FINDING SIMILAR MUSIC ARTISTS FOR RECOMMENDATION

ABSTRACT

Music information retrieval had become an interesting research subject to be explored. The development of information clustering leads the user to find related contents and interests more easily. In this paper, we present a recommendation of similar music artists based on the music genre classification, artist's era, and social rating information. The algorithm is performed in three steps: compute similarity measure on music genre; apply the user rating factor to the artist; and finalize the similarity by selecting artists who have the same period of music activities. The Jaccard's coefficient and Nearest-Neighbor search have been used in the computation. The experiment shows that we can obtain better results using the proposed method.

KEYWORDS

Information retrieval, music contents, artist similarity; user rating.

1. INTRODUCTION

Recently, video sharing web sites such as YouTube has been popular greatly. As the number of video contents increase exponentially, it is very important to recommend videos which the users want to see among tons of video contents. Most recommendation algorithms try to find similar videos based on textual and visual similarity. However, the user would like to search videos by recommendation using additional information related to keywords in addition to video content similarity.

In this paper, we suggest a new method to recommend music artists by computing the artist similarity. Based on the similar artist list, the user can find music contents in which he/she may has interest. The algorithm constructs music artist list and genre structure using an external DB, Yahoo! Music, compute the similarity and group music artists based on genre and era, and evaluates artist reputation based on social rating information from Yahoo! Webscope dataset. We expect that the combination of similarity measure and artist reputation can improve the searching result.
We have conducted some experiments on YouTube to verify that the designed algorithm can make better recommendation. It turns out that the algorithm shows better results.

The paper is organized as follows: Section 2 describes the related works. Section 3 describes the music information and the model approach. The proposed method computing artist similarity is presented in Section 4. In Section 5 the experiment results are presented to evaluate the proposed method. And, finally some conclusions and future works are drawn in Section 6.

2. RELATED WORK

There have been numbers of interesting research in music information retrieval especially in artist similarity computation. Hong et al. presents the similarity measure that utilizes tag and tag co-occurrence, importing the tags from Last.fm (http://www.last.fm), then compute the genre classification based on the previous similarity score between artists. Another work has been introduced by Li et al. They cluster the music with some features from different resources. A bimodal clustering framework for integrating the features based on minimizing disagreement is used. The term bimodal must have a complete feature representation, consists of the acoustic features which summarize the sound, and text features which summarize the words put into the music. A paper from Geleijnse et al. suggests the use of community-based data for artist tagging and artist similarity. Tags which are community-based hence give a description of a product through a community rather than an expert opinion. In addition, tags which are collected from Last.fm shows to be consistent and descriptive. These works that have been presented are almost similar, that they use tags which are provided by the users in Last.fm to describe the music. However, these tags can be very either too general or too specific. Thus we need to find steadier factor in order to acquire a better artist similarity and identification. Through this study, we endeavor to improve the artist similarity using more specific data related to artist itself rather than tags.

3. MUSIC INFORMATION AND MODEL APPROACH

The purpose of this similarity study is to create a group clustering based on parameters of activities. By choosing the music category as a field of experiment, and artists as the object, it is expected that the final results will give improvement to the previous method.

3.1 Artist Information and Genre

The artists information and genres are collected from the Yahoo! Music web service (http://developer.yahoo.com/music/) that is available with API. Compare to other API, Yahoo! Music provides the most accessible and more complete data. The API that is used provides access to the Yahoo! Music Catalog of artists, album, tracks, videos, and more. It provides numerous ways to browse the catalog: through charts, search, similarities, genres, artists, and user recommendations and ratings.

In using the Yahoo! Music API, an application ID is needed to be used as our identification when accessing the data. The API is a HTTP REST-based API (http://www.xfront.com/REST-Web-Services.html) that returns data in any format, including XML, JSON, and RSS (XML by default). The Yahoo! Music API service is limited to 5,000 queries per day per IP address.

The data have been collected include:

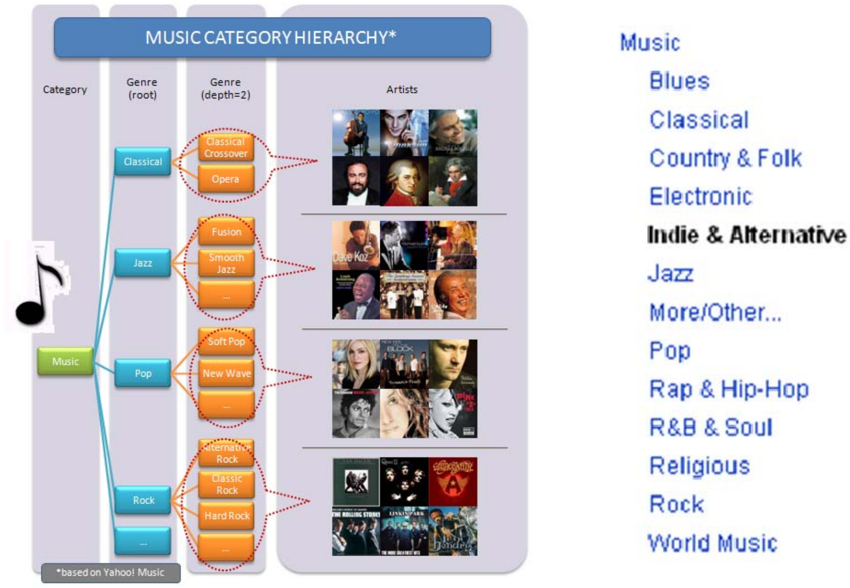
a. artists information,
b. category,
c. releases and releases album,
d. videos,
e. top similar artists,
f. radio stations,
g. top tracks, and
h. events

3.2 Music Category Hierarchy

Figure 1(a) shows that there are music classifications based on the music genre into several depth levels, different to YouTube that only have limited classification on music category, with only one level of category as we can see in Figure 1(b).

As we see in the Yahoo! Music data, it is shown that the results of similar artists can be very different from user interests and expectation. One example is shown in the Figure 2 below.

Both Enya and Alanis Morissette are sharing the same one genre which is pop (soft pop is sub genre of pop), but the other two are completely different. It’s the same case when Louis Armstrong irrelevantly to be found as the similar artist to Enya. Therefore, the items grouping should be improved.



(a) (b)

Figure 1. (a) Yahoo! Music Hierarchy vs. (b) YouTube Music Category Hierarchy

GENRE of Enya

ID	NAME	TYPE
39469069	Celtic	Genre
7318646	New Age	Genre
7318718	Soft Pop	Genre

GENRE of Alanis Morissette

ID	NAME	TYPE
7318736	Adult Alternative	Genre
7318734	Mainstream Rock	Genre
7318647	Pop	Genre

SIMILAR ARTISTS from Enya

ID	NAME
258777	Alanis Morissette
259099	Barenaked Ladies
290443	Dido
314079	Enigma
303614	Eric Clapton
271612	Fatboy Slim
266172	James Taylor
252708	Jewel
282175	Josh Groban
252147	Louis Armstrong

1 2

Figure 2. Similar artists to Enya

3.3 Rating data

The artists' ratings are collected from the Yahoo! Webscope dataset (http://research.yahoo.com) which provide reference library of datasets for non commercial use. These datasets have been reviewed to conform to Yahoo!'s data protection standards, including strict controls on privacy.

The dataset that is used here is R1, Yahoo! Music User Ratings of Musical Artists, version 1.0. This dataset represents a snapshot of the Yahoo! Music community's preferences for various musical artists. The dataset contains 11,557,943 ratings of musical artists given by Yahoo! Music over the course of a one month period sometime prior to March 2004, and 97,812 artists are listed.

3.4 Model-Based Approach

The idea of items grouping comes from a model of distance learning environment, proposed by Pollalis, et al. They define the similarity coefficient between users and learning objects on automatic creation of properly matching collaborating groups, by selecting the appropriate learning objects to form a corresponding educational package, and the proper formation of the groups of learner. Breese et al. introduced a model based collaborative filtering, which uses the user database to estimate or learn a model, which is then used for predictions. They calculate the expected value of a vote, given what we know about the user (from probabilistic perspective). By applying both of the role models introduced previously, in computing the similarity measure using model based algorithm to the items (artists). The combination of Jaccard coefficient and Nearest-Neighbor search will be performed to get the optimum results.

4. ARTIST SIMILARITY

We now consider about the genre and album releases year as the basic parameters to be used in our computation. By using that information, we can build certain knowledge field which will be used in our algorithm. In general, we assume that all artists have these information, thus comparison between artists can be done respectively. The similarity measure on distance and Nearest-Neighbor search are proposed in this paper. We implement three steps of computation in order to obtain more detail results, as defined below:

a. Applying comparison between artists' genre using distance measure
b. Applying user rating to artists
c. Applying Nearest-Neighbor search on artists' releases album

In music directory, we define some properties for artist and genre information.

Table 1. Parameters properties

Properties	Description
$A = \{a_x\}$, x = 1, 2, … , k	the set of artist
$G = \{g_y\}$, r = 1, 2, … , m	the set of genre

Here we describe how we construct our method.

4.1 Similarity measure on distance

The Jaccard's coefficient is found to be the most stable similarity coefficient among 20 similarity coefficients according to Yin et al. A model for distance learning environment defined by Pollalis et al. also adapt the Jaccard's similarity coefficient in order to measure similarities between learners and the learning objects. Using the same analogy, as the same characteristics of artist and genre, we use the same analogy to calculate the similarity of artist.
The *Similarity Level* (SL) between the Artist and Genre is the Jaccard's coefficient between $|a_x|$ and a_c as defined in formula below.

$$SL(|a_x|, a_c) = \frac{\gamma}{\alpha + \beta + \gamma}. \qquad (1)$$

where α represents the total number of genres that is not presented in $|a_x|$ but appears in a_c, β represents the total number of genres that is not presented in a_c but appears in $|a_x|$, γ represents the total number of genres presented in both $|a_x|$ and a_c. While $|a_x(g_y)| = \{a_x(g_1), a_x(g_2), \ldots, a_x(g_n)\}$.

$$a_x(g_y) = \begin{cases} 1, \text{ if artist belongs to genre } g_y \\ 0, \text{ otherwise} \end{cases}$$

$|a_x|$ represents the genre which the artist belongs to, where $a_x(g_y) = 1$

4.2 User rating

We calculate the user rating data collected from the Yahoo! Webscope dataset by computing the average of user rating grouped by artist denoted by Ra_x. Suppose that we have a set of ru_a_x of users, who give rating to artist a_x, and na_x is the number of users in ru_a_x, then we can define Ra_x as follows:

$$Ra_x = AVG(ru\_a_x)_{na_x} = \frac{ru_1 a_x + ru_2 a_x + \cdots + ru_p a_x}{na_x}. \qquad (2)$$

Thus, by applying weight to both similarity level and user rating, Step 2 of the computation can be defined as follows:

$$SL_2 = (W_{SL1} * SL(|a_x|, a_c)) + (W_{Ra} * NORM(Ra_x)). \qquad (3)$$

where W_{SL1} is weight for the corresponding similarity level in Equation 1, and W_{Ra} is the weight for the corresponding artist rating in Equation 2.

4.3 Nearest-Neighbor Search

Nearest-Neighbor search also known as proximity search or closest point search in metric spaces. The query finds the closest object to the given query object, that is the nearest neighbor of q. The concept can be generalized to the case where we want to find the k nearest neighbors, in the equation as follows (Zezula et al.):

$$kNN(q) = \{R \subseteq X, |R| = k \wedge \forall x \in R, y \in X - R : d(q,x) \leq d(q,y)\}. \qquad (4)$$

where k NN(q) query retrieves the k nearest-neighbors of the object q, and in the distance range (r) searching, where p Є S with d(q, p) ≤ r.

5. EXPERIMENT RESULTS

In Section 4, we have described a strategy for computing the artist similarity. In this Section, we focus on the proposed method and evaluate their performance using the dataset available. Figure 3 below is the example of artist similarity application that we can search the artist name, the *SL* threshold, and also the distance (r) in years, the period on the artist released his album.

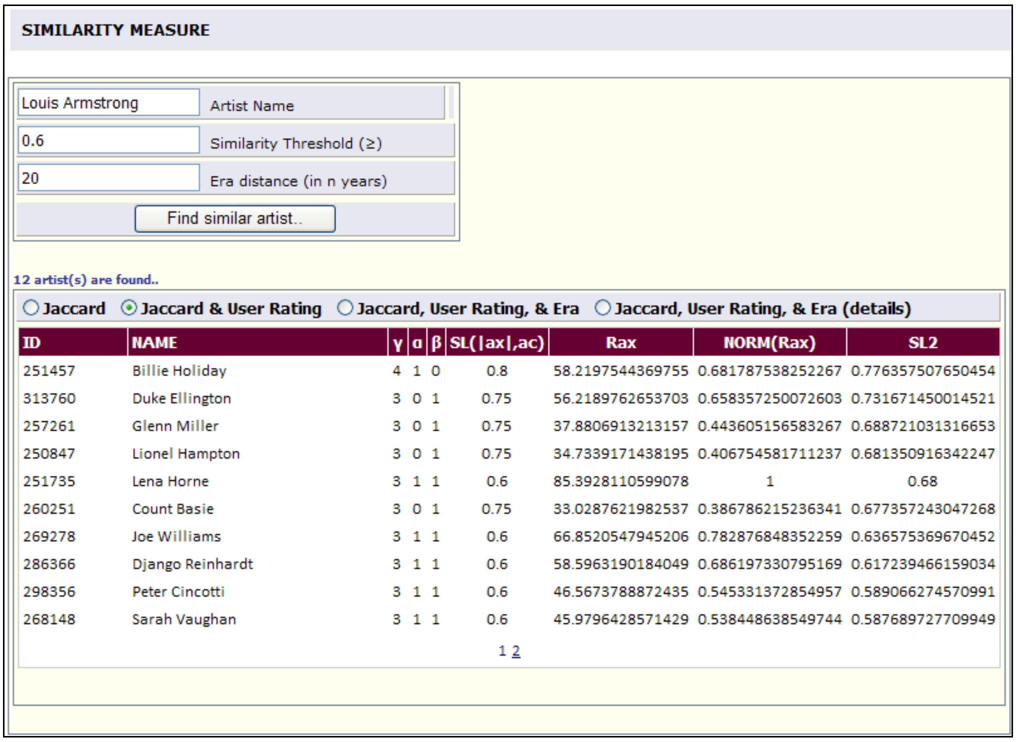


Figure 3. Application example of artist similarity after applying the user rating

The result for every step artist similarity computation is shown in Table 3, Table 4, and Table 5. All artist genres are compared to the "Louis Armstrong" genre which is shown in Table 2. The first step of computation results 13 artists similar to Louis Armstrong as shown in Table 3. In second step after applying the user rating from Yahoo! music user rating database, it is shown in Table 4 that the list of artists is decreasing, known that Harry Connick, Jr. is not on the list, because the user rating for this artist is not available. In the last step, we get more comprehensive result after applying the nearest neighbor search of the artist's era factor. Louis Armstrong's first release according to Yahoo! Music web service is in the year 1925, titled "Hot Fives". With the distance r = 20 years (1925 + r), we get releases from the artists listed in Table 5, and group the results by artists as shown in Table 6.

Table 2. Genre of Louis Armstrong

ID	Name	Type
39468850	**Big Band/Swing**	Genre
7318643	**Jazz**	Genre
39469134	**Jazz Classics**	Genre
39469081	**Vocal Jazz**	Genre

Table 3. Step 1: Similar artists to Louis Armstrong, with $SL \geq 0.6$

Artist Name	γ	α	β	$SL(\|a_x\|, a_c)$
Louis Armstrong	**4**	**0**	**0**	**1**
Billie Holiday	4	1	0	0.8
Harry Connick, Jr.	4	1	0	0.8
Lionel Hampton	3	0	1	0.75
Glenn Miller	3	0	1	0.75
Count Basie	3	0	1	0.75
Duke Ellington	3	0	1	0.75
Lena Horne	3	1	1	0.6

Wynton Marsalis	3	1	1	0.6
Chet Baker	3	1	1	0.6
Sarah Vaughan	3	1	1	0.6
Joe Williams	3	1	1	0.6
Django Reinhardt	3	1	1	0.6
Peter Cincotti	3	1	1	0.6

Table 4. Step 2: Similar artist to Louis Armstrong, with $SL \geq 0.6$ and user rating to artists, $Wj = 0.8$, $Wr = 0.2$

Artist Name	γ	α	β	SL($\|a_x\|$, a_c)	Ra_x	NORM(Ra_x)	TOTAL
Billie Holiday	4	1	0	0.8	58.2197544369755	0.681787538252267	0.776357507650454
Duke Ellington	3	0	1	0.75	56.2189762653703	0.658357250072603	0.731671450014521
Glenn Miller	3	0	1	0.75	37.8806913213157	0.443605156583267	0.688721031316653
Lionel Hampton	3	0	1	0.75	34.7339171438195	0.406754581711237	0.681350916342247
Lena Horne	3	1	1	0.6	85.3928110599078	1	0.68
Count Basie	3	0	1	0.75	33.0287621982537	0.386786215236341	0.677357243047268
Joe Williams	3	1	1	0.6	66.8520547945206	0.782876848352259	0.636575369670452
Django Reinhardt	3	1	1	0.6	58.5963190184049	0.686197330795169	0.617239466159034
Peter Cincotti	3	1	1	0.6	46.5673788872435	0.545331372854957	0.589066274570991
Sarah Vaughan	3	1	1	0.6	45.9796428571429	0.538448638549744	0.587689727709949
Wynton Marsalis	3	1	1	0.6	35.7785758259799	0.418988148790174	0.563797629758035
Chet Baker	3	1	1	0.6	34.3095869647594	0.401785425949841	0.560357085189968

Table 5. Step 3: Similar artist to Louis Armstrong, era added (r = 20 years)

Artist Name	Release Title	Year
Duke Ellington	1928	1928
Django Reinhardt	1935	1935
Count Basie	One O'Clock Jump (MCA Jazz)	1937
Glenn Miller	Live At The Paradise Restaurant	1939
Glenn Miller	The Carnegie Hall Concert	1939
Count Basie	Volume 2	1939
Duke Ellington	Fargo, North Dakota--November 7, 1940	1940
Glenn Miller	1942 Chesterfield Shows	1942
Glenn Miller	Planet Jazz: Glenn Miller	1942
Count Basie	And His Orchestra (1944)	1944
Count Basie	Beaver Junction (1944-1946)	1944

Table 6. Final result: Similar artist to Louis Armstrong

Artist Name
Duke Ellington
Glenn Miller
Count Basie
Django Reinhardt

We compare the result of Yahoo! Music Web Service to our proposed method as shown in Table 7. Artists such as Dave Koz, Diana Krall, and George Benson are not included in our result, even after we apply the $SL \geq 0.5$. From these figures, we can see that genre has very important factor in similarity measurement. Consequently, we argue that our proposed method can improve the result.

Table 7. Similar artist: comparison between Yahoo! Music Web Service and Step 2 Proposed Method ($SL \geq 0.6$)

Yahoo! Music	Proposed method
Billie Holiday	**Billie Holiday**
Charles Mingus	**Duke Ellington**
Charlie Parker	Glenn Miller
Chick Corea	Lionel Hampton
Chris Botti	Lena Horne
Count Basie	**Count Basie**
Dave Koz	Joe Williams
Diana Krall	Django Reinhardt
Dizzy Gillespie	Peter Cincotti
Duke Ellington	Sarah Vaughan
Ella Fitzgerald	Wynton Marsalis
George Benson	Chet Baker
John Coltrane	
Miles Davis	

Nina Simone
Pat Metheny
Rick Braun
Stan Getz

6. CONCLUSION

In this paper, we study the similarity measure applied to cluster the artists based on genre. The experimental results on similarity measure show that the proposed method can perform more accurate results. Additional factor of user ratings from Yahoo! Webscope dataset helps to eliminate the artists whose song or artist ratings factors are low, and with the era added, the more detail result can be obtained.

In this study, we can conclude that the similarity measure is possible to be performed not only in music category but also in other categories, as long as the model based and the data structures are available to be constructed, thus the similar hierarchy as in music category can be used as the basis of similarity measure.

REFERENCES

Breese, J.S. et al, 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, Madison, USA.

Geleijnse, G. et al, 2007. The Quest for Ground Truth in Musical Artist Tagging in the Social Web Era. *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*. Vienna, Austria, pp. 525-530.

Hong, J. et al, 2008. Tag-based Artist Similarity and Genre Classification. *Proceedings of IEEE International Symposium on Knowledge Acquisition and Modelling Workshop.* Wuhan, China, pp. 628-631.

Li, T. et al, 2009. Music Clustering with Features From Different Information Sources. *In IEEE Transactions on Multimedia*, Vol. 11, No. 3, pp. 477-485.

Pollalis, Y.A. and Mavrommatis, G., 2009. Using similarity measures for collaborating groups formation: A Model for distance learning environments. *In European Journal of Operational Research*, Vol. 193, No. 2, pp. 626-636.

Yahoo! Music web service with API, http://developer.yahoo.com/music/.

Yahoo! Research Alliance Webscope Program: Yahoo! Music User Ratings of Musical Artists, version 1.0, http://research.yahoo.com.

Yin, Y. and Yasuda, K., 2005. Similarity coefficient methods applied to the cell formation problem: a comparative investigation. *In Computers & Industrial Engineering*, Vol. 48, No. 3, pp. 471-489.

Zezula, P. et al, 2006. *Similarity Search: The Metric Space Approach*. Springer Publishers, New York, USA.