USER-CENTERED MEASURES VS. SYSTEM EFFECTIVENESS IN FINDING SIMILAR SONGS

Xiao Hu
Faculty of Education
The University of Hong Kong
xiaoxhu@hku.hk

Noriko Kando
National Institute of Informatics
Japan
kando@nii.ac.jp

ABSTRACT

User evaluation in the domain of Music Information Retrieval (MIR) has been very scarce, while algorithms and systems in MIR have been improving rapidly. With the maturity of system-centered evaluation in MIR, time is ripe for MIR evaluation to involve users. In this study, we compare user-centered measures to a system effectiveness measure on the task of retrieving similar songs. To collect user-centered measures, we conducted a user experiment with 50 participants using a set of music retrieval systems that have been evaluated by a system-centered approach in the Music Information Retrieval Evaluation eXchange (MIREX). The results reveal weak correlation between user-centered measures and system effectiveness. It is also found that user-centered measures can disclose difference between systems when there was no difference on system-effectiveness.

1. INTRODUCTION

With the rapid growth of digital music, research on Music Information Retrieval (MIR) has been flourishing in recent years. Many algorithms and systems have been developed to facilitate searching and retrieving music pieces automatically. As a crucial aspect of system development, evaluation of MIR systems has attracted continuous attention among researchers. However, so far, MIR evaluation has been dominated by system-oriented approaches, while users, whom MIR systems would ultimately serve, have rarely been considered in MIR evaluation.

The system-centered evaluation approach, also known as the Cranfield evaluation [3], has been adopted by the Music Information Retrieval Evaluation eXchange (MIREX), a community-based annual evaluation campaign for MIR. Since its inception in 2005, MIREX has evaluated and compared more than a thousand systems on various MIR tasks such as genre classification, artist identification, query-by-humming, etc. [5]. MIREX not only greatly enhances the development of MIR, but also provides rich evaluation data on system effectiveness.

Despite its long tradition and popularity, system-centered approach has been criticized for excluding users from the evaluation process. Researchers argue that the goal of MIR systems is to facilitate users' music information tasks, and thus the evaluation of MIR should inevitably take users into consideration [8]. Furthermore, since music appreciation is more or less a subjective process, users' perceptions about whether a MIR system is useful might be different from a system-centered point of view. However, there have been no formal studies investigating whether there are correlations between user-centered measures and system effectiveness measures in the MIR domain. Thus, people remain puzzled when they see the precision and recall numbers of certain systems. Would they be helpful to users? Would users be satisfied with them? This study aims to fill the research gap and answer the following research question: *to what extent is system effectiveness related to user-centered measures?*

In particular, this study focuses on one MIR task, audio music similarity and retrieval where systems search for songs similar to a given query song. There are two major reasons for choosing this task. First, finding similar songs, as a query-by-example scenario, is a prevailing music information need. For instance, many people search for similar songs to build playlists [4]. Second, the MIREX has the same task and thus provides system-centered measures needed in this study.

2. RELATED WORK

2.1 User Evaluation in MIR

There have been very few studies on formal user evaluation MIR systems. The Philips Research Laboratories in the Netherlands is the leader on this topic. During 2002 to 2005, they conducted a series of controlled user experiments to evaluate their playlists generation systems [9][10][14]. The general approach was to recruit 22 to 24 participants to use a novel system they developed as well as one or two control systems for the task of generating playlists in some pre-defined music listening situations (e.g., "soft music" and "lively music"). The experiments may consist of one to four sessions. The researchers then compared the novel system to control systems using user-centered measures including users' ratings on playlist quality, time spent on the task, number of button presses in accomplishing the task, as well as perceived usefulness, ease-of-use and preference reported by the users.

A more recent study by Hoashi and colleagues [7] compared the effectiveness of three visualization methods for a content-based MIR system. Besides user effective-

ness and user satisfaction, they also employed self-reported usability measures on perceived system accuracy, explicitness and enjoyability. It is particularly valuable that the authors also advocated for the user-centered approach as necessary to evaluate MIR systems.

2.2 User and System Effectiveness in Text Information Retrieval

Evaluating retrieval systems from the users' perspective has been active in the domain of text Information Retrieval (IR). Studies have been done to examine the relationship between user-centered measures and system-effectiveness. Hersh et al. investigated this question in the tasks of instance recall [6] and question answering [12]. They conducted user experiments and found user effectiveness and system effectiveness did not yield the same conclusion. More recently, Turpin and Scholer [13] evaluated systems with large differences in system effectiveness and again found no significant relationship between user and system effectiveness for precision-based tasks and a weak relationship for recall-based tasks.

In contrast, there were also studies finding significant correlations. For instance, Allen et al. [1] studied the task of text passage retrieval and found user effectiveness (as measured by task completion time and number of relevant passages) was correlated with system effectiveness when the latter was either low or high, but not in the middle. Last but not least, Al-Maskari et al. [2] controlled the variance of system effectiveness and reported significant correlations between multiple user-centered measures and system effectiveness.

As a first study investigating the relationship between user-centered measures and system effectiveness in the MIR domain, this study is inspired by the aforementioned previous work in text IR. Many of these studies used TREC (Text Retrieval Conference) evaluation results to select systems to be evaluated by users and to obtain data on system effectiveness. In this study, we resort to MIREX, the counterpart of TREC in the MIR domain, for obtaining system effectiveness measures and the underlying MIR systems.

3. METHOD AND RESEARCH DESIGN

3.1 The AMS Task in MIREX

The MIREX has included the Audio Music Similarity and Retrieval (AMS) task every year except for 2008. In this task, systems are given a number of queries (i.e. audio song clips) and a large collection of music audio clips. The goal of the systems is to retrieve clips from the collection that sound similar to the queries. In 2010, the AMS task had 100 queries sampled from 10 different genres, and the candidate collection contained 7,000 music clips also evenly sampled from the 10 genres. There were eight systems evaluated in this task[1] and all of them were considered in this study except for one system (RZ1) which was a random baseline and performed poorly. For each query, the top five song clips retrieved by each system were collected for similarity judgment by human experts. This is much like the pooling method for relevance judgment in TREC [15]. However, unlike TREC, the pooled candidates were deliberately randomized when presented to the human judges so as to eliminate any cues given by the order of candidates. In the 2010 cycle of the AMS task, each query candidate pair was judged by one assessor. Based on the similarity judgments on the pooled candidates, system effectiveness measures were calculated for evaluating and comparing the systems.

This study is built upon the 2010 cycle of the MIREX AMS task. For each query, two systems were selected and user-centered measures on both systems were collected in formal user experiments. Then the research question is answered by comparing the user-centered measures to system effectiveness.

3.2 The Systems

In selecting systems, we adopted the approach proposed by Al-Maskari et al. [2]: different systems are selected for different queries so that the differences of system effectiveness between systems can be better controlled. As there are no previous studies of this kind in the MIR domain, we also followed [2] in using average precision (AP) as the system effectiveness measure. AP is calculated as the mean of precisions at the point of each relevant document in the ranked sequence. This measure rewards relevant documents retrieved at high ranks. In MIREX AMS task, human judges evaluated each candidate using ternary relevance: very similar, somewhat similar, and not similar. In calculating AP scores in this study, we convert the judgments into binary relevance by combining "very similar" and "somewhat similar" into "similar". In the future we will evaluate system effectiveness measures based on ternary relevance.

The AP scores of the seven participating systems vary across queries. For 79 out of the 100 queries, the differences of AP values among systems are from 0.01 to 0.6. For the rest of 21 queries, the systems had exactly the same AP. Unlike [2] where the best and worst performing systems were chosen for each query, we choose the best performing system (denoted as the "superior" system in this paper) and the second best system (denoted as the "inferior" system in this paper) for each query. This is because the systems tend to have lower AP than those in [2] (99% are lower than 0.5 in this study), and the difference between the best and worst performing systems can be very obvious, making it a trivial task to decide system preference. In addition, the worst performing systems sometimes are so bad that our pilot testers felt it was boring to listen to songs very dissimilar to the queries. Finally, using systems with different but close AP scores makes it possible to investigate whether user-centered measures can differentiate system quality when system effectiveness had little difference. For the 21 queries without system difference, two systems were randomly

[1] The MIREX 2010 AMS task results: http://www.music-ir.org/mirex/wiki/2010:Audio_Music_Similarity_and_Retrieval_Results

selected. Figure 1 shows the AP scores of the two selected systems across queries where the queries are ordered by the difference of AP scores between the two systems.

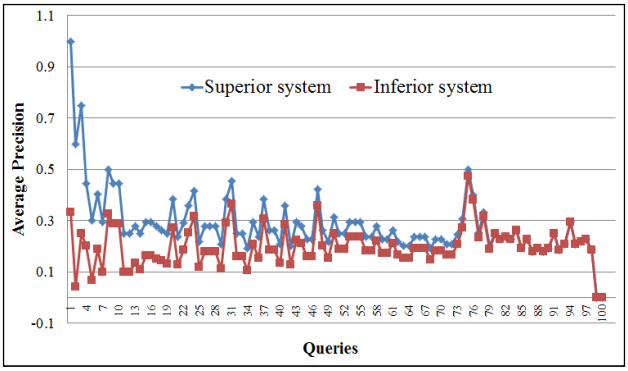

Figure 1. System AP scores across queries.

3.3 Participants

50 Japanese undergraduate and graduate students from 13 different universities were recruited, including 24 females and 26 males. Their average age was 21.7 years old (standard deviation was 4.30, range from 18 to 50). Their majors ranged from engineering, medicine to social sciences and humanities. Statistics of participants' background on music knowledge, computer and English skills, as well as familiarity with the genres of the songs are shown in Table 1. Self-reported English abilities were collected because some of the songs had English lyrics and the pre- and post- experiment questionnaires were written in English. As the songs were associated with American genre classification system, the participants' familiarity levels with the genres were collected.

	Median	Max.	Min.
Music knowledge*	4	6	2
Expertise with computers*	4	6	3
Expertise with online searching*	5	6	2
Ability in reading English*	5	7	3
Ability in listening to English*	4	7	2
Familiarity with the genres ♮	3	5	0

*: in a Likert scale from 1 to 7. 1: novice, 7: expert; ♮: in a Likert scale from 1 to 5. 1: very unfamiliar, 5: very familiar

Table 1. Statistics of participants' background.

3.4 Tasks

All of the 100 queries in the MIREX AMS tasks are included in this study. The queries are evenly distributed into ten different genres, namely Baroque, Romantic, Classical, Country, Jazz, Blues, RocknRoll, Rap/HipHop, Metal, and Edance. Each participant was assigned ten queries with one in each genre. Since there are 50 participants, each query was evaluated by five participants. The orders of the ten query genres were distributed to participants using a Latin Square design so as to reduce the effect of genre order on results.

For each query, a participant evaluated the list of candidate songs retrieved by the two selected systems. Specifically, a participant needed to play and listen to the query song and indicate his/her familiarity level with it, as well as his/her personal preference on the query. Then he or she proceeded to listen to each of the five songs returned by one system and indicate whether it sounded similar to the query. Just like in MIREX, we used ternary similarity scale: participants needed to indicate whether a candidate was very similar, somewhat similar or not similar to the query. In this experiment, the candidate songs were presented to users in the original ranked order as retrieved by the systems. This setting mirrors a real life retrieval system where a higher ranked item is expected to be more relevant. In contrast, MIREX human judgments had no information on the rank of the candidates nor did they know which candidates were retrieved by the same system.

After evaluating all the five songs returned by one system, the participant was asked to indicate his/her satisfaction level towards this system and the perceived task easiness in a Likert scale. Then the participant proceeded to the other system. The relative difference of the systems was not revealed to the user and the order of two systems was randomized. After listening to candidates from both systems, the participant was asked to indicate his or her preference between the systems and how easy it was to compare the two systems. The audio of each song (either query or candidate) was 30 second long, but it could be paused and/or replayed at any time. A participant could also change answers when working with one system. However, once proceeding to the other system or the next query, a participant could not go back to change answers. This is to prevent influences from other systems on users' judgments. A screenshot of the working interface is shown in Figure 2.

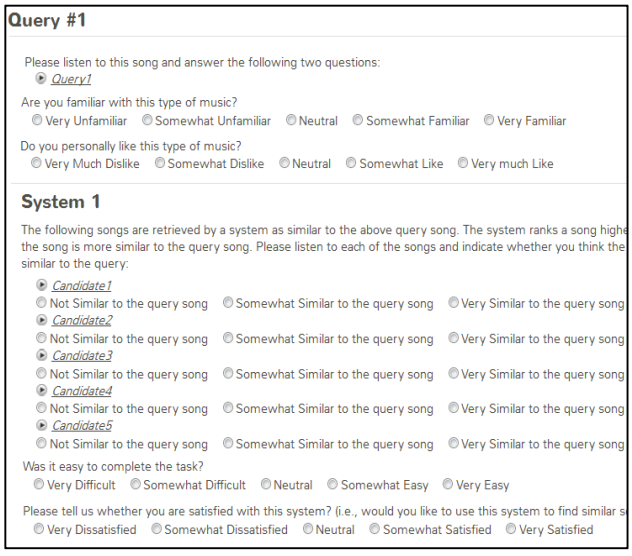

Figure 2. Screenshot of the evaluation interface.

3.5 Procedure

The experiment was conducted in a batch manner, with 5 to 7 subjects in each batch performing the tasks at the same time. Before the experiment started, each subject read and signed a consent form. After that, she or he filled an online pre-experiment questionnaire with regard to demographic information, music background and

search experience. Then, the experiment facilitator introduced the experiment system and the experiment procedure in Japanese. The training sessions lasted about 10 minutes.

According to our pre-tests of the procedure, the participants were given 55 minutes to finish all the 10 assigned queries. Most participants finished the process within 45 minutes. During the process, the experiment system recorded users' interactions including play and pause queries and candidates, answers to each question as well as changes of answers. After all queries were finished, each subject filled an online post-experiment questionnaire which asked for his or her general impression on the evaluated music retrieval systems and the experiment in general. The entire procedure lasted about 1.5 hours and each participant was paid 2000 yen for their participation.

3.6 User-centered Measures

The following user-centered measures were collected and compared to system effectiveness.

User effectiveness:

1) Number of similar songs found using each system. A candidate is "similar" to a query if the user chooses "very similar" or "somewhat similar" option.
2) Task completion time: time spent on making judgments on all candidates of one system.
3) Time spent in finding the first similar song using each system. If there is no similar song found among the five candidates, the time is assumed to be 3 minutes which is the time needed for listening to 6 candidates in full length.
4) Rank of the first similar song using each system. If there is no similar song found among the five candidates, the rank is assumed to be 6.

User perception:

1) Task easiness in evaluating results of each system.
2) User satisfaction with each system.
3) Easiness in comparing two systems.

Each of these measures was on a Likert scale from 1 to 5, with 1 indicating very difficult/very dissatisfied and 5 indicating very easy/very satisfied.

User preference:

1) The system a user prefers: the superior one, inferior one or neither.

3.7 Hypotheses

To answer our research question, we compared AP scores of the two systems to the aforementioned user-centered measures by testing a series of hypotheses:

H1: When the systems' AP scores were different, users were more effective and more satisfied with the superior systems than the inferior systems;

H2: When the systems' AP scores were the same, users were similarly effective and satisfied with both systems;

H3: When the systems' AP scores were different, users preferred the superior systems to the inferior systems;

H4: When the systems' AP scores were the same, users did not have a preference between the systems;

H5: User perceived higher easiness level when comparing systems with AP score difference than comparing those without AP score difference.

To test the correlation between user-centered measures and AP score, we examined the following hypotheses:

H6: User-centered measures are highly correlated with AP score;

H7: When the difference of systems' AP scores gets larger, users would tend to prefer the superior systems and feel it is easier to compare the two systems.

4. RESULS AND DISCUSSIONS

4.1 User Effectiveness and Satisfaction

Table 2 presents means and standard deviations (in parenthesis) of AP scores and user effectiveness and perception measures for the superior and inferior systems on the 79 queries where the two systems had different AP scores. In order to test *H1*, we employed the non-parametric Wilcoxon signed rank sum test because studies have shown that system performance data rarely comply with normal distribution [5] and the Wilcoxon test does not assume normal distribution of tested variables.

Measure	Superior	Inferior	*p* value
average precision	0.30 (0.13)	0.20(0.08)	< 0.001*
number of similar songs	3.53 (0.99)	3.00 (1.07)	< 0.001*
task completion time (seconds)	76.75 (21.99)	77.09 (25.44)	0.688
time finding first similar song (seconds)	19.91 (14.42)	28.55 (26.74)	0.042*
rank of first similar song	1.48 (0.63)	1.72 (0.99)	0.047*
task easiness	3.47 (0.55)	3.48 (0.53)	0.714
user satisfaction	3.44 (0.73)	3.04 (0.77)	< 0.001*

N=79. *: significant at $p < 0.05$ level

Table 2. Measures for queries with different AP scores.

As shown in Table 2, the difference between the AP scores of the superior and inferior systems was significant across the 79 queries. The user-centered measures indicate that, using the superior systems, users found more similar songs, spent less time in finding the first similar song which had a higher rank, and were more satisfied than using the inferior systems. However, there was little difference on the time used to judge all the five candidates of each system. In addition, users perceived the tasks were about the same easiness level when using both systems.

Therefore, hypothesis *H1* is partially supported by four out of six user-centered measures under consideration. In other words, when the AP scores were significantly different, some user-centered measures could also differentiate the systems. The little differences on task completion time and perceived task easiness are related to each other. If a task is difficult, it will likely take more time. The insignificant result indicates systems with

higher AP scores did not make the task of music similarity judgment easier.

Table 3 presents means and standard deviations (in parenthesis) of the aforementioned measures and Wilcoxon test results for the two systems on the 21 queries where the two systems had the same AP scores. As can be seen from Table 3, *H2* is also partially supported by four out of six user-centered measures. That is, the two systems had no significant difference on number of similar songs found, task completion time, rank of first similar songs and perceived task easiness. However, the two systems were significantly different in terms of time spent finding the first similar songs and users' satisfaction towards systems, even though the AP scores of the two systems were exactly the same. This evidences that some user-centered measures can tell the differences between systems that system-effectiveness cannot. In particular, the difference on user satisfaction on systems with the same AP scores is remarkable since user satisfaction has been called by IR researchers as a main criterion of IR system evaluation (e.g., [11]).

Measure	System 1	System 2	*p* value
average precision	0.20 (0.07)	0.20 (0.07)	-
number of similar songs	3.59 (0.95)	3.46 (1.01)	0.470
task completion time (seconds)	78.61 (21.19)	78.83 (24.49)	0.776
time finding first similar song (seconds)	14.37(10.83)	24.50 (18.27)	0.009*
rank of first similar song	1.27 (0.47)	1.50 (0.63)	0.197
task easiness	3.22 (0.43)	3.12 (0.54)	0.616
user satisfaction	3.37 (0.64)	3.09 (0.61)	0.032*

N=21. *: significant at $p < 0.05$ level

Table 3. Measures for queries with same AP scores.

4.2 User Preference

Statistics on user preferences between the systems are shown in Table 4. It is interesting to see that users perceived no difference between the two systems 25% of the time while the AP scores of the systems were different. In contrast, for queries where the two systems had exactly the same AP scores, 80% of users thought the systems were different. A Wilcoxon test was conducted on each set of the queries to see if the differences on number of votes of the two systems were significant. The results support both *H3* and *H4*: users preferred the superior systems when the systems' AP scores differed ($p < 0.001$) while users did not have significant preferences when the systems had the same AP scores ($p = 0.06$). However, the low *p* value and the low percentage of "no preference" votes on queries with the same AP scores (20%) indeed suggest that the difference on AP scores may not be a good indicator of system preference.

4.3 Perceived Easiness in Comparing Systems

The test results on hypothesis *H5*, perceived easiness level in comparing the two systems are shown in Table 5. The average easiness score is 3.06 (easier) across the 79 queries with AP difference and 2.81 (harder) across the 21 queries without AP difference. As the two sample sizes are not equal, a two sample unequal variance *t*-test was employed to test the significance of the difference on easiness level. The test result is significant and thus hypothesis *H5* is supported: users perceived it was easier to compare the two systems when there were AP differences between the systems. So far, the results of the analysis generally support our hypotheses *H1* to *H5*. That is, user-centered measures tend to agree with system-effectiveness (as measured by AP score). However, the exceptions in *H1* and *H2* are also noteworthy. In the next subsection, we continue to investigate the correlation between user-centered measures and AP scores.

79 queries with different AP scores	**Superior**	**Inferior**	**No pref.**	**Total**
Number of pref. votes	190	105	100	395
Percentage of preference	48.10%	26.58%	25.32%	100%
21 queries with same AP scores	**System 1**	**System 2**	**No pref.**	**Total**
Number of pref. votes	48	36	21	105
Percentage of preference	45.71%	34.29%	20.00%	100%

Table 4. Votes of system preferences.

	With AP difference	**Without AP difference**	***p* value**
difficulty level	3.06 (1.26)	2.81(1.29)	0.002*
sample size	395	105	

*: significant at $p < 0.05$ level

Table 5. Perceived difficulty in comparing systems

4.4 Correlation between User-centered Measures and System Effectiveness

To test Hypothesis *H6*, Pearson's correlation coefficients were calculated for measures on interval scales: number of similar songs, task completion time and time of finding the first similar songs. For measures on ordinal scales such as rank of first similar songs, task easiness and user satisfaction, Spearman's rank correlation coefficient was calculated. The results are shown in Table 6.

Measure	Coefficient	*p* value
number of similar songs	0.111 (Pearson)	0.059
task completion time	-0.069 (Pearson)	0.177
time finding first similar song	-0.142 (Pearson)	0.022*
rank of first similar song	-0.163 (Spearman)	< 0.021*
task easiness	0.057 (Spearman)	0.423
user satisfaction	0.246 (Spearman)	< 0.001*

N = 200. *: significant at p < 0.05 level

Table 6. Correlation between user-centered measures and AP score.

Number of similar songs, task completion time and task easiness have no significant correlation with AP score while the correlation between AP score and other user-centered measures are fairly weak despite being significant. Our hypothesis *H6* is not supported. The fact that there is no significant relationship between perceived

task easiness and AP score confirms an earlier finding in Section 4.1 that higher AP scores did not make the task of music similarity judgment easier.

Table 7 shows Spearman's correlation coefficients between the AP scores difference of the two systems and the two user-centered measures related to system comparison: system preference and easiness in system comparison. System preference is encoded as an ordinal variable with values 1, 0, and -1 indicating preferring the superior system, no preference, and preferring the inferior system, respectively. From Table 7 we can see that hypothesis *H7* is not supported: the correlations are either insignificant or fairly weak. The insignificance between system preference and AP score difference helps explain an earlier observation that 80% of the users indicated system preference when there was no difference on the AP scores.

Measure	Correlation with AP difference	*p* value
system preference	0.080 (Spearman)	<0.053
easiness in system comparison	0.174 (Spearman)	<0.001*

N=100. *: significant at $p < 0.05$ level.

Table 7. Correlation between user-centered measures and AP score difference

5. CONCLUSIONS AND FUTURE WORK

This paper presents a user experiment on evaluating results of music similarity retrieval systems in the AMS task in MIREX 2010, with the goal of comparing a well-accepted system effectiveness measure to user-centered measures. Such comparison has rarely been explored in the MIR domain. The results revealed none or weak correlations between system effectiveness and eight user-centered measures. In particular, significant differences on two user-centered measures, including user satisfaction, were found between systems with the same system effectiveness. As a first study on user-centered vs. system-centered measures in MIR, this research prompts many interesting observations for future research. More user behavior measures can be examined such as number of times a query song was played, number of changes a user made to his or her answers, as well as measures based on ternary relevance judgment. In addition, similar evaluations could be done for other MIR tasks such as genre and mood classification in the future.

6. ACKNOWLEGEMENTS

The research is partially supported by the JSPS Grand-in-Aid (#21300096) and the Non-MOU Grant funded by the National Institute of Informatics in Japan, and was conducted when the first author was in the University of Denver. We also thank the IMIRSEL in the University of Illinois for providing the MIREX AMS data.

7. REFERENCES

[1] J. Allan, B. Carterette and J. Lewis: "When will information retrieval be 'good enough'? User effectiveness as a function of retrieval accuracy," *Proceedings of the ACM SIGIR Conference on Information Retrieval*, pp. 433-440, 2005.

[2] A. Al-Maskari, M. Sanderson, P. Clough and E. Airio: "The good and the bad system: Does the test collection predict users' effectiveness?"*Proc. of the ACM SIGIR Conference*, pp. 59-66, 2005.

[3] C. W. Cleverdon and E. M. Keen: "Factors determining the performance of indexing systems," Vol. 1: Design. Vol 2: Results. Cranfield, U.K: *Aslib Cranfield Research Project*, 1966.

[4] S. J. Cunningham, D. Bainbridge and A. Falconer: "'More of an art than a science': supporting the creation of playlists and mixes," *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, 2006.

[5] J. S. Downie: "The Music Information Retrieval Evaluation eXchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, 29(4), pp. 247–255. 2008.

[6] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kraemer, L. Sacherek, and D. Olson: "Do batch and user evaluations give the same results?" *Proc. of the ACM SIGIR Conference*, pp. 17–24, 2000.

[7] K., Hoashi, S. Hamawaki, H. Ishizaki, Y. Takishima, and J. Katto: "Usability evaluation of visualization interfaces for content-based music retrieval systems," *Proc. of ISMIR*, 2009.

[8] X. Hu, and J. Liu: "Evaluation of music information retrieval: towards a user-centered approach," *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR)*, 2010.

[9] S. Pauws, and B. Eggen: "PATS: Realization and evaluation of an automatic playlist generator," *Proc. of ISMIR*, 2002.

[10] S. Pauws, and S. van de Wijdeven: "User evaluation of a new interactive playlist generation concept," *Proc. of ISMIR*, 2005.

[11] K. Spärck Jones: *Information Retrieval Experiment*, London, Butterworths. 1981.

[12] A. Turpin and W. Hersh: "Why batch and user evaluations do not give the same results," *Proc. of the ACM SIGIR Conference*, pp. 225-231, 2001.

[13] A. Turpin and F. Scholer: "User performance versus precision measures for simple search tasks," *Proc. of the ACM SIGIR Conference*, pp.11-18, 2006.

[14] F. Vignoli and S. Pauws: "A music retrieval system based on user driven similarity and its evaluation," *Proc. of ISMIR*, 2005.

[15] E. M. Voorhees and D. K. Harman: *TREC: Experiments in Information Retrieval Evaluation*, MIT Press, 2005.