EVALUATION OF ALGORITHMS USING GAMES: THE CASE OF MUSIC TAGGING

Edith Law
CMU
edith@cmu.edu

Kris West
IMIRSEL/UIUC
kris.west@gmail.com

Michael Mandel
Columbia University
mim@ee.columbia.edu

Mert Bay J. Stephen Downie
IMIRSEL/UIUC
mertbay, jdownie@uiuc.edu

ABSTRACT

Search by keyword is an extremely popular method for retrieving music. To support this, novel algorithms that automatically tag music are being developed. The conventional way to evaluate audio tagging algorithms is to compute measures of agreement between the output and the ground truth set. In this work, we introduce a new method for evaluating audio tagging algorithms on a large scale by collecting set-level judgments from players of a human computation game called TagATune. We present the design and preliminary results of an experiment comparing five algorithms using this new evaluation metric, and contrast the results with those obtained by applying several conventional agreement-based evaluation metrics.

1. INTRODUCTION

There is a growing need for efficient methods to organize and search for multimedia content on the Web. This need is reflected in the recent addition of the audio tag classification (ATC) task at MIREX 2008, and the introduction of new music tagging algorithms [1, 2]. The conventional way to determine whether an algorithm is producing accurate tags for a piece of music is to compute the level of agreement between the output generated by the algorithm and the ground truth set. Agreement-based metrics, e.g. accuracy, precision, F-measure and ROC curve, have been long-time workhorses of evaluation, accelerating the development of new algorithms by providing an automated way to gauge performance.

The most serious drawback to using agreement-based metrics is that ground truth sets are never fully comprehensive [3]. First, there are exponentially many sets of suitable tags for a piece of music – creating all possible sets of tags and then choosing the best set of tags as the ground truth is difficult, if not impossible. Second, tags that are appropriate for a given piece of music can simply be missing in the ground truth set because they are less salient, worded differently (e.g. *baroque* versus *17th century classical*), or that they do not facilitate the objectives of the particular annotator. For example, a last.FM user who wants to showcase his expertise on jazz music may tag the music with highly obscure and technical terms. In output-agreement games such as MajorMiner [2] and the Listen Game [4], where the scoring depends on how often players' tags match with one another, players are motivated to enter (or select) tags that are common, thereby omitting tags that are rare or verbose. Furthermore, because an exhaustive set of negative tags is impossible to specify, when a tag is missing, it is impossible to know whether it is in fact inappropriate for a particular piece of music.

Agreement-based metrics also impose restrictions on the type of algorithms that can be evaluated. To be evaluated, tags generated by the algorithms must belong to the ground truth set. This means that audio tagging algorithms that are not trained on the ground truth set, e.g. those that use text corpora or knowledge bases to generate tags, cannot be evaluated using agreement-based metrics.

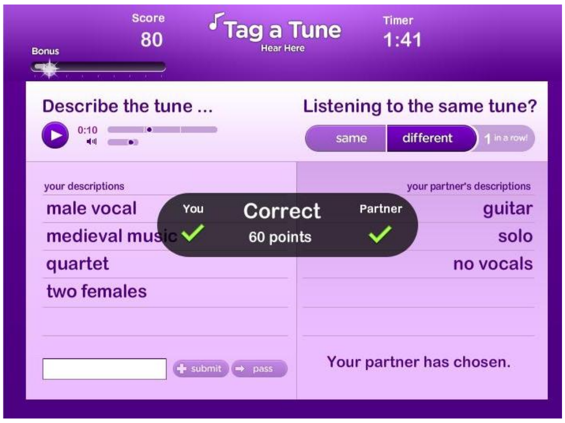
To be useful, tags generated by audio tagging algorithms must, from the perspective of the *end user*, accurately describe the music. However, because we do not yet fully understand the cognitive processes underlying the representation and categorization of music, it is often difficult to know what makes a tag "accurate" and what kinds of inaccuracies are tolerable. For example, it may be less disconcerting for users to receive a *folk* song when a *country* song is sought, than to receive a *sad, mellow* song when a *happy, up-beat* song is sought. Ideally, an evaluation metric should measure the quality of the algorithm by implicitly or explicitly capturing the users' differential tolerance of incorrect tags generated by the algorithms. The new evaluation metric we are proposing in this paper has exactly this desired property.

The problems highlighted above suggest that music tagging algorithms, especially those used to facilitate retrieval, would benefit enormously from evaluation by human users. Manual evaluation is, however, often too time-consuming or costly to be feasible. Human computation is a new area of research that studies how to build systems, such as simple casual games, to collect annotations from human users. In this work, we investigate the use of a human computation game called TagATune to collect evaluations of algorithm-generated music tags. In an off-season MIREX [5] evaluation task, we compared the performance of five audio tagging algorithms under the newly proposed metric, and present in this paper the preliminary findings.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2009 International Society for Music Information Retrieval.

2. TAGATUNE AS AN EVALUATION PLATFORM

TagATune [6] is a two-player online game that collects music tags from players. In each round of the game, two players are either given the same music clip or different music clips, and are asked to type in tags for their given music clip. After seeing each other's tags, players must then decide whether they were given the same music clip or not.



male vocal
medieval music
quartet
two females

guitar
solo
no vocals

Correct
60 points

Your partner has chosen.

Figure 1. The TagATune interface

When a human partner is not available, a player is paired with a computer bot, which outputs tags that have been previously collected by the game for the particular music clip served in each round. This so-called *aggregate bot* serves tags that are essentially the ground truth, since they were provided by human players.

The key idea behind TagATune as an evaluation platform is that the aggregate bot can be replaced by an *algorithm bot*, which enters tags that were previously generated by an algorithm. An interesting by-product of playing against an algorithm bot is that by guessing same or different, the human player is essentially making a judgment on the appropriateness of the tags generated by the algorithm. Unlike the conventional evaluation metrics where a tag either matches or does not match a tag in the ground truth set, this evaluation method involves set-level judgments and can be applied to algorithms whose output vocabulary is ***arbitrarily different*** from that of the ground truth set.

2.1 Special TagATune Evaluation

To solicit submissions of audio tagging algorithms whose output can be used to construct the TagATune algorithm bots, a "Special TagATune Evaluation" was run off-season under the MIREX rubric. Participating algorithms were asked to provide two different types of outputs:

1. a binary classification decision as to whether each tag is relevant to each clip.

2. a real valued estimate of the 'affinity' of the clip for each tag. Larger values of the affinity score indicate that a tag is more likely to be applicable to the clip.

2.1.1 The Dataset

In the context of the off-season MIREX evaluation task, we trained the participating algorithms on a subset of the TagATune dataset, such that the tags they generated could be served by the algorithm bots in the game. The training and test sets comprise of 16289 and 100 music clips respectively. The test set was limited to 100 clips for both the human evaluation using TagATune and evaluation using the conventional agreement-based metrics, in order to facilitate direct comparisons of their results. Each clip is 29 seconds long, and the set of clips are associated with 6622 tracks, 517 albums and 270 artists. The dataset is split such that the clips in the training and test sets do not belong to the same artists. Genres include Classical, New Age, Electronica, Rock, Pop, World, Jazz, Blues, Metal, Punk etc. The tags used in the experiments are each associated with more than fifty clips, where each clip is associated only with tags that have been verified by more than two players independently.

2.1.2 Participating Algorithms

There were five submissions, which we will refer to as Mandel, Manzagol, Marsyas, Zhi and LabX[1] from this point on. A sixth algorithm we are using for comparison is called AggregateBot, which serves tags from a vocabulary pool of 146 tags collected by TagATune since deployment, 91 of which overlap with the 160 tags used for training the algorithms. The inclusion of AggregateBot demonstrates the utility of TagATune in evaluating algorithms that have different tag vocabulary.

2.1.3 Game-friendly Evaluation

An important requirement for using human computation games for evaluation is that the experiment does not significantly degrade the game experience. We describe here a few design strategies to maintain the enjoyability of the game despite the use of algorithm bots whose quality cannot be gauged ahead of time.

First, a TagATune round is randomly chosen to be used for evaluation with some small probability x. This prevents malicious attempts to artificially boost or degrade the evaluation of particular algorithms, which would be easy to do if players can recognize that they are playing against an algorithm bot. Second, while it may be acceptable to use half of the rounds in a game for evaluating good algorithms, one round may be one too many if the algorithm under evaluation always generates completely wrong tags. Since we do not know ahead of time the quality of the algorithms being evaluated, x must be small enough such that the effects of bad algorithms on the game will be minimized. Finally, using only a small portion of the game for evaluation ensures that a wide variety of music is served, which is especially important when the test set is small.

[1] The LabX submission was identified as having a bug which negatively impacted its performance, hence, the name of the participating laboratory has been obfuscated. Since LabX essentially behaves like an algorithm that randomly assigns tags, its performance establishes a lower bound for the TagATune metric.

Despite the small probability of using each round for evaluation, the game experience can be ruined by an algorithm that generates tags are contradictory (e.g. *slow* followed by *fast*, or *guitar* followed by *no guitar*) or redundant (e.g. *string*, *violins*, *violin*). Our experience shows that players are even less tolerant of a bot that appears "stupid" than of one that is wrong. Unfortunately, such errors occur quite frequently. Table 1 provides a summary of the number of tags generated (on average) by each algorithm for the clips in the test set, and how many of those are removed because they are contradictory or redundant.

Algorithm	Generated	Contradictory or Redundant
Mandel	36.47	16.23
Marsyas	9.03	3.47
Manzagol	2.82	0.55
Zhi	14.0	5.04
LabX	1.0	0.00

Table 1. Average number of tags generated by algorithms and contradictory/redundant ones among the generated tags

To alleviate this problem, we perform the following post-processing step on the output of the algorithms. First, we retain only tags that are considered relevant according to the binary outputs. Then, we rank the tags by affinity. Finally, for each tag, starting from the highest affinity, we remove lower affinity tags with which it is mutually exclusive. Although this reduces the number of tags available to the algorithm bots to serve in the game, we believe that this is a sensible post-processing step for any tag classification algorithms.

An alternative method of post-processing would be to first organize the "relevant" tags into categories (e.g. genre, volume, mood) and retain only the tag with the highest affinity score in each category, thereby introducing more variety in the tags to be emitted by the algorithm bots. We did not follow this approach because it may bias performance in an unpredictable fashion and favour the output of certain algorithms over others.

2.1.4 Evaluation Using The TagATune Metric

During an evaluation round, an algorithm is chosen to emit tags for a clip drawn from the test set. The game chooses the algorithm-clip pair in a round robin fashion but favors pairs that have been seen by the least number of unique human players. In addition, the game keeps track of which player has encountered which algorithm-clip pair, so that each evaluator for a given algorithm-clip pair is unique.

Suppose a set of algorithms $\mathcal{A} = \{a_i, \ldots, a_{|\mathcal{A}|}\}$ and a test set $\mathcal{S} = \{s_j, \ldots, s_{|\mathcal{S}|}\}$ of music clips. During each round of the game, a particular algorithm i is given a clip j from the test set and asked to generate a set of tags for that clip. To be a valid evaluation, we only use rounds where the clips given to the human player and the algorithm bot are the same. This is because if the clips are different, an algorithm can output the wrong tags for a clip and actually *help* the players guess correctly that the clips are different.

A human player's guess is denoted as $G = \{0, 1\}$ and the ground truth is denoted as $GT = \{0, 1\}$, where 0 means that the clips are the same and 1 means that the clips are different. The performance P of an algorithm i on clip j under TagATune metric is as follows:

$$P_{i,j} = \frac{1}{N} \sum_{n}^{N} \delta(G_{n,j} = GT_j) \qquad (1)$$

where N represents the number of players who were presented with the tags generated by algorithm i on clip j, and $\delta(G_{n,j} = GT_j)$ is a Kronecker delta function which returns 1 if, for clip j, the guess from player n and the ground truth are the same, 0 otherwise. The overall score for an algorithm is averaged over the test set S:

$$P_i = \frac{1}{S} \sum_{j}^{S} P_{i,j} \qquad (2)$$

2.1.5 Evaluation Using Agreement-Based Metrics

We have chosen to compute the performance of the participating algorithms using a variety of agreement-based metrics that were included in the 2008 MIREX ATC task, as a comparison against the TagATune metric. These metrics include precision, recall, F-measure [7], the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the accuracy of the positive and negative example sets for each tag. We omitted the "overall accuracy" metric, as it is a very biased statistics for evaluating tag classification models where there is a large negative to positive tag ratio.

As the TagATune game and metric necessarily focus on the first few tags returned by an algorithm (i.e. tags that have the highest affinity scores), we chose to also calculate the Precision-at-N ($P@N$) score for each algorithm. This additional set of statistics allows us to explore the effect of sampling the top few tags on the performance of the algorithms.

2.1.6 Statistical Significance

Friedman's ANOVA is a non-parametric test that can be used to determine whether the difference in performance between algorithms is statistically significant [5].

For each algorithm, a performance score is computed over the test set. Using the TagATune metric, this performance score is the percentage of unique players that correctly judged that the clips are the same or not using the tags emitted by the algorithm, computed using equation (1) and (2). For automated statistical evaluations, such as those performed during the MIREX ATC task, these may be the F-measure or $P@N$ for the "relevant" tags generated for each clip, or the AUC-ROC for the "affinity" scores. These scores can be viewed as a rectangular matrix, with the different tagging algorithms represented as the columns and the clips (or the tags, in the case of F-measure aggregated over each tag) forming the rows.

To avoid having variance introduced by different tags affecting the scaling and distribution of the scores, Friedman's test replaces the performance scores with their ranks amongst the algorithms under comparison.

Algorithm	**TagATune metric**	+ve Example Accuracy	-ve Example Accuracy	Precision	Recall	**F-measure**
AggregateBot	**93.00%**	–	–	–	–	–
Mandel	**70.10%**	**73.13%**	80.29%	0.1850	**0.7313**	0.2954
Marsyas	68.60%	45.83%	96.82%	**0.4684**	0.4583	**0.4633**
Manzagol	67.50%	13.98%	98.99%	0.4574	0.1398	0.2141
Zhi	60.90%	40.30%	93.18%	0.2657	0.4030	0.3203
LabX	26.80%	0.33%	**99.36%**	0.03	0.0033	0.0059

Table 2. Evaluation statistics under the TagATune versus agreement-based metrics

Algorithm	Precision at N					Precision for 'relevant' tags	AUC-ROC
	3	6	9	12	15		
Mandel	0.6133	0.5083	0.4344	0.3883	0.3387	0.1850	0.8514
Marsyas	**0.7433**	**0.5900**	**0.4900**	**0.4308**	**0.3877**	**0.4684**	**0.9094**
Manzagol	0.4767	0.3833	0.3222	0.2833	0.2520	0.4574	0.7521
Zhi	0.3633	0.3383	0.3100	0.2775	0.2480	0.2657	0.6697
LabX	–	–	–	–	–	0.03	–

Table 3. Precision and AUC-ROC statistics collected for each algorithm

Friedman's ANOVA is used to determine if there exists a significant difference in performance amongst a set of algorithms. If a difference is detected, then it is common to follow up with a Tukey-Kramer Honestly Significant Difference (TK-HSD) test to determine which pairs of algorithms are actually performing differently. This method does not suffer from the problem that multiple t-tests do where the probability of incorrectly rejecting the null hypothesis (i.e. that there is no difference in performance) increases in direct proportion to the number of pairwise comparisons conducted.

3. RESULTS

Tables 2 and 3 provide summaries of the evaluation statistics collected for each algorithm under the TagATune metric as well as agreement-based metrics. Each of the summary results was computed over the 100 clips in the test set, while the statistical significance tests were computed over the results for each individual clip. The following sections detail additional statistics that were collected by the TagATune evaluation.

3.1 Algorithm Ranking

According to the TK-HSD test on the TagATune metric results, AggregateBot's performance is significantly better than all the others. A second group of equally performing algorithms consists of Mandel, Manzagol, Marsyas, and Zhi. LabX is the sole member of the worst performing group. Figure 2 highlights these TK-HSD performance groupings.

Several authors have speculated on the possibility of a "glass-ceiling" on the performance of current music classification and similarity estimation techniques. As identified by Aucouturier [8], many of these techniques are based on 'bag-of-frames' approaches to the comparison of the audio streams. Hence, the lack of a significant difference among the performances of the correctly functioning algorithms is not surprising.

The TK-HSD ordering of the algorithms using the F-measure scores (Table 2 and Figure 3) is different from that produced by the TagATune scores. Notably, the Marsyas algorithm significantly outperforms the other algorithms and the Zhi algorithm has improved its relative rank considerably.

These differences may be attributed to the fact that the performance of the Marsyas and Zhi algorithm is more balanced in terms of precision and recall than the Mandel algorithm (which exhibits high recall but low precision) and the Manzagol algorithm (which exhibits high precision but low recall). This conclusion is reinforced by the positive and negative accuracy scores, which demonstrate the tendency of the Mandel algorithm to over-estimate and Manzagol to under-estimate relevancy. Metrics that take into account the accuracies of all tags (e.g. F-measure) are particularly sensitive to these tendencies, while metrics that consider only the top N tags (e.g. the TagATune metric and $P@N$) are affected little by them.

These results suggest that the choice of an evaluation metric or experiment must take into account the intended application of the tagging algorithms. For example, the TagATune metric may be most suitable for evaluating retrieval algorithms that use only the highest ranked tags to

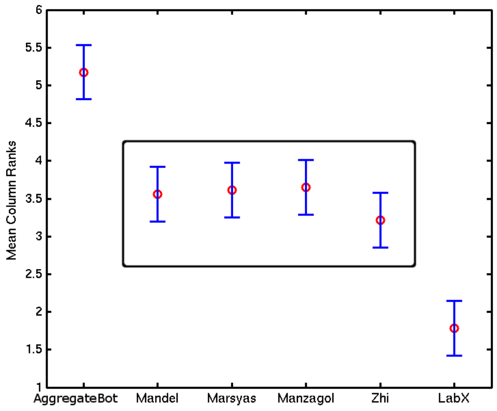


Figure 2. Tukey-Kramer HSD results based on the TagATune metric

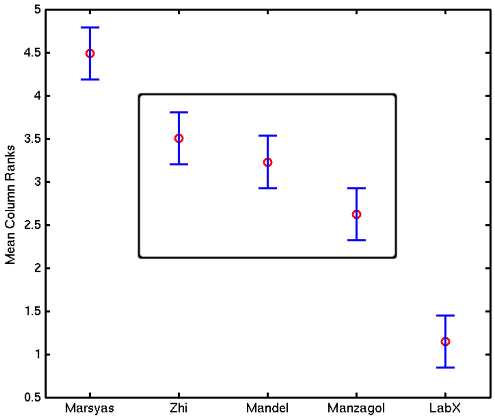


Figure 3. Tukey-Kramer HSD results based on the F-measure metric

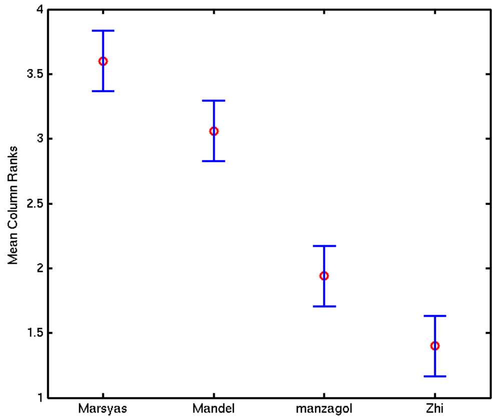


Figure 4. Tukey-Kramer HSD results based on the AUC-ROC metric

compute the degree of relevance of a song to a given query. However, for applications that consider the all relevant tags regardless of affinity, e.g. unweighted tag clouds generators, the TagATune metric is not necessarily providing an accurate indication of performance, in which case the F-measure might be a better candidate.

3.2 Game Statistics

In a TagATune round, the game selects a clip from the test set and serves the tags generated by a particular algorithm for that clip. For each of the 100 clips in the test set and for each algorithm, 10 unique players were elicited (unknowingly) by the game to provide evaluation judgments. This totals to 5000 judgments, collected over a one month period, involving approximately 2272 games and 657 unique players.

3.2.1 Number of tags reviewed

One complication with using TagATune for evaluation is that players are allowed to make the decision of guessing same or different at any point during a round. This means that the number of tags reviewed by the human player varies from clip to clip, algorithm to algorithm. As a by-product of game play, players are motivated to guess as soon as they believe that they have enough information to guess whether the clips are the same or different. Figure 5, which shows that players reviewed only a small portion of the generated tags before guessing, reflects this situation.

3.2.2 Correlation with precision

Figure 6 shows the average number of tags reviewed by players and how many of the reviewed tags are actually true positive tags (according to the ground truth) in success rounds (where the human player made the correct guess) versus failed rounds (where the human player made the wrong guess). Results show that generally the number of true positive tags reviewed is greater in success rounds than in failed rounds, suggesting that players are more likely to fail at guessing when there are more top-affinity tags that are wrong. Additionally, the average number of tags reviewed before guessing is fewer in the failed rounds than in the success rounds, with the exception of Mandel, possibly due to outliers and the much greater number of tags that this algorithm returns. This suggests that players make their guesses more hastily when algorithms make mistakes.

3.2.3 Detectable errors

A natural question to ask is whether one can detect from game statistics which of the reviewed tags actually caused players to guess incorrectly.

System	failed round	success round
Mandel	86.15%	49.00%
Marsyas	80.49%	45.00%
Manzagol	76.92%	33.33%
Zhi	84.38%	70.10%
LabX	100.0%	95.77%

Table 4. Percentage of the time that the last tag displayed before guessing is wrong in a failed round versus success round

To investigate this question, we consult the game statistics for the most frequent behavior of human players in terms of the number of tags reviewed before guessing, in the case when the guess is wrong. For example, we might find that most players make a wrong guess after reviewing n tags for a particular algorithm-clip pair. The hypothesis is that the last tag reviewed before guessing, i.e. the n^{th} tag, is the culprit.

Table 4 shows the percentage of times that the n^{th} tag is actually wrong in failed rounds, which is above 75% for all algorithms. In contrast, the probability of the last tag being wrong is much lower in success rounds, showing that using game statistics alone, one can detect problematic tags that cause most players to make the wrong guess in the game. This trend does not hold for LabX, possibly because players were left guessing randomly due to the lack of information (since this algorithm generated only one tag per clip).

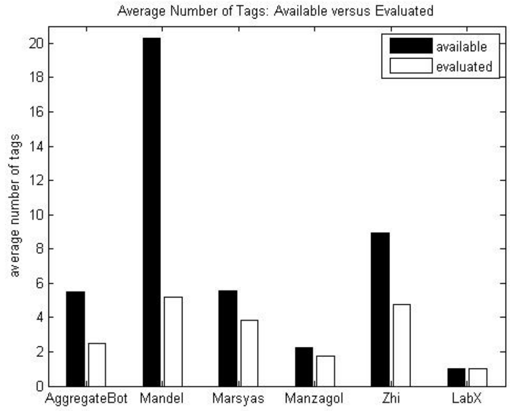


Figure 5. Number of tags available and reviewed by players before guessing

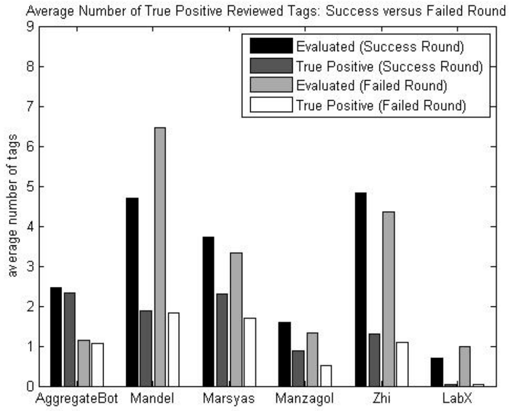


Figure 6. Number of overall and true positive tags evaluated in success and failed rounds

4. CONCLUSION

This paper introduces a new method for evaluating music tagging algorithms and presents the results of a proof-of-concept experiment using a human computation game as an evaluation platform for algorithms. This experiment has also been used to explore the behavior of conventional agreement-based metrics and has shown that averaged retrieval statistics, such as F-measure, can be sensitive to certain tendencies (e.g. imbalanced performance in terms of precision versus recall) that do not affect the TagATune metric, which considers the accuracies of only the top most relevant tags.

While there are many benchmarking competitions for algorithms, little is said about the level of performance that is acceptable for real world applications. In this work, we have shown that the use of aggregate data in the bot provides a performance level against which the algorithms can be judged. Specifically, human players can correctly guess that the music are the same 93% of the times when paired against the aggregate bot, while only approximately 70% of the times when paired against an algorithm bot.

Finally, our work has shown that TagATune is a feasible and cost-effective platform for collecting a large number of evaluations from human users in a timely fashion. This result is particularly encouraging for future research on using human computation games to evaluate algorithms in other domains, such as object recognition and machine translation.

5. ACKNOWLEDGMENT

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grants NSF IIS-0328471 and NSF IIS-0327371. Edith Law is generously funded by a Microsoft Graduate Research Fellowship. Many thanks to those who participated in the MIREX Special TagATune Evaluation.

6. REFERENCES

[1] Doug Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *SIGIR*, pages 439–446, 2007.

[2] Michael I. Mandel and Daniel P. W. Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[3] Edith Law. The problem of accuracy as an evaluation criterion. *ICML Workshop on Evaluation Methods in Machine Learning*, 2008.

[4] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert Lanckriet. A game-based approach for collecting semantic annotations of music. In *ISMIR*, pages 535–538, 2007.

[5] J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[6] Edith Law and Luis von Ahn. Input-agreement: A new mechanism for data collection using human computation games. *CHI*, pages 1197–1206, 2009.

[7] Keith Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979.

[8] Jean-Julien Aucouturier. *Ten experiments on the modelling of polyphonic timbre*. PhD thesis, University of Paris 6, France, June 2006.