

Comparison of User Ratings of Music in Copyright-free Databases and On-the-market CDs

Keiichiro Hoashi
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka
Saitama 356-8502 Japan
hoashi@kddilabs.jp

Kazunori Matsumoto
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka
Saitama 356-8502 Japan
matsu@kddilabs.jp

Naomi Inoue
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka
Saitama 356-8502 Japan
inoue@kddilabs.jp

ABSTRACT

We have been conducting research to develop a music information retrieval system which retrieves music based on user preferences. In order to conduct evaluation experiments, it is necessary to accumulate an experiment set of music data, and collect user ratings of each music data included in this data set. Two music data sets were prepared for user evaluation: the RWC Popular Music Database, which is a collection of 100 copyright-free pop songs, and a collection of songs recorded in on-the-market CDs. Comparison of user ratings towards the songs in each data set shows that user ratings of the RWC songs were far lower than that of the songs in the other music data set. Based on this experience, the artists believe that the construction of a standard list of on-the-market CDs is necessary for the development of an entertaining MIR system.

General Terms

Experimentation.

1. INTRODUCTION

The main task of conventional music retrieval systems is to retrieve a music data, which matches the request of a user. The importance of such systems is expected to increase due to the rapid spread of digital music data formats such as MP3. However, conventional systems can only be used to search a particular song from a database.

We have been conducting research regarding a music information retrieval (MIR) method based on music preferences of users[1][2]. Such a system will enable users to *discover* good songs which they have never heard before, from a large music database.

User preferences of music songs are necessary to evaluate the effectiveness of proposed algorithms. Therefore, we collected user ratings for songs included in two music data sets: one is a collection of copyright-free pop songs, and the

other is a collection of songs from CDs on the market. A significant difference was observed between the user ratings of the songs in the two data sets.

Based on the results of user rating data collection experiments, we make a proposal to the MIR community in this paper to build a standard collection of on-the-market CDs which can be used for MIR-related research purposes. In the next section, a brief explanation of our MIR algorithm is presented. This explanation is followed by a description of the music data sets used for the user rating data collection experiments, and an explanation of the experiments. Our proposal follows the experiment descriptions.

2. MUSIC INFORMATION RETRIEVAL BASED ON USER PREFERENCES

In this section, we will make a brief explanation of our research, which projects the development of an MIR algorithm based on user preferences.

2.1 Tree-based vector quantization

Our MIR algorithm is based on the tree-structured vector quantization method (TreeQ), developed by Foote[3]. The approach of the TreeQ method is to “train” a vector quantizer instead of modelling the sound data directly. Figure 1 illustrates the concept of the TreeQ method.

As illustrated in Figure 1, each audio datum in the training data set, which is a collection of audio data associated with a class such as artist or genre, is first parameterized into a spectral representation, by calculating mel-frequency cepstral coefficients (MFCC)[4]. More specifically, each audio waveform, sampled at 44.1 kHz, is transformed into a sequence of 13-dimensional feature vectors (12 MFCC coefficients plus energy).

Once all training data has been parameterized by the calculation of MFCC, a quantization tree is generated offline based on these MFCCs, and category data labeled to each item in the training data set. The resulting tree is optimized so that it attempts to put samples from different training classes into different bins (leaves) as much as possible. A histogram of an audio file can be generated by looking at the relative frequencies of samples in each quantization bin. The relative frequency can be considered as the probability of a data sample to end up in a certain leaf. For example, if 20 of 100 samples input into the quantization tree are classified in leaf i , the relative frequency of leaf i is 0.20. If the resulting histograms are considered as vectors, typical vector similarity measures such as the cosine measure can

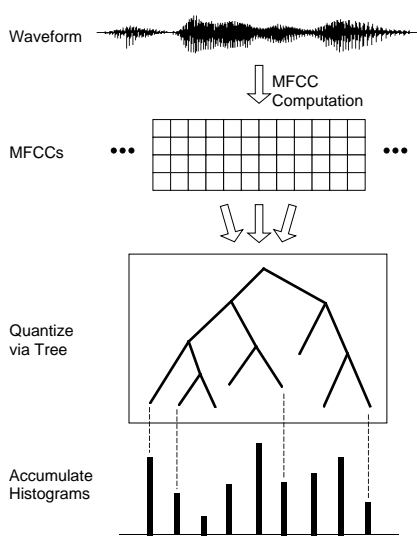


Figure 1: Outline of tree-structured vector quantization method

be applied to calculate the similarity between any incoming audio data and category vectors. This method has been used for music and audio retrieval experiments. In Reference [3], experiments were conducted to retrieve short audio data such as oboe sounds and human laughter. The same paper also reports experiments to retrieve music data of a specified musical genre, such as jazz, pop, rock, etc. as possible.

Our method is to associate user preferences as category information of the training music data set. For example, N “good” songs and M “bad” songs are input to the tree generation process to build the VQ tree. Next, user profiles, i.e., vectors which represent “good” and “bad” songs, are generated by inputting all “good” or “bad” songs through the VQ tree. Vectors of all other music data are also generated by inputting the data through the VQ tree. Scores of each music data are calculated by measuring the vector similarity between each data and the user profiles.

2.2 Relevance feedback

Due to the complexity and ambiguity of music and users’ musical preferences, and the limited amount of learning data, it is not realistic to expect the previously described music retrieval methods to achieve satisfactory performance consistently for all users. Therefore, there may be situations where users are unsatisfied with the system’s retrieval results, and feel the need to provide additional information regarding their preferences. The most efficient way to implement this is to collect relevance feedback from the user, and update the user profile based on the collected relevance feedback information, which is a widely used method in text IR.

As described in Section 2.1, each category C is expressed by a vector $\vec{C} = (c_1, \dots, c_n)$, where n is the number of histogram bins, and c_i is the relative frequency of samples in bin i . Since c_i is considered as the probability of a training data sample to be categorized in bin i , c_i can be calculated by the following formula:

$$c_i = \frac{|c_i|}{\sum_{i=1}^n |c_i|} \quad (1)$$

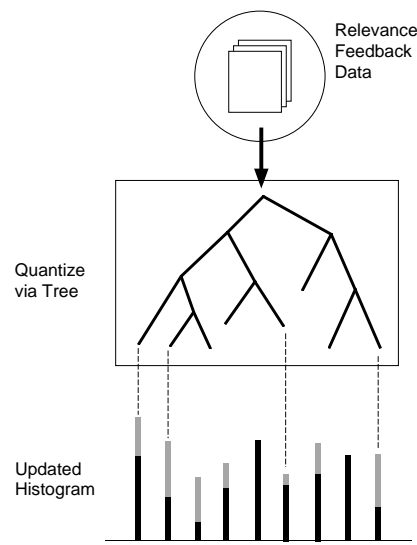


Figure 2: Outline of relevance feedback method

where $|c_i|$ expresses the number of data samples classified to bin i in the learning phase.

Relevance feedback is implemented by adding “relevant” data to each category vector. For example, if a song retrieved by the pilot search is preferred by the user, this song can be considered as relevant to category C_g . If the relevant song K_g is expressed as $\vec{K}_g = (k_{g1}, \dots, k_{gn})$, and the updated category vector is expressed as $\vec{C}'_g = (c'_{g1}, \dots, c'_{gn})$, relevance feedback is implemented by the following formula:

$$c'_{gi} = \frac{|c_{gi}| + |k_{gi}|}{\sum_{i=1}^n (|c_{gi}| + |k_{gi}|)} \quad (2)$$

In other words, the updated category vector is obtained by accumulating the relevant data set on the original category histogram. This concept is illustrated in Figure 2.

Songs which the user dislikes can also be used to update category C_g , as in the γ factor of Rocchio’s algorithm[5], where information derived from non-relevant documents are applied to relevance feedback. The following formula defines the implementation of a non-relevant song K_b to update category vector \vec{C}_g .

$$c'_{gi} = \frac{|c_{gi}| + |k_{gi}| - |k_{bi}|}{\sum_{i=1}^n (|c_{gi}| + |k_{gi}| - |k_{bi}|)} \quad (3)$$

As in Rocchio’s algorithm, information from non-relevant data is simply subtracted from the original vector. The resulting bin value is set to 0, if the bin value resulting from Formula 3 is negative. A similar approach can be implemented to update category vector \vec{C}_b , by considering “bad songs” as relevant, and “good songs” as non-relevant. This method is expressed by the following formula:

$$c'_{bi} = \frac{|c_{bi}| + |k_{bi}| - |k_{gi}|}{\sum_{i=1}^n (|c_{bi}| + |k_{bi}| - |k_{gi}|)} \quad (4)$$

The updated category vectors are then used to recalculate scores of songs in the test data set. Details of the evaluation experiments of our methods are to be published in Reference [2].

3. MUSIC DATA SETS

A collection of user ratings for songs is necessary to evaluate the previously described MIR algorithms. We have constructed a user rating data collection based on two music data sets: the *RWC Music Database: Popular Music*, and a collection of songs recorded in CDs on the market. The following sections describe the two music data sets.

3.1 RWC Music Database: Popular Music

“RWC Music Database: Popular Music” is a part of a music data collection available to music researchers[6]¹. The RWC Popular Music Database consists of 100 songs. All songs were composed and recorded for the purpose of inclusion in the database. Furthermore, in order to represent the various styles of music, the developers of the database prepared as many professional composers, lyric writers, arrangers, singers, instruments, etc.

3.2 HMV data collection

The other music data collection was constructed based on weekly ranking data of CD albums at HMV Japan², a major Japanese online music store. HMV Japan provides an archive of weekly CD sales rankings on their Web site, which was used to accumulate the list of CDs used for our experiments.

First, all CD albums which were ranked in the weekly top 10 rankings in the year 2001 was accumulated. We divided this CD collection into two subsets: CDs ranked from January to June 2001, and from July to December 2001. Furthermore, all “best” albums and “compilation” albums (i.e., collections of hit songs by various artists) were omitted from the list, in order to reduce the number of highly popular songs. Next, all songs contained in the resulting CD collections were extracted to construct the experiment data set. As a result, a total of 756 songs were derived from the Jan-Jun data set, which consists of 60 CDs, and 812 songs were derived from the Jul-Dec data set, which consists of 61 CDs.

4. USER RATING COLLECTION EXPERIMENT

4.1 Method

In order to collect user ratings data, 12 subjects applied subjective ratings ranging from 1 to 5 (Bad:1 ~ Good:5) for all songs in each experiment data set. Songs of the RWC Popular Database were rated first, followed by the two HMV data sets. Songs were presented to each subject in random order within each data set. Furthermore, metadata information such as song titles and artist names were removed so that the subjects could not rate songs without actually listening to them. However, the subjects were allowed to fast-forward (or rewind) through each song to listen to any portion of the song.

4.2 Results

The ratio of all user ratings for each data set is written in Table 1.

¹Distribution of the RWC Music Database outside of Japan has recently been started.

²<http://www.hmv.co.jp/>, Contents in Japanese only.

Table 1: Ratio of user ratings per music data set

Rating	RWC	HMV(Jan-Jun)	HMV(Jul-Dec)
5	7.3%	14.0%	13.8%
4	20.7%	26.4%	25.7%
3	25.7%	30.9%	30.1%
2	28.8%	21.5%	21.4%
1	17.5%	7.2%	9.0%

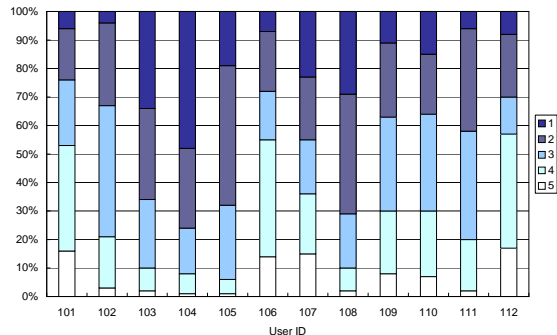


Figure 3: Ratio of user ratings for RWC data

Results in Table 1 clearly show that user ratings are generally higher for the HMV music data, compared to the RWC data. If songs with ratings (1 or 2), 3, and (4 or 5) are regarded as “bad”, “fair”, and “good” songs, respectively, the average ratio of “good” songs in the HMV data set is approximately 40%, while the same ratio for the RWC data set is 28%. Furthermore, the ratio of “bad” songs in the HMV set is just below 30%, while the “bad” song ratio for the RWC set is about 44%.

For further analysis of user ratings, the ratio of ratings in the RWC and HMV data sets for each user is illustrated in Figures 3 and 4, respectively.

Figures 3 and 4 show the wide variety of ratings between subjects. However, these results also show that not one subject has provided higher ratings to the RWC data compared to the HMV data. Furthermore, there were 4 subjects who rated less than 10% of all RWC songs as “good” songs.

Rating results for the RWC data set made it difficult to conduct MIR experiments based on the RWC data, since our algorithm requires learning data to generate a VQ tree, and relevance feedback data to improve precision of retrieval. It is obvious that the number of relevant (i.e., “good”) songs is insufficient for the 4 “fussy” subjects illustrated in Figure 3. Therefore, we were forced to run our experiments based on the HMV data set.

5. PROPOSAL

In order to achieve further advancement on research of MIR based on user preferences, it is obvious that a more large-scaled experiment is necessary. However, the manual construction of a large-scaled music data set is not only time-consuming, but also a very difficult task, if we are to make an objectively *fair* archive of music. We do believe that our method to extract a list of CDs based on weekly sales ranking data is a rather fair way to generate a music data set. However, the accumulated CD list can be considered as a biased data set, since consumers of CD shops may be biased

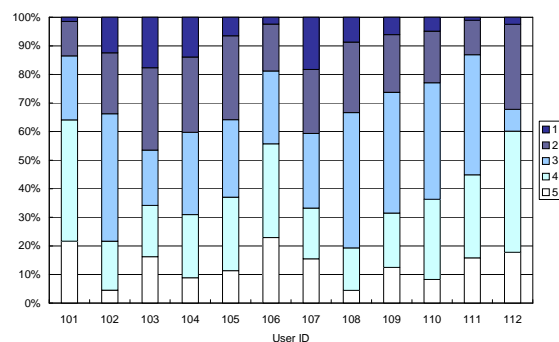


Figure 4: Ratio of user ratings for HMV data

towards young people who prefer new music over traditional types of music.

While copyright-free music data collections such as the RWC Database are definitely fair, our user rating experiments show that typical users simply do *not* like such music. Therefore, we do not expect the expansion of copyright-free music data (which is, of course, beneficial for conventional MIR research, though) will contribute to the development of user preference based MIR systems.

Based on our experiences, our proposal to the MIR community is to build a large list of on-the-market CDs which can be used as a standard set of music for MIR research. The number of CDs in this standard data set should be in the order of thousands, in order to provide extensive data for various research goals.

Generation of a music data set of this scale will certainly require cooperation from the music industry. Unfortunately, due to recent situations regarding the illegal distribution of copyrighted music on the Net, many major record companies are becoming more and more restrictive on the use of their contents. We wish the MIR community will make a push towards the music industry to provide an abundant data set for the development of MIR research.

6. CONCLUSION

In this paper, we presented the results of user rating collection experiments based on a copyright-free music data set, and a music data set derived from weekly sales ranking data of a major online CD shop. Results of our experiments clearly show that typical users apply higher ratings for songs extracted from on-the-market CDs, compared to songs in copyright-free music data sets. Based on this experience, we make a proposal to the MIR community to make a standard list of on-the-market CDs. We expect such a standard data set to provide a major contribution not only for MIR based on user preferences, but also for other potential MIR-related research issues.

7. REFERENCES

- [1] Hoashi, Zeitler, Inoue: “Implementation of relevance feedback for content-based music retrieval based on user preferences”, Proceedings of ACM-SIGIR 2002, pp 385-286, 2002.
- [2] Hoashi, Matsumoto, Inoue: “Personalization of user profiles for content-based music retrieval based on relevance feedback”, *to be published in Proceedings of ACM Multimedia 2003*, 2003.

- [3] Foote: “Content-based retrieval of music and audio”, Proceedings of SPIE, Vol 3229, pp 138-147, 1997.
- [4] David, Mermelstein: “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. Acoustic, Speech, Singal Proc., ASSP-28(4), 1980.
- [5] Rocchio: “Relevance Feedback in Information Retrieval”, in “The SMART Retrieval System – Experiments in Automatic Document Processing”, Prentice Hall Inc., pp 313-323, 1971.
- [6] Goto, Hashiguchi, Nishimura, Oka: “RWC Music Database: Popular, classical and jazz music databases”, Proceedings of ISMIR 2002, pp 287-288, 2002.