

Emphasizing the Need for TREC-like Collaboration Towards MIR Evaluation

Shyamala Doraisamy
Department of Computing
180 Queen's Gate
London SW7 2BZ
+44-(0)20-7594-8180
sd3@doc.ic.ac.uk

Stefan M Ruger
Department of Computing
180 Queen's Gate
London SW7 2BZ
+44-(0)20-7594-8355
srueger@doc.ic.ac.uk

ABSTRACT

The need for standardized large-scale evaluation of music information retrieval (MIR) and music digital library (MDL) methodologies is being addressed with the recent resolution calling for the construction of the infrastructure necessary to support MIR/MDL research. The methodology of our MIR study investigating the use of n-grams for polyphonic music retrieval has been based on a small-scale test collection of around 10,000 polyphonic MIDI files developed by us. A review of MIR studies show that a number of researchers have similarly developed individual state-of-the-art test collections for the purpose of metric scientific evaluation. However, there are a number of potential problems that are generic to these various small-scale test collections. These include the lack of consistency and the completeness of the relevance judgements used. This paper discusses a number of test collections developed by various researchers and ourselves, and describes how many of the limitations generic to all these studies could be overcome through the development of standardized large-scale test collections.

1. INTRODUCTION

With the advancement in multimedia and network technologies, the interest in music information retrieval (MIR) and music digital libraries (MDL) has grown noticeably over the past few years. There already exist an appreciable number of MIR systems that are commercially viable and of a high degree of sophistication. However, with the lack of standardized, generally agreed-upon test collections, tasks and metrics for MIR evaluation, researchers and developers are facing difficulties in benchmarking and endorsing the performances of their systems. The need for standardized large-scale evaluation of MIR methodologies has been identified only very recently with the resolution for the construction of the infrastructure necessary to support MIR/MDL research [1].

One of the approaches proposed towards MIR evaluation is the use of test collections based on the Cranfield and TREC models [1, 14] that have been used for many years in the text retrieval community. A test collection for information retrieval (IR) encompasses (i) a set of documents, (ii) a set of queries, and (iii) a set of relevance judgments, and aims to model real-world situations, so that one could expect the performance of a retrieval system on the test collection to be a good approximation of its performance in practice. Of the two series of studies conducted at Cranfield University, UK, it is the second series conducted in the

1960s that became the exemplar for experimental evaluation of IR systems (with the first conducted in the 1950s) [9,10]. The Cranfield 2 test collection consists of 1,400 documents, mainly in the field of aerodynamics with 221 search questions [26]. Because of computer storage and processing costs, it was not until the 1980s, however, that large-scale testing became possible. In the early 90s the Cranfield 2 paradigm was gradually replaced by TREC (Text REtrieval Conference), a major initiative in IR evaluation with an emphasis on large collection size and completeness of relevance judgements. It made possible the large-scale, robust evaluation of text retrieval methodologies and has been running successfully since [5,9].

An important factor in the success of an evaluation is the task description and the metrics used to score the quality of a response. "Intuitively understandable metrics that map to commercially significant problems" have been described as one desirable feature of an evaluation [5, 22]. Query by melody (QBM), a task useful to most classes of MIR users (music librarians, disc-jockeys, music scholars, etc.), has been the research focus of a large number of MIR studies. There are various interfaces to these: query by humming (QBH), text-boxes for contour or absolute note letter-names input, or a graphically visualized keyboard. In our study of full-music indexing of polyphonic music, we are concerned with two tasks: QBM for monophonic queries and query by example (QBE) for polyphonic queries [16].

Relevance judgments are what turn a set of documents and topics into a test collection. Deciding which documents are relevant given a particular query constitutes an important theoretical and practical challenge. Among the more immediate questions it raises are: would performances with the melody transposed or played in a different speed be considered similar? Must performances to be retrieved be in the same genre? Was the melody heard as part of the accompaniment? How does the computer intuitively decide which is the accompaniment? Can the melody be slightly varied? If yes, what is degree of variation allowed? Must the query be complete in structure musically (theme, motive, phrase, fugal subject, etc)? Is a query with an arbitrary number of notes possibly not complete musically, analogous to half a word or sentence with text? It would be difficult to define relevant documents if the query is not even considered valid. These are only a few of the issues that need to be addressed in order to define relevance, a notion that takes a central position in IR evaluation.

In developing a list of relevant documents for a query, TREC uses a pooling technique, whereby diverse retrieval systems suggest documents for human evaluation. Documents that are not in the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

pool, because all systems failed to rank the document high enough, are assumed to be irrelevant to the topic. Human assessors then judge the relevance of the documents in the pool. The quality of the judgements created using the pooling technique should be assessed, in particular with respect to their completeness and the consistency of the relevance judgements. Completeness measures the degree to which all the relevant documents for a topic have been found while consistency measures the degree to which the human assessor has marked all relevant documents as relevant and the irrelevant documents as irrelevant [5].

With MIR, one of the first steps taken towards standardized music test collections was to list candidate MIR test collections. These collections are useful by themselves for a number of research projects but would be even more useful if they were accompanied by a set of well-defined queries and relevance judgements. The Uitdenbogerd and Zobel collection is a notable exception as it does come with a set of (however incomplete) human relevance judgments [14].

The paper is structured as follows: Section 2 discusses some of the MIR studies that have based the evaluation on relatively large test collections. The discussion includes the relevance judgments and effectiveness measures used. Section 3 discusses the experiments of our MIR study on polyphonic music indexing and retrieval with n-grams using a test collection we developed. Section 4 discusses potential problems of the various test collections and re-emphasizes the need for TREC-like collaboration.

2. BACKGROUND

In this section, we discuss a number of MIR studies for which researchers have developed individual small-scale test collections (using between 3000 to 10,000 documents) for evaluation [3,12,21, 23, 24]. We will focus on issues such as complexity of music data and queries, relevance judgements and effectiveness measures used.

2.1 COLLECTIONS AND QUERIES

There are a number of additional problems that MIR researchers face when dealing with music data for computer-based MIR systems when compared to text. Music data could come as simple monophonic sequences, where a single musical note is sounded at any one time, or as polyphonic sequences, where several notes may be sounded at any one time. Music data is multi-dimensional with musical sounds commonly described in terms of pitch, duration, dynamics and timbre. Music data can be encoded in multiple formats: highly structured, semi-structured or highly unstructured. A few studies have included additional pre-processing modules to deal with these various aspects of music data, for the data collection and/or queries. The collection and queries used in the studies by Downie [3] and Sødning et al [23] were monophonic. Melody extraction algorithms were used for the studies by Uitdenbogerd [12] and Kosugi et al [21] to preprocess a polyphonic collection. The collection comprised the monophonic sequences obtained from the preprocessing step along with monophonic queries. The study by Pickens et al [24] used both polyphonic queries and source collection. The collection was encoded in a highly structured format and a

prototype polyphonic audio transcription system was integrated to transcribe polyphonic queries in the audio format.

Collection sizes varied between 3000 and 10,000 music files. The studies by Downie [3] and Sødning et al [23] used the NZDL [14] collection of about 10,000 folksongs in the monophonic format. Uitdenbogerd [12] and Kosugi et al [22] both used around 10,000 MIDI files where the former downloaded music of various genres from the Internet and the latter obtained the collection from a company in Japan. Kosugi et al [22] chose MIDI as the format for their collection as there is a large amount of MIDI available in Japan, where the popularity of karaoke ensures easy access to all the latest pop hits. Most karaoke recordings store the melody data on one MIDI channel, which makes it easy to recognize the melody [21]. The test collection developed by Pickens et al [24] used data provided by CCARH (<http://www.ccarh.org>). It consists of around 3000 files of separate movements from polyphonic full-encoded music scores by a number of classical composers (including Bach, Beethoven, Handel and Mozart). Three additional sets of variations, from which queries were extracted, were added with the final collection comprising a total of 3150 documents.

For query acquisition, approaches taken were either simulation/automatic or manual. Faced with the difficulty in obtaining real-world queries, many researchers simulate queries by extracting excerpts from pieces within the collection, and then using error models to generate erroneous queries. One such study is by Downie [3] and consists of two phases. In the first phase, 100 songs of a variety of musical styles were selected and queries of lengths 4, 6 and 8 were extracted from the incipits of each song. Thirty randomly selected pieces from the collection and a sub-string of length 11 from various locations in the piece were used in the second phase of the study. An error model based on the study by McNab et al [18] was used for error simulation.

Both the automatic and manual query acquisition approach were used in the study by Uitdenbogerd [12]. Automatic queries were selected from the collection by assuming that versions of a given piece formed a set of relevant documents. Versions were detected by locating likely pieces of music via the filenames and then verifying by listening to these pieces. One of the pieces from each set of versions was randomly chosen to extract an automatic query. Manual queries were obtained by asking a musician to listen to pieces that were randomly chosen from the set of pieces with multiple versions obtained from the automatic approach, and then to generate a query melody.

In the study by Sødning et al [23], 50 real music fragments were generated manually on a keyboard by a person with music knowledge before being transcribed into Parson's¹ notation to form the query set. 258 tunes hummed by 25 people were used as candidate queries for the study by Kosugi et al [21]. 186 tunes from these were recognized as melodies available in the database, and hence adopted as the query set. For the study by Pickens et al [24], the audio version of one variation was selected from each of the three sets of variations that had been added for the purpose of query extraction, was used as the query for the QBE task. They were unable to get human performances of all these variations.

¹ An encoding that reflects directions of melodies.

Instead, queries were converted to MIDI and a high-quality piano soundfont was used to create an audio “performance”.

Looking at a few test collections, we can see the diversity in size, genres, formats, complexity and query acquisition approaches that have already been used in MIR studies.

2.2 RELEVANCE JUDGEMENTS

TREC has almost always used binary relevance judgements (a document is relevant to the topic or not). There have been studies investigating the use of multiple relevance levels [8]. The most recent web track used a three point relevance scale: not relevant, relevant and highly relevant [6]. To overcome the difficulties in obtaining agreement on relevance, TREC uses the pooling technique to obtain a repository of candidate relevant documents and a number of human assessors to judge the relevance of these. In defining relevance for the assessors, the assessors are told: *Assume that you are writing a report on the particular topic. If the document would provide helpful information then mark the entire document relevant, otherwise mark it irrelevant. A document is to be judged relevant regardless of the number of other documents that contain the same information* [5].

With the need to evaluate MIR systems, a number of relevance definitions have been assumed. With the known item search used in the first phase of Downie’s [3] study, the document from which 100 query sequences were extracted was considered relevant and all remaining documents considered non-relevant. In the second phase, the set of relevant documents for a given query was defined as being the set of those songs in which the query’s progenitor string was found intact.

In the study by Uitdenbogerd [12], “automatic” and manual relevance judgements were used. Versions of a piece were considered to be relevant. Pieces were only considered to have distinct versions if there were obvious differences in the arrangement, such as (i) being in a different key, (ii) using different instruments or (iii) having differences in the rhythm, dynamics, or structure. All arrangements of the piece were assumed to be relevant versions and all other pieces assumed to be irrelevant, thus giving “automatic” relevance judgements. For manual relevance judgements, six judges were asked to listen to the pieces returned by the retrieval system for relevance assessment.

The pooling approach was adopted by Sødring et al [23] to obtain a set of relevant documents. The answer set obtained from submitting the 50 manually generated queries to their MIR system, was used as the relevant document set for the various experiments in their study. This approach was seen to be useful in generating a list of relevant documents for a set of queries without the need for human relevance judgements.

Audio queries were transcribed and submitted for retrieval for the study by Pickens et al [24] and Kosugi et al [21]. Being one of the first studies to use an audio polyphonic query, the study by Pickens et al [24] used one variation from each of the three query sets of variations as a query, and any variation within each corresponding set was considered relevant. With the monophonic hummed queries for the study of Kosugi et al [21], relevant documents were identified by song names.

In this section, we have shown that despite the difficulties in defining relevance for MIR where human music perception has to

be addressed, relevance had been defined within the scope of the various MIR studies, based on the need to evaluate the respective systems.

2.3 EVALUATION MEASURES

Many measures of retrieval effectiveness have been defined and a number of these have been adopted for MIR studies. All the measures are based on the notion of relevance. The measures assume that, given a document collection and a query, some documents are relevant to the query and others are not. The objective of an IR system is to retrieve relevant documents and to suppress the retrieval of non-relevant documents [9]. Notable performance measures that were used for the Cranfield tests and continue to be widely used in IR are precision and recall. Recall is the proportion of relevant documents retrieved and precision is the proportion of retrieved documents that are relevant. Formally, given a query q , a set of retrieved documents $A(q)$ and a set of relevant documents $R(q)$, then recall r and precision p are defined as

$$r = \frac{|A(q) \cap R(q)|}{|R(q)|} \text{ and } p = \frac{|A(q) \cap R(q)|}{|A(q)|}.$$

The official TREC reports several variants of precision and recall, such as the mean precision at various cut-off levels and a recall-precision graph. The mean average precision is often used as a single summary evaluation statistic [7]. Another measure that has been used is based on the rank of known-item search. The quality of the retrieval mechanism is judged by the reciprocal rank of the known item – e.g., if the known (and only relevant) item is retrieved at 5th rank, a quality of 0.2 would be assigned for this query. By repeating this process with many queries, a mean reciprocal rank (MRR) is obtained to assess a particular retrieval and indexing method, averaged over the number of queries. The MRR measure is between 0 and 1 where 1 indicates perfect retrieval.

With the complexities of music data [2], other benchmarking measures beyond precision and recall have been proposed. These include evaluating based on aspects of retrieval efficiency (e.g. speed of processing), software quality metrics and human computer interaction (HCI) and user interface (UI) features [13].

For the first phase of Downie’s [3] study, a modified measure of precision was used. Precision was defined as: $P = 1/\text{number of song titles retrieved}$. Queries were extracted from 100 songs. For the purpose of the study, a non-relevant hit was any song title retrieved other than that from which the query was extracted. The second study used normalized precision and recall. The normalized precision (NPREC) and normalized recall (NREC) metrics capture how closely a ranking system performs relative to the ideal by including in the calculation, information about the ranks at which relevant documents are listed. A NPREC or NREC value of 1 indicates that the ideal has been realized while a value of 0 indicates the worst case.

The NPREC and NREC metrics are defined as [25]:

$$\text{NPREC} = 1 - \frac{\sum_{m=1}^{REL} \log Rank_m - \sum_{m=1}^{REL} \log m}{\log(N!/(N - REL)!REL!)}$$

$$NREC = 1 - \frac{\sum_{m=1}^{REL} RANK_m - \sum_{m=1}^{REL} m}{REL(N - REL)}$$

Where N is the number of documents in the database, REL the number of relevant documents contained in the database, and $RANK_m$ the rank assigned to relevant document m [3, 11].

Other standard measures adopted for MIR studies include: (i) eleven-point precision averages (recall and precision can be averaged at fixed recall levels to compute an overall eleven-point recall-precision average) [12], (ii) precision at k pieces retrieved (number of relevant melodies amongst the first k retrieved) [12, 23], (iii) precision/recall graphs [23], (iv) mean average precision and mean precision at the top 5 retrieved documents [24].

For the study by Kosugi et al [21], where relevance was based on the song names, the percentage of songs retrieved within a given rank number formed the basis for their evaluation.

With relevance judgements defined to a certain extent within the scope of each particular study, standard evaluation measures (modified in some cases) have already proved useful for MIR evaluation. Whether or not such measures are useful for large-scale evaluation in a MIR context is a question that needs to be investigated further.

3. TEST COLLECTION

This section describes the series of experiments we performed to investigate the use of n -grams for polyphonic music retrieval. Upon surveying the candidate music test collection [14] for our MIR study, one that was closely appropriate was the one developed by Uitdenbogerd and Zobel. However, we had difficulties obtaining this collection as the web site had been discontinued and permission to use the downloaded files failed due to copyright problems. We describe our approach towards the development of a test collection that was needed for our study. In particular, the test collection, i.e. the set of documents, queries, and relevance judgements, used in the various experiments are discussed.

A collection of almost 10,000 polyphonic MIDI performances that were mostly classical music performances had been obtained from the Internet [<http://www.classicalarchives.com>]. These were organized by composers – Bach, Beethoven, Brahms, Byrd, Chopin, Debussy, Handel, Haydn, Liszt, Mendelssohn, Mozart, Scarlatti, Schubert, Schumann and Tchaikovsky. Other composers' works were organised in directories alphabetically (midi-a-e, midi-f-m, etc). The rest of the collection was categorized as aspire, early, encores and others. A smaller collection of around 1000 MIDI files of various categories of popular tunes - tv and movie themes, pop, oldies and folksongs was collected from the Internet (no longer available). Files that converted to text formats with warning messages on the validity of the MIDI file such as 'no matching offset for a particular onset', by the midi-to-text conversion utility, were not considered for the test collection. Various subsets of this collection formed the document-set for the different experiments. The various tasks and relevance used in these experiments are discussed.

3.1 EXPERIMENT 1 – PRELIMINARY INVESTIGATION

A preliminary study on the feasibility of our approach of pattern extraction from polyphonic music data for full-music indexing was performed. N -grams were constructed based on a gliding window approach using all possible patterns of polyphonic music data [15]. Data analysis was performed to study the frequency distribution for the directions and distances of pitch intervals and ratios of the onset time differences that occur within the data set. For query simulation, polyphonic excerpts were extracted from 30 randomly selected musical documents similar to the study by Downie [3]. The only relevant document for this type of query is the music piece from which the query was extracted, and not variants or otherwise similar pieces. In simulating a variety of query lengths, lengths of the excerpts extracted from the randomly selected files were of 10, 30 and 50 onset times. With no error models available with polyphonic music queries, the Gaussian error model was used in generating erroneous queries from the perfect queries extracted.

3.2 EXPERIMENT 2 – QUERY BY MELODY, FAULT-TOLERANCE AND COMPARATIVE STUDY

The second experiment was performed to test the feasibility of querying a polyphonic music collection with a monophonic sequence [17]. Monophonic queries are thought to resemble query by melody. In particular, we focus on QBH systems and the fault-tolerance of the n -gram approach was examined based on QBH error models. In order to simulate ad-hoc queries, where the collection is kept constant but the information need changes, we hand-crafted ten monophonic queries. These were popular tunes of various genres. The list of songs and relevant documents are listed in Table 1.

With pieces from the classical collection (Songs ID 1, 4, 6, 7, 8 and 10), using the filename and composer directory, one performance of each tune was identified. Each of these performance files was edited using a midi sequencer, jazz-4.1.3, to extract a polyphonic excerpt containing the theme. Pitches for the themes were referenced using the Dictionary of Musical Themes [20]. Using the retrieval approach and the optimal parameters identified from the first experiment, these polyphonic excerpts containing the theme were used as queries to form a relevant document pool. MIR studies that include the pooling approach to relevant document acquisition includes the studies by Uitdenbogerd [12] and Sødning et al [23]. The documents retrieved were listened to and its relevance was judged based on assumptions similar to Uitdenbogerd [12] (see section 2). From this polyphonic query excerpt, the theme was extracted manually as a monophonic sequence, using midi to text and vis-à-vis conversion utilities.

With the popular music pieces (Songs ID 3 and 9), only one version of each was available in the collection. These were used to extract monophonic queries using the midi sequencer. Versions for these were obtained by performing a search on the Internet using the same relevance assumption. Lastly, Happy Birthday (Song ID 2), assumed to be a tune that everybody knew, was added. This was not available in our collection at all, therefore as many versions possible, based on the relevance

assumption defined above, were searched for on the Internet and one of the versions with basic chords accompanying the tune was selected to extract the monophonic query. Query lengths varied between 15-25 notes for eight of the songs. The query for Beethoven's Symphony No. 5 had just 8 notes and that for Hallelujah was the most elaborate with 285 notes.

Table 1. Song list

Song ID	Song Title	No. relevant
1	Alla Turka (Mozart)	5
2	Happy Birthday	4
3	Chariots of Fire	3
4	Etude No. 3 (Chopin)	1
5	Eine kleine Nachtmusik (Mozart)	5
6	Symphony No. 5 in C Minor, (Beethoven)	8
7	The WTC, Fugue 1, Bk 1 (Bach)	2
8	Für Elise (Beethoven)	3
9	Country Gardens	2
10	Hallelujah (Händel)	7

This test collection enabled us to perform further investigation on the use of n-grams towards full-music indexing of polyphonic music and a comparative study of various n-gramming strategies. QBH error models were surveyed and used to investigate the fault-tolerance of the indexing approach. For performance evaluation, we used the precision-at-15 measure, in which the performance of a system is measured by the number of relevant melodies amongst the first k retrieved, with k=15 in our case.

3.3 EXPERIMENT 3 – ROBUSTNESS AND ENVELOPES

This study performed a more extensive investigation on the robustness of QBM and QBE tasks. The study also proposed and evaluated an approach to reduce the number of musical words generated with the n-gram approach to full-music indexing of polyphonic music [16]. Two types of queries were extracted from the music pieces: monophonic queries, where only one note per onset time was extracted and polyphonic queries, where a polyphonic subsequence of events was extracted from the music piece. The approach of extracting the highest pitch from the several possible pitches for each onset was the best of the several melody extraction algorithm investigated by Uitdenbogerd [12] and was therefore used to extract the monophonic queries as query melodies. The MRR measure was used as the evaluation metric for this experiment.

3.4 EXPERIMENT 4 – PROXIMITY ANALYSIS

This section describes the test collection development for our on-going work on proximity analysis. Term position information

with indexes is known to improve the retrieval performance. An approach to obtain polyphonic term positions of “overlying” musical words based on the nature of polyphonic music data is introduced. Query formulation using proximity operators and a ranking function that addresses the adjacency and concurrency of musical words generated from polyphonic data for better retrieval precision are being investigated.

In order to evaluate this approach, we have extended the query set to 50 queries (40 added to the 10 used in the third experiment). These were thought to be sufficient, based on the notion of using 50 topics from which queries are generated with TREC. These additional queries were considered as queries that were amongst the ‘pop’ of classical music. This query list can easily be extended to a more comprehensive list, possibly including real-world queries such as collection of queries from music libraries, music stores, etc [2]. Query acquisition from a Dictionary of Themes [20] we think is a useful repository. The query list, that comprises works of various composers and periods of music within the scope of tonal music, is deemed sufficiently comprehensive for the QBM task. A separate task may need to be defined for contemporary music with its own set of specific problems, such as difficulty to define a melody in this class of music [19].

Exhaustive judging, where relevance is determined for each document, is feasible for a collection of this size. The relevance assumption of versions used by Uitdenbogerd [12] was adopted and the first few seconds of each performance were listened to for a judgement to be made. This seemed a reasonable approach, as the author was sufficiently familiar with the query melodies to be able to recognize a performance within the first few seconds.

4. TREC-LIKE COLLABORATION

It is clear from the MIR studies discussed in Section 2 and our work in Section 3, that the use of test collections has enabled MIR researchers to evaluate their work. Standardized test collections have been a useful benchmark for text retrieval evaluation for around 40 years now [5]. Manual relevance judgements, the pooling approach, known item searches, relevance by song titles/filenames and exhaustive judging have all been used with MIR test collections. All of these approaches, however, have potential problems such as completeness and consistency of the relevance judgements. These problems have been addressed with the large TREC collections and will need to be addressed when moving away from the small-scale test collections towards large-scale evaluation.

With so much research evaluated using test collections, the stability of the test collection is an important issue and has been addressed by TREC. Relative effectiveness of two retrieval strategies should be insensitive to slight changes in the relevant document set in order to reflect the true merit of the retrieval strategy being evaluated. The reasons as given by [4] for stability of system rankings despite differences in relevance judgements have been further discussed by [7]. The reasons for stability are as follows:

1. Evaluation results are reported as averages over many topics
2. Disagreements among judges affect borderline documents, which in general are ranked after documents that are unanimously agreed upon

3. Recall and precision depend on the relative position of the relevant and non-relevant documents in the relevance ranking, and changes in the composition of the judgement sets may have only a small effect on the ordering as a whole

It has been argued that the third reason may not apply to the large collection sizes of TREC where there could be hundreds of relevant documents. It has been shown, however, that first and the second reason appear to hold for the TREC collections [7]. In the context of MIR, it is unclear to what extent the second reason will prove valid, for it could be argued that human music perception would generate much larger differences in relevance judgments. This question clearly needs to be investigated further.

The investigation into the stability of system rankings with different sets of relevance assessments was carried out by NIST with the following tests [7]:

1. The use of the overlap of the relevant document sets to quantify the amount of agreement among different sets of relevance assessments [4]. Overlap is defined as the size of intersection of the relevant document sets divided by the size of the union of the relevant document sets
2. As a different view of how well assessors agree with one another, one set of judgements, say set Y, can be evaluated with respect to another set of judgements, set X. Assume the documents judged relevant in set Y are the retrieved set; then the recall and precision of that retrieved set using the judgements in X can be calculated.
3. The correlation can be quantified by using a measure of association between the different system rankings. A correlation based on Kendall's τ as the measure of association between two rankings can be used. Kendall's τ computes the distance between two rankings as the minimum number of pair-wise adjacent swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is -1.0 , and the expected correlation of two rankings chosen at random is 0.0.

When looking at completeness, it is also necessary to assess the degree of selection bias² that occurs. Relevance judgments need to be unbiased, i.e. it does not matter how many or how few judgments are made, but the documents that are judged should not be correlated with the documents in a particular retrieval method. Having complete judgments ensures that there is no selection bias, but pooling with sufficiently diverse pools has been shown to be a good approximation [5].

TREC-like collaboration is clearly needed to conduct the extensive tests required to address stability issues of MIR test collections and to obtain sufficiently diverse pools. Such tasks are formidably difficult for individual researchers to accomplish at a smaller scale.

² Selection bias occurs when the subjects studied are not representative of the target population about which conclusions are to be drawn.

5. CONCLUSION

With no available standardized test collection for MIR evaluation, we have shown how individual researchers have developed state-of-the-art test collections for the purpose of metric scientific evaluation. However, there are a number of potential problems that are generic to these various small-scale test collections and we believe that these can be overcome through TREC-like collaboration. Collaboration could alleviate problems in the following areas:

1. Resources (Collection and queries)

The collection sizes used are a fraction of real-world music repositories [2] and much larger collection sizes are needed. The difficulties in query acquisition may be overcome obtaining real-world queries such as those available with music libraries, radio stations, scanning or encoding documented themes. Extensive studies in obtaining real-world error models of music queries would be one approach towards generating large music query repository. A collaborative effort would be needed to identify potentially relevant documents (via pooling) and more resources to assess the relevance of the pooled documents. Extensive testing of the stability of the system rankings is required. A test bed created in this manner and made available would further the whole research field of Music IR

2. Relevance

Perhaps within the QBM task, relevance assumptions such as those already used in the studies thus far could be expanded. A comprehensive representation of user classes may be needed to reach a consensus on the relevance based on a task. Various tasks need to be identified and the relevance defined accordingly. Relevance based on a scale is one possibility, with human musical perception of similarity is notoriously difficult to model.

3. Copyright

TREC mainly uses old newspaper articles, which have no real commercial value, and hence distribution is not problematic. Music pieces have a value, and the test bed would need to be protected, either through drastic licenses specifying the legal use of the data, or perhaps by storing the test collection in a centralized secure environment that offers computing services such as the downloading of search engine indexing and retrieval code, which is then executed at the repository and receives the ranked lists. Alternatively, one could get preprocessed "features" of music pieces which are commercially not relevant (not the original music piece from the audio): a midi-like representation, the volume footprint, a rhythm or pitch extract etc, or the test collection consists of "half" music pieces, the first 60 seconds of each 120 seconds segment of the music piece etc.

4. Generic modules

Collaboration does not have to end with the creation of test collection. With the complexities of music data, a collaborative effort to develop modules that create features, extract melodies, preprocess data that requires expertise from various disciplines would certainly ease MIR research tasks. For example, groups with an expertise in IR of symbolic representation might well benefit from preprocessing that translates raw audio into symbolic forms such as MIDI.

6. ACKNOWLEDGMENTS

This work is partially supported by the EPSRC, UK.

7. REFERENCES

- [1] J. Stephen Downie, *Panel on Music Information Retrieval Evaluation Frameworks*, 3rd International Conference on Music Information Retrieval, ISMIR2002, Paris, France, pgs 303-304.
- [2] J. Stephen Downie, *Music Information Retrieval*, Annual Review of Information Science and Technology 37:295-340.
- [3] J. Stephen Downie, *Evaluating A Simple Approach to Music Information Retrieval: Conceiving Melodic N-Grams as Text*, PhD Thesis, University of Western Ontario, 1999.
- [4] Lesk, M., & Salton, G, *Relevance Assessments and Retrieval System evaluation*. Information Storage and Retrieval, 4, 343-359, 1969.
- [5] Ellen M. Voorhees, *Wither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC*, Panel, JCDL 2002 Workshop on MIR evaluation.
- [6] Ellen M. Voorhees, *Evaluation by Highly Relevant Documents*, SIGIR 01, pgs 74-81.
- [7] Ellen M. Voorhees, *Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness*, Information Processing and Management, 36: 697-716, 2000.
- [8] Amanda Spink and Howard Greisdorf, *Regions and Levels: Measuring and Mapping Users' Relevance Judgements*, Journal of the American Society for Information Science and Technology, 52(2): 161-173, 2001.
- [9] Stephen P. Harter and Carol A. Hert, *Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods*, Annual Review of Information Science and Technology, Volume 32, 1997
- [10] Cyril Cleverdon, *The Significance of the Cranfield Tests on Index Languages*, SIGIR '98.
- [11] C.J. van Rijsbergen, *Information Retrieval*, online book.
- [12] Alexandra Uitdenbogerd, *Music Information Retrieval Technology*, Phd Thesis, Royal Melbourne Institute of Technology, 2002.
- [13] *The MIR/MDL Evaluation Project White Paper Collection*, Edition #2, <http://www.music-ir.org>
- [14] Don Byrd, *Candidate Music Test Collections*, Background document for ISMIR 2000 on Music Information Retrieval Evaluation, The First International Symposium on Music Information Retrieval, ISMIR 2000, Plymouth, Massachusetts, USA, Oct 23rd – 25th 2000.
- [15] Shyamala Doraisamy and Stefan R ger, *An Approach Towards A Polyphonic Music Retrieval System*, 2nd International Symposium on Music Information Retrieval, ISMIR 2001, Indiana, USA, pgs 187-193.
- [16] Shyamala Doraisamy and Stefan R ger, *Robust Polyphonic Music Retrieval with N-Grams*, Journal of Intelligent Information Systems, 21:1, 53-70, 2003.
- [17] Shyamala Doraisamy and Stefan R ger, *A Comparative and Fault-Tolerance Study of the Use of N-Grams with Polyphonic Music*, 3rd International Conference on Music Information Retrieval, ISMIR2002, Paris, France, pgs 101-106.
- [18] Rodger J. McNab, Lloyd A. Smith, Ian H. Witten, Clare L. Henderson and Sally Jo Cunningham, *Towards the Digital Music Library: tune Retrieval from Acoustic Input*, DL '96, Bethesda MD, USA.
- [19] Alain Bonardi, *IR for Contemporary Music: What the Musicologist Needs*, ISMIR 2000.
- [20] Harold Barlow and Sam Morgenstern, *A Dictionary of Musical Themes*, London: Ernest Benn, 1949.
- [21] Naoko Kosugi, Yuichi Nishihara, Tetsuo Sakata, Masahi Yamamuro and Kazuhiko Kushima, *A Practical Query-By-Humming for A Large Music Database*, ACM Multimedia 2000, Los Angeles, CA., Nov. 2000.
- [22] L. Hirshman. *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.
- [23] Thomas S dring and Alan F. Smeaton, *Evaluating a Music Information Retrieval System – TREC Style*, Panel Discussion, 3rd International Conference on Music Information Retrieval, ISMIR2002, Paris, France.
- [24] Jeremy Pickens, Juan Pablo Bello, Giuliano Monti, Tim Crawford, Matthew Dovey, Mark Sandler and Don Byrd, *Polyphonic Score Retrieval Using Polyphonic Audio Queries: A Harmonic Modeling Approach*, 3rd International Conference on Music Information Retrieval, ISMIR2002, Paris, France pgs 140-149.
- [25] Salton, Gerard, and Michael J. McGill. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.
- [26] Cyril Cleverdon, *Evaluation Tests of Information Retrieval Systems*, Journal of Documentation, Vol 26, No. 1, March 1970, pgs 55-67.