Video Test Collection with Graded Relevance Assessments

Weng Qiying
University of Strathclyde
Department of Computer and Information Sciences
Glasgow, United Kingdom.
weng-qiying.2014@uni.strath.ac.uk

Martin Halvey
University of Strathclyde
Department of Computer and Information Sciences
Glasgow, United Kingdom.
martin.halvey@strath.ac.uk

Robert Villa
University of Sheffield
Information School
Sheffield, United Kingdom.
r.villa@sheffield.ac.uk

ABSTRACT
Relevance is a complex, but core, concept within the field of Information Retrieval. In order to allow system comparisons the many factors that influence relevance are often discarded to allow abstraction to a single score relating to relevance. This means that a great wealth of information is often discarded. In this paper we outline the creation of a video test collection with graded relevance assessments, to the best of our knowledge the first example of such a test collection for video retrieval. To directly address the shortcoming above we also gathered behavioural and perceptual data from assessors during the assessment process. All of this information along with judgements are available for download. Our intention is to allow other researchers to supplement the judgements to help create an adaptive test collection which contains supplementary information rather than a completely static collection with binary judgements.

CCS Concepts
• Information systems~Test collections • Information systems~Relevance assessment

Keywords
Video; Variable; Graded; Relevance; Assessment; Test Collection.

1. INTRODUCTION
Judgment of relevance is a heavily studied topic within Information Retrieval. Relevance judgment is important to both the search process itself [14], and in the creation of test collections [3; 12]. With regard to the latter, there is a considerable body of work which has investigated the criteria assessors use to judge relevance for the creation of test collections [3; 12]. However the generation of test collections is not without its drawbacks. On one hand, when generating relevance assessments for test collections, the behavior of assessors is not normally considered as important [8], beyond the overall time taken to create a set of relevance judgments [12; 16]. Given the importance of relevance assessment to the information seeking process, the relative lack of research studying assessors is perhaps surprising. On the other hand, often by necessity, relevance assessments are reduced down to a binary decision which ignores the many facets of relevance [11]. There are some exceptions to this. TREC HARD considered multiple levels of relevance in their assessments [1]. Whilst in the area of image retrieval, previous ImageCLEF tasks have considered the variance within tasks [4; 5]. Although these initiatives have gone some of the way to addressing this shortcoming for text and image retrieval to the best of our knowledge there has been no such effort for video retrieval. Thus in this paper we attempt to address the issue of a lack of details on the creation assessments and provide a test collection with variable levels of relevance assessment for video retrieval, this is addressed in our 2 research objectives:

1. Create a video test collection with graded relevance assessments
2. To capture behavioural and perceptual information about the judgment process to augment the judgments.

In relation to the second objective, all data captured as part of the judgment process described in this paper is available for download (http://dx.doi.org/10.15129/7f16e19f-794a-4cd5-a9d7-caba4e2bcf2a). It is hoped this repository will not be a static resource, but rather we welcome other researchers to download and use this resource, which will be updated with future results from future evaluations of others. The remainder of this paper is organised as follows. In the next section we describe the corpus used for evaluation. We follow this with a description of the topics used for assessment. We follow this with an outline of the judgement process and information gathered. Finally we provide a conclusion section.

2. CORPUS
The British Universities Film and Video Council Roundabout Collection (http://bufvc.ac.uk/newsonscreen/roundabout) was used as a corpus. This collection contains a total of 660 videos showcasing Britain and Asia during the 1960s to the late 1970s. The videos have a mixed time duration, from about 44 secs up to about 10 minutes in duration. The videos are categorised based on their topics and each video contains additional metadata including a title, keywords, location, date and summary (see Table 1).

3. TOPICS
To identify topics all of the text from all of the documents was downloaded and placed into a tag cloud (see Figure 1) to visualise commonly occurring words and themes. This resulted in an initial list of 17 topics. To help narrow down the topics we created HITs on CrowdFlower which required users to create search terms. We wanted to see how diverse the range of search terms would be. After discussion of the topics amongst the research team and also analysis of the search terms returned, a final list of 10 topics was decided upon, those topics (and descriptions are):

1. Manufacturing in Britain: Identify products that Britain was manufacturing during the 1960s and 70s

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA © 2016 ACM.
ISBN 978-1-4503-3751-9/16/03...$15.00
DOI: http://dx.doi.org/10.1145/2854946.2854980

2. Outdoor Water Activities: Identify water sports well-known to the world or equipment used for the water sports
3. South East Asia Country Development: Identify countries in South East Asia and their development in areas of either education, industry, agriculture or society in the 1960s and 70s
4. British Exhibitions: Identify famous exhibitions you can see in Britain (including Scotland, England and Wales)
5. Asian Culture Events: Identify events, festivals or cultural exchanges between Asia and Britain
6. Industrial Research: Identify industries in Britain that have students studying and learning, Industrial experts investigating their industry through site visits can also be included.
7. Aviation Display: Identify demonstrations and events associated with aircraft.
8. Educational Visit for Children: Identify events where children or students visit other countries or take part in educational events such as learning new skills and supporting educational campaigns.
9. Transport for People: Identify the modes of transport commonly used by people in Asia and Britain.
10. Military Events: Identify well-known military ceremonies or events in Britain.

Table 1: Example of metadata provided for each video.

ID	327511
URL	http://bufvc.ac.uk/newsonscreen/search/index.php/story/327511
Start Time	17
End Time	166
Title	Treble One Squadron
Series Name	Roundabout
Issue No.	1
Date Released	May-62
Story within the issue	01-Mar
Summary	COI synopsis: Treble One Squadron, famous aerobatic unit of the Royal Air Force making its last public appearance before reconstituting as a fighter unit, in a breathtaking display of precision flying.
Keywords	Displays; Aviation; Air force
COI Ref	MI 1072/1

Figure 1: Tag cloud describing metadata from Roundabout Collection.

4. JUDGEMENT PROCESS

4.1 Procedure

Gathering assessments involved two stages. The first stage involved collecting search terms and the second involved collecting relevance assessments. For the first stage each participant was sent an online form. This form contained each of the search topics, for each topic the participants were asked to provide 4 queries that they would use to satisfy the information need exemplified in the topic. The participants were also asked to rate the perceived the difficulty of the topic on a 5 point scale.

In the second stage participants were assigned to groups. Each group was given a set of 2 topics and for each topic had to provide a relevance assessment for the top 5 ranked videos (see next section for explanation of ranking). Each set of relevance assessments was collected individually in our lab. The participants were asked to rate the relevance of each video on a four point scale. A four point scale has previously been proposed for gathering relevance assessments [9]. For our scale we used the 3 options from the TREC HARD track, namely not relevant, partially relevant and highly relevant. We also gave the participants the option to say that they are not sure.

After each judgment the participants were presented with a questionnaire. This questionnaire asked questions about effort which are inspired by the NASA TLX [7], which has previously been used to measure effort for both TREC HARD [15] and ImageCLEF [6]. Following the procedure of Kelly and Azzopardi [10] we used a 7 point Likert scale for our questions rather than the standard TLX scale. We also removed the question about performance, as this question reverses the scale. After each topic we also asked participants to judge their knowledge of the topic and how interesting the topic was on a 7 point Likert scale. An additional benefit of gathering this information is that it allows us to make some comparisons between the effort involved in making those assessments for video retrieval with the effort for text [15] and image [6] assessment.

4.2 Ranking Videos

To rank the videos we used the search terms gathered in the first part of the assessment. Each search query was used to rank the documents using TFIDF (based on the metadata associated with each video), the each document in top 10 documents of each rank was assigned a score from 10 for top rank down to 1 for 10th rank. These scores was aggregated across all queries and the videos with the 5 highest scores were used for assessment. The variance in search terms gathered in the first stage of the assessment insures a variance in returned videos. In this way we simulate the pooling that normally occurs when creating a test collection [13].

4.3 Participants

In total 20 participants were recruited as assessors. 12 male and 8 female, with an average age of 30. 10 were International students studying at the University of Strathclyde, 8 workers, 1 self-employed and 1 homemaker. For stage 1, a total of 80 distinct search terms were contribute by the participants across all 10 topics. For stage 2, a total of 50 videos were judged across all 10 topics, resulting in a total of 200 relevance judgements.

5. DATA GATHERED

In this section we outline some of the data gathered during the relevance assessments.

5.1 Relevance Assessments

In Table 2 we can see that there is a reasonably even distribution of assessments throughout the three relevance categories. We can also see differences in the results returned from our pooling approach, topics 2,3 and 10 have a high number of highly relevant documents, whereas topics 8 has only 1. Overall for only 2 videos were individual assessors unable to make a judgement. In terms of assessor agreement we plot agreement (see Figure 2) in a similar way to Alonso and Mizzaro [2], where we plot error distribution. For ground truth we consider the majority decision. Column 0 represents complete agreement, +1 represents 1 assessor having a positive assessment outside the consensus, -1 represents 1 assessor having a negative assessment outside the consensus. In our figure we have to additional columns, a tie represents when an equal number of assessors have differing assessments, n/a represents where there is a majority (2) and 2 other assessors disagree, 1 being more positive and 1 more negative in assessment. We can see only for a small number of documents is there a consensus (32%). This helps demonstrate the difficulty in getting a consensus and perhaps why multiple assessments a range of assessments may be beneficial in test collections.

Table 2: Choice of relevance assessment per topic.

Topic	Not Sure	Not Relevant	Partially Relevant	Highly Relevant
1	0	6	7	7
2	0	4	5	11
3	0	2	7	11
4	0	5	9	6
5	0	10	5	5
6	1	1	11	7
7	0	13	3	4
8	1	12	6	1
9	0	8	5	7
10	0	6	3	11
Total	2	67	61	70

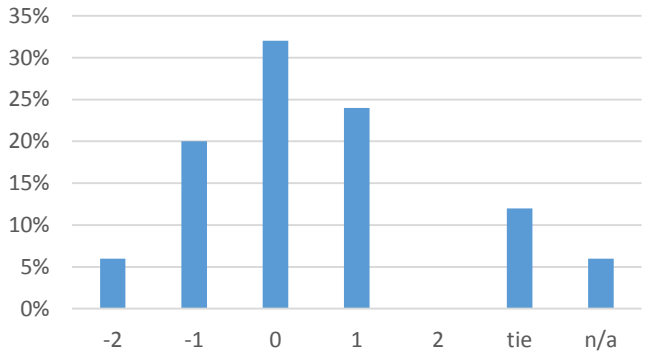

Figure 2: Agreement between reviewers Table 1 the number of responses given for each of the topics.

5.2 Time

Table 3 shows the average duration of videos judged for each topic. In comparison we present the average judgement time per video. It can be seen that for most judgements that the assessors do not watch the entire video. We can also see high variance in judgement time, with some topics taking a lot less time than other topics.

Table 3: Average duration for all videos in each topic

Topic Number	Avg. Duration (secs)	Avg. Judgement (secs)
1	373	140
2	264	54
3	352	156
4	364	163
5	214	58
6	401	85
7	294	126
8	212	145
9	262	55
10	329	64

Table 4: Video duration classification, sample mean and sample standard deviation

Time duration (sec)	Number of Videos judged	Mean (Std Dev.)
Short (0-200 sec)	68	63.49 (5.62)
Medium (201-400 sec)	56	121.61 (11.01)
Long (>401sec)	76	130.16 (12.27)

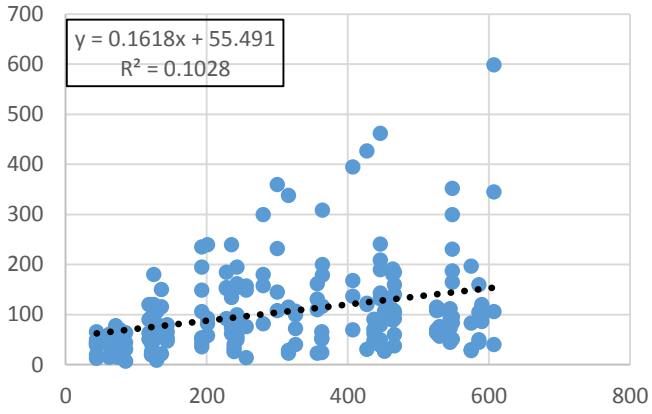

Figure 3: Scatter plot of video duration against judgement time per participant across all topics

We also considered if video length has any impact on assessment time. Initially we segmented videos in short (0-200 secs), medium (201-400 secs) and (401-600 secs) segments (see Table 4). It was found that longer videos took longer to judge on average in comparison with shorter videos. This finding is in keeping with previous research on text documents [15]. To look at this in more detail we plotted time duration of the video and the time per judgement in a scatter plot (see Figure 3). We can observe from the plot that there is only a weak correlation ($R^2 = 0.1028$) between the video duration and the judgement time. This suggests that as video time increase in video duration does not always produce a linear

increase in the time taken for relevance judgement. Based on the scatter plot, most relevance judgements across the collection are made within 200 secs while some judgements took more than 300 seconds to complete.

5.3 Workload

Table 5 presents the average responses for the 7 point Likert scales that measure work load and interest. The not sure category only has responses from 2 participants, but it clearly has higher workload and lower interest than any other assessment. In terms of workload we see that the more clear categories i.e. not relevant and highly relevant, have lower workload in almost all categories in comparison to partially relevant. Again this is in keeping with other research which has found that less clear relevance assessments have higher workloads [15].

Table 5: Average responses for post judgement questionnaires. All on a 7 point Likert scale. Higher=better.

Question	Not Sure	Not Relevant	Partially Relevant	Highly Relevant
Mental	4.5	3.0	3.4	2.29
Physical	5	2.15	1.94	1.9
Temporal	5.5	2.42	2.68	2
Effort	5.5	3.05	3.37	2.41
Frustration	4	2.86	3.06	1.94
Interest	3.5	3.77	3.94	4.59

6. CONCLUSION

In our introduction we set out two research objectives, namely:

1. Create a video test collection with graded relevance assessments
2. To capture behavioural and perceptual information about the judgment process to augment the judgments.

In this paper we have described the topics, document collection and judgment process that were used to create our test collection. As part of the judgment process we gathered feedback from the assessors on their assessments as well as measuring time taken to make judgments. All of this data is made available to other researchers. Researchers can use this test collection for video search evaluations, to the best of our knowledge a video test collection with variable relevance judgments is not available. We will also maintain this test collection and encourage supplementary data to be gathered by researchers.

7. ACKNOWLEDGEMENTS

This work was funded in part by a UK Arts and Humanities Research Council (grant AH/L010364/1) grant to the second and third authors.

8. REFERENCES

[1] ALLAN, J., 2005. HARD track overview in TREC 2003 high accuracy retrieval from documents. DTIC Document.

[2] ALONSO, O. and MIZZARO, S., 2012. Using crowdsourcing for TREC relevance assessment. Information Processing & Management 48, 6, 1053-1066.

[3] CARTERETTE, B., ALLAN, J., and SITARAMAN, R., 2006. Minimal test collections for retrieval evaluation. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval ACM, 268-275.

[4] GRUBINGER, M., 2007. Analysis and evaluation of visual information systems performance Victoria University.

[5] GRUBINGER, M., CLOUGH, P., HANBURY, A., and MÜLLER, H., 2008. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In Advances in Multilingual and Multimodal Information Retrieval Springer, 433-444.

[6] HALVEY, M. and VILLA, R., 2014. Evaluating the effort involved in relevance assessments for images. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval ACM, 887-890.

[7] HART, S.G. and STAVELAND, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology 52, 139-183.

[8] HASLER, L., HALVEY, M., and VILLA, R., 2015. Augmented Test Collections: A Step in the Right Direction. arXiv preprint arXiv:1501.06370.

[9] KEKÄLÄINEN, J. and JÄRVELIN, K., 2002. Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology 53, 13, 1120-1129.

[10] KELLY, D. and AZZOPARDI, L., 2015. How many results per page?: A Study of SERP Size, Search Behavior and User Experience. In Proceedings of the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2767732, 183-192.

[11] SARACEVIC, T., 1975. Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science 26, 6, 321-343.

[12] SORMUNEN, E., 2002. Liberal relevance criteria of TREC-: Counting on negligible documents? In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval ACM, 324-330.

[13] SPARCK JONES, K. and VAN RIJSBERGEN, C.J., 1976. Information retrieval test collections. Journal of Documentation 32, 1, 59-75.

[14] TANG, R. and SOLOMON, P., 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. Information Processing & Management 34, 2, 237-256.

[15] VILLA, R. and HALVEY, M., 2013. Is relevance hard work?: evaluating the effort of making relevant assessments. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval ACM, 765-768.

[16] WANG, J., 2011. Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In Proceedings of SIGIR Information Retrieval for E-Discovery Workshop, 1.