A Fair Approach to Music Recommendation Systems Based on Music Data Grouping

Ye-In Chang, Chen-Chang Wu and Meng-Chang Tsai

***Abstract*—How recommending the music that user is interested in from a wide variety of music is the development intentions of the music recommendation system MRS (Music Recommendation System). Chen *et al.* have proposed the Content-based (CB) and Collaborative (COL) methods for music recommendation. The CB method is to recommend the music objects that belong to the music groups the user is recently interested in and the COL method is to provide unexpected findings due to the information sharing between relevant users. But the CB method will lead to the result that the group weight of music group B which appears once in the later transaction is larger than the group weight of the music group A which appears many times in the earlier transaction. The COL method will lead to the result that the supports of the groups which have different densities are the same, and then the users may be grouped together. Therefore, in this paper, to be fair, we propose the TICI (Transaction-Interest-Count-Interest) method to improve the CB method, and propose the DI (Density-Interest) method to improve the COL method. Our DI method calculates the supports of music groups and consider the distributions of appearances of the music group. From our simulation results, we show that our TICI method could provide better performance than the CB method. Moreover, our DI method also could provide better performance than the COL method.**

***Index Terms*—music recommendation system, user interest, transaction, count, weight.**

I. Introduction

In recent years, the music becomes more popular due to the evolution of the technology. Various kinds of music around us become more complexity and huge. In addition to searching expected music objects for users, it becomes necessary to develop a music recommendation service. The Music Recommendation System (MRS) is a website which provides the service of music recommendation [14].

There have been many researches in the field of MRS, such as content-based music filtering system with editable user profile [6], the sensitivities of user profile information in music recommender systems [10] and the music recommendation based on music data grouping and user interests [5], [4]. A content-based music filtering system with an editable user profile [6] is using a decision tree in a content-based music filtering system [11]. The sensitivities of user profile information is describing empirical research into the factors influencing the trade-off between the perceived benefits of personalization and the privacy ’costs’ experienced by individuals [10]. Instead of textual descriptions, the music recommendation based on music data grouping and user interests considers the perceptual properties of music objects, such as pitch, duration, and loudness, which can be directly extracted from the music objects [5].

TABLE I
A SAMPLE OF THE ACCESS HISTORY H1

Transaction	Music Group
T1	AA
T2	AC
T3	DEF
T4	GHI
T5	JK
T6	B

Arbee L.P. Chen *et al.* have proposed an alternative way of music recommendation [5]. Instead of textual descriptions, they consider the perceptual properties of music objects, such as pitch, duration, and loudness, which can be directly extracted from the music objects. For users, the preferences are derived from the access histories and recorded in profiles. Two recommendation methods are proposed to approach the corresponding goals:

(1) **The CB Method**. The purpose of the CB method is to recommend the music objects that belong to the music groups the user is recently interested in. To capture the recent interests of the user, they analyze the latest transactions in the access history as follows. Each transaction is assigned a different weight, where the latest transaction has the highest weight.

(2) **The COL Method**. The CB method tends to provide expected and interesting music objects for users. Based on the collaborative approach, the purpose of the COL method is to provide unexpected findings due to the information sharing between *relevant users*. They compute the Euclidean distance between two users and apply the clustering algorithm to group users. To make a recommendation for a user, the weights of each music group associated with the relevant users in the same group will be averaged. These average weights will be recorded in a *reference table* for the user.

The CB method recommends recently hot music to users according to the access history of users. But in the CB method, the formula of computing music group weight pays much attention to the weight of the transaction occurring time. Table I is an example of access history H1 of a user. In Table I, we focus on the group weight of groups A and B. We find group A appearing many times in the early transactions. On the other hand, group B appears one time in the latest transaction. But the group weight of group B is larger than the group weight of group A in the CB method, the result is not conventional. Observing the result of the CB method, we can find when the music group B appears once in the

Manuscript received August 8, 2011.

Y.-I. Chang is with the Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C. (e-mail: changyi@cse.nsysu.edu.tw).

C.-C. Wu is with the Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C. (corresponding author; e-mail: wucc5501@gmail.com).

M.-C. Tsai was with the Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, R.O.C. (e-mail: tsaimc@db.cse.nsysu.edu.tw).

TABLE II
THE ACCESS HISTORY OF USER U1

Transaction	Music Group in Transaction
T1	ABC
T2	BEF
T3	ABE
T4	BDF
T5	ACD
T6	AG

TABLE III
THE ACCESS HISTORY OF USER U2

Transaction	Music Group in Transaction
T1	ABC
T2	AEF
T3	ABE
T4	ADF
T5	BCD
T6	BG

later transaction, it will have larger group weight than the group weight of the music group A which appears many times in the earlier transaction. This result may be not good for some users, because the purpose of the CB method is to recommend the music object which the users are interested. When the count of music group is large in the user's access history, it means that this user is interested in this group, too.

The COL method uses *large-1 itemsets* and *large-2 itemsets* to be the user interests and behaviors, the result that they do not care the last transaction which the target group shows may let the users that have different interests be grouped together. Note that an itemset is large if the support of the item is larger than minimum support. Large-1 itemset means the size of the item is 1. For example, Table II and Table III show the access histories of user U1 and U2, respectively. In Table II, we let the density of the appearance of group B is larger than group A of user U1. That is, the transactions which group B appears are closer than those of group A in access history of user U1. On the other hand, we let the density of the appearance of group A is larger than group B of user U2. In the COL method, the support for each music group of these two users are the same. However, the density of group A of user U2 is larger than the density of group A of user U1. On the other hand, the density of group B of user U1 is larger than the density of group B of user U2. This result is not suitable for these two users in this example obviously.

Therefore, in this paper, to be fair, we propose the TICI (Transaction-Interest-Count-Interest) method to improve the performance of the CB method. In our TICI method, for the same access history shown in Table I, we can decide the rank of the music group weight between groups A and B. And we put two parameters: Count-Interest CI and Transaction-Interest TI in our TICI method to let user choose which weight they want to emphasize. We also propose the DI (Density-Interest) method to improve the performance of the COL method. In our DI method, we can distinguish the difference between two access histories of two users respectively shown in Table II and Table III by calculating the supports of music groups and considering the distributions of appearances of music group. From the simulation results, we show that our TICI method could provide better performance than the CB method in terms of the weight differences. That is, our TICI method can decide the rank of the group weight precisely. Moreover, we also show that our DI method also could provide better performance than the COL method in terms of the Hamming distance. That is, our DI method can distinguish the users who have different access behaviors obviously.

TABLE IV
A SAMPLE OF THE ACCESS HISTORY

Access Time	Object ID	Music Group	Transaction
2001/4/06 AM 11:47:03	1	B	T1
2001/4/06 AM 11:47:03	23	C	T1
2001/4/12 AM 10:11:25	7	D	T2
2001/4/12 AM 10:11:25	5	C	T2
2001/4/12 AM 10:11:25	32	B	T2
2001/4/16 AM 09:51:33	16	A	T3
2001/4/16 AM 09:51:33	19	B	T3
2001/4/16 AM 09:51:33	42	A	T3
2001/4/20 AM 08:31:12	31	D	T4
2001/4/20 AM 08:31:12	63	C	T4
2001/4/20 AM 08:31:12	26	A	T4
2001/4/22 AM 10:24:49	53	B	T5
2001/4/22 AM 10:24:49	12	A	T5

The rest of the paper is organized as follows. Section 2 gives a survey of some music recommendation systems. Section 3 presents the proposed TICI method. Section 3 presents the proposed DI method. Section 5 makes a comparison between our TICI method and CB method. Finally, Section 6 gives the conclusions.

II. RELATED WORK

The music objects in the database of the Music Recommendation System (MRS), as well as the incoming music objects, are candidates for music recommendation. When a new music object is inserted in the database of the MRS, it goes through the track selector and the feature extractor. According to the extracted features, the incoming music object is properly assigned to certain music group by the classifier function block. When the user accesses a music object from the list of music objects or the recommendation results, the profile manager will record the object information into the access history. An example of the access history is shown in Table IV. In Table IV, the information of each accessed music object, *i.e.*, the access time, the object ID, the corresponding music group which the object belongs to, and the corresponding transaction is recorded in the access history. Note that the transaction ID is monotonically increasing.

In this Section, we introduce Music Recommendation Based on Music Data Grouping and User Interests [5] that can use music data grouping and user interests for music recommendation. There are two recommendation methods. First, we describe the CB method briefly. Then, we describe the COL method.

A. The CB Method

Arbee L.P. Chen *et al.* have proposed the CB method to recommend the music objects that belong to the music groups the user is recently interested in [5]. Instead of textual descriptions, they consider the perceptual properties of music objects, such as pitch, duration, and loudness, which can

TABLE V
THE PREFERENCE TABLE FOR THE USER

Music Group	Weight
A	3.08
B	2.5616
C	1.7216
D	1.312

TABLE VI
NUMBER OF MUSIC OBJECTS TO THE RECOMMENDED IN EACH GROUP

Music Group	Number of Recommended Music Objects
A	8
B	6
C	4
D	4

be directly extracted from the music objects. For users, the preferences are derived from the access histories and recorded in profiles.

To capture the recent interests of the user, they analyze the latest transactions in the access history as follows. Each transaction is assigned a different weight, where the latest transaction has the highest weight. The weight GW_i of music group G_i is computed as follows:

$$GW_i = \sum_{j=1}^{n} TW_j * MO_{j,i} \tag{1}$$

where TW_j is the weight of transaction T_j, n is the number of latest transactions used for analysis, $M0_{j,i}$ is the number of music objects which belong to music group G_i in transaction T_j. These weights will be recorded in a preference table for the user. The MRS ranks all the music groups.

To avoid recommending a large number of music objects to users, the MRS limits the number of music objects for recommendation. The number of music objects R_j from each music group is decided as follows:

$$R_i = \lceil N * \frac{GW_i}{\sum_{k=1}^{M} GW_k} \rceil \tag{2}$$

where N is the number of music objects in the recommendation list, GW_i is the weight of the target group, M is the total number of music groups in MRS. In the same music group, the latest music object will be first recommended.

Example. Take the user's access history shown in Table IV as an example. Assign the weights 0.4096, 0.512, 0.64, 0.8, and 1 to Tl, T2, T3, T4, and T5, respectively. The weight for each music group is calculated, as shown in Table V.

According to Table V, the total weight of all music groups is 8.6752. Suppose the number of music objects to be recommended is 20. The result of recommending music objects is shown in Table VI.

B. The COL Method

The CB method tends to provide expected and interesting music objects for users. Based on the collaborative approach, the purpose of the COL method is to provide unexpected findings due to the information sharing between *relevant users*.

In the COL method, they group the users first. They apply the technique proposed in [16] for user grouping. The *large-l itemsets* derived from transactions in the access history are used for user interests and the *large-2 itemsets* are used for user behaviors.

TABLE VII
THE ACCESS HISTORY OF A USER

Transaction	Music Group in Transaction
T1	A,C,E
T2	B,C,E,F
T3	C,D,E,F
T4	B,C,D,F
T5	A,G

TABLE VIII
THE INTEREST TABLE

Music Group	Count	First Transaction (FT)	Last Transaction (LT)
A	2	T1	T5
B	2	T2	T4
C	4	T1	T4
D	2	T3	T4
E	3	T1	T3
F	3	T2	T4
G	1	T5	T5

Example. Suppose there are five transactions in the access history as shown in Table VII. They construct the interest table from the access history of the corresponding user as shown in Table VIII. The support of a music group is calculated by the following formula:

$$support = \frac{count}{T_c - FT + 1}$$

where T_c denotes the current transaction number. Suppose the T_c is 5. The support for each music group is shown in Table IX. If the minimum support is 75%, there will be three large-l itemsets, *i.e.*, music groups C, F and G, which form the *interest profile* for the user.

Example. Take the access history shown in Table VII for example. They construct the behavior table and compute the support of each music group pair as shown in Table X.

If the minimum support is 0.65, there will be four large-2 itemsets, *i.e.*, pairs AG, CD, CF and DF, which form the the *behavior profile* {AG, CD, CF, DF} is derived for the user. And then they construct an I-B matrix and transform it into an I-B vector. The I-B matrix of the user is shown in Table XI. Then, they transform the I-B matrix to the I-B vector (0000001 000000 11010 0010 000 10 1). Therefore, each user has a corresponding I-B vector. According to the I-B vector, they compute the Euclidean distance between two users and apply the clustering algorithm to group users. To make a recommendation for a user, the weights of each music group associated with the relevant users in the same group

TABLE IX
THE SUPPORT OF THE MUSIC GROUPS

Music Group	Support
A	0.4
B	0.5
C	0.8
D	0.67
E	0.6
F	0.75
G	1

TABLE X
THE BEHAVIOR TABLE WITH THE CORRESPONDING SUPPORT

Music Group Pair	Count	FT	LT	Support
AC	1	T1	T1	0.2
AE	1	T1	T1	0.2
AG	1	T5	T5	1
BC	2	T2	T4	0.5
BD	1	T4	T4	0.5
BE	1	T2	T2	0.25
BF	2	T2	T4	0.5
CD	2	T3	T4	0.67
CE	3	T1	T3	0.6
CF	3	T2	T4	0.75
DE	1	T3	T3	0.33
DF	2	T3	T4	0.67
EF	2	T2	T3	0.5

TABLE XI
THE I-B MATRIX

	A	B	C	D	E	F	G
A	0	0	0	0	0	0	1
B		0	0	0	0	0	0
C			1	1	0	1	0
D				0	0	1	0
E					0	0	0
F						1	0
G							1

will be averaged. These average weights will be recorded in a *reference table* for the user. The following example shows the process to construct a reference table and to make recommendation using the COL method.

Example. There are three persons U1, U2, and U3 in user group U. Table XII shows the partial access histories of U1, U2 and U3. Assign the weights 1, 0.8, 0.64, 0.512, and 0.4096 to the latest five transactions, respectively. They apply the equation 1 in the CB method, the result is shown in Table XIII. To make a recommendation for U1 for example, the reference table for U1 is constructed as shown in Table XIV. The weight for each music group in the reference table is subtracted from that in the preference table, and the result is shown in Table XV. The music group with zero or negative weight difference will not be recommended to the user. Therefore, they recommend music groups C, E, and F to U1. Apply equation 2 , the recommendable result for U1 is shown in Table XVI. Note that the M in equation is set to 3 in this case.

TABLE XII
THE LATEST FIVE TRANSACTIONS IN THE ACCESS HISTORIES OF USERS U1, U2, AND U3

Partial access history of user U1

Music Group	Transaction
B	T8
C	T8
D	T9
C	T9
B	T9
A	T10
B	T10
A	T10
D	T11
C	T11
A	T11
B	T12
A	T12

Partial access history of user U2

Music Group	Transaction
E	T13
F	T13
A	T14
A	T14
B	T15
C	T15
A	T16
D	T16
B	T17
A	T17
B	T17
E	T17

Partial access history of user U3

Music Group	Transaction
A	T11
C	T11
B	T12
B	T12
A	T13
A	T13
D	T14
C	T14
F	T14
A	T15
C	T15
B	T15
D	T15
C	T15

TABLE XIII
THE PREFERENCE TABLES FOR USERS U1, U2 AND U3

Preference table of user U1

Music Group	Weight
A	3.08
B	2.5616
C	1.7216
D	1.312

Preference table of user U2

Music Group	Weight
A	2.824
B	2.64
C	0.64
D	0.8
E	1.4096
F	0.4096

Preference table of user U3

Music Group	Weight
A	2.6896
B	2.024
C	3.2096
D	1.8
F	0.8

TABLE XIV
THE REFERENCE TABLE FOR USER U1

Music Group	Weight
A	2.7568
B	2.332
C	1.9248
D	1.3
E	0.7048
F	0.6048

TABLE XV
THE TABLE OF WEIGHT DIFFERENCES

Music Group	Weight Difference
A	-0.3232
B	-0.2296
C	0.2032
D	-0.012
E	0.7048
F	0.6048

III. THE TRANSACTION-INTEREST-COUNT-INTEREST METHOD

In [5], Chen *et al.* proposed a CB method for recommending music object. Although their CB method can find the recently hot music group according to the user's access history, the result is not fair. They pay much attention to the weight of time. Therefore, we propose a fair formula which emphasizes both the weight of time and the weight of count of music group. In this section, we will present our proposed Transaction-Interest-Count-Interest(TICI) method. First, we will give some assumptions and notations. Next, we will introduce our proposed formula, which can emphasize both counting and time.

First, we give initial conditions of music recommendation system [5]. When a user accesses a music object from the list of music objects or the recommendation results, the profile manager will record the object information into the access history. An example of the access history H1 is shown in Table IV.

As shown in Table IV, the information of each accessed music object, *i.e.*, the access time, the object ID, the corresponding music group which the object belongs to, and the corresponding transaction are recorded in the access history. Note that the transaction ID is monotonically increasing. Each transaction is assigned a different weight,

TABLE XVI
NUMBER OF MUSIC OBJECTS TO BE RECOMMENDED IN EACH GROUP FOR USER U1

Music Group	Number of Recommended Music Objects
C	2
E	10
F	8

TABLE XVII
DESCRIPTION OF PARAMETERS

Parameter	Description
TID	The target transaction ID
$FTID$	The first transaction ID
Nt	The number of transactions
TW_j	The weight of transaction T_j
GW_i	The weight of music group G_i
CI	The interest of the count
TI	The interest of the transaction
$GA_{j,i}$	The number of appearances of music group i in transaction j

where the latest transaction has the largest weight. Moreover, the music group containing more accessed music objects in a transaction has a larger weight than other groups in the same transaction. According to the weights of music groups, different numbers of music objects from the music groups will be recommended. For music group G_i, we select the latest R_i music objects which have not been accessed by the user. In the recommendation list, the music objects will be sorted by the corresponding group.

Although the CB method can find the recently hot music group according to the user's access history [5], the result is not fair. They pay much attention to the weight of time. Therefore, we propose a fair formula which emphasizes both the weight of time and the weight of count of music group.

To simplify the comparison, first, we assume that two same groups will not appear in one transaction. We use the example that is an access history H1 which has six transactions to compare the results between our formula and the CB method. Table XVII shows the parameters used in our proposed method.

The formula for the group weight of group i (GW_i) of CB method is

$$GW_i = \sum_{j=1}^{n} TW_j * GA_{j,i} \tag{3}$$

where TW_j is the weight of transaction j and $GA_{j,i}$ is the number of music objects which belong to music group G_i in transaction T_j. In this formula, the equation of TW_j is not given in [5]. Therefore, we give a new equation of TW_j:

$$TW_j = \frac{TID_j - FTID + 1}{Nt} \tag{4}$$

where TID_j is the target transaction ID, $FTID$ is the first transaction ID and Nt is the number of transactions. Note that the following TW_j is calculated by using this equation.

According to Formula 3, when the music group A appears once in the later transaction, it will have larger weight than the weight of the music group B appears many times in the earlier transaction. This result is not good for some users, because the purpose of the CB method is to recommend the music object which the users are interested. When the count of music group is large in the user's access history, it means that this user is interested in this group, too. Therefore, we propose a new formula to avoid this problem:

$$GW_i = \sum_{j=1}^{n} TW_j * TI + GA_{j,i} * CI \tag{5}$$

TABLE XVIII
THE ACCESS HISTORY OF USER U1

Transaction	Music Group in Transaction
T1	ABC
T2	BEF
T3	ABE
T4	DF
T5	CD
T6	AG

TABLE XIX
THE ACCESS HISTORY OF USER U2

Transaction	Music Group in Transaction
T1	ABC
T2	AEF
T3	ABE
T4	DF
T5	CD
T6	BG

where Count-Interest CI ($0 \leq CI \leq 1$) and Transaction-Interest TI ($0 \leq TI \leq 1$) are assigned by users, TI = 1 - CI. According to each user's preferences, our formula adds two parameters CI and TI to let users decide the importance of the time and count.

IV. The Density-Interest Method

Chen 's COL method [5] used the following formula:

$$support = \frac{C_i}{T_C - T_S + 1} \tag{6}$$

to calculate the support of a music group, where C_i is the count of appearances of the target group i, T_C is the current transaction number and T_S is the first transaction number that the target group shows. But in this formula, we can find that they do not care the last transaction which the target group shows. Because they use *large-1 itemsets* and *large-2 itemsets* to be the user interests and behaviors, the result that they do not care the last transaction which the target group shows may let the users that have different interests be grouped together. For example, music group A appears in transaction T1, T3 and T6 in the access history of the user U1, and group A appears in transaction T1, T2 and T3 in the access history of the user U2. The distributions of appearances of group A in this two access histories are very different, but the supports of A for two users are the same by Formula 6. Therefore, we propose a new formula that added with the density of appearance of the target group:

$$support = \frac{C_i}{T_C - T_S + 1} * (1 - DI) + \frac{C_i}{T_E - T_S + 1} * DI \tag{7}$$

where T_E is the last transaction number that the target group shows, DI is the interest of density and C_i is the count of the appearances of the target group. We use the $\frac{C_i}{T_E - T_S + 1}$ to stand for the density of the appearance of the music group. We set DI = 0.5 in the following comparison of the situation with the COL method.

To simplify the comparison, first, we assume that two same groups will not appear in one transaction. Table XVIII and Table XIX show the access history of user U1 and U2.

In Table XVIII, we let the density of the appearance of group B is larger than group A of user U1. On the other

TABLE XX
THE SUPPORT OF THE MUSIC GROUPS BY THE COL METHOD

Music Group	Support
A	$\frac{3}{6}$
B	$\frac{3}{6}$
C	$\frac{2}{6}$
D	$\frac{2}{3}$
E	$\frac{2}{5}$
F	$\frac{2}{5}$
G	1

TABLE XXI
THE SUPPORT OF THE MUSIC GROUP PAIRS OF USER U1 BY THE COL METHOD

Music Group Pair	Support
AB	$\frac{2}{6}$
AC	$\frac{1}{6}$
AE	$\frac{1}{4}$
AG	1
BC	$\frac{1}{6}$
BE	$\frac{2}{5}$
BF	$\frac{1}{5}$
CD	$\frac{1}{2}$
DF	$\frac{1}{3}$
EF	$\frac{1}{5}$

hand, we let the density of the appearance of group A is larger than group B of user U2. In the COL method, we can find the support for each music group that is calculated by Formula 6 are the same. The support is shown in Table XX. We use a new equation to calculate the minimum support:

$$MinSupport = \frac{\sum_{i \in GS} GSupport_i}{GN} \quad (8)$$

where $GSupport_i$ is the support of music group i, GN is the number of music groups and GS is the set of music groups. In this example, the minimum support is $\frac{114}{210}(\approx 0.543)$. That is, $(\frac{3}{6}+\frac{3}{6}+\frac{2}{6}+\frac{2}{3}+\frac{2}{5}+\frac{2}{5}+1)/7 = \frac{114}{210}$. There will be two large-1 itemsets, *i.e.*, music groups D and G.

Table XXI and Table XXII show the supports of the music group pairs of user U1 and U2, respectively.

We use a formula to calculate the support of the music group pairs:

$$MinSupport = \frac{\sum_{j \in GPS} GPSupport_j}{GPN} \quad (9)$$

TABLE XXII
THE SUPPORT OF THE MUSIC GROUP PAIRS OF USER U2 BY THE COL METHOD

Music Group Pair	Support
AB	$\frac{2}{6}$
AC	$\frac{1}{6}$
AE	$\frac{2}{5}$
AF	$\frac{1}{5}$
BC	$\frac{1}{6}$
BE	$\frac{1}{4}$
BG	1
CD	$\frac{1}{2}$
DF	$\frac{1}{3}$
EF	$\frac{1}{5}$

TABLE XXIII
THE I-B MATRIX OF USER U1 BY THE COL METHOD

	A	B	C	D	E	F	G
A	0	0	0	0	0	0	1
B		0	0	0	1	0	0
C			0	1	0	0	0
D				1	0	0	0
E					0	0	0
F						0	0
G							1

TABLE XXIV
THE I-B MATRIX OF USER U2 BY THE COL METHOD

	A	B	C	D	E	F	G
A	0	0	0	0	1	0	0
B		0	0	0	0	0	1
C			0	1	0	0	0
D				1	0	0	0
E					0	0	0
F						0	0
G							1

where $GPSupport_j$ is the support of music group pair j, GPN is the number of music group pairs and GPS is the set of music group pairs. In this example, the minimum supports of users U1 and U2 are also equal to $\frac{71}{200}(\approx 0.355)$. There will be three large-2 itemsets for users U1 and U2, *i.e.*, music group pairs AG, BE and CD are large-2 itemsets for user U1 and music group pairs AE, BG and CD are large-2 itemsets for user U2. The I-B matrix of users U1 andU2 are shown in Table XXIII and Table XXIV, respectively. The I-B vector of user U1 is (0000001 000100 01000 1000 000 00 1) and the I-B vector of user U2 is (0000100 000001 01000 1000 000 00 1).

Then, we use Formula 7 to calculate the support of this example. Note that DI is 0.5. The supports of users U1 and U2 are shown in Table XXV and Table XXVI, respectively.

By Formula 8, the minimum supports of users U1 and U2 are equal to 0.7. Therefore, music groups B, D, E and G are large-1 itemsets for user U1 and music groups A, D, E and G are large-1 itemsets for user U2. Table XXVII and Table

TABLE XXV
THE SUPPORT OF THE MUSIC GROUPS OF USER U1 BY THE DI METHOD

Music Group	Support
A	$\frac{1}{2}$
B	$\frac{3}{4}$
C	$\frac{7}{12}$
D	$\frac{5}{6}$
E	$\frac{7}{10}$
F	$\frac{8}{15}$
G	1

TABLE XXVI
THE SUPPORT OF THE MUSIC GROUPS OF USER U2 BY THE DI METHOD

Music Group	Support
A	$\frac{3}{4}$
B	$\frac{1}{2}$
C	$\frac{7}{12}$
D	$\frac{5}{6}$
E	$\frac{7}{10}$
F	$\frac{8}{15}$
G	1

TABLE XXVII
THE SUPPORT OF THE MUSIC GROUP PAIRS OF USER U1 BY THE DI METHOD

Music Group Pair	Support
AB	$\frac{2}{3}$
AC	$\frac{7}{12}$
AE	$\frac{5}{8}$
AG	1
BC	$\frac{7}{12}$
BE	$\frac{7}{10}$
BF	$\frac{3}{5}$
CD	$\frac{3}{4}$
DF	$\frac{2}{3}$
EF	$\frac{3}{5}$

TABLE XXVIII
THE SUPPORT OF THE MUSIC GROUP PAIRS OF USER U2 BY THE DI METHOD

Music Group Pair	Support
AB	$\frac{2}{3}$
AC	$\frac{7}{12}$
AE	$\frac{7}{10}$
AF	$\frac{3}{5}$
BC	$\frac{7}{12}$
BE	$\frac{5}{8}$
BG	1
CD	$\frac{3}{4}$
DF	$\frac{2}{3}$
EF	$\frac{3}{5}$

XXVIII shows the supports of the music group pairs of user U1 and U2, respectively.

By Formula 9, the minimum supports of users U1 and U2 are equal to 0.6775. Therefore, music group pairs AG, BE, and CD are large-2 itemsets for user U1 and music groups AE, BG, and CD are large-2 itemsets for user U2. The I-B matrix of users U1 andU2 are shown in Table XXIX and Table XXX, respectively. The I-B vector of user U1 is (0000001 100100 01000 1000 100 00 1) and the I-B vector of user U2 is (1000100 000001 01000 1000 100 00 1).

According to the I-B vector, we use the Hamming distance [9] to compute the Euclidean distance between two I-B vectors which belong to two users respectively. The number of dissimilar bits is the Hamming distance between two vectors. As the number of dissimilar bits is larger, the Hamming distance is larger. When the Hamming distance is large, it means that these two users have different access behaviors. In the COL method, the Hamming distance is 4; that is, there are four dissimilar bits. On the other hand, the Hamming distance is 6 in our DI method. In this example, the Hamming distance of our DI method is larger than the Hamming distance of the COL method, the probability of our DI method group these two users together is smaller than that of the COL method. This result can prove that our DI method can distinguish the users who have different access behaviors obviously.

TABLE XXIX
THE I-B MATRIX OF USER U1 BY THE DI METHOD

	A	B	C	D	E	F	G
A	0	0	0	0	0	0	1
B		1	0	0	1	0	0
C			0	1	0	0	0
D				1	0	0	0
E					1	0	0
F						0	0
G							1

TABLE XXX
THE I-B MATRIX OF USER U2 BY THE DI METHOD

	A	B	C	D	E	F	G
A	1	0	0	0	1	0	0
B		0	0	0	0	0	1
C			0	1	0	0	0
D				1	0	0	0
E					1	0	0
F						0	0
G							1

TABLE XXXI
PARAMETERS USED IN THE EXPERIMENT

Parameters	**Meaning**
N	The number of transactions in the access history
M	The number of music groups
MinT	The minimum length of the transaction
MaxT	The maximum length of the transaction

V. PERFORMANCE

In this section, we study the performance of the proposed TICI and DI methods. We also make a comparison with the CB method and COL method. The simulation was performed on an Intel Pentium Core2 1.86G Hz CPU computer with 1GB of RAM, and the operation system is Microsoft Windows XP service pack 3.

A. Generation of Synthetic Data

We generated synthetic access histories to evaluate the performance of the methods. The parameters used in the generation of the synthetic data are shown in Table XXXI. The length of a transaction is chosen randomly between *MinT* and *MaxT*. For the TICI method, the *MinT* and *MaxT* is 2 and 5, respectively. In the comparison between the CB method and the TICI method, the music group will appear one time in a transaction or appear more than one time in a transaction. Therefore, for the music group appears more than one time in a transaction, we choose the music group in the set of music group randomly. For the music group appears one time in a transaction, we use the flag to record the appearance of the music group so that the music group will not appear again in a transaction. This way can achieve the goal that the music group appears one time in a transaction. For generating synthetic data, we assign an *occur rate Orate*, and we generate a random real number which is between 0 and 1. If the random number is larger than *Orate*, the generation runs normally. If the random number is smaller than or equal to *Orate*, we let the music group which never appears in the earlier transactions appear in the last transaction. The larger *Orate* is, the larger repeatability of the music group is. We call this synthetic data *DataType1*.

For the DI method, the *MinT* and *MaxT* is 5 and 10, respectively. In the comparison between the COL method and the DI method, the music group will appear only one

TABLE XXXII
THE ACCESS HISTORY OF USER U1

Transaction	Music Group
T1	A,B,C,D,E
T2	A,C,E
T3	A,B,C,D,E
T4	C,E,K
T5	A,D,E,M

TABLE XXXIII
THE ACCESS HISTORY OF USER U2

Transaction	Music Group
T1	A,B,C,D,E
T2	B,C,E
T3	A,B,C,D,E
T4	C,E,K
T5	B,D,E,M

time in a transaction. Therefore, we use the flag to record the appearance of the music group so that the music group will not be chosen again in a transaction. And we generate two user access histories. First, we generate one access history. In generating synthetic data, we choose two target groups, for example, A and B. We assign two *density values* d_A, d_B to decide the appearances of the target groups, and we generate two random real numbers, r_A and r_B, which are between 0 and 1. If r_A is smaller than or equal to d_A, A appears. Similarly, r_B is smaller than or equal to d_B, B appears. Specially, we let two target groups must appear in the first transaction. When one access history is generated, the second access history is generated by changing the appearances of two target groups and other groups are the same. For example, Table XXXII and Table XXXIII are two access histories generating by our method. We call this synthetic data *DataType2*.

B. Simulation Results of Synthetic Data

In this subsection, first, we discuss the simulation results of the TICI method and the CB method under data *DataType1*. Then, we discuss the simulation results of the DI method and the COl method under data *DataType2*.

1) TICI $vs.$ CB: In this subsection, we make a comparison of our TICI method with the CB method by using the synthetic data *DataType1*. We study the impact of five parameters on Table XXXI. We let *M* be 50 because the numbers of music groups are not more than 50 in the current environment and we let *MinT* = 2 and *MaxT* = 5. Moreover, we make the comparison between the TICI method and the CB method under the two cases. One case is that the music group can appear more than one time in a transaction, another case is that the music group can only appear one time in a transaction. A comparison of the music group weight difference in the TICI method and CB method is shown in Figure 1 and Figure 2.

Note that our TICI method has three cases: CI = 0.3, CI = 0.5 and CI = 0.7 to compare the change of the result that emphasizes weight of transaction and the weight of count. We use the group weight difference to be our performance measure. The group weight difference is the difference between the group weights of the group weight rank which are decided by the methods and we add all group weight differences to be the results of comparison between our TICI method and the CB method. When the group weight difference is larger, it means that the method can decide the rank of the group weight clearly.

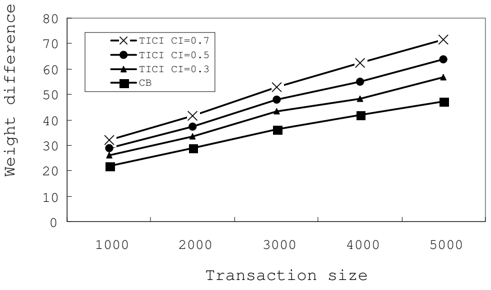

Fig. 1. A comparison of the group weight difference under the case that the music group appears more than one time in a transaction

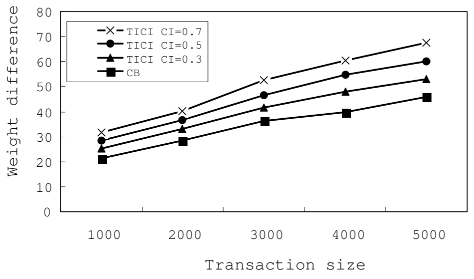

Fig. 2. A comparison of the group weight difference under the case that the music group only appears one time in a transaction

In Figure 1, the range of N is set to 1000, 2000, 3000, 4000 and 5000, while the other parameters are kept as their base values. Under changing the value of N, we can find that the group weight differences of our TICI method are larger than that of the CB method. Because the CB method only emphasizes the weight of transaction, the impact of the count of music group decreases when the transaction size increases. For example, when the transaction size is 1000, the transaction weight of Transaction 5 is $\frac{5}{1000}$ and the transaction weight of Transaction 900 is $\frac{900}{1000}$. The group weight of group A that appears five times in Transaction 5 is still smaller than the group weight of group B which appears one time in Transaction 900. Therefore, the group weights are usually the same; that is, the CB method usually can not decide the rank of the group weight so that the group weight difference of the CB method is small.

In the three cases of our TICI method, we can find that the rank of the group weight difference is CI = 0.7, CI = 0.5 and CI = 0.3. According to the result, we can find when we emphasizes the weight of transaction, *i.e.*, CI = 0.3 and TI = 0.7, the group weight difference is smaller than other cases. When we emphasizes the weight of count, *i.e.*, CI = 0.7 and TI = 0.3, the group weight difference is larger than other cases. The reason is that the impact of the count is larger than the impact of the transaction weight. For example, the transaction size is 1000 and the transaction weight is between $\frac{1}{1000}$ and $\frac{1000}{1000}$. On the other hand, the music group at least appears one time, the count is larger than or equal to the transaction weight. Therefore, the impact of the count is always larger than the impact of

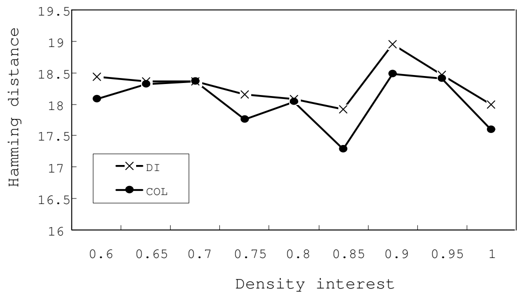

Fig. 3. A comparison of the Hamming distance under the case that the transaction size is 100

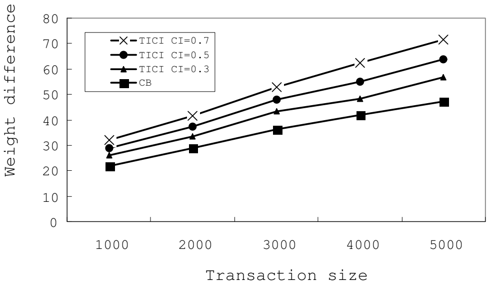

Fig. 4. A comparison of the Hamming distance under the case that the density interest is 1

TABLE XXXIV
A COMPARISON OF THE HAMMING DISTANCE UNDER THE CASE THAT THE TRANSACTION SIZE IS 100

Density Interest	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
DI	18.44	18.36	18.36	18.16	18.08	17.92	18.96	18.46	18
COL	18.08	18.32	18.36	17.76	18.04	17.28	18.48	18.4	17.6

TABLE XXXV
A COMPARISON OF THE HAMMING DISTANCE UNDER THE CASE THAT THE DENSITY INTEREST IS 1

Transaction Size	1000	1500	2000	2500	3000
DI	6.0645	3.625	2.1111	1.3684	0.5143
COL	6	3.4375	1.8888	1.2105	0.4571

the transaction weight. When the transaction size increases, the impact difference between the count and the transaction weight increases. Therefore, the larger the transaction size is, the larger the group weight difference is.

From Figure 1 and Figure 2, we can find the group weight difference in Figure 1 is larger than the group weight difference in Figure 2. Because the data in Figure 1 allows that the music group can appear more than one time in a transaction, and the data in Figure 2 only allows that the music group appear one time in a transaction. The count of the music group in Figure 1 is larger than the count of the music group in Figure 2. Therefore, the group weight difference in Figure 1 is larger than the group weight difference in Figure 2.

2) DI vs. COL: In this subsubsection, we make a comparison of our DI method with the COL method by using the synthetic data *DataType2*. We study the impact of five parameters on Table XXXI. We let *M* be 50 and we let *MinT* = 5 and *MaxT* = 10. We use the Hamming distance to be our performance measure. The Hamming distance is the counts of the different bits between two bit strings. Because our DI method and COL method will use a bit string to stand for a user. Therefore, if the Hamming distance is large, the counts of the different bits between two bit strings are large, which also means the difference between two users is large. Figure 3 shows a comparison of the Hamming distance in the DI method and the COl method under the case that the transaction size is 100. Note that the density interest in the DI method is from 0.6 to 1. The details of this result are shown in Table XXXIV. Figure 4 shows a comparison of the Hamming distance in the DI method and the COl method under the case that the density interest is 1. The details of this result are shown in Table XXXV.

In Figure 3 and Table XXXIV, we can find that when the density interest increases, the Hamming distance of our DI method is larger than that of the COL method. The reason is that when the density interest increases, it means that we emphasize the density of the appearance of the music group. Therefore, we can distinguish the difference between two access histories of two users respectively clearly. In Figure 4 and Table XXXV, we can find that when the transaction size becomes large, the Hamming distance of our DI method is larger than that of the COL method. Note that when the transaction size become large, the Hamming distances of our DI method and the COL method become small. The reason is that the density is calculated by the counts of the target group or group pair divided by the last transaction that the target group or group pair appears in minus the first transaction which the target group or group pair. Therefore, when the transaction size increases, the impact of the density decreases.

VI. CONCLUSION

In this paper, we first have proposed the TICI (Transaction-Interest-Count-Interest) method for the music recommendation in music databases. The TICI method can improve the performance of the CB method by change the formula which calculates the weight of music group. Then, we have proposed the DI method to improve the performance of the COL method by change the formula which calculates the supports of music group. We also have studied the performance of the TICI method and the CB method, and the DI method and the COL method. The simulation results have shown that the performance of the TICI method is better than that of the CB method in terms of the weight difference. This is because the TICI method considers the count of the music group and the time of appearance of the music group so that the TICI method can decide the rank of the group weight more precisely than the CB method. The simulation results have also shown that the performance of the DI method is better than that of the COL method in terms of the Hamming distance, because the DI method considers the density of the appearance of the music group so that the DI method can distinguish the users who have different access behaviors more obviously than the COL method.

References

[1] D. B. And, D. Billsus and M. J. Pazzani, “A Hybrid User Model for News Story Classification,” in *Proc. of the 7th Int. Conf. on User Modeling*, Banff, Canada, 1999, pp. 99-108.

[2] M. Balabanovic and Y. Shoham, “Content-based, Collaborative Recommendation,” *Commun. ACM*, vol. 40, no. 3, pp. 66-72, 1997.

[3] C. Basu, H. Hirsh and W. W. Cohen, “Recommendation as Classification: Using Social and Content-Based Information in Recommendation,” in *Proc. of National Conf. on Artificial Intelligence*, Madison, Wisconsin, 1998, pp. 714-720.

[4] Y.-I. Chang, C.-C. Wu and M.-C. Tsai, “A User-Interests Approach to Music Recommendation,” *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2011*, WCE 2011, 6-8 July, 2011, London, U.K., pp. 1813-1817.

[5] H. C. Chen and A. L. P. Chen, “A Music Recommendation System Based on Music Data Grouping and User Interests,” in *Proc. of the 10th Int. Conf. on Information and Knowledge Management*, Atlanta, Georgia, 2001, pp. 231-238.

[6] Y. Hijikata, K. Iwahama and S. Nishida, “Content-based Music Filtering System with Editable User Profile,” in *Proc. of ACM Symp. on Applied Computing*, Kennesaw, Georgia, 2006, pp. 1050-1057.

[7] B. Krulwich and C. Burkey, “Learning User Information Interests Through Extraction of Semantically Significant Phrases,” in *Proc. of the AAAI Spring Symposium on Machine Learning in Information Access*, Technical Papers, Stanford, 1996.

[8] K. Lang, “NewsWeeder: Learning to Filter Netnews,” in *Proc. of the 12th Int. Machine Learning Conf.*, Tahoe City, California, 1995, pp. 331-339.

[9] A. X. Liu, K. Shen, E. Torng, “Large scale Hamming distance query processing,” in *Proc. of the IEEE 27th Int. Conf. on Data Engineering*, Hannover, Germany, 2011, pp. 553-564.

[10] E. Perik and B. de Ruyter, P. Markopoulos and B. Eggen, “The Sensitivities of User Profile Information in Music Recommender Systems,” in *Proc. of the 2nd Annual Conf. on Privacy, Security and Trust*, New Brunswick, Canada, 2004, pp. 137-141.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*, 1993.

[12] J. Rucker, and M. J. Polanco, “Siteseer: Personalized Navigation for the Web,” *Commun. ACM*, vol. 40, no. 3, pp. 73-76, 1997.

[13] U. Shardanand and P. Maes, “Social Information Filtering: Algorithms for Automating “Word of Mouth”,” in *Proc. of ACM Conf. on Human Factors in Computing Systems*, Denver, Colorado, 1995, pp. 210-217.

[14] C. W. Shih, M. Y. Chen, H. C. Chu, and Y. M. Chen, “Recommendation System Using Information Needs Radar Model,” *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2011*, WCE 2011, 6-8 July, 2011, London, U.K., pp. 1901-1904.

[15] B. Smyth and P. Cotter, “A Personalized Television Listings Service,” *Communications of the ACM*, vol. 43, no. 8, pp. 107-111, 2000.

[16] Y. H. Wu and Y. C. Chen and A. L. P. Chen, “Enabling Personalized Recommendation on the Web Based on User Interests and Behaviors,” in *Proc. of IEEE Int. Workshop on Research Issues in Data Eng.*, Heidelberg, Germany, 2001, pp. 17-24.