

A Task-Oriented Approach for the Development of a Test Collection for Music Information Retrieval

Massimo Melucci

Department of Information Engineering
University of Padova
Via Gradenigo 6/a, 35131 Padova, Italy
+39-049-827-7500

massimo.melucci@unipd.it

Nicola Orio

Department of Information Engineering
University of Padova
Via Gradenigo 6/a, 35131 Padova, Italy
+39-049-827-7500

nicola.orio@unipd.it

ABSTRACT

This paper addresses the design of a test collection for CBMR, by applying a task-oriented approach. After having created a test document set, it is proposed to provide assessors with a number of topics and to build musical queries starting from music documents, or documents excerpts, being particularly relevant to the topic. Relevance judgments are then built based on the similarity between this document and the other documents in the collection. Being far to solve all the problematic issues related to the design of a test collection, this paper aims to give some insights and procedures to carry on the test collection design process, which requires many human resources and different types of expertise.

1. INTRODUCTION

Content-Based Music Retrieval (CBMR) is a quite new field in Information Retrieval (IR) research. Given an end user's query expressing an information need, the aim of a CBMR system is to retrieve music documents being relevant to the information need on the basis of their musical content, without using textual bibliographic data. Insofar, a number of approaches have been proposed and a few prototypes already exist, e.g. [2, 3]. In our presentation we will address the design and implementation of a *test collection* for CBMR.

A test collection essentially consists of three parts: a set of *documents*, a set of *queries*, and a set of *relevance judgments*. The evaluation model based on a test collection is a valid and viable solution to make techniques and systems comparable under the same conditions. Indeed, test collections have permitted researchers in IR to significantly improve average performance of the systems tested at TREC [7]. The use of test collections is somehow a necessary condition to test and compare different CBMR systems, especially at the current state-of-the-art that includes different approaches, techniques and researchers working in the field.

The peculiarities of music language make the development of a test music collection a rather different task from the development of a test collection of textual documents. The concept of information need itself may vary dramatically depending on the end user. Music has an auditory and temporal nature that inevitably conditions retrieval experiments, while its characteristics require us to deal with requests and relevance judgments covering all possible musical dimensions, such as melody, harmony, rhythm, or structure. While test collections

permit researchers to compare systems, the user-centered approach to evaluation would permit researchers to analyze the aspects being related to the mentioned musical dimensions.

Our contribution aims to present a methodology to build a test collection that takes into account the characteristics of music with respect to IR.

2. DESIGN OF A TEST COLLECTION

Determining test collection features has been a cornerstone since early "classic" textual IR experimentation. The vast and well established literature on the design of test collections – see for instance [5] and [6] – helps define standard procedures to set different test collections for diverse application domains, media and languages. However, there are these standard procedures need to be adapted to specific data type or task. For example, a task-oriented approach of image retrieval for the building of a test collection has been proposed in [4]. The schema is the following: A professional illustrator is asked to perform an illustration task, that is to retrieve a set of photos that can be related to an illustration idea; he retrieves the photos by compiling a textual query describing the idea and he chooses the ones considered as pertinent; retrieved photos judged as relevant by the user are then considered as relevant to the query.

We propose a task-oriented approach to the creation of a test collection for CBMR. We will discuss the analysis of requirements, and specifically of: the set of test documents; the various types of end users and of information needs; the topics representing the various types of information needs.

3. REQUIREMENT ANALYSIS

We started the requirement analysis by collecting a set of documents. The choice of collecting documents before topics, and therefore before analyzing topics and users, has been driven by the need to precisely define a representative sample of the domain to which we were dealing with in terms of quantity, format, and source of documents.

According to [5], a test document set should show both homogeneity and variety, which can be obtained by organizing it as a collection of sub-collections. Sub-collections may be related to music genres, which in turn are partially related to the degree of complexity and to end users. With regards to the size of the document set, it should be noted that the number of objects to be indexed and retrieved has different aspects from text retrieval: music documents can be long and include many different parts which may be managed in different ways. The differences between sub-collection sizes should be both the degree of

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.

diversification within each musical genre, by emphasizing those being more prolific and heterogeneous, and the potential interest of the end users, as can be observed from the proportion of music works of each type being publicly available on the Web. We propose the following list of genres and dimensions of the sub-collections:

1. classical – 30%,
2. jazz, blues – 20%,
3. pop, rock – 30%,
4. folk, ethnic – 20%.

It is important to identify which are the possible types of the end users accessing an IR system. We propose to classify end users by applying a criterion based on the level and type of expertise in music. A possible classification may be include:

- Base users are fond of music, fans or more generally people that listen music for pleasure or hobby without necessarily a deep knowledge of the genre of the music they listen.
- Intermediate Users are music reviewers, music critics, or soundtrack composers, who have a good expertise in music in relation with the professional tasks they have to carry out.
- Expert Users are musicians, composers, scholars, and all people with a deep, both theoretical and practical knowledge in music.

Table 1. Allocation of search tasks to the different types of end user

Search task	User types		
	base	intermediate	expert
melody	•	•	•
form	•	•	•
orchestration		•	•
Harmony		•	•
structure			•
rhythm			•

As a first approximation, the information requirements which an end user can search for may be summarized by the following list of tasks: author, genre or title; melody, e.g. query-by-humming; orchestration and lead instruments; music form; rhythm; harmonic progression; music structure. From the proposed list, it is possible to draw a cross-reference table between search tasks and user types as we have done in Table 1. Following the guidelines reported in [4], a solution to the problem of defining the test topics can be described as follows: (i) select a group of users with different levels of expertise to be provided with a set of test topics covering all the music dimensions; (ii) prepare some *textual* topics for each category of users, according to their level of expertise.

The provision of a textual topic to the user is an important design choice: It aims to make the representation of the information need clearer than the representation that would be possible if a musical topic were provided. The musical part of the topic should however

be an integral part of the topic since it carries some information that can hardly be represented in textual form and because it will eventually be the musical query to be used for experimentation. Other textual fields can support the representation of the information need, e.g. to indicate which dimensions should be considered when assessing relevance. The end users are then asked to: (i) choose one music work from the provided database to be used as a musical query, the query can be both the complete work and an excerpt; (ii) choose additional works being "similar" to the query and relevant to the same information need.

We propose a TREC-style scheme to represent information needs as topics. Topics being designed for CBMR experimentation are virtual topics, because they are used to produce a musical query: It is the musical query that will be submitted to an experimental system, whereas the textual fields are ignored at query generation time, yet still used to formulate relevance judgments. Then, a topic for a test music collection can be structured as follows:

```

<num>      identification number
<title>    title, e.g. the title of a music work that will be either
            the eventual music query or a music work
            representing the query
<context>  query context, i.e. whether search should be
            performed by paying attention to one or more music
            dimensions
<desc>    complete description, if necessary
<narr>    description of the features so that a document is
            considered as relevant to the information need

```

With regards the preparation of the relevance judgments, should regard all the dimensions, using a non-binary scale to represent the multidimensional and subjective nature of the process of relevance assessment in the musical context. With regards the assignment of topics to users, one could follow the procedural scheme proposed in [6] or the pooling method proposed to evaluate the retrieval effectiveness if very large document databases are used in laboratory-based experiments, as done at TREC [7]. Clearly, the types of users should be taken into account regardless to the specific method being employed.

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4. FUTURE WORK

This paper reports some results of a larger project on test collection design and implementation being in progress. We are working on the scheme to assign topics to assessors and on the database schema representing a set of test collections.

5. ACKNOWLEDGMENTS

We thank Tommaso Gobbato for discussion, cooperation and support.

6. REFERENCES

- [1] S. Benedusi. *Progetto e realizzazione di una base di dati per la gestione di collezioni sperimentali di Information Retrieval* (Design and implementation of a database to manage Information Retrieval test collections). Tesi di laurea

- (laurea *thesis*), supervisor: Maristella Agosti, Università di Padova, Facoltà di Ingegneria, Padova, Italy, 1998/1999. In Italian.
- [2] R. J. McNab, L. A. Smith, D. Bainbridge, and I. H. Witten. The New Zealand Digital Library MELody inDEX. Technical Report may97-witten, D-Lib Magazine, May 15, 1997.
- [3] M. Melucci and N. Orio. SMILE: a system for content-based musical information retrieval environments. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO) Conference*, pages 1246-1260, Paris, France, April 2000.
- [4] E. Sormunen, M. Markkula, and K. Järvelin. The perceived similarity of photos – A test collections based evaluation framework for the content-based image retrieval algorithms. In S.W. Draper, M.D. Dunlop, and C.J. van Rijsbergen, editors, *Proceedings of Mira*, Evaluating Interactive Information Retrieval, Glasgow, Scotland, UK, April 1999. Electronic Workshops in Computing. <http://www.ewic.org.uk/>.
- [5] K. Sparck Jones and C.J. van Rijsbergen. Information Retrieval test collections. *Journal of Documentation*, 32(1):59-75, March 1976.
- [6] J. Tague-Sutcliffe. The pragmatics of Information Retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467-490, 1992.
- [7] E. Voorhees and D. Harman, editors. *Special Issue on the Sixth Text Retrieval Conference (TREC-6)*, volume 36(1) of *Information Processing & Management*, 2000.