HOW SIGNIFICANT IS STATISTICALLY SIGNIFICANT? THE CASE OF AUDIO MUSIC SIMILARITY AND RETRIEVAL

Julián Urbano
University Carlos III of Madrid
J. Stephen Downie
University of Illinois at Urbana-Champaign
Brian McFee
University of California at San Diego
Markus Schedl
Johannes Kepler University Linz

ABSTRACT

The principal goal of the annual Music Information Retrieval Evaluation eXchange (MIREX) experiments is to determine which systems perform well and which systems perform poorly on a range of MIR tasks. However, there has been no systematic analysis regarding how well these evaluation results translate into real-world user satisfaction. For most researchers, reaching statistical significance in the evaluation results is usually the most important goal, but in this paper we show that indicators of statistical significance (i.e., small p-value) are eventually of secondary importance. Researchers who want to predict the real-world implications of formal evaluations should properly report upon practical significance (i.e., large effect-size). Using data from the 18 systems submitted to the MIREX 2011 Audio Music Similarity and Retrieval task, we ran an experiment with 100 real-world users that allows us to explicitly map system performance onto user satisfaction. Based upon 2,200 judgments, the results show that absolute system performance needs to be quite large for users to be satisfied, and differences between systems have to be very large for users to actually prefer the supposedly better system. The results also suggest a practical upper bound of 80% on user satisfaction with the current definition of the task. Reflecting upon these findings, we make some recommendations for future evaluation experiments and the reporting and interpretation of results in peer-reviewing.

1. INTRODUCTION

Evaluation experiments are the main research tool in Information Retrieval (IR) to determine which systems perform well and which perform poorly for a given task [1]. Several effectiveness measures are used to assign systems a score that estimates how well they perform. The assumption underlying these evaluations is that systems with better scores are actually perceived as more useful by the users and therefore are expected to bring more satisfaction.

Researchers are usually interested in the comparison between systems: is system A better or worse than system B? After running an experiment with a test collection, researchers have a numeric answer to that question that measures the effectiveness difference between systems. Statistical procedures are then used to check whether that difference is statistically significant or not. Statistical significance is usually thought of as a sort of bulletproof evidence that one system really is better than another. Teams usually follow one or another research line based solely on statistical significance, and it has also become an essential requirement for publication in peer-reviewed venues.

However, there are several misconceptions regarding statistical significance [2, 11]. In the case of IR evaluation experiments, null hypotheses about differences in performance are false by definition, so observing a small p-value to conclude significance is just a matter of meeting certain conditions in the experiment. On the other hand, very little attention is paid to the effect-sizes and their implications in practical terms. In fact, even if statistical significance is present, the difference between two systems may very well be so subtle that users do not note the difference.

However, IR evaluations are traditionally focused on the algorithmic aspect of the systems, and whether the results do predict user satisfaction or not is very seldom studied [8]. Evaluation experiments make different assumptions regarding the operational settings and the needs and behavior of the users, so the extent to which results can be extrapolated should be questioned [9].

In this paper we focus on the evaluation of the Audio Music Similarity and Retrieval task (AMS), as carried out in the annual Music Information Retrieval Evaluation eXchange (MIREX). AMS is one of the tasks that most closely resemble a real-world music retrieval scenario, and it is also one of the tasks that receives most attention from the research community. We carried out an experiment with 100 users that allowed us to map system effectiveness onto user satisfaction, providing a new perspective in the interpretation of evaluation results. We also argue that researchers should not only focus on achieving statistical significance in effectiveness differences, but also on the size and practical implications of such differences.

2. SYSTEM EFFECTIVENESS AND USER SATISFACTION

In the MIREX AMS evaluation experiments, the similarity of a document to a query is assessed by humans and based on two different scales. The Broad scale has three levels: 0 (not similar), 1 (somewhat similar) and 2 (very similar). The Fine scale has 101 levels, from 0 (not similar at all) to 100 (identical to the query). Only one measure is reported to assess the effectiveness of the participating systems: $AG@5$ (Average Gain after 5 documents retrieved):

$$AG@k = \frac{1}{k}\sum_{i=1}^{k} gain_i$$

where $gain_i$ is the gain of the i-th document retrieved (the similarity score assigned). Two versions of $AG@5$ are actually reported, following the Broad and Fine scales.

$AG@k$ assumes that a document retrieved at rank 3 is as useful as a document retrieved at rank 30. A measure with a more realistic user model is $nDCG@k$ (Normalized Discounted Cumulated Gain after k retrieved) [3]:

$$nDCG@k = \frac{\sum_{i=1}^{k} gain_i / \log_2(i+1)}{\sum_{i=1}^{k} gain_i^* / \log_2(i+1)}$$

where $gain_i^*$ is the gain of the i-th document in the ideal ranking (i.e. $\forall i : gain_i^* \geq gain_{i+1}^*$). The gain contribution of a document is discounted with the logarithm of the rank at which it is retrieved, thus penalizing late arrival of relevant documents. Also, the gain contribution of documents is divided by the ideal contribution, bounding the measure between 0 and 1. Therefore, and for the sake of simplicity when comparing results across measures, we normalize $AG@k$ between 0 and 1 too:

$$nAG@k = \frac{1}{k \cdot l^+}\sum_{i=1}^{k} gain_i$$

where l^+ is the maximum similarity score allowed by the scale: 2 in the Broad scale and 100 in the Fine scale.

2.1 Interpretation of Effectiveness Scores

After running an evaluation of AMS systems, researchers interpret the results and make design decisions accordingly [9]. The ultimate goal is answering this question: what system would yield more user satisfaction? But we need to ask something else first: what measure and what similarity scale are better to predict user satisfaction?

Intuitively, if a system obtained a $nAG@5$ or $nDCG@5$ score of 1, our interpretation would be that an arbitrary user would be 100% satisfied with the results of the system, or satisfied 35% of the times if the effectiveness score achieved were 0.35. On the other hand, if system A obtained an effectiveness score larger than the one obtained by system B, we should expect users to prefer A. By choosing one or another measure, researchers make different assumptions as to the behavior and needs of the final users, and by choosing one or another similarity scale they follow different criteria to differentiate satisfying from unsatisfying results. To the best of our knowledge, none of these assumptions has been validated in the literature so far.

3. EXPERIMENTAL DESIGN

We devised an experiment with actual users that allowed us to map system effectiveness onto user satisfaction. Subjects were presented with a query clip and two ranked lists of five results each, as if retrieved by two different AMS systems A and B [8]. They had to listen to the clips and then select one of the following options: system A provided better results, system B did, they both provided *good* results, or they both returned *bad* results (see Figure 1). From these we can differentiate 4 judgments:

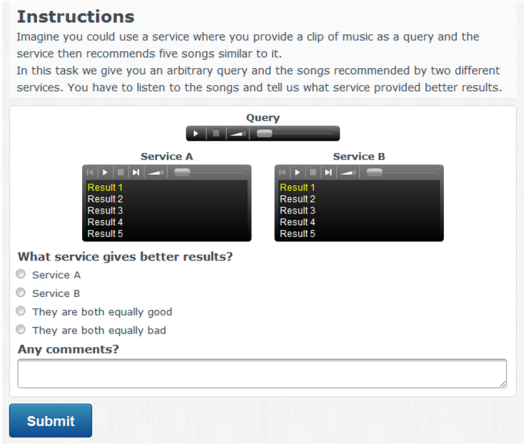

Figure 1. Task template used in the experiment.

- **Positive preference**, if the subject selected the system whose results yield *larger* effectiveness.
- **Negative preference**, if the subject selected the system whose results yield *smaller* effectiveness.
- **Good nonpreference**, if the subject indicated both systems are equally *good*.
- **Bad nonpreference**, if the subject indicated both systems are equally *bad*.

Such a design allows us to analyze the results from two different approaches: evaluation of a single system and comparison of two systems. Subjects indicating that both systems are *good* suggest that they are satisfied with both ranked lists. That is, their answer serves as an indication that the effectiveness measured for those systems translates into user satisfaction. If, on the other hand, they indicate that both systems are *bad*, we can infer that those effectiveness scores do not translate into user satisfaction. Subjects indicating a preference for one ranked list over the other one suggest that there is a difference between them large enough to be noted. That is, their answer serves as an indication that the difference in effectiveness between the systems translates into users being more satisfied with one system than with the other.

3.1 Data

We used the similarity judgments collected for the 2011 edition of the MIREX AMS task: a total of 18 systems by 10 research teams were evaluated with 100 queries, leading to a total of 6,322 unique similarity judgments. This is the largest edition as of the writing of this paper [1].

According to the definition of $nAG@k$ with Broad judgments, the difference between two systems is always a multiple of 0.1. For each difference $\Delta \in \{0, 0.1, ..., 1\}$, we selected 200 random queries and artificially created two random ranked lists of 5 documents such that their difference in $nAG@5$ would be Δ according to the Broad judgments made for that query in MIREX 2011. Therefore, we have a total of 2,200 examples. Note that for the extreme value $\Delta = 1$ we need at least 5 very similar documents and 5 not

[1] http://www.music-ir.org/mirex/wiki/2011:MIREX2011_Results

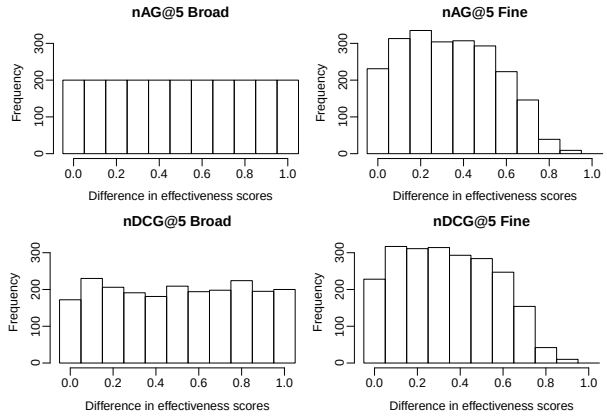

Figure 2. Distribution of effectiveness differences in all 2,200 examples, for $nAG@5$ (top) and $nDCG@5$ (bottom), and Broad (left) and Fine (right) judgments.

similar documents for the query. Due to this restriction, we could actually use only 73 of the total 100 queries. Across all 2,200 examples, we had 2,869 unique ranked lists of results, with 3,031 unique clips (including the 73 queries).

Figure 2 shows the distributions of effectiveness differences in the 2,200 examples. As mentioned, differences for $nAG@5$ with Broad judgments follow a uniform distribution, but with the Fine judgments there are very few examples with large differences. We note though that this is an artifact of the Fine scale itself and not a sampling flaw: for $\Delta = 0.9$ we need 5 documents with very high similarity scores (90 to 100) and 5 documents with very low scores (0 to 10); however, assessors very rarely assign such small and large scores. Therefore, it is very rare to observe differences that large when using the Fine scale.

These 2,200 pairs of artificial ranked lists can also be evaluated as per $nDCG@5$. As Figure 2 shows, the distributions of differences in $nDCG@5$ are very similar to $nAG@5$. Our examples do therefore cover the wide range of possible evaluation outcomes.

3.2 Procedure

All 2,200 judgments were collected via crowdsourcing. Previous work by Lee [6] and Urbano et al. [10] demonstrated that music similarity judgments gathered through crowdsourcing platforms are very similar to the ones collected with experts, with fast turnaround and low cost. Another advantage of using crowdsourcing for our experiment is that it offers a large and diverse pool of subjects around the globe. Using a controlled group of students or experts would probably bias our results, but using a diverse pool of workers allows us to draw conclusions that should generalize to the wider population of users.

However, using crowdsourcing has other issues. The quality of judgments via crowdsourcing can be questioned because some workers are known to produce spam answers and others provide careless answers to profit without actually doing the task. We decided to use the platform Crowdflower to gather the judgments, which delegates the work to other platforms such as Amazon Mechanical Turk. It also provides a quality control layer at the process level that separates good from bad workers by means of trap examples [5,8]: some of the examples shown to workers have known answers (provided by us) that are used to estimate worker quality. Workers that show low quality on the trap examples are rejected, and those that show high agreement are allowed to participate. We provided Crowdflower with 20 such trap examples (5 for each of the four answers), assigning each of them a subjective level of difficulty based on the answers by two experts.

3.3 Task Design

Figure 1 shows the task template we used. A first section listed the task instructions, and then a Flash player permitted subjects to listen to the query clip. Below, they could find the two ranked lists of 5 results each, followed by radio buttons to select the answer. Finally, a textbox was provided for workers to optionally leave feedback. All 3,031 audio clips were uploaded to our servers, and served upon request. The order in which examples are shown to workers is random, as is the assignment of the ranked lists as system A or system B. Also, we limited the maximum number of answers by a single worker to 50, minimizing the possible bias due to super-workers.

We collected all answers in four batches of 550 examples each. Lee collected similarity judgments paying \$0.20 for 15 query-document pairs [6], while Urbano et al. collected preference judgments paying \$0.02 for each query-document-document [10]. In both studies workers were therefore paid approximately \$0.007 per audio clip. Music-related tasks are known to be enjoyable by workers, and given that quality does not significantly degrade when decreasing wages [7], we decided to pay \$0.03 for each example, leading to approximately \$0.003 per clip. Adding the corresponding feeds to Crowdflower, all 2,200 judgments were collected for a grand total of \$100.

4. RESULTS [2]

The four batches were completed in less than 24 hours. We collected answers from 881 unique workers from 62 countries and 7 different crowdsourcing markets. These workers provided a grand total of 6,895 answers, from which Crowdflower accepted 3,393 (49%) as trustworthy. Note that the extra answers are due to repeatedly showing trap examples to workers. Only 100 workers were responsible for these trusted answers, so 781 workers (87%) were rejected. The average quality of these 100 workers, as computed by Crowdflower [5], ranges from 60% to 100%, with an average of 95%. In fact, 27 of our 2,200 examples contained the exact same documents, in the exact same order, in both ranked lists. Only twice did we not get, as should have, an unsigned preference in these cases. Therefore, the results reported herein comprise 2,200 answers by 100 different users who, apparently, provided honest responses.

4.1 Evaluation of a Single System

For 884 of the 2,200 examples (40%) we received a non-preference (i.e. subjects judged both systems as equally good or bad). Therefore, we have 1,768 ranked lists that subjects considered equally satisfying. Figure 3 shows the

[2] All data can be downloaded from http://julian-urbano.info.

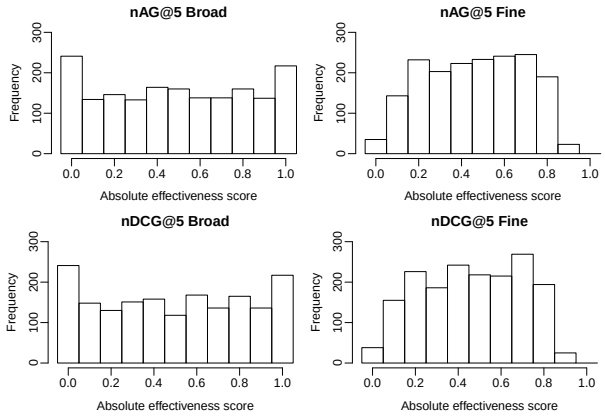

Figure 3. Distribution of absolute effectiveness scores in the 884 examples with unsigned preferences, for $nAG@5$ (top) and $nDCG@5$ (bottom), and Broad (left) and Fine (right) judgments.

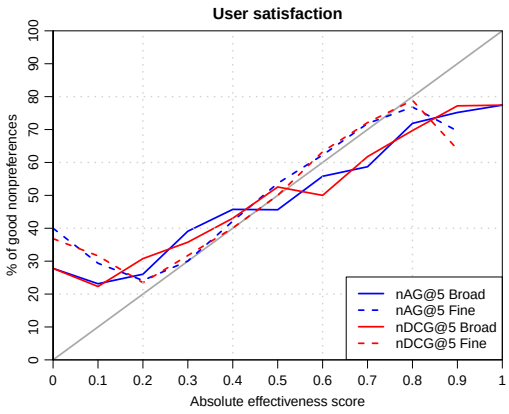

Figure 4. Ratio of good and bad nonpreferences in 884 examples, as a function of absolute system effectiveness, for $AG@5$ and $nDCG@5$ combined with the Broad and Fine judgments.

distributions of absolute effectiveness scores. As can be seen, a wide range of scores are covered, following a somewhat uniform distribution as well. The number of good and bad nonpreferences was almost the same too: 440 vs. 444.

Figure 4 shows the ratio of good nonpreferences observed in these 884 examples as a function of absolute effectiveness. As expected, there is a very tight positive correlation between effectiveness and user satisfaction. In fact, the relationship appears to be nearly linear. There is no appreciable difference between measures, but the Fine scale seems to adhere better to the diagonal than the Broad scale does. Note that the deviations from the trend with the Fine judgments ($\Delta < 0.2$ and $\Delta > 0.8$) are just an artifact of the very small number of observations in that range (see Section 3.1 and Figure 3).

Figure 4 shows a pretty straightforward mapping between $nAG@k$ and $nDCG@k$ scores and user satisfaction. However, the Broad scale seems to reveal a practical lower bound of 20% and an upper bound of 80% on user satisfaction. This could be merely due to noise in the crowdsourced data or a fault in the measures or scales. But given the symmetry, we believe these bounds are due to the natural diversity of users: some might consider something a very good result while others do not [4]. This means that even if a system obtains a $nAG@5$ score of 0, about 20% of the users will like the results (or dislike if $nAG@5 = 1$).

This is evidence of the room for improvement through personalization. Therefore, the AMS evaluations should include a user factor, possibly through user profiles, so that systems can attempt to reach 100% satisfaction on a per user basis. Otherwise, the final user satisfaction should not be expected to pass 80% for arbitrary users.

4.2 Evaluation of Two Systems

For 1,316 of the 2,200 examples (60%) we did receive a preference (i.e. subjects indicated that one system provided better results than the other one). Whether those user preferences were positive or negative (i.e. agreeing with the effectiveness difference or not), depends on the combination of measure and scale used. Figure 5 shows the ratio of preference signs across all 2,200 examples.

In terms of *positive* preferences (left plot), ideally we would want users to show a preference for the better system whenever we observe an effectiveness difference in the evaluation, regardless of how large this difference is. But there is a very tight positive correlation instead: the larger the difference in effectiveness, the more likely for users to prefer the supposedly better system. The relationship is again nearly linear, though this time we can observe a very clear difference between the Broad and Fine scales: for the same magnitude of the difference, the Fine judgments are always closer to the ideal 100% of positive user preferences. In fact, the Broad scale seems to indicate once again an upper bound of 80%. In addition, the plot shows that for users to prefer the supposedly better system more than the random 50% of the times, a difference of at least 0.3 in the Fine scale is needed, or 0.5 in the Broad scale. Note that the deviations from the trend with the Fine judgments ($\Delta > 0.8$) are also here just an artifact of the very small number of observations in that range.

As a consequence, there is a very clear negative correlation in terms of *nonpreferences* (middle plot): the larger the differences between systems, the more likely for users to prefer one of them. Again, the Fine scale seems to behave better than the Broad scale.

As the right plot shows, all four combinations of measure and similarity scale yield very similar ratios of *negative* preferences. There is a very slight negative correlation with difference in effectiveness, but in general about 5-10% of the user preferences disagree with the sign of the effectiveness difference. That is, about 5-10% of the times users prefer the supposedly worse system.

5. UNDERSTANDING EVALUATION RESULTS

The effectiveness of IR systems is assessed with different measures such as $nAG@k$ and $nDCG@k$. These measures are used to assign systems a score that represents how well they would satisfy users. For an arbitrary system A a measure M defines a distribution of effectiveness scores Y_A, describing the effectiveness of the system for an arbitrary query. The goal of evaluation experiments is usually finding the mean of that distribution: y_A.

Computing the parameter y_A allows researchers to assess how well the system performs and what is the expected user satisfaction according to the user model un-

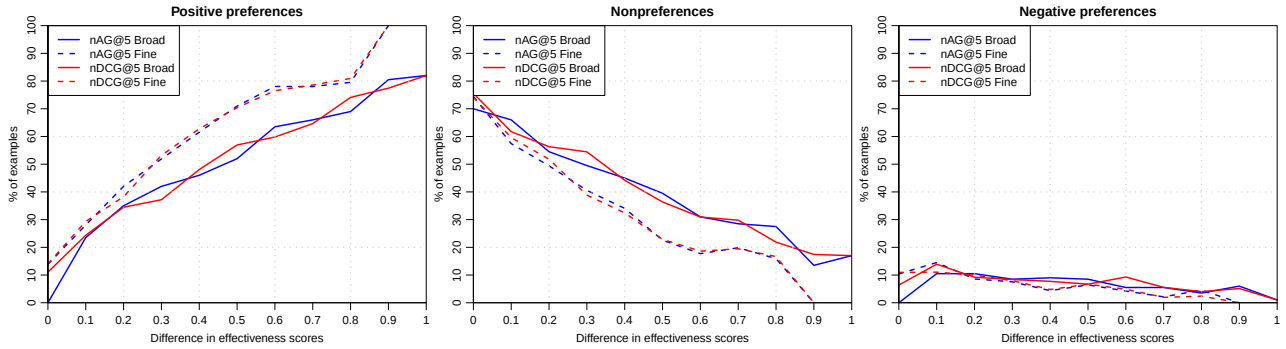

Figure 5. Ratio of positive preferences (left), nonpreferences (middle) and negative preferences (right) observed in the 2,200 examples, as a function of system effectiveness difference, for $nAG@5$ and $nDCG@5$ combined with the Broad and Fine scales.

derlying M. However computing this distribution would require running the system for the universe of all queries, which is clearly impossible. Instead, IR evaluation experiments are run with a sample of queries $\mathcal{Q}$, so they are used as estimators of the true y_{A}. The average effectiveness across queries, $\overline{y}_{\mathsf{A}}$, is used as the estimate $\hat{y}_{\mathsf{A}}$. Like any other estimate, $\hat{y}_{\mathsf{A}}$ bears some uncertainty, so statistical techniques such as confidence intervals should be employed to report the confidence on the estimation.

When comparing two systems, say A and B, one is usually interested in the distribution of the difference D_{AB}, representing the *paired* difference in effectiveness between A and B for an arbitrary query. Again, a comparative IR evaluation experiment only provides an estimate $\hat{d}$, whose sign indicates which system is expected to perform better.

5.1 Statistical Significance: p-values

Given that $\overline{d}$ is an estimate, the immediate question is: how confident can we be of this difference? The observed $\overline{d}$ could be just a random and rare observation due to the particular sample of queries used. Again, statistical techniques are needed to compute some sort of confidence on the difference. The most popular is hypothesis testing.

In a statistical hypothesis testing procedure, a *null* hypothesis H_0 is defined, such as $H_0 : d = 0$. The *alternative*, or research hypothesis, is then defined as the opposite: $H_1 : d \neq 0$. All hypothesis testing procedures are based on probability distributions, so there is always some degree of uncertainty when estimating parameters such as d. Thus, researchers may commit one of two errors: a Type I error if they conclude H_0 is not true when it actually is, or a Type II error if they conclude H_0 is true when it is not. The maximum probability of committing a Type I error is known as the *significance level*, usually $\alpha = 0.05$. The probability of committing a Type II error is denoted with the letter β, and $1-\beta$ is known as the *power* of the test: the probability of detecting a difference if there really is one.

The result of a hypothesis testing procedure is a probability called *p-value*. These are usually mistaken with the probability of H_0 being true [2, 11], but they are actually the probability of observing the difference $\overline{d}$ (or one larger) under the assumption that H_0 is true. That is, p-values are the probability of the data given the hypothesis, not the probability of the hypothesis given the data. If the reported p-value is smaller than the significance level α, we then reject the null hypothesis in favor of the alternative, and say that the difference is *statistically significant*. But it is important to note that the test does *not* tell anything about H_0 being true or false: that dichotomous interpretation is made by *us* based on the p-value and α, not by the test.

This is the ultimate goal of an IR evaluation: reaching significance. However, observing a statistically significant difference between two systems is usually misinterpreted as having high confidence that one system *really* is better than the other one because H_0 was rejected [2, 11]. In fact, all these null hypotheses are false by definition: any two different systems produce a distribution of differences with $d \neq 0$. What is important is the magnitude of d: differences of 0.0001, for instance, are probably irrelevant, but differences of 0.8 definitely are. However, a difference of just 0.0001 will always be statistically significant under certain experimental conditions, so focusing on statistical significance alone becomes, at some point, meaningless.

5.2 Practical Significance: effect-sizes

The most popular procedure to test such hypotheses about population means is the paired t-test. In IR evaluation, the hypotheses use to be $H_0 : d \leq 0$ and $H_1 : d > 0$. The test statistic is then computed as (note that in our case $d = 0$):

$$t = \frac{\overline{d} - d}{s_d / \sqrt{|\mathcal{Q}|}} \qquad (1)$$

where s_d and $\overline{d}$ are the standard deviation and mean of the sample of D_{AB} computed with the set of queries $\mathcal{Q}$ in the test collection. Using the t-distribution's cumulative distribution function, the p-value is then calculated as the area that is to the right of t. If p-value $< \alpha$, *we* reject the null hypothesis and plainly conclude $d > 0$.

Examining Eqn. (1) we can see that the test is more likely to come up significant with larger observed differences $\overline{d}$ and smaller deviations s_d. But most important is to note that the power of the test is also directly proportional to the sample size $|\mathcal{Q}|$: the more queries we use to evaluate systems, the more likely to observe a significant difference. This shows that focusing on significance alone is eventually meaningless: all a researcher needs to do in order to obtain significance is evaluate with more queries.

Increasing the sample size (number of queries) increases the power of the test to detect ever smaller differences because the standard error on the mean, $s_d/\sqrt{|\mathcal{Q}|}$, decreases. Thus, observing a statistically significant difference does

not mean that the systems really are different, in fact *they always are*. It just means that the observed difference and the sample size used were large enough to conclude *with confidence* that the true difference is larger than zero.

What really matters is how far apart from zero d is. This is the effect-size, which measures the *practical* significance of the result. As shown in Section 4.2, large differences in effectiveness scores (large effect-sizes) do predict more user satisfaction, but small differences do not really. However, with a sufficiently large number of queries we may be able to detect a statistically significant difference whose effect-size is extremely small, having no value for real users. In such a case we would have statistical significance, but no practical significance at all.

5.3 Reporting and Interpreting Results

We showed above that obtaining small p-values (statistical significance) should not be the sole focus of researchers when running evaluation experiments. The focus should really be on obtaining large effect-sizes (practical significance). The easiest way to report effect-sizes is just to report the effectiveness difference between systems or the absolute score of a single system. But these figures are just estimates of population means, and therefore subject to error. A better way to report effect-sizes is with confidence intervals, computed as $\overline{d} \pm t_{\alpha/2} \cdot s_d/\sqrt{|\mathcal{Q}|}$. Confidence intervals for the absolute effectiveness of a single system are computed likewise, but using the $\overline{y}$ and s_y estimates.

Along with the results in Section 4, these confidence intervals can be used to interpret evaluation results from the ultimate perspective of user satisfaction. For instance, the HKHLL1 system in MIREX AMS 2011 obtained a $nAG@5$ score of 0.422 for the Fine judgments, with a 95% confidence interval ranging from 0.376 to 0.468. According to the results in Figure 4, this system is expected to satisfy an arbitrary user from about 35% to 45% of the times.

On the other hand, the difference between SSPK2 and DM2 was found to be statistically significant. The magnitude of the difference was just 0.082, with the 95% confidence interval ranging from 0.051 to 0.112. According to Figure 5 though, such difference is hardly ever noted by the users. Indeed, substituting in Eqn. (1) we find that any $\overline{d}$ larger than 0.031 would have been deemed as statistically significant for these two systems. This is an example of a statistically significant difference that makes no practical difference for arbitrary users.

In summary, we suggest to report not only the observed scores but also their confidence intervals, and the actual p-values rather than an indication of significance. For instance, a proper report for a single system would read as $nAG@5 = 0.584 \pm 0.023$. For the difference between two systems, we suggest $\Delta nAG@k = 0.037 \pm 0.031 (p = 0.02)$. By reporting the p-value we leave the interpretation of significance to the reader and his operational context: a large effect-size (e.g. $\overline{d} = 0.43$), even if not statistically significant (e.g. p-value $= 0.06$), is definitely worth implementing. After all, the levels $\alpha = 0.05$ and $\alpha = 0.01$, despite widely accepted, are completely arbitrary. People generally consider p-value $= 0.054$ as significant, while others request p-value < 0.005. It depends on the context of the reader and factors such as the cost of committing a Type I error or the cost of implementing one or another technique.

6. CONCLUSIONS

Reaching statistical significance in IR evaluation experiments is usually the most important goal for researchers. A difference between systems is usually regarded as important if significance is involved, when in reality all systems are different. With the development of ever larger test collections, statistical significance can easily be misunderstood, suggesting large differences between systems when they are actually very similar. To predict the real-world implications of these differences, researchers need to focus on effect-sizes as indicators of practical significance. That is, it does not matter whether there is a difference or not (in fact, there always is), what matters is how large it is. Final user satisfaction is only predicted with effect-sizes. Statistical significance serves just as a measure of confidence.

However, even when reporting on the magnitude of effectiveness differences, there is no established relationship with final user satisfaction. To fill this gap we carried out a user study with 100 real users in the context of the Audio Music Similarity and Retrieval task, where subjects indicated their preferences between different system outputs. Our results allow researchers to map observed absolute scores and relative effectiveness differences directly onto expected user satisfaction. In addition, they suggest room for improvement if considering personalization, as well as further work on the development of measures and evaluation criteria that more closely capture the user model underlying the task.

7. REFERENCES

[1] D. Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2011.

[2] J. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2005.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Systems*, 2002.

[4] M. Jones, J. Downie, and A. Ehmann. Human similarity judgments: implications for the design of formal evaluations. In *ISMIR*, 2007.

[5] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution. In *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

[6] J. Lee. Crowdsourcing music similarity judgments using Mechanical Turk. In *ISMIR*, 2010.

[7] W. Mason and D. Watts. Financial incentives and the performance of crowds. In *ACM SIGKDD Workshop on Human Computation*, 2009.

[8] M. Sanderson, M. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *ACM SIGIR*, 2010.

[9] J. Urbano. Information retrieval meta-evaluation: challenges and opportunities in the music domain. In *ISMIR*, 2011.

[10] J. Urbano, J. Morato, M. Marrero, and D. Martín. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

[11] S. Ziliak and D. McCloskey. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008.