Towards Minimal Test Collections for Evaluation of Audio Music Similarity and Retrieval

Julián Urbano
University Carlos III of Madrid
Department of Computer Science
Leganés, Spain

jurbano@inf.uc3m.es

Markus Schedl
Johannes Kepler University
Department of Computational Perception
Linz, Austria

markus.schedl@jku.at

ABSTRACT

Reliable evaluation of Information Retrieval systems requires large amounts of relevance judgments. Making these annotations is quite complex and tedious for many Music Information Retrieval tasks, so performing such evaluations requires too much effort. A low-cost alternative is the application of Minimal Test Collection algorithms, which offer quite reliable results while significantly reducing the annotation effort. The idea is to incrementally select what documents to judge so that we can compute estimates of the effectiveness differences between systems with a certain degree of confidence. In this paper we show a first approach towards its application to the evaluation of the Audio Music Similarity and Retrieval task, run by the annual MIREX evaluation campaign. An analysis with the MIREX 2011 data shows that the judging effort can be reduced to about 35% to obtain results with 95% confidence.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation/ methodology; H.3.3 [**Information Search and Retrieval**]; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness).

General Terms

Algorithms, Experimentation, Measurement, Performance.

Keywords

Music information retrieval, evaluation, test collections, relevance judgments.

1. INTRODUCTION

The evaluation of Information Retrieval (IR) systems requires a test collection, usually containing a set of documents, a set of task-specific queries, and a set of annotations that provide information as to what results a system should return for each query. Depending on the task, the set of queries may comprise the collection of documents itself, and the type of annotations can differ widely. In the field of Music IR (MIR), building these collections is very problematic due to the very nature of the musical information, legal restrictions upon the documents, etc. [4]. In addition, annotating a test collection is a very time-consuming and expensive process for some MIR tasks. For instance, annotating a single clip for Melody Extraction can take

several hours. As a result, test collections for MIR tasks use to be very small, and they are unlikely to change from year to year, posing serious problems for the proper evolution of the field [6].

The Music Information Retrieval Evaluation eXchange (MIREX) started in 2005 as an international venue to promote and perform evaluation of MIR systems for various tasks [5]. MIREX was developed following the principles and methodologies that have made the Text REtrieval Conference (TREC) such a successful forum for evaluating Text IR systems [9]. However, since its inception in 2005, the MIREX campaigns have evolved in parallel to TREC, practically ignoring all recent developments in the evaluation of IR systems [6]. In fact, the last five years have witnessed several works on reliable and low-cost evaluation of IR systems. One of these works is the development of methodologies for evaluation with Minimal Test Collections (MTC) [3][2].

The idea behind MTC is that the results of evaluating IR systems may be estimated with high confidence even if the set of annotations is very incomplete. In a typical Text IR setting, it means that we do not need to judge all documents retrieved for a topic, but only a small fraction of it, to estimate with high confidence which of two systems is better. In this paper we study the application of MTC to the evaluation of Audio Music Similarity and Retrieval (AMS) systems, as it is the task that most closely resembles the ad hoc Text IR scenario: for a given audio clip (the query), an AMS system returns a list of music pieces deemed to be similar to it. AMS is one of the most important tasks in MIR, and it has been run in MIREX in five of the seven editions so far (see Table 1).

Year	Teams	Systems	Queries	Results	Judgments	Overlap
2006	5	6	60	1,800	1,629	10%
2007	8	12	100	6,000	4,832	19%
2009	9	15	100	7,500	6,732	10%
2010	5	8	100	4,000	2,737	32%
2011	10	18	100	9,000	6,322	30%

Table 1. Summary of MIREX AMS editions.

Each edition of the AMS task requires the work of dozens of volunteers to perform similarity judgments, telling how similar two 30 second audio clips are. In the last edition, in 2011, 6,322 of these judgments were needed, meaning that at least 53 hours of assessor time were needed to complete the judging task. In practice, though, collecting all these judgments takes several days. But AMS is one of the couple of tasks for which a new set of queries and relevance judgments are put together every year. Most of the tasks just use the same collections over and over again because they are too expensive to build, especially in terms of judging or annotation effort. Therefore, the study of low-cost

evaluation methodologies is imperative for the development of proper test-collections to reliably evaluate MIR systems and properly advance the state of the art [6].

Developing low-cost evaluation methodologies is essential for private, in-house evaluations too. A researcher investigating several improvements of an existing MIR technique is not really interested in knowing how well they perform for the task (which is highly dependent on the test collection anyway), but in which one performs better. That is, she is interested in the *comparative* evaluation of systems. MTC is specifically designed for these cases: it minimizes the annotation effort needed to find a difference between systems, incrementally selecting for judging those documents that are more informative to figure out the difference between systems, and reusing previous judgments when available.

The remainder of the paper is organized as follows. Section 2 details the methodology currently followed to evaluate AMS systems in MIREX. Section 3 develops the methodology based on MTC to evaluate with incomplete judgments, and Section 4 shows the main results. Section 5 discusses the estimation of significance of the system comparisons and Section 6 concludes with final remarks and lines for future work.

2. AMS EVALUATION

Audio Music Similarity systems are evaluated according to an effectiveness measure that assesses how well they would satisfy a user for a given query. In order to generalize the results of an evaluation experiment to an arbitrary query, the MIREX evaluations use a random sample Q of 100 queries. Each system is run for every query, returning a list of all documents in the collection $\mathcal{D}$, ranked by their similarity to the query. The effectiveness measure used in MIREX is Average Gain of the top k documents retrieved ($AG@k$), with $k = 5$. For an arbitrary system A, $AG@k$ is defined as:

$$AG@k = \frac{1}{k}\sum_{i\in\mathcal{D}} G_i \cdot I(A_i \leq k)$$

where G_i is the gain of document i, A_i is the rank at which system A retrieved document i, and $I(x)$ is a boolean indicator function that evaluates to 1 if the expression x is true and to 0 otherwise. Therefore, the summation adds the gain of all documents in the collection that were ranked by A in the top k.

The gain of a document is a measure of how much information the user will gain from inspecting that result. In MIREX, there are two different scales: the BROAD scale is a 3-point graded scale where a document is considered either not similar to the query (gain 0), somewhat similar (gain 1) or very similar (gain 2); and the FINE scale, where the gain of a document ranges from 0 (not similar at all) to 100 (identical to the query)[1]. These gain scores are assessed by humans, who make similarity judgments between queries and documents. After all the judging is done, every system gets an $AG@k$ score for each query, and then they are ranked by their mean score across all queries.

To minimize random effects due to the particular sample of queries chosen, the Friedman test is run with the Average Gain

[1] In some editions of the MIREX AMS task it was defined from 0 to 10, with one decimal digit. Both definitions are equivalent.

scores of every system to look for significant differences across them. The Tukey's HSD test is then used to correct the experiment-wide Type I error rate [7]. The grand results of the evaluation are therefore pairwise comparisons between systems, telling which one is better for the current set of queries Q, and whether the observed difference was found to be significant.

3. EVALUATION WITH INCOMPLETE JUDGMENTS

The evaluation methodology used in MIREX is expensive in the sense that a complete set of similarity judgments is needed: the top k documents retrieved by every system have to be judged for every query. However, we may investigate how to compare systems so that we do not need to judge all documents and still be confident on the result of an evaluation experiment.

Let G_i be a random variable representing the gain of document i. The distribution of G_i is multinomial and depends on the similarity scale used: for the BROAD scale G_i can take one of three values, and for the FINE scale it can take one of 100 values. For now, let us assume that G_i follows a uniform distribution, that is, every similarity level is equally likely. The expectation and variance of G_i are as follows:

$$E[G_i] = \sum_{l\in\mathcal{L}} P(G_i = l) \cdot l$$

$$Var[G_i] = \sum_{l\in\mathcal{L}} P(G_i = l) \cdot l^2 - E[G_i]^2$$

where $\mathcal{L}$ is the set of possible relevance levels: $\mathcal{L}_{BROAD} = \{0, 1, 2\}$ and $\mathcal{L}_{FINE} = \{0, 1, ..., 100\}$. Given this definition of the gain of an arbitrary document, we can now define the $AG@k$ of an arbitrary system as a random variable too. Whenever document i is judged and assigned a gain l, the expectation and variance are fixed to $E[G_i] = l$ and $Var[G_i] = 0$; that is, no uncertainty about G_i.

Under the assumption that the gain of one document is independent of the others, expectation and variance of $AG@k$ are:

$$E[AG@k] = \frac{1}{k}\sum_{i\in\mathcal{D}} E[G_i] \cdot I(A_i \leq k)$$

$$Var[AG@k] = \frac{1}{k^2}\sum_{i\in\mathcal{D}} Var[G_i] \cdot I(A_i \leq k) \tag{1}$$

Having $AG@k$ defined this way allows us to estimate its value from an incomplete set of judgments. With no judgments at all, the variance of the estimator would be maximum, but as judgments are made the variance decreases. With all k documents judged, the variance is zero and the estimation equals the actual score.

3.1 Difference in AG@k

Using equations (1) we can estimate the $AG@k$ score of a system. But we are really interested in knowing which of two systems is better, that is, the sign of their difference in $AG@k$. For two arbitrary systems A and B:

$$\Delta AG@k = \frac{1}{k}\sum_{i\in\mathcal{D}} G_i \cdot I(A_i \leq k) - \frac{1}{k}\sum_{i\in\mathcal{D}} G_i \cdot I(B_i \leq k)$$

$$= \frac{1}{k}\sum_{i\in\mathcal{D}} G_i \cdot \big(I(A_i \leq k) - I(B_i \leq k)\big) \tag{2}$$

If $\Delta AG@k$ is positive, we can conclude system A performed better than system B (worse if negative) for the query. We can see that only documents retrieved by one system and not by the other will contribute to $AG@k$: documents retrieved by both systems will contribute $G_i - G_i = 0$. Therefore, judging these documents will not tell us anything about the difference. Thus, the larger the overlap between the systems' results, the fewer judgments are necessary to figure which one is better. Because the two systems are independent of each other, the expectation and variance are[2]:

$$E[\Delta AG@k] = \frac{1}{k}\sum_{i\in\mathcal{D}} E[G_i] \cdot \left(I(A_i \leq k) - I(B_i \leq k)\right)$$

$$Var[\Delta AG@k] = \frac{1}{k^2}\sum_{i\in\mathcal{D}} Var[G_i] \cdot \left(I(A_i \leq k) - I(B_i \leq k)\right)^2 \quad (3)$$

Now that we can compute an estimate of the difference for one query, let us generalize to a set Q of queries, computing the mean of the $\Delta AG@k$ scores for all them. As they are sampled randomly[3], queries are independent of each other, so the expectation and variance are:

$$E[\overline{\Delta AG@k}] = \frac{1}{|Q|}\sum_{q\in Q} E[\Delta AG@k_q]$$

$$Var[\overline{\Delta AG@k}] = \frac{1}{|Q|^2}\sum_{q\in Q} Var[\Delta AG@k_q] \quad (4)$$

With these estimates we can rank all systems by their difference in $AG@k$. For a given set of judgments, we can compute $P(\overline{\Delta AG@k} \leq 0)$, that is, the probability of system A performing worse than system B. If $P(\overline{\Delta AG@k} \leq 0) < \alpha$ then we can conclude that system A performs worse than B with α confidence ($1 - \alpha$ confidence of B being worse than A). If, while judging documents, we reached a certain confidence on the sign, say 95%, we could stop judging.

3.2 Distribution of ΔAG@k

To compute the confidence in the sign, we need to know the distribution of $\overline{\Delta AG@k}$. For a relevance scale with only two levels (similar and not similar), $AG@k$ is basically the same as $P@k$ (precision at k), which can be approximated by a normal distribution under a binomial or uniform prior distribution of G_i [1]. In our case, the BROAD scale has 3 possible levels, and the FINE scale has 101 levels.

Let G be a random variable representing the gain of the top 5 documents retrieved by a system for all possible queries, and let the set $\{X_1, \dots, X_{|Q|}\}$ be a random sample of size $|Q|$ where each X_i is the mean gain of k documents sampled from G. By the Central Limit Theorem, as $|Q| \to \infty$ the distribution of the sample average $\bar{X} = \sum X_i/|Q|$ approximates a normal distribution, regardless of the underlying distribution of G. Every X_i follows the definition of $AG@k$ for an arbitrary query i, and so $\bar{X}$ follows the definition of $\overline{AG@k}$ for a set of queries Q. Therefore, $\overline{\Delta AG@k}$ is normally distributed for a large number of queries, because it is the sum of two variables distributed normally.

---

[2] The indicator functions are squared in the variance so all documents have a positive contribution to the total variance.
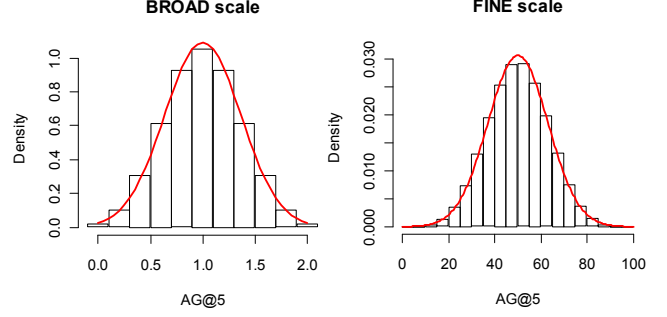[3] Note that this is rarely true in Text Information Retrieval.

Figure 1. Distribution of $AG@5$ assuming a uniform distribution of gain values for the BROAD (left) and FINE (right) scales. The red lines are normal distributions with means $E[AG@5]$ and variances $Var[AG@5]$.

Let us define Γ^k as the set of all $|\mathcal{L}|^k$ possible assignments of gain that can be made for k documents. Then, the probability of $AG@k$ being equal to a value z is:

$$P(AG@k = z) := \sum_{\gamma^k \in \Gamma^k} P(AG@k = z|\gamma^k) \cdot P(\gamma^k)$$

$$= \frac{1}{|\Gamma^k|}\sum_{\gamma^k \in \Gamma^k} I\left(\sum_{\gamma_i^k \in \gamma^k} \gamma_i^k = kz\right)$$

that is, the fraction of possible similarity assignments for which the average gain equals z. The left plot in Figure 1 shows the histogram of $AG@5$ scores observed in all $3^5=243$ possible assignments with the BROAD scale; and the right plot shows the scores observed in a random sample of 1 million assignments out of the 101^5 possibilities with the FINE scale. The red lines are normal distributions with means $E[AG@k]$ and variances $Var[AG@k]$. We can see that the normal distributions do indeed approximate very well.

Using the normal cumulative density function Φ we can easily compute the confidence on the sign of the difference as:

$$P(\overline{\Delta AG@k} \leq 0) = \Phi\left(\frac{-E[\overline{\Delta AG@k}]}{\sqrt{Var[\overline{\Delta AG@k}]}}\right) \quad (5)$$

3.3 Document Selection

Equations (3) and (4) can be used to estimate the difference between two systems with an incomplete set of judgments, but the problem is: what documents should we judge? Ideally, we want to judge only those that are most informative to know the sign of the difference in $AG@k$. For just two systems it is obvious from equation (2): only documents retrieved by one system and not by the other one are informative. For an arbitrary number of queries, we can just refer to a query-document pair as a single document (i.e. the gain of a document *for* a particular query).

For an arbitrary number of systems, a particular document could be informative for more than just one of the system comparisons. Therefore, we can assign a weight w_i to every query-document i, equal to the number of pairwise system comparisons for which judging query-document i would affect the estimate of $\Delta AG@k_q$. At any given time, we will want to judge the query-documents with largest weight because they will have the largest effect.
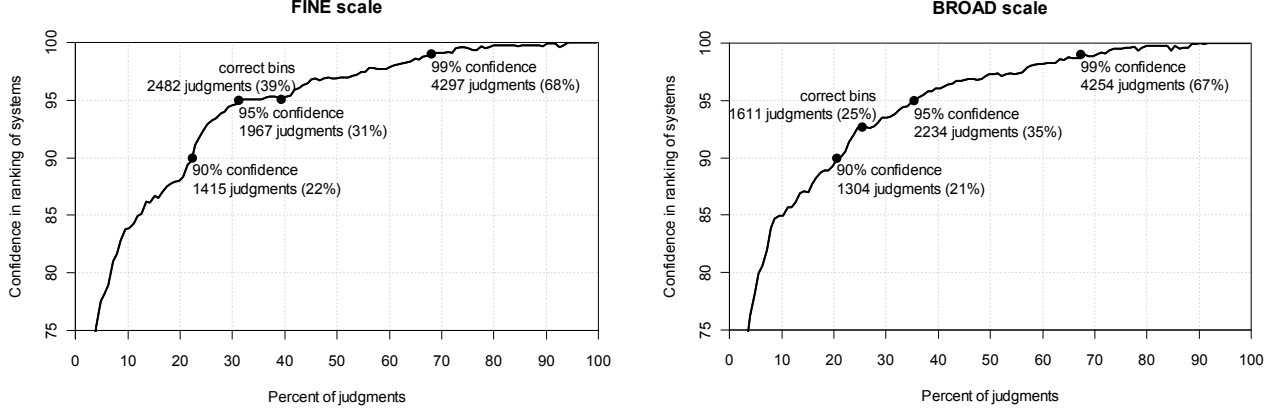
Figure 2. Confidence in the ranking of systems as the number of judgments increases, with the BROAD (left) and FINE (right) similarity scales. "correct bins" marks the point at which all true significant pairwise comparisons have a correct estimation of the sign of $\overline{\Delta AG@k}$.
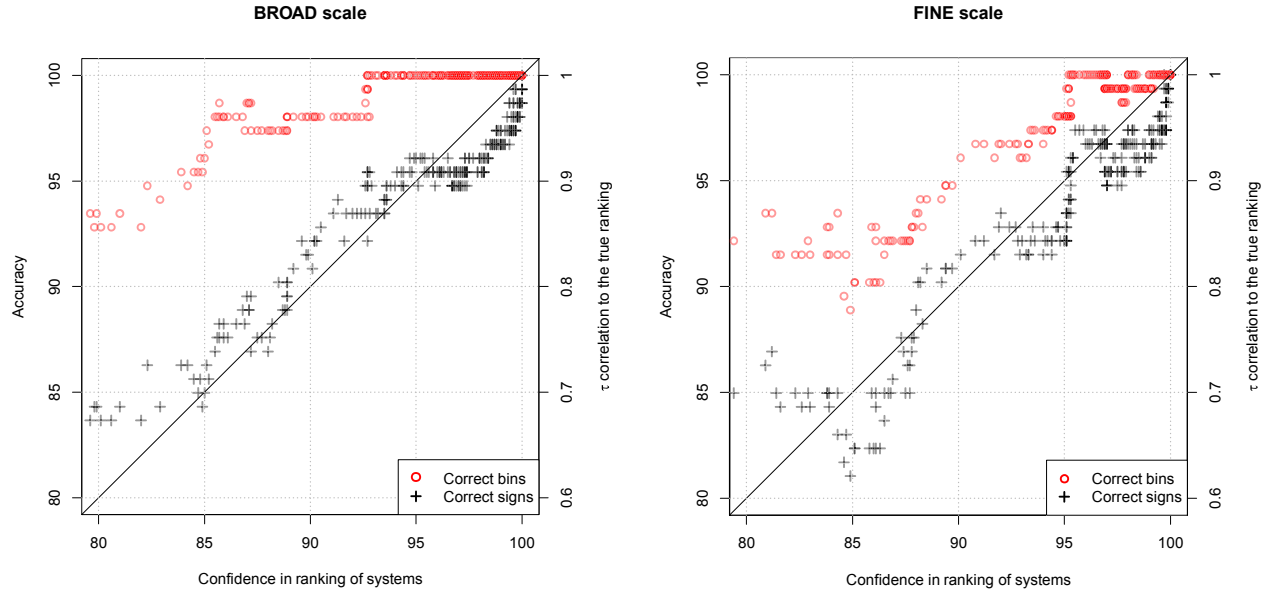

Figure 3. Accuracy of the sign of $\overline{\Delta AG@k}$ estimates and Kendall's τ correlation to the true ranking as confidence increases, for the BROAD (left) and FINE (right) similarity scales. Pairs with wrong sign estimates are considered correct under "correct bins" if they are not significantly different.

But if we were already highly confident about the difference between two systems, we would not need to judge another one of their query-documents. For two arbitrary systems A and B, let us define the confidence on the sign of their difference, as per Equation (5), as $C_{A,B}$. Being $\mathcal{S}$ the set of all system pairs, at any point we can compute the subset $\mathcal{R} \subset \mathcal{S}$ as the subset of pairs for which we are already highly confident on the sign of $\overline{\Delta AG@k}$. Ignoring these, the weight of every query-document is:

$$w_i = \sum_{(A,B)\in\mathcal{S}-\mathcal{R}} \left(1 - C_{A,B}\right) \cdot \left(I(A_i \leq k) - I(B_i \leq k)\right)^2 \quad (6)$$

That is, the contribution of a pairwise system comparison to the weight of a query-document is inversely proportional to the confidence in the sign of their difference.

At this point, we can define the MTC algorithm for $\Delta AG@k$:

Algorithm 1: MTC for $\Delta AG@k$

1: **while** $\frac{1}{|\mathcal{S}|} \sum_{(A,B)\in\mathcal{S}} C_{A,B} \leq 1 - \alpha$ **do**
2:      $\quad i^* \leftarrow argmax_i \, w_i$ for all unjudged query-document pairs
3:      $\quad$ judge query-document i^* (obtain true $gain_{i^*}$)
4:      $\quad E[G_{i^*}] \leftarrow gain_{i^*}$
5:      $\quad Var[G_{i^*}] \leftarrow 0$
6: **end while**

For the stopping condition we compute the mean confidence across all system pairs. If it is sufficiently large, we stop judging altogether. We call this the confidence on the ranking. Equation (6) ensures that the judging effort will be put into the less confident pairs.

4. RESULTS

We simulated the use of MTC to evaluate all systems from the MIREX 2011 Audio Music Similarity and Retrieval task. This is

the largest edition so far, where 18 systems were evaluated with 100 queries for a total of 6,322 judgments (see Table 1). There are thus 153 pairwise comparisons between systems.

Figure 2 shows how the confidence in the ranking of systems increases as more judgments are made: with no judgments confidence is 50% (i.e. one system is equally likely to be better than another one than it is to be worse), and with a complete set of judgments confidence is 100%. The pattern is quite similar for both similarity scales: 90% confidence is reached with about one fifth of the total judgments, 95% confidence with one third of the judgments; and 99% confidence with two thirds. This ranking confidence can be interpreted as the confidence in $\overline{\Delta AG@k}$ of any two systems picked at random.

We can see that high confidence levels can be achieved with considerably fewer judgments, but how good are the estimates of the sign of $\overline{\Delta AG@k}$? Figure 3 plots the accuracy of the estimates as a function of the ranking confidence. Accuracy is defined as the ratio of correct sign estimates across all 153 system comparisons. An estimate is considered correct under "correct bins" if it has the same sign as the true difference or the true difference is not statistically significant anyway. An estimate is considered correct under "correct signs" only if it has the same sign as the true difference regardless of the significance.

The accuracy of the estimated bins is always better than the ranking confidence, and for more than 90% confidence on the ranking most significant differences seem to be identified. If we look at the accuracy regardless of true significance ("correct signs"), it is again highly correlated with the ranking confidence, but it is sometimes lower than expected. This is caused by pairs of systems that are very similar, making the estimates swap from positive to negative values with very few judgments. Note that these swaps were considered correct under "correct bins".

A traditional way of comparing the estimated ranking and the true ranking is to compute the Kendall's τ correlation coefficient between the two. Rankings with correlations above 0.9 are usually considered equivalent if we account for the effect of having one or another person make the judgments [8]. Formally, 0.9 Kendall correlation corresponds to 95% accuracy under "correct signs". We can see in Figure 3 that rankings with more than 95% confidence do indeed have a very high correlation with the true ranking. Virtually all ranking estimates have a correlation over 0.9, although in the case with the FINE judgments a higher confidence seems necessary, as there are some lower observations around 95% confidence.

5. STATISTICAL SIGNIFICANCE

The results in the previous section show that we can estimate the ranking of systems with a fraction of the total judgments, and that this estimated ranking is very similar to the true ranking or equivalent for all practical purposes. The next problem would be to figure out whether differences between systems are statistically significant.

With the traditional methodology, after completing all the judgments and computing the true differences in $AG@k$ scores, systems are compared to each other to see whether the difference observed with the current set of queries would be expected with a different set of queries [7]. So far, the MTC algorithm allows us to estimate the difference for the *current* set of queries, but it does not allow us to generalize to a *different* set. The main problem at

this point is that we have an estimate of $\Delta AG@k$ for each query, but we do not have the variance of the sample of true $\Delta AG@k$'s for every query. Estimating this sample variance and adapting a statistical test accordingly is far from trivial. Instead, we work with the best and worst cases for each $\Delta AG@k$.

Let π and $\bar{\pi}$ be the sets of all judged and unjudged documents so far; and let l^+ and l^- be the maximum and minimum similarity levels permitted by the scale (2 and 0 for BROAD; and 100 and 0 for FINE). For an arbitrary pair of systems A and B, we compute the upper and lower bounds on $\Delta AG@k$:

$$[\Delta AG@k] = \frac{1}{k}\sum_{i\in\pi} G_i \cdot \big(I(A_i \le k) - I(B_i \le k)\big) +$$
$$+ \frac{1}{k}\sum_{i\in\bar{\pi}} l^+ \cdot I(A_i \le k) - \frac{1}{k}\sum_{i\in\bar{\pi}} l^- \cdot I(B_i \le k \wedge A_i > k)$$
$$[\Delta AG@k] = \frac{1}{k}\sum_{i\in\pi} G_i \cdot \big(I(A_i \le k) - I(B_i \le k)\big) +$$
$$+ \frac{1}{k}\sum_{i\in\bar{\pi}} l^+ \cdot I(B_i \le k) - \frac{1}{k}\sum_{i\in\bar{\pi}} l^- \cdot I(A_i \le k \wedge B_i > k)$$

The upper bound corresponds to the $\Delta AG@k$ score in the best case for system A: the first summation accounts for the information due to documents that have already been judged; the second summation assumes that all unjudged documents retrieved by A have the highest gain allowed by the scale; and the third summation assumes that all unjudged documents retrieved by B, but not by A, have the lowest gain allowed by the scale. The same (opposite) rationale follows for the lower bound. In our case, the third summations can be ignored, as the minimum gain in the BROAD and FINE scales is 0.

At any iteration of the algorithm we can compute the upper and lower bounds for every pair of systems and follow these rules:

1. If in the upper bound (best case for A), A would still be significantly worse than B, it does not matter which judgments we do next: we can conclude that A is significantly worse than B.
2. If in the lower bound (best case for B), B would still be significantly worse than A, we can similarly conclude that B is significantly worse than A.
3. If in the upper bound A would still not be significantly better than B, and in the lower bound B would still not be significantly better than A, we conclude they are not significantly different.

These rules only become useful with a relatively large amount of judgments: the upper and lower bounds are *very* high overestimations because the true performance of systems is far from the assumed perfect cases. With a large degree of incompleteness we can use another heuristic:

4. If the estimated difference is larger than a threshold t we naively conclude the systems are significantly different.

This threshold may be fixed based on power analysis. We can compute the effect size detectable by a paired 1-tailed t-test for a particular number of queries, variance of the sample of $\Delta AG@k$'s and permitted Type I and Type II error rates. We estimate the variance based on previous MIREX data ($\sigma = 0.248$), set the sample size to 100 queries, and set the Type I and Type II error rates to typical values $\alpha = 0.05$ and $\beta = 0.15$ (power = 0.85),
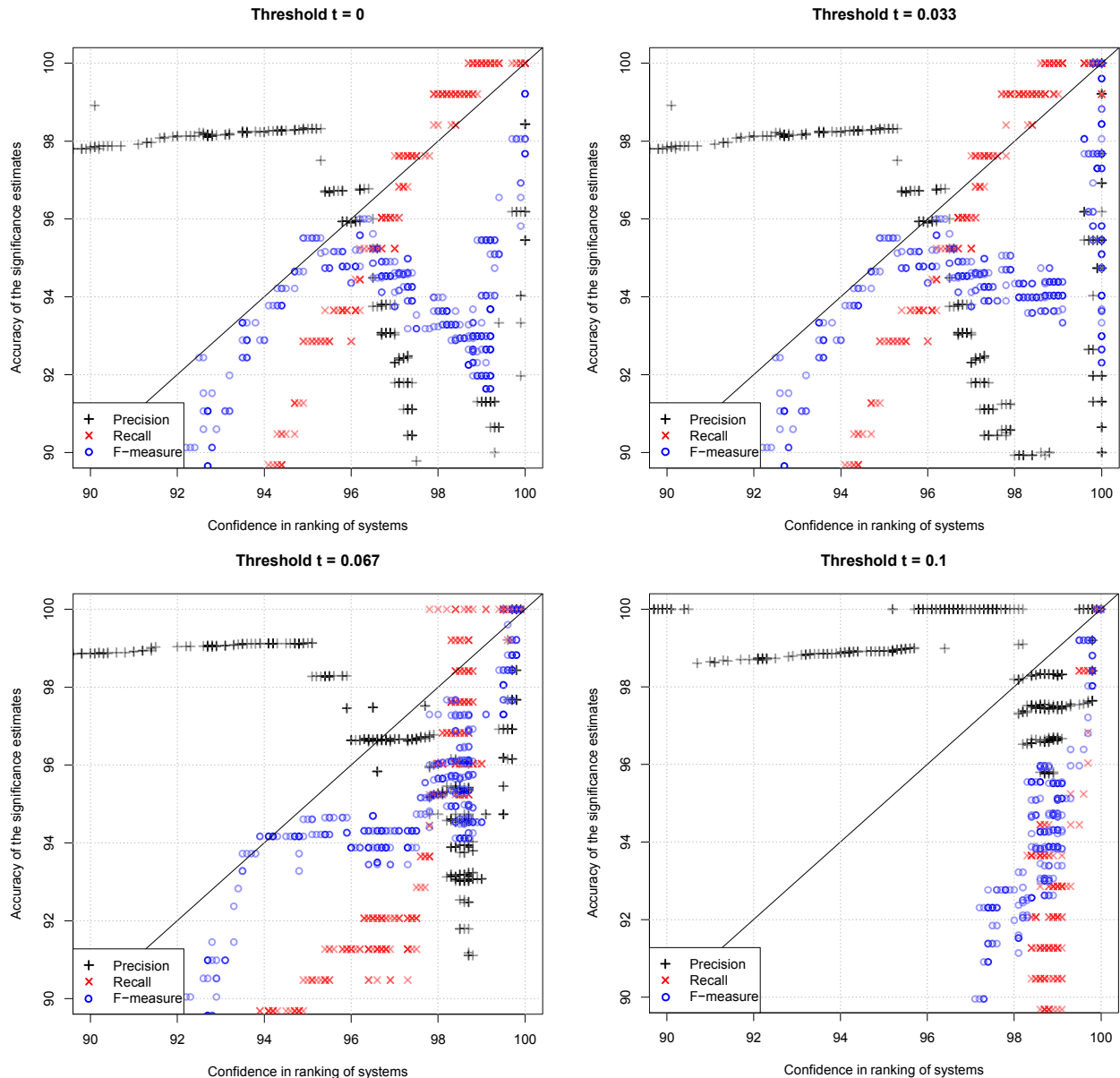
Figure 4. Accuracy of the significance estimates in terms of precision, recall and F-measure as confidence increases, for thresholds $t = 0$ (top left), $t = 0.033$ (top right), $t = 0.067$ (bottom left) and $t = 0.1$ (bottom right).

respectively. The observable true difference in this case would be ≈ 0.067. This is the value to which we fix the threshold t.

We can evaluate the accuracy of these rules with typical precision-recall ratios. If the rules estimate that a pair of systems is significantly different we count it as a positive result, negative otherwise. If the prediction is correct (according to the complete set of judgments), we count a true estimate, false otherwise. Figure 4 shows precision, recall, and F-measure for various thresholds.

The effect of the threshold t is clear: the larger the minimum difference between systems (large t), the fewer estimates turn out significant, so precision is high and recall is low. With $t = 0$ (i.e. the magnitude of the estimated differences is ignored), precision is above expected and recall is below; but when 95% confidence is

reached they begin to swap. At this point, we are confident of the sign of about 95% of the pairs, so about 95% of them will be (over) estimated as significant because $t = 0$ in rule 4. As the confidence increases, more comparisons are overestimated, but when approaching the complete set of judgments (over 99% confidence), rules 1 to 3 reduce the amount of false estimates again.

The overall accuracy of the significance estimates, as per the F-measure, corresponds fairly well with the confidence in the ranking until 95% confidence is reached. Nonetheless, it is always above 90%. When using the threshold $t = 0.067$ computed above, the F-measure improves considerably for very high confidence levels (bottom left plot), and gradually diminishes as the threshold gets larger (bottom right).

In general, we can see that the statistical significance of the differences can be estimated fairly well despite the incompleteness of judgments and the uncertainty on the true $\Delta AG@k$ scores.

6. CONCLUSIONS

We have shown how to adapt the Minimal Test Collections (MTC) family of algorithms for the evaluation of the Audio Music Similarity and Retrieval task. Assuming a uniform distribution of similarity judgments, we showed that the distribution of $AG@k$ scores is normally distributed, which allows us to look at it as a random variable whose expectation may be estimated with a certain level of confidence. This confidence is proportional to the number of similarity judgments available, and MTC ensures that the set of judgments we make to reach some confidence level is minimal.

Using the data from the MIREX 2011 AMS evaluation, we simulated MTC and found that with just one third of the judgments the correct ranking of systems can be estimated with 95% confidence, and with no swap between significantly different systems. We also showed some simple rules to estimate the significance of the differences, which work reasonably well for 95% confidence but tend to overestimate significance when higher confidence is achieved in the ranking of systems.

Three clear lines for future work can be identified. First, this paper assumed that the distribution of similarity judgments was uniform, that is, that all assignments of similarity were equally likely. However, it is clear this assumption does not hold in reality: the distribution of similarity judgments is skewed towards the highly similar or the not similar side, depending on how well the system performs. Having better estimates of the true distribution would make the algorithm perform better in terms of effort and accuracy of the estimates. These estimations of the true distribution could be approximated from previous MIREX AMS data, or with a model fitted with the judgments we make for the very systems we are evaluating, learning the true distribution on a per system or per query basis. Second, our estimates of the significance of the differences were based on very simple rules that assumed the (very unrealistic) best cases. Developing a comprehensive mathematical framework for testing significance is another clear line for future work.

The most important direction for further research is the study of low-cost evaluation methodologies for other MIR tasks. In accordance with previous work [7], we have shown that the effort in evaluating a set of AMS systems can be greatly reduced, leaving open the possibility of building brand new test collections for other tasks for which creating annotations is very expensive. For instance, the group of volunteers requested by MIREX for the yearly evaluation of the AMS and SMS tasks could be better employed if some of them were instead dedicated to incrementally add new annotations for the other tasks.

Another clear setting for the application of low-cost methodologies is that of a researcher evaluating a set of systems with a private document collection, a scenario very common in MIR given the legal restrictions on sharing music corpora. Those researchers, and in most cases public forums too, do not have the possibility of requesting large pools of external volunteers for annotating their collections. Thus, being able to evaluate systems

with the minimal effort is paramount. To this end, low-cost evaluation methodologies must be investigated for the wealth of MIR tasks.

In most of these tasks researchers rely on test collections annotated *a priori*, which can be very expensive and time consuming to build. However, we have seen that not all annotations are necessary to evaluate systems. For instance, if two Audio Melody Extraction algorithms predict the same F0 (fundamental frequency) in a given frame, whether that F0 prediction is correct or not is not useful to know which of the two systems is better. The adoption of *a posteriori* evaluation methodologies such as MTC can take advantage of this to greatly reduce the annotation cost or allow the use of larger collections. Getting to that point, though, requires a shift in the current evaluation practices. But given the benefits of doing so, both in terms of cost and reliability, we strongly encourage the MIR community to study these evaluation alternatives and progressively adopt them for a more rapid and stable development of the field.

REFERENCES

[1] B. Carterette. *Low-Cost and Robust Evaluation of Information Retrieval Systems*. Ph.D. dissertation, Department of Computer Science, University of Massachusetts Amherst, 2008.

[2] B. Carterette. Robust Test Collections for Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55-62, 2007.

[3] B. Carterette, J. Allan, and R. Sitaraman. Minimal Test Collections for Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268-275, 2006.

[4] J.S. Downie. The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. *Computer Music Journal*. 28(2): 12-23, 2004.

[5] J.S. Downie, A.F. Ehmann, M. Bay, and M.C. Jones. . The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. In *Advances in Music Information Retrieval*, W.R. Zbigniew and A.A. Wieczorkowska, eds. Springer. 2010, 93-115.

[6] J. Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *International Society for Music Information Retrieval Conference*, pages 609-614, 2011.

[7] J. Urbano, D. Martín, M. Marrero, and J. Morato. Audio Music Similarity and Retrieval: Evaluation Power and Stability. In *International Society for Music Information Retrieval Conference*, pages 597-602, 2011.

[8] E.M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management*. 36(5): 697-716, 2000.

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