Lyrics Based Song Genre Classification

Aarti Girase, Apurva Advirkar, Chandrika Patil, Dhanshri Khadpe, Amruta Pokhare
Department of Information Technology, Mumbai University
Atharva college of engineering, Malad(w),India
aartigirase2012@hotmail.com,advirkar.apurva@gmail.com,chandrika.patil@ymail.com,
dh.khadpe@gmail.com,amrutapokhare09@gmail.com

Abstract— **Nowadays the demand of text classification is increasing tremendously. The problem of text classification finds applications in a wide variety of domains in text mining such as news filtering and organization, music classification based on lyrics, email classification and spam filtering and opinion mining. This paper focuses on lyrics based song classification into genres. Songs can be divided into many categories based on styles, rhythm, and even cultural background. The styles are what we call the "genres". The classification can be achieved using a hybrid model that consists of Association rule, Naive Bayes rule and part of Genetic algorithm.**

Keywords— **Text classification, Association rule, Apriori algorithm, confidence, support, frequent item sets, Naive Bayes classifier, Genetic Algorithm**

I. INTRODUCTION

A. Huge Size of Digital Music Database

Since the rapidly increasing size of database, digital music becomes much more popular. An automatic and faster system is thus demanding for enhancing the searching efficiency and quality, music genre classification, music emotion classification, beat tracking, preference recommendation and etc.

B. Methods for Song Categorization

Traditionally, the genre of each song was tagged by experts, called expert system, which was subjective and time-consuming. To address this issue, low-level features were exploited for classification, which are extracted directly from music signal which required actual recording of songs.On the other hand, different data-mining algorithms, including supervised and unsupervised classification, are proposed to classifying music data based on genre, emotion, beat tracking, preference recommendation and etc.

C. Relevance of High Level Features

Since the flourishing development of music signal processing, such as music emotional classification and singer recognition, many research areas in music can no more depend only on low-level features. To address these issues, high-level features are exploited, and music genre is one of these high-level features. For example, we can only get the signal information from low-level features, but not "how the song sounds like for human" or "does the song say a happy or a sad story".[3]

II. RELATED WORK

Song sentiment classification has been investigated since 1990s in audio signal processing community and research works are mostly found relying on audio signal to make a decision using machine learning algorithms (Li and Ogihara, 2006; Lu et al., 2006). Luet al. (2006) proposes the hierarchical framework to perform song sentiment classification with two steps. In the first step the energy level is detected with intensity features and the stress level is determined in the second step with timbre and rhythm features. It is proved difficult to detect stress level using audio as classification proof. Song sentiment classification using lyric as proof is recently investigated by Chen et al. (2006).

They adopt the hierarchical framework and make use of song lyric to detect stress level in the second step. In fact, many literatures have been produced to address the sentiment analysis problem in natural language processing research. Three approaches are dominating, i.e. knowledge-based approach (Kim and Hovy, 2004), information retrieval-based approach (Turney and Littman, 2003) and machine learning approach (Pang et al., 2002), in which the last approach is found very popular. [9]

III. DETAILED VIEW

A. Preparing Text for Classification

Song lyrics are considered as training document for developing a model for classifying new song of unknown class.

Each lyrics file is considered as a transaction in the dataset. So number of lyrics file is equal to the number of transactions in the transaction set. The next step is to remove unnecessary words leaving only the keywords. As we know, highly frequent words, such as determiners and prepositions, are not considered to be content words because they appear in virtually every song. Unlike considering all words in a text we have considered only those words that are related to the subject of the text. These keyword extraction processes drops the common unnecessary words like am, is, are, to, from...etc. and also drops all kinds of punctuations and stop words. Singular and plural form of a word is considered same and keeping the singular form in the text. Finally, the remaining frequent words are considered as a single transaction data in the set of database transaction. This process is applied to all lyrics files before applying association mining to the transaction database.

Consider the following transaction:

My dream is to fly
Over the rainbow, so high!
My dream is to fly
Over the rainbow, so high!
Eh... eh...

Rise up,
Don't falling down again
Rise up,
Love like I broke the chains

I tried to fly a while so high
Direction: sky!
I tried to fly a while so high
Direction: sky!

My dream is to fly
Over the rainbow, so high!
My dream is to fly
Over the rainbow, so high!
My dream is to fly
Over the rainbow, so high!
My dream is to fly
Over the rainbow, so high!
Eh...

Rise up, rise up, rise up, rise up
We'll be the game
Rise up, rise up, rise up, rise up
For my mind and my brain

'Cause I tried to fly a while so high
Direction: sky!
I tried to fly so high
Direction: sky!

Keywords extracted are: rise, dream, up, fly, high, sky.
This keyword extraction process is applied to all the lyrics.

B. *Deriving Associated Word Sets*

Considering each keyword set i.e. each lyrics file as a transaction, we generated a list of maximum length sets applying the Apriori algorithm. The support and confidence is set to 0.05 and 0.75 respectively. A partial list of the generated large word set with their occurrence frequency is illustrated in the following table.

TABLE I
Word Set with Occurrence Frequency

Maximum Length Set	No of Occurrences		
	Inspirational	Fun	Romantic
Rise, high, fly, up, sky	5		
Love, heart			6
Spend ,life,			2
Faith, alive	2		
Hope, heart	3		
Survived, strength	2		
Crazy, life, party		2	

C. *Setting Associated Word Set with Probability Value*

From the generated word set after applying association mining on training data we have found the following information:total number of word set is 22, total number of word set from Inspirational is 12, Fun is 2 and Romantic is 8 respectively.

Now we can use the Naïve Bayes classifier for probability calculation. The calculation of first term is based on the fraction of each target class in the training data. Prior probability for the categories are 0.545,0.909 and 0.363 respectively. The second term is calculated according to the equation,

$$V_{NB} = \text{argmax } P(Vj)\Pi P(aj \mid Vj) \qquad (1)$$

VNB is the classification that maximizes the probability of observing the words that were actually found in the example documents, subject to the usual Naive Bayes independence assumption. The first term can be estimated based on the fraction of each class in the training data. For estimating the second term the following equation is used:

$$\frac{n_k + 1}{n + |vocabulary|} \qquad (2)$$

Where n is the total number of word positions in all training examples whose target value is Vj, nk is the number of items that word is found among these n word positions, and |vocabulary| is the total number of distinct words found within training data.[10]

TABLE II
Word Set with Probability Value

Maximum Length Set	Probability		
	Inspirational	Fun	Romantic
Rise, high, fly, up, sky	0.024	0.043	0.567
Love, heart	0.217	0.234	0.356
Spend ,Life	0.005	0.443	0.155
Faith, alive	0.567	0.176	0.024
Hope, heart	0.078	0.333	0.237
Survived, strength	0.760	0.167	0.237
Crazy, life, party	0.176	0.256	0.098

D. Last Classifier Circuit/Testing Algorithm

n = number of class,
m = number of associated sets
1. For each class i = 1 to n do
2. Set pval = 0, nval = 0, p = 0, n = 0
3. For each set s = 1 to m do
4. If the probability of the class (i) for the set (s) is maximum then increment pval else increment nval
5. If 50% of the associated set s is matched with the keywords set do step 6 else do step7
6. If maximum probability matches the class i then increment p
7. If maximum probability does not match the class i increment n
8. If (s<=m) go to step 3
9. Calculate the percentage of matching in positive sets for the class i
10. Calculate the percentage of not matching in negative sets for the class i
11. Calculate the total probability as the summation of the results obtained from step 9 and 10 and also the prior probability of the class i in set s
12. If (i<=n) go to step 1
13. Set the class having the maximum probability value as the result

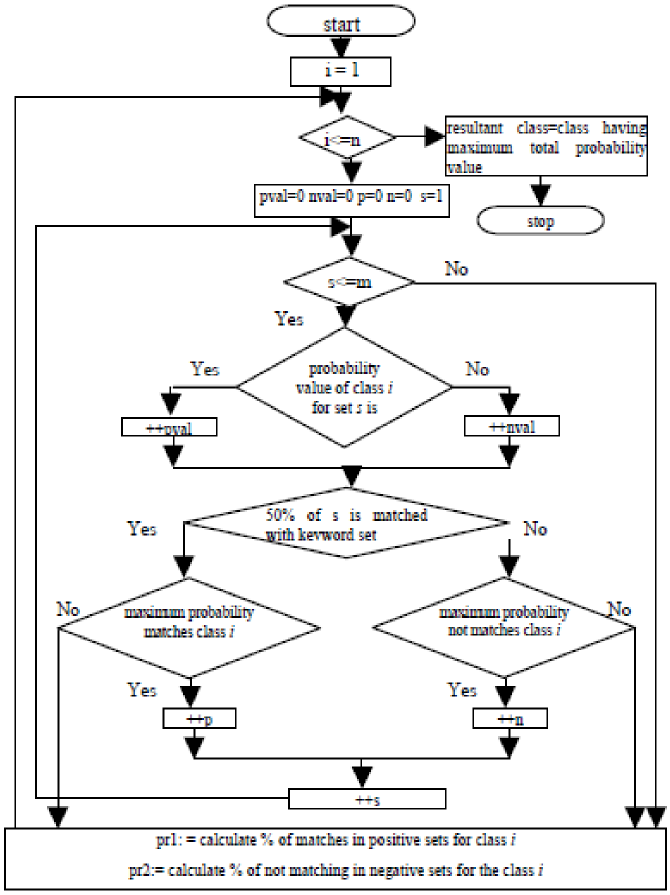

Fig 1. Flow chart of the classifier circuit

IV. COMPARATIVE STUDY

Association Rule and Naïve Bayes Classifier:

The following result shows comparison between Association Rule with Naive Bayes Classifier and Hybrid model. The result shows that proposed approach works well using only 50% training data. [1]

TABLE III
Comparison of Classifier Circuit With Text Classifier Using Association Rule and Naïve Bayes Classifier.

% of Training Data	% of Accuracy	
	Association Rule with Naïve Byes Classifier	Proposed Method
10	36	13
20	40	76
30	63	85
40	63	71
50	31	78

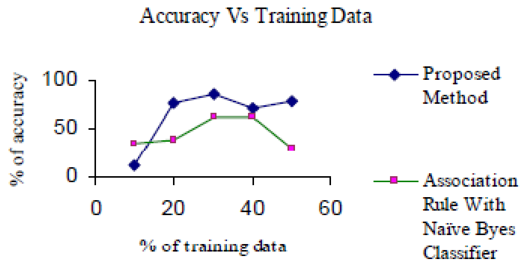

Fig 2. Accuracy Vs Training Curve

V. CONCLUSION

Classifying music into genres based on lyrics is an interesting problem in the field of Music Information Retrieval that presents several challenges. Despite these difficulties, however, there is still much potential for song lyric analysis. The use of hybrid model is an efficient technique for text classification. The existing techniques require more data for training as well as the computational time of these techniques is also large. In contrast to the existing one, the testing algorithm requires less training data and less computational time. Also the use of text classification techniques in categorizing songs based on lyrics into genres is not just useful for music retrieval and categorization but also for music emotional classification, singer recognition and many research areas in music.

REFERENCES

[1] S.M. Kamruzzaman, Farhana Haider And Ahmed Ryadh Hasan , "Text Classification Using Association Rule With A Hybrid Concept Of Naive Bayes Classifier And Genetic Algorithm"

[2] "A SURVEY OF TEXT CLASSIFICATION ALGORITHMS" Charu C. Aggarwal IBM T. J. Watson Research Centeryorktown Heights, NY, Chengxiang Zhaiuniversity Of Illinois At Urbana-Champaign Urbana, IL

[3] Mei-Lan Chu, R97945048," Automatic Music Genre Classification" Graduate Institute Of Biomedical Electronics And Bioinformatics,National Taiwan University, Taipei, Taiwan Abstract.

[4] Xiao Hu, J. Stephen Downie, Andreas F. Ehmann, "LYRIC TEXT MINING IN MUSIC MOOD CLASSIFICATION", 10th International Society For Music Information Retrieval Conference (ISMIR 2009)

[5] Adam Sadovsky, Xing Chen ,"Song Genre And Artist Classification Via Supervised Learning From Lyrics", CS 224N Final Project

[6]Music.DataMining, Http://Michael.Hahsler.Net/SMU/8331/Tutorials/Music_Mining_Slides.Pdf,

[7] Rudolf Mayer, Robert Neumayer, And Andreas Rauber, "RHYME AND STYLE FEATURES FOR MUSICAL GENRE CLASSIFICATION BY SONG LYRICS", Vienna University Of Technology, Department Of Software Technology And Interactive Systems Favoritenstraße 9-11Vienna, Austria

[8] Cory Mckay, John Ashley Burgoyne, Jason Hockman, Jordan B. L. Smith,Gabriel Vigliensoni And Ichiro Fujinaga "EVALUATING THE GENRE CLASSIFICATION PERFORMANCE OF LYRICAL FEATURES RELATIVE TO AUDIO, SYMBOLIC AND CULTURAL FEATURES" Centre For Interdisciplinary Research In Music Media And Technology (CIRMMT) Mcgill University, Montréal, Québec, Canada

[9] Sentiment Vector Space Model For Lyric-Based Song Sentiment Classification Yunqing Xia Linlin Wang Center For Speech And Language Tech State Key Lab Of Intelligent Tech. And Sys.Kam-Fai Wong Mingxing Xudept. Of SE&EM Dept. Of CST The Chinese University Of Hong Kong Tsinghua University

[10] S. M. Kamruzzaman , "Text Classification Using Artificial Intelligence", Department Of Information And Communication Engineering University Of Rajshahi, Rajshahi-6205, Bangladesh

