GENDER IDENTIFICATION AND AGE ESTIMATION OF USERS BASED ON MUSIC METADATA

Ming-Ju Wu
Computer Science Department
National Tsing Hua University
Hsinchu, Taiwan
brian.wu@mirlab.org

Jyh-Shing Roger Jang
Computer Science Department
National Taiwan University
Taipei, Taiwan
roger.jang@mirlab.org

Chun-Hung Lu
Innovative Digitech-Enabled Applications
& Services Institute (IDEAS),
Institute for Information Industry,
Taipei, Taiwan
enricoghlu@iii.org.tw

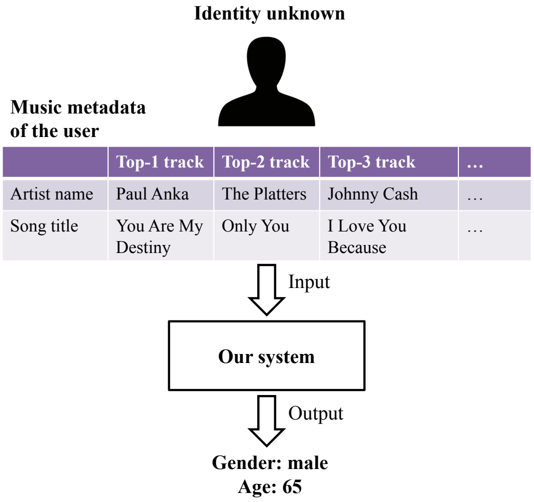
ABSTRACT

Music recommendation is a crucial task in the field of music information retrieval. However, users frequently withhold their real-world identity, which creates a negative impact on music recommendation. Thus, the proposed method recognizes users' real-world identities based on music metadata. The approach is based on using the tracks most frequently listened to by a user to predict their gender and age. Experimental results showed that the approach achieved an accuracy of 78.87% for gender identification and a mean absolute error of 3.69 years for the age estimation of 48403 users, demonstrating its effectiveness and feasibility, and paving the way for improving music recommendation based on such personal information.

1. INTRODUCTION

Amid the rapid growth of digital music and mobile devices, numerous online music services (e.g., Last.fm, 7digital, Grooveshark, and Spotify) provide music recommendations to assist users in selecting songs. Most music-recommendation systems are based on content- and collaborative-based approaches [15]. For content-based approaches [2, 8, 9], recommendations are made according to the audio similarity of songs. By contrast, collaborative-based approaches involve recommending music for a target user according to matched listening patterns that are analyzed from massive users [1, 13].

Because music preferences of users relate to their real-world identities [12], several collaborative-based approaches consider identification factors such as age and gender for music recommendation [14]. However, online music services may experience difficulty obtaining such information. Conversely, music metadata (listening history) is generally available. This motivated us to recognize users' real-world identities based on music metadata. Figure 1 illustrates the proposed system. In this preliminary study, we focused on predicting gender and age according to the most listened songs. In particular, gender identification was treated as a binary-classification problem, whereas age estimation was considered a regression problem. Two features were applied for both gender identification and age estimation tasks. The first feature, TF*IDF, is a widely used feature representation in natural language processing [16]. Because the music metadata of each user can be considered directly as a document, gender identification can be viewed as a document categorization problem. In addition, TF*IDF is generally applied with latent semantic indexing (LSI) to reduce feature dimension. Consequently, this serves as the baseline feature in this study.



	Top-1 track	Top-2 track	Top-3 track	…
Artist name	Paul Anka	The Platters	Johnny Cash	…
Song title	You Are My Destiny	Only You	I Love You Because	…

Figure 1. Illustration of the proposed system using a real example.

The second feature, the Gaussian super vector (GSV) [3], is a robust feature representation for speaker verification. In general, the GSV is used to model acoustic features such as MFCCs. In this study, music metadata was translated into proposed hotness features (a bag-of-features representation) and could be modeled using the GSV. The concept of the GSV can be described as follows. First, a universal background model (UBM) is trained using a Gaussian mixture model (GMM) to represent the global music preference of users. A user-specific GMM can then be obtained using the maximum a posteriori (MAP) adaptation from the UBM. Finally, the mean vectors of the user-specific GMM are applied as GSV features.

The remainder of this paper is organized as follows: Section 2 describes the related literature, and Section 3 introduces the TF*IDF; the GSV is explained in Section 4, and the experimental results are presented in Section 5; finally, Section 6 provides the conclusion of this study.

2. RELATED LITERATURE

Machine learning has been widely applied to music information retrieval (MIR), a vital task of which is content-based music classification [5, 11]. For example, the annual Music Information Retrieval Evaluation eXchange (MIREX) competition has been held since 2004, at which some of the most popular competition tasks have included music genre classification, music mood classification, artist identification, and tag annotation. The purpose of content-based music classification is to recognize semantic music attributes from audio signals. Generally, songs are represented by features with different aspects such as timbre and rhythm. Classifiers are used to identify the relationship between low-level features and mid-level music metadata.

However, little work has been done on predicting personal traits based on music metadata [7]. Figure 2 shows a comparison of our approach and content-based music classification. At the top level, user identity provides a basic description of users. At the middle level, music metadata provides a description of music. A semantic gap exists between music metadata and user identity. Beyond content-based music classification, our approach serves as a bridge between them. This enables online music services to recognize unknown users more effectively and, consequently, improve their music recommendations.

3. TF*IDF FEATURE REPRESENTATION

The music metadata of each user can be considered a document. The TF*IDF describes the relative importance of an artist for a specific document. LSI is then applied for dimensionality reduction.

3.1 TF*IDF

Let the document (music metadata) of each user in the training set be denoted as

$$d_i = \{t_1, t_2, \cdots, t_n\},\ d_i \in D \tag{1}$$

where t_n is the artist name of the top-n listened to song of user i. D is the collection of all documents in the training set. The TF*IDF representation is composed of the term frequency (TF) and inverse document frequency (IDF). TF indicates the importance of an artist for a particular document, whereas IDF indicates the discriminative power of an artist among documents. The TF*IDF can be expressed as

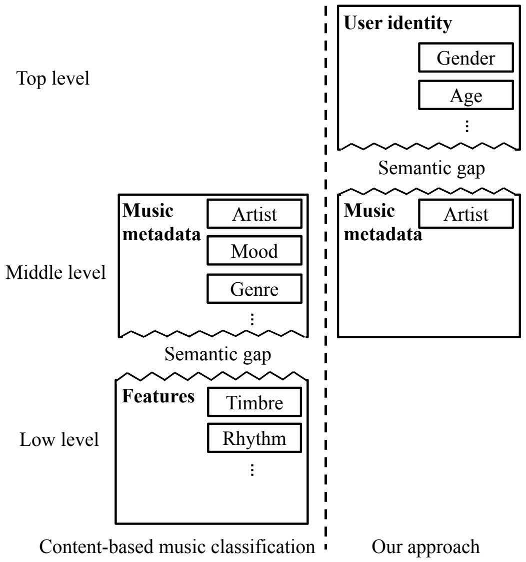


Figure 2. Comparison of our approach and content-based music classification.

$$tfidf_{i,n} = tf_{i,n} \times \log\left(\frac{|D|}{df_n}\right) \tag{2}$$

where $tf_{i,n}$ is the frequency of t_n in d_i, and df_n represents the number of documents in which t_n appears.

$$df_n = |\{d : d \in D \text{ and } t_n \in d\}| \tag{3}$$

3.2 Latent Semantic Indexing

The TF*IDF representation scheme leads to high feature dimensionality because the feature dimension is equal to the number of artists. Therefore, LSI is generally applied to transform data into a lower-dimensional semantic space. Let W be the TF*IDF reorientation of D, where each column represents document d_i. The LSI performs singular value decomposition (SVD) as follows:

$$W \approx U\Sigma V^T \tag{4}$$

where U and V represent terms and documents in the semantic space, respectively. Σ is a diagonal matrix with corresponding singular values. $\Sigma^{-1}U^T$ can be used to transform new documents into the lower-dimensional semantic space.

4. GSV FEATURE REPRESENTATION

This section introduces the proposed hotness features and explains how to generate the GSV features based on hotness features.

4.1 Hotness Feature Extraction

We assumed each artist t_n may exude various degrees of hotness to different genders and ages. For example, the count (the number of times) of Justin Bieber that occurs in users' top listened to songs of the training set was 845, where 649 was from the female class and 196 was from the male class. We could define the hotness of Justin Bieber for females as 76.80% (649/845) and that for males as 23.20% (196/845). Consequently, a user tends to be a female if her top listened to songs related mostly to Justin Bieber. Consequently, the age and gender characteristics of a user can be obtained by computing the hotness features of relevant artists.

Let D be divided into classes C according to users' genders or ages:

$$\begin{cases} C_1 \cup C_2 \cup \cdots \cup C_p = D \\ C_1 \cap C_2 \cap \cdots \cap C_p = \emptyset \end{cases} \tag{5}$$

where p is the number of classes. Here, p is 2 for gender identification and 51 (the range of age) for age estimation. The hotness feature of each artist t_n is defined as h_n:

$$h_n = \begin{bmatrix} \frac{c_{n,1}}{\alpha} \\ \frac{c_{n,2}}{\alpha} \\ \vdots \\ \frac{c_{n,p}}{\alpha} \end{bmatrix} \tag{6}$$

where $c_{n,p}$ is the count of artist t_n in C_p, and α is the count of artist t_n in all classes.

$$\alpha = \sum_{l=1}^{p} c_{n,l} \tag{7}$$

Next, each document in (1) can be transformed to a $p \times n$ matrix x, which describes the gender and age characteristics of a user:

$$x = [h_1, h_2, \cdots, h_n] \tag{8}$$

Because the form of x can be considered a bag-of-features, the GSV can be applied directly.

4.2 GSV Feature Extraction

Figure 3 is a flowchart of the GSV feature extraction, which can be divided into offline and online stages. At the offline stage, the goal is to construct a UBM [10] to represent the global hotness features, which are then used as prior knowledge for each user at the online stage. First, hotness features are extracted for all music metadata in the training set. The UBM is then constructed through a GMM estimated using the EM (expectation-maximization) algorithm. Specifically, the UBM evaluates the likelihood of a given feature vector x as follows:

$$f(x|\theta) = \sum_{k=1}^{K} w_k \mathrm{N}(x|m_k, r_k) \tag{9}$$

where $\theta = (w_1, ..., w_K, m_1, ..., m_K, r_1, ..., r_K)$ is a set of parameters, with w_k denoting the mixture gain for the kth mixture component, subject to the constraint $\sum_{k=1}^{K} w_k = 1$, and $\mathrm{N}(x|m_k, r_k)$ denoting the Gaussian density function with a mean vector m_k and a covariance matrix r_k. This bag-of-features model is based on the assumption that similar users have similar global artist characteristics.

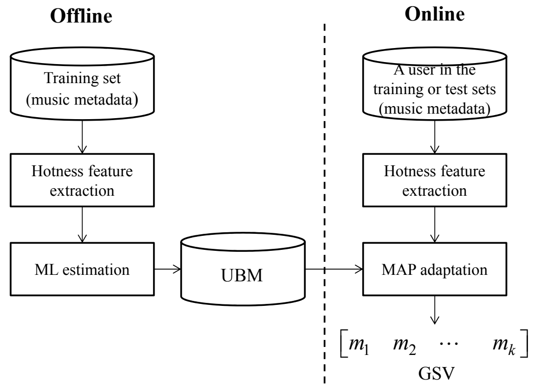

Offline
Online
Training set (music metadata)
A user in the training or test sets (music metadata)
Hotness feature extraction
Hotness feature extraction
ML estimation
UBM
MAP adaptation
$[m_1 \;\; m_2 \;\; \cdots \;\; m_k]$
GSV

Figure 3. Flowchart of the GSV feature extraction.

At the online stage, the MAP adaptation [6] is used to produce an adapted GMM for a specific user. Specifically, MAP attempts to determine the parameter θ in the parameter space Θ that maximizes the posterior probability given the training data x and hyperparameter ω, as follows:

$$\theta_{MAP} = \arg\max_{\theta} f(x|\theta)\, g(\theta|\omega) \tag{10}$$

where $f(x|\theta)$ is the probability density function (PDF) for the observed data x given the parameter θ, and $g(\theta|\omega)$ is the prior PDF given the hyperparameter ω.

Finally, for each user, the mean vectors of the adapted GMM are stacked to form a new feature vector called GSV. Because the adapted GMM is obtained using MAP adaptation over the UBM, it is generally more robust than directly modeling the feature vectors by using GMM without any prior knowledge.

5. EXPERIMENTAL RESULTS

This section describes data collection, experimental settings, and experimental results.

5.1 Data Collection

The Last.fm API was applied for data set collection, because it allows anyone to access data including albums, tracks, users, events, and tags. First, we collected user IDs through the *User.getFriends* function. Second, the *User.getInfo* function was applied to each user for obtaining their age and gender information. Finally, the *User.getTopTracks* function was applied to acquire at most top-50 tracks listened to by a user. The track information included song titles and artist names, but only artist names were used for feature extraction in this preliminary study.

The final collected data set included 96807 users, in which each user had at least 40 top tracks as well as complete gender and age information. According to the users' country codes, they were from 211 countries (or

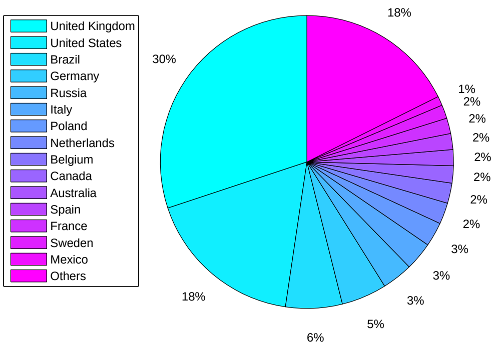


Figure 4. Ratio of countries of the collected data set.

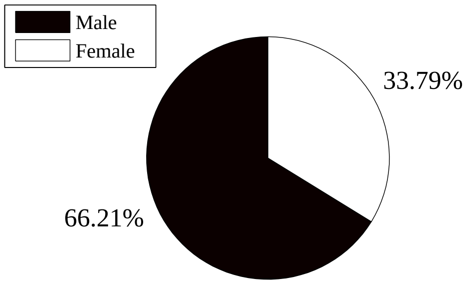


Figure 5. Gender ratio of the collected data set.

regions such as Hong Kong). The ratio of countries is shown in Figure 4. The majority were Western countries. The gender ratio is shown in Figure 5, in which approximately one-third of users (33.79%) were female and two-thirds (66.21%) were male. The age distribution of users is shown in Figure 6. The distribution was a skewed normal distribution and most users were young people.

Figure 7 shows the count of each artist that occurred in the users' top listened songs. Among 133938 unique artists in the data set, the ranking of popularity presents a pow-law distribution. This demonstrates that a few artists dominate the top listened songs. Although the majority of artists are not popular for all users, this does not indicate that they are unimportant, because their hotness could be discriminative over ages and gender.

5.2 Experimental Settings

The data set was equally divided into two subsets, the training (48404) and test (48403) sets. An open source tool of *Python*, *Gensim*, was applied for the TF*IDF and LSI implementation. followed the default setting of *Gensim* that maintained 200 latent dimensions for the TF*IDF. A support vector machine (SVM) tool, LIBSVM [4], was applied as the classifier. The SVM extension, support vector regression (SVR) was applied as the regressor, which has been observed in many cases to be superior to existing regression approaches. The RBF kernel with $\gamma = 8$ was applied to the SVM and SVR. For the UBM parameters, two Gaussian mixture components were experimentally applied (similar results can be obtained when using a different number of mixture components). Consequently, the numbers of dimensions of GSV features for gender identification and age estimation were 4 (2×2) and 102 (2×51), respectively.

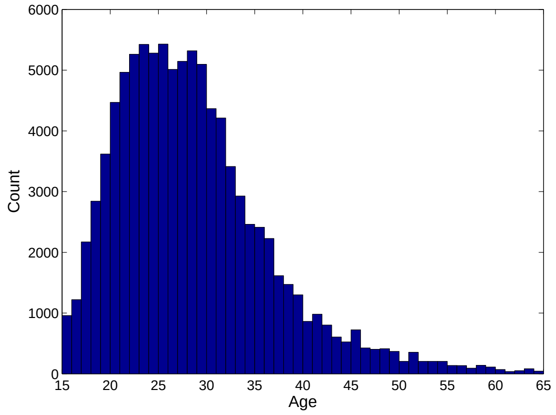


Figure 6. Age distribution of the collected data set.

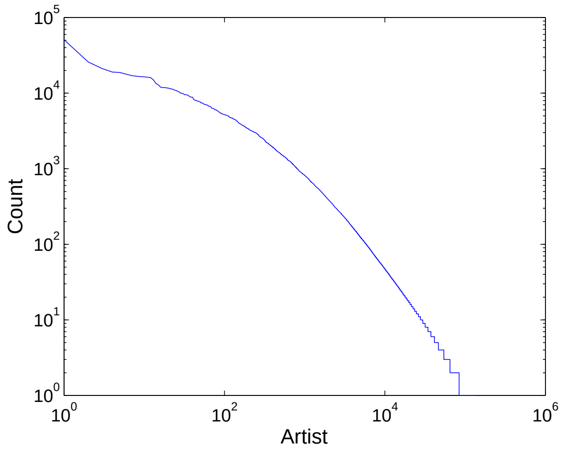


Figure 7. Count of artists of users' top listened songs. Ranking of popularity presents a pow-law distribution.

5.3 Gender Identification

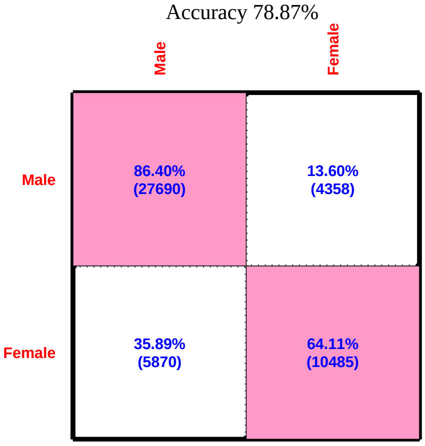
The accuracy was 78.87% and 78.21% for GSV and TF*IDF + LSI features, respectively. This indicates that both features are adequate for such a task. Despite the low dimensionality of GSV (4), it was superior to the high dimensionality of TF*IDF + LSI (200). This indicates the effectiveness of GSV use and the proposed hotness features. Figures 8 and 9 respectively show the confusion matrix of using GSV and TF*IDF + LSI features. Both features yielded higher accuracies for the male class than for the female class. A possible explanation is that a portion of the females' were similar to the males'. The classifier tended to favor the majority class (male), resulting in many female instances with incorrect predictions. The age difference can also be regarded for further analysis. Figure 10 shows the gender identification results of two features over various ages. Both features tended to have lower accuracies between the ages of 25 and 40 years, implying that a user whose age is between 25 and 40 years seems to have more blurred gender boundaries than do users below 25 years and above 40 years.

5.4 Age Estimation

Table 1 shows the performance comparison for age estimation. The mean absolute error (MAE) was applied as the performance index. The range of the predicted ages of the SVR is between 15 and 65 years. The experimental results show that the MAE is 3.69 and 4.25 years for GSV and TF*IDF + LSI, respectively. The GSV describes the age characteristics of a user and utilizes prior knowledge from the UBM; therefore, the GSV features are superior to those of the TF*IDF + LSI.

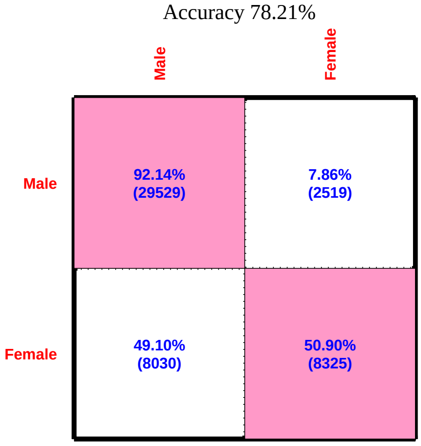
Method	MAE	MAE (male)	MAE (female)
GSV	3.69	4.31	2.48
TF*IDF+LSI	4.25	4.86	3.05

Table 1. Performance comparison for age estimation.


Accuracy 78.87%

	Male	Female
Male	86.40% (27690)	13.60% (4358)
Female	35.89% (5870)	64.11% (10485)

Figure 8. Confusion matrix of gender identification by using GSV features.


Accuracy 78.21%

	Male	Female
Male	92.14% (29529)	7.86% (2519)
Female	49.10% (8030)	50.90% (8325)

Figure 9. Confusion matrix of gender identification by using TF*IDF + LSI features.

For further analysis, gender difference was also considered. Notably, the MAE of females is less than that of males for both GSV and TF*IDF + LSI features. In particular, the MAE differences between males and females are approximately 1.8 for both features, implying that females have more distinct age divisions than males do.

6. CONCLUSION AND FUTURE WORK

This study confirmed the possibility of predicting users' age and gender based on music metadata. Three of the findings are summarized as follows.

- GSV features are superior to those of TF*IDF + LSI for both gender identification and age estimation tasks.
- Males tend to exhibit higher accuracy than females do in gender identification, whereas females are more predictable than males in age estimation.
- The experimental results indicate that gender identification is influenced by age, and vice versa. This suggests that an implicit relationship may exist between them.

Future work could include utilizing the proposed approach to improve music recommendation systems. We will also explore the possibility of recognizing deeper social aspects of user identities, such as occupation and education level.

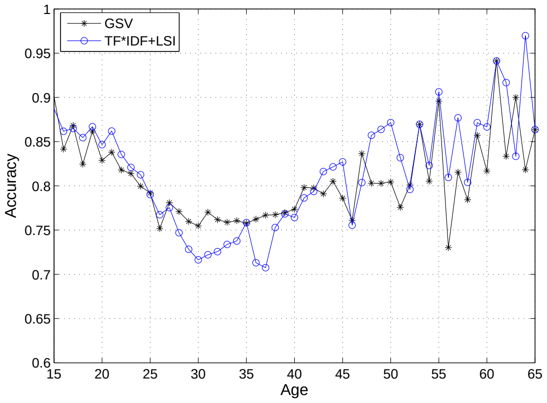


Figure 10. Gender identification results for various ages.

7. ACKNOWLEDGEMENT

This study is conducted under the ”NSC 102-3114-Y-307-026 A Research on Social Influence and Decision Support Analytics” of the Institute for Information Industry which is subsidized by the National Science Council.

8. REFERENCES

[1] L. Barrington, R. Oda, and G. Lanckriet. Smarter than genius? human evaluation of music recommender systems. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 357–362, 2009.

[2] D. Bogdanov, M. Haro, F. Fuhrmann, E. Gomez, and P. Herrera. Content-based music recommendation based on user preference examples. In *Proceedings of the ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery*, 2010.

[3] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, May 2006.

[4] C. C. Chang and C. J. Lin. Libsvm: A library for support vector machine, 2010.

[5] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Trans. Multimedia.*, 13(2):303–319, Apr. 2011.

[6] J. L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Audio, Speech, Lang. Process.*, 2(2):291–298, Apr. 1994.

[7] Jen-Yu Liu and Yi-Hsuan Yang. Inferring personal traits from music listening history. In *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM ’12, pages 31–36, New York, NY, USA, 2012. ACM.

[8] B. McFee, L. Barrington, and G. Lanckriet. Learning content similarity for music recommendation. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(8):2207–2218, Oct. 2012.

[9] A. V. D. Orrd, S. Dieleman, and B. Benjamin. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, 2013.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Process*, 10(13):19–41, Jan. 2000.

[11] B. L. Sturm. A survey of evaluation in music genre recognition. In *Proceedings of the Adaptive Multimedia Retrieval*, 2012.

[12] A. Uitdenbogerd and R. V. Schnydel. A review of factors affecting music recommender success. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 204–208, 2002.

[13] B. Xu, J. Bu, C. Chen, and D. Cai. An exploration of improving collaborative recommender systems via user-item subgroups. In *Proceedings of the 21st international conference on World Wide Web*, pages 21–30, 2012.

[14] Billy Yapriady and AlexandraL. Uitdenbogerd. Combining demographic data with collaborative filtering for automatic music recommendation. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3684 of *Lecture Notes in Computer Science*, pages 201–207. 2005.

[15] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 296–301, 2006.

[16] W. Zhang, T. Yoshida, and X. Tang. A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.