TRECVID as a Re-Usable Test-Collection for Video Retrieval

Thijs Westerveld
CWI
PO Box 94079, 1090 GB Amsterdam
The Netherlands
thijs@cwi.nl

ABSTRACT

TRECVID has been running as a video retrieval benchmarking platform for a number of years now. Some progress seems to be made in the area of video retrieval, but also it has been shown that many of the differences in scores between tested approaches are non-significant [8]. This paper studies the reliability of the TRECVID search collections for measuring video retrieval effectiveness and investigates how useful the collections are for re-use.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Keywords

Multimedia Retrieval, Evaluation, Pool quality

1. INTRODUCTION

Until recently no commonly used evaluation methodology existed for content-based image and video retrieval. An important reason for this is that for a long time, the field has been merely a showcase for computer vision techniques. Many papers in the field 'proved' the technical merits and usefulness of their approaches to image processing by showing a few well-chosen, and well-performing examples. Since 1996 the problem of systematically evaluating multimedia retrieval techniques has gained more and more interest. In that year, the MIRA (Multimedia Information Retrieval Applications) working group was formed [3, 4]. The group, consisting of people from the fields of information retrieval, digital libraries, and library science, studied user behaviour and information needs in multimedia retrieval situations. Based on their findings, they developed performance measures. Around the same time, in the multimedia community the discussion on proper evaluation started, and Narasimhalu et al. [12] proposed measures for evaluating content-based information retrieval systems. These measures are based on comparing a system's ranked list of documents to the perfect, or ideal, ranking. However, they do not specify how to obtain such a perfect ranking, nor do they propose a common test set. A year later, Smith [19] proposed to look at the text retrieval community, and to use measures from TREC [1] for image retrieval evaluation. Again, no dataset was proposed. At the start of the 21st century, the evaluation problem gained more attention within the content-based image retrieval community, with the publication of three papers discussing benchmarking in visual retrieval [5, 9, 11]. These three papers call for a common test collection and evaluation methodology and a broader discussion on the topic. The BENCHATHLON network[2] was started to discuss the development of a benchmark for image retrieval. Then, in 2001, TREC started a video track [17, 18] that evolved into the workshop now known as TRECVID [15, 16].

2. LABORATORY TESTS IN INFORMATION RETRIEVAL

Information retrieval is interactive. In web search, for example, queries are often changed or refined after an initial set of documents has been retrieved. In multimedia retrieval, where browsing is common, interactivity is perhaps even more important. Saracevic [14], and Sparck Jones and Willett [23] argue that evaluation should take interactivity into account, and measure user satisfaction. Tague-Sutcliffe [24] called evaluation of a system as a whole in an interactive setting an *operational test*. Such tests measure performance in a realistic situation. Designing such an operational test is difficult and expensive: many users are needed to free the experiment of individual user effects, the experimental setup should not interfere with the user's natural behaviour, and learning effects need to be minimised. Also, because there are many free variables, it is hard to attribute observations to particular causes. In contrast to these tests in fully operational environments, Tague-Sutcliffe defined *laboratory tests* as those tests in which possible sources of variability are controlled. Thus, laboratory tests can provide more specific information, even though they are further away from a realistic setting. Also, laboratory tests are cheaper to set up, because the interactive nature is ignored, and the role of the user is reduced to judging for relevance. Laboratory tests measure the quality of the document ranking instead of user satisfaction.

2.1 The Cranfield tradition

Most current evaluation procedures, including TRECVID, are laboratory tests, based on the CRANFIELD paradigm [1]. This section provides a short introduction to this paradigm. A thorough review of the fundamental assumptions behind CRANFIELD style experiments can be found in [26].

The term *laboratory tests* will be used to refer to tests following this paradigm. A test collection for laboratory tests consists of a fixed set of documents, a fixed set of topics, and a fixed set of relevance judgements. Documents are the basic elements to retrieve, topics are descriptions of the information needs, and relevance judgements list the set of relevant documents for each topic.

[1] The yearly text retrieval benchmark [27].

[2] `http://www.benchathlon.net`

The focus in laboratory tests is on comparative evaluation. Different approaches are tested, and their relative performance is measured. The process is as follows. Each approach produces a ranked list of documents for each topic. The quality of the ranked lists is measured based on the positions of the relevant documents in the list. The results are averaged across all topics to obtain an overall quality measure.

2.2 TRECVID

TRECVID is a laboratory test for evaluating the effectiveness of video retrieval systems based on the CRANFIELD tradition.

TRECVID defines various tasks: shot boundary detection, scene detection, feature detection and general information search. This paper concentrates on the search task, where the goal is to find as many relevant shots as possible for a given topic, within a predefined search collection. Topics consist of a short textual description of the information need and one or more still images or video examples. Figure 1 shows an example.

The test collections for TRECVID 2003 and 2004 consisted of broadcast news material. Each of these collections contained over 30,000 shots and for each year around 25 search topics where available. Participants in TRECVID submit their top N results for each topic. These submissions are judged by human accessors[3] and systems get scored based on their ability to retrieve relevant shots for the topics. The metric most commonly used to compare systems is mean average precision (MAP). Average precision is the average of the precision values measured after each relevant document that is retrieved, using zero as the precision value for non-retrieved documents. MAP is the mean across topics of these Average precision scores.

2.3 Reliability of Laboratory tests

A number of aspects influences the reliability of evaluation results. First, a sufficiently large set of topics is needed. Sparck Jones and van Rijsbergen [22] suggest a minimum of 75. Second, the measures should be stable. This means it should not be influenced too much by chance effects. Clearly, measures based on few observations are less stable than measures based on many observations. For example, precision at rank 1 –is the first retrieved document relevant?– is not a very stable measure. Third, there needs to be a reasonable difference between two approaches before deciding one approach is better than the other. Sparck Jones [20] suggests a 5% difference is noticable, and a difference greater than 10% is material. Finally, the relevance judgements on which all measures are based should be reliable. The following sections discuss the details of these four conditions in the context of TRECVID.

3. RELIABLE MEASURES, NUMBER OF TOPICS AND DIFFERENCE IN SCORES

The first three reliability conditions mentioned in the previous section (reliable measures, enough topics, difference in scores) are clearly interrelated. For example, when stable measures are used, fewer topics are needed; and when many topics are used, a smaller difference in scores can lead to the conclusion that two approaches are different.

Hauptmann and Lin [8] analyse TRECVID 2003 and 2004 results to assess the reliability of results obtained with these collections. They measure a Retrieval Experiment Error Rate (REER), which is defined as the probability that system A is judged more effective than system B on one topic set while the effectivity judgement is reversed on another set of topics. Based on average precision values for the runs submitted for TRECVID 2003 and 2004, they conclude that for REER to be smaller than 0.05, a difference in MAP scores should be greater than 0.02 before concluding one submissions is more effective than the other. A theoretic analysis of the REER [10] shows that factors influencing this error rate are number of topics, the difference in scores and the variance in scores (across topics). REER is shown to decrease when more topics are used, when the difference between scores for two submissions is large or when the variance across topics for the submissions are low.

In separate experiments, Hauptmann and Lin [8] perform analysis of variance (ANOVA) tests and pair-wise significance tests on the TRECVID results. They find large sets of submissions for which no significant difference in effectiveness is found. Hauptmann and Lin's findings in both papers indicate that either more than 25 topics are needed to be able to draw reliable conclusions, or a larger difference between score should be observed before concluding one approach is better than another.

4. RELIABLE JUDGEMENTS

The main assumptions with regard to relevance in laboratory tests are the following. First, relevance is approximated by topical similarity: a document is relevant if it is on topic, i.e., if it discusses the topic of the query in a text retrieval setting or if it shows the topic in video retrieval. This means the information need is assumed not to change over time. It also means relevance is judged independently for each document. If a document contains information that is on topic, but all this information is already present in other documents, the document is still regarded relevant. The second assumption is that relevance judgements are representative of a user population. Although the judgements are a single person's opinion, they are assumed to be representative of the typical user. Third, judgements are assumed to be complete. For each topic, all relevant documents in the collection are identified. Finally, judgements are often assumed to be binary, i.e., a document is either relevant to a topic or it is not. The original CRANFIELD experiments used graded relevance judgements on a five-point scale, but most modern laboratory tests assume binary judgements.

Clearly, these assumptions do not hold. Relevance judgements from a single user do not represent the opinion of a whole population, topical similarity is not the same as utility, and in many cases it is impossible to identify all relevant documents in a collection. However, the goal in laboratory tests is to compare retrieval strategies, not to find an indication of their absolute performance. Therefore, even though the assumptions may not be strictly true, laboratory tests may be useful. The concern is not so much about the truth in the assumptions, but about the influence of the assumptions on relative scores. Below we look at several aspects of relevance judgements for the TRECVID collection. We start with investigating the effects of incomplete judgements on comparative results. Then we test if there is a bias against systems that did not contribute to the assessments, i.e., is the test collection re-usable for evaluating (new) approaches. Finally, we discuss the representativeness of the judgements.

4.1 Incompleteness

Ideally full relevance judgements would be available, i.e., the relevance value for each document-topic pair would have been judged manually by some assessor. In practise, however, this is impossible. With a document collection of a realistic size (30,000+ documents for TRECVID) it is unfeasible for somebody to assess each document for a given topic. Instead, TRECVID uses a pooling method for creating judgements [21].

[3]More on the process of obtaining relevance judgements follows in Section 4.

vt0110: Find shots of a person diving into some water.

Figure 1: Example topic from the TRECVID 2003 collection.

Pooling is the process of forming a pool, or set, of the top ranked documents from a variety of different approaches. As explained before, participants in TRECVID submit their top N results for each topic in the search task. For each topic a pool is constructed, consisting of the union of the top $K < N$ retrieved documents from all submissions. Only the documents in the pools are judged for relevance. Before showing these documents to the assessors, they are randomised to make sure the assessors have no knowledge of the number of approaches that retrieved a given document or at what rank it was retrieved. Documents not retrieved within any approach's top K are assumed not relevant. The idea behind this approach is that documents that are not retrieved at a high rank by any system are unlikely to be relevant. This assumption may not be valid. Indeed both Harman [6] and Zobel [29] show that in the TREC collections some of the unjudged documents are in fact relevant. This could potentially influence the results since systems are usually evaluated on a top $N > K$. However, if the pool is large and sufficiently diverse, that is, if many different techniques contributed to the pool, then the fact that some relevant documents are missing is assumed to be of little consequence. For text retrieval, pool quality has been intensively studied. Zobel [29] found that incomplete judgements do not influence comparative results, i.e., the relative ranking of the approaches does not change. For content-based image and/or video retrieval, similar tests of pool quality have not been performed yet. This section investigates the pool quality of the TRECVID 2003 collection.

In TRECVID 2003, for each topic the pool has been created by taking the top K results from each of the submissions. The resulting set of documents is then manually inspected for relevance. This set of documents together with the relevance judgement for each of these (either relevant or not relevant), is known as the *qrels*. The pool depth K for TRECVID 2003 was either 50 or 100, depending on the number of relevant documents found in the top 50 (if many were found, the depth was increased to 100).

To test the effect of pool depth on the the measurements, we looked at smaller pool sizes. We re-evaluated all submissions on qrels obtained from pool depths varying from 1 to 50 (these modified qrels can easily be obtained by assuming all documents that are not retrieved within *any* top K for a given topic are not relevant for that topic). Figure 2 shows the MAP for all submissions based on the original qrels (circles), and for the different pool sizes (dots), the submissions are sorted by decreasing original MAP. The figure shows, that the scores based on qrels from the smaller pools follow the trend of the original scores. This means the ranking of systems is not influenced much by the pool depth. Still, some submissions may swap positions when a different pool depth is used for evaluation. To quantify how often such swaps occur, and how severe they are, we measure the correlations between system rankings using Kendall's τ, a measure of the correlation between two ordered lists [e.g., 2]. Kendall's τ is based on the minimum number of adjacent

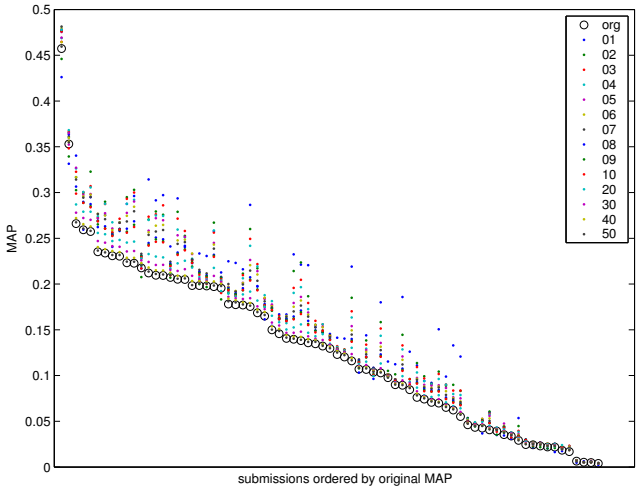

Figure 2: MAP for TRECVID 2003 submissions based on original qrels and qrels obtained from smaller pool depths.

swaps needed to turn one ranking into another. Two identical rankings would produce a τ of 1.0, a ranking and it's perfect inverse would produce $\tau = -1.0$, and for two random rankings one can expect tau to be 0.0. Figure 3 shows the correlation between the original system ranking and rankings based on smaller pool depths. Rankings obtained from smaller pools are highly correlated to the scores obtained from the original full pools. Even a pool depth of $K = 3$, (i.e., only the first three documents of each submission get judged), shows high correlation to the original ranking $\tau > 0.90$.

4.2 Bias

Another concern with incomplete relevance judgements, is their usefulness for evaluating approaches that have not contributed to the pool. The set of relevance judgements could potentially be biased against them. Zobel [29] and Voorhees and Harman [28] show that for the text retrieval collections used at TREC, this is not the case. A first indication that the TRECVID pools are not biased in favour of approaches that have contributed is that the number of documents uniquely retrieved by a single system is low, both for TRECVID 2003 and 2004. That means it is likely that the relevant documents found by a new approach which did not contribute to the pool are in the pool already. This section studies the quality of the pool for re-use in detail.

For each submission, we compute the MAP based on the original pool, and the MAP based on a modified pool from which we removed documents that are uniquely contributed by the submission under study. A third MAP was computed based on a modified pool from which we removed the documents uniquely contributed

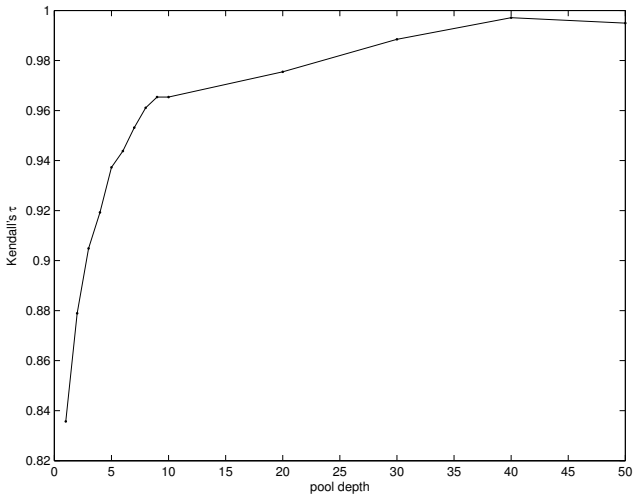

Figure 3: Kendall's τ between system rankings based on MAP obtained from different pool depths and MAP obtained from the original pool.

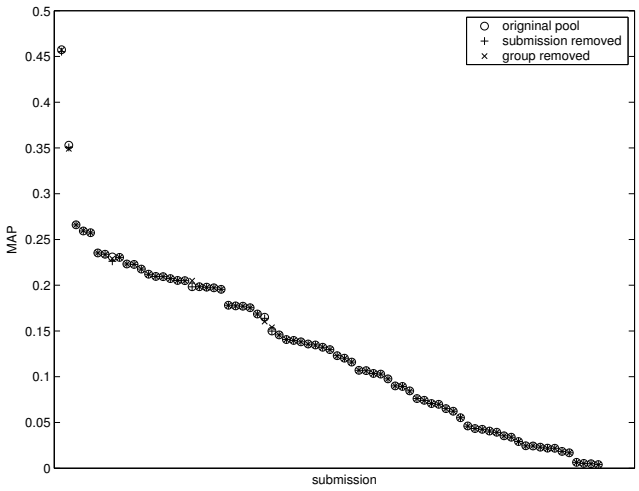

Figure 4: MAP for TRECVID 2003 submissions based on original qrels and qrels after removing submission or submission's group from the pool.

by that submission's group. This third MAP mimics the situation in which the group had not participated at TRECVID at all. Figure 4 shows for each submission the original MAP scores, and the ones obtained after removing that submission or the submission's group from the pool. The results based on the modified pools follow the original results almost perfectly. The correlation between system rankings based on original MAPs and modified MAPs is very high: $\tau > 0.98$

4.3 Subjectivity

It is well known that relevance judgements are subjective. Different judges will have different opinions on the relevance of documents [e.g., 7]. Since the focus is on comparative results this is not necessarily problematic. As Voorhees [25] states:

> For a test collection, the important question is not so much how well assessors agree with one another, but how evaluation results change with the inevitable differences in assessment.

Voorhees [25] investigates the influence of difference in assessments on evaluation results by having topics judged by multiple assessors. The different approaches have been evaluated using different combinations of judgements, and ranked by mean average precision. Voorhees finds the resulting rankings are highly correlated, and concludes comparative results are stable with regard to the subjectivity in relevance judgements.

For multimedia retrieval, such multi-assessor studies have not been conducted yet. Clearly, judgements in visual information retrieval are subjective as well. However, judging visibility as is done in TRECVID, is arguably more objective than judging aboutness or topicality in TREC. The assessor guidelines for TRECVID state [13]:

> When a topic says a shot must "contain x" that is short for "contain x to a degree sufficient for x to be recognisable as x to a human". This means among other things that unless explicitly stated, partial visibility or audibility may suffice.

The requested item is either visible or not, there is little room for discussion. Thus, agreement on visibility in TRECVID can be expected to be relatively high. Still, experiments with multiple assessors judging the same topics are needed to verify this as well as to investigate the effects on comparative results.

5. DISCUSSION

The findings of Hauptmann and Lin [8] invalidate conclusions based on small differences in MAP at TRECVID. Based on the current TRECVID results we need to conclude that most common approaches are equally effective. Still, a significant difference can (often) be observed between interactive approaches and non-interactive approaches, between approaches that do and do not incorporate textual information, and between interactive systems used by expert and by novice users.

One explanation for not finding significant differences between the most common approaches is the relatively small number of topics that is used in the evaluations. But, another could be that there really are no big differences in effectiveness. The only way to find that out would be to evaluate the approaches using a larger topic set, but that requires a larger effort from the assessors.

The analysis of the pool quality shows that comparative results are very stable against (small) changes in the pools. Both reducing the depth of the pool and removing uniquely contributed documents have only minor effects on the relative ordering of systems by their MAP scores. This means that the TRECVID collections are valuable test collections that can be re-used without the need of repeating the assessment process. A danger is that pooling effects appear to be small because many similar approaches contribute to the pool. When the contributions of one approach are removed, the pool contents hardly changes since a similar approach (from another group) is bound to have found almost the same set of documents. Therefore, fixed test collections may be most useful for evaluating variants of existing techniques. It remains unclear how the evaluation of revolutionary techniques will be influenced. Still, even the interactive runs –clearly different in technique from the manual ones– do not suffer much from not contributing to the pool. Even when all documents that are found by interactive submissions only are removed from the pool, the resulting ranking of submission is very close to the ranking based on the original pool ($\tau = 0.99$). The high quality of the pool could perhaps be attributed to the relatively

small size of the document collection. With a collection of only 30,000 shots, and a fair number of submissions, it is likely that the judgements are nearly complete.

The re-usability of the collections creates the possibility of comparing systems based on larger topic sets without the drawback of the additional assessor effort. If the document collection could be fixed for a number of years in a row, than the topics and relevance judgements for these years could be merged to create a bigger test collection, allowing for more reliable conclusions regarding the effectiveness of the various approaches. Unfortunately, so far not only the topics, but also the document collection has changed yearly. In any case, results on two different test collections can never be compared directly, thus it is impossible to compare results from two different editions of TRECVID. Nevertheless, the collections used in 2003 and 2004 are highly comparable (both are news broadcasts from ABC and CNN news), and could therefore be combined, thus creating a new test collection. Running systems on the combined collection, and computing aggregate scores like MAP over the union of the topics for the two years would already allow for a more reliable comparison of approaches.

References

[1] Cyril W. Cleverdon. The cranfield tests on index language devices. *Aslib Proceedings*, pages 173–192, 1967.

[2] W. J. Conovar. *Practical Non-Parametric Statistics*, pages 249–250. John Wiley and Sons, 1980.

[3] S. W. Draper, M. D. Dunlop, I. Ruthven, and C. J. van Rijsbergen, editors. *Proceedings of Mira 99: Evaluating Interactive Information Retrieval*, Electronic Workshops in Computing, Glasgow, Scotland, 1999. British Computer Society.

[4] Mark Dunlop. Reflections on mira: interactive evaluation in information retrieval. *Journal of the American Society for Information Science*, 51(14):1269–1274, 2000. ISSN 0002-8231.

[5] Neil J. Gunther and Giordano Beretta. A benchmark for image retrieval using distributed systems over the internet: BIRDS-I. Technical Report HPL-2000-162, HP Laboratories, 2000.

[6] Donna K. Harman. Overview of the fourth text retrieval conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference, TREC-4*, pages 1–23. NIST Special Publications, 1996.

[7] Stephen P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Sociecty for Information Science*, 47(1):37–49, 1996. ISSN 0002-8231.

[8] Alex Hauptmann and Wei-Hao Lin. Assessing effectiveness in video retrieval. In *Proceedings of The International Conference on Image and Video Retrieval (CIVR2005)*, 2005. to appear.

[9] Clement H. C. Leung and Horace Ho-Shing Ip. Benchmarking for content-based visual information search. In Robert Laurini, editor, *Advances in Visual Information Systems, 4th International Conference, VISUAL 2000, Lyon, France, November 2-4, 2000, Proceedings*, volume 1929 of *Lecture Notes in Computer Science*. Springer, 2000. ISBN 3-540-41177-1.

[10] Wei-Hao Lin and Alex Hauptmann. Revisiting the effect of topic set size on retrieval error. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005. to appear.

[11] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters (Special Issue on Image and Video Indexing)*, 22(5):593–601, 2001. URL `http://www.elsevier.nl/gej-ng/10/35/61/49/29/36/abstract.html`. H. Bunke and X. Jiang Eds.

[12] A. Desai Narasimhalu, Mohan S. Kankanhalli, and Jiankang Wu. Benchmarking multimedia databases. *Multimedia Tools and Applications*, 4(3):333–356, 1997. ISSN 1380-7501.

[13] Paul Over. personal communication, 2004.

[14] Tefko Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146. ACM Press, 1995. ISBN 0-89791-714-6.

[15] Alan F. Smeaton, Wessel Kraaij, and Paul Over. TRECVID 2003 - an introduction. In Alan F. Smeaton, Wessel Kraaij, and Paul Over, editors, *TRECVID 2003 Workshop*, Gaithersburg, MD, USA, 2003. NIST, NIST Special Publications.

[16] Alan F. Smeaton and Paul Over. The TREC-2002 video track report. In Ellen M. Voorhees and Lori P. Buckland, editors, *The Eleventh Text Retrieval Conference, TREC 2002*. National Institute of Standards and Technology, NIST Special Publications, 2003.

[17] Alan F. Smeaton, Paul Over, Cash J. Costello, Arjen P. de Vries, David Doermann, Alexander Hauptmann, Mark E. Rorvig, John R. Smith, and Lide Wu. The TREC-2001 video track: Information retrieval on digital video information. In *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings*, volume 2458 of *Lecture Notes in Computer Science*, pages 266–275, Rome, Italy, September 2002. Springer. ISBN 3-540-44178-6.

[18] Alan F. Smeaton, Paul Over, and Ramazan Taban. The TREC-2001 video track report. In Ellen M. Voorhees and Donna K. Harman, editors, *The Tenth Text Retrieval Conference, TREC 2002*, volume 10. National Institute of Standards and Technology, NIST Special Publications, 2002.

[19] John R. Smith. Image retrieval evaluation. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, 1998.

[20] Karen Sparck Jones. Automatic indexing. *Journal of Documentation*, 30:393–432, 1974.

[21] Karen Sparck Jones and Cornelis J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. Technical Report British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[22] Karen Sparck Jones and Cornelis J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32 (1):59–75, 1976.

[23] Karen Sparck Jones and Peter Willett. Evaluation. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, chapter 4, pages 167–174. Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-454-5.

[24] Jean Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992. ISSN 0306-4573.

[25] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323. ACM Press, 1998. ISBN 1-58113-015-5.

[26] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, and Michael Kluck Julio Gonzalo, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406, pages 355–370. Springer-Verlag, 2002.

[27] Ellen M. Voorhees and Lori P. Buckland, editors. *Proceedings of the Thirteenth Text Retrieval Conference, (TREC-2004)*, 2004. NIST Special Publications.

[28] Ellen M. Voorhees and Donna K. Harman. Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of the Eighth Text Retrieval Conference, TREC-8*. NIST Special Publications, 2000.

[29] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM Press, 1998. ISBN 1-58113-015-5.