THE 2007 MIREX AUDIO MOOD CLASSIFICATION TASK: LESSONS LEARNED

Xiao Hu[1] J. Stephen Downie[1] Cyril Laurier[2] Mert Bay[1] Andreas F. Ehmann[1]
[1]International Music Information Retrieval System Evaluation Laboratory
University of Illinois at Urbana-Champaign
{xiaohu, jdownie, mertbay, aehmann}@uiuc.edu
[2]Music Technology Group, Universitat Pompeu Fabra claurier@iua.upf.edu

ABSTRACT

Recent music information retrieval (MIR) research pays increasing attention to music classification based on moods expressed by music pieces. The first Audio Mood Classification (AMC) evaluation task was held in the 2007 running of the Music Information Retrieval Evaluation eXchange (MIREX). This paper describes important issues in setting up the task, including dataset construction and ground-truth labeling, and analyzes human assessments on the audio dataset, as well as system performances from various angles. Interesting findings include system performance differences with regard to mood clusters and the levels of agreement amongst human judgments regarding mood labeling. Based on these analyses, we summarize experiences learned from the first community scale evaluation of the AMC task and propose recommendations for future AMC and similar evaluation tasks.

1. INTRODUCTION

With the goal of systematically evaluating state-of-the-art algorithms for Music Information Retrieval (MIR) systems, the Annual Music Information Retrieval Evaluation eXchange (MIREX) included an Audio Mood Classification (AMC) task for the first time in 2007. It is inspired by MIR researchers' growing interest in classifying music by moods (e.g. [4][7][8]), and the difficulty in the evaluation of music mood classification caused by the subjective nature of mood. Most previous experiments on music mood classification used different mood categories and datasets, raising a great challenge in comparing systems. MIREX, as the largest evaluation event in the MIR community, is a good venue to build an available audio dataset and ground-truth for AMC and to facilitate collaborations among MIR researchers around the world.

The first AMC task in MIREX was a success. A ground-truth set of 600 tracks distributed across five mood categories was built based on metadata analysis and human assessments. A total of nine systems from Europe and North America participated in the task. Resultant accuracies ranged from 25.67% to 61.50%, with an average of 52.65%, a median of 55.83% and a standard deviation of 11.19%.

In this paper, we examine the evaluation process and explore the datasets in detail. We also analyze the system performances with a special focus on the possible effects that the data creation and evaluation process might have. The findings will help in organizing similar MIREX-style evaluations in the future.

The rest of the paper is organized as follows: Section 2 describes critical issues in preparing and carrying out the AMC evaluation task. Section 3 analyzes the human assessments made on candidate audio pieces for the purpose of building a ground-truth set. Statistical analyses on system performances from multiple angles are presented in Section 4, and Section 5 wraps up with discussions and recommendations based on the findings from the analyses.

2. EVALUATION SETUP

2.1. Dataset

A standard dataset used in evaluating a classification task should include a set of categories, a collection of samples distributed across these categories and ground-truth labels that ideally are given by agreements among multiple human judges. The AMC task adopted the set of five mood clusters proposed in [6] which effectively reduce the mood space into a manageable set. For clarity purposes, we present the mood clusters in Table 1. The words in each cluster collectively define the "mood spaces" associated with the cluster.

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Rowdy	Amiable/	Literate	Witty	Volatile
Rousing	Good natured	Wistful	Humorous	Fiery
Confident	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense/anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Table 1. Five mood clusters used in the AMC task

2.1.1. Audio Track Collection

Our previous study [6] shows that music mood, as a descriptive metadata type, is independent of music genre. Thus, it is desirable to evaluate mood classification algorithms against a collection of music pieces in a variety of genres. Furthermore, the ideal collection should have not been used in previous evaluations so as to avoid the overfitting problem caused by repetitive uses of the same collection. Keeping these criteria in mind, we chose the libraries of Associated Production Music (APM) as the candidate pool. The APM collection is "the world's leading production music library… offering every imaginable music genre from beautiful classical music recordings to vintage rock to current indie band sounds"[1]. This collection is made available to the MIR community under a contract of academic use between APM and the IMIRSEL lab. It contains 206,851 tracks pooled from 19 libraries produced by APM and covers 27 genres. Each track is annotated with rich, descriptive metadata including instrument, tempo, style, etc. One such metadata field is called "category" which contains a list of descriptors including 32 mood-related ones (e.g. "Moods-Aggressive", "Moods-Quirky"). Such descriptors are particularly useful in building the evaluation dataset using the following steps:

Step 1. Eliminate tracks without genre information, so that we can explicitly select tracks in diversified genres;

Step 2. A set of rules are manually designed according to statistics of the mood-related descriptors, and tracks matching the rules are selected and pre-labeled with the mood clusters specified by the rules. An exemplar rule is "if 'Moods-Quirky' ∈ Song1.category, then Song1 ∈ Cluster 4";

Step 3. Eliminate tracks shorter than 50 seconds, as a means to reduce the chances of including non-music content in the extracted 30 seconds clips (see below);

Step 4. Eliminate tracks from the same CDs or same libraries, for diversity purposes.

The result is a collection of 1250 tracks with 250 pieces in each mood cluster. These tracks were then truncated into 30-second clips and were taken as candidates to be judged by human evaluators.

The choice of using 30-second clips over whole tracks is due to several considerations. First, it lessens the burden on our human evaluators. Second, it reduces the runtime needed in audio processing. Finally, it alleviates variation caused by the changing nature of music mood, i.e., some pieces may start from one mood but end in another. The clips are extracted from the middle of the tracks which are assumed to be more representative than other parts.

2.1.2. Ground-truth and Human Assessment

Our goal was to build a ground-truth set of 600 clips with 120 in each mood cluster. These numbers were decided by polling opinions from potential participants via the AMC task wiki [2]. We planned to have multiple human assessments on each of the 1250 candidates selected by the metadata statistics. Based on a pilot experiment on selecting exemplar songs, we estimated at least half of the candidates would achieve majority agreements on their mood labels among human assessments. We used the Evalutron 6000 [5], a Web-based device designed for collecting human judgments for MIREX evaluations. As music mood is very subjective and judging music mood is a new task for human evaluation, we designed concise but explicit instructions to help control variations among the human assessors. The most important instructions include:

1) Ignoring lyrics. Many clips have lyrics which often express certain moods and thus could affect listeners' judgments. However, state-of-the-art audio music processing technology has not yet been developed to sufficiently transcribe lyrics. Hence, we should try our best to mitigate possible bias imposed by lyrics.

2) Mini-training on exemplar songs. In order to better articulate what the mood clusters mean, we prepared a set of exemplar songs in each mood cluster that were unanimously judged by 6 IMIRSEL members. We added special settings to the Evalutron to ensure an assessor cannot complete registration until finishing listening to the excerpts of the exemplar songs.

3) "Other" category. A human assessor can change the pre-selected label on a candidate piece if she does not agree with it. To better capture assessors' opinions, we designed an "Other" category for the cases when none of the five clusters seems appropriate to the assessors.

2.2. Evaluation Method

As in many other evaluations on classification tasks, the submitted systems were trained, tested and evaluated using cross-validation. In particular, the AMC task was evaluated using 3-fold cross-validation. As a starting point for evaluating on music mood classification, the AMC task was defined as a single-label classification problem, i.e., each song can only be classified into one mood cluster. The classification results were evaluated and compared using classification accuracy as the measure. There is also an alternative evaluation approach proposed during task preparation that is discussed in Section 5. The audio format used in the task was 22KHz mono-channel wav files, as voted on by the majority of the potential participants on the AMC task wiki.

[1] www.apmmusic.com/pages/aboutapm.html

[2] http://www.music-ir.org/mirex/2007/index.php/AMC

3. HUMAN ASSESSMENT ANALYSIS

3.1. Statistics on the Evalutron Data

Human assessment data were collected between 1 Aug. and 19 Aug., 2007, from volunteer assessors from the MIR/MDL research community, representing 5 different countries from Europe and the U.S. Among the 21 volunteers registered on the Evalutron, 15 of them provided at least one judgment and 8 of them finished assessing all assigned 250 clips while each of the remaining assessors completed 6 to 140 assessments.

Table 2 shows the number of human agreements for clips with at least two judgments for each of the mood clusters (denoted as "C1" to "C5"), produced by early September 2007. To build the ground-truth, we needed clips with at least two agreed judgments. As can be seen from the table, there were not enough such clips (< 120) for Cluster 1, 2 and 4. As a means to quickly compensate for the missing judgments, the IMIRSEL lab collected clips with only one judgment in these clusters and organized an in-house assessment on these clips.

	# of clips with 3 or 2 agreements	# of clips with no agreement	Total
C1	102 (57.6%)	75 (43.4%)	177
C2	105 (64.0%)	59 (36.0%)	164
C3	147 (79.9%)	37 (20.1%)	184
C4	95 (60.5%)	62 (39.5%)	157
C5	137 (75.3%)	45 (24.7%)	182
Total	586 (67.8%)	278 (32.2%)	864

Table 2. Agreements on clips with 3 or 2 judgments

The fact that Cluster 3 and 5 had enough agreed judgments and also allowed the highest agreement rate (> 70%) reflects that the clips pre-selected in these two clusters caused fewer confusions among human assessors. This possibly means the pre-labeling methods worked better for Cluster 3 and 5 than Cluster 1 (where agreement rate was the lowest) and/or that Cluster 3 and 5 were easier to assess than Cluster 1. In Section 4, we will see the participating systems performed better in Cluster 3 and 5 as well.

3.2. Effects on Reclassification

As human assessors were asked to change the pre-assigned labels whenever appropriate, it is interesting to see how aggregated human opinions differ from pre-assigned labels. We only consider pieces with at least two agreed human judgments. Table 3 shows how the agreed assessments concurred with or changed the pre-assigned labels. For example, among the clips pre-assigned to Cluster 1, 59.8% were kept in Cluster 1 by agreed human judgments while 40.2% of them were reclassified to other clusters including 17.6% to the "other" category (far higher than other mood clusters). For other clusters, more than 84% of the clips kept their pre-assigned labels. Cluster 1 seemed to have caused the greatest confusion with other clusters, according to the human assessments.

	C1	C2	C3	C4	C5	Other
C1	59.8%	11.8%	17.6%	6.9%	0.0%	17.6%
C2	1.0%	89.5%	5.7%	1.0%	0.0%	2.9%
C3	2.0%	1.4%	92.5%	0.7%	0.7%	2.7%
C4	0.0%	7.4%	1.1%	84.2%	2.1%	5.3%
C5	4.4%	0.0%	1.5%	1.5%	89.8%	2.9%

Table 3. Percentage distribution of agreed judgments (rows: pre-assigned labels, columns: human assessment)

3.3. Three Judgment Sets

Among the 600 audio pieces eventually used in the AMC task, 153 of them were agreed upon by 3 human judges (denoted as "Set 1"), 134 were assessed by 3 judges but only 2 of the judges reached agreement ("Set 2"), and 313 pieces were assessed by 2 judges who agreed on the mood label assignments ("Set 3"). Table 4 demonstrates the distribution of clips in these sets across mood clusters. In a later analysis (Section 4.4), we will investigate if the number of agreed judgments made a difference in system performances.

	C1	C2	C3	C4	C5	Total
Set 1	21	24	56	21	31	153
Set 2	41	35	18	26	14	134
Set 3	58	61	46	73	75	313
Total	120	120	120	120	120	600

Table 4. Number of audio pieces in different judgment sets across mood clusters

4. SYSTEM PERFORMANCE ANALYSIS

Nine systems participated in the AMC 2007 task. The average classification accuracy scores over a 3-fold cross-validation, as well as run time information were published on the MIREX results wiki[1]. Table 5 presents the accuracy results ($\in [0, 1]$) for each system broken down by fold, along with its overall average accuracy.

In this section, we examine the performance results from several different angles. We are interested in knowing the answers to the following questions:

1. Which system(s) performed better than the others in a statistically significant way? Is there any difference between systems a) across folds; or, b) across mood clusters?
2. Are there significant performance differences between the folds or between the mood clusters?

[1] http://www.music-ir.org/mirex/2007/index.php/MIREX2007_Results

3. Does the division of ground-truth into judgment Set 1, 2 and 3 (Section 3.1) have an effect on performance?

	GT	CL	TL	ME1	ME2	IM2	KL1	IM1	KL2	Avg.
Fold 1	0.70	0.66	0.67	0.63	0.62	0.58	0.52	0.51	0.22	0.57
Fold 2	0.56	0.59	0.58	0.57	0.51	0.55	0.52	0.45	0.26	0.51
Fold 3	0.58	0.57	0.56	0.53	0.55	0.54	0.46	0.46	0.29	0.50
Avg.	0.61	0.61	0.60	0.58	0.56	0.56	0.50	0.47	0.26	0.53

Legend: GT: George Tzanetakis; CL: Cyril Laurier and Perfecto Herrera; TL: Thomas Lidy, Andreas Rauber, Antonio Pertusa and José Manuel Iñesta; ME1, ME2: Michael I. Mandel and Daniel P. W. Ellis; IM1, IM2: IMIRSEL M2K; KL1, KL2: Kyogu Lee

Table 5. System accuracies across folds

4.1. Comments on Statistical Testing

Exploratory data analyses of the result datasets consistently indicated that these data failed to meet the necessary assumptions of parametric statistical tests in that they exhibit non-normal distributions and non-equal variances. Therefore, we adopted the non-parametric Friedman's ANOVA test [1] as our general tool to determine if significant differences were present among the groups of interest (e.g., among systems, among mood clusters, among judgment sets, etc.). If and only if a Friedman's overall test showed the presence of a statistically significant difference (at $p < 0.05$) somewhere among the groups, would we then turn to the Tukey-Kramer Honestly Significantly Different (TK-HSD) analyses [1] to determine where the significant differences actually existed among the groups. The TK-HSD is used rather than the commonly mis-used multiple *t*-tests because TK-HSD properly controls for the experiment-wise Type I error rate whereas the naïve adoption of multiple *t*-tests does not. Failure to adjust for this experiment-wise inflation in Type I error in situations such as ours, where many pair-wise comparisons are made, all but guarantees that the null hypotheses of no differences in means between pairs of interest (i.e., H_0: μ (rank x) = μ (rank y) will be falsely rejected somewhere within the set of comparisons being made.

4.2. System Performance Comparison

To explore question 1 outlined above, we looked at the results from two different viewpoints corresponding to questions 1.a (fold-based) and 1.b (cluster-based). In test 1.a, we ran the Friedman's test using the accuracies shown in Table 5. In test 1.b we used accuracy scores in Table 6. Both tests proved significant (1.a: χ^2 (8, 16) = 21.42, $p < 0.01$; 1.b: χ^2 (8, 32) = 15.61; $p = 0.048$).

We then conducted the follow-up TK-HSD analysis on the two sets, to see exactly how the system performances differed. The results are displayed in Table 7 (for set 1.a) and Table 8 (for set 1.b). The shaded areas in the tables represent the groups of systems which did not show significant differences within each group.

	C1	C2	C3	C4	C5	All
GT	0.43 (7)	0.53 (1)	0.80 (3)	0.52 (4)	0.80 (1)	0.62 (1)
CL	0.46 (5)	0.50 (2)	0.83 (2)	0.53 (2)	0.71 (4)	0.61(2)
TL	0.53 (2)	0.49 (3)	0.75 (4)	0.53 (3)	0.69 (6)	0.60 (3)
ME	0.52 (3)	0.46 (4)	0.70 (5)	0.55 (1)	0.67 (7)	0.58 (4)
IM_2	0.51 (4)	0.45 (5)	0.68 (7)	0.45 (6)	0.70 (5)	0.56 (5)
ME2	0.55 (1)	0.43 (6)	0.65 (8)	0.51(5)	0.66 (8)	0.56 (5)
KL1	0.25 (8)	0.37 (7)	0.84 (1)	0.25 (8)	0.78 (3)	0.50 (7)
IM_1	0.45 (6)	0.22 (9)	0.70 (5)	0.19 (9)	0.80 (1)	0.47 (8)
KL2	0.15 (9)	0.25(8)	0.33 (9)	0.27 (7)	0.28 (9)	0.26 (9)

Table 6. Accuracies cross mood clusters. Numbers in parenthesis are ranks within each cluster.

	GT	CL	TL	ME1	ME2	IM2	KL1	IM1	KL2
Group 1									
Group 2									

Table 7. System groups: fold-based performances

	CL	GT	TL	ME1	ME2	IM2	KL1	IM1	KL2
Group 1									
Group 2									

Table 8. System groups: cluster-based performances

The fact that KL2 is not grouped together with GT and CL for set 1.a (Table 7) indicates the two system pairs (GT, KL2) and (CL, KL2) are significantly different in the fold-based analysis. It is noteworthy that among all systems, CL has the best ranks across all mood clusters despite its average accuracy being the second highest. As the TK-HSD is based on ranks rather than the raw scores, CL is part of the only pair with difference in set 1.b, the cluster-based analysis.

4.3. Effects of Folds and Mood Clusters

In order to see whether there were any significant differences among folds or mood clusters, we transposed the datasets used in test 1.a and 1.b and conducted Friedmans' tests on the transposed sets. Again, both tests were significant: test 2.a: χ^2 (2, 16) = 6.22, $p < 0.01$; test 2.b: χ^2 (4, 32) = 27.91; $p < 0.01$).

The follow-up TK-HSD analysis showed that Fold 1 and Fold 3 were significantly different. In general, using more folds would help alleviate the impact of one fold on the overall performances. In regard to the five clusters, there were two pairs of difference: (Cluster 3, Cluster 1) and (Cluster 5, Cluster 1). This is consistent with what we saw in human assessment analysis: Cluster 3 and 5 reached the best agreements among assessors and best performances among systems while Cluster 1 caused the most confusion both among assessors and systems.

4.4. System Performance and Human Agreement

In this section, we investigate whether the number of agreed judgments on the audio pieces affects system performances. We calculated accuracy scores on each of the three judgment sets described in Section 3.1 and present them in Table 9.

	GT	CL	TL	ME1	ME2	IM2	KL1	IM1	KL2	Avg.
Set 1	0.72	0.67	0.67	0.67	0.65	0.66	0.54	0.46	0.28	0.59
Set 2	0.41	0.43	0.37	0.4	0.39	0.47	0.39	0.36	0.19	0.38
Set 3	0.64	0.62	0.65	0.59	0.56	0.55	0.49	0.51	0.26	0.54

Table 9. System accuracies across three judgment sets

From the table, we can see that across all systems, accuracies on Set 1 are consistently the highest and those on Set 2 are always the lowest. A Friedman's test on the scores in Table 9 indicated the existence of significant difference ($\chi 2$ (2, 16) = 16.22, $p < 0.01$). The follow-up TK-HSD analysis showed two pairs of significant difference: (Set 1, Set 2) and (Set 3, Set 2). This means that the systems performed significantly better on Set 1 and Set 3 than on Set 2. Set 2 consisted of audio pieces involving disagreement among human judges while the other two sets contained only agreed judgments. This suggests pieces with discrepant human judgments impose greater challenges to the systems. In order to reduce such inherent ambiguity in the dataset, our recommendation for future evaluations is to exclude pieces with mixed labels when three judgments are collected for each piece.

Now we investigate the interactions between the 3 judgment sets and the 3 folds in cross-validation. The breakdown of the sets and folds is shown in Table 10. We can see that Fold 1 was dominated by Set 1, the set with best system performances, while Fold 3 contained none of the Set 1 pieces[1]. Also, Fold 1 included fewer pieces from the problematic Set 2 than Fold 3 did. These factors at least partially explain the observation that systems performed significantly better on Fold 1 than Fold 3. Again, this shows that the unanimity among human assessors on the dataset had an important influence on system performances, and suggests that one way to reduce variations among folds is to stratify audio pieces with different numbers of human agreements when splitting the dataset into training and testing sets.

	Set 1	Set 2	Set 3	Total
Fold 1	116	22	62	200
Fold 2	37	47	116	200
Fold 3	0	65	135	200
Total	153	134	313	600

Table 10. Number of audio pieces across judgment sets and cross-validation folds

[1] The folds were split according to the result order of MySql queries on the collected human assessments. The impact of the levels of human agreements was not recognized until after the evaluation.

5. DISCUSSION AND RECOMMENDATIONS

5.1. Techniques Used in Systems

Comparing the abstracts of the submissions available on the task result wiki page, we found that most of the systems adopted Support Vector Machines (SVM) as the classifier[2], but with different implementations, including the libsvm library [2], SMO algorithm [9] implemented in the WEKA machine learning software [12], and DAG-SVM for efficient n-way classification [10]. Table 11 shows the systems grouped by the classifiers they used, as well as the average accuracy within each group.

	Weka SMO	LibSVM	DAG-SVM	KNN
System(s)	IM2, TL	CL, GT	ME1, ME2	IM1
Avg. Acc.	0.58	0.61	0.57	0.47

Table 11. System groups w.r.t. classifiers (KNN signifies a K-nearest neighbor classification model)

The systems utilized a variety of audio and symbolic features, but spectral features were included in all systems. Among them, GT and ME2 were exclusively based on spectral features. ME1 added temporal features to ME2 and improved the overall accuracy by 2%. CL utilized fairly rich music descriptors including temporal, tonal, loudness and high level features (e.g., danceability) while TL added symbolic features capturing pitches and note durations. According to the analysis in Section 4, there were no significant differences among these systems. Thus, the (high-level) features other than the basic spectral ones did not show any advantage in this evaluation. One possible reason is that the training set is too small (for each fold, only 80 training clips exist in each mood cluster) to optimize models with a large feature space. In future AMC tasks, a larger training set is desirable for allowing possible improvements by using large feature spaces.

5.2. Evaluation Approaches

During the evaluation setup process, two evaluation approaches were proposed. One was to evaluate on a closed dataset with ground-truth labels, as adopted in AMC 2007. The advantages of this approach include rigorous evaluation metrics and the ability to conduct cross-validation. However, since labeling ground-truth is human labor intensive, both the sizes of training and testing sets are limited in this approach. The other proposed approach was to train systems on a labeled dataset, and subsequently test them on an unlabeled audio pool. After each system returns a ranked list of song candidates for each mood cluster, human assessors make judgments only on the top ranked candidates. This

[2] As abstracts of KL1 and KL2 were not published on the task result wiki page, we do not know what techniques were used in these systems.

approach adopts the pooling mechanism first proposed in TREC, the annual evaluation event in the domain of text-focused information retrieval [11], and has been used in the Audio Similarity and Retrieval task in both MIREX 2006 and 2007. The obvious advantage is that the size of the testing set can be arbitrarily large because the number of judgments required is independent of the size of the test set. However, one downside of this approach is that traditional evaluation metrics (e.g. precision, recall, F-measure, accuracy) can only be used after careful modifications. For example, one cannot measure the absolute "recall" metrics, but can only compare systems using "relative recall" scores calculated by assuming all unjudged samples are irrelevant [11].

The AMC 2007 participants chose the first approach by voting on it: 8 voted for the first approach while only 1 voted for the second exclusively. However, 4 voters indicated they would prefer to have their systems evaluated using both approaches. Considering the second approach's advantage of being able to test systems on a large scale dataset, we recommend adopting this approach in future AMC tasks and perhaps in other classification tasks as well.

5.3. Recommendations on Task Organization

Based on the analysis on the three judgment sets, we recommend that future AMC tasks exclude pieces where judges disagree. This would reduce ambiguity in the dataset and help measure the systems more accurately.

The methods of acquiring more ground-truth of high quality need to be validated and improved in the future. Formal studies are needed to guide choices such as whether to provide human assessors with pre-selected labels and exemplar songs.

During the debates on how the evaluation should be done, the polls conducted through the AMC wiki drew votes from potential participants and helped make critical decisions. We recommend this polling method as it can collect opinions directly from participants in a timely manner, and can clearly present the distribution of all votes. In fact, 6 out of the 11 tasks in MIREX 2007 used polls to aid decision making.

6. ACKNOWLEDGEMENTS

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grants NSF IIS-0340597 and NSF IIS-0327371.

7. REFERENCES

[1] Berenson, M. L., Goldstein, M. and Levine, D. *Intermediate Statistical Methods and Applications: A Computer Package Approach*. Prentice-Hall, 1983.

[2] Chang, C. and Lin, C. "LIBSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm 2001.

[3] Downie, J. S. "The music information retrieval evaluation exchange (MIREX)", *D-Lib Magazine* Vol. 12(12), 2006.

[4] Feng, Y., Zhuang, Y. and Pan, Y. "Popular music retrieval by detecting mood", *Proceedings of the. 26th annual international ACM SIGIR conference*, Toronto, Canada, 2003.

[5] Gruzd, A. A., Downie J. S., Jones, M. C. and Lee, J. H. "Evalutron 6000: collecting music relevance judgments", *Proceedings of the Joint Conference on Digital Libraries (JCDL)*, Vancouver, Canada, 2007.

[6] Hu, X. and Downie, J. S. "Exploring mood metadata: relationships with genre, artist and usage metadata", *Proceedings of the 8th International Conference on Music Information Retrieval*, *ISMIR'07,* Vienna, Austria, 2007.

[7] Lu, L., Liu, D. and Zhang, H. "Automatic mood detection and tracking of music audio signals", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol.14 (1), 2006.

[8] Mandel, M. Poliner, G. and Ellis, D. "Support vector machine active learning for music retrieval", *Multimedia Systems*, Vol.12 (1), 2006.

[9] Platt, J. "Machines using sequential minimal optimization", In B. Schoelkopf and C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998.

[10] Platt, J., Cristianini, N. and Shawe-Taylor, J. "Large margin DAGs for multiclass classification", In S.A. Solla, T.K. Leen, and K.-R. Mueller, editors, *Advances in Neural Information Processing Systems* Vol. 12, MIT Press, 2000.

[11] Voorhees, E. and D. Harmon. "The text retrieval conference," in E. Voorhees and D. Harmon, editors *TREC Experiment and Evaluation in Information Retrieva*l. MIT Press, 2005.

[12] Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.