INFORMATION RETRIEVAL META-EVALUATION: CHALLENGES AND OPPORTUNITIES IN THE MUSIC DOMAIN

Julián Urbano
University Carlos III of Madrid
Department of Computer Science
jurbano@inf.uc3m.es

ABSTRACT

The Music Information Retrieval field has acknowledged the need for rigorous scientific evaluations for some time now. Several efforts were set out to develop and provide the necessary infrastructure, technology and methodologies to carry out these evaluations, out of which the annual Music Information Retrieval Evaluation eXchange emerged. The community as a whole has enormously gained from this evaluation forum, but very little attention has been paid to reliability and correctness issues. From the standpoint of the analysis of experimental validity, this paper presents a survey of past meta-evaluation work in the context of Text Information Retrieval, arguing that the music community still needs to address various issues concerning the evaluation of music systems and the IR cycle, pointing out directions for further research and proposals in this line.

1. INTRODUCTION

Information Retrieval (IR) is a highly experimental discipline, and IR Evaluation (IRE) experiments are the main research tool to scientifically compare IR systems and algorithms to advance the state of the art through careful examination and interpretation of their results. IRE has been used and studied in Text IR for over 50 years now, since the Cranfield 2 experiments [18], with successful evaluation forums such as TREC, CLEF, NTCIR or INEX. Until 2006, these evaluations were not usual at all in Music IR (MIR), although there was general concern about specific needs and resources for a fruitful beginning of evaluation campaigns in the Music domain.

The "ISMIR 2001 resolution on the need to create standardized MIR test collections, tasks, and evaluation metrics for MIR research and development" was drafted and signed by many members of the community as a demonstration of the general concern [20]. A series of three workshops then followed between July 2002 and August 2003, were researches begun this long-needed work for evaluation in Music IR [20]. There was some general agreement that evaluation frameworks for Music IR would need to follow the steps of the Text REtrieval Conference (TREC) [53][56], although it was clear too that special care was to be taken not to oversimplify the TREC evaluation model [19], because Music IR differs greatly from Text IR in many aspects that affect evaluations [21]. The general outcome of these workshops, and many other meetings, was the realization by the Music IR community that these evaluations were clearly necessary, and that a lot of effort and commitment was needed to establish a periodic evaluation forum for Music IR systems. Finally, in 2005 the first edition of the Music Information Retrieval Evaluation eXchange (MIREX) took place, and ever since it has evaluated over a thousand Music IR systems for many different tasks on a yearly basis [23].

The impact of MIREX has been without doubt beneficial for the Music IR community, not only for fostering these experiments, but also for studying and establishing specific evaluation frameworks for the Music domain. But now that it is widely accepted, it seems that the community has settled down in the belief that we finally have what we wanted. It is our belief though, that while we are on the right path, there is still a lot of work to do in Music IR Evaluation. These experiments are anything but easy and straightforward [54][26], so much that a whole area therein is concerned with their reliability and correctness: Information Retrieval Meta-Evaluation. The Text IR literature has been flooded with meta-evaluation studies for the past two decades, showing year after year that IRE has its very own issues and proposing different approaches and techniques to cope with them. While the MIR community has inherited good evaluation practices by adopting TREC-like frameworks, some are already outdated, and others still lack appropriate analyses. We agree that not everything from the Text IR community applies to Music IR, but a lot of meta-evaluation studies do. In fact, since the inception of MIREX in 2005 several landmark studies have taken place in the context of TREC, specially focused on large-scale evaluation, robustness and reliability, none of which has even been considered for Music IR.

In this paper we approach meta-evaluation from the point of view of the analysis of experimental validity of IR Evaluation experiments. We show different aspects of IRE affected by these validity considerations, and survey the Text IR literature outlining how these problems are dealt with in evaluation forums such as TREC. Finally, we show the current shortcomings in MIR evaluation and propose lines for further work, as a starting point for what we hope begins a tradition of periodic Music IR Evaluation studies.

2. IR EVALUATION

IR evaluation experiments follow the traditional Cranfield paradigm conceived by Cleverdon in the late 50's [18]. The main element needed for these evaluations is a test collection, which is made up of three basic pieces: a document collection, a set of information needs and the relevance judgments telling what documents are relevant to these information needs (the ground truth or gold standard). These test collections are built in the context of a particular task defining the intent of the information needs, and several measures are used to rank the systems following different criteria, always from the point of view of a user model with assumptions and restrictions as to the potential real users of the systems being evaluated.

Although some variations exist, a typical IRE experiment goes as follows [54][26]. First, the task is identified and well-defined, normally seeking the agreement between several researchers. Depending on the task, a document collection is either put together or reused from another task, and a set of information needs is selected, often given as direct input queries. The systems to evaluate return their results for the particular query set and document collection, and these results are evaluated using several measures that attempt to assess how well the systems would have satisfied a real user. This assessment employs the relevance judgments in the ground truth, made before or after running the systems, depending on the task and other factors.

3. IR META-EVALUATION

Experimental validity establishes how well an experiment meets the well-grounded requirements of the scientific method [30][35][36]. That is, whether the results obtained do fairly and actually assess what the experimenter attempted to measure. Validity of experiments is usually assessed from different points of view, depending on what aspects of the scientific method are at stake.

	Task	User model	Documents	Queries	Ground truth	Systems	Measures
Construct	x	x					x
Content	x	x	x	x		x	
Convergent		x			x		x
Criterion				x	x		x
Internal			x	x	x	x	x
External			x	x	x	x	
Conclusion		x		x	x	x	x

Table 1. The effect of Experimental Validity on Information Retrieval Evaluation experiments.

Information Retrieval Evaluation experiments, as scientific experiments themselves, are also subject to validity analysis. Meta-evaluation can be viewed as the analysis of this experimental validity, highlighting that the evaluation is itself being evaluated. Next, we discuss several types of experimental validity and show how they affect IR evaluation experiments (see Table 1).

3.1 Construct Validity

Construct validity evaluates the extent to which the variables of an experiment correspond to the theoretical meaning of the concept they purport to measure. For example, an experiment to assess the quality of the results given by a Web search engine would not have construct validity if quality were measured as the number of visits to the site, because this actually measures its popularity. Thus, an experiment acquires construct validity by thorough selection and justification of the variables used.

In the case of IRE, construct validity is concerned mainly with the evaluation measures and the user model considered for the particular task [16]. For instance, in a traditional ad hoc retrieval task, binary set-based measures such as Precision and Recall do not resemble a real user who wants not only relevant documents, but highly relevant ones at the top of the results list [42]. Instead, rank-based measures such as Average Precision, graded relevance judgments [52][31], or the combination [29], are more appropriate.

3.2 Content Validity

Content validity evaluates the extent to which the experimental units reflect and represent the elements of the domain under study. For example, an experiment measuring the reading comprehension of students would not have content validity if only science-fiction stories were employed. Thus, an experiment acquires content validity by careful selection of the experimental units included.

In IR evaluation, it is imperative that the task resembles as closely as possible the real-world settings it represents, and that the systems evaluated fulfill as much as possible the needs of the real users. However, evaluating under such conditions would introduce a heavy user component very difficult to manage and control, so a more system-oriented approach is usually followed [54][18]. As such, the actual value of the systems in real settings is many times overlooked [34], and sometimes it can be questioned [45].

Likewise, the documents in the collection must resemble as closely as possible the documents that would be found in a real-world setting of the task, and have a sufficiently large sample so as to be representative of the domain. Also, the particular queries used should be carefully selected to represent a diverse and wide range of possible use cases, while being reasonable for the document collection in use [54][12]. Moreover, some queries are more helpful than others to differentiate between systems [25][38].

3.3 Convergent Validity

Convergent validity evaluates the extent to which the results of an experiment agree with other results, theoretical or experimental, they should be related with. For example, the results of a study measuring the mathematical skills of students should be correlated with other studies on abstract thinking. Thus, an experiment acquires convergent validity by careful examination and confirmation of the relationship between its results and others supposedly related.

Ground truth data is a much debated part of IR evaluation because of the subjectivity in the very concept of relevance. Several studies show that documents are judged differently by different people in terms of their relevance to some specific information need, even by the same people over time. As such, the validity of IRE experiments can be questioned because different results are obtained depending on the people that make the relevance judgments. Several studies have shown that absolute figures do indeed change, but the relative differences between systems stand still for the most part [51]. For very large-scale experiments though, these differences can have a large impact on the results [13].

Effectiveness measures are usually categorized as precision- or recall-oriented. Therefore, it is expected for precision-oriented measures to yield effectiveness scores correlated with other precision-oriented measures, and likewise with recall-oriented ones. However, this does not always happen [39][31], and some measures are even better correlated with others than with themselves [57], evidencing predictability problems. In general, all these measures should be correlated with user satisfaction in the particular task [42], so alternatives such as rank-based measures, different forms of ground truth data [4] or relevance discount functions [29] are usually considered.

3.4 Criterion Validity

Criterion validity evaluates the extent to which the results of an experiment are correlated with those of other experiments already known to be valid. For example, a study to evaluate if a new product would have as good sales as an old one would lack criterion validity if subjects were just asked whether they like the new one, instead of whether they like it even more: the context changed in the second case. Thus, an experiment acquires criterion validity by careful examination and confirmation of the correlation between its results and others previously established.

As real-world systems need to manage more and more amounts of information, modern IR evaluation studies have focused on practical large-scale methodologies, mainly through a technique called pooling [8]. This permits the use of large collections while requiring somewhat reasonable effort in relevance judging by assuming that documents not retrieved by any system are indeed not relevant. More recent studies analyze the use of non-experts for relevance judging [3], crowdsourcing platforms such as Amazon Mechanical Turk [1][17], requiring fewer judgments to give an estimate of the absolute effectiveness scores of the systems [59][60], selecting what judgments better tell the difference between systems [10][11], or even using no relevance judgments at all [44]. All these improvements allow for an increase on content validity as the effort per query diminishes. The results of all these methodologies are usually compared with the results of traditional ones, in terms of criterion validity, to see whether they are really viable or not. That is, whether the results they produce not only require less effort, but also agree with those of previous, accepted methodologies.

3.5 Internal Validity

Internal validity evaluates the extent to which the conclusions of an experiment can be rigorously drawn from the experimental design followed, and not from other factors unaccounted for. For example, a study on the usability of two word processors would not have internal validity if the subjects were already familiar with one of the products. Thus, an experiment acquires internal validity by careful identification and control of possible confounding variables and selection of experimental designs.

In IR evaluation, observed differences between systems could be the result of the particular people that do the relevance judgments, as their personal notion of relevance could be more beneficial for some systems than for others [13], let alone if the ground truth data has inconsistencies. Likewise, if a pooling method were used, systems more alike would reinforce each other, while a system with a novel technology would not be able to contribute that much to the pool: it is more likely for the former systems to have more of their documents included in the pool than for the latter [62]. In general, the non-relevancy assumption affects both the measures [40] and the overall results [9].

The particular queries used could also be unfair if some systems were not able to fully exploit their characteristics. This is of major importance for machine learning tasks where systems are first tuned with a training collection: if the query characteristics were very different between the training and evaluation collections, systems could be misguided. On the other hand, if the same collections were used from year to year, an increase in performance could be just due to overfitting and not to a real improvement [54]. Also, some evaluation measures could be unfair to some systems if accounting for information they cannot provide.

3.6 External Validity

External validity evaluates the extent to which the results of an experiment can be generalized to other populations and experimental settings. For example, a study on the effects of some cancer treatment would not have external validity if most patients in the sample were teenage males, as it would not be clear what the effect of the drug is in, say, elder women. Thus, and experiment acquires external validity by careful experimental design and justification of sampling and selection methods.

This is probably the weakest point of IR evaluation [54]. As mentioned, it is very important that the document collection and query set is representative of the domain being studied. On the other hand, having large collections means that the completeness of the ground truth is compromised: it is just not feasible to judge every query-document pair [8][62]. As mentioned, the usual solution is to pool the first k results of the participating systems and judge only those, assuming that all others are not relevant. This is an obvious problem because the very test collection (documents, queries and ground truth), which is in its own a product of the experiment, might not be reusable for

subsequent evaluations of new systems [14][15]. The validity of the latter experiments could be compromised.

Likewise, it is not justified to compare two systems evaluated with different test collections, because the results in each case are very dependent on the query set, relevance judgments, measures, etc. [6][54]. Indeed, it is known that different systems can perform very differently when evaluated with different collections, especially if machine learning techniques are involved. This highlights the lack of external validity in IRE experiments, and the importance of always interpreting the results in terms of pairwise system comparisons rather than absolute performance figures [54]. That is, comparisons across collections and claims about the state of the art based on a single collection, are not justified. Nonetheless, very rough comparisons between two systems across collections could be made if reporting the results of well-established baseline systems for those collections and their relative difference with the systems of interest [2].

3.7 Conclusion Validity

Conclusion validity evaluates the extent to which the conclusions drawn from the results of an experiment are justified. For example, a study might claim that people has better access to the Internet in China than in the U.S. because there are more users connected, when in fact the percentage of people connected, over the total population, is much less. Thus, an experiment acquires content validity by careful selection of the measuring instruments and the statistical methods used to draw de grand conclusions.

Two important characteristics of the effectiveness measures used in IR Evaluation are their stability and sensitivity. The results should be stable under different conditions, such as relevance judgments made by different people or different sets of queries, so the results do not vary significantly and alter the conclusions as to what systems are better [7]. Also, they are desired to discriminate between systems if they actually perform differently [55][39], and to do so with the minimum effort [41]. Likewise, they are desired to not discriminate between systems that actually perform very similarly. Note that these performance differences must be considered always in the context of the task and its underlying user model.

Given a set of systems and the scores they obtained for different queries according to some measure, they are usually compared in terms of their mean effectiveness score. Not until recently, statistical methods have been systematically employed and analyzed to compare systems by their score distribution rather than just their mean score [43][58]. At this point, it is very important to interpret correctly the results and understand the very issues of hypothesis testing; and most importantly, distinguish between statistical and practical significance: even if one system is found to be significantly better than another one, the difference might be extremely small to be noticed by users. In fact, the tiniest practical difference will turn out statistically significant with a sufficient number of queries.

4. CHALLENGES IN MUSIC IR EVALUATION

Research in IR follows a cycle that ultimately leads to the development of better systems. First, in the Development phase researchers build a system for a particular task, and to assess how good it is, there is an Evaluation phase. Once the experiments are finished, researchers then enter a phase of Interpretation of the results, which leads to a phase of Learning why the system worked well or bad and under what circumstances. Finally, with the new knowledge gained researchers get into an Improvement phase to try and make their system better, going back over to the Evaluation phase. Unfortunately, current evaluation practices in Music IR seem to fall short in this cycle.

Development. The task intent and its underlying user model are sometimes unclear or its real-world applicability uncertain. For instance, is it realistic that while the queries to the Query by Humming task are in audio format, the document collection is in symbolic form? Or, in the similarity tasks, is it realistic that the queries are actual items contained in the collection? Likewise, are 30 second clips realistic for all tasks?

Evaluation. Several tasks, such as Audio Chord Detection or Symbolic Melodic Similarity, use document collections either too small or biased toward some genre or time period [46][48], which jeopardizes the validity of the results. Moreover, the lack of standardized and public collections results in research groups using their personal, private, often undescribed and rarely analyzed collections, which precludes other researchers to compare systems or validate and replicate results, hindering the overall development of the field and often leading to wrong conclusions. In this line, the lack of standard evaluation software that all researchers can use, thus minimizing the likelihood of bugs and incorrect results, should be addressed too, especially with new or undocumented measures specific of Music IR.

Interpretation. Some effectiveness measurers, such as Normalized Recall at Group Boundaries, are used without description, references or source code, making them impossible to interpret or use in private evaluations. Also, widely-accepted baseline systems are very rarely included in evaluations, and when they are, they use to be implemented as random systems, having no useful value as a lower bound to which compare new systems. Another point that needs discussion is the set of statistical procedures used, or the lack thereof. Given the small-scale evaluations usually carried out in the Music IR field, it is imperative that statistical significance procedures be used, and certainly that the ones used are thoroughly selected and analyzed, for wrong conclusions can easily be drawn from incorrect procedures or incorrect interpretation [50].

Learning. When the results of an evaluation experiment are calculated and interpreted, the next step would be to figure out what happened and for what reasons. But there is a great problem here: most of the times the raw musical material is not available to experimenters, the actual queries used are unknown, and not even their characteristics are

published. Researchers cannot analyze the evaluation results and improve their systems: if they had very bad results for some queries, there is no way of knowing why. They can only use their private collections over and over again, ultimately leading to overfitting and misleading results.

Improvement. There is another reason why researchers are forced to use their private collections all along: current test collections put together in collective evaluation forums are hardly reusable. As seen, the incompleteness of ground truth data depends largely on the number of participating systems, and with the current low participation level, a new system would be highly penalized with the collection as is. The reusability is of course null if these data were not publicly available, as happens with some tasks. As such, researchers have no option but to blindly improve their systems and wait for another evaluation round, with no way of comparing cross-edition results due to the lack of data.

5. OPPORTUNITIES IN MUSIC IR EVALUATION

Although not easily, these shortcomings of current evaluation practices in Music IR can be overcame. To this end, we list several proposals to ease the way through the IR research and development cycle.

Collections. The document collections need to be large, move beyond the handful of songs currently being used in several tasks; and try to include heterogeneous material in terms of genre, time period, artist, etc. This is not hard to achieve, but when making such a collection open to other researchers, copyright issues immediately arise [21]. A possibility is to publish feature vectors and metadata, such as in the recent Million Song Dataset [5], although this still poses problems if researchers wanted to study a new feature or analyze specific items for which their system worked better or worse. In any case, these collections should be standard and used throughout the community, across tasks if possible, for a better comparison and understanding of the improvements between systems.

Raw Data. For a successful execution of the Learning and Improvement phases, raw musical material is needed. An alternative is to use music free of copyright restrictions, such as that provided by services like Jamendo, but the possible biases this might introduce are subject for further research. In this line, the use of artificial material, such as synthesized or error-mutated queries, should be revised [37].

Evaluation Model. Having publicly accessible and standardized collections would allow for a change in the current execution model employed in MIREX. Researchers could be in charge of executing their systems and producing the runs to submit back to MIREX, relieving them from a good deal of workload and bringing researchers reluctant to give their algorithms away to third parties. This data-to-algorithm model is used in the recent MusiCLEF forum [32], and in fact it is the only viable way of moving to large scale evaluations, not only in terms of data but also in terms of wider participation. The current algorithm-to-data model is in our view unsustainable in the long run, let alone if IMIRSEL finally stops receiving funds [24], and platforms like MIREX-DIY under NEMA [61] would still not permit a full execution of the IR cycle.

Organization. The current organization of MIREX rests heavily on the IMIRSEL team, who plan, schedule and run a good number of tasks each year. We propose a 2nd tier organization below, for each particular task, and by leading third-party researchers. These organizers would deal with all the logistics, planning, evaluation, troubleshooting and so on, diminishing the workload of IMIRSEL, which would act as a sort of steering meta-organization tier providing the necessary resources and general planning. This is the format successfully adopted by major Text IR forums like TREC or CLEF, which has helped in smoothing the process and developing tasks to push the state of the art in each edition.

Specific Methodologies. Both new methodologies [46][48][27][22] and effectiveness measures [47] have been proposed for Music IR tasks, needing meta-evaluation studies in the near future to keep improving the evaluations. Some work has studied the reduction of effort needed to evaluate through the use of crowdsourcing platforms [49][33], and further studies should follow this line given the usual restrictions the Music IR field has as to availability of resources. Another line is the study of human effects on ground truth data and evaluation results [28].

Overview Publications. The organization proposal would also benefit the community if by the end of each MIREX edition the organizers published an overview paper thoroughly detailing the process followed, data, results, and discussion to boost the Interpretation and Learning phases. Such a publication would be the perfect wrap-up to the participant-papers that describe the systems but rarely investigate and elaborate on the results. In fact, many of these participant-papers are not even drafted.

Software Standardization. It is not rare to find incorrect evaluation results due to software bugs. With the development and acceptance of a software package to evaluate systems we would gain in reliability within and between research groups, speeding up experiments and guiding novice researchers. Also, it would further serve as documentation of the measures and processes used, for the implementation of some details is unknown or subject to different interpretations; and it would call for the standardization of data formats to speed up the IR cycle.

Baselines. The establishment of baseline systems to serve as a lower bound on effectiveness would help in assessing the overall progress in the field. With the standardization of formats, public software, public collections of raw music material and the supervision of task-specific organizers, the inclusion of baselines in these experiments would greatly benefit the execution of the IR cycle and the measurement of the state of the art.

Commitment. In general, the current problems of Music IR Evaluation need to be acknowledged by researchers. Now that we have a well-established evaluation forum like MIREX, we need to start questioning the validity of the

experiments, with the sole purpose of making them better and more striking. Current IR experiments seem to stop at the Evaluation phase of the IR cycle, but the next phases are often ignored or impossible to engage into.

6. CONCLUSIONS

We have presented a survey of the Text IR literature on studies tackling the problem of IR Evaluation experiments. From the point of view of the analysis of experimental validity, this survey shows different aspects of IR Evaluation that have been overlooked and need special attention in the Music IR domain. From the point of view of the IR research and development cycle a researcher follows in Music IR, we have also shown that current evaluation practices force researchers to stop early in the cycle. Evaluation experiments release good amounts of numbers and plots, but there is a lack of proper interpretation and discussion due in part to the lack of public and standardized resources, usually leaving researchers blind to improve their systems. In this line, several proposals are made to engage researchers in these last phases of the cycle, which should ultimately lead to a more rapid development of the field.

We hope this paper makes the case for MIR Meta-Evaluation studies and the fact that they *are* actual MIR research, playing a central role in which researchers should engage to begin a tradition of evaluation articles in ISMIR.

REFERENCES

[1] Alonso et al., Can We Get Rid of TREC assessors? Using Mechanical Turk for Relevance Assessment, *SIGIR Workshop on the Future of IR Evaluation*, 2009.

[2] Armstrong et al., Improvements that Don't Add Up: Ad-Hoc Retrieval Results since 1998, *CIKM*, 2009.

[3] Bailey et al., Relevance Assessment: Are Judges Exchangeable and Does it Matter?, *SIGIR*, 2008.

[4] Bennett et al., Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments, *SIGIR Forum*, 2008.

[5] Bertin-Mahieux et al., The Million Song Dataset, *ISMIR*, 2011.

[6] Bodoff et al., Test Theory for Assessing IR Test Collections, *SIGIR*, 2007.

[7] Buckley et al., Evaluating Evaluation Measure Stability, *SIGIR*, 2000.

[8] Buckley et al., Retrieval Evaluation with Incomplete Information, *SIGIR*, 2004.

[9] Buckley et al., Bias and the Limits of Pooling for Large Collections, *Journal of IR*, 2007.

[10] Carterette et al., Minimal Test Collections for Retrieval Evaluation, *SIGIR*, 2006.

[11] Carterette, Robust Test Collections for Retrieval Evaluation, *SIGIR*, 2007.

[12] Carterette et al., If I Had a Million Queries, *ECIR*, 2009.

[13] Carterette et al., The Effect of Assessor Error on IR System Evaluation, *SIGIR*, 2010.

[14] Carterette et al., Measuring the Reusability of Test Collections, *WSDM*, 2010.

[15] Carterette et al., Reusable Test Collections Through Experimental Design, *SIGIR*, 2010.

[16] Carterette, System Effectiveness, User Models, and User Utility: A General Framework for Investigation, *SIGIR*, 2011.

[17] Carvalho et al., Crowdsourcing for Search Evaluation, *SIGIR Forum*, 2010.

[18] Cleverdon, The Significance of the Cranfield Tests on Index Languages, *SIGIR*, 1991.

[19] Downie, Interim Report on Establishing MIR/MDL Evaluation Frameworks: Commentary on Consensus Building, *ISMIR Panel on Music Information Retrieval Evaluation Frameworks*, 2002.

[20] Downie, The MIR/MDL Evaluation Project White Paper Collection, 3rd ed, 2003.

[21] Downie, The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future, *Computer Music Journal*, 2004.

[22] Downie et al., Audio Cover Song Identification: MIREX 2006-2007 Results and Analysis, *ISMIR*, 2008.

[23] Downie et al., The Music Information Retrieval Evaluation eXchange: Some Observations and Insights, in *Advances in Music IR*, Springer, 2010.

[24] Downie, MIREX Next Generation, *music-ir email list*, 2011. Available at: http://listes.ircam.fr/wws/info/music-ir.

[25] Guiver et al., A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation, *ACM Trans. Inf. Sys.*, 2009.

[26] Harman, Information Retrieval Evaluation, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2011.

[27] Hu et al., The 2007 MIREX Audio Mood Classification Task: Lessons Learned, *ISMIR*, 2008.

[28] Jones et al., Human Similarity Judgments: Implications for the Design of Formal Evaluations, *ISMIR*, 2007.

[29] Järvelin et al., Cumulated Gain-Based Evaluation of IR Techniques, *ACM Trans. Inf. Sys.*, 2002.

[30] Katzer et al., Evaluating Information: A Guide for Users of Social Science Research, 4th ed., 1998.

[31] Kekäläinen, Binary and Graded Relevance in IR Evaluations: Comparison of the Effects on Ranking of IR Systems, *Inf. Proc. Mngt.*, 2005.

[32] Lartillot et al., MusiClef: A Benchmark Activity in Multimodal Music Information Retrieval, *ISMIR*, 2011.

[33] Lee, Crowdsourcing Music Similarity Judgments using Mechanical Turk, *ISMIR*, 2010.

[34] Marchionini, Exploratory Search: from Finding to Understanding, *Communications of the ACM*, 2006.

[35] Mitchell et al., Research Design Explained, 7th ed., 2009.

[36] Montgomery, Design and Analysis of Experiments, 7th ed., 2009.

[37] Niedermayer et al., On the Importance of 'Real' Audio Data for MIR Algorithm Evaluation at the Note-Level: A comparative Study, *ISMIR*, 2011.

[38] Robertson, On the Contributions of Topics to System Evaluation, *ECIR*, 2011.

[39] Sakai, On the Reliability of Information Retrieval Metrics Based on Graded Relevance, *Inf. Proc. Mngt.*, 2007.

[40] Sakai et al., On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Journal of IR*, 2008.

[41] Sanderson et al., Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *SIGIR*, 2005.

[42] Sanderson et al., Do User Preferences and Evaluation Measures Line Up?, *SIGIR*, 2010.

[43] Smucker et al., A Comparison of Statistical Significance Tests for Information Retrieval Evaluation, *CIKM*, 2007.

[44] Soboroff et al., Ranking Retrieval Systems Without Relevance Judgments, *SIGIR*, 2001.

[45] Turpin et al., Why Batch and User Evaluations Do Not Give the Same Results, *SIGIR*, 2001.

[46] Typke et al., A Ground Truth for Half a Million Musical Incipits, *Journal of Digital Inf. Mngt.*, 2005.

[47] Typke et al., A Measure for Evaluating Retrieval Techniques based on Partially Ordered Ground Truth Lists, *IEEE Int. Conf. on Multimedia and Expo*, 2006.

[48] Urbano et al., Improving the Generation of Ground Truths based on Partially Ordered Lists, *ISMIR*, 2010.

[49] Urbano et al., Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks, *SIGIR Workshop Crowdsourcing for Search Evaluation*, 2010.

[50] Urbano et al., Audio Music Similarity and Retrieval: Evaluation Power and Stability, *ISMIR*, 2011.

[51] Voorhees, Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, *Inf. Proc. Mngt.*, 2000.

[52] Voorhees, Evaluation by Highly Relevant Documents, *SIGIR*, 2001.

[53] Voorhees, Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC, *in* [20], 2002.

[54] Voorhees, The Philosophy of Information Retrieval Evaluation, *CLEF*, 2002.

[55] Voorhees et al., The Effect of Topic Set Size on Retrieval Experiment Error, *SIGIR*, 2002.

[56] Voorhees et al., TREC: Experiment & Evaluation in Information Retrieval, 2005.

[57] Webber et al., Precision-At-Ten Considered Redundant, *SIGIR*, 2008.

[58] Webber et al., Statistical Power in Retrieval Experimentation, *CIKM*, 2008.

[59] Yilmaz et al., Estimating Average Precision with Incomplete and Imperfect Information, *CIKM*, 2006.

[60] Yilmaz et al., A Simple and Efficient Sampling Method for Estimating AP and NDCG, *SIGIR*, 2008.

[61] Zhu et al., MIREX-DIY under NEMA, *ISMIR*, 2010.

[62] Zobel, How Reliable are the Results of Large-Scale Information Retrieval Experiments?, SIGIR, 1998.