Evaluating the Effort Involved in Relevance Assessments for Images

Martin Halvey
Interactive and Trustworthy Technologies Group
School of Engineering and Built Environment
Glasgow Caledonian University
martin.halvey@gcu.ac.uk

Robert Villa
Information Retrieval Group
Information School
University of Sheffield
r.villa@sheffield.ac.uk

ABSTRACT

How assessors and end users judge the relevance of images has been studied in information science and information retrieval for a considerable time. The criteria by which assessors' judge relevance has been intensively studied, and there has been a large amount of work which has investigated how relevance judgments for test collections can be more cheaply generated, such as through crowd sourcing. Relatively little work has investigated the process individual assessors go through to judge the relevance of an image. In this paper, we focus on the process by which relevance is judged for images, and in particular, the degree of effort a user must expend to judge relevance for different topics. Results suggest that topic difficulty and how semantic/visual a topic is impact user performance and perceived effort.

Categories and Subject Descriptors
H.3.3 [**Information search and retrieval**]: Search process

Keywords
Relevance; images; effort; judgment; assessment

1. INTRODUCTION

The assessment of relevance is one of the many central issues in Information Retrieval, both as a part of a user's search process [7], and as part of the creation of relevance assessments for test collections [12]. The study of relevance has a long history in information retrieval and information science, there being a considerable number of papers that have defined and modelled relevance [9, 10, 13]. Recent papers, such as the work of Al-Harbi and Smucker [1], have revisited the question of how assessor's judge documents for relevance. In a typical IR test collection, relevance judgments are normally reduced to a list of relevant/not-relevant judgments, with issues surrounding how the judgments were made being removed or ignored. A new interest is growing in filling this gap, investigating more closely how judgments come about. In this paper we investigate one important aspect; what is the degree of effort required to judge relevance: i.e. how much work must an assessor exert to judge the relevance of an information item? This work is a follow up to the study by Villa and Halvey [15], which looked at the effort involved in judging the relevance of text documents. Rather than text, which is the focus of much recent work [1, 9, 13], in this paper we investigate the assessment of images. The problem of classifying and assessing images is well-known, e.g. there has been considerable research looking at the criteria individuals use to judge relevance for test collection creation [3, 5]. Similar to the case of text documents, work investigating the effort involved in judging the relevance of images is rare. Work in image retrieval sometimes informally assumes that the effort required to judge images is less when compared to text, reflected in the many image (and video) retrieval systems which take advantage of the fact that users can judge the relevance of a large number of images relatively quickly [13]. Work such as the Extreme Browsing paradigm which uses Rapid Serial Visual Presentation [7] would suggest that users are fast at judging the relevance of images/videos. Past work has also investigated the effect of image size on the ability of users to interpret the image, and found that even small images, as small as 32 by 32 pixels can work well [8, 14]. While these studies have looked at judgment accuracy at different thumbnail sizes, little attempt has been made to estimate the "effort" required, e.g. do users slow down when forced to interpret a smaller image? In this paper, we wish to consider the effort required to judge image relevance in isolation in a formal user study, investigating the following three research questions:

RQ1: Does the size of an image being judged affect the effort and accuracy of the judging process?
RQ2: Does the degree of difficulty of a search topic affect the effort and accuracy of the image judging process?
RQ3: Does the visual or semantic nature of a topic affect the effort and accuracy of the image judging process?

In all cases we are interested in two main responses: the effort required (including perceived effort), and the accuracy of the relevance assessments made. In RQ1 the focus is on image size (e.g. is the effort required to judge small thumbnail style images similar to full sized images?) By varying image size we aim to reflect the fact that searchers are often presented with different sized images at different stages of their search process, from small thumbnails which represent a set of results, to larger images in more detailed views. RQ2 considers difficulty of the underlying image search topic (e.g. are images from "difficult" topics more difficult to judge?). RQ3 considers a second aspect of topic, whether the topic is visually or semantically oriented [6] e.g. are images easier to judge when considering visually oriented topics? Our work differs from recent work in this area as we consider the search topic rather than only document attributes [15] or the search system [2]. We also focus on images rather than text documents, a first step in expanding the types of information investigated by such studies.

2. EXPERIMENTAL DESIGN

2.1 Design

In our experiment we manipulated four independent variables: image size (small, medium, large), relevance level (relevant, not relevant), topic difficulty (easy, medium, difficult, very difficult) and topic visuality (visual, medium, semantic). The latter three variables were based on the topic classifications defined in the ImageCLEF 2007 [4, 5] and allow us to investigate how the

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright © 2014 ACM 978-1-4503-2257-7/14/07…$15.00.
http://dx.doi.org/10.1145/2600428.2600735

relevance judgment task varies by topic type. The study was designed as a relevance judgment task only, with users being presented with a search topic, and then six different images for that topic, sequentially, one by one. The task for the user was to judge the relevance of each image to a topic. As users judged images, the system would record the users' actions, and after a block of judgments the user's perception of topic effort was gathered via a NASA TLX [6]. All combinations of independent variables were presented, with each combination of topic (3 visuality x 4 difficulty) being presented randomly, and then for each topic all combinations of image size and relevance level (3 sizes x 2 relevance levels) were presented randomly as a block.

2.2 Data and Topics

The ImageCLEF 2007 collection is a set of 20,000 images, 60 search topics, and associated relevance judgments. The topics are categorised into a number of different categories, including: easy/hard (topic "difficulty"), semantic/visual (topic "visuality"), and geographic/general [4]. For this evaluation we used the easy/hard and semantic/visual categories. Easy/hard has 4 categories; easy, medium_hard, difficult and very difficult (which here after we refer to easy, medium, difficult and very difficult). The semantic/visual category has 3 categories; visual, medium and semantic. One of the combinations (very_difficult and visual) had no topics so was not used for the evaluation, which leaves us with 11 possible combinations of topic type. Another of the topics (topic 45) contained images that our ethics review process deemed potentially upsetting to some, and so this topic was removed. Each image in this collection has a size of approximately 480 by 320 pixels (depending on orientation). This size was defined as "large" in this study, with medium and small images being 2/3 and 1/3 of the size horizontally and vertically (circa 320 x 240, and 160 x 120 pixels respectively). The smallest size falls roughly half way between Bing and Google image search thumbnails.

2.3 Procedure

The study was implemented online, and was distributed to staff and students at the University of Sheffield, UK as well as via social media. The webpages were made up of a short demographic questionnaire, a set of instructions, followed by eleven topics. The system first displayed the topic, which consisted of a short title and three example images, along with a button which was used to start the display of the images for that topic. The topics were randomly selected from the possible topics available for the given difficulty and visuality combination. Most combinations contained multiple topics, with the exception of easy/semantic, easy/medium visual, and very difficult/medium visual. Six different images were shown to the participant for each topic, the images varied for each combination of size and relevance, for that topic. At the bottom of each image participants could select the relevance of the image (not relevant or relevant), and then click to move on to the next image. Respondents were not able to move to the next image without judging the relevance of each image, but a button did allow the participant to review the topic: they could move between the topic and judging images as many times as they wished by using the "view topic" button. In total each participant made 66 relevance assessments: 11 topics (combinations of difficult and visual categories) x 3 sizes (large, medium, small) x 2 relevance levels (relevant, irrelevant). After making relevance judgments for the 6 images for a topic, a NASA TLX questionnaire would be displayed. Only part one of the questionnaire was utilised, similar to the approach adopted by other IR researchers [2, 15]. Part one of the NASA TLX is composed of six semantic differentials; mental, physical, and temporal demand, performance, effort and frustration; all are rated between 0 and 100, where lower is better (i.e. less effort), the exception being performance where a higher value equates to better perceived performance. After completing this questionnaire the next topic would be displayed, and this process would continue for each of the 11 topics. No payment was made for participation. The study webpage was designed to control the size of the page which would be viewed by the participant, as far as possible. On starting the study a new browser window opened with a fixed width and height, and a simple page design was used to ensure consistency between browsers. If the browser window was resized it would result in a logged event. While control of the pixel size of images could be made, we could not control the zoom level used by users. A range of other events were also tracked, such as page scrolling, and button presses. The dependent variables were: (1) relevance assessment accuracy, (2) judgment time (how long it took the user to make a judgment), (3) topic views (number of times the "view topic" button was pressed), and (4) subjective effort to make a judgment (via NASA TLX).

3. RESULTS

3.1 Participants

In total 110 participants completed the experiments: 58 females, 51 males, with one participant declining to indicate their gender. Participants mean age was 25 (SD=9). The participants were multinational, with 55 participants self-identifying as being from the UK. Most participants rated their English as good, with only one participant rating their English as poor. In total the participants judged 7260 images across all 59 topics, with 4441 unique images judged. As much of the data analysed showed significant differences for Levene's Test, non-parametric statistical tests were used. Friedman tests were used, with pairwise comparisons made using Wilcoxon sign ranked tests (bonferroni adjusted alpha for difficulty = 0.0125, bonferroni adjusted alpha for visual-semantic and size = 0.0167).

3.2 Performance

Performance was compared using accuracy of judgment, true positive rate (TPR) and false positive rate (FPR), following the approach of Smucker and Clarke [11], these are shown in the first 3 columns of Table 1. For accuracy, both difficulty and visual-semantic were found to be significant ($X^2(3)=33.444$, $p<0.001$ and $X^2(2)=110.351$, $p<0.001$ respectively). Pair wise testing found a significant difference between semantic and both the medium and visual levels ($z=-7.126$, $p<0.001$ and $z=-10.050$, $p<0.001$ respectively). A significant difference was also found between the visual and medium levels ($z=-4.129$, $p<0.001$). For difficulty, post-hoc tests found that accuracy differed significantly between the easy level and medium, difficult, and very difficult levels ($z=-2.608$, $p=0.009$; $z=-4.051$, $p<0.001$; and $z=-5.568$, $p<0.001$ respectively). A significant difference was also found between medium and very difficult ($z=-3.760$, $p<0.001$). It was found that size did not significantly affect the judgment accuracy of participants ($X^2(2)=3.715$, $p=0.156$). A significant difference was found between the accuracy of relevance and non-relevant images ($z=-9.722$, $p<0.001$) with participants identifying non-relevant images correctly with higher accuracy.

3.3 Effort

Table 1, last 2 columns, show the objective measures of effort (time taken to judge an image, and number of times the "view topic" button was pressed), split by topic difficulty and visual-semantic. Looking at time first, a significant difference was found for topic difficultly ($X^2(3)=69.112$, $p<0.001$). Pair-wise tests found that time to judge was significantly different between easy topics and those difficult or very difficult ($z=-6.107$, $p<0.001$ and $z=-6.355$, $p<0.001$). Significant differences were also found between medium difficulty and difficult and very difficult topics ($z=-3.718$, $p<0.001$ and $z=-5.250$, $p<0.001$). For view topic counts, difficulty was again found to be significant ($X^2(3)=8.032$, $p=0.045$), with pair wise comparisons finding a significant difference between easy and very difficult ($z=-2.837$, $p=0.005$).

Table 1: Performance for different topic types and image characteristics (TPR = True positive rate, FPR = False positive rate) and measures of objective effort, time and number of view topic clicks, Mean (Std. Deviation).

	Accuracy	TPR	FPR	Time (secs)	View Topic
All	0.872	0.834	0.090	4.97 (7.05)	0.02 (0.14)
Topic difficulty					
Easy	0.908	0.854	0.038	4.60 (4.97)	0.01 (0.11)
Medium	0.883	0.848	0.083	4.79 (7.19)	0.01 (0.12)
Difficult	0.867	0.830	0.096	5.31 (8.65)	0.02 (0.15)
Very difficult	0.811	0.789	0.168	5.29 (6.76)	0.03 (0.18)
Visual-semantic					
Visual	0.932	0.903	0.038	4.88 (9.36)	0.02 (0.13)
Mixed	0.884	0.845	0.078	4.87 (6.49)	0.02 (0.13)
Semantic	0.816	0.771	0.140	5.14 (5.35)	0.02 (0.15)
Image Size					
Small	0.862	0.801	0.078	4.80 (7.51)	0.02 (0.13)
Medium	0.877	0.845	0.091	4.87 (7.26)	0.02 (0.14)
Large	0.877	0.855	0.101	5.24 (6.31)	0.02 (0.14)
Image Relevance					
Relevant	0.834	0.834	--	5.03 (7.41)	0.02 (0.14)
Non-relevant	0.910	--	0.090	4.91 (6.66)	0.02 (0.14)

Looking next at the visual-semantic category and time, a significant difference was again found ($X^2(2)=49.140$, $p<0.001$). Pair wise comparisons found significant differences between semantic and both medium and visual ($z=-5.319$, $p<0.001$ and $z=-7.348$, $p<0.001$). A significant difference was also found between the visual and medium levels ($z=-3.364$, $p=0.001$). The view topic measure was found to be not significant for the visual-semantic topic category. Considering image size, this varied significantly by time ($X^2(2)=76.002$, $p<0.001$). Pair wise comparisons found significant differences between small and large images ($z=-5.690$, $p<0.001$) and medium sized and large ($z=-5.687$, $p<0.001$). The view topic measure was not found to be significant for image size ($X^2(2)=0.683$, $p=0.711$). Considering relevance, no significant interactions were found for either time or view topic ($z=-0.647$, $p=0.518$ and $z=-0.259$, $p=0.796$ respectively).

3.4 Subjective Effort

In addition to the objective measures of effort, a NASA TLX was also used to measure subjective effort (Table 2, values range from 0-100, where lower is better with the exception of performance). Considering subjective effort with regard to topic difficulty, all six NASA TLX scales were found to vary significantly ($p < 0.001$ for all scales). For each of these scales, significant differences were found between easy and all other difficulties (medium hard, difficult, and very difficult). For mental demand, temporal demand, and performance, significant differences were also found between the medium and difficult and medium and very difficult levels. For the visual-semantic topic categorisation all scales had significant differences ($p<0.001$ for all). Pairwise comparisons found significant differences between semantic-visual and medium-visual for all scales. The semantic-medium differential was also found to be significant for mental, performance, effort, and frustration. No other significant differences were found.

Table 2: Subjective effort for task difficulty and visuality, from 0 (low) to 100 (high), lower is better with the exception of performance; Mean (Std. Deviation)

	Task Difficulty		Task Visuality	
Effort	Easy	21.52 (18.45)	Semantic	20.27 (19.45)
	Med	23.85 (20.73)	Medium	18.96 (18.48)
	Diff	24.86 (20.68)	Visual	17.99 (19.14)
	V.Diff	24.70 (19.93)		
Frust-ration	Easy	16.88 (16.48)	Semantic	25.33 (20.31)
	Med	20.09 (20.11)	Medium	22.77 (19.42)
	Diff	20.24 (20.16)	Visual	22.12 (20.20)
	V. Diff	20.21 (19.61)		
Mental	Easy	24.09 (19.75)	Semantic	28.46 (22.18)
	Med	28.02 (23.02)	Medium	27.05 (21.14)
	Diff	28.62 (21.80)	Visual	25.23 (21.52)
	V. Diff	29.41 (22.57)		
Perfor-mance (higher = better)	Easy	77.05 (22.52)	Semantic	71.52 (23.47)
	Med	73.55 (23.48)	Medium	74.50 (22.62)
	Diff	71.33 (23.72)	Visual	75.91 (23.76)
	V. Diff	72.14 (23.31)		
Physical	Easy	14.73 (14.93)	Semantic	15.60 (16.02)
	Med	16.47 (16.96)	Medium	16.14 (16.47)
	Diff	16.08 (16.46)	Visual	15.65 (16.37)
	V. Diff	16.05 (16.85)		
Temp.	Easy	18.74 (16.86)	Semantic	20.68 (18.18)
	Med	20.76 (18.77)	Medium	20.74 (18.27)
	Diff	21.18 (18.03)	Visual	19.11 (17.60)
	V. Diffs	20.48 (18.68)		

4. Discussion

Considering RQ1, which looked at the relationship between image size and both effort and accuracy, accuracy was not significantly affected by image size. However, it was found that image size did have a significant effect on the time taken to judge an image, with larger images taking longer to judge. In some ways this finding is counter intuitive as it might seem that larger images should be easier to judge. However, this finding is in keeping with results for text documents in similar experiments [15]. For image search, this would suggest that smaller images as typically shown in search results do not reduce the accuracy of users in determining relevance, but will have an impact on the length of time it takes a user to judge the relevance of the results. It should be noted, however, that in this experiment images were presented individually, rather than in a list.

Considering RQ2, which looked at topic difficulty, and first considering accuracy, there were significant differences between easy topics and all other categories of difficulty. There was also a significant difference between the most difficult topics and the second easiest topics. The general trend was that as topic difficulty increases accuracy of judgement decreases. In terms of judgement time, as topic difficulty increases so does the time required to make a judgement, with the exception of the difficult and very difficult levels. In terms of subjective effort, for all of the differentials there were significant differences. In all cases the easy topics were significantly different to all others levels. It cannot be said that subjective effort increases with topic difficulty, however, as in some cases difficult was different to medium hard and very difficult, thus there was not a completely linear relationship. But it can be seen that easy topics have lower subjective effort than any other topic category.

Lastly, looking at RQ3 which focuses on the semantic-visual dimension of the topics, it can be seen that more visual topics have higher accuracy than more semantic topics. This higher accuracy is also achieved in less time than semantic topics, although there is not as large a difference as with medium topics. The subjective effort results also indicate that visual topics require less effort to judge in terms of subjective effort, for example it was found that participants believed they had better performance for visual topics, while for semantic topics, the perceived mental workload and effort was greater. Image relevance was also considered to be a factor for this experiment. In terms of accuracy of judgment, the participants were significantly more accurate at judging non-relevance when compared with judging relevant images. This may be an artefact of design, or the qrels, as the non-relevant images were randomly selected from the non-relevant qrels for each topic. Previous work has shown that ambiguity in the document to be judged for text documents can result in lower accuracy [15]. In terms of all other measures no significant differences were found.

5. CONCLUSIONS AND FUTURE WORK

From the results presented here, we can make the following three conclusions: (1) image size does affect the time required to judge an image, with larger images taking more time, but does not affect the accuracy or perceived effort; (2) The degree of topic difficulty affects accuracy, time to judge and effort: the trend is for accuracy to decrease as difficulty increases, time to judge and perceived effort increase as difficulty increases; (3) For image search as topics move from being visual to semantic, accuracy decreases, time to judge and perceived effort increase. There are a number of implications for these findings. First, while no relationship was found between image size and performance, it was found that users did take longer to judge the relevance of smaller images. This would suggest a possible size/time trade-off when judging the relevance of images in search results. Further investigation of how this manifests itself in the judging of image search results is required, to investigate if there are benefits to showing smaller numbers of larger images in search results, in regards to the length of time required by users to find and judge relevant images. The relationship with topic difficulty and topic "visuality" may also have implications for image search result presentation, with images for "easier" ImageCLEF topics being quicker to judge and more likely to be accurately judged. Future work may investigate how different search result presentations could be changed based on the type of topic or query, increasing or reducing the size of search results based on the query type. Simulations and evaluation metrics should take account of both topic difficulty and topic nature when simulating users. In addition as image size does not appear to impact accuracy and perceived effort it may not always be essential to present the largest image possible to users. In future work we aim to consider how the results of this study can be integrated into simulations of the search process. In particular we would like to integrate these results with other studies of effort which have looked at document types [15] and interface design/query cost [2].

6. ACKNOWLEDGEMENTS

This work was funded by the UK Arts and Humanities Research Council (grant AH/L010364/1).

7. REFERENCES

[1] Al-Harbi, A. L. and Smucker, M. D. User Expressions of Relevance Judgment Certainty. HCIR 2013.

[2] Azzopardi, L., Kelly, D. and Brennan, K. How query cost affects search behavior. SIGIR 2013, 23-32.

[3] Burford, B., Briggs, P. and Eakins, J. P. A taxonomy of the image: on the classification of content for image retrieval. Visual Communication, 2, 2 (2003), 123-161.

[4] Grubinger, M. Analysis and evaluation of visual information systems performance. PhD Thesis 2007.

[5] Grubinger, M., Clough, P., Hanbury, A. and Müller, H. Overview of the ImageCLEFphoto 2007 photographic retrieval task. CLEF 2007.

[6] Hart, S. G. and Stavenland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. .*Human Mental Workload.* Elsevier, 1988, 139-183.

[7] Hauptmann, A. G., Lin, W., Yan, R., Yang, J. and Chen, M. Extreme video retrieval: joint maximization of human and computer performance. ACM Multimedia 2006, 385-394.

[8] Hürst, W., Snoek, C. G., Spoel, W. and Tomin, M. Size matters! how thumbnail number, size, and motion influence mobile video retrieval. MMM 2011, 230-240.

[9] Mizzaro, S. Relevance: The whole history. JASIS, 48, 9 (1997), 810-832.

[10] Saracevic, T. RELEVANCE: A review of and a framework for the thinking on the notion in information science. JASIS, 26, 6 (1975), 321-343.

[11] Smucker, M. D. and Clarke, C. L. A. Time-based calibration of effectiveness measures. SIGIR 2012, 95-104.

[12] Sormunen, E. Liberal relevance criteria of TREC -: counting on negligible documents? SIGIR 2002, 324-330.

[13] Spink, A., Greisdorf, H. and Bateman, J. From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management, 34, 5 (1998), 599.

[14] Torralba, A., Fergus, R. and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30, 11 (2008), 1958-1970.

[15] Villa, R. and Halvey, M. Is relevance hard work?: evaluating the effort of making relevant assessments. SIGIR 2013, 765-768.