Evaluating ADM on a Four-Level Relevance Scale Document Set from NTCIR

(DRAFT - Not for Quotation)

Vincenzo Della Mea[1], Luca Di Gaspero[2], Stefano Mizzaro[1]

[1] Department of Mathematics and Computer Science University of Udine
Via delle Scienze, 206 I 33100 Udine, Italy
{dellamea|mizzaro}@dimi.uniud.it
http://www.dimi.uniud.it/~{dellamea|mizzaro}
Ph.: +39 0432 55{8461|8456}

[2] Department of Electrical, Management, and Mechanical Engineering University of Udine
Via delle Scienze, 208 I 33100 Udine, Italy
l.digaspero@uniud.it
http://www.diegm.uniud.it/digaspero
Ph.: +39 0432 558242

Abstract. Most common effectiveness measures for Information Retrieval (IR) systems are based on the assumptions of binary relevance (either a document is relevant to a given query or it is not) and binary retrieval (either a document is retrieved or it is not). These assumptions are often questioned, since almost everybody agrees that relevance and retrieval are matter of degree (three or more categories, if not a continuum). However, the standard practice in IR systems evaluation remains based on the use of precision and recall (and related measures), thus hindering IR development and evaluation.

We recently questioned these assumptions, and proposed a new measure named ADM (Average Distance Measure), in order to pass from binary to continuous relevance and retrieval (Mizzaro 2001, Della Mea & Mizzaro 2004). In this paper we describe the basic idea on which ADM is based and present the experimental results on two different test collections, namely TREC and NTCIR. TREC features 2-levels human relevance judgments and IR systems that rank the retrieved documents, whereas NTCIR features 4-levels human relevance judgments and IR systems that assign a numeric score to the retrieved documents.

Keywords Information retrieval evaluation, average distance measure.

1. Introduction

In the *Information Retrieval* (*IR*) field, most common measures of the effectiveness of an *Information Retrieval System* (*IRS*) are based on binary relevance (either a document is relevant to a given query or it is not) and binary retrieval (either a document is retrieved or it is not). These assumptions can, and need to, be questioned: relevance might be not binary, and IRSs usually rank the retrieved documents and, sometimes, show their weights (e.g., all the Web search engines, let alone the vector space based IR system existing since the 70es).

In previous articles (Mizzaro 2001, Della Mea & Mizzaro 2004) we have proposed and validated on TREC data a new IR effectiveness evaluation measure, that is based on nonbinary views of relevance and retrieval. In this paper we describe the basic idea on which this measure (named ADM for Average Distance Measure) is based and present the experimental results on two different test collections, namely TREC and NTCIR. TREC features 2-levels relevance judgments and IR systems that rank the retrieved documents, whereas NTCIR features 4-levels human relevance judgments and IR systems that assign a numeric score to the retrieved documents. Experimental results on TREC are more stable and have already been published in (Della Mea & Mizzaro 2004); on the other side, the analysis of NTCIR-4 data is still work in progress, and the results presented here are very preliminary ones.

The paper is structured as follows. In the following section, we recall the definition of ADM (Average Distance Measure), a new measure of retrieval effectiveness based on a continuous view of relevance and

© 2004 National Institute of Informatics

retrieval, and we also recall the results of a previous experiment on TREC data (full details are available in our previous articles). In Section 3 we present preliminary experimental results on using ADM to evaluate the IRSs participating in NTCIR-4 Workshop. The last section concludes the paper and sketches some future developments.

2. The Average Distance Measure

To describe ADM we need to identify two figures. We define the *User Relevance Score* (*URS*) as a value in the [0,1] range that measures the real (i.e., user determined) relevance of a document with respect to an information need. URS assumes the maximum value (i.e., 1) for "totally relevant" documents, it assumes a 0 value for "totally nonrelevant" items, and it assumes intermediate values for documents with various degrees of "partial" relevance. Conversely, the retrieval measure is named *System Relevance Score* (*SRS*): the score of the relevance of a document to a query given by the IRS. SRS has the same behavior as URS: it is in the [0,1] range, and 1 is its maximum value.

ADM is a new retrieval effectiveness measure based on the average distance, or difference, between URSs (the actual relevance of documents) and SRSs (their estimates by the IRS). For a given query *q*, we define two relevance weights for each document d_i in the database *D*: the SRS for d_i with respect to *q* (denoted by $SRS_q(d_i)$), and the URS for d_i with respect to *q* ($URS_q(d_i)$). ADM for the query *q* is then defined as the average distance between $SRS_q(d_i)$ and $URS_q(d_i)$:

$$ADM_q = 1 - \frac{\sum_{d_i \in D} \left| SRS_q(d_i) - URS_q(d_i) \right|}{|D|} \qquad (1)$$

(where the denominator is the number of documents in the database *D*). ADM_q is in the [0,1] range, with 0 representing the worst performance. By averaging ADM_q on some queries we obtain *ADM*, a measure of IR effectiveness. The set of the queries on which to average ADM_q is left unspecified by now, and this is a subtle issue as we will see in the following; the reader can think of averaging on all the documents in the database (even if this will not be the solution adopted in practice).

We can graphically understand ADM in the following way. Let us assign to each document in the database its own SRS and URS values (in the [0,1] range) and plot these values on a standard Cartesian diagram in the $[0,1]^2$ square (see Figure. 1). Each document is therefore a point in the *URS-SRS* plane; the closer the point to the ideal *SRS = URS* line (the dotted line in the figure), the better the estimate by the IRS (the points on the line are represented by filled circles in figure). The last thing we need to define is the distance between a point and the ideal line. Since the URS value is predefined and cannot be changed as a result of the retrieval of a document,[1] the distance is not the standard distance between a point and a line (i.e., the length of an orthogonal line from the point to the line), but the distance between the point representing the document and the point on the line with the same abscissa (represented by the arrows in figure). This is the definition used in Eq. (1).

Specialized forms of ADM can be also defined for systems based on ordinal categories for relevance and retrieval (even binary). ADM can be specialized into an $ADM_{(2)}^{(2)}$ measure to handle the binary relevance binary retrieval view: in this case, all the points in the URS-SRS plane turn out to be in either (0,0), (0,1), (1,0), or (1,1) and, therefore, the distances from the ideal line are either 0 or 1. When it is possible to associate a numeric value to ordinal categories, it is also straightforward to define: $ADM_{(N)}^{(M)}$, based on *N* categories of relevance and *M* categories of retrieval (i.e., URSs assume one of *N* values, and SRSs assume one of *M* values); $ADM^{(M)}$ (with *M* categories of retrieval and continuous relevance); and $ADM_{(N)}$ (with *N* categories of relevance and continuous retrieval).[2] $ADM^{(Rank)}$ represents ADM computed on the basis of

[1] In this paper we do not take into account the subjective and dynamic nature of relevance (Mizzaro, 1998; Schamber et al., 1990), and we assume that the user is able to determine the "real" relevance value. However, our results can be extended in a straightforward way to the more general case of the user view of relevance.

[2] The assignment of numerical value to ordinal categories can present subtle problems, that have been brought to our attention by Steve Robertson. As a matter of fact, the "linear scale assumption", i.e., the naïve assumption that the categories correspond to

the rank of retrieved documents (e.g., if the retrieved documents are 1000, the 1st ranked has SRS = 1.0, the 2nd has SRS = 0.999, and so on until the 1000th with SRS = 0.001). $ADM_{(N)}^{(Rank)}$ is $ADM^{(Rank)}$ with N categories of relevance.

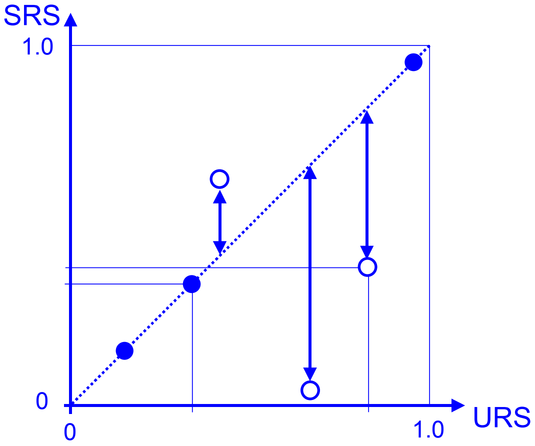


Figure 1. Graphical representation of ADM.

Given the features of NTCIR (4-levels relevance scale and IRSs evaluated on the basis of their score), the measures used in this paper will mainly be $ADM_{(4)}$ and $ADM_{(4)}^{(Rank)}$.

In our previous papers we have shown that, from a conceptual point of view, ADM is adequate for measuring the effectiveness of IRSs, in some respect even more adequate than classical precision and recall. A comparison between ADM on the one side and precision and recall on the other shows how rough precision and recall are. First, precision and recall are highly (too) sensitive to the thresholds chosen and to the documents close to the borders between sectors. Figure 2(a) shows how three documents might be judged by three hypothetical IRSs (circles represent IRS1, crosses IRS2, and squares IRS3). Clearly, the three systems are extremely similar, or at least evaluate the three documents in very similar ways. However, the values for precision, recall, E-measure (assuming again that the two thresholds, between relevant and nonrelevant and between retrieved and nonretrieved, are 0.5), and ADM (Table 1(a)) show that classical measures are rather different, whereas ADM is more stable.

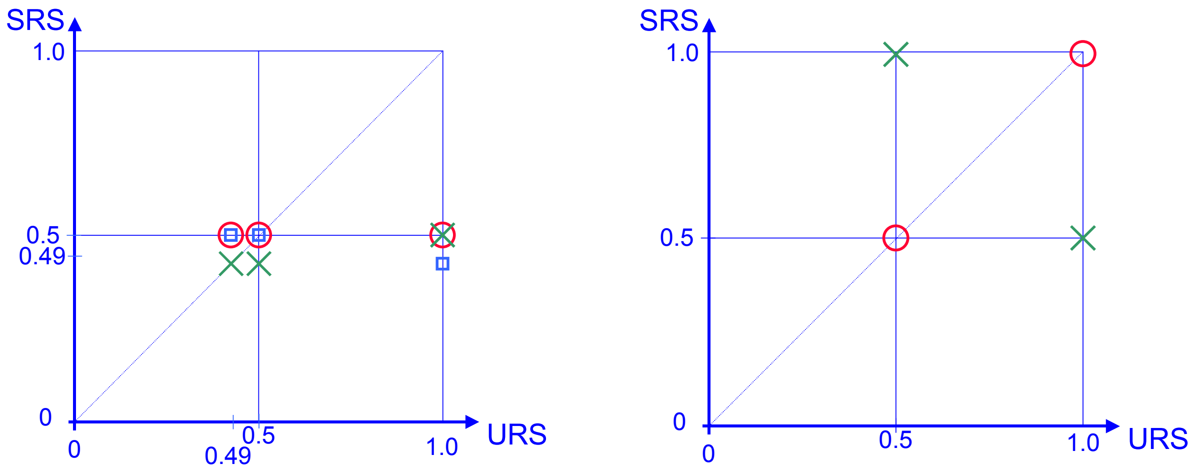
The second problem is that precision and recall are not sensitive enough to important differences between systems. Figure 2(b) shows how two documents might be judged by two hypothetical systems (circles stand for IRS1, crosses for IRS2). Clearly, the two systems evaluate the two documents in rather different ways. The values for precision, recall, E-measure, and ADM (Table 1(b)) show that classical measures are completely unable to grasp the difference, whereas ADM clearly differentiates the effectiveness of the two systems.

Therefore, the two problems about precision and recall are: first, small differences in the SRS can lead to very different precision, recall, and E-measure figures, whereas small differences do not affect ADM; second, big differences in SRS can lead to very similar (even identical) precision, recall, and E-measure figures, whereas big differences do affect ADM.

Both problems are relieved in real IRS evaluation, since precision and recall figures are obtained by averaging many queries retrieving many documents. However, they might be one reason for the high variation of precision and recall among different queries (often higher than the variation among different IRSs) (Harter, 1996). Moreover, looking at it from a different perspective, by using ADM in place of precision and recall, information retrieval experiments may be carried out on smaller data sets (less queries), and the effectiveness for queries with very few relevant documents is measured in a more reliable way.

equally distant URS (or SRS) values, can be easily questioned. This can be seen by means of a simple example. If we have three categories labeled “relevant”, “partially relevant”, and “not relevant”, it seems rather natural to give them 1, 0.5 and 0 values. But why should this assignment be preferred to, say, the 1, 0.6, 0 choice? Moreover, the symmetry considerations that might help in this case do not hold if the labels of the three categories are “highly relevant”, “relevant”, and “not relevant”, for which the values are even more arbitrary. Anyway, any solution seems better than collapsing the intermediate relevance categories into “relevant” or “not relevant”: this latter choice is the one with the highest error rate.

Moreover, binary relevance measures are not capable of taking into account multigraded relevance judgments, as the NTCIR case clearly shows.



(a) (b)

Figure 2. Small (a) and big (b) differences in SRS values.

	P	R	E	ADM
IRS1 ○	0.67	1	0.84	0.83
IRS2 ✕	1	0.5	0.75	0.83
IRS3 □	0.5	0.5	0.5	0.826

(a)

	P	R	E	ADM
IRS1 ○	1	1	1	1
IRS2 ✕	1	1	1	0.5

(b)

Table 1. Effectiveness measures for Figs. 2(a) and 2(b).

In (Della Mea & Mizzaro 2004) we also obtained experimental data to support our conceptual observations, by applying ADM to the ad-hoc track (both manual and automatic runs) of TREC-8 (Voorhees & Harman 2000). Since TREC-8 data do not contain reliable continuous SRS and URS values, we had to introduce some simplification, in order to compute two continuous URS from the available data, from which we then obtained two ADMs, to be compared with traditional measures.

Full details of the comparison are available in the paper (Della Mea & Mizzaro 2004); to summarize, the evaluation results allow to state that ADM evaluates IRS effectiveness in a way similar to that given by the measures used in TREC (Rel-Ret, AvgPrec and R-Prec), and the number of documents needed for evaluation can be lower.

3. NCTIR Data

In order to evaluate ADM on IR systems not based on the binary relevance/binary retrieval paradigm, we have been working on data from NTCIR-4, which is composed by documents judged on a four-level scale, with IR systems providing continuous retrieval values. This allowed us to test our measures on data more suited to the ADM approach.

The data set included results from 14 IRS, working to retrieve documents for 50 queries in four languages (Chinese, English, Japanese, and Korean). In the submission form adopted by NTCIR, IRSs query results are composed by a list of 1000 documents for each system, with a continuous score for each document. Thus, the SRS is continuous.

However, different IRSs have SRSs ranging over different values; we had therefore to normalize them in the [0..1] range. We adopted a simple linear normalization, in which the original maximum SRSs is mapped to 1 and the minimum to 0. We experimented on normalizing both *by run* (i.e., choosing the maximum and minimum values among all the SRS expressed by an IRS on all the queries) and *by query* (i.e., choosing the maximum and minimum values among all the SRS expressed by an IRS on a single query).[3] Normalization is another subtle issue on which we will come back in the following.

Some among the first documents retrieved by the IRSs are pooled and judged by human assessors on a four level scale:

- totally relevant (S);
- relevant (A);
- partially relevant (B);
- not relevant (C).

Since judgments are expressed in categories, we had to convert them to real values to obtain USR. This is somehow an arbitrary choice; we chose S = 7/8, A = 5/8, B = 3/8, C = 1/8. These values are chosen to have the [0..1] range split into four sub-intervals of equal length (i.e., those centered on the four values: [0..2/8), [2/8.. 4/8), [4/8..6/8), and [6/8..1]).

4. ADM Experiments on NTCIR-4 Data

Given the short time (we have been able to work on the data for a couple of weeks only) we can report rather preliminary results only, that can be thought of as the kind of results we might find, rather than the actual results. All evaluations have been made using the R statistics package, by means of suitable scripts.

First of all, we evaluated ADM starting from URS and SRS as above defined. Then, we compared ADM with the traditional effectiveness measures used by NTCIR-4 as well as TREC. We focused on the following reference measures: Average Precision (AvgPrec) and R-Precision (R-Prec) (Voorhees & Harman 2000). We based the comparison on the Kendall correlation, as we did in our previous work (Della Mea & Mizzaro 2004). A good correlation between ADM (in one of its variants) and a traditional measure might imply that ADM is able to measure IRS effectiveness as usual measures do. However, we also aim at having ADM measuring something different from classical IR measures, so a low correlation might be, under some circumstances, interesting.

Following the approach in (Della Mea & Mizzaro 2004), we also evaluated ADM using less data. The idea is that we may use less documents for the evaluation if the ADM, measured on a limited number of documents, correlates well with the ADM measured on all the topics and with the standard measures.

While doing such basic experiments, we also had to study the IRS score distributions, and starting from that we introduced further evaluations and comparisons. The next subsections report on experiments and results, either positive or negative.

4.1. Unexpected Results and Problems

Let's start with some unexpected results. $ADM_{(4)}$ does not correlate with gold standards. Also $ADM_{(4)}^{(Rank)}$ shows correlation values that, even if slightly higher, are far from the expected ones. Table 2 shows Kendall's correlations among ADM measures and Average precision and R-Prec (we use the "relaxed" version; the figures for rigid ones are similar when not otherwise explicitly stated).

	R-Prec	ADM	$ADM^{(Rank)}$
AvgPrec	0.91	0.03	0.35
R-Prec	1	0.05	0.37
ADM		1	0.25

Table 2. Kendall correlations among Standard measures and ADM.

The reasons can be understood by analyzing the score distributions. Let's start from the ADM (score) case. Figure 3 shows (black lines) the URS step function for a sample (and representative) topic. The height of the steps depends on the numerical scores assigned as URS to the 4 relevance levels S, A, B, and C (as above said, 1/8, 3/8, 5/8, and 7/8). The x-axis is truncated at about the 500th document.

[3] Actually, to avoid a too high dependency on the outliers, we did not choose the maximum and the minimum, but the 5th and 995th SRS, thus leaving 5 on each side.

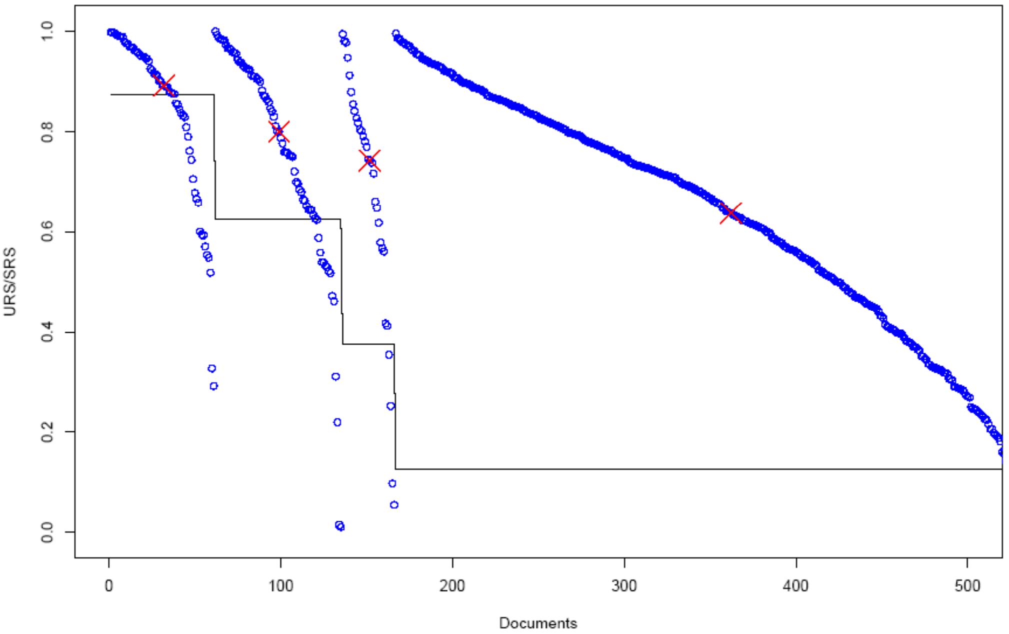


Figure 3. URS and SRS distributions for a given topic and an IRS.
The red "X" are the median values within each relevance category.

Figure 3 also shows (blue small cricles) the typical score distribution for an IRSs with high effectiveness according to standard effectiveness measure (this is the IRS with highest R-Prec) and rather low ADM, whereas Figure 4 shows a typical score distribution for a system with high ADM and rather low R-Prec. Comparing the two figures, we note that:

- The IRS in Figure 3 does a better job in discriminating the documents in the four categories of decreasing relevance S, A, B, and C, whereas the IRS in Figure 4 seems less effective in discriminating the 4 categories in the proper order. In other terms, the IRS in Figure 3 ranks the retrieved documents in a more effective way, whereas the IRS in Figure 4 does a very bad job in ranking the documents.
- However, from a different standpoint, the IRS in Figure 4 does a better job than the IRS in Figure 3: it approximates in a better way the step distribution of the URSs. Since for each topic the relevant (S, A, and B) documents are much fewer than the nonrelevant ones (C), the effects of SRSs on S, A, and B documents are negligible, and ADM depends on the SRSs assigned to C documents only. Indeed, on average, A documents are about 15, B documents are about 15, C documents are about 20 and C, or not assessed, documents are then about 1000 - (15+15+20) = 950.[4] Therefore, the ADM value for a given IRS depends on the average distance on the C assessed, or not assessed, documents in the right part of Figures 3 and 4, and this measure is not correlated to the effectiveness of an IRS measured with standard measures. For instance, an IRS with a steep decreasing and convex curve (as the one in Figure 4) has a higher ADM than a linear decreasing curve, which in turn has a higher ADM than a concave and mildly decreasing curve (as the one in Figure 3). While one may retain these as sterile, purely mathematical observations, they have also a practical effect on the usability of the score as actual measure of relevance for a specific document after a specific query. In fact, from a numerical point of view, a system like that shown in Figure 3 gives similar scores for relevant as well as most non relevant documents, even if it is able to discriminate among them. This means that differences in score value are not well related to differences in relevance.

A similar, though slightly different, argument holds for $ADM^{(Rank)}$: again, the ADM value mainly depends on the C-assessed documents, with the difference that this value is almost the same for all IRSs (since with $ADM^{(Rank)}$ the SRSs have the same slope for all IRSs). This means that the ADM values tend to be more similar to each other, thus lowering the correlation between ADM and golden measures. Indeed, the standard deviation is, on average, 0.16 for $ADM^{(Score)}$ and 0.07 for $ADM^{(Rank)}$ and IQR has similar values.

[4] The URS and SRS scores distributions in the two figures do not respect the average number of S, A, B, and C documents. We choose those two figures to have a better graphical understanding.

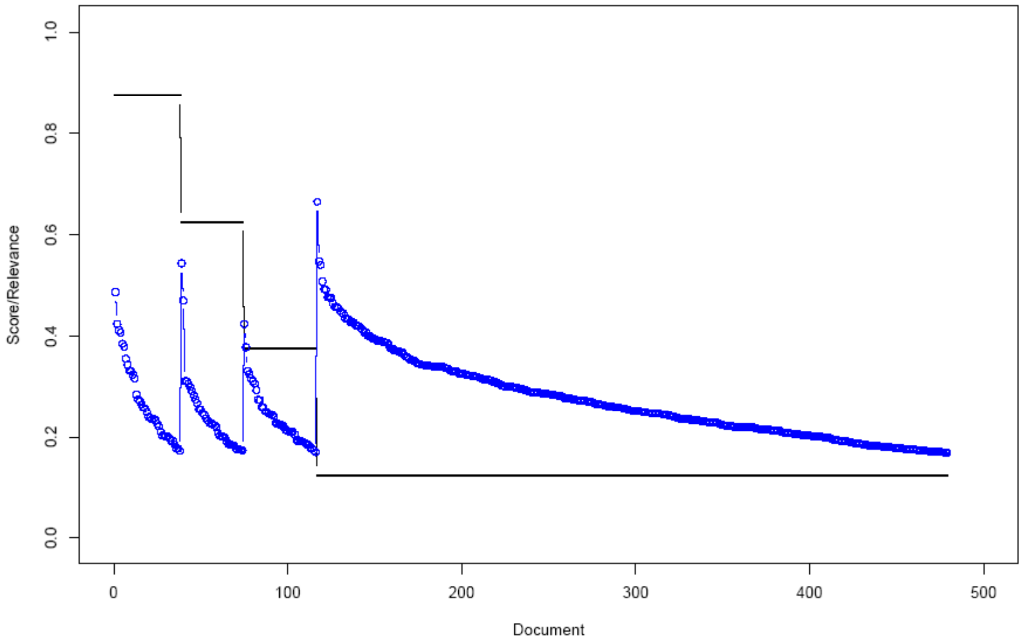


Figure 4. URS and SRS distributions for a given topic and an IRS.

4.2. Correlations with Standard IR Measures

Therefore, ADM on the whole set of retrieved documents turns out to be not effective. However, by restricting that set, in order to remove the documents on the right hand side of Figures 3 and 4 that lead to the above explained problems, we got more interesting results. We also relied on $ADM^{(Rank)}$ because the high dependency on the kind of the score distribution. We therefore defined $ADM_{(4)}^{(Rank)}[N]$ as $ADM_{(4)}^{(Rank)}$ measured using, for each IRS, only the first N documents that it has retrieved (and that have been assessed), with $N = 5, 10, 20, 50, 100, 200$.

Correlations among $ADM_{(4)}^{(Rank)}$, $ADM_{(4)}^{(Rank)}[5]$ $ADM_{(4)}^{(Rank)}[10]$, $ADM_{(4)}^{(Rank)}[20]$, $ADM_{(4)}^{(Rank)}[50]$, $ADM_{(4)}^{(Rank)}[100]$, and $ADM_{(4)}^{(Rank)}[200]$ on the one side and AvgPrec and R-Prec are shown in Table 3 and are of the same order of magnitude than correlation among the gold standard measures. This means that $ADM_{(4)}^{(Rank)}[N]$, with N = 5, 10, 20, 50, 100 could be another candidate measure of retrieval effectiveness.

	AvgPrec	R-Prec
$ADM_{(4)}^{(Rank)}[5]$	0.747	0.755
$ADM_{(4)}^{(Rank)}[10]$	0.792	0.802
$ADM_{(4)}^{(Rank)}[20]$	0.8	0.816
$ADM_{(4)}^{(Rank)}[50]$	0.788	0.799
$ADM_{(4)}^{(Rank)}[100]$	0.718	0.724
$ADM_{(4)}^{(Rank)}[200]$	0.129	0.126
$ADM^{(Rank)}$	0.35	0.37

Table 3. Kendall correlations among standard measures and ADM variants**.**

Data in Table 3 also show how the correlation between an ADM variant and standard measures drops suddenly after the first 100 documents. We have not yet an explanation for this phenomenon, that deserves further attention.

The usefulness of having four relevance levels is witnessed by $ADM_{(2)}^{(Rank)}[5]$ and $ADM_{(2)}^{(Rank)}[10]$, both relaxed and rigid, giving a lower correlation than the corresponding measures on 4 relevance levels, as shown in Table 4 (see also Table 3).

	AvgPrec	R-Prec
$ADM_{(2)}^{(Rank)}[5, relax]$	0.5	0.5
$ADM_{(2)}^{(Rank)}[10, relax]$	0.492	0.5
$ADM_{(2)}^{(Rank)}[5, rigid]$	0.439	0.45
$ADM_{(2)}^{(Rank)}[10, rigid]$	0.361	0.369

Table 4. Kendall correlations among Standard measures and ADM variants

5. Discussion

Results are not so clear as we hoped, but these are just preliminary findings and there is much future work to be done. Anyway, some of the data teach some useful lessons.

$ADM_{(2)}^{(Rank)}[N]$ is an interesting measure since it allows to measure IR effectiveness with very few documents, and with a much smaller pool than the one used in classical IR evaluation experiments. The information given by the four relevance levels can be usefully exploited.

Following the approach in (Della Mea & Mizzaro 2004), we also plan to evaluate whether ADM is sufficiently sensitive and stable to be safely measured using less data. The idea is that we may use less topics for the evaluation if the ADM, measured on a limited number of topics, correlates well with the ADM measured on all the topics and with the standard measures.

Given the limitations shown above, we also intend to perform further experiments with a pooling approach. For a given query, we might pool the first N (i.e., 5, 10, and so on) documents retrieved by each IRS and compute the ADM of each IRS on all the documents in the pool, with the assumption that if an IRS does not retrieve a document, then the SRS is zero. The rationale behind this is to avoid an ADM measure depending mainly (in practice, only) on the C assessed and not assessed documents, i.e., on the nonrelevant documents. We also avoid the paradoxical case of an IRS that gives a zero value for the SRSs of all the documents in the database, obtaining an ADM = 1.

From a more general viewpoint, it seems clear that IRSs do not carefully determine their own SRSs. This is not strange, since the effectiveness measures used so far are not sensible to variations in SRSs that preserve the rank (e.g., an IRS with a linear decreasing SRS distribution and an IRS with a quadratic decreasing SRS distribution get exactly the same evaluation by AvgPrec and R-Prec). However, to encourage the improvement of IRSs, it is important to arrive at a better estimation of the URS distribution.

It is important to understand that normalization is a crucial issue, and that normalization functions chosen by the designers of an IRS are more likely to be effective than those chosen by evaluators, as we did. Some IRSs seem to work with a SRS limited in an interval, whereas other IRSs work with an additive scheme, with no upper (or lower) limit for SRSs. These two kinds of IRSs need two different normalization strategies, aiming at either [0..1] or [0..+∞). Then, the latter can be mapped into the former with, e.g., the well known logistic transformation. Furthermore, the maximum and minimum SRS values within the same IRS are highly query dependent, thus making even more difficult to rely on SRS scores. Let us remark that this hinders the possible use of scores by the user as a quantitative indicator or relevance (or not) for the documents retrieved after a query. In fact, we may suppose users could look at the score to have an idea of the results quality; but in the examined IRSs, scores are rather inadequate to this aim.

Another interesting issue is whether the designer of an IRS participating in NTCIR should exploit the information that a four levels relevance scale is used, and therefore aim at a four levels distribution for the SRSs of his own engine. Indeed, according to ADM, the "perfect" (i.e., ADM = 1) IRS within NTCIR would have a step distribution of the SRSs. But, given the actual effectiveness of IRSs, collapsing an SRS to the most adequate relevance level might lead to higher distances and lower measured performance. Also, at least two approaches to the transformation of an SRS distribution into a step distribution can be foreseen: (i) a first one in which the relevance level closer to the SRS is chosen, and (ii) a second one in which the average number of documents in each level is exploited to get a better approximation. These are issues deserving further study.

Finally, as we mentioned in (Della Mea & Mizzaro 2004), SRSs are important not only for evaluation, but also for fusion of the results from different IRS (as it is done in some meta-search engines). We suggest to ask all the groups participating in next NTCIRs to have their IRSs to normalize their SRS in the [0..1] range: each normalization that preserves the rank will not modify the effectiveness evaluation according to standard measures, but ADM is capable of measuring the goodness of the distribution, and we believe this is an important result for the community.

References

Della Mea, V. & Mizzaro, S. (2004). Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):530-543.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37-49.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, 10(3):305-322.

Mizzaro, S. (2001). A new measure of retrieval effectiveness (Or: What’s wrong with precision and recall), In T. Ojala editor, *International Workshop on Information Retrieval (IR'2001)*, 43-52. Infotech Oulu, Oulu, Finland.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition, *Information Processing & Management*, 26(6), 755-776.

Voorhees, E. M. & Harman, D. (2000). Overview of the Eighth Text Retrieval Conference (TREC-8), *The 8th Text Retrieval Conference (TREC-8)*, 1-24, NIST SP-500-246, http://trec.nist.gov/.