of some criteria across the information seeking process.

Hung et al. [9] investigated the relevance criteria elicited by ten undergraduate students from Department of Journalism and Media Studies at Rutgers University. The participants' relevance judgements were observed by assigning them three different image searches (specific, general, abstract) using the ACCUNET/AP Photo Archive database system. During the search process, participants were asked to save selected photos for later evaluation. After completing all the three search tasks, participants were then interviewed once and asked to describe the relevance criteria that they had used in selecting the photos. Their study identified several common relevance criteria which were used across all three search tasks with *typicality, emotion* and *aesthetic* as the most frequently mentioned.

In a similar follow-up study involving thirty subjects who have photo-editing experience recruited from newspaper and magazine companies, the searchers applied 32 relevance criteria in the specific search, 26 relevance criteria in the general search, and 23 relevance criteria in the subjective search. After comparing the relevance criteria mentioned in the three searches, 37 types of relevance criteria were identified [8]. This includes the previously identified common criteria [9]. The top ten core criteria were *symbol, composition, consequence, emotion, interest, text, topicality, context, implication* and *facial expression*. His findings also showed that there was a difference in using relevance criteria among the three search tasks.

Sedghi et al. [15] investigated relevance criteria used by twenty six health care professionals when searching for medical images. The participants were asked to specify and perform medical image searches as they would normally do in their daily activities. During the search, they would de-

scribe the relevance criteria that they had applied. They found that *visual relevancy*, *background information* and *image quality* were the three most frequently used relevance criteria. From the interview, they also found that the health care professionals perform image search for different reasons based on their medical image information need. The medical image information need was deemed as the most influential factor in making relevance judgements.

In conclusion, regardless of the experimental setup, users in all these studies apply similar criteria such as *topicality*, *accuracy/visual relevancy*, *textual information* and *technical attributes of images*.

3 Methodology

3.1 Relevance Criteria

During any image search process, it is the user who ultimately decides if the retrieved images are useful or relevant in satisfying their information needs. This decision or assessment of relevance is often influenced by many different criteria. Research by Barry and Schamber [2] suggest that there exist a finite set of criteria which are applied consistently across different types of information users. Although they maybe different in terms of terminology, the criteria seemed to have a common, consistent meaning to users and can also be categorised.

In the relevance criteria study we conducted, we used a subset of the criteria identified in previous image retrieval studies [5, 8]. We selected ten criteria as follows. First, seven criteria (*topicality*, *accuracy*, *suggestiveness*, *completeness*, *appeal of information*, *technical attributes of images* and *textual description*) was selected from [5]. We only selected these criteria because they are applicable for all search tasks and not just historical tasks (*time frame* and *novelty*). Secondly, from [8] we selected six criteria (*topicality*, *composition*, *consequence*, *emotion*, *interest* and *text*). These criteria was selected as they were the core criteria elicited from users when making image relevance judgements for different types of search tasks. Other criteria were not chosen as we did not want to confuse the participants as some criteria can be similar (*symbol*, *context* and *implication*) or too specific (*facial expression*). Of the thirteen criteria selected from the two studies, three criteria were overlapping. Therefore, for our study, we applied these ten relevance criteria and adapted them for the post-session questionnaires as follows:

1. I selected an image if it was relevant to my search topic (*Topicality*) [5, 8].
2. I selected an image if it was an accurate representation of what I was looking for (*Accuracy*) [5].

3. I selected an image if it gave me new ideas or new insights (*Suggestiveness*) [5].
4. I selected an image if it was interesting (*Appeal of information/interest*) [5, 8].
5. I selected an image if it contained the kinds of details I could use to clarify important aspects of my search topic (*Completeness*) [5].
6. Technical attributes (such as colour, perspective, or angle) were important to me in making my selections for this search task (*Technical attributes of images*) [5].
7. I selected an image if it evoked an emotional response in me regarding the search topic (*Emotion*) [8].
8. Text descriptions of the images were useful in making my selections for this search topic (*Textual information*) [5, 8].
9. I selected an image if it contained consequences or implications of the search topic (*Consequence*) [8].
10. I selected an image if it has strong visual impact (*Composition*) [8].

3.2 Experimental Design

We are interested in understanding users' behaviour when performing image search and aim to identify factors that might be important to a user when they perform image search. Therefore, in designing the experiment, three types of image search tasks were created based on Shatford's image analysis [16]. These include specific, general and abstract image search tasks.

- **Specific Task:** You are interested in entering a World Cup 2010 contest. One of the contest conditions is that you have to find 6-8 images that best depicts the 2006 World Cup final match in Germany. Your task is to make a selection from a large collection of images from the World Wide Web and save those that in your opinion would most effectively fulfil the contest's condition.
- **General Task:** As a fashion design student, you are required to create a portfolio showcasing the traditional fabrics of different cultural heritages. Your portfolio will include several different traditional fabrics and one of them is entitled "Timeless Songket". Your task is to make a selection from a large collection of images from the World Wide Web and save 6-8 images that in your opinion would most effectively highlight its uniqueness.

- **Abstract Task:** You and your classmates are preparing a report on the topic ‘Justice and Equality’ and your task is to make a selection from a large collection of images from the World Wide Web and save 6-8 images those that in your opinion would most effectively illustrate the meaning of ‘justice’.

In our exploratory experiment, we made use of a within-subjects experimental design [10]. We recruited 12 people as volunteers to participate in our study as the subjects of the experiments. All of them are either undergraduate or postgraduate students from RMIT who were approached and recruited via posters, electronic forums and face-to-face recruitment after lecture sessions. The participants were met one at a time, each on a separate occasion. The experiment was conducted anonymously, so that responses could not be traced back to individual participants. For each subject, our procedure was as follows:

1. an introductory orientation session;
2. a pre-search questionnaire;
3. a training session to familiarise the subject on how the task was to be performed;
4. a written instruction for the first task;
5. a search session in which the subject perform the first task;
6. a post-session questionnaire about the first task;
7. steps 4 to 6 were repeated for the remaining two tasks;
8. a final exit questionnaire.

Similar to Hung et al. [9], we used a simulated real work task situation [4] to place our participants in a work task scenario. This scenario allows the participants to fashion their information needs in the same manner as they would when performing an actual search session. The participants were instructed to make a selection of images from the World Wide Web, that in their opinion would be most appropriate for the particular task type. In the course of the search, the participants were allowed to submit as many separate queries as they needed. They could also delete any of the images that they had selected if they changed their mind about the suitability of a particular image. In determining the order of tasks which the participants were to perform, we employed a mathematical factorial design with two users for each of the six permutations of the three tasks. This controls for order effects from learning that participants might acquire from one search task to the next.

The experiment used Google Images¹ search engine to perform image search and retrieval. The experiment was carried out over several weeks and during that time, Google Images changed the way they present image search results. These changes include removing the metadata below the image and having it pop up whenever the user put the cursor on it, which creates a mosaic of images and an infinite scrolling page that presents up to 1000 results per “page” [6]. Only three participants performed their search using the old search interface, while the remaining nine participants performed the tasks using the new interface.

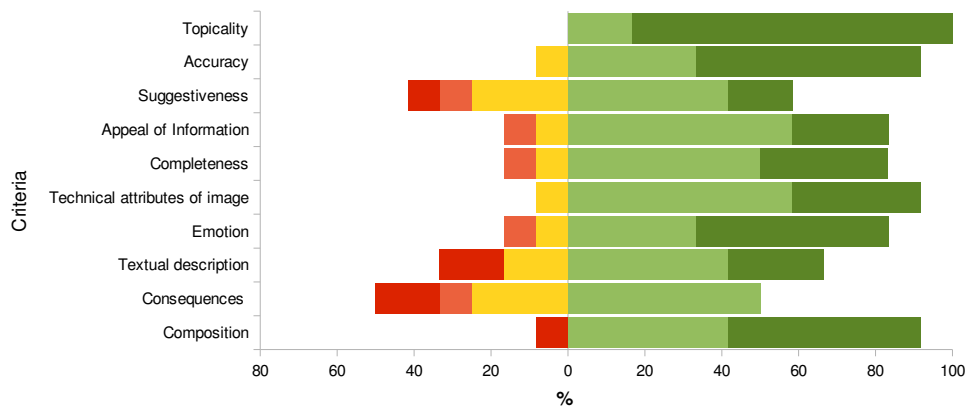
Data for the study was collected through questionnaires and participants’ screen capture recordings. Questionnaires were used as it was found to be more effective for users to communicate their response as compared to interview [11]. According to Kelly et al. [11], although users may express more ideas, many of these ideas are similar; they seem to be repeating it rather than providing new ideas. The pre-search questionnaire was used to collect participant’s prior experience with image search such as frequently used search engines, search frequency, and search expertise. There were two types of relevance criteria questionnaires: the post-session and the exit questionnaire.

The post-session questionnaire has two sets of closed-ended questions. The first set, asks participants to rate their agreement on the reasons they selected images for the search task that they had just performed based on a selected set of relevance criteria while the second set asked to rate other aspects of the task such as topic familiarity, ease of navigation and result satisfaction. The post-session questionnaire allowed us to collect data and have a better understanding of users’ perception of relevance criteria for each task they performed. Finally, open-ended questions were used in the exit questionnaire to collect information regarding the users’ whole search experience and any other issues that may have an effect on how they judge image relevance such as what justifies a relevant image, what makes judging relevance difficult (if any) and how to make it easier.

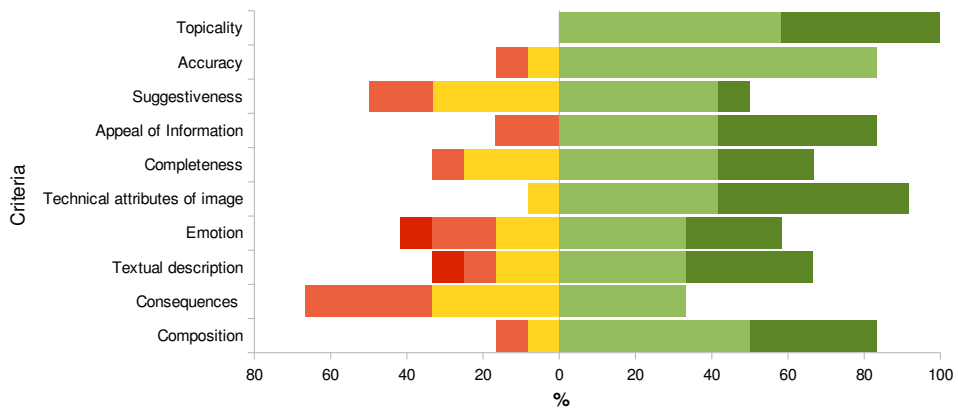
4 Results

Quantitative data from the post-session questionnaires were analyzed using descriptive statistics by assigning numerical values for each agreement rating. This is to determine the average scores of each criteria for relevance judgements and to measure how widely spread the scores were. Another way of showing this information is by calculating the percentage of agreement between users on the cri-

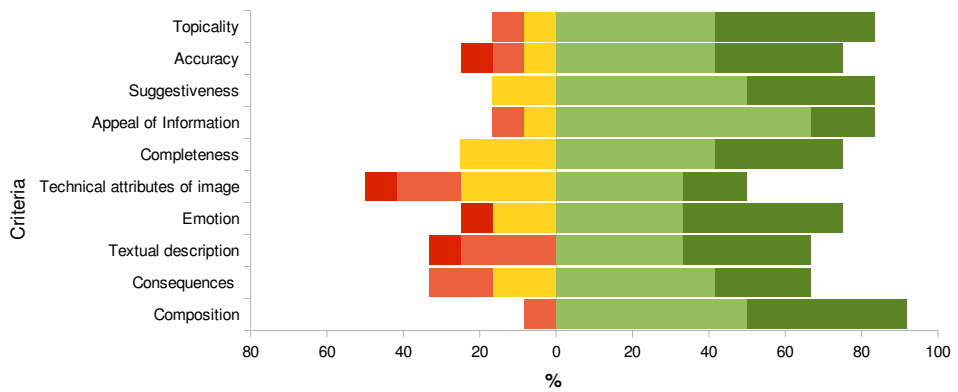
¹<http://www.google.com.au/imghp?hl=en&tab=wi>



a) Specific task



b) General task



c) Abstract task

■ Strongly Disagree
 ■ Disagree
 ■ Neutral / Undecided
 ■ Agree
 ■ Strongly Agree

Figure 1: Percentage of agreement between users on the criteria utilised while making relevance judgements for three different type of task.

Table 1: The mean, standard deviation, number of users' in agreement and Chi-Square's p -value for each relevance criteria across search tasks

Relevance criteria	Statistics	Specific Task	General Task	Abstract Task
Topicality	μ	4.83	4.42	4.17
	σ	0.39	0.51	0.94
	# agree	12	12	10
	p -value	0.0005	0.0005	0.0209
Accuracy	μ	4.5	3.75	3.83
	σ	0.67	0.62	1.27
	# agree	11	10	9
	p -value	0.0039	0.0209	0.0832
Suggestiveness	μ	3.5	3.42	4.17
	σ	1.17	0.9	0.72
	# agree	7	6	10
	p -value	0.5637	1.0000	0.0209
Appeal of information	μ	4	4.08	3.92
	σ	0.85	1.08	0.79
	# agree	10	10	10
	p -value	0.0209	0.0209	0.0209
Completeness	μ	4.08	3.83	4.08
	σ	0.9	0.94	0.79
	# agree	10	8	9
	p -value	0.0209	0.2482	0.0832
Technical attributes of image	μ	4.25	4.42	3.33
	σ	0.62	0.67	1.23
	# agree	11	11	6
	p -value	0.0039	0.0039	1.0000
Emotion	μ	4.25	3.5	4
	σ	0.96	1.31	1.21
	# agree	10	7	9
	p -value	0.0209	0.5637	0.0832
Textual information	μ	3.58	3.75	3.58
	σ	1.38	1.29	1.44
	# agree	8	8	8
	p -value	0.2482	0.2482	0.2482
Consequence	μ	3.08	3	3.75
	σ	1.16	0.85	1.06
	# agree	6	4	8
	p -value	1.0000	0.2482	0.2482
Composition	μ	4.08	4.42	4.25
	σ	1.14	0.9	0.87
	# agree	11	10	11
	p -value	0.0039	0.0209	0.0039

teria that they find were important when searching, and making image relevance judgement (Figure 1).

Although topicality and accuracy is important across all search tasks, it is more important in the specific and general search with twelve users (100%) for both tasks agree that their image selection was based on the topic of search while eleven users (91.6%) and ten users (83.3%) respectively agree they selected images that is the accurate match of the search. In performing a specific search, the user usually has detailed information about what he/she is looking for. Thus, selecting images that matches the information as accurately as possible. Composition was also a common criteria that users find important across all search tasks. Meanwhile, suggestiveness (83.3%) and consequence (66.7%) is more important, while technical attributes of images is the least important criteria in an abstract search with only six users (50%) as compared to specific (91.6%) and general search (91.7%). A reason for this could be that an abstract image can be represented in so many ways and not easily described like an object, place or action.

On the other hand, it is interesting to learn that a few users would select or judge an image as relevant even if the image does not appeal to them and would not consider technical attributes of the image as an important criteria. This shows that relevance is subjective and each user have different ways of making relevance judgements. The image search tasks was performed on a text-based web search engine by submitting textual queries. Therefore, the returned results will be images that are described by that text. However, across all three image search tasks, there are a few users who disagree that textual description is an important criteria while making relevance judgements. The reason could be that the textual description does not always represent the image that the user was looking for and consequently proved that there is ambiguity when using text to describe images.

In order to examine whether there are statistically significance differences in the attitudes of the participants in regards to the importance of certain criteria while making image relevance judgements, a Chi-Square analysis was done. The p -value is calculated based on two categories which are (i) combination of strongly agree and agree, and (ii) combination of strongly disagree, disagree and neutral/undecided. For the purpose of this study, it was decided to adopt a minimum significance level of $p < 0.05$. Table 1 shows the mean value of each relevance criteria for the three search tasks.

From the table, we can see that the importance of relevance criteria varies between type of tasks and those with higher mean values and number of users who are in agreement (agree and strongly

agree) are more widely seen as important when making relevance judgements. This was also shown in the results of the Chi-Square analysis for criteria with a p -value < 0.05 . It was found that *topicality*, *appeal of information* and *composition* are important criteria in determining relevance for all search tasks. In contrast, *textual information* and *consequence* are not considered important to users in determining relevance. *Accuracy* and *technical attributes of image* are important for both specific and general tasks. As for the remaining criteria, *suggestiveness* is more important for an abstract search while *completeness* and *emotion* are for a specific task.

As for the exit questionnaire, users were asked to comment on issues regarding image relevance. When asked, "What factors influenced your decision on whether an image was relevant or not", their responses included: "images related to the topic"; "connection or relationship between image and topic"; "images that reflects the search" and "accurate representation of what I believe the image should look like". These responses were in accord with responses to another question: "In your opinion, what justifies an image as relevant?". The users commented: "relevant images should be which will give exact idea about subject of search even if someone doesnt know about it"; "if it describes the topic theme" and "if it is related with the query and it represents the meaning of that query". Thus, images which satisfy these justifications were considered much more useful or of value and relevant to the users. In addition, users were also asked "Did you find it difficult to decide whether some particular images were relevant or not? If so, what made it difficult?". All participants agreed that at some point, it can be difficult to decide whether an image is relevant or not and some of their reasons were "sometimes if the query is not true", "the returned results were not what I expected from the query entered" and "because I knew little or nothing on the topic besides the keywords to search with". Therefore, although users' judge relevance based on certain set of criteria, there are other factors that could make passing judgement difficult such as knowledge on the search topic or the context in which the search should be performed. Further analysis on users' screen capture recordings might reveal more information on how users judge relevance.

Overall, from the ten selected criteria identified from previous studies [5, 8, 9], not all were important to users when judging image relevance. Our results show that users apply more criteria when judging image relevance for specific task as compared to general and abstract task.

5 Conclusion

In this study, 12 participants were asked to rate their agreement about the relevance criteria that they think is important for searching and selecting specific, general and abstract images. Ten relevance criteria were selected from the criteria set identified from previous studies. The results indicate that users do not find all of the criteria important when making image relevance judgements. Different sets of criteria were used to make relevance judgements for specific, general and abstract images. The three common criteria used were *topicality*, *appeal of information* and *composition*. However, the order of importance for the criteria differ between the type of tasks. This shows that different search tasks affects how users' judge image relevance. Nonetheless, it is acknowledged that since only one task of each type is used, we may be observing individual task effects rather task type effects. Therefore, further research on a bigger sample with multiple tasks of each type is needed to show the effects of relevance criteria on task type and also to perform statistical tests such as factor analysis for significance of results. Further analysis of results and screen capture recordings will also be done, particularly on the process of users searching and selecting relevant images to find out factors that might have an effect when performing image search.

Acknowledgements This research was made possible with a scholarship and study leave granted by the Universiti Tun Hussein Onn Malaysia and has been approved by the RMIT University Ethics Committee (BSEHAPP 08-10 HAMID). We would also like to thank the students who participated in the study and the anonymous reviewers of this paper.

References

- [1] C. Barry. User-defined relevance criteria: An exploratory study. *The American Society For Information Science*, Volume 45, Number 3, pages 149–159, 1994.
- [2] C. L. Barry and L. Schamber. Users criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, Volume 34, Number 2–3, pages 219–236, 1998.
- [3] P. Borlund. The concept of relevance in IR. *The American Society For Information Science and Technology*, Volume 54, Number 10, pages 913–925, 2003.
- [4] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Documentation*, Volume 53, Number 3, pages 225–250, 1997.
- [5] Y. Choi and E. M. Rasmussen. Users' relevance criteria in image retrieval in american history. *Information Processing and Management*, Volume 38, Number 5, pages 695–726, 2002.
- [6] M. Hachman. Google images gets revamped interface, more relevant results. <http://www.pcmag.com/article2/0,2817,2366736,00.asp>, July 2010.
- [7] S. G. Hirsh. Childrens relevance criteria and information seeking on electronic resources. *The American Society For Information Science*, Volume 50, Number 14, pages 1265–1283, 1999.
- [8] T.-Y. Hung. *Search Strategies For Image Retrieval in The Field of Journalism*. Ph.D. thesis, School of Communication, Information and Library Studies, Rutgers University, 2006.
- [9] T.-Y. Hung, C. Zoeller and S. Lyon. Relevance judgments for image retrieval in the field of journalism: A pilot study. In *Digital Libraries: Implementing Strategies and Sharing Experiences*, Volume 3815 of *Lecture Notes in Computer Science*, pages 72–80. Springer Berlin/Heidelberg, 2005.
- [10] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, Volume 3, Number 1–2, pages 1–224, 2009.
- [11] D. Kelly, D. J. Harper and B. Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management*, Volume 44, Number 1, pages 122–141, 2008.
- [12] T. K. Park. The nature of relevance in information retrieval: An empirical study. *Library Quarterly*, Volume 63, Number 3, pages 318–351, 1993.
- [13] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *The American Society For Information Science and Technology*, Volume 58, Number 13, pages 2126–2144, 2007.
- [14] L. Schamber. Users criteria for evaluation in a multimedia environment. In *Proceedings of the 54th Annual Meeting of the American Society for Information Science*, pages 126–133, Washington, DC, October 1991.
- [15] S. Sedghi, M. Sanderson and P. Clough. A study on the relevance criteria for medical images. *Pattern Recognition Letters*, Volume 29, Number 15, pages 2046–2057, 2008.
- [16] S. Shatford. Analyzing the subject of a picture: A theoretical approach. *Cataloging and Classification Quarterly*, Volume 6, Number 3, pages 39–62, 1986.
- [17] T. Volkmer, J. A. Thom and S. M. M. Tahaghoghi. Exploring human judgement of digital imagery. In *ACSC '07: Proceedings of the thirtieth Australasian conference on Computer science*, pages 151–160, Darlinghurst, Australia, January-February 2007.