

# A MEASURE FOR EVALUATING RETRIEVAL TECHNIQUES BASED ON PARTIALLY ORDERED GROUND TRUTH LISTS

Rainer Typke, Remco C. Veltkamp, Frans Wiering

Universiteit Utrecht  
Padualaan 14  
3584 CH Utrecht, The Netherlands

## ABSTRACT

For the RISM A/II collection of musical incipits (short extracts of scores, taken from the beginning), we have established a ground truth based on the opinions of human experts. It contains correctly ranked matches for a set of given queries. These ranked lists contain groups of documents whose ranks were not significantly different. In other words, they are only partially ordered. To make use of the available information for measuring the quality of retrieval results, we introduce the “average dynamic recall” (ADR) that averages the recall among a dynamic set of relevant documents, taking into account the fact that the ground truth reliably orders groups of matches, but not always individual matches. Dynamic recall measures how many of the documents that should have appeared before or at a given position in the result list actually have appeared. ADR at a given position averages this measure up to the given position. Our measure was first used at the MIREX 2005 Symbolic Melodic Similarity contest.

## 1. INTRODUCTION

The ground truth from [1] was used at the “1st Annual Music Information Retrieval Evaluation eXchange” (MIREX) 2005 for comparing various methods for measuring melodic similarity for notated music. In order to compare different algorithms, a measure was necessary that compares every algorithm’s performance with the ground truth. The ground truth does not give one single correct order of matches for every query. One reason is that limited numbers of experts do not allow statistically significant differences in ranks for every single item. Also, for some alternative ways of altering a melody, human experts simply do not agree on which one changes the melody more. See Figure 1 for an example. In cases like this, even increasing the number of experts might not always avoid situations where the ground truth contains only *groups* of matches whose correct order is reliably known, while the correct order of matches within the groups is not known. Here, the 31 experts we asked do not

**Table 1.** Ground truth for Winter: “Domus Israel speravit”.



**Query:** Peter von Winter (1754-1825): Domus Israel speravit, RISM A/II signature: 600.054.278



1. Peter von Winter: Domus Israel speravit, 600.054.278



2. Peter von Winter : Domus Israel speravit, 600.055.822



3. Anonymus: Offertories, 450.040.980

agree on whether the second or the third incipit is more similar to the query. The third incipit is shorter, but otherwise identical to the query, while the second one contains more musical material from the query, but two ties are missing.

*Related work:* For situations where relevance is known on a scale that is finer than binary, Kekäläinen and Järvelin suggested graded relevance assessment measures based on cumulated gain [2], [3], which are related to traditional measures such as expected search length [4], average search length [5], and normalized recall [6], [7].

*Contribution:* We propose a measure (called “average dynamic recall”) that measures, at any point in the result list, the recall among the documents that the user should have seen so far. Unlike Kekäläinen’s and Järvelin’s measures [3], this measure only requires a partially ordered result list as ground truth, but no similarity scores, and it works without a binary relevance scale. It does not have any parameters that can be chosen arbitrarily, and it is easy to interpret. Our measure was used for the first time at the MIREX 2005 competition for symbolic melodic similarity.

## 2. MOTIVATION

Because the ground truth we used is not based on a finite relevance scale and does not contain relevance scores for the documents, we are proposing a new measure for our comparison. We try to meet the following criteria:

1. To make comparisons easy, the measure should deliver one number, for example in the range from 0 to 1, where 0 denotes a completely useless result and 1 a result that completely agrees with the ground truth.
2. In the ground truth, we know only the correct order of groups of matches, not necessarily of every single match. The measure should be able to use the existing information without requiring the ground truth to be completely ordered.
3. There are no relevance scores known for the documents in the ground truth, which only consists of a partially ordered list. The measure should therefore not depend on relevance scores.
4. The measure should not have any parameters one could use to dramatically alter the results (such as a freely chosen discount function for the purpose of rewarding returning highly relevant matches early, arbitrarily chosen thresholds, and the like).
5. The measure should reward putting matches in the right order, as far as that order is known. Therefore, differences in the order within groups should not influence the result, but differences in the order across group boundaries should.
6. In a similar fashion, violations of the correct order should be punished if they happen across group boundaries.
7. False positives in the result should lead to a lower measure, even if the order of the true positives is correct.
8. Both true and false positives that occur close to the beginning of the result list should have a higher influence on the measure than those occurring closer to the end of the list.
9. Since the group sizes do not mean much (they are influenced, for example, by the threshold for statistical significance that was chosen when the groups were established [1]), they should not have a high influence on the measure.

We are not aware of an existing measure that fulfills all of these criteria and therefore introduce the "average dynamic recall".

## 3. DEFINITION

Our measure is the average recall over the first  $n$  documents, where  $n$  is the number of items in the ground truth, and the recall is calculated over a dynamic set of relevant documents. Because of this, we call it "average dynamic recall". At the beginning of the result list, only the most similar document is counted as relevant (or all documents of which it is not known that they are less similar than the most similar one). The set of relevant documents grows with the position in the result list. Since there are groups of documents in the ground truth where no differences in relevance are known, the dynamic set of relevant documents does not always grow just by one single new relevant document. Rather, at each group boundary it grows by all elements of the next group, and it does not grow between group boundaries. However, at each position in the result list, we still divide the number of found relevant items at that position by the position number, not by the number of all items that would count as relevant.

More formally, consider a result list

$$\langle R_1, R_2, \dots \rangle$$

and a ground truth of  $g$  groups of items

$$\langle (G_1^1, G_2^1, \dots, G_{m_1}^1), (G_1^2, \dots, G_{m_2}^2), \dots, (G_1^g, \dots, G_{m_g}^g) \rangle$$

(with  $m_i$  denoting the number of members of group  $i$ ) where we know that  $\text{rank}(G_j^i) < \text{rank}(G_l^k)$  if and only if  $i < k$ , but we do not know whether  $\text{rank}(G_j^i) < \text{rank}(G_p^i)$  for any  $i$  (unless  $j = p$ ). We propose to calculate the result quality as follows. Let  $n = \sum_{i=1}^g m_i$  be the number of matches in the ground truth and  $c$  the number of the group that contains the  $i$ th item in the ground truth ( $\sum_{v=1}^c m_v \geq i \wedge \sum_{v=1}^{c-1} m_v < i$ ). Then we can define  $r_i$ , the recall after the item  $R_i$ , as:

$$r_i = \frac{\#\{R_w | w \leq i \wedge \exists j, k : j \leq c \wedge R_w = G_k^j\}}{i}.$$

The result quality is then defined as:

$$ADR = \frac{1}{n} \sum_{i=1}^n r_i.$$

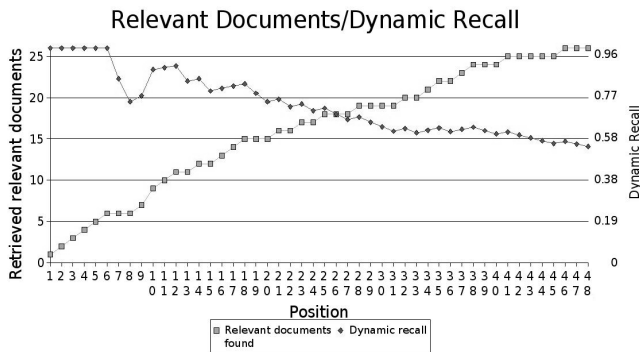
As an example, consider  $\langle (1, 2), (3, 4, 5) \rangle$  as ground truth and the result list  $\langle 2, 3, 1, 5, 7, 8, 9, 4 \rangle$ . That is, while we do not know whether item 1 or item 2 should be at the top of the list, we know that both should be ranked higher than any of the items 3, 4, and 5. In this case, the result quality is calculated as follows:

Pos.	encountered	relevant	#found	recall
1	2	1, 2	1	1
2	2, 3	1, 2	1	0.5
3	2, 3, 1	1, 2, 3, 4, 5	3	1
4	2, 3, 1, 5	1, 2, 3, 4, 5	4	1
5	2, 3, 1, 5, 7	1, 2, 3, 4, 5	4	0.8

The overall result quality here is  $ADR = (1+0.5+1+1+0.8)/5 = 0.86$ .

If there was an additional false positive at position 2, say,  $\langle 2, 10, 3, 1, 5, 7, 8, 9, 4 \rangle$ , the result quality would be lower: 0.7433. False positives lower the result quality in two ways: by shifting subsequent true positives to lower ranks and possibly by shifting true positives out of the scope altogether. Both true and false positives have higher impacts if they occur closer to the beginning of the result list since they influence all subsequent recall values. This illustrates how the criteria number 7 and 8 are met. Criterion 1 is obviously met, and so are criteria 2, 3, and 4. Criteria 5 and 6 are met because of the way  $r_i$  is defined: at every group boundary, the set of items that count as relevant is extended by all elements in the next group. Therefore, it does not matter in which order group members are found, as long as they are found before the group boundary.

A more complex example can be found in Figure 1, which shows the ADR for a sample result of our Earth Mover’s Distance-based algorithm for measuring melodic similarity [8].



**Fig. 1.** Number of retrieved documents (left scale) and dynamic recall (right scale) for a distorted version of incipit 240.001.397-1 from [9] using a variant of our algorithm from [8]. The average dynamic recall here is 0.74. The dynamic recall curve goes down whenever the retrieved documents curve stays flat. An ideal dynamic recall curve would be 1 from position 1 to 48, where the ground truth ends. At position 48, the dynamic recall is the same thing as recall, here  $26/48 \approx 0.54$ .

#### 4. COMPARISON WITH NORMALIZED DISCOUNTED CUMULATIVE GAIN

The average dynamic recall (ADR) shares many advantages with the cumulative gain measures introduced by Järvelin and Kekäläinen [3], who state that their measures are, among other things, obvious to interpret, are based on recall bases instead of only on retrieved lists, systematically combine document rank and degree of relevance, and, in their normalized forms, support the analysis of performance differences:

- ADR is obvious to interpret: at any number of retrieved items, it gives the average recall among the documents that the user should have seen so far. It can be calculated not only for the first  $n$  documents, if  $n$  is the number of items in the ground truth, but also for other numbers of documents.
- ADR is based on an absolute ground truth, not on retrieved lists alone, and therefore does not vary uncontrollably if the considered retrieved lists change.
- ADR systematically combines actual document rank and desired document rank.
- ADR supports the analysis of performance differences of different IR methods since it is normalized.

An important difference between ADR and cumulated gain-based measures is that ADR does not rely on relevance scores and therefore does not take them into consideration. This avoids the problem of correctly choosing a discount function for a discounted cumulative gain measure. By choosing the discount function for the normalized discounted cumulative gain (nDCG) [3] accordingly, one can sometimes invert the result of performance analyses. Different discount functions put, for instance, different emphasis on the beginnings of result lists. Because of this, it is possible to construct pairs of result lists that differ at the beginning in a way such that with, for example,  $\log_2$  as discount function, the first list gets a better nDCG score than the second one. With  $\log_3$  as the discount function and the same pair of lists, the nDCG score of the second list can be better than that of the first list.

Besides the discount function, the relative differences between relevance scores also have a high impact on nDCG results. Changing the relevance scores can also lead to opposite comparison results. So, to make nDCG results meaningful, one has to know exactly how the value of a relevant item decreases with a growing position in the result list – this determines the discount function –, and also exactly how relevant every document is in relation to the other documents. The ADR, on the other hand, only requires a partially ordered list as a ground truth for delivering meaningful results.

A weakness of the ADR is that situations can arise where different documents are both counted as relevant or both as irrelevant, with no distinction between them, although it is known which one of the two should be ranked higher.

As an illustration of this problem, consider a ground truth of  $\langle(1), (2), (3), (4)\rangle$  and the result lists  $\langle 4, 3, 5, 6 \rangle$  and  $\langle 3, 4, 5, 6 \rangle$ . It would be nice if the second result list would get a better score because it is known that item 3 should be ranked higher than item 4. But the ADR does not distinguish between them since at the second position, both item 3 and item 4 are not yet in the dynamic set of relevant documents, and at the third position, it is too late to treat them differently because both item 3 and item 4 are already in the set of encountered documents. In a similar way, one can construct examples where pairs of relevant items from different groups in the ground truth are encountered so late in a result list that both are counted as relevant, no matter in which order they appear, although it is known which one of the two should be ranked higher.

Problems like this can be caused in two ways during the calculation of the ADR: by items which are first counted as irrelevant and later as relevant (like item 3 in the example above), or by items which are encountered at a higher position than the end of the group to which they belong in the ground truth. Therefore, one could break ties like this by calculating an ADR score based on a list containing those problematic items and an inverted ground truth. In this constructed list, all other items are replaced with one item from the most highly ranked group.

In the example above, items 3 and 4 fulfill the condition for inclusion in the constructed list, while items 5 and 6 do not, so we would construct the lists  $\langle 4, 3, 1, 1 \rangle$  and  $\langle 3, 4, 1, 1 \rangle$ . If we now calculate the ADR on these constructed lists using the inverted ground truth (here:  $\langle(4), (3), (2), (1)\rangle$ ), the problem with items being treated the same although it is known that they should be ranked differently cannot occur anymore (because of the way the list was constructed). The ADR calculated from these constructed lists and the inverted ground truth could be used to break ties. However, to have a measure that is obvious to interpret, we simply used the ADR as described in Section 3 for our comparison of melodic similarity algorithms.

## 5. CONCLUSIONS

The introduced evaluation measure opens new ways to assess group-based rankings. MIREX 2005 has shown that our ground truth for incipits from the RISM A/II collection in combination with our proposed “average dynamic recall” measure can serve as a basis for a benchmark for evaluating information retrieval systems. The ground truth, along with the sets of queries, candidates, and experimental re-

sults, can be found at <http://give-lab.cs.uu.nl/orpheus>. We encourage music retrieval researchers to apply their favourite methods to the RISM A/II collection and compare their results to our ground truth.

The combination of our method for building a ground truth and the ADR measure would probably also give interesting insights into the quality of search results for data other than music. Often, relevance cannot be easily captured with a binary scale, for example in image or video retrieval. In some cases, for example text retrieval on the internet, a large number of relevant documents makes it desirable to take even subtle relevance differences into consideration. Typical users of an internet search engine look only at the first ten matches, even if there are hundreds of relevant documents.

## 6. REFERENCES

- [1] R. Typke, M. den Hoed, J. de Nooijer, F. Wiering, and R. C. Veltkamp, “A ground truth for half a million musical incipits,” *Journal of Digital Information Management*, vol. 3, no. 1, pp. 34–39, 2005, Ground truth data available at <http://give-lab.cs.uu.nl/orpheus/>.
- [2] J. Kekäläinen and K. Järvelin, “Using graded relevance assessments in IR evaluation,” *Journal of the American Society for Information Science and Technology*, vol. 53, no. 13, pp. 1120–1129, 2002.
- [3] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.
- [4] W. S. Cooper, “Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems,” *Journal of the American Society for Information Science and Technology*, vol. 19, no. 1, pp. 13–41, 1968.
- [5] R. M. Losee, *Text retrieval and filtering: analytic models of performance*, Kluwer Academic, Boston, 1998.
- [6] J. J. Rocchio, *Document retrieval systems - Optimization and evaluation. PhD dissertation.*, Harvard, 1966.
- [7] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- [8] R. Typke, R. C. Veltkamp, and F. Wiering, “Searching notated polyphonic music using transportation distances,” in *Proceedings of the ACM Multimedia Conference*, 2004, pp. 128–135.
- [9] *Répertoire International des Sources Musicales (RISM). Serie A/II, manuscrits musicaux après 1600.*, K. G. Saur Verlag, München, Germany, 2002.