

Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899
ellen.voorhees@nist.gov

ABSTRACT

Benchmark tasks are a powerful method for advancing the state of the art in a field. The music information retrieval community recently acknowledged the utility of such tasks by resolving to create an evaluation framework for music information retrieval (MIR) tasks and music digital libraries (MDL). This paper describes the processes used in the Text REtrieval Conference (TREC) evaluations to create information retrieval evaluation infrastructure, reviews assessments of how appropriate the evaluation methodology is for TREC tasks, and makes suggestions regarding the development of an MIR/MDL evaluation framework based on TREC experience.

1. INTRODUCTION

The music information retrieval (MIR) and music digital library (MDL) community has recently issued a resolution calling for the construction of the infrastructure necessary to support MIR/MDL research [8]. The vision is to define benchmark tasks with standardized collections of music to provide researchers with the means to make comparisons among different approaches to MIR problems. One suggestion for realizing the vision is to create the benchmarks through a series of evaluations similar in spirit to the Text REtrieval Conference (TREC) workshops.

This paper provides a history of TREC with an emphasis on the lessons learned regarding retrieval evaluation. While TREC started with just two document-oriented tasks, it has grown to encompass a variety of additional tasks including question answering, retrieval of speech recordings, and content-based access to digital video. Some of these new tasks have required different evaluation methodologies from the Cranfield paradigm used extensively in document retrieval evaluation. Similar approaches may prove useful for MIR/MDL tests.

2. HISTORY OF TREC

The Text REtrieval Conference (TREC) was started in 1992 to support the text retrieval industry by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. At that time, the Cranfield tradition of using retrieval experiments on test collections was already well-established [15], but progress in the field was hampered by the lack of realistically large test collections. Large test collections did exist, but they were proprietary, with each collection usually the result of a single company's efforts. The proprietary nature of the collections biased them in various ways, in addition to rendering them unavailable to most of the retrieval research

community. TREC was conceived as a way to address this need for large, unbiased test collections.

TREC began with the following goals:

- to encourage research in text retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

Most of these goals have been accomplished for the particular task of "ad hoc" document retrieval. (An ad hoc task investigates the performance of systems that search a static set of documents using new queries. This task is similar to how a researcher might use a library—the collection is known but the questions likely to be asked are not known.) Through the pooling process (described below) used in TREC, eight ad hoc test collections have been created, each collection consisting of approximately 800,000 documents, 50 information need statements, and the associated relevance judgments¹. Participation in TREC has grown each year, starting with 25 research groups in 1992 and growing to 87 groups from 21 different countries in TREC 2001. Researchers have adapted techniques demonstrated to be successful in TREC to their own systems, resulting in a doubling of ad hoc document retrieval performance in the first six years of TREC [22].

The initial TREC contained two main tasks, the ad hoc task and the routing task. Starting in TREC-3 (1994), TREC introduced additional tasks called "tracks" which focused research on particular facets of retrieval. The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and the track creates the necessary infrastructure to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups. Table 1 lists the different

This paper is authored by an employee of the United States Government and is in the public domain.

¹ These collections are available to the research community at <http://trec.nist.gov/data.html>.

tracks that were in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participate in each TREC. (More details about the tracks run in a particular TREC can be found in the overview paper for that year in the TREC proceedings. The TREC proceedings are available on the TREC web site, <http://trec.nist.gov>.)

The set of tracks run in any particular year depends on the interests of the participants and sponsors, as well as on the suitability of the problem to the TREC environment. Some initial tracks have been discontinued because the goals of the track were met. For example, the Spanish track, an ad hoc task in which

both topics and documents are in Spanish, was discontinued when the results demonstrated that current retrieval systems can retrieve Spanish documents as effectively as English documents. Other tracks, such as the interactive track, have been run each year, but have changed their focus in different years. For all of its tasks, TREC provides a common test set to focus research on the particular retrieval task, yet actively encourages participants to do their own experiments within the umbrella task. These individual experiments broaden the scope of the research that is done within TREC and makes TREC more attractive to individual participants.

Table 1. Number of participants per track and total number of distinct participants in each TREC

Track	TREC									
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Ad Hoc	18	24	26	23	28	31	42	41	—	—
Routing	16	25	25	15	16	21	—	—	—	—
Interaction	—	—	3	11	2	9	8	7	6	6
Spanish	—	—	4	10	7	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—
Database Merging	—	—	—	3	3	—	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15	19
Chinese	—	—	—	—	9	12	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—
Cross-Language	—	—	—	—	—	13	9	13	16	10
High Precision	—	—	—	—	—	5	4	—	—	—
Very Large Corpus	—	—	—	—	—	—	7	6	—	—
Query	—	—	—	—	—	—	2	5	6	—
Question Answering	—	—	—	—	—	—	—	20	28	36
Web	—	—	—	—	—	—	—	17	23	30
Video	—	—	—	—	—	—	—	—	—	12
Total participants	25	31	33	36	38	51	56	66	69	87

3. THE CRANFIELD PARADIGM

One of the goals of TREC from the beginning was to use TREC as a vehicle for exploring retrieval evaluation techniques. Text retrieval already had a well-established evaluation tradition, the Cranfield paradigm, at the time TREC started, and TREC has continued in that tradition. Nevertheless, TREC's scale—both in terms of the size of the collections and the number of retrieval results—provides an unique opportunity for exploring the limits of the evaluation methodology.

This section first describes the Cranfield paradigm and examines the assumptions inherent in it. Since the main component of the Cranfield paradigm in the test collection, the second subsection describes how test collections are built in TREC. The section concludes by summarizing a set of experiments that confirm the utility of TREC test collections as laboratory tools.

3.1 Cranfield Assumptions

The Cranfield paradigm of text retrieval evaluation is based on an abstraction of the retrieval process known as a test collection. A test collection consists of three components, a document set, a set of information need statements (called "topics" in TREC), and a set of relevance judgments. The relevance judgments are a list of which documents should be retrieved for each topic. Test collections are laboratory tools that allow experimenters to control some of the variables that affect retrieval performance thereby increasing the power of comparative experiments. They are necessarily abstractions to the total retrieval process, providing control at the expense of realism by substituting a static set of relevance judgments for the complex interactions of a user. In much the same way as the medical community uses laboratory research to identify promising treatments before undergoing the expense of human clinical trials, IR researchers can use test collections to identify promising retrieval

technologies that can then be tested under operational conditions.

Laboratory testing of retrieval systems was first done in the Cranfield 2 experiment [5], where the Cranfield experiments were a series of investigations into which of several alternative indexing languages was best. The experimental design of Cranfield 2 called for the same set of documents and same set of information needs to be used for each language, and for the use of both precision and recall to evaluate the effectiveness of the search. (Recall is the proportion of relevant documents that are retrieved while precision is the proportion of retrieved documents that are relevant.) Relevance was based on topical similarity where the judgments were made by domain experts (i.e., aeronautics experts since the document collection was an aeronautics collection).

The Cranfield experiments made three major simplifying assumptions. The first assumption was that relevance can be approximated by topical similarity. This assumption has several implications: that all relevant documents are equally desirable, that the relevance of one document is independent of the relevance of any other document, and that the user information need is static. The second assumption was that a single set of judgments for a topic is representative of the user population. The final assumption was that the list of relevant documents for each topic is complete (all relevant documents are known). The vast majority of test collection experiments since then have also assumed that relevance is a binary choice, though the original Cranfield experiments used a five-point relevance scale.

Of course, in general these assumptions are not true, which makes laboratory evaluation of retrieval systems a noisy process. The primary consequence of the noise is the fact that evaluation scores computed from a test collection are valid *only* in comparison to scores computed for other runs using the exact same collection. A second consequence of the noise is that there is an (unknown) amount of error when comparing two systems on the same collection. Researchers have evolved a standard experimental design to decrease the noise, and this design has become an intrinsic part of the Cranfield paradigm. In the design, each retrieval strategy to be compared produces a ranked list of documents for each topic in a test collection, where the list is ordered by decreasing likelihood that the document should be retrieved for that topic. The effectiveness of a strategy for a single topic is computed as a function of the ranks at which the relevant documents are retrieved. The effectiveness of the strategy on the whole is then computed as the average score across the set of topics in the test collection.

This design contains three interrelated components—the number of topics used, the evaluation measures used, and the difference in scores required to consider one method better than the other—that can be manipulated to increase the reliability of experimental findings [3]. Since retrieval system effectiveness is known to vary widely across topics [1], the greater the number of topics used in an experiment the more confident the experimenter can be in its conclusions. TREC uses 25 topics as a minimum and 50 topics as the norm, though a recent empirical investigation of past TREC results suggests that even 25 topics is probably too few [20]. A wide variety of different evaluation measures have been developed (see van Rijsbergen [18] for a summary), and some are inherently less stable than others. For

example, measures based on very little data such as precision at one document retrieved (i.e., is the first retrieved document relevant?) are very noisy, and the mean average precision measure, which measures the area underneath the entire recall-precision curve, is much more stable. Requiring a larger difference between scores before considering the respective retrieval methods to be truly different increases reliability at the cost of not being able to discriminate between as many methods.

A major departure from the original Cranfield 2 experiments and TREC test collections is the relative emphasis given to collection size and the completeness of relevance judgments. While Cleverdon's experience suggested that the size of the document collection was not important so that every document in the collection could be judged for each topic [6], further experience has shown that collection size (i.e., the number of documents in the collection *does* matter. Text retrieval is challenging because of the large number of different ways the same concept can be expressed in natural language, and larger collections are generally more diverse. Unfortunately, even a moderately large collection cannot possibly have complete relevance judgments. Instead, TREC uses a technique called pooling [14] to create a subset of the documents (the "pool") to judge for a topic. Before discussing the effect of pooling on retrieval evaluation, it is necessary to understand how the test collections are built.

3.2 Building TREC Test Collections

NIST provides a document set and a set of topics to the TREC participants. Each participant runs the topics against the documents using their retrieval system, and returns to NIST a ranked list of the top 1000 documents per topic. NIST forms pools from the participants' submissions, which are judged by human relevance assessors. Each submission is then evaluated using the resulting relevance judgments, and the evaluation results are returned to the participant.

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. The TREC ad hoc collections contain mostly newspaper or newswire articles, though some government documents (the *Federal Register*, a small collection of patent applications) are also included in some of the collections to add variety. These collections contain about 2 gigabytes of text (between 500,000 and 1,000,000 documents). The document sets used in various tracks have been smaller and larger depending on the needs of the track and the availability of data.

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. The format of a topic statement has evolved since the beginning of TREC, but it has been stable for the past several years. A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative. A sample topic is shown in Figure 1.

```
<num> Number: 409
<title> legal, Pan Am, 103

<desc> Description:
What legal actions have resulted from the destruction of
Pan Am Flight 103 over Lockerbie, Scotland, on
December 21, 1988?

<narr> Narrative:
Documents describing any charges, claims, or fines
presented to or imposed by any court or tribunal are
relevant, but documents that discuss charges made in
diplomatic jousting are not relevant.
```

Figure 1: A sample TREC topic statement

TREC topic statements are created by the same person who performs the relevance assessments for that topic. Each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the document collection to determine the likely number of relevant documents per candidate topic. The NIST TREC team selects the final set of topics from among these candidate topics based on the estimated number of relevant documents and balancing the load across assessors.

The relevance judgments are what turn a set of documents and topics into a test collection. TREC has almost always used binary relevance judgments—either a document is relevant to the topic or it is not. To define relevance for the assessors, the assessors are told to assume that they are writing a report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

The judgment pools are created as follows. NIST selects the maximum number of runs that can be contributed to the pools by a single participating group; each group contributes this many runs to the pools unless they submitted fewer runs to NIST (in which case all their retrieval runs contribute to the pools). When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST merges the runs into the pools respecting this preferred ordering. For each selected run, the top X documents (usually, $X = 100$) per topic are added to the topics' pools. Since the retrieval results are ranked by decreasing similarity to the query, the top documents are the documents most likely to be relevant to the topic. Many documents are retrieved in the top X for more than one run, so the pools are generally much smaller than the theoretical maximum of $X \times \text{the-number-of-selected-runs}$ documents (usually about 1/3 the maximum size). Each pool is sorted by a document identifier so assessors cannot tell if a document was highly ranked by some system or how many systems (or which systems) retrieved the document.

Each document in the pool for a topic is judged for relevance by the topic author. Documents that are not in the pool (because no system ranked the document high enough) are assumed to be irrelevant to that topic.

3.3 TREC Collections as Laboratory Tools

To be viable as a laboratory tool, a test collection must reliably rank different retrieval variants according to their true effectiveness. Given the critical role relevance judgments play

in a test collection, it is important to assess the quality of the judgments created using the pooling technique. In particular, both the *completeness* and the *consistency* of the relevance judgments are of interest. Completeness measures the degree to which all the relevant documents for a topic have been found; consistency measures the degree to which the assessor has marked all the "truly" relevant documents relevant and the "truly" irrelevant documents irrelevant.

Inconsistency—the fact that different relevance assessors produce different relevance sets for the same topics—has been the main perceived problem with test collections since the initial Cranfield experiments [16, 7, 11]. The main gist of the critics' complaint is that relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [13]. Critics question how valid conclusions can be drawn when the evaluation process is based on something as volatile as relevance.

To study the effect of different relevance judgments on the stability of comparative evaluation results, NIST obtained multiple independent sets of assessments for two topic sets [19]. The results confirmed earlier work in that the assessments did differ, and different assessments caused the absolute scores of effectiveness measures to change. However, the changes are very highly correlated across systems. That is, if a particular system gets a relatively high score with a particular relevance judgment set, then it is very likely that the other systems will also get a relatively high score with that judgment set. As a result, the relative scores among systems are very stable, and the use of test collections to compare systems is supported.

Incompleteness is a relatively recent criticism of the Cranfield paradigm since the original test collections were complete. The concern is that systems that retrieve unjudged documents cannot be evaluated fairly because some of those documents might be relevant. In particular, critics worry that the evaluation scores for methods that did not contribute to the pools will be deflated relative to methods that did contribute.

It is true that pooling does not capture all of the relevant documents. For example, a test of the TREC-2 and TREC-3 collections demonstrated the presence of unjudged relevant documents [10]. In this test, relevance assessors judged the documents in new pools formed from the second 100 documents in the ranked results submitted by participants. On average, the assessors found approximately one new relevant document per run. The distribution of the additional relevant documents was roughly uniform across runs, but was skewed across topics—topics that had many relevant documents initially also had many more new relevant documents.

Zobel also found this pattern of un-judged documents in his analysis of the effect of pooling on retrieval results [24]. He showed that the pool depth and the diversity of the runs that contribute to the pools are important factors in the quality of the resulting test collection, but that with adequate controls during the pooling process the test collections can be used to fairly compare retrieval systems. He reached this last conclusion by introducing a test for the stability of retrieval results using collections created by pooling. A similar (more stringent) test has been used to analyze the TREC collections built since then. In this test, each run that contributed to the TREC pools is evaluated first using the official set of relevant documents

published for that collection, and then using set of relevant documents produced by removing the relevant documents uniquely retrieved by the organization that submitted the run. Not surprisingly, groups that perform manual runs generally have the largest number of uniquely retrieved documents and thus the largest percentage differences in mean average precision when evaluated by the two different relevance sets (the largest difference was about 10% for the TREC-8 ad hoc collection [21]). But given that the manual runs' contributions are in the pool, the difference in evaluation results for automatic runs is negligible. For the TREC-8 ad hoc collection, every automatic run that had a mean average precision score of at least 0.1 had a percentage difference of less than 1%, a smaller difference than is caused by changing relevance assessors.

In the end, the concern regarding completeness is a red herring. Since test collections support only comparative evaluations, the important question is whether the relevance judgments are unbiased. That is, it does not matter how many or how few judgments are made as long as the documents that are judged are not correlated with the documents a particular retrieval method will retrieve. Having complete judgments ensures that there is no bias in the judgments, but pooling with sufficiently diverse pools has been shown to be a good approximation.

4. LESSONS FOR MIR/MDL EVALUATION

The dominating factor in the success of an evaluation is the task description, by which I mean both a precise definition of what is to be done and the metrics used to score the quality of a response. Tichy [17] notes that

The most subjective and therefore weakest part of a benchmark test is the benchmark's composition...Hence, benchmark composition is always hotly debated.

and Hirschman [12] lists "intuitively understandable metrics that map to commercially significant problems" as one desirable feature of an evaluation. TREC benefited enormously from the fact that the Cranfield experiment paradigm was already accepted when TREC began; yet as TREC has evolved to include many different retrieval tasks it has had to face the task definition problem. This section examines whether some of the task definitions used in TREC may serve as models for MIR/MDL tasks.

4.1 Ad Hoc Searches

The Cranfield paradigm is insensitive to what constitutes a document and how information needs are expressed. A document can be any uniquely identifiable unit of information that is to be retrieved, and the information need can be expressed in whatever form intended users of the system would express requests. For example, the task in the TREC speech track used recordings of individual news stories from broadcast news reports as documents, and the TREC video track task used video shots as documents. The speech track used standard TREC textual topic statements to express the information need, while the topic statements in the video track consisted of text plus optional other objects such as speech recordings, still images, and video clips. In both tracks, the systems were to retrieve the documents that satisfied the information need

expressed in the topic statement, and assessors judged whether or not the retrieved documents did in fact meet that information need.

Some MIR problems seem to be equally good fit with the Cranfield paradigm. For example, Downie and Nelson give "In which compositions can we find the following note sequence anywhere in the composition?" and "Given a melody identify the work." as example MIR tasks of interest [9]. Such problem statements are neutral with respect to *how* the task should be accomplished, which provides lots of room for many approaches to the problem. Systems can return ranked lists of musical works that they believe meet the criteria, and (presumably) human assessors can judge whether or not the criteria are met. Of course, the particular implementation of such an evaluation will materially affect what can be learned from it. Even though the problem statement is neutral with regards to music representation, if the collection of musical works to be searched contains only MIDI files, then approaches based on full musical notation cannot be tested. Similarly, if the collection contains only monophonic music, then nothing will be learned regarding approaches for polyphonic music [4]. Further, the criteria used to determine a match will affect the kind of assessors that are required (and hence the cost of the evaluation). For example, criteria that involve technical musical distinctions will require trained musicologists as assessors.

4.2 Known-item Searches

Closely related to the ad hoc search task assumed by the Cranfield paradigm is a known-item task. In a known-item task the target of the search is a particular document that the searcher knows to exist in the collection and wants to find again. The task generally assumes that the searcher's recollection of the document is not perfect (otherwise, they'd be able to find the document easily!), so exact match is not a viable approach. The retrieval system's response is usually a ranked list of documents as in the ad hoc search, and the system is evaluated by the rank at which the target document is retrieved. Known-item searches have been used in TREC in the confusion track (so-called because the text of the documents were corrupted or "confused" as a result of being the output of an optical character recognition process), the speech track, and the video track.

The classic "query by humming" MIR problem is a good candidate for a known-item search task. As in the ad hoc search case, the particular make-up of the music collection and the kinds of information included in a query will determine the difficulty and salience of the evaluation. A known item task has the advantage of not requiring relevance judgments after the fact (though topic development is somewhat more complicated than in an ad hoc task), so it is relatively inexpensive to evaluate.

4.3 Searches That Don't Retrieve Documents

All of the tasks discussed so far assumed there was an identifiable unit of information that can act as a document. This assumption is not true in the TREC question answering (QA) track where systems return small text snippets that contain answers rather than entire documents. The size of the snippet is not a problem for the standard Cranfield evaluation, but the lack of an unique identifier is. The main benefit test collections provide is that they allow researchers to run their own experiments and receive rapid feedback as to the quality of

alternative retrieval methods. This not only advances the state of the art more quickly, but amortizes the cost of building the collection across a much larger base. Unfortunately, different QA runs very seldom return exactly the same answer strings, and it is quite difficult to determine whether the difference between a new string and a judged string is significant with respect to the correctness of an answer. Several partial solutions to this problem have been developed such as using word recall or answer patterns based on judged string to approximate judgments on new strings [2, 23]. While such approaches are better than nothing, more satisfactory solutions are required to realize the full benefits of benchmark tests for QA tasks.

It is not clear to me whether MIR/MDL evaluation has tasks analogous to the QA task in that human judgments for one set of results would not support the evaluation of a different set of results. If so, careful consideration needs to be given to how the results of such an evaluation would support further research.

5. CONCLUSION

The MIR/MDL research community has recognized a need to create an evaluation framework for the field. Common evaluation tasks—benchmarks—can be a powerful method for advancing the state of the art, but they need to be carefully constructed. A benchmark that represents a skewed view of the real-world problem can waste research time and resources. A single, static benchmark may cause an entire field to overfit to one particular data sample. Test collections are a kind of benchmark that has been used extensively for text retrieval evaluation. The forty year history of test collection use has demonstrated that test collections are viable laboratory tools. That is, retrieval results obtained from test collections are stable and the results are predictive of behavior in operational settings².

A test collection consists of a set of documents, a set of information need statements, and a set of relevance judgments that specify which documents meet which information needs. With appropriate definitions of document, information need, and need satisfaction, the test collection paradigm can cover a broad range of retrieval tasks, including some MIR tasks. By using the test collection paradigm for initial evaluation tasks in MIR/MDL, the field can benefit from the experience of the text retrieval community.

6. REFERENCES

[1] D. Banks, P. Over, and N. F. Zhang. Blind men and elephants: Six approaches to TREC data. *Information Retrieval* 1:7-34, 1999.

[2] E. Breck, J. Burger, L. Ferro, L. Hirschman, D. House, M. Light, and I. Mani. How to evaluate your question answering system every day ...and still get real work done. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, volume 3, pages 1495-1500, 2000.

[3] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In N. Belkin, P. Ingwersen, and M. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33-40, 2000.

[4] D. Byrd and T. Crawford. Problems of music information retrieval in the real world. *Information Processing and Management*, 38(2):249-272, 2002.

[5] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173-192, 1967. (Reprinted in *Readings in Information Retrieval*, K. Sparck-Jones and P. Willett, editors, Morgan Kaufmann, 1997).

[6] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 3-12, 1991.

[7] C. A. Cuadra and R. V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4):291-303, 1967.

[8] M. Dovey. ISMIR 2001 Resolution. <http://music-ir.org/mirbib2/resolution>, 2001.

[9] S. Downie and M. Nelson. Evaluation of a simple and effective music information retrieval method. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73-80, 2000.

[10] D. Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1-23, October 1996. NIST Special Publication 500-236.

[11] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37-49, 1996.

[12] L. Hirschman. Language understanding evaluations: Lessons learned from MUC and ATIS. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 117-122, Granada, Spain, May 1998.

[13] L. Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3-48, 1994.

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

[15] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1997.

[16] M. Taube. A note on the pseudomathematics of relevance. *American Documentation*, 16(2):69-72, April 1965.

² For example, basic components of modern commercial retrieval systems such as full-text indexing, term weighting, and relevance feedback were first developed using test collections.

- [17] W. F. Tichy. Should computer scientists experiment more? *Computer*, 31(5):32-40, May 1998.
- [18] C. van Rijsbergen. *Information Retrieval*, chapter 7. Butterworths, 2 edition, 1979
- [19] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697-716, 2000.
- [20] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002. To appear.
- [21] E. M. Voorhees and D. Harman. Overview of the eighth Text REtrieval Conference (TREC-8). In E. Voorhees and D. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1-24, 2000. NIST Special Publication 500-246. Electronic version available at <http://trec.nist.gov/pubs.html>.
- [22] E. M. Voorhees and D. Harman. Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3-35, January 2000.
- [23] E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200-207, July 2000.
- [24] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307-314, Melbourne, Australia, Aug. 1998. ACM Press, New York.