CONTENT-BASED MUSIC RETRIEVAL USING QUERY INTEGRATION FOR USERS WITH DIVERSE PREFERENCES

Keiichiro Hoashi **Hiromi Ishizaki** **Kazunori Matsumoto** **Fumiaki Sugaya**
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Fujimino-shi, Saitama 356-8502, JAPAN
e-mail:{hoashi, ishizaki, matsu, fsugaya}@kddilabs.jp

ABSTRACT

This paper proposes content-based music information retrieval (MIR) methods based on user preferences, which aim to improve the accuracy of MIR for users with "diverse" preferences, *i.e.*, users whose preferences range in songs with a wide variety of features. The proposed MIR method dynamically generates an optimal set of query vectors from the sample set of songs submitted by the user to express their preferences, based on the similarity of the songs in the sample set. Experiments conducted on a music collection with subjective user ratings verify that our proposal is effective to improve the accuracy of content-based MIR. Furthermore, by implementing a two-step MIR algorithm which utilizes song clustering results, the efficiency of the proposed MIR method is significantly improved.

1 INTRODUCTION

Recent popularity of online music distribution services have provided the opportunity to access to millions of songs, and also have enabled common users to accumulate a large-scaled music collection. This rapid growth of both online and personal music collections has made it increasingly difficult for users to efficiently find songs which they want to listen to. Development of an effective music information retrieval (MIR) system is, therefore, essential to realize satisfactory music distribution services, and improve usability of music applications.

Due to these recent developments, various research have been conducted in the area of content-based MIR based on user preferences. Logan has proposed a content-based MIR method which extracts the acoustic features from a set of songs, *e.g.*, an album, which is defined as an expression of user preferences[8]. Grimaldi *et al.* have extended feature extraction techniques that have been proved effective for music genre classification, to conduct retrieval of music based on user preferences[5]. Furthermore, the authors have proposed a content-based MIR method, which generates a vector expression of user preferences from a sample set of songs submitted by the user[6][7].

Although the above content-based MIR methods have been proved to be reasonably effective, the effectiveness of the methods are limited on conditions that the preferences of the users are focused. For example, the evaluations in [8] have utilized individual albums, which typically consist of similar songs, as an expression of user preferences. Furthermore, evaluations in [5] conclude that their MIR method is ineffective when user preferences are *not* focused on a specific genre.

The main objective of this paper is to propose a method which is capable of providing accurate content-based MIR results to users with diverse preferences. Namely, this paper proposes a method which automatically generates an optimal number of queries from a sample set of songs submitted by the user, based on the similarity between the songs in the set. Effectiveness of the proposed method is verified by experiments conducted on a music collection with subjective user ratings.

2 CONVENTIONAL MIR METHODS

In [6], the authors have proposed a content-based MIR method, which retrieves songs that fit music preferences based on a small set of example songs (hereafter referred as the "sample set") provided by the user. Our MIR method applies the tree-based vector quantization (TreeQ) algorithm developed by Foote[2]. Furthermore, in [7], the authors have proposed a feature space modification (FSM) method, which utilizes song vector clustering results to automatically generate a training set for TreeQ, from any music collection. By implementing this method, the MIR system can build a feature space optimized to the songs in the music collection.

By utilizing the tree-based vector quantizer (hereafter referred to as the "VQ tree") generated by the above methods, our system is able to retrieve a list of songs which fit user preferences, based on the sample set of "good" songs submitted by the user. The system first generates a vector expression of the users' preferences (hereafter referred to as the "user profile"), by calculating the vector sum of all vectors of the songs in the sample set. Scores of all songs in the collection are calculated based on the cosine similarity between the user profile and song vectors, and songs with high similarity to the query are presented to the user as the MIR result.

3 PROBLEMS

As previously described, existing content-based MIR methods are able to achieve reasonable success in retrieving songs which fit user preferences. However, the accuracy of existing MIR methods, including our methods in Section 2, are dependent on the information submitted by the user. For instance, if the user inputs his/her preferences to the MIR system by specifying a single song or musical genre, it is fairly easy to derive accurate MIR results, since the user preference is well-expressed by the submitted query. It is obvious, though, that the preferences of users are not always focused on a specific genre and/or song. If anything, user preferences are usually diverged in multiple genres. In such cases, the accuracy of existing content-based MIR methods are expected to be degraded.

For example, consider a situation where a noisy rock song and a soothing classical song are both included as "good" songs in the sample set. As mentioned in Section 1, most existing MIR methods are not capable to output accurate MIR results for users with such diverse preferences. Our previously proposed methods in [6] and [7] are capable of conducting MIR, even from such a diverse sample set. However, the user profile generated from the sample set is the sum vector of the two songs, meaning that the user profile points to the area in the middle of the two songs. Naturally, the MIR result based on the query is expected to consist of songs located in that area. In other words, the system will not be able to retrieve songs that are similar to either rock or classical songs, which are assumably a better representation of the user preferences than the songs that will be retrieved by these methods.

A naive way to solve this problem is to utilize the vectors of all songs in the sample set as independent queries, and merge all MIR results obtained from each query. However, this approach obviously will increase the computational cost to conduct MIR.

4 QUERY INTEGRATION

In order to solve the previously described problems, we propose the *query integration* method. The objective of this method is to automatically determine an optimal number of queries to be generated from the sample set, based on the similarity of the features of the songs in the set.

A conceptual illustration of the proposed method is shown in Figure 1. Figure 1 illustrates a situation where the sample set consists of six songs, $S_1, \cdots, S_6$. There are two sets of songs in the sample set that are highly similar to each other, $\{S_1, S_2, S_3\}$, and $\{S_4, S_5\}$. In this case, the intuitionally optimal set of queries can be generated by integrating the songs in these two sets to generate a single query, which represents each set respectively, and utilizing S_6 as an independent query.

The proposed query integration method is implemented by conducting hierarchical clustering of the song vectors included in the sample set. First, the similarity between all song vectors in the sample set are calculated. Next, the two songs with the highest similarity are extracted. If the similarity between these two songs exceed a predefined threshold τ_q, the songs are integrated to a single vector, by summing the vectors of the two songs. This procedure is repeated until all song vectors in the sample set are integrated to a single vector, or when the maximum similarity fails to exceed τ_q.

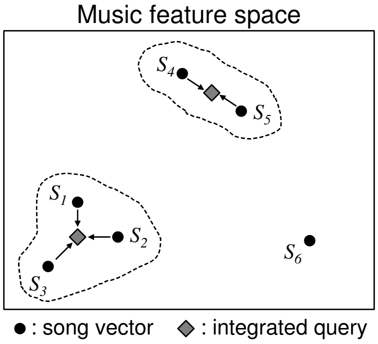

Figure 1. Conceptual illustration of query integration

Query integration is expected to solve the diverse user preference problem, since the generated queries are able to represent the features of the songs in the sample set if they are diverse, and, simultaneously, can generate a focused query whenever appropriate. Furthermore, by integrating vectors of highly similar songs, this method can reduce the increase of computational cost, compared to conducting MIR individually for all songs in the sample set.

5 EXPERIMENTS

In order to evaluate the accuracy and efficiency of the proposed method, an experiment is conducted based on a large music collection with subjective user ratings. Details of the experiments are as follows.

5.1 Experimental setup

5.1.1 Data

The music collection used for our experiments is the same as the collection used in the experiments in [7], which is constructed by combining the experiment data set used in the experiments in [6], with songs included in the CDs listed in the *uspop2002* music data set constructed by Ellis[1]. The total number of songs in the music collection is 6863.

Rating data for our MIR experiments, which are used for the evaluation of MIR accuracy, are collected by inviting 20 subjects to apply subjective ratings to the songs in the above music data set, ranging from 1 to 5 (Bad:1 $\sim$ Good:5). The rated data is then classified into three categories, according to the ratings applied to each song. The three categories are: "good songs" (C_g), "bad songs" (C_b), and "fair songs" (C_f). Categories C_g, C_b, and C_f consist of songs rated (4 or 5), (1 or 2), and 3, respectively. Table 1 shows the average ratio of songs per rating and category for all subjects.

Table 1. Summary of user rating data

Category	Rating	Ratio(%)
C_g	5	16.9
	4	20.1
C_f	3	31.3
C_b	2	20.4
	1	11.2

5.1.2 VQ tree construction

The VQ tree, which is used in the following experiments, is constructed based on the FSM method described in [7]. First, the initial feature space is generated by using the RWC Genre Database [4] as training data. Next, the feature space is modified by cluster-based FSM. The number of clusters, and the number of songs that are extracted as training data from each cluster for FSM, are set to 10 and 3, respectively. These values are determined empirically, based on the results of preliminary experiments.

5.2 Method

In this experiment, a sample set of songs for each subject is first generated by randomly extracting $N = 5$ songs that belong to C_g. This sample set is utilized to generate two user profiles, one based on query integration, and the other by the conventional MIR method, where the sum vector of all N songs in the sample set is utilized as the user profile. Next, MIR is conducted using the two user profiles. If the query integration method generates multiple queries from the sample set, the MIR result is generated by merging the results from each generated query, by sorting songs based on their score in each MIR result. The final result is obtained by extracting the top Num songs based on their score. In order to compensate the randomness of the sample set generation process, this query generation process is repeated five times per subject.

The threshold for the query integration method, τ_q, is determined by the following formula: $\tau_q = \mu_{sim} + \alpha \cdot \sigma_{sim}$, where μ_{sim} and σ_{sim} denote the average, and standard deviation of all song-to-centroid similarity values, respectively. α is a coefficient which is defined to adjust the value of τ_q. In the following experiments, the value of α is set as $\alpha = \{0.5, 1.0\}$.

5.3 Evaluation measures

In order to evaluate MIR accuracy, we calculate the ratio of songs in the MIR result which belong to categories C_g, C_f, and C_b, and use the average ratio of all experiments as the final evaluation measure. A high ratio of songs in C_g indicate that more preferable songs have been retrieved, thus is considered as a superior result.

Furthermore, in order to measure MIR efficiency, we count the number of similarity calculations executed in each experiment, and calculate the ratio to the number of songs in the music collection (hereafter referred to as *CalcRate*). Note that, in the following experiments, the number of queries apply an effect to *CalcRate*. Namely, when $K > 1$ queries are generated as a result of query integration, the *CalcRate* of the MIR process is $K \times 100\%$.

Table 2. Average MIR accuracy and efficiency of query integration and conventional methods ($Num = 50$)

Method	C_g	C_f	C_b	*CalcRate*
QI($\alpha = 0.5$)	0.668	0.195	0.137	291.3%
QI($\alpha = 1.0$)	0.683	0.187	0.130	364.3%
Conv	0.585	0.244	0.170	100.0%
No-QI	0.702	0.176	0.122	500.0%

5.4 Results

Comparison of the MIR accuracy of the proposed query integration method (*QI*) and the conventional method (*Conv*) is shown in Table 2, where the average ratio of song categories in all MIR results when $Num = 50$ is listed. For additional comparison, we also present the results of experiments when no query integration is conducted (*No-QI*), *i.e.*, where all $N = 5$ songs in the sample set are treated as individual queries. Furthermore, the average *CalcRate* of all experiments are also written in this Table.

Results in Table 2 indicate that all methods have accurately retrieved songs that fit user preferences, since the ratio of songs in C_g in the MIR results all exceed the overall ratio of songs in C_g (=37.0%, from Table 1). It is also clear that the *QI* methods are superior to *Conv*, based on comparison of the C_g ratio of the two methods. In terms of MIR accuracy, the *No-QI* method has achieved the best results among our experiments. However, the *CalcRate* results indicate that the *QI* methods are capable of improving accuracy, without increasing the amount of computation as much as *No-QI*. Under extreme conditions where the query consists of hundreds of songs, the cost of the query integration process may apply a significant effect to the overall performance of the MIR process. However, in a realistic situation where N is a small number, we can conclude that the *QI* method is capable of achieving accurate MIR, while limiting increase of computational cost.

Next, we conduct analysis to confirm the effectiveness of the query integration method for users with diverse preferences, as assumed in our hypotheses. First, we define the "diversity" of a query as the average similarity between the songs included in the query (hereafter referred to as *AvgQSim*). If the *AvgQSim* of a query is low, the preference expressed by the query is considered as diverse. On the contrary, queries with high *AvgQSim* are considered to be focused, since all songs in the query are similar to each other. In the following analysis, queries are categorized to the following three classes: *LowSim*, *MidSim*, and *HiSim*, which consist of queries whose *AvgQSim* is below 0.3, between 0.3 and 0.5, and over 0.5, respectively.

Table 3 lists the average rate of songs in C_g for the *Conv* and *QI* results. These results indicate that the MIR accuracy for queries in the *LowSim* and *MidSim* classes have been significantly improved by *QI*, verifying that, as

Table 3. Average ratio of songs in C_g for each query class ($Num = 50$)

Query class	*Conv*	*QI*($\alpha = 0.5$)	*QI*($\alpha = 1.0$)
LowSim	0.562	0.699	0.715
MidSim	0.598	0.678	0.702
HiSim	0.558	0.565	0.564

hypothesized, *QI* is especially effective for generating accurate MIR results for users with diverse preferences.

Generally, the experimental results indicate that a high τ_q is beneficial for improvement of MIR accuracy. However, utilization of a fixed τ_q does not necessarily improve MIR results for all queries, as can be observed from the low improvement rate of *HiSim* queries in Table 3. Therefore, a method which can determine an optimal threshold for query integration, based on the features of the songs in the sample set, is assumed to be necessary for further improvement of MIR accuracy.

6 CLUSTER SELECTION MIR

While the previous experiment has proved the effectiveness of query integration, it is also true that the method increases the computational cost of the MIR process, as indicated by the high *CalcRate* values in Table 2. In order to improve MIR efficiency, we propose a selective MIR method, which utilizes the clustering results of the songs in the collection.

The cluster selection MIR method utilizes the clustering results of the VQ tree construction phase described in Section 5.1.2. All songs in the collection are associated to a cluster, by selecting the cluster whose centroid has the highest similarity to the song vector. Next, in order to select the target cluster set for a given query, we calculate the similarity between the query and each cluster centroid, and select the top m clusters based on the query-centroid similarity. MIR is then conducted only for the songs which belong to the clusters in the target set. The combination of cluster selection MIR with query integration is implemented by selecting the target set of clusters for each query generated by the query integration process.

A simple experiment is conducted to evaluate the accuracy and efficiency of combining cluster selection MIR with query integration. Table 4 lists the average accuracy and *CalcRate* of all conducted experiments. Note that "*CS*" in Table 4 expresses the cluster selection MIR runs.

By comparison of the results in Tables 2 and 4, it is clear that MIR accuracy similar to that of the original *QI* method can be achieved when m is set to 4 for cluster selection MIR, while significantly reducing *CalcRate*. Furthermore, even if the conditions of cluster selection are extreme, *e.g.*, when $m = 1$, the MIR accuracy still outperforms that of *Conv*, while reducing *CalcRate* down to 31.3%. These results indicate that the increase of computational cost necessary for query integration can be easily reduced by implementing cluster selection MIR.

Table 4. Average accuracy and efficiency of query integration combined with cluster selection MIR

Method	C_g	C_f	C_b	*CalcRate*
QI+*CS*($\alpha = 0.5, m = 1$)	0.606	0.221	0.173	31.3%
QI+*CS*($\alpha = 0.5, m = 2$)	0.645	0.199	0.156	69.1%
QI+*CS*($\alpha = 0.5, m = 3$)	0.661	0.198	0.141	109.7%
QI+*CS*($\alpha = 0.5, m = 4$)	0.666	0.195	0.139	148.7%
QI+*CS*($\alpha = 1.0, m = 1$)	0.609	0.221	0.171	40.9%
QI+*CS*($\alpha = 1.0, m = 2$)	0.658	0.193	0.149	89.6%
QI+*CS*($\alpha = 1.0, m = 3$)	0.674	0.190	0.136	141.6%
QI+*CS*($\alpha = 1.0, m = 4$)	0.680	0.187	0.133	190.3%

7 CONCLUSION

In this paper, we have proposed a query integration method, which aims to achieve accurate MIR for users who have diverse preferences. The overall results of evaluation experiments conducted on a music collection with subjective user ratings have proved that the proposed method is capable to provide accurate MIR results for such users. Furthermore, the implementation of cluster selection MIR has proved to be effective to reduce computational cost with minimal sacrifice of MIR accuracy.

8 ACKNOWLEDGMENTS

This research used the TreeQ package[3] developed by Jonathan T. Foote.

9 REFERENCES

[1] D. Ellis: "The *uspop2002* Pop Music data set", List available at http://www.ee.columbia.edu/%7Edpwe/research/musicsim/uspop.html, 2003.

[2] J. Foote: "Content-based retrieval of music and audio", Proceedings of SPIE, Vol 3229, pp 138-147, 1997.

[3] J. Foote: "TreeQ software," http://treeq.sourceforge.net/

[4] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka: "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *Proc. of ISMIR*, pp.229-230, Baltimore, MD, USA, 2003.

[5] M. Grimaldi, P. Cunningham: "Experimenting with music taste prediction by user profiling," *Proc. of 6th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, pp 173-180, 2004.

[6] K. Hoashi, K. Matsumoto, N. Inoue: "Personalization of user profiles for content-based music retrieval based on relevance feedback", Proc. of ACM Multimedia 2003, pp. 110-119, 2003.

[7] K. Hoashi, K. Matsumoto, F. Sugaya, H. Ishizaki, J. Katto: "Feature space modification for content-based music retrieval based on user preferences," Proc. of ICASSP 2006, pp. 517-520, 2006.

[8] B. Logan: "Music recommendation from song sets", *Proceedings of ISMIR*, Barcelona, Spain, 2004.