Journal of Intelligent Information Systems

Evaluation in Music Information Retrieval

Julián Urbano · Markus Schedl · Xavier Serra

Received: 19 November 2012 / Accepted: 9 May 2013

Abstract The field of Music Information Retrieval has always acknowledged the need for rigorous scientific evaluations, and several efforts have set out to develop and provide the infrastructure, technology and methodologies needed to carry out these evaluations. The community has enormously gained from these evaluation forums, but we have reached a point where we are stuck with evaluation frameworks that do not allow us to improve as much and as well as we want. The community recently acknowledged this problem and showed interest in addressing it, though it is not clear what to do to improve the situation. We argue that a good place to start is again the Text IR field. Based on a formalization of the evaluation process, this paper presents a survey of past evaluation work in the context of Text IR, from the point of view of validity, reliability and efficiency of the experiments. We show the problems that our community currently has in terms of evaluation, point to several lines of research to improve it and make various proposals in that line.

Keywords Music Information Retrieval · Text Information Retrieval · Evaluation and Experimentation · Survey

1 Introduction

Information Retrieval (IR) is a highly experimental discipline. Evaluation experiments are the main research tool to scientifically compare IR techniques and advance the state of the art through careful examination and interpretation of

M. Schedl is supported by the Austrian Science Fund (FWF): P22856.

Julián Urbano
Department of Computer Science - University Carlos III of Madrid, Leganés, Spain
E-mail: jurbano@inf.uc3m.es

Markus Schedl
Department of Computational Perception - Johannes Kepler University, Linz, Austria
E-mail: markus.schedl@jku.at

Xavier Serra
Music Technology Group - Universitat Pompeu Fabra, Barcelona, Spain
E-mail: xavier.serra@upf.edu

their results. Despite being a quite young field of research, Music IR is not an exception. In its early years, our community mirrored Text IR in terms of evaluation practices, but there has been little research studying whether that mirroring should be fully applied and, when it should not, what alternatives work better. These are very important questions to deal with, because reaching wrong conclusions from evaluation experiments may not only hamper the proper development of our field, but also make us follow completely wrong research directions. Some presentations and discussions at the recent ISMIR (International Society for Music IR) 2012 conference[1] showed the general concern of the Music IR community in this matter, but also the lack of clear views to improve the situation.

In what follows we show the importance and impact of research on IR Evaluation and how it has evolved for the past fifty years in Text and Music IR. Our main argument is that our community has missed a great deal of research actually devoted to improve evaluation experiments. In this paper we formalize and discuss the IR Evaluation process and show where our experiments may fail. We review the Text IR Evaluation literature and discuss how it tackles different issues of these experiments; namely their validity, reliability and efficiency. Our review is intended as a starting point for Music IR researchers to engage in this discussion and improve the currently used evaluation frameworks. We conclude by identifying some current challenges in this area and discussing proposals for future work.

1.1 Importance and Impact of IR Evaluation Research

Evaluation is recognized as one of the key areas in Information Retrieval research. In 2002, a workshop gathering world-wide leading IR researchers identified Evaluation as one of the seven grand challenges in the field (Allan and Croft, 2003). This meeting turned into the SWIRL series of workshops, which explore the long-range issues in IR, recognize key challenges and identify past and future research directions. Reflecting upon the history of IR research, the first workshop collected in 2004 a list of 47 recommended readings for IR researchers (Moffat et al, 2005), where as many as 9 (19%) were devoted to analyzing or improving evaluation methods, clearly showing the importance of this topic. The second meeting took place in 2012, and Evaluation was still recognized as one of the six grand challenges in Information Retrieval (Allan et al, 2012). Even the 2012 ACM Computing Classification System[2], which updates the previous 1998 version, reflects the importance of Evaluation by listing it as one of the eight main areas in the IR field.

In the Music IR side, the recent MIReS project (Roadmap for Music Information ReSearch), funded by the 7th Framework Programme of the European Commission, is an international and collective attempt at recognizing the challenges and future directions of the field. Evaluation is also listed here as one of the seven technical-scientific grand challenges in Music IR research[3]. This recognition was also explicit during ISMIR 2012, where a discussion panel on Evaluation in Music IR was held along with a late-breaking session (Peeters et al, 2012).

To quantitatively measure the importance and impact of evaluation studies in IR, we analyzed the proceedings of the two major conferences on Text IR and

[1] http://ismir2012.ismir.net

[2] http://www.acm.org/about/class/2012

[3] http://mires.eecs.qmul.ac.uk/wiki/index.php/MIR_Challenges

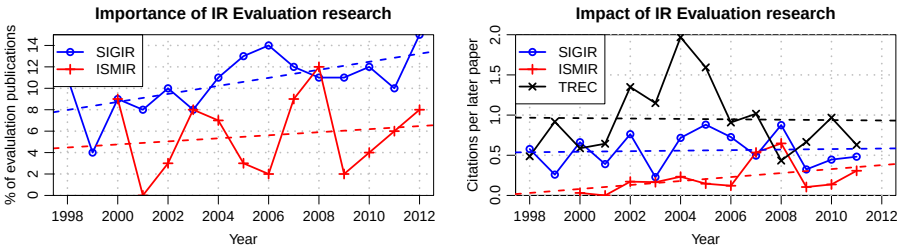

Fig. 1 Importance (left) of publications in SIGIR and ISMIR proceedings related to IR Evaluation; and their impact (right) along with TREC overview papers. The dashed lines plot a linear fit on the data points.

Music IR: the ACM SIGIR and ISMIR conferences. For each edition since 1998, we examined the proceedings and counted the number of publications devoted to analyzing or improving evaluation methods. Figure 1-left shows that on average Evaluation comprised 11% of research in SIGIR, while in ISMIR this goes down to 6%. In fact, it is very interesting to see that the relative difference between both trends has been twofold over the years. To measure the impact of that research, we also checked the number of citations received by evaluation papers for each year, and then divided the citation counts by the total number of papers (related to evaluation or not) published later and in the same venue. Figure 1-right shows that SIGIR papers on evaluation are cited an average of 0.56 times for each paper published later. Impact seems much lower in ISMIR, although the positive trend shows that the community is indeed becoming aware of this research. These figures serve as a rough indication that Evaluation is in fact a very important topic of research which might not be receiving enough attention from the Music IR community yet. Another indicator of this mismatch can be found in the best paper awards: from the 17 papers awarded in SIGIR, 4 (24%) are related to evaluation. To the best of our knowledge, this has never been the case in ISMIR.

1.2 History of IR Evaluation Research

Information Retrieval Evaluation has attracted a wealth of research over the years (Harman, 2011; Robertson, 2008) (see Figure 2). The Cranfield 2 experiments (Cleverdon, 1991), carried out by Cyril Cleverdon between 1962 and 1966, are often cited as the basis for all modern IR evaluation experiments, and even as the birthplace of the IR field altogether (Harman, 2011). Cleverdon established the so-called Cranfield paradigm for IR Evaluation based on test collections (see Section 2). From 1966 to 1967, the MEDLARS (Medical Literature Analysis and Retrieval System) study (Lancaster, 1968) focused on the evaluation of a complete system from a user perspective, taking into consideration the user requirements, response times, required effort, etc. The SMART project (Lesk et al, 1997) was directed by Gerard Salton from 1961 until his death in 1995. One of the results of the project was the development of several test collections, procedures and measures that allowed researchers to perform batch evaluation experiments in a systematic fashion. Meanwhile, the ACM SIGIR conference started in 1978 as the premier venue for Text IR research.

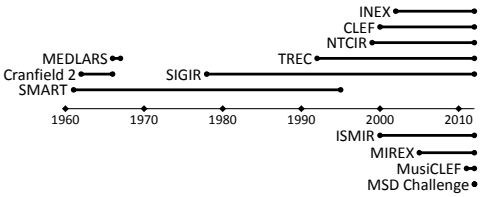

Fig. 2 Timeline of Evaluation in Text IR (top) and Music IR (bottom).

Very successful IR Evaluation forums have followed ever since. TREC[4] (Text REtrieval Conference) started in 1992 to provide infrastructure necessary for evaluations based on large-scale test collections (Voorhees and Harman, 2005). NTCIR[5] (National Institute of Informatics–Testbeds and Community for Information access Research) started in 1999 to provide similar infrastructure for Asian languages. CLEF[6] (Conference and Labs of the Evaluation Forum) started in 2000 with an emphasis on multilingual and multimodal information, and INEX[7] (INitiative for the Evaluation of XML retrieval) focuses on structured information since 2002.

On the Music IR side, the ISMIR conferences started in 2000. Reflecting upon this very long tradition of IR Evaluation research, the "ISMIR 2001 resolution on the need to create standardized MIR test collections, tasks, and evaluation metrics for MIR research and development" was drafted during ISMIR 2001, and signed by many members of the Music IR community as a demonstration of the general concern regarding the lack of formal evaluations (Downie, 2003). A series of three workshops then followed between July 2002 and August 2003, where researches engaged in this long-needed work for evaluation in Music IR (Downie, 2003). There was some general agreement that evaluation frameworks for Music IR would need to follow the steps of TREC (Voorhees, 2002b), although it was clear too that special care had to be taken not to oversimplify the TREC evaluation model (Downie, 2002), because Music IR differs greatly from Text IR in many aspects that affect evaluation (Downie, 2004). The general outcome of these workshops and many other meetings was the realization by the Music IR community that a lot of effort and commitment was needed to establish a periodic evaluation forum for Music IR systems. The ISMIR 2004 Audio Description Contest stood up as the first international evaluation project in Music IR (Cano et al, 2006). Finally, the first edition of the Music Information Retrieval Evaluation eXchange[8] (MIREX) took place in 2005, organized by IMIRSEL (International Music IR Systems Evaluation Laboratory) (Downie et al, 2010), and ever since it has evaluated over 1,500 Music IR systems for many different tasks on a yearly basis. More recent evaluation efforts have appeared in the Music IR field, namely the MusiClef[9] campaign in 2011 (Lartillot et al, 2011) (now part of the MediaEval series) and the Million Song Dataset Challenge[10] in 2012 (McFee et al, 2012).

[4] http://trec.nist.gov
[5] http://research.nii.ac.jp/ntcir/
[6] http://www.clef-initiative.eu
[7] http://inex.mmci.uni-saarland.de
[8] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[9] http://www.multimediaeval.org/mediaeval2012/newtasks/music2012/
[10] http://labrosa.ee.columbia.edu/millionsong/challenge

1.3 Motivation

The impact of MIREX has been without doubt positive for the Music IR community (Cunningham et al, 2012), not only for fostering these experiments, but also the study and establishment of specific evaluation frameworks for the Music domain. For some time the community accepted MIREX as "our TREC", but we are just now becoming aware of its limitations (Peeters et al, 2012).

Evaluation experiments in IR are anything but trivial (Harman, 2011; Sanderson, 2010; Voorhees, 2002a; Tague-Sutcliffe, 1992; Saracevic, 1995). Figure 1 shows that for the past fifteen years the Text IR literature has been flooded with studies showing that they have their very own issues, proposing different approaches and techniques to improve the situation. While the Music IR community has inherited good evaluation practices by adopting TREC-like frameworks, some are already outdated, and most still lack appropriate analysis. We agree that not everything from the Text IR community applies to Music IR, but *many evaluation studies do*. In fact, our evaluation frameworks and body of knowledge are based on research up to the early 2000's, but nearly 250 evaluation papers have been published in SIGIR alone, and several landmark studies have taken place in the context of TREC since MIREX started in 2005. These studies are particularly focused on large-scale evaluation, robustness and reliability, and none of them has even been considered for Music IR. In our view, this is where our community should start.

Therefore, Evaluation is not only a cornerstone in IR for allowing us to quantitatively measure which techniques work and which do not, but also a very active area of research receiving a lot of attention in recent years. We have seen this tendency in Text IR with a series of indicators which, at the same time, show that the Music IR field does not seem to pay as much attention as it probably should.

2 The Cranfield Paradigm for IR Evaluation

Batch evaluation experiments in IR usually follow the traditional Cranfield paradigm conceived by Cyril Cleverdon half a century ago (Cleverdon, 1991). The main element needed for these evaluations is a test collection, which is made up of three basic components (Sanderson, 2010): a collection of documents, a set of information needs and the relevance judgments telling what documents are relevant to what information needs (the ground truth or gold standard). These test collections are built within the context of a particular task, which defines the expected behavior of the systems, the intent of the information needs, and the characteristics of the documents to be considered relevant. Several effectiveness measures are used to score systems following different criteria, always from the point of view of a user model with assumptions as to the potential real users of the systems.

A typical IR research scenario goes as follows (Harman, 2011; Voorhees, 2002a). First, the task is identified and 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 trying to mimic the potential requests of the final users. The systems to evaluate return their results for the particular query set and document collection, and these results are then evaluated using several effectiveness measures. Doing so, we attempt to assess how well the systems would have satisfied a real user at

different levels. This framework promotes rapid development and improvement of systems because it allows researchers to systematically and iteratively evaluate and compare alternative algorithms and parametrizations. In that line, it also allows to repeat and reproduce results across research groups.

Music IR tasks[8] such as *Audio Music Similarity* or *Query by Humming* clearly fit into this classic retrieval setting. In other cases such as *Audio Melody Extraction* and *Audio Chord Estimation* a slightly different procedure is followed. Instead of retrieving documents in response to a query, systems provide annotations for different segments of this query item; and therefore the ground truth data does not provide information about document-query pairs, but rather about different segments of the queries. Other tasks such as *Audio Mood Classification* and *Audio Genre Classification* are similar to annotation tasks, but instead of providing annotations for different segments of the query, systems provide tags for the query itself. Therefore, in all our tasks systems are provided with some kind of query item and they return different output data in response.

3 Formalizing the IR Evaluation Process

The ultimate goal of evaluating an IR system is to characterize the usage experience of the users who will employ it. We may consider several facets. For example, given an arbitrary query input, we may be interested in knowing how likely it is for a user to be satisfied by the system results or the interface used to show them, or how long it would take to complete the task defined by the query. We can formalize these user-measures by employing random variables. For example, we can define the variable U_1, that equals 1 if the user is satisfied by the system and 0 otherwise. This variable is defined by a probability distribution function f_{U_1}, specified by a vector of parameters $\boldsymbol{\theta}_{U_1}$. We could consider another variable U_2, equal to the task completion time in the interval $(0, \infty)$, similarly defined by a probability distribution function f_{U_2} with parameters $\boldsymbol{\theta}_{U_2}$.

This multifaceted characterization of the system usage allows researchers to assess the performance of the system from different perspectives, such as the probability of user satisfaction, the minimum time needed to complete 50% of the tasks, the probability that at least 80% of users will find the system satisfactory, etc.

3.1 Modeling Users

Unfortunately, the are several problems to know what the f_{U_i} distributions look like. First, including real users in experiments is expensive and complex, and there are ethical issues to consider (e.g. privacy and wages). Second, involving users makes it harder to tune system parameters due to the cost of running an evaluation trial. Third, it is hard to reproduce experiments that involve human subjects, so system comparisons across research groups is difficult. To minimize these problems, Cleverdon came up with the idea of removing actual users from the experiment but including a static user component: the ground truth. He controlled the experiment and reduced all sources of variability to just the systems themselves, so it became possible to iteratively compare them in a systematic, fast and inexpensive way.

Therefore, when evaluating a system following the Cranfield framework we are actually characterizing the system response rather than the user experience. The ground truth provides us with information on how good or accurate that response is, but it does not provide information on the user-system interaction, let alone on user-specific characteristics such as perceived easiness in using the system. Likewise, each of the system-based measures used in the experiment provides us with a description of the system from different perspectives, each of which can again be modeled with random variables. For instance, when evaluating music similarity systems we may use a variable S_1 to refer to the similarity of the items returned by the system, and another variable S_2 might refer to the rank at which the system retrieves the first similar item. These variables are computed with effectiveness measures (e.g. *Average Gain* and *Reciprocal Rank*), and they are also defined by functions f_{S_1} and f_{S_2} with parameters $\boldsymbol{\theta}_{S_1}$ and $\boldsymbol{\theta}_{S_2}$. The assumption underlying Cranfield is that S_i is correlated with U_i, and therefore the distribution defined by f_{S_i} can somehow be used to describe the distribution defined by f_{U_i}.

3.2 Parameter Estimation

Computing the parameters $\boldsymbol{\theta}_{S_i}$ is clearly impossible. It requires to evaluate our system with the universe of all queries, that is, all existing queries and all queries yet to exist, which is a potentially infinite task. Instead, we use a sample of queries. When we evaluate a system according to a measure i, we compute an effectiveness score for each query. When we repeat the process with all queries in the sample, we are actually estimating the distribution f_{S_i} that defines the random variable associated with the effectiveness measure. That is, we are estimating the parameters $\boldsymbol{\theta}_{S_i}$, and because we assume the correlation between S_i and U_i, we treat those as estimates of the $\boldsymbol{\theta}_{U_i}$ parameters that define the user-based distribution.

In most cases there is really no theoretical basis for using one distribution family or another to describe these variables (e.g. Log-Normal or Gamma). In reality, what is of most value to a researcher is just knowing the first and second moments of the distributions: the mean and the variance. These provide us with estimates of the average performance and how much variability there is.

3.3 Validity, Reliability and Efficiency

In summary, we can look at an evaluation experiment as just an estimator of the true parameters defining a user-based distribution. An effectiveness measure is our measurement instrument, whose system-based distribution is assumed to perfectly correlate with our target user-based distribution. As such, there are three aspects

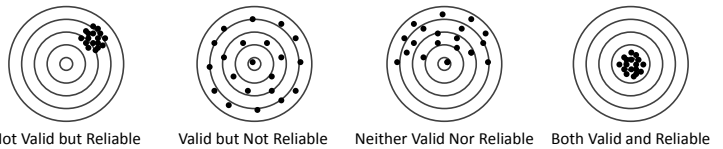

Fig. 3 Validity and Reliability. Adapted from (Trochim and Donnelly, 2007).

of these evaluations that must be considered: validity, reliability and efficiency (Tague-Sutcliffe, 1992) (further discussion is provided in Sections 4 to 6):

Validity. Do our effectiveness measures and ground truth data really define system-distributions that match the intended user-distributions? We assume there is a function mapping S_i to U_i, and therefore f_{S_i} to f_{U_i}. In fact, researchers somehow assume $U_i = S_i$. In a more relaxed form, validity can be reformulated as: are we really measuring what we want to?

Reliability. How many query items are needed in the evaluation so that the estimates can be trusted? The more queries we use, the smaller the random error we have in our estimates, but the higher the cost too. Therefore, evaluation experiments must find a tradeoff between reliability and effort. In a more relaxed form, reliability can be formulated as: how repeatable are our results?

Efficiency. Creating a ground truth set is usually a very expensive and tedious task, and some forms of ground truth data can be prohibitive for a large number of query items. Therefore, the efficiency of the ground truth annotation process directly impacts the reliability of the evaluation. On the other hand, an efficient annotation process might be inaccurate, lowering the validity of the results. Therefore, evaluation experiments must also find a balance between validity and reliability and the cost of the annotation process.

Figure 3 illustrates validity and reliability with the metaphor of a target (Trochim and Donnelly, 2007). Imagine our goal is the center of the target (i.e. the mean of U_i), and each shot we take is our measurement with a different test collection. In the first and fourth examples we have an instrument that is very reliable, but in the first case we are clearly off the target. In the second and third examples our instrument is not reliable, but in the second case we still manage to hit around the target so that our measure is correct on average. In this example, efficiency can be thought of as the cost of the weapon: rifle, bow, handgun, etc.

In Statistics terms, validity refers to the *accuracy* and *bias* of the estimates, and reliability refers to their *precision* or *variance* (Lehmann and Casella, 1998). That is, how close they are to the true parameters and how much uncertainty there is in those estimates. In Machine Learning terms, validity refers to the *bias* of a learner, and reliability refers to its *variance* (Geman et al, 1992). That is, the average difference over training datasets between the true values and the predictions, and how much they vary across training datasets. They can also be linked to the concepts of systematic and random error in measurement (Taylor, 1997).

4 Validity

Validity is the extent to which an experiment actually measures what the experimenter intended to measure (Shadish et al, 2002; Trochim and Donnelly, 2007; Tague-Sutcliffe, 1992). Validity is frequently divided in four types that build upon each other, addressing different aspects of an experiment. *Conclusion Validity* relates to the relationship found between our experimental treatments (systems) and our response variables (user-measures). Can we conclude that the systems are different? How much different? *Internal Validity* relates to confounding factors that might cause the differences we attribute to the systems. Are those differences caused by specific characteristics of the annotators or the queries? *External Validity* relates to the generalization of that difference to other populations. Would

system differences remain for the wider realm of all genres and artists? *Construct Validity* relates to the actual relationship between the system-measures and the user-measures. Do differences in system-measures directly translate to the same differences in user-measures? How do those differences affect end users?

4.1 Conclusion Validity

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 (Sakai, 2007; Kekäläinen, 2005), and some measures are even better correlated with others than with themselves (Webber et al, 2008b), evidencing problems when predicting user-measures. In general, system-measures should be correlated with user-measures, but observing a difference between two systems according to some system-measure does not necessarily mean there is a noticeable difference with end users. For example, it can be the case that relatively large differences need to appear between systems for users to actually note them (Urbano et al, 2012).

At this point it is important to note that in most situations systems are not provided with any kind of user information (Järvelin, 2011; Schedl and Flexer, 2012), and therefore our results should be interpreted as if targeting *arbitrary* users. As such, even if our system-measures corresponded perfectly to user-measures, the system distributions estimated with an evaluation experiment would not correspond perfectly to the expected user distributions because we are not accounting for user factors in the ground truth data (Voorhees, 2000; Urbano et al, 2012).

It is also important to recall that an evaluation experiment provides an estimate of a true population mean, which bears some degree of uncertainty due to sampling. Confidence intervals should always be calculated when drawing conclusions from an experiment, to account for that uncertainty and provide reliable reports of effect sizes (Cormack and Lynam, 2006). Depending on the experimental conditions, it might be the case that such interval is too wide to draw any accurate conclusion regarding the true performance of systems. In this line, it is important to distinguish between *confidence intervals*, used as estimators of distribution parameters such as the true mean; and *prediction intervals*, which serve as estimators of the expected performance on any new query item.

4.2 Internal Validity

Ground truth data is a much debated part of IR Evaluation because of the subjectivity component it usually has. 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 (Schamber, 1994). As such, the validity of evaluation experiments can be questioned because different results are obtained depending on the people that make the annotations. Nevertheless, it is generally assumed that ground truth data is invariable, and user-dependent factors are ignored (Järvelin, 2011; Schedl and Flexer, 2012). Several studies have shown that absolute scores do indeed change, but that relative differences between

systems stand still for the most part (Voorhees, 2000). For domain-specific tasks results may have large variations (Bailey et al, 2008), and for very large-scale experiments different assessor behaviors may also have a large impact on the results (Carterette and Soboroff, 2010), let alone if the ground truth has inconsistencies.

Likewise, if a low-cost evaluation method were used with an incomplete ground truth (see Section 6), systems more alike could reinforce each other, while systems with novel technology might be harmed (Zobel, 1998). In general, making assumptions about missing annotations affects both the measures (Buckley and Voorhees, 2004; Sakai and Kando, 2008) and the overall results (Buckley et al, 2007). 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 (Carterette et al, 2010a,b).

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 repeatedly, an increase in performance could be just due to overfitting and not to a real improvement (Voorhees, 2002a). Also, some evaluation measures could be unfair to some systems if accounting for information they cannot provide.

4.3 External Validity

In IR Evaluation it is very important to clearly define what our *target populations* are. That is, who our final users are, the music corpora they will work with, etc. When we carry out an experiment to evaluate a system, we are interested in the distributions of user-measures for those populations. The problem is that we might not be able to get access to those users (e.g. anonymous users of an online music service, music artists, etc.) or those corpora (e.g. copyrighted material or songs yet to exist). Therefore, we often have access only to restricted and biased subsets of those populations. These are the *accessible populations*. To reduce costs, we draw a *sample* from those accessible populations and carry out the experiment. Our assumption when doing this is that the results obtained with our samples can be generalized back to the target populations. In particular, for an arbitrary system-measure we assume that the sample mean is an unbiased estimator of the true population mean because our sample is *representative* of the target population.

This is probably the weakest point in IR Evaluation (Voorhees, 2002a). In order to get a sample representative of the accessible population we generally want that sample to be large: the more elements we draw the better our estimates will be. This poses obvious problems in terms of cost. Having large corpora means that the completeness of the ground truth is compromised: it is just not feasible to judge every query-document pair or annotate every single segment of every query (Buckley and Voorhees, 2004; Zobel, 1998). As a result, collections contain too few query items or their corpus is too small to be realistic.

In addition, we want the sample to be random in order to eliminate biases. In Text IR, this has been a problem since the early days, because there was no pool of queries to draw a sample from; they were made up on demand for the evaluations (Voorhees and Harman, 2005). Because of this, the Text IR literature has always

emphasized that results with a single test collection must be taken with a grain of salt because results are highly dependent on document collections and query sets (Robertson, 2011; Voorhees, 2002a); that is, systems may work very well with a test collection but significantly worse with a different one (Poibeau and Kosseim, 2001), especially if Machine Learning algorithms are involved. This is also emphasized in that results should be interpreted in terms of relative pairwise system differences rather than absolute. That is, comparisons across collections and claims about the state of the art based on a single collection are not justified.

To partially overcome this problem with non-random samples, the Text IR community has traditionally sought very large collections. In the last decade though, several sources of information, such as query logs from commercial search engines, are used to draw random samples and slightly reduce the cost. This has the additional advantage that queries are likely to be representative of the final user needs. A similar problem arises in Music IR because the accessible population is hardly representative of the target population, so even if we have a very large sample we still can not generalize back as we would like. Recent research has studied query selection methods that try to avoid queries that do not provide useful information to differentiate between systems (Guiver et al, 2009; Robertson, 2011).

4.4 Construct Validity

In IR experiments, Construct Validity is concerned mainly with the system-measures used, their underlying user model (Carterette, 2011), and their correlation with user-measures. Unlike batch experiments where the only user component is the ground truth, some studies carried out experiments with actual users interacting with IR systems. They found little correlation between system-measures and user-measures, questioning the whole Cranfield paradigm (Hersh et al, 2000; Turpin and Hersh, 2001). But the problem strives in seeking correlations between measures that are not really supposed to be related (Smucker and Clarke, 2012a). For instance, *Precision* is not designed as an indicator of task completion time; *Reciprocal Rank* is. Various alternatives have been studied, such as using different relevance thresholds on a per assessor basis (Scholer and Turpin, 2008), carefully normalizing effectiveness scores (Al-Maskari et al, 2007), or including other factors in the measurement of relevance (Smucker and Clarke, 2012b; Huffman and Hochster, 2007). Later work further explored this issue, finding clear correlations between system effectiveness and user satisfaction (Allan et al, 2005; Sanderson et al, 2010). Similar studies have appeared recently at ISMIR 2012, showing little relationships (Hu and Kando, 2012) and tight correlations (Urbano et al, 2012).

The development of appropriate system-measures that closely capture the user experience is thus very important. 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 (Sanderson et al, 2010). Instead, measures that take the rank into account (Moffat and Zobel, 2008), graded relevance judgments (Voorhees, 2001; Kekäläinen, 2005), or a combination of them (Järvelin and Kekäläinen, 2002; Robertson et al, 2010; Chapelle et al, 2009; Kanoulas and Aslam, 2009), are more appropriate. Other forms of ground truth can also be studied (Bennett et al, 2008).

5 Reliability

Reliability is the extent to which the results of the experiment can be replicated (Trochim and Donnelly, 2007; Tague-Sutcliffe, 1992). Will we obtain similar results if we repeat the experiment with different sets of queries and annotators?

As mentioned, it is very important that our samples be representative of the target populations. It is important not only because we want our estimates to correspond to the true population parameters, but also because our results would otherwise be unreliable: with one sample system A is better than system B, but with another sample it is the other way around. That is, we can not reproduce results. There are three main factors that influence reliability: the effectiveness measures, the size of our samples and the agreement between human annotators.

Two important characteristics of the effectiveness measures used in IR Evaluation are their stability and sensitivity. The results should be stable under different annotators and query sets, so the results do not vary significantly and alter the conclusions as to what systems are better (Buckley and Voorhees, 2000). They are also desired to discriminate between systems if they actually perform differently (Voorhees and Buckley, 2002; Sakai, 2007), and to do so with the minimum effort (Sanderson and Zobel, 2005). However, they are desired to not discriminate between systems that actually perform very similarly. These performance differences must always be considered in the context of the task and its user model.

In general, the more queries we use the more stable the results and therefore the more reliable, because we compute estimates closer to the true values. Estimating how many queries are "enough" to reach some level of reliability is a quite tedious process if following a data-based approach as (Buckley and Voorhees, 2000; Voorhees and Buckley, 2002; Sakai, 2007; Sanderson and Zobel, 2005; Urbano et al, 2011b). A simpler yet more powerful approach can be followed with *Generalizability Theory* (Bodoff and Li, 2007; Carterette et al, 2009; Salamon and Urbano, 2012). It allows to measure the stability of a test collection *while* it is being developed. It can also be used to estimate the stability of a different experimental design, or to estimate the point at which it is more reliable to employ more annotations and the current query set rather than just including more queries.

Given a set of systems and the resulting distributions obtained with different queries according to some system-measure, they are usually compared in terms of their mean effectiveness score. This can be problematic, because those means are just estimates of the true population means, and are therefore subject to random error due to sampling. Not until relatively recently, statistical methods have been systematically employed to compare systems by their score distribution rather than just their sample mean score (Carterette, 2012; Sakai, 2006; Carterette and Smucker, 2007; Webber et al, 2008a). It is also very important to study which statistical methods are more appropriate, because their assumptions are known to be violated in IR Evaluation (Smucker et al, 2007; Zobel, 1998). 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 (Urbano et al, 2012). Even if one system is found to be statistically significantly better than another one, the difference might be extremely small; too small to be noticed by users. On the other hand, the tiniest practical difference will turn out statistically significant with a sufficient number of queries.

6 Efficiency

Efficiency is the extent to which the experimenter reaches a valid and reliable result at a low cost (Trochim and Donnelly, 2007; Tague-Sutcliffe, 1992). Are there other annotation procedures and alternative evaluation methods that result in a more cost-effective experiment?

Annotations for test collections are usually made by experts, which increases the cost of building large datasets. Some recent work examined the use of non-experts for relevance judging (Bailey et al, 2008), and found that in general there are no noticeable differences in the evaluation results, although clear differences exist when the task is very specialized. Others explore the use of paid crowdsourcing platforms such as Amazon Mechanical Turk (Alonso and Mizzaro, 2012; Carvalho et al, 2010) to gather annotations for a very low cost. The problem in these cases is the potential low quality of the results. Some quality control techniques are based on known answers (Sanderson et al, 2010), redundant answers to compute consensus (Ipeirotis et al, 2010; Snow et al, 2008) or trying to detect neglecting behavior (Kittur et al, 2008; Urbano et al, 2011a; Rzeszotarski and Kittur, 2011).

Other research focused on the use of *incomplete* ground truth data where not all annotations are present in the test collections. A first approach to reduce the number of annotations in retrieval tasks was the pooling technique (Buckley and Voorhees, 2004). When evaluating a set of systems, annotating all documents retrieved by all systems is very expensive. Instead, a pool with the top-k results from all systems is formed, and only those are annotated; all documents outside the pool are then assumed to be non-relevant. This technique has been used in Text IR for many years, and it has been repeatedly shown to be reliable despite the non-relevance assumption, permitting the use of large collections by reducing the annotation cost to about 35%. With very large collections though, it is shown to have problems (Buckley et al, 2007). Different modifications of the basic pooling technique have been proposed via interactive annotation processes (Zobel, 1998; Cormack et al, 1998), meta-search models (Aslam et al, 2003), intelligent selection of documents to judge (Moffat et al, 2007) and ignoring them altogether (Buckley and Voorhees, 2004; Sakai and Kando, 2008). Other alternatives studied the evaluation of systems even when annotations are not available at all (Soboroff et al, 2001), which es useful as a lower bound on evaluation reliability.

More recent work has focused on the inference of annotations based on a *very* incomplete set of previous annotations, using a more probabilistic view of evaluation. Some techniques focus on sampling theory (Aslam and Yilmaz, 2007), document similarities (Carterette and Allan, 2007) or meta-search (Carterette, 2007). The inferred data are then used to estimate effectiveness scores based on random samples of annotations (Yilmaz and Aslam, 2006; Yilmaz et al, 2008); or to estimate the ranking of systems by annotating only those documents that are more informative to tell the difference between systems (Carterette et al, 2006; Carterette, 2007). These low-cost techniques have been studied mainly in the TREC Million Query Track between 2007 and 2009, offering very reliable results for a very low cost of annotation. In fact, they allowed a dramatic increase in the number of queries from a few dozens to over a thousand (Carterette et al, 2009).

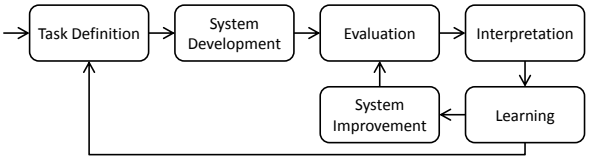

Fig. 4 The IR Research and Development cycle.

7 Challenges in Music IR Evaluation

Most research in Information Retrieval follows a cycle that ultimately leads to the development of better systems thanks to evaluation experiments (see Figure 4). First, a research problem is identified and an IR task is defined to evaluate different approaches to solve it. In the Development phase researchers build a new system for that task or adapt a previous one, and to assess how good it is they then go through an Evaluation phase. Once experiments are finished the Interpretation of results is carried out, 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 go through an Improvement phase to try and make their system better, going back over to the Evaluation phase. In some cases, and especially when the task is new, the first evaluation rounds lead to a re-definition of the task to better capture the real user scenario. Unfortunately, current evaluation practices in Music IR seem to fall short in this cycle, as we detail next.

Definition. To define a proper research problem, and thus a proper evaluation task, is the single most important part of carrying out a successful research project. Most often the definition of a task is the result of the research methodology being developed, not of the identification of a real industrial or user need. However, many tasks currently studied in Music IR do not really evaluate system or application level issues. In fact, MIREX tasks are often initiated by graduate students who build a dataset to support their work and then donate it to be used in MIREX.

Development. The task intent and its underlying user model are sometimes unclear or its real-world applicability uncertain. In Music IR many of the concepts used are either very subjective or very much context-dependent. Tasks to evaluate concepts such as melody, similarity, or emotion are developed without defining and taking into account the proper context in which the particular task has to be evaluated. The application use is not clearly defined in this phase.

Evaluation. One of the major problems in Music IR evaluations is the lack of proper test collections with which to run the experiments. Typically there is a mismatch between the final application of the task and the data used in the actual evaluations. Collections are often either too small or biased (e.g. same genre or time period), jeopardizing the external validity of the results. Moreover, the lack of standardized and public collections results in researchers using their personal, private, often insufficiently described and rarely analyzed collections, which prevents other researchers from comparing systems or validating and replicating results, hindering the overall development of the field and often leading to misleading conclusions. There is also a lack of sufficiently standard evaluation procedures and software tools available to and used by the majority of researchers.

Interpretation. Given the subjectivity of many of the tasks evaluated in Music IR, the effectiveness measures used tend to be very particular, subjective or without a clear user model. Many measures are task-dependent, and without proper analysis it is unclear how they behave. Other measures are not documented, and they are reported without description, references or source code, making them impossible to interpret or use in private evaluations (e.g. *Normalized Recall at Group Boundaries*, used in *Symbolic Melodic Similarity*). Also, widely-accepted baseline systems are very rarely included in evaluations, and when they are, they are often not strong enough or 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 Music IR, 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, most often, incorrect interpretation.

Learning. The goal of performing an evaluation experiment is to learn form it, but in most Music IR evaluation frameworks this is barely possible. For example, the raw musical material is usually not available to participants, and the actual queries used are unknown. Even in some cases only the average scores are reported, so researchers can not analyze performance on a per query basis: if they had very bad results in some cases 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. This issue clearly differentiates Music IR from Text IR: new text information is created at a dramatic rate, for example in the Internet. This makes most text corpora easy to obtain, but this does not hold for music repositories. Indeed, copyright restrictions are a huge problem that hinders the creation of public music evaluation corpora of wide acceptance.

Improvement. Once we learn from the evaluation we should be able to improve the system being developed. This is hardly possible in some MIREX tasks mainly because the test collections are not reusable, and in most cases they are just not publicly available. As such, researchers have no option but to blindly improve their systems and wait for another MIREX round, with no way of comparing cross-edition results due to the lack of data and proper baselines. The recent Million Song Dataset Challenge and the MusiClef benchmarking campaign alleviate this problem by providing a more open experimentation and evaluation setting, highly encouraging the use of multimodal material related to music.

8 Opportunities in Music IR Evaluation

Although not easily, the mentioned shortcomings of current evaluation practices in Music IR can be overcome. To this end, we list several proposals to ease the way through the IR Research and Development cycle. However, we have to first mention that Music IR does not only encompass the traditional ad-hoc retrieval setting, where the input is a query and the output is a list of items sorted according to some relevance model that ideally corresponds to the user perceived match to his or her information or entertainment need. There are different ways to access music collections (Schedl et al, 2012), such as direct querying, query-by-example, browsing, music recommendation, metadata-based search, etc. Direct querying means

that the query is given in the same form as the features computed from the audio, for instance, melodic search via notes as input when the features are actual MIDI representations of the score. Query-by-example refers to the case where the input is a (possibly short and noisy) audio representation of a music piece. A popular commercial system supporting this kind of query is Shazam[11]. Browsing refers to the process of digging into music collections, sifting through them, and exploring them, typically to find interesting new songs. Also intended to find interesting and novel songs, music recommendation is given as query the listening history of the user (sometimes enhanced with demographic data), and should ideally provide serendipitous music recommendations. Eventually, in metadata-based search query consists of a text description such an artist or song name, and the system is supposed to return pieces whose editorial metadata match the query.

Corpora. IR evaluation needs large corpora if we pursue external validity and generalization of results (Voorhees, 2002a). We need to go 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 corpus open to other researchers copyright issues immediately arise (Downie, 2004). A possibility is to publish feature vectors and metadata, such as in the recent Million Song Dataset (Bertin-Mahieux et al, 2011). However, in this case features were computed with algorithms that are not open, and it still poses problems if researchers want to develop a new audio feature or analyze specific items for which their system worked better or worse in previous runs (Rauber et al, 2012). These corpora should be standardized so they can be used *throughout the community* and *across tasks*. This would allow us to compare and better understand improvements between systems and tasks, besides offering clear advantages in terms of distribution, licensing, etc. We admit though that if tasks are too heterogeneous, using only one music corpus for all of tasks is infeasible. We hence suggest a more holistic view on corpus generation. Taking into account the multimodality of possible representations and descriptions of music, such as editorial meta-data, symbolic MIDI, signal-based features, collaborative tags, playlist co-occurrences, music video clips, and even images of album covers or band photographs; we should opt likewise to establish multimodal music corpora.

Raw audio data. For most Music IR tasks we need shared access to the encoded audio signals of the music corpora used. An alternative to closed commercial corpora that cannot be openly distributed is to use music free of copyright restrictions, such as music provided by services like Jamendo[12], the Internet Archive[13] or the RWC database (Goto et al, 2003). However, this may potentially introduce threats to external validity that are subject to study. Despite this possible bias, copyright-free music is a perfectly viable alternative for many tasks, so we should seek the collaboration of free content providers to put together *controlled* corpora to distribute throughout the community. In this line, the use of artificial material such as synthesized or error-mutated queries (Niedermayer et al, 2011), or the use of clips instead of full songs (Salamon and Urbano, 2012), should be reconsidered.

Annotations. Music IR researchers are used to evaluating their algorithms in MIREX with collections that are not publicly available. This has been justified

[11] http://www.shazam.com
[12] http://www.jamendo.com
[13] http://archive.org

by copyright issues on the audio corpora and by the need to hide the annotations so that researchers are prevented to cheat or overfit. But only with performance scores there is really no way to improve systems and analyze results to know why they work or why they fail. MIREX is sometimes considered a *contest* instead of a collaborative evaluation forum, and the unavailability of annotations surely is the prime cause: avoiding cheating and overfitting is perceived by participants as the necessary requirement for a contest. On the other hand, collections are usually built by individuals or concrete research groups. Given that annotations for Music IR tasks can be quite expensive, researchers are understandably reluctant to share their annotations because it gives them an edge over the rest. But it is very important to realize that this situation does not benefit the community as a whole. In our view, further fostering research based on private data should be discouraged because it is impossible to analyze validity and reliability, besides breaking two pillars of Science: repeatability and reproducibility. We should promote collaborative efforts to *incrementally* build publicly accessible datasets, by and for the community, employing the low-cost techniques described in Section 6. In the meantime, collections that are apparently larger than needed for reliable evaluation may be split in a public half for training and a hidden half for testing (Salamon and Urbano, 2012)

Raw evaluation data. In order to improve our evaluation frameworks we need to share as much of the generated data as possible. The raw, unedited system output is a very valuable resource for IR Evaluation research, as it allows us to investigate "what-if" alternative evaluation scenarios and possible improvements of the evaluation process itself (Zobel et al, 2011). Making all these data publicly available would undoubtedly boost this research. We emphasize the need for *raw* data. For instance, if a system returns a list of 50 items but only the top 5 are evaluated, we should still publish the full list with all 50 items. Likewise, if a system returns an annotation every 5ms, we should make all these annotations available, not just one every 20ms or so. Asking for these data in particular papers surely is unrealistic, but it should be immediate in community evaluations like MIREX.

Evaluation model. Having publicly accessible and standardized corpora would allow for a change in the execution model currently employed in MIREX. Researchers should be in charge of executing their systems and producing the runs to submit back to MIREX, relieving the IMIRSEL group from a great deal of workload and motivating researchers reluctant to give their algorithms away to third parties. This data-to-algorithm model is followed by the recent Million Song Dataset Challenge and MusiClef campaigns, and in our view it is in fact 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 in the current situation where IMIRSEL has finally stopped receiving funds. The community is exploring alternatives (Page et al, 2012; Mayer and Rauber, 2012), like providing automated online platforms that allow researchers to run batch experiments on demand, such as in MIREX-DIY (Do-It-Yourself) (Ehmann et al, 2007). However, this evaluation model would still not permit a full execution of the IR cycle because of the lack of fully accessible data. While this could probably help researchers in improving their systems, for the yearly MIREX rounds we suggest that decentralization goes a step further, where participants run their systems and submit their *raw* output to a third party that scores the systems.

Organization. The current organization of MIREX rests heavily on the IMIRSEL team, who plan, schedule and run a good number of tasks every year. An alternative based on two organization tiers was proposed in ISMIR 2011 and further discussed during ISMIR 2012. This additional tier should be task-dependent and comprise third-party leading researchers. These organizers or task leaders would deal with all the logistics, planning, evaluation, troubleshooting, etc. As of now IMIRSEL is responsible for almost everything involved in running a MIREX task, so developing these tasks year after year to make them more challenging is hardly expected because of the work required. Adopting a second tier of task-dependent organizers would diminish the workload in 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. Annual rounds of MIREX use to just replicate whatever task designs and datasets were used in previous years, which clearly limits the development and improvement of algorithms and discourages researchers to participate in increasingly unchallenging tasks. In our view, this effect is to some extent beginning to appear in MIREX.

Overview publications. The inclusion of task organizers would also benefit the community if by the end of each MIREX edition they published an overview paper thoroughly detailing the evaluation process followed, data, results and, most importantly, discussion to boost the Interpretation and Learning phases of the IR cycle. Such a publication would be the perfect wrap-up to the extended abstracts where participants describe their systems but very rarely investigate and elaborate on the results (Cunningham et al, 2012). In fact, many of these participant-papers are not even drafted. The current work overload in IMIRSEL does not help at all in this matter. A sign of this can be found in the year-specific web pages describing the tasks[8], which use to be just replicates from previous years and hardly ever reflect the changes introduced, which then go undocumented and produce erroneous interpretations and conclusions. Task overview publications are a very valuable source of information in other forums such as TREC. As Figure 1 shows, these papers receive about twice as many citations as regular evaluation papers, suggesting that they have an impact not only on evaluation research but also on the wider audience. Unfortunately, this kind of publications do not exist in MIREX. The result is that many of the changes introduced are communicated in a word of mouth manner and erroneous information is still up online for the unaware reader.

Specific methodologies. Both new methodologies and effectiveness measures have been proposed for Music IR tasks (e.g. (Typke et al, 2005; Urbano et al, 2010a; Hu et al, 2008; Downie et al, 2008; Typke et al, 2006; Poliner et al, 2007)). Following the principles of the work described in Sections 4 and 5, we need to study the extent to which they are valid and reliable. Some work has studied the reduction of effort needed to annotate through the use of crowdsourcing platforms (Urbano et al, 2010b; Lee, 2010) or games (Law et al, 2007), and further studies should follow this line, given the usual restrictions the Music IR field has to face with respect to availability of resources. Another line is the study of human effects on ground truth data, evaluation results and task design (Järvelin, 2011; Jones et al, 2007; Schedl and Flexer, 2012). Additionally, the low-cost evaluation techniques mentioned in Section 6 should definitely be studied for the wealth of Music IR tasks. An example

has already been proposed for *Audio Music Similarity*, reducing annotation cost to less than 5% (Urbano and Schedl, 2013).

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 with 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. In fact, the suggested release of raw evaluation data from MIREX would allow researchers to use strong baselines to compare their systems with and publish their results. It is very important that we agree on the use of strong baseline systems and compare private results with the best annual figures provided by evaluation forums like MIREX, because authors often publish improvements over weak baselines that in reality do not outperform the stronger and well-known baselines (Armstrong et al, 2009).

Software standardization. It is not rare to find published results that are incorrect because of software bugs. With the development and wide acceptance of a common 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 call for the standardization of data formats to speed up the IR cycle; and serve as explicit documentation of the evaluation measures and processes used by the community, for the implementation of some details is unknown or subject to different interpretations. These tools are available and widely used by the Text IR community, such as the `trec_eval`[14] and `ntcireval`[15] packages. In fact, a close look at the source code of these two tools reveals that they follow different implementations for heavily used measures like *Average Precision*. This is a clear example of the paramount need to standardize evaluation tools, and require their use and explicit mention in publications if we strive for robust and repeatable research. If not, one paper might claim improvements over another paper that are in reality attributed to different evaluation software. Another clear example is the measure *DCG*, widely used in Web Retrieval and Learning to Rank. The formulation of this measure has suffered several modification with the years, and while the de facto formulation follows the principles of the original one, they are quite different in reality. Nevertheless, the original publication (Järvelin and Kekäläinen, 2002) keeps being cited despite the actual measure used is different. A novice researcher may easily be unaware of this generally undocumented practice.

Openness towards other communities. If we want to succeed in actively pushing forward the Music IR field, in particular evaluation aspects, we should seek discussion and collaboration with related communities, such as traditional Text IR, Multimedia IR, Signal Processing and Recommendation Systems. There is certainly an interest in Music IR from researchers in other communities, as initiatives such as the Million Song Dataset Challenge or the KDD Cup 2011[16] on music recommendation showed. One possible way to position Music IR as a prominent and interesting field among related communities is to establish multimodal corpora and run multimodal evaluation tasks accessible to the wider non-music audience. In addition, a common criticism of Music IR work from other fields such

[14] http://trec.nist.gov/trec_eval/

[15] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

[16] http://www.sigkdd.org/kdd2011/kddcup.shtml

as Text IR is that our evaluations do not meet the standards of venues like TREC and SIGIR. This is further evidence that we need to push this issue forward.

Commitment of the community. In general, the current problems in Music IR Evaluation need to be fully acknowledged by researchers. Now that we have a well-established evaluation forum like MIREX, we need to start questioning the validity, reliability and efficiency of the experiments, with the sole purpose of making it better and more challenging. Current Music IR experiments seem to stop at the Evaluation phase of the IR cycle, but the subsequent Interpretation and Learning phases are often ignored or impossible to engage into. MIREX should not only be a place to evaluate our systems, but a place to *improve how we evaluate* those systems; it needs to be a place to experiment with alternative evaluation methods and validate the current ones. This endeavor is the responsibility of not only MIREX and similar campaigns, but of the whole ISMIR community.

Support from the ISMIR society. We believe the ISMIR society should provide organizational and financial support for the development of test collections following the above proposals. In the financial side, it was proposed during ISMIR 2012 to slightly increase the ISMIR registration fee, maybe voluntarily, so that by the end of each conference there are some funds to hire annotators. With the low-cost evaluation techniques mentioned in Section 6, the annotation effort can be greatly reduced, with the possibility of incrementally adding new annotations when necessary, by hired annotators or members of the community. In the organizational side, the home `http://www.ismir.net` website can be the home to a centralized repository of test collections built in this manner. If they are thoroughly designed, described and controlled (Peeters and Fort, 2012), we will reach a point where evaluation resources are publicly available to all the community. Finally, a centralized repository of publicly available systems would make it much easier for researchers to include widely accepted baselines in their experiments.

9 Conclusions

Evaluation is a very important area of research in Information Retrieval that has received a lot of attention in recent years. However, it seems that the Music IR field has not been, until very recently, aware of the need to analyze the evaluation frameworks used. 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 experimental validity, reliability and efficiency, we show different aspects of IR Evaluation that have been overlooked and need special attention in Music IR. This survey is intended as a start point for the Music IR community to engage in this research topic and begin a hopefully fruitful tradition in ISMIR.

From the point of view of the IR Research and Development cycle a researcher follows, we have also shown that current evaluation practices do not allow us to fully carry out our research activity. Evaluation experiments produce great 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, we make several proposals to improve the situation and engage researchers in these last phases of the cycle, which should ultimately lead to a more rapid development of the field.

References

Al-Maskari A, Sanderson M, Clough P (2007) The Relationship between IR Effectiveness Measures and User Satisfaction. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 773–774

Allan J, Croft B (2003) Challenges in Information Retrieval and Language Modeling. ACM SIGIR Forum 37(1):31–47

Allan J, Carterette B, Lewis J (2005) When Will Information Retrieval Be 'Good Enough'? In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 433–440

Allan J, Croft B, Moffat A, Sanderson M (2012) Frontiers, Challenges and Opportunities for Information Retrieval: Report from SWIRL 2012. ACM SIGIR Forum 46(1):2–32

Alonso O, Mizzaro S (2012) Using Crowdsourcing for TREC Relevance Assessment. Information Processing and Management 48(6):1053–1066

Armstrong TG, Moffat A, Webber W, Zobel J (2009) Improvements that Dont Add Up: Ad-Hoc Retrieval Results since 1998. In: ACM International Conference on Information and Knowledge Management, pp 601–610

Aslam JA, Yilmaz E (2007) Inferring Document Relevance from Incomplete Information. In: ACM International Conference on Information and Knowledge Management, pp 633–642

Aslam JA, Pavlu V, Savell R (2003) A Unified Model for Metasearch, Pooling and System Evaluation. In: ACM International Conference on Information and Knowledge Management, pp 484–491

Bailey P, Craswell N, Soboroff I, Thomas P, de Vries AP, Yilmaz E (2008) Relevance Assessment: Are Judges Exchangeable and Does it Matter? In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 667–674

Bennett PN, Carterette B, Chapelle O, Joachims T (2008) Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments. ACM SIGIR Forum 42(2):53–58

Bertin-Mahieux T, Ellis DP, Whitman B, Lamere P (2011) The Million Song Dataset. In: International Society for Music Information Retrieval Conference

Bodoff D, Li P (2007) Test Theory for Assessing IR Test Collections. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 367–374

Buckley C, Voorhees EM (2000) Evaluating Evaluation Measure Stability. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 33–34

Buckley C, Voorhees EM (2004) Retrieval Evaluation with Incomplete Information. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 25–32

Buckley C, Dimmick D, Soboroff I, Voorhees EM (2007) Bias and the Limits of Pooling for Large Collections. Journal of Information Retrieval 10(6):491–508

Cano P, Gómez E, Gouyon F, Herrera P, Koppenberger M, Ong B, Serra X, Streich S, Wack N (2006) ISMIR 2004 Audio Description Contest. Tech. Rep. MTG-TR-2006-02, Universitat Pompeu Fabra

Carterette B (2007) Robust Test Collections for Retrieval Evaluation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 55–62

Carterette B (2011) System Effectiveness, User Models, and User Utility: A General Framework for Investigation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 903–912

Carterette B (2012) Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM Transactions on Information Systems 30(1)

Carterette B, Allan J (2007) Semiautomatic Evaluation of Retrieval Systems using Document Similarities. In: ACM International Conference on Information and Knowledge Management, pp 873–876

Carterette B, Smucker MD (2007) Hypothesis Testing with Incomplete Relevance Judgments. In: ACM International Conference on Information and Knowledge Management, pp 643–652

Carterette B, Soboroff I (2010) The Effect of Assessor Error on IR System Evaluation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 539–546

Carterette B, Allan J, Sitaraman R (2006) Minimal Test Collections for Retrieval Evaluation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 268–275

Carterette B, Pavlu V, Kanoulas E, Aslam JA, Allan J (2009) If I Had a Million Queries. In: European Conference on Information Retrieval, pp 288–300

Carterette B, Gabrilovich E, Josifovski V, Metzler D (2010a) Measuring the Reusability of Test Collections. In: ACM International Conference on Web Search and Data Mining, pp 231–240

Carterette B, Kanoulas E, Pavlu V, Fang H (2010b) Reusable Test Collections Through Experimental Design. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 547–554

Carvalho VR, Lease M, Yilmaz E (2010) Crowdsourcing for Search Evaluation. ACM SIGIR Forum 44(2):17–22

Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected Reciprocal Rank for Graded Relevance. In: ACM International Conference on Information and Knowledge Management, pp 621–630

Cleverdon CW (1991) The Significance of the Cranfield Tests on Index Languages. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3–12

Cormack GV, Lynam TR (2006) Statistical Precision of Information Retrieval Evaluation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 533–540

Cormack GV, Palmer CR, Clarke CL (1998) Efficient Construction of Large Test Collections. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 282–289

Cunningham SJ, Bainbridge D, Downie JS (2012) The Impact of MIREX on Scholarly Research (2005-2010). In: International Society for Music Information Retrieval Conference, pp 259–264

Downie JS (2002) Interim Report on Establishing MIR/MDL Evaluation Frameworks: Commentary on Consensus Building. In: ISMIR Panel on Music Information Retrieval Evaluation Frameworks, pp 43–44

Downie JS (2003) The MIR/MDL Evaluation Project White Paper Collection, 3rd edn. URL http://www.music-ir.org/evaluation/wp.html

Downie JS (2004) The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. Computer Music Journal 28(2):12–23

Downie JS, Bay M, Ehmann AF, Jones MC (2008) Audio Cover Song Identification: MIREX 2006-2007 Results and Analysis. In: International Conference on Music Information Retrieval

Downie JS, Ehmann AF, Bay M, Jones MC (2010) The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. In: Zbigniew WR, Wieczorkowska AA (eds) Advances in Music Information Retrieval, Springer, pp 93–115

Ehmann AF, Downie JS, Jones MC (2007) The Music Information Retrieval Evaluation eXchange "Do-It-Yourself" Web Service. In: International Conference on Music Information Retrieval, pp 323–324

Geman S, Bienenstock E, Doursat R (1992) Neural Networks and the Bias/Variance Dilema. Neural Computation 4(1):1–58

Goto M, Hashiguchi H, Nishimura T, Oka R (2003) RWC Music Database: Popular, Classical and Jazz Music Databases. In: International Conference on Music Information Retrieval, pp 287–288

Guiver J, Mizzaro S, Robertson S (2009) A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation. ACM Transactions on Information Systems 27(4):1–26

Harman DK (2011) Information Retrieval Evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services 3(2):1–119

Hersh W, Turpin A, Price S, Chan B, Kraemer D, Sacherek L, Olson D (2000) Do Batch and User Evaluations Give the Same Results? In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 17–24

Hu X, Kando N (2012) User-Centered Measures vs. System Effectiveness in Finding Similar Songs. In: International Society for Music Information Retrieval Conference, pp 331–336

Hu X, Downie JS, Laurier C, Bay M, Ehmann AF (2008) The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: International Conference on Music Information Retrieval

Huffman SB, Hochster M (2007) How Well does Result Relevance Predict Session Satisfaction? In: International ACM SIGIR Conference on Research and Development in Information

Retrieval, pp 567–573
Ipeirotis PG, Provost F, Wang J (2010) Quality Management on Amazon Mechanical Turk. In: ACM SIGKDD Workshop on Human Computation, pp 64–67
Järvelin K (2011) IR Research: Systems, Interaction, Evaluation and Theories. ACM SIGIR Forum 45(2):17–31
Järvelin K, Kekäläinen J (2002) Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems 20(4):422–446
Jones MC, Downie JS, Ehmann AF (2007) Human Similarity Judgments: Implications for the Design of Formal Evaluations. In: International Conference on Music Information Retrieval, pp 539–542
Kanoulas E, Aslam JA (2009) Empirical Justification of the Gain and Discount Function for nDCG. In: ACM International Conference on Information and Knowledge Management, pp 611–620
Kekäläinen J (2005) Binary and Graded Relevance in IR Evaluations: Comparison of the Effects on Ranking of IR Systems. Information Processing and Management 41(5):1019–1033
Kittur A, Chi EH, Suh B (2008) Crowdsourcing User Studies With Mechanical Turk. In: Annual ACM SIGCHI Conference on Human Factors in Computing Systems, pp 453–456
Lancaster F (1968) Evaluation of the MEDLARS Demand Search Service. Tech. rep., U.S. Department of Health, Education, and Welfare
Lartillot O, Miotto R, Montecchio N, Orio N, Rizo D, Schedl M (2011) MusiClef: A Benchmark Activity in Multimodal Music Information Retrieval. In: International Society for Music Information Retrieval Conference
Law EL, von Ahn L, Dannenberg RB, Crawford M (2007) TagATune: A Game for Music and Sound Annotation. In: International Conference on Music Information Retrieval, pp 361–364
Lee JH (2010) Crowdsourcing Music Similarity Judgments using Mechanical Turk. In: International Society for Music Information Retrieval Conference, pp 183–188
Lehmann E, Casella G (1998) Theory of Point Estimation. Springer
Lesk M, Harman DK, Fox EA, Wu H, Buckley C (1997) The SMART Lab Report. ACM SIGIR Forum 31(1):2–22
Mayer R, Rauber A (2012) Towards Time-Resilient MIR Processes. In: International Society for Music Information Retrieval Conference, pp 337–342
McFee B, Bertin-Mahieux T, Ellis DP, Lanckriet G (2012) The Million Song Dataset Challenge. In: WWW International Workshop on Advances in Music Information Research, pp 909–916
Moffat A, Zobel J (2008) Rank-Biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems 27(1)
Moffat A, Zobel J, Hawking D (2005) Recommended Reading for IR Research Students. ACM SIGIR Forum 39(2):3–14
Moffat A, Webber W, Zobel J (2007) Strategic System Comparisons via Targeted Relevance Judgments. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 375–382
Niedermayer B, Widmer G, Böck S (2011) On the Importance of Real Audio Data for MIR Algorithm Evaluation at the Note-Level: A comparative Study. In: International Society for Music Information Retrieval Conference
Page K, Fields B, de Roure D, Crawford T, Downie JS (2012) Reuse, Remix, Repeat: the Workflows of MIR. In: International Society for Music Information Retrieval Conference, pp 409–414
Peeters G, Fort K (2012) Towards a (Better) Definition of the Description of Annotated MIR Corpora. In: International Society for Music Information Retrieval Conference, pp 25–30
Peeters G, Urbano J, Jones GJ (2012) Notes from the ISMIR 2012 Late-Breaking Session on Evaluation in Music Information Retrieval. In: International Society for Music Information Retrieval Conference
Poibeau T, Kosseim L (2001) Proper Name Extraction from Non-Journalistic Texts. Language and Computers - Studies in Practical Linguistics 37:144–157
Poliner GE, Ellis DP, Ehmann AF, Gómez E, Streich S, Ong B (2007) Melody Transcription From Music Audio: Approaches and Evaluation. IEEE Transactions on Audio, Speech and Language Processing 15(4):1247–1256
Rauber A, Schindler A, Mayer R (2012) Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset. In: International Society for Music Information Retrieval

Conference, pp 469–474
Robertson S (2008) On the History of Evaluation in IR. Journal of Information Science 34(4):439–456
Robertson S (2011) On the Contributions of Topics to System Evaluation. In: European Conference on Information Retrieval, pp 129–140
Robertson S, Kanoulas E, Yilmaz E (2010) Extending Average Precision to Graded Relevance Judgments. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 603–610
Rzeszotarski J, Kittur A (2011) Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In: ACM Symposium on User Interface Software and Technology
Sakai T (2006) Evaluating Evaluation Metrics Based on the Bootstrap. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 525–532
Sakai T (2007) On the Reliability of Information Retrieval Metrics Based on Graded Relevance. Information Processing and Management 43(2):531–548
Sakai T, Kando N (2008) On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments. Journal of Information Retrieval 11(5):447–470
Salamon J, Urbano J (2012) Current Challenges in the Evaluation of Predominant Melody Extraction Algorithms. In: International Society for Music Information Retrieval Conference, pp 289–294
Sanderson M (2010) Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval 4(4):247–375
Sanderson M, Zobel J (2005) Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 162–169
Sanderson M, Paramita ML, Clough P, Kanoulas E (2010) Do User Preferences and Evaluation Measures Line Up? In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 555–562
Saracevic T (1995) Evaluation of Evaluation in Information Retrieval. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 138–146
Schamber L (1994) Relevance and Information Behavior. Annual Review of Information Science and Technology 29:3–48
Schedl M, Flexer A (2012) Putting the User in the Center of Music Information Retrieval. In: International Society for Music Information Retrieval Conference, pp 385–390
Schedl M, Stober S, Gómez E, Orio N, Liem CC (2012) User-Aware Music Retrieval. In: Müller M, Goto M, Schedl M (eds) Multimodal Music Processing, Dagstuhl Publishing, pp 135–156
Scholer F, Turpin A (2008) Relevance Thresholds in System Evaluations. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 693–694
Shadish WR, Cook TD, Campbell DT (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin
Smucker MD, Clarke CL (2012a) The Fault, Dear Researchers, is Not in Cranfield, But in Our Metrics, that They Are Unrealistic. In: European Workshop on Human-Computer Interaction and Information Retrieval, pp 11–12
Smucker MD, Clarke CL (2012b) Time-Based Calibration of Effectiveness Measures. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 95–104
Smucker MD, Allan J, Carterette B (2007) A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In: ACM International Conference on Information and Knowledge Management, pp 623–632
Snow R, OConnor B, Jurafsky D, Ng AY (2008) Cheap and FastBut is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In: Conference on Empirical Methods in Natural Language Processing, pp 254–263
Soboroff I, Nicholas C, Cahan P (2001) Ranking Retrieval Systems Without Relevance Judgments. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 66–73
Tague-Sutcliffe J (1992) The Pragmatics of Information Retrieval Experimentation, Revisited. Information Processing and Management 28(4):467–490
Taylor JR (1997) An Introduction Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books

Trochim WM, Donnelly JP (2007) The Research Methods Knowledge Base, 3rd edn. Atomic Dog Publishing

Turpin A, Hersh W (2001) Why Batch and User Evaluations Do Not Give the Same Results. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 225–231

Typke R, den Hoed M, de Nooijer J, Wiering F, Veltkamp RC (2005) A Ground Truth for Half a Million Musical Incipits. Journal of Digital Information Management 3(1):34–39

Typke R, Veltkamp RC, Wiering F (2006) A Measure for Evaluating Retrieval Techniques based on Partially Ordered Ground Truth Lists. In: IEEE International Conference on Multimedia and Expo, pp 1793–1796

Urbano J, Schedl M (2013) Minimal Test Collections for Low-Cost Evaluation of Audio Music Similarity and Retrieval Systems. International Journal of Multimedia Information Retrieval 2(1):59–70

Urbano J, Marrero M, Martín D, Lloréns J (2010a) Improving the Generation of Ground Truths based on Partially Ordered Lists. In: International Society for Music Information Retrieval Conference, pp 285–290

Urbano J, Morato J, Marrero M, Martín D (2010b) Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks. In: ACM SIGIR Workshop on Crowdsourcing for Search Evaluation, pp 9–16

Urbano J, Marrero M, Martín D, Morato J, Robles K, Lloréns J (2011a) The University Carlos III of Madrid at TREC 2011 Crowdsourcing Track. In: Text REtrieval Conference

Urbano J, Martín D, Marrero M, Morato J (2011b) Audio Music Similarity and Retrieval: Evaluation Power and Stability. In: International Society for Music Information Retrieval Conference, pp 597–602

Urbano J, Downie JS, Mcfee B, Schedl M (2012) How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval. In: International Society for Music Information Retrieval Conference, pp 181–186

Voorhees EM (2000) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. Information Processing and Management 36(5):697–716

Voorhees EM (2001) Evaluation by Highly Relevant Documents. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 74–82

Voorhees EM (2002a) The Philosophy of Information Retrieval Evaluation. In: Workshop of the Cross-Language Evaluation Forum, pp 355–370

Voorhees EM (2002b) Whither Music IR Evaluation Infrastructure: Lessons to be Learned from TREC. In: JCDL Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation, pp 7–13

Voorhees EM, Buckley C (2002) The Effect of Topic Set Size on Retrieval Experiment Error. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 316–323

Voorhees EM, Harman DK (2005) TREC: Experiment and Evaluation in Information Retrieval. MIT Press

Webber W, Moffat A, Zobel J (2008a) Statistical Power in Retrieval Experimentation. In: ACM International Conference on Information and Knowledge Management, pp 571–580

Webber W, Moffat A, Zobel J, Sakai T (2008b) Precision-At-Ten Considered Redundant. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 695–696

Yilmaz E, Aslam JA (2006) Estimating Average Precision with Incomplete and Imperfect Information. In: ACM International Conference on Information and Knowledge Management, pp 102–111

Yilmaz E, Kanoulas E, Aslam JA (2008) A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 603–610

Zobel J (1998) How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 307–314

Zobel J, Webber W, Sanderson M, Moffat A (2011) Principles for Robust Evaluation Infrastructure. In: ACM CIKM Workshop on Data infrastructures for Supporting Information Retrieval Evaluation