VERIFYING TAG ANNOTATIONS THROUGH ASSOCIATION ANALYSIS

Tom Arjannikov
University of Lethbridge
tom.arjannikov@uleth.ca

Chris Sanden
University of Lethbridge
sanden@cs.uleth.ca

John Z. Zhang
University of Lethbridge
zhang@cs.uleth.ca

ABSTRACT

Music tags provide descriptive and rich information about a music piece, including its genre, artist, emotion, instrument, etc. While many work on automating it, at present, tag annotation is largely a manual process. It often involves judgements and opinions from people of different background and level of musical expertise. Therefore, the resulting tags are usually subjective, ambiguous, and error-prone. To deal with this situation, we seek automatic methods to verify and monitor this process. Furthermore, because multiple tags can annotate each music piece, our task lends itself to multi-label methods which capture the inherent associations among annotations in a given music repository. In this paper, we propose a novel approach to verify the quality of music tag annotations via association analysis. We demonstrate the effectiveness of our approach through a series of simulations using four publicly available music datasets. To our knowledge, our work is among the initial efforts in verifying music tag annotations.

1. INTRODUCTION

Due to the advances in technology, such as data storage and compression, media processing, information retrieval, and the Internet, digital music collections have been growing enormously in volume. Millions of songs previously in physical formats are now readily available through on-line accesses, presenting many new challenges in related fields.

Among them is *Music Information Retrieval* (*MIR*), an interdisciplinary area that attracts practitioners from computer science, cognitive science, information retrieval, musicology, psychology, etc. One of its current tasks is the design and implementation of algorithmic approaches for managing large collections of digital music; these include but are not limited to music classification, automatic tag annotation, recommendation and playlist generation. [6]

Currently, it is a common strategy to organize and access digital music collections via *textual meta-data*, usually *tags*, such as *genre*, *style*, *mood*, *artist*, etc. The process of annotating a music piece with appropriate meta-data is referred to as *music tag annotation*. It relies heavily on music experts as well as amateurs [19]. Due to the ambiguity and subjectivity introduced in the annotation process, music tags can be inconsistent, incomplete and sometimes even error-prone, making it difficult to maintain them in large music collections [12]. Moreover, the manual annotation process can be complex, involving substantial financial and labor costs [5]. These are among the many issues, which automatic approaches aim to tackle.

1.1 Previous work

Automatic music tag annotation is an important problem in MIR with numerous applications. For instance, an accurate prediction of tags is often the first step toward making music recommendations. As of recently, music tag annotation has received considerable attention and many related techniques have been proposed.

Turnbull *et al.* [18] employ a generative probabilistic model and propose one of the first automatic tag annotation systems. In addition, they create a music dataset called *The Computer Audition Lab 500* (*CAL500*), which has since become a *de facto* benchmark for evaluating the performance of tag annotation systems.

Hoffman *et al.* [8] present another probabilistic model, called the *Codeword Bernoulli Average*, which attempts to predict the probability that a tag can be associated with a music piece. In addition, Bertin-Mahieux *et al.* [3] propose *Autotagger*, a model that uses ensemble learning schemes to associate tags with music pieces. Ness *et al.* [10] describe how *stacked generalization* of the probabilistic outputs of a *Support Vector Machine* (*SVM*) can be used to improve the performance of automatic tag annotation.

Shen *et al.* [16] propose a framework called *MMTagger* that combines advanced feature extraction techniques and high-level semantic concept modeling for music tag annotation. The proposed framework uses a multilayer architecture that gathers multiple *Gaussian mixture models* and SVMs.

Sanden and Zhang [13] treat music tag annotation as a multi-label classification problem and discuss various issues related to tag annotation through extensive experiments using a set of a multi-label classifiers and a set of evaluation measures.

Neubarth *et al.* [11], use association analysis to find different, musicologically motivated, associations between various folk music genres and their geographical distribution.

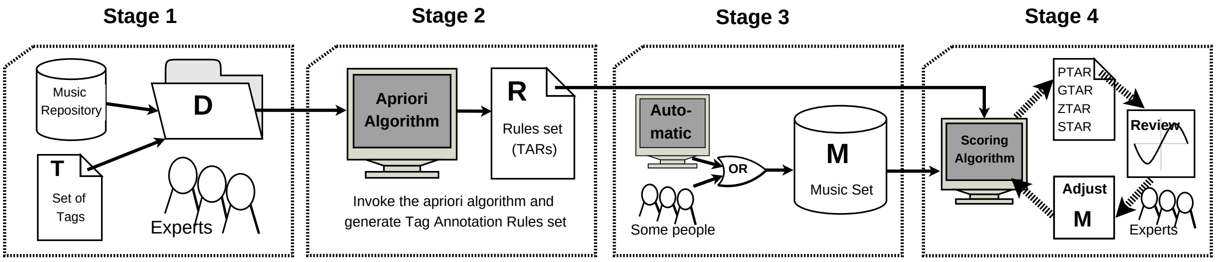

Figure 1: The four stages of our proposed approach to verification of tag annotations through association analysis.

1.2 Tag Annotation

We follow Ness *et al.* [10] in their formulation of the music tag annotation process as follows. Given a set of tags $T = \{t_1, t_2, ..., t_A\}$ and a set of music pieces $M = \{m_1,$ m_2, ..., $m_R\}$, then each music piece $m_j \in M$ is an annotation vector $A = (a_1, a_2, \cdots, a_A)$, where $a_i > 0$ if tag t_i has been associated with the piece, and $a_i = 0$, otherwise. These a_i's, referred to as *semantic weights*, describe the strength of the semantic correspondence between a tag and its music piece. When mapped to a binary assignment of $\{0,1\}$, the semantic weights can be interpreted as class labels, i.e., whether a tag is assigned to the music piece or not. Naturally, each music piece can have multiple tags [10, 13].

Tags for a given music piece reveal the inherent musical nature that it attempts to convey and express. As a coherent expression, these tags represent features that distinguish this music piece from others. This expression intuitively shows a strong association of these tags to the music piece or to a set of similar music pieces in terms of their musical nature. We work toward this intuition and aim to capture associations between tags and utilize them to verify the annotation process.

1.3 Association Analysis

Association analysis attempts to discover the inherent relationships among data objects in an application domain. Such relationships are represented as association rules. An example of such application domain is the basket analysis in supermarkets, where one tries to discover the relationships among commodities in baskets. For example, the association rule $\{milk\}$,$\{eggs\} \rightarrow \{bread\}$ implies that, if *milk* and *eggs* are bought together by a customer, then *bread* is likely to be bought as well, i.e., they have some inherent statistical relationships.

Let $I = \{i_1, i_2, \cdots, i_n\}$ be a set of n items and $T = \{t_1, t_2, \cdots, t_m\}$ be a transactional database, where each transaction $t_i \in T$ is a nonempty subset of I. An association rule is of the form $A \rightarrow B$, where A and B are called *itemsets* and $A \subseteq I$ and $B \subseteq I$. It is required that $A \bigcap B = \phi$. A is called the *antecedent* of the rule while B is its *consequent*. We say that the rule $A \rightarrow B$ holds for T with *support* s, where s is the percentage of transactions in T that contain both A and B, and with *confidence* c, where c is the percentage of transactions containing A that also contain B, i.e, $support(A \bigcup B)/support(A)$.

It should be noted that our work presented in this paper focuses on the application of association analysis to tag anoation in music. We employ one of the many algorithms that find the association rules in a transactional database, the *Apriori* algorithm [1, 2], which generates association rules that satisfy user-specified *minimum support* and *minimum confidence*. For a simple but illustrative example of association rule generation, see [7].

2. VERIFYING TAG ANNOTATIONS

In our experience, we observe that tag annotation by music experts, though at a high cost, is relatively more accurate, due to their music expertise, as compared to the one performed by amateurs. Based on this, we propose to take advantage of the judgments and opinions of music experts for a tag annotation task. It is our belief that tags of high quality reveal inherent associations among music pieces in a repository and can be utilized in a tag annotation process. Our proposed approach to verifying tag annotations is conducted in four stages, as shown in Figure 1.

First, given an annotation task, a group of experts with extensive music background and experience are solicited. They examine the repository and select a set D comprised of those songs that best represent the repository as a whole. Then, they manually annotate this set using a predetermined and fixed set of tags. Consequently, the tag annotations contain expert knowledge and show authoritative judgments and opinions of those experts regarding the representative music pieces.

In the second stage, we treat the selected music pieces as a transactional database, where tags are the transactional items and music pieces are the transactions. We then invoke the *Apriori* [1] algorithm to find the association rules in it under the current minimum support and minimum confidence, both of which are specified in advance. We denote the resulting association rule set $R = \{r_1, r_2, \cdots, r_k\}$, where each r_i is called a *tag annotation rule*.

During the third stage, we introduce new music pieces whose annotations are to be verified. They could be samples from the same music repository, which would be the case where the task is to verify the annotations of a given music repository. They could also be an addition to the existing repository, which would happen when the repository grows and the newly added music pieces need to contain annotations that are consistent with the existing repository. We collectively call these new music pieces $M = \{m_1,$

m_2,$\cdots$, m_n}; they are annotated manually or automatically as discussed in Section 1.1. We represent each m_i as a set of tags.

In the last stage, we verify and adjust the annotations of M. First, we run the Scoring Algorithm and generate a set of evaluation measures, both of which are discussed in the following two subsections. Then, we examine the evaluation and adjust M accordingly. Once M reaches a desired level of agreement with R, and hence the repository, the whole process of verifying tag annotations comes to completion.

2.1 Scoring Algorithm

The *annotation score* $S(m_i)$ of a music piece m_i represents the number of rules in R that m_i satisfies. It is calculated as follows. For each tag annotation rule $A \rightarrow B$, if the song m_i contains both, the antecedent ($A \in m_i$) and the consequent ($B \in m_i$), then we increment the song's score $S(m_i)$ by 1. However, there could be situations where $A \rightarrow B$ and $A \rightarrow C$ coexist in R, representing the multi-label nature of tag annotation. If a music piece misses the first rule but satisfies the second one, its score still increments by 1 instead of 0. To achieve this, we iterate through R twice. At first, we build a list of rules (hit_list), which are satisfied by the song m_i, and increment $S(m_i)$ accordingly. Then we look for the rules which are not satisfied by the song's annotations, such as $A \rightarrow B$, where $A \in m_i$ but $B \notin m_i$. If their antecedents are not found in the hit_list, then we decrement the song's score $S(m_i)$ by 1 for each problematic rule.

Scoring Algorithm: used together with R during Stage 4.

```
for each m_i ∈ M {
  S(m_i) = 0;
  for each rule r_i = (A → B) ∈ R
    if A ∈ m_i then
      if B ∈ m_i then {
        record the rule in hit_list
        S(m_i)++
      }
  for each rule r_i = (A → B) ∈ R
    if A ∈ m_i then
      if B ∉ m_i then
        if A ≠ any antecedent in the hit_list then
          S(m_i)--
}
```

2.2 Evaluation Measures

If a music piece has a positive score, we say that it has a *sound tag annotation* (*STA*). Otherwise, we say that it has a *problematic tag annotation* (*PTA*). Furthermore, we attempt to distinguish music pieces that have a certain degree of ambiguity and subjectivity. A music piece has a *gray tag annotation* (*GTA*) if its annotation score is between [l,h], where l and h are user-specified range values. For instance, they can be -1 and +1 respectively. In our simulations below we use $[-2, 0]$ for this range. A music expert, depending on her/his musical expectation and experience, may set a different range.

We calculate four measures. The first is the *Problematic Tag Annotation Rate* (*PTAR*) for the annotation process. It is the ratio between the number of music pieces with PTA and the total number of music pieces. The second is the *Sound Tag Annotation Rate* (*STAR*), which is the ratio between number of STA music pieces and the total number of music pieces. These two rates represent the quality of the tag annotation process. It is obvious that the higher PTAR is, the worse the tag annotation quality; while for STAR it is the opposite. In addition, we calculate the *Gray Tag Annotation Rate* (*GTAR*), which is the ratio between the number of GTA music pieces and the total number of music pieces in the dataset. It represents the uncertainty in the annotation process. The fourth measure that we calculate is the *Zero Tag Annotation Rate* (*ZTAR*), which represents the percentage of music pieces that do not contain any of the tag annotation rules. These measures divide the whole music set into partitions and add up to 1.

3. SIMULATIONS

We are facing three major challenges when examining the effectiveness of our proposed approach to tag annotation verification.

It would be ideal to examine our approach using the process as depicted by Figure 1. However, such a process involves a great amount of financial and labor costs. Given our current situation and circumstances, it would be extremely hard, if not impossible, for us to deploy such a process, although involving a group of experts with extensive music background and experience could prove invaluable, as they would inevitably provide important feedback about our approach.

To make the situation worse, it is often the case that different music repositories and datasets vary in the sets of tags used to annotate their music pieces. Furthermore, association analysis captures each dataset's own associations that do not necessarily translate to others. Therefore, it is possible for two datasets to be completely incompatible in such a way that one could not be used to verify another.

Moreover, there is a lack of good quality datasets that could be used for benchmarking, although the MIR community is making efforts to come up with such datasets, which is evidenced by the Million Song Dataset [4, 9] and the Million Song Dataset Benchmarks [14].

Taking these issues into consideration, we design and implement a series of simulations to demonstrate the effectiveness of our proposed approach. Through these simulations, we aim to achieve three goals: (G_1) demonstrate that our approach is stable, in that it will not behave arbitrarily when given different music datasets or, put in another way, given similar music datasets, it should behave similarly; (G_2) assess the four music datasets using our evaluation measures and confirm that they maintain different relationships among their tags due to the differences in their annotation processes; and (G_3) confirm that, when the quality of annotations in a music dataset improves, our proposed measures reflect this improvement.

We run our simulations with different minimum support and confidence value pairs; for each minimum support value ranging from 5 to 95 we use a different minimum confidence ranging from 5 to 95.

3.1 Music Datasets

For our simulations, we use four datasets described below and presented in Table 1. They have been previously annotated with tags by music experts and amateurs.

The *CAL500* [18] dataset, denoted as S_{CAL500}, is a collection of 502 Western songs recorded by different artists. Each song is manually annotated by at least three human annotators using a vocabulary of 174 tags, which is distributed across six categories: Mood, Genre, Instrument, Song, Usage, and Vocal. All tags are manually generated under controlled experimental conditions. In our work, we use the "hard" annotations found in CAL500, which give a binary value for all tags in every song indicating whether a tag applies to a song.

The *CAL10K* [17] dataset, S_{CAL10K}, is comprised of meta-data collected from the Pandora website. It consists of 10,886 songs annotated by expert musicologists who maintain a high level of agreement. This dataset contains 1053 unique tags. Furthermore, it contains a set of 55 tags in common with S_{CAL500}. We use this information to find the same subset of tags in the remaining datasets.

The *Magnatagatune* [15], S_{MAGNA}, contains annotations of 21,642 music clips with a subset of 188 different tags. The annotations are collected through an on-line game, referred to as "TagATune", developed to collect tags for music and sound clips. Each clip, 29 seconds in length, is an excerpt of music provided by (magnatune.com) and (freesound.org). All of the tags in the collection have been verified, i.e. a tag is associated with a clip only if it is generated independently by more than two players. Moreover, only those tags that are associated with more than 50 clips are included in the collection. It is important to note that Magnatagatune has not been used as widely as CAL500 due to its size and skewed tag distribution.

The *Million Song Dataset* (*MSD*) [4], is the largest MIR dataset publically available to date. Its purpose is to encourage research of scalable solutions and to provide a reference dataset for benchmarking. Unlike all other datasets, MSD is a conglomeration of complimentary datasets. For our simulations, we use the portion provided by LastFM (www.lastfm.com) excluding the known duplicates. For this reason, we denote it as S_{LASTFM}.

Since the S_{CAL500} dataset is too small to produce good sample size, we perform our simulation 10 times and obtain an arithmetic mean for each of our evaluation measures, similar to the 10-fold cross validation. Each time we choose the training set at random, and the remainder of the dataset becomes the testing set. This is done to preserve the 30/70 split used in our simulations, as described below. The other three datasets are large enough and do not require this kind of repeated random sub-sampling validation.

Dataset name	Number of songs	Number of labels	Label cardinality	Songs with at least 2 tags
S_{CAL500}	502	174	26.04	502
S_{CAL10K}	10886	1053	11.88	10886
S_{MAGNA}	25863	188	3.46	18097
$S_{\text{MAGNA-CLN}}$	25863	188	4.74	18097
$S_{\text{MAGNA-ADJ}}$	25863	166	5.15	13991
S_{LASTFM}	449503	522366	15.94	449503
$S_{\text{LASTFM-CLN}}$	449503	1046	8.76	280922
$S_{\text{LASTFM-ADJ}}$	449503	287	7.13	315254

Table 1: Music datasets and their statistics. The *label cardinality* of a dataset is the arithmetic mean of the number of labels per music piece in the dataset. Datasets denoted by CLN undergo simple data cleaning, such as the removal of songs with label cardinality less than 1. The datasets denoted by ADJ undergo the adjustment process depicted by Stage 4 in Figure 1 and discussed in Simulation 3.

3.2 Simulation 1

This simulation demonstrates G_1, that our approach is stable. Here, we randomly split a dataset in half and see if the resulting halves, H_1 and H_2, behave similarly in terms of our evaluation measures as outlined in Section 2.2. In this simulation, for each half, we go through all of the sages depicted by Figure 1 except the review and adjust steps in stage 4. First, we split each half into two subsets at random: we call one (30%) the *training set*, it corresponds to D, and the other (70%) corresponds to M and we call it the *testing set*. Then we compare the results from the scoring Algorithm between H_1 and H_2.

3.3 Simulation 2

Similarly to Simulation 1, we randomly divide each music dataset into a training set (30%), from which we derive association rules, and a testing set (70%), which we score against the association rules. They respectively correspond to D and M in Figure 1. Then we compare our evaluation measures between different datasets and hope to confirm G_2 that they maintain different relationships among their tags.

In this simulation we also use one dataset to verify another. Since CAL10K provides 55 tags that appear in all four datasets, we reduce all datasets to just those tags. Then we use one dataset as the training set D and another as the testing set M and perform all steps outlined in Figure 1 except the adjustment cycle.

3.4 Simulation 3

To fully demonstrate stage 4, we adjust the two datasets, S_{MAGNA} and S_{LASTFM}, by amalgamating some equivalent tags into one, thus reducing the diversity of tags in the datasets. For example, we convert several tags like $\{$*instrument singer male*$\}$, $\{$*male singer*$\}$ and $\{$*male voices*$\}$ into one tag $\{$*male vocals*$\}$. After this adjustment, we run the scoring Algorithm and hope to observe that the new PTAR is lower while the new STAR is higher.

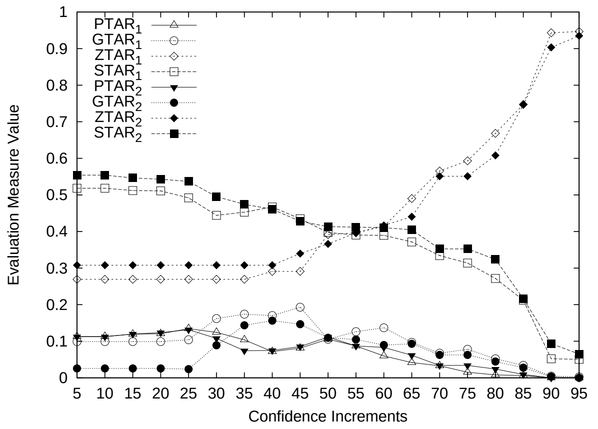

Figure 2: Comparison between H_1 (denoted by subscript 1) and H_2 (denoted by subsript 2) for S_{CAL10K} H_1 using our four evaluation measures.

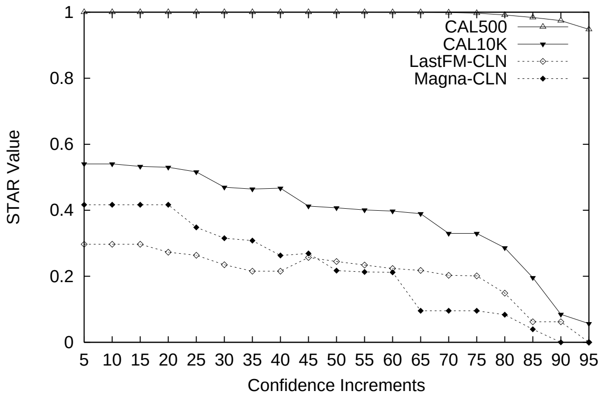

Figure 3: Comparison of STAR between four datasets.

4. RESULTS AND DISCUSSIONS

Simulation 1 demonstrates that our association-based tag annotation verifiction is stable. Figure 2 illustrates the different measure values that we obtain when we apply our approach to S_{CAL10K}. Here, the minimum support is set to 5% while the confident increases from 5% to 95%. We observe that STAR values are very similar at all confidence levels. The same applies to all other measures that we discuss in Section 2.2. We find similar results in the other three datasets across various minimum support thresholds and thus conclude our work toward G_1, i.e., our approach is stable. Due to space limit, we do not show all of the figures in this paper.

Simulation 2 applies our approach to all four music datasets. We obtain the values of our proposed four measures and report the STAR values for all four datasets in Figure 3. The figure shows that S_{CAL500} clearly achieves the highest STAR values, when compared to the other three datasets. The same argument is applied to other measures. For instance, we have observed lower PTAR values on S_{CAL500} but higher ones on $S_{\text{LASTFM-CLN}}$. Due to the space consideration, we do not report all of the figures here.

Our simulations clearly show that dataset S_{CAL500} has a better tag annotation quality than the other three datasets. Some similar observations have also been made in previous works, as mentioned in Section 1.1, and further suggested by the annotation processes as discussed in Section 3.1. Therefore, we used S_{CAL500} as the representative dataset D in Figure 1 to evaluate the other datasets, each considered as M in the same figure. As clearly shown in Figure 4, except for a few lower confidence ranges, S_{CAL10K} outperforms the other two datasets in terms of STAR. The same applies to the other measures, which are not shown due to space limitation.

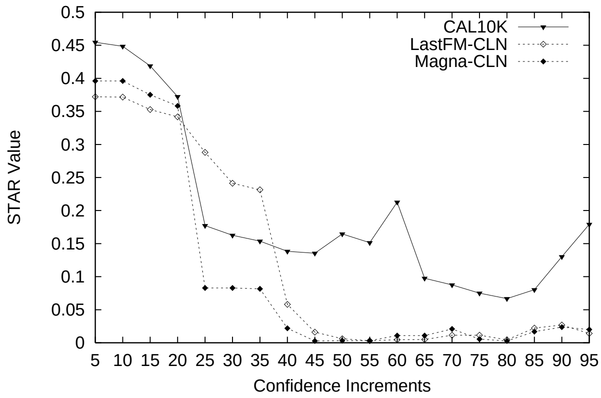

Figure 4: Using S_{CAL500} to verify the other datasets.

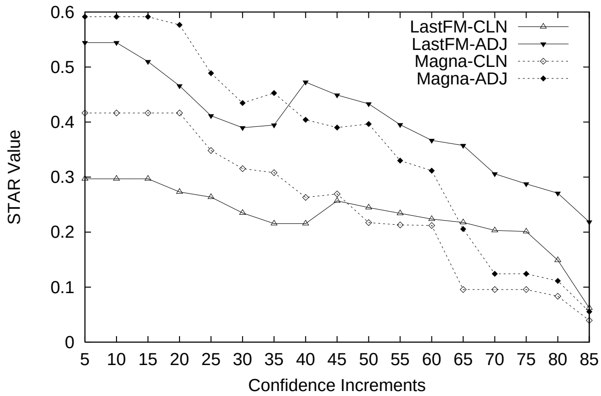

Figure 5: Comparison of STAR in two datasets before and after the adjustment step.

Towards our goal (G_3), we adjusted the two datasets that were evaluated as the worst in Simulation 2, namely the $S_{\text{LASTFM-CLN}}$ and the $S_{\text{MAGNA-CLN}}$. We then applied our approach to them again, as outlined in Stage 4 in Figure 1, hoping to see some improvements in terms of our evaluation measures. As presented in Figure 5, we clearly see that improvement. After the adjustment step, both datasets show better performance than before in terms of STAR. The same can be said about the other measures as well, which are not shown for the sake of space.

5. CONCLUSION

In this work, we have presented a novel approach to the verification of music tag annotation. We believe that there exist inherent associations among music tags that can be further utilized to verify and monitor a tag annotation process. The simulations presented here show the effectiveness of our approach.

There are several directions along which we can extend our work. It would be very interesting to explore whether we can use content-based information, such as MFCC or ZCR [5], in our analysis and verification process. We conjecture that this additional information will help improve our approach. Furthermore, we also plan to examine the individual rules that were generated for each music piece.

In addition, we could calculate the tag annotation rate for a specific category, such as style, mood and instrument. Furthermore, we could consider the representative music pieces of a single tag. For example, we could examine the tag annotation rules of the music genre *pop*. These rules may provide more insight into the nature of this genre and why a music piece is associated with it as opposed to others.

6. REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, volume 22, pages 207–216. ACM, 1993.

[2] Rakeshl Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, volume 1215, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.

[3] Thierry Bertin-Mahieux, Douglas Eck, Francois Maillet, and Paul Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596. University of Miami, 2011.

[5] Antti Eronen. *Signal Processing Methods for Audio Classification and Music Content Analysis*. PhD thesis, Tampere University of Technology, Tampere, Finland, June 2009.

[6] Joe Futrelle and Stephen J. Downie. Interdisciplinary communities and research issues in music information retrieval. In *Proceedings of the 3rd International Society for Music Information Retrieval Conference*, pages 121–131, 2002.

[7] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., the second edition, 2006.

[8] Matthew D Hoffman, David M Blei, and Perry R Cook. Easy as CBA: A simple probabilistic model for tagging music. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, volume 9, pages 369–374, 2009.

[9] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference Companion on World Wide Web*, pages 909–916, 2012.

[10] Steven R Ness, Anthony Theocharis, George Tzanetakis, and Luis Gustavo Martins. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 705–708. ACM, 2009.

[11] Kerstin Neubarth, Izaro Goienetxea, Colin Johnson, and Darrell Conklin. Association mining of folk music genres and toponyms. In *ISMIR*, pages 7–12, 2012.

[12] Franois Pachet. Content management for electronic music distribution. *Communications of the ACM*, 46(4):71–75, 2003.

[13] Chris Sanden and John Z. Zhang. An empirical study of multi-label classifiers for music tag annotation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 717–722, 2011.

[14] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pages 469–474, 2012.

[15] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. Automatic music tag classification based on block-level features. In *Proceedings of the Sound and Music Computing Conference*, pages 126–133, 2010.

[16] Jialie Shen, Wang Meng, Shuichang Yan, HweeHwa Pang, and Xiansheng Hua. Effective music tagging through advanced statistical modeling. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 635–642. ACM, 2010.

[17] Youngmoo E. Kim Tingle, Derek and Douglas Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the international conference on Multimedia Information Retrieval*, pages 55–62. ACM, 2010.

[18] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467–476, February 2008.

[19] Kris West. *Novel Techniques for Audio Music Classification and Search*. PhD thesis, University of East Anglia, UK, September 2008.