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

**Vincenzo Della Mea**
Department of Mathematics and Computer Science, University of Udine
Via delle Scienze, 206 – I 33100 Udine, Italy
dellamea@dimi.uniud.it
http://www.dimi.uniud.it/dellamea/

**Luca Di Gaspero**
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/

**Stefano Mizzaro**
Department of Mathematics and Computer Science, University of Udine
Via delle Scienze, 206 – I 33100 Udine, Italy
mizzaro@dimi.uniud.it
http://www.dimi.uniud.it/mizzaro/

## 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 [21, 6]. In this paper we describe the idea on which ADM is based, a conceptual analysis of this new measure, and the results of a first experimental validation on TREC data, which feature 2-levels human relevance judgments and IR systems that rank the retrieved documents. Furthermore, we present some experimental results on NTCIR-3 data, which feature 4-levels human relevance judgments and IR systems that assign a numeric score to the retrieved documents. Both conceptual and experimental results show that ADM might be potentially adequate, provided that IR system designers take some care in system implementation.*

**Keywords:** *Information retrieval evaluation, Average Distance Measure, non-binary relevance.*

## 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 [21, 6] we have proposed and validated on TREC data a new IR effectiveness evaluation measure, which 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 [6]; on the other side, the analysis of NTCIR-3 data is still work in progress, and the results presented here are very preliminary ones.

The paper is structured as follows. In Section 2, we survey the problem of evaluating retrieval effectiveness, emphasizing some problems and the underlying assumptions of dichotomous conception of both relevance and retrieval. Then we propose a novel standpoint based on non binary relevance and, in Section 3, we define ADM (Average Distance Measure), a new measure of retrieval effectiveness based on a continuous view of relevance and retrieval. In Section 4 we discuss ADM adequacy from a conceptual standpoint (highlighting that ADM overcomes some problems inherent in the effectiveness measures usually adopted in retrieval evaluation, namely precision and recall) and we also summarize the positive results of a previous experiment on TREC data. In Section 5 we present preliminary experimental results on using ADM to evaluate the IRSs participating in NTCIR-3 Workshop. In Section 6 we summarize some lessons learned from this preliminary work and sketch what needs to be done in order to be able to use ADM in next NTCIRs. The last section concludes the paper and sketches some future developments.

## 2 Measuring retrieval effectiveness

### 2.1 Some problems in measuring IR effectiveness

Traditionally, given an information need and the corresponding query, the database of documents is partitioned in two ways, as it is graphically represented in Fig. 1(a), adapted from [23]: (i) between relevant and not relevant items, and (ii) between retrieved and not retrieved items. A reason of this approach is historical: the first IRSs were boolean, and they indeed provided a clear cut distinction between retrieved and nonretrieved documents; from that, it is (and, probably, has been) a small step to assume that relevance is binary as well, and, given the binary conceptions of relevant and retrieved documents, the definition of precision (i.e., the proportion of retrieved documents that are relevant) and recall (i.e., the proportion of relevant documents that are retrieved) is (has been) a logical consequence.

Actually, the two underlying assumptions (binary relevance and binary retrieval) have been questioned for long time. On the one side, after the first IRSs based on the vector space and probabilistic models, it has been clear that an IRS does not "either retrieve or not retrieve a document", but it rather ranks all the documents in the database on the basis of some system-assigned weight. This is widely known today, since everybody has experienced some search engine. On the other side, the long record of research on relevance [19] indicates that neither relevance is binary, nor binary judgments seem the most adequate method of expression [3, 7, 9, 12, 14, 15].

Indeed, some measures that go beyond the binary relevance, binary retrieval view have been proposed, most of them are well known (and described in standard IR textbooks, see, e.g., [29, Ch. 7], [23, Ch. 5]; [17, Ch. 8]), and are sometimes used. Let alone the other measures based on the same assumptions (i.e., fallout, generality factor, E-measure, mean average precision, and so on), by discarding the binary retrieval assumption we obtain measures based on the ranking induced by the IRS (i.e., normalized precision and recall, expected search length) or even on a continuous measure provided by the IRS (e.g., Swets's E-measure). If we also discard the binary relevance assumption, we obtain measures that can be classified in three groups:

- Measures based on categories of relevance and the rank produced by the IRS, e.g., Ranked Half Life [2] or Discounted Cumulative Gain [16].
- Measures that compare the ranking induced by the IRS with the ideal one, e.g., ndpm [32] or usefulness measure [10].
- Measures that evaluate the IR effectiveness using continuous values of relevance and retrieval, like the sliding ratio.

However, precision and recall have survived all these discussions, and are still widely used as the standard measures of IR evaluation. Still today, the standard practice is to evaluate IRSs by precision and recall, and, therefore, on the basis of the binary relevance and retrieval assumptions: in IR evaluation, often (if not usually) IRSs are meant to either retrieve or not retrieve a document, and human relevance judgments are dichotomous ones. The well known TREC experiment series is an example of this approach, even if in TREC the binary retrieval view is in some way smoothed by the procedure requiring 1000 ranked documents being returned by each system, and the adopted effectiveness measures are derivations of precision and recall.

The standard practice is so deeply rooted that, even when human relevance judgments are not dichotomous (i.e., they are expressed either by means of a scale of categories, or on a continuum), often precision and recall cause a "binarization" of the judgments, and NTCIR is not an exception to this practice. For example, it is often assumed that, on a three levels scale (i.e., nonrelevant, partially relevant, and relevant), the partially relevant items collapse into relevant ones [24, 27] and/or (less frequently) into nonrel-

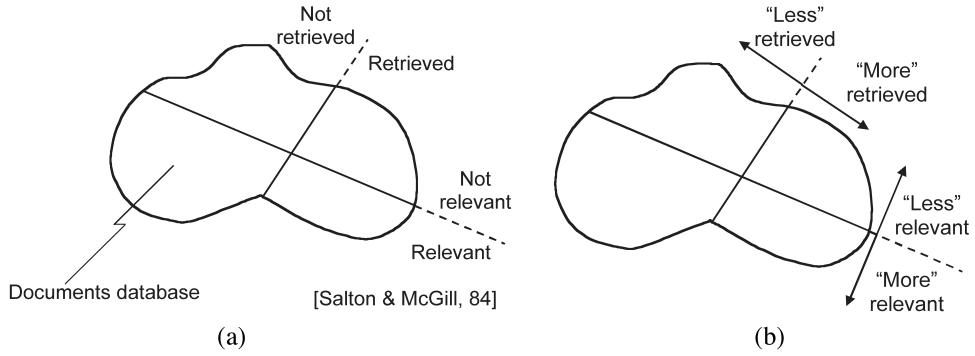


**Figure 1. From binary relevance and retrieval to continuous relevance and retrieval.**

evant ones [30]; also continuous judgments are binarized [8, 9, 22, 24]. Even if there is some experimental motivation for preferring "relevant" to "nonrelevant" when collapsing "partially relevant" [7, 8, 13], there is absolutely no reason for binarizing the relevance judgments (apart from being able to compute precision and recall). Similar phenomena do happen on the retrieval side too, where it is common to speak of "the retrieved documents", or of "the first page of documents retrieved by the search engine X". Moreover, the error rates for commonly used measures are far from being negligible, so that, for a reliable IR evaluation experiment, 50 queries are needed, and for having a significant difference between two IRSs, a 10% difference in IR performances is needed [4, 5, 26]. Finally, the measures obtained starting from artificially binarized figures are eventually averaged on many data, thus obtaining the rather peculiar result of continuous values.

Therefore, the IR field is in a curious situation: on the one side we have a "conceptual" standard, since almost everybody agrees that relevance and retrieval are matter of degree (three or more categories, if not a continuum); on the other side, we have an old habit, the "precision-recall old standard", that relies on the assumption of binary relevance and retrieval. This situation has the consequence that most of the evaluation experiments disregard the "conceptual" standard, thus hindering IR development and evaluation.

This dissonance is dangerous since researchers risk: (i) to evaluate in the wrong way the IRSs that they are developing; (ii) to develop "wrong" IRSs, i.e., IRSs that are evaluated as effective by the wrong measures, but that are not so effective; and (iii) to make more effort than needed for evaluating IR effectiveness.

We propose a novel approach.

## 2.2 From binary to continuous relevance and retrieval

We generalize Fig. 1(a) as shown in Fig. 1(b): in place of two clear cut partitions, we have gradients of relevance and retrieval. By going one step further, we make explicit the two figures that measure relevance and retrieval. As far as relevance is concerned, 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. Boolean IRSs give either SRS = 0 or SRS = 1.[1]

On the basis of the definitions of URS and SRS, we can slightly change the graphical representation in Fig.1(b), obtaining Fig. 2(a), that shows a URS-SRS plane, in which each document is a point with its own URS and SRS values (in the figure, u and s are these values for one document, represented by the point in the lower right corner).

This representation emphasizes how the dichotomies relevant-nonrelevant and retrieved-nonretrieved correspond to the (somewhat artificial and hardly justifiable) choice of two thresholds on the SRS and URS values. Fig. 2(b) is yet another representation of the same scenario, with the color shading representing the two gradients. In this figure, the ellipses show which documents concur to determining precision (P) and recall (R). Indeed, on the basis of Fig. 2, one might define precision, recall, fallout, and generality factor in the following way:

$$P = \frac{|\beta|}{|\alpha| + |\beta|}, \quad R = \frac{|\beta|}{|\beta| + |\delta|},$$

[1] SRS is similar to Retrieval Status Value (RSV) [1], but there is a difference: RSVs are used only to rank the documents and, therefore, any transformation of a RSV distribution that preserves the ranking is another equivalent RSV distribution. This is not the case for SRS, as we will discuss in the following examples. The difference stems from the underlying notion of relevance: RSV is based on binary relevance, SRS on continuous.

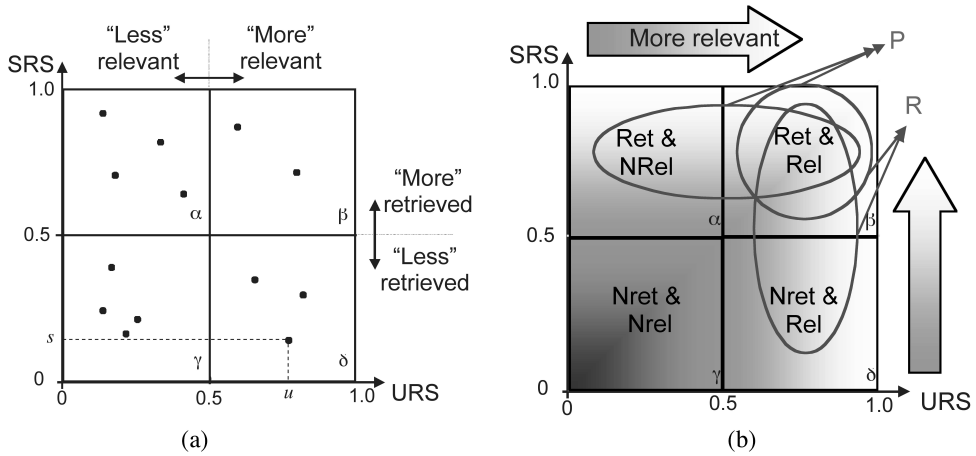


**Figure 2. The URS-SRS plane.**

$$F = \frac{|\alpha|}{|\alpha| + |\gamma|}, \quad G = \frac{|\beta| + |\delta|}{|\alpha| + |\beta| + |\gamma| + |\delta|}$$

where $|\alpha|$, $|\beta|$, $|\gamma|$, and $|\delta|$ are the numbers of points, i.e., documents, in the $\alpha$, $\beta$, $\gamma$, and $\delta$ sectors, respectively. Of course, one might choose on each axis two thresholds (or, in general, $n$ thresholds) and single out, in this way, nine regions (or, in general, $(n+1)^2$ regions). However, an even more general case is the continuous one, that we exploit to define a new measure of retrieval effectiveness, as it is shown in the next section.

Of course, there is the problem of collecting URS and SRS values. On the one hand, obtaining URSs seems feasible in various ways. One could simply use standard dichotomous — or category — relevance judgments: by averaging several such judgments, by different judges, on the same document-query pair, a continuous value is obtained. Or, one could use magnitude estimation techniques: line length and force hand grip have been used in the past rather effectively [3, 12, 14, 22].

On the other hand, to have IRSs computing true SRSs requires new IR models and a new approach to IRSs implementation. At a first stage, one might think of using probabilistic and vector space IRSs, but it is important to note that both the estimation of the probability of relevance given by a probabilistic IRS and the distance between the query and document vectors given by a vector space IRS are not estimation of the amount of relevance of a document to a query. To obtain such an estimation, new IRSs based on new IR models are needed.

A last important issue that we mention is the apparent arbitrary of URSs and SRSs. Even though URS might seem arbitrary at first, they turn out to be not arbitrary at all if they can be elicited reliably and consistently from human relevance assessors. And the above cited studies on magnitude estimation techniques [3, 12, 14, 22] are some first positive results in this direction. Now, if URS are not arbitrary, SRSs turn out not to be arbitrary too: the correct SRS for a document with respect to a query is the URS of that document for that query. This natural observation leads to the evaluation measure proposed in the next section.

## 3 The Average Distance Measure

We propose a new retrieval effectiveness measure, named *Average Distance Measure* (ADM), based on the average distance, or difference, between URSs (the actual relevance of documents) and SRSs (their estimates by the IRS). To have 0 as the minimum value, and 1 as the maximum value (as it is common in IR evaluation), we subtract the average distance from 1. In a more formal way, 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 $\text{SRS}_q(d_i)$), and the URS for $d_i$ with respect to $q$ ($\text{URS}_q(d_i)$). ADM for the query $q$ is then defined as the average distance between $\text{SRS}_q(d_i)$ and $\text{URS}_q(d_i)$:

$$\text{ADM}_q = 1 - \frac{\Sigma_{d_i \in D} \, |\text{SRS}_q(d_i) - \text{URS}_q(d_i)|}{|D|} \qquad (1)$$

(where the denominator is the number of documents in the database $D$). $\text{ADM}_q$ is in the $[0, 1]$ range, with 0 representing the worst performance. By averaging on some queries we obtain ADM, a measure of IR effectiveness.

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 Fig. 3). 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,[2] 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.

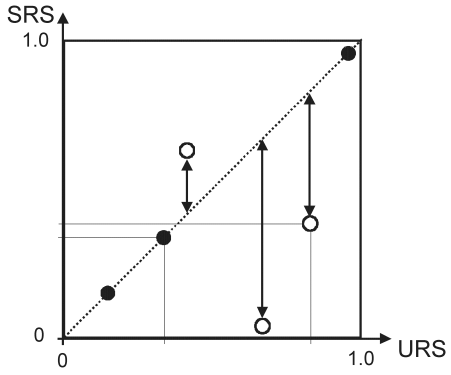


**Figure 3. Graphical representation of ADM.**

Let's see an example. Tab. 1 shows three hypothetical documents, with their URSs and the corresponding SRSs for three different IRSs. The last four columns of the table contain the values for precision, recall, E-measure (defined here as the mean between precision and recall), and ADM for the three IRSs, under the assumption that both the thresholds, between relevant and nonrelevant, and between retrieved and nonretrieved, are 0.5 (values $\geq 0.5$ are bold in the table). See also Fig. 4, where circles are IRS1 points, crosses are IRS2 points, and squares are IRS3 points.

| Docs. | $d_1$ | $d_2$ | $d_3$ | P | R | E | ADM |
|---|---|---|---|---|---|---|---|
| URS | **0.8** | 0.4 | 0.1 | | | | |
| IRS1 | **0.9** | **0.5** | 0.2 | 0.5 | 1 | 0.75 | 0.9 |
| IRS2 | **1.0** | **0.6** | 0.3 | 0.5 | 1 | 0.75 | 0.8 |
| IRS3 | **0.8** | 0.4 | **1.0** | 0.5 | 1 | 0.75 | 0.7 |

**Table 1. An example.**

Let us briefly analyze this example (more detailed discussion about ADM follows in the next section).

[2]In this paper we do not take into account the subjective and dynamic nature of relevance [20, 25], 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.

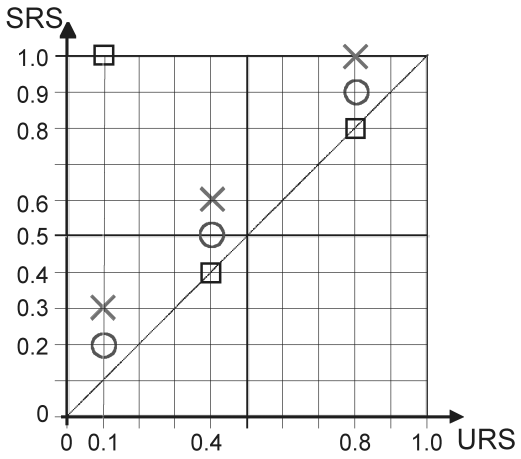


**Figure 4. Graphical representation of the example in Tab. 1.**

System IRS1 performs constantly better than IRS2 (each circle is closer to the ideal SRS=URS line than the corresponding cross); this is not reflected in the values of the three classical measures, whereas ADM captures the difference in effectiveness. Systems IRS1 (circles) and IRS3 (squares) are more difficult to compare, since IRS3 performs better than IRS1 on all but one of the documents ($d_3$), but on $d_3$ the SRS by IRS3 is really wrong. Precision, recall, and E-measure for IRS1 and IRS3 do not differ, whereas there is a difference in the two ADM values.[3]

Also, specialized forms of ADM can be defined. ADM can be specialized into an $\text{ADM}^{(2)}_{(2)}$ 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 $\text{ADM}^{(M)}_{(N)}$, 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), $\text{ADM}^{(M)}$ (with $M$ categories of retrieval and continuous relevance), and $\text{ADM}_{(N)}$ (with $N$ categories of relevance and continuous retrieval).[4]

[3]The SRSs given by IRS1 and IRS2 lead to the same ranking of the three documents. Therefore, they are equivalent if interpreted as RSV (see Footnote 1). However, if the IRSs have the aim of finding the best approximation of URSs, IRS1 is more effective than IRS2

[4]The assignment of numerical value to ordinal categories can present subtle problems (this issue has 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 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

Finally, ADM can be tuned in a very simple way, just by selecting the sample of documents used for its computation. For example, if only the most relevant documents are used, one measures the accuracy of the IRS in estimating the user relevance on the highly relevant documents only, and this seems a very important measure from the user point of view [16, 30].

# 4 Adequacy of ADM

## 4.1 Conceptual analysis

In this section we show, from a conceptual point of view, how ADM is adequate for measuring the effectiveness of IRSs, in some respect even more adequate than classical precision and recall.

ADM satisfies the classical four desirable properties proposed by Swets [28] and reported also in [29, Ch. 7]: it measures the effectiveness only, isolating it from efficiency and cost; it expresses the discrimination power of IRSs, independently of any acceptance criterion employed; it is a single number; and it allows complete ordering of different performances. Of course, ADM is not the only IR effectiveness measure that satisfies these properties (e.g., the E-measure does), nor these four properties guarantee that ADM is a good measure, since they are necessary and not sufficient conditions.

ADM adequacy is clearly shown when we compare it with other IR effectiveness measures usually adopted in retrieval evaluation. What follows concerns mainly precision and recall, but it can be generalized to other measures as well. This comparison, besides being useful for discussing ADM adequacy, will also lead us to reconsider the classical effectiveness measures, by highlighting their intrinsic limitations.

We can compare ADM with precision and recall on the basis of Fig. 2. ADM is, in some sense, more general, since:

- Precision and recall take into account the documents in some of the four sectors only (e.g., precision is based on sectors a and b only). If, in Fig. 2(a), some points were added to the $\gamma$ sector, either close to the ideal line or far from it, neither precision nor recall would be affected. However, if the points were close to (far from) the ideal SRS = URS line, this would mean that the IRS has correctly (wrongly) estimated the relevance of the corresponding documents, and therefore its effectiveness measure should increase (decrease). This is also a justification for preferring the recall-fallout pair to the recall-precision one: the former covers the whole $[0, 1]^2$ sector, while the latter covers just 75% of it ($\alpha$, $\beta$, and $\delta$), and the 75% with less documents in it, since most of them will be in the $\gamma$ sector (in general, given a query, most of the documents are neither relevant nor retrieved).

- Precision and recall do not use the full-fledged distance from the ideal line used in Eq. 1, since all the documents within each sector ($\alpha$, $\beta$, $\gamma$, and $\delta$) are considered as equivalent (the distance used is 0 if the document is in sector $\beta$ or $\gamma$, 1 if the document is in sector $\alpha$ or $\delta$: the same limitation of $\mathrm{ADM}^{(2)}_{(2)}$).

This comparison between ADM on the one side and precision and recall on the other shows how rough precision and recall are. The second point above also reveals two further problems. First, precision and recall are highly (too) sensitive to the thresholds chosen and to the documents close to the borders between sectors. Fig. 5(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 (Tab. 2(a)) show that classical measures are rather different, whereas ADM is more stable.

| | 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 2. Effectiveness measures for Figs. 5(a) and (b).**

The second problem is that precision and recall are not sensitive enough to important differences between systems. Fig. 5(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 (Tab. 2(b)) show that classical measures are completely unable to grasp the difference, whereas

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.

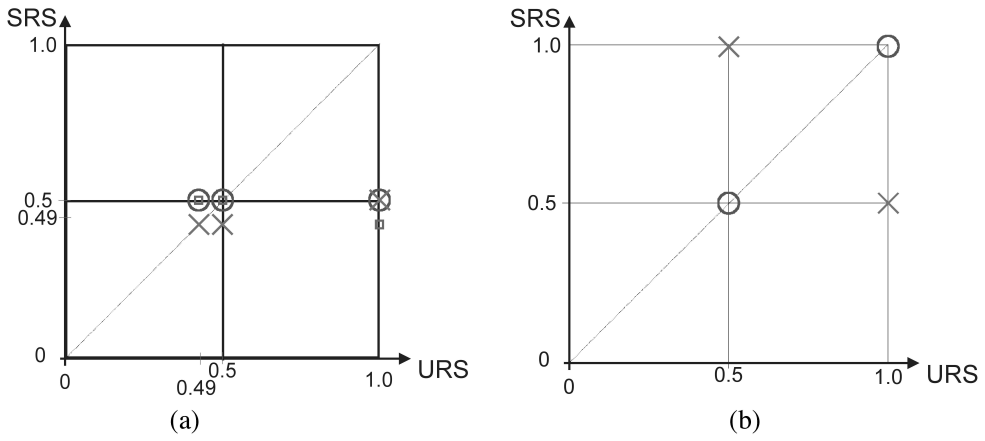


**Figure 5. Small (a) and big (b) differences in SRS values.**

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) [11]. 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.

Both problems depend on the thresholds on SRS and URS. The second one, however, has a further component: the equal status given to documents within each sector ($\alpha$, $\beta$, $\gamma$, and $\delta$, see Fig. 2(a)) in the calculation of precision and recall. Indeed, it seems not fair to consider all the documents in, say, $\beta$ sector simply as "retrieved and relevant"; a fairer categorization might be the one shown in Fig. 6(a), where the documents in the brighter area $\alpha 1$ (closer to the ideal line) are considered as correctly evaluated (their URS-SRS distance is below a given threshold value, $t$ in figure), whereas the document in the darker areas $\beta 1$ and $\gamma 1$ are not correctly evaluated. Correct evaluation, of course, leads to higher IR effectiveness.

On the basis of this categorization, one could define two substitutes for precision and recall. Let us start by noticing that recall can be considered an inverse measure of relevance under-evaluation (regarding relevant, i.e., 1, as higher than nonrelevant, i.e., 0), since it depends on the number of relevant documents considered not relevant. In the same way, precision is an inverse measure of relevance over-evaluation.

Starting from these properties, we may remark that under-evaluation can be expressed also as an IRS assigning SRS values that are lower than the URS values (leading to lower recall values): the points (documents) that should be placed in the upper right corner tend to be moved in the lower right corner of the URS-SRS plane, and not retrieved (think, for example, of a document that is relevant, i.e., URS close to 1, but is under-evaluated as not relevant, i.e., SRS close to 0). Conversely, over-evaluation, can be described as an IRS assigning SRS values higher than the URS values, leading to lower precision values, since more documents (points) are retrieved (moved toward the upper zone of the URS-SRS plane).

On the basis of these remarks, two hypothetic measures replacing precision and recall might be defined as:

$$P^* = \frac{|\alpha 1|}{|\alpha 1| + |\gamma 1|}, \quad R^* = \frac{|\alpha 1|}{|\alpha 1| + |\beta 1|},$$

(where, as usual, $|\alpha 1|$ is the number of documents in the $\alpha 1$ sector, and the same for the other sectors). $P^*$ inversely measures the number of over-evaluated documents, in the same way as precision inversely characterizes the number of nonrelevant documents retrieved by the IRS. Similarly, $R^*$ inversely measures the number of under-evaluated documents, i.e., relevant documents "forgotten" by the IRS.

However, these two measures are threshold based, and can be the subject of exactly the same critiques above presented about precision and recall (though the thresholds are chosen in a more sensible way). More specifically, which value to choose for the threshold $t$? The ideal zero is not feasible because most of the points will not lie on the SRS = URS line, and any other value is completely arbitrary. To avoid these

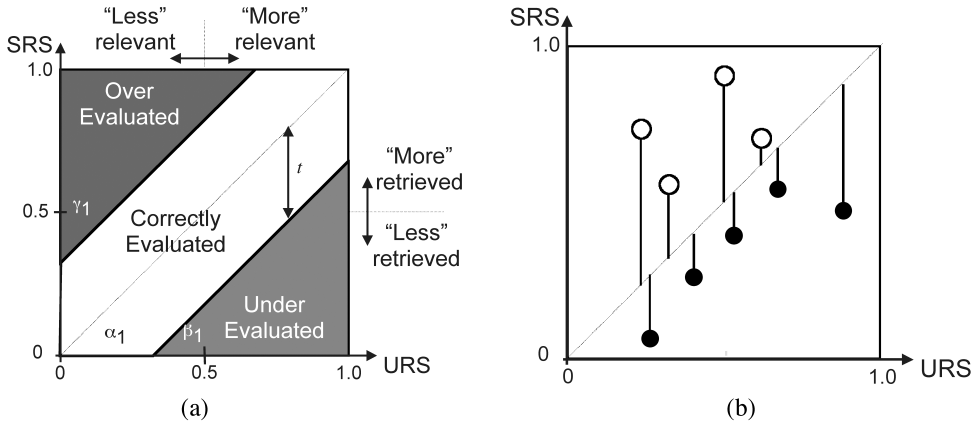


**Figure 6. A better categorization than the classical one in Fig. 2(b).**

critiques, and by exploiting the full potential of the continuous ADM measure, we can instead define the following two ADM measures, reflecting the original precision-recall pair: (i) *Average Distance Precision* (*ADP*), that is ADM computed on the over-evaluated documents only (i.e., those above the ideal SRS = URS line), and (ii) *Average Distance Recall* (*ADR*) that is ADM computed on the under-evaluated documents only. In formulæ, for each query $q$:

$$\mathrm{ADP}_q = 1 - \frac{\Sigma_{d_i \in \mathrm{DO}} |\mathrm{SRS}_q(d_i) - \mathrm{URS}_q(d_i)|}{|D|}$$

$$\mathrm{ADR}_q = 1 - \frac{\Sigma_{d_i \in \mathrm{DU}} |\mathrm{SRS}_q(d_i) - \mathrm{URS}_q(d_i)|}{|D|}$$

(where DO and DU are the sets of all the over- and under-evaluated documents, respectively, and the average distance is subtracted from 1 to have 1 as the value of higher effectiveness). Their graphical representation is shown in Fig. 6(b): $\mathrm{ADP}_q$ is the ADM computed only on the white points above the ideal SRS = URS line, whereas $\mathrm{ADR}_q$ is ADM computed only on the black points below the SRS = URS line.

ADP and ADR are continuous versions of precision and recall, respectively. Moreover, with these definitions, we have the nice property that, for each query $q$,

$$\mathrm{ADM}_q = \mathrm{ADP}_q + \mathrm{ADR}_q - 1.$$

### 4.2 Experimental analysis on TREC data

In [6] 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 [31]. 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 [6]; Tab. 3 shows the main results: Kendall correlations among standard measures on the one side and ADM measure on the other side are of the same order of magnitude than Kendall correlations among standard measures themselves. Moreover, the last two rows of the table show that ADM computed on subsets of retrieved, relevant, and topics are high enough to support the above presented conceptual analysis. 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.

| | R-Prec | Rel-Ret | ADM |
|---|---|---|---|
| AvgPrec | 0.90 | 0.82 | 0.88 |
| R-Prec | | 0.81 | 0.84 |
| Rel-Ret | | | 0.89 |
| ADM' (100%,100%,50%) | | | 0.85 |
| ADM" (100%,0%,100%) | | | 0.94 |

**Table 3. Kendall correlations among ADM and standard measures on TREC data.**

## 5 ADM experiments on NTCIR-3 data

In order to evaluate ADM on IRSs not based on the binary relevance/binary retrieval paradigm, we have been working on data from NTCIR-3, which features documents judged on a four-level scale and IRSs providing continuous retrieval values, thus allowing us to test ADM on more suited data.

We evaluated ADM starting from URSs and SRSs as above defined. Then, we compared ADM with two of the traditional effectiveness measures used by

NTCIR-3 and TREC, i.e., Average Precision (AvgPrec) and R-Precision (R-Prec). We based the comparison on the Kendall correlation, as we did in our previous work [6]. 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 [6], 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 performing these 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.

### 5.1 NTCIR-3 data

The data set included results from about 50 queries (44 on average) and 14 IRSs working to retrieve documents within and across four languages (Chinese, English, Japanese, and Korean) for a total of 114 runs. In the submission form adopted by NTCIR, 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.

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"), and not relevant ("C").

Since judgments are expressed in categories, we had to convert them into numeric URSs, i.e., into real values in the $[0..1]$ range. This is somehow an arbitrary choice (see Footnote 4); we chose to use a linear scale (alternatives like logarithmic scales are of course possible), and to split the $[0..1]$ range into four subintervals of equal length (i.e., $[0..\frac{2}{8})$, $[\frac{2}{8}..\frac{4}{8})$, $[\frac{4}{8}..\frac{6}{8})$, and $[\frac{6}{8}..1]$). This leads to using the following values: $S = \frac{7}{8}$, $A = \frac{5}{8}$, $B = \frac{3}{8}$, $C = \frac{1}{8}$.

### 5.2 Rank-based ADM

As a first experiment, we adopted the same schema used on TREC data: we did not take into account the SRSs given by the IRSs; rather, we obtained SRSs by converting the rank into a normalized measure of the position and we computed $\text{ADM}^{\text{rank}}_{(4)}$ (ADM computed on the basis of the rank given by the IRSs to the retrieved documents using the four relevance levels). Being 1000 documents retrieved by each system for each query, the first ranked documents were assigned SRS = 1, the second ranked ones SRS = 0.999, until position 1000, with value 0.001; zero for all other documents. We also defined $\text{ADM}^{\text{rank}}_{(4)}@N$ as measured using, for each IRS, only the first $N$ documents that it has retrieved (and that have been assessed). We experimented with the values $N = 5, 10, 20, 50, 100, 200$.

Correlations among $\text{ADM}^{\text{rank}}_{(4)}@N$ and $\text{ADM}^{\text{rank}}_{(4)}$ on the one side and AvgPrec and R-Prec on the other side are shown in Tab. 4. We used the relaxed version; the figures for rigid ones are similar when not otherwise explicitly stated. Correlation values are of the same order of magnitude (although about 0.1 lower) than correlation among the standard measures. This confirms the results obtained on TREC data : $\text{ADM}^{\text{rank}}_{(4)}@N$, with $N = 5, 10, 20, 50, 100$ could be another candidate measure of retrieval effectiveness. However, $\text{ADM}^{\text{rank}}_{(4)}@200$ and the full-fledged $\text{ADM}^{\text{rank}}_{(4)}$ show surprising low correlations, and the sudden decrease after $N = 100$ is surprising as well. Possible reasons of this behavior are discussed in the next subsection.

| | AvgPrec | R-Prec |
|---|---|---|
| $\text{ADM}^{\text{rank}}_{(4)}@5$ | 0.75 | 0.76 |
| $\text{ADM}^{\text{rank}}_{(4)}@10$ | 0.79 | 0.80 |
| $\text{ADM}^{\text{rank}}_{(4)}@20$ | 0.80 | 0.82 |
| $\text{ADM}^{\text{rank}}_{(4)}@50$ | 0.79 | 0.80 |
| $\text{ADM}^{\text{rank}}_{(4)}@100$ | 0.72 | 0.72 |
| $\text{ADM}^{\text{rank}}_{(4)}@200$ | 0.13 | 0.13 |
| $\text{ADM}^{\text{rank}}_{(4)}$ | 0.35 | 0.37 |

**Table 4. Kendall correlations between $\text{ADM}^{\text{rank}}_{(4)}$ and standard measures.**

Using four relevance levels in place of the classical two levels used in TREC is useful, as witnessed, for intance, by $\text{ADM}^{\text{rank}}_{(2)}@5$ and $\text{ADM}^{\text{rank}}_{(2)}@10$, both relaxed and rigid, giving a lower correlation than the corresponding measures on 4 relevance levels (see Tab. 5 and compare to the first two rows in Tab. 4).

| | AvgPrec | R-Prec |
|---|---|---|
| $\text{ADM}^{\text{rank}}_{(2)}@5$[relax] | 0.51 | 0.51 |
| $\text{ADM}^{\text{rank}}_{(2)}@10$[relax] | 0.54 | 0.54 |
| $\text{ADM}^{\text{rank}}_{(2)}@5$[rigid] | 0.52 | 0.53 |
| $\text{ADM}^{\text{rank}}_{(2)}@10$[rigid] | 0.58 | 0.59 |

**Table 5. Kendall correlations among $\text{ADM}^{\text{rank}}_{(2)}$ variants and standard measures.**

### 5.3 SRS-based ADM

From the ADM viewpoint, NTCIR data are interesting since they feature continuous SRSs, to which we turned our attention. Since different IRSs have SRSs ranging over different values (some IRSs return values in the $[0..1]$ range; others return an SRS between $10,000$ and $9,000$; other ones give even negative values, and so on), we had to normalize SRSs in the $[0..1]$ range. We adopted a simple linear normalization, in which the original minimum SRSs is mapped to $0$ and the maximum to $1$:

$$\text{SRS}_N = \frac{\text{SRS}_O - min(\text{SRS}_O)}{max(\text{SRS}_O) - min(\text{SRS}_O)}$$

(where $\text{SRS}_O$ is the original SRS and $\text{SRS}_N$ is the normalized value). We experimented on normalizing both within run (i.e., choosing the maximum and minimum values among all the SRSs expressed by an IRS on all the queries) and within query (i.e., choosing the maximum and minimum values among all the SRSs expressed by an IRS on a single query), obtaining similar results. Normalization is another subtle issue on which we will come back in the following.

Correlations among, on the one side, $\text{ADM}^{\text{score}}_{(4)}@N$ and $\text{ADM}^{\text{score}}_{(4)}$ and, on the other side, AvgPrec and R-Prec are shown in Tab. 6. $\text{ADM}^{\text{score}}_{(4)}$ does not correlate with standard measures. The reasons can be understood by analyzing the distributions of both SRSs and URSs. Fig. 7 shows (black lines) the URS step function for a sample query. 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, $\frac{7}{8}$, $\frac{5}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$). The x-axis is truncated at about the 500th document. The query is representative, i.e., other queries show similar distributions; actually, for graphical reasons, we chose a query with a number of S, A and B documents higher than usual.

| | AvgPrec | R-Prec |
|---|---|---|
| $\text{ADM}^{\text{score}}_{(4)}@5$ | 0.60 | 0.61 |
| $\text{ADM}^{\text{score}}_{(4)}@10$ | 0.4 | 0.41 |
| $\text{ADM}^{\text{score}}_{(4)}@20$ | 0.22 | 0.23 |
| $\text{ADM}^{\text{score}}_{(4)}@50$ | 0.09 | 0.09 |
| $\text{ADM}^{\text{score}}_{(4)}@100$ | -0.03 | -0.02 |
| $\text{ADM}^{\text{score}}_{(4)}@200$ | -0.17 | -0.16 |
| $\text{ADM}^{\text{score}}_{(4)}$ | 0.04 | 0.05 |

**Table 6. Kendall correlations between $\text{ADM}^{\text{score}}_{(4)}$ and standard measures.**

Fig. 7 also shows (blue small circles) the typical score distribution for an IRS with high effectiveness according to standard effectiveness measure (this is the IRS with highest R-Prec) and rather low ADM, whereas Fig. 8 shows a typical score distribution for a system with high ADM and rather low R-Prec (on a different query). Comparing the two figures, we can make two remarks.

The first is that the IRS in Fig. 7 ranks the retrieved documents in a more effective way (median values are decreasing), whereas the IRS in Fig. 8 does a very bad job in ranking the documents. In other terms, the IRS in Fig. 7 does a better job in discriminating the documents in the four categories of decreasing relevance S, A, B, and C, whereas the IRS in Fig. 8 seems less effective in discriminating the 4 categories in the proper order.

The second remark is that, taking a different standpoint, the IRS in Fig. 8 does a better job than the IRS in Fig. 7: 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 (or mainly). Indeed, on average, for each query S documents are about 20, A documents are about 40, B documents are about 40 and C documents are about 2500, let alone the non assessed documents. Therefore, ADM value for a given IRS depends on the average distance on the C assessed, or not assessed, documents in the right part of Figs. 7 and 8 (remember that we chose a query with a number of S, A and B documents higher than usual), and this measure is not correlated to the effectiveness of an IRS measured with standard measures. In other terms, we can say with good approximation that ADM measures the area between the SRS and URS curves on the right hand side of the figures. For instance, an IRS with a steep decreasing and convex curve (as the one in Fig. 8) has a higher ADM than a linear decreasing curve (as the one in Fig. 7), which in turn has a higher ADM than a concave and mildly decreasing curve.

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 Fig. 7 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 $\text{ADM}^{\text{rank}}$: 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 $\text{ADM}^{\text{rank}}$ the SRSs have the same —linear— 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 standard measures. In-

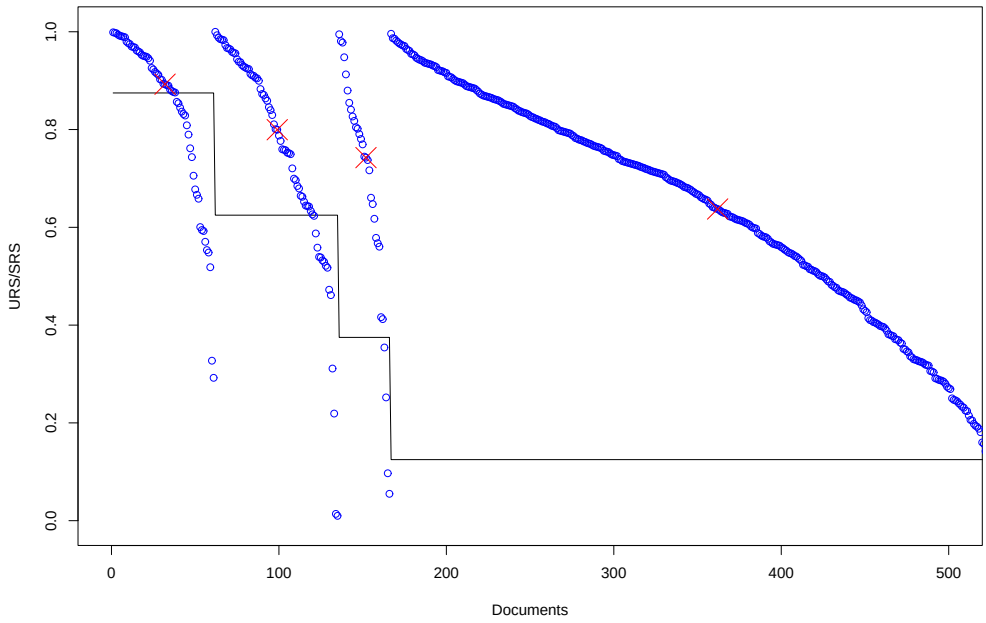


**Figure 7. URS and SRS distributions for a given topic and an IRS. The red Xs are the median values within each relevance category.**

deed, the standard deviation is, on average, 0.16 for $\text{ADM}^{\text{score}}$ and 0.07 for $\text{ADM}^{\text{rank}}$ and IQR has similar values.

The correctness of the above argument is witnessed by the higher correlations obtained when using less documents (see Tabs. 4 and 6): ADM on the whole set of retrieved documents turns out to be not effective; however, by restricting that set, and thus removing the documents on the right hand side of Figs. 7 and 8 that lead to the above explained problems, we got the higher correlation values. Moreover, by plotting the values of, e.g., $\text{ADM}^{\text{score}}_{(4)}@10$, versus AvgPrec, as is done in Fig. 9, one can also realize that the correlation is higher on some subsets of IRSs, i.e., those IRSs featuring similar score distributions (e.g., the IRSs in the lower area of Fig. 9 have a linear distribution of SRSs).

Actually, there is another phenomenon still lacking an explanation. Let us remark that the influence of the SRS distribution decreases when less documents are used to compute ADM. Indeed, $\text{ADM}^{\text{score}}_{(4)}@N$ is higher for low $N$ values, but not as high as $\text{ADM}^{\text{rank}}_{(4)}$. Therefore, the lower correlations in Tabs. 6 with respect to Tab. 4 must depend also on other reasons. We see several issues here. First, by using $\text{ADM}^{\text{rank}}_{(4)}$ we rely on the same information used in computing the standard measures, whereas by using $\text{ADM}^{\text{score}}_{(4)}$ we exploit information that is neglected by standard measures. Having said that, it is not surprising that correlation values decrease. Second, it is likely that the normalization scheme that we adopted favors some IRSs and hinders other ones. Third, language (of both topic and collection) and, perhaps, the part of topics used to formulate the query are other independent variables, that have an effect deserving further study.

## 6 Lessons learned

A first lesson is that SRS normalization must be done by system designers: they have all the necessary knowledge to choose the normalization schema that maximizes system performances. Indeed, some IRSs seem to work with an 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 $[0..+\infty)$. 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.

A second important results is that IRSs should not only try to optimize the rank of the retrieved documents; they should also try to approximate at their best the URSs distribution. After a preliminary analysis, it seems that it has a negative exponential shape; a least square fitting of the URS data leads to a function like:

$$f(x) \quad = \quad Ae^{-\frac{x}{\mu}} + C$$

with different parameter values for each language.[5]

A simple minded approach for IRS providing, for instance, a linear SRS distribution is a straightforward

[5] Let us remark that the variability across languages is a confirmation of the third issue raised at the end of the Section 5.3.

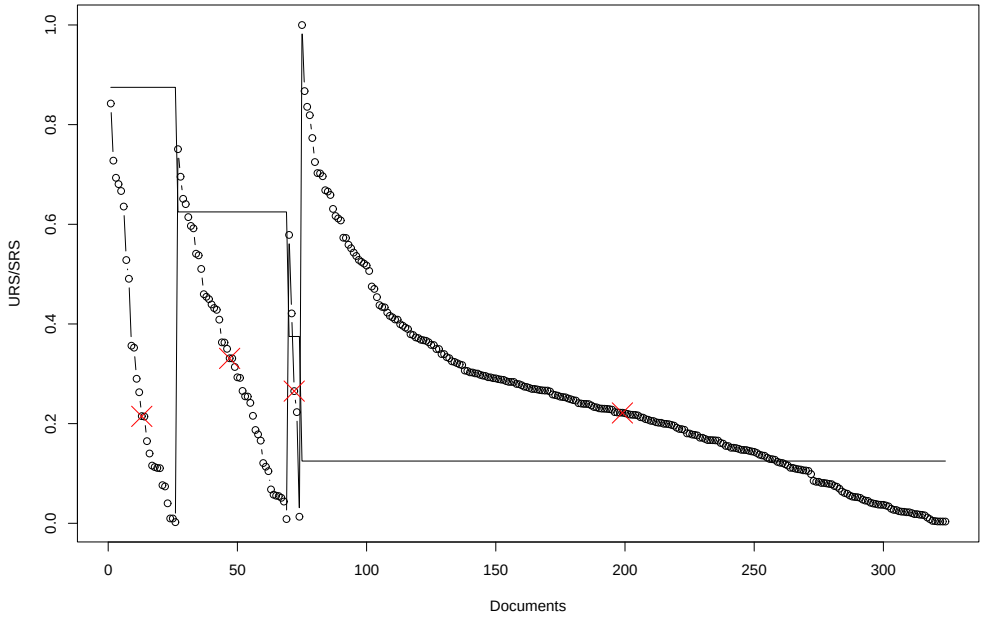


**Figure 8. URS and SRS distributions for a given topic and an IRS.**

normalization; more sophisticate solutions, e.g., based on a more careful study of the URS distribution, perhaps on a query-by-query basis, might lead to better approximations.

## 7 Conclusions and future work

We have discussed the issue of IR evaluation (Section 2) and presented the ADM measure, based on continuous views of relevance and retrieval (Section 3. After having recalled previous results (both conceptual and experimental on TREC data, Section 4), we have described the novel experimental results obtained on NTCIR-3 data. These novel 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 (Section 6).

We obtained a confirmation that $\text{ADM}^{\text{rank}}_{(4)}@N$ 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. This makes ADM a potentially interesting measure for terabyte collections, where the set of all the relevant documents is obviously unavailable. Another result is that the information given by the four relevance levels can be usefully exploited. $\text{ADM}^{\text{score}}$, conversely, presents some phenomena that are not fully understood yet.

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. Indeed, 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 would an important contribution to the IR community.

It is important to understand that normalization is a crucial issue, and that normalization functions chosen by the designers of an IRS are likely to be more effective than those chosen by evaluators, as we did in this paper. Let us also remark that SRSs might be exploited both in IRS output combination [18] and by users as a quantitative indicator of relevance or nonrelevance for the documents retrieved after a query (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

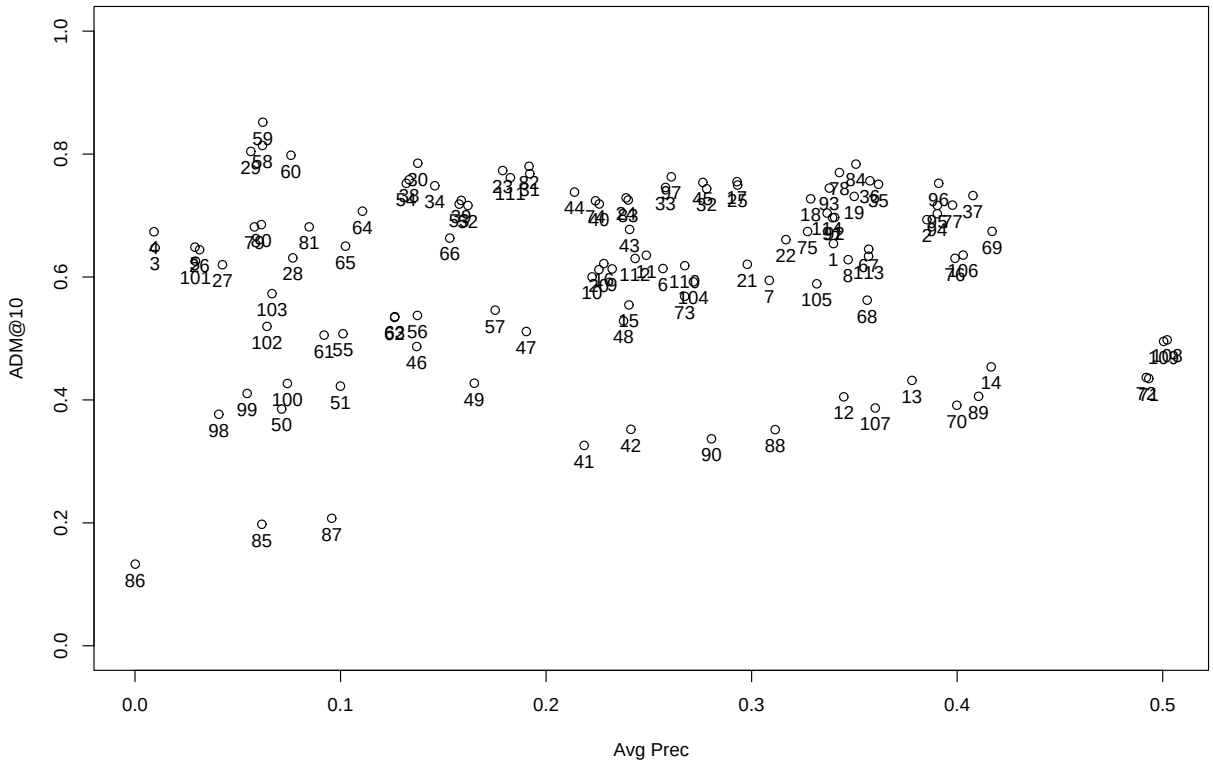


**Figure 9. Correlation between $ADM^{score}_{(4)}@10$ and AvgPrec.**

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 is likely to 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.

Given the limitations discussed 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.

Finally, it is important to get some new data free from the above problems. To this aim, we intend to evaluate IRSs participating in the next NTCIR-5 by ADM too. In this way, participants will be prompted to design IRSs that compute $[0..1]$ normalized SRSs and that try to approximate the URS distribution. The effort required to participants is limited, and we will provide `ADM Eval`, a software package to calculate ADM, which will be similar to the well known `TREC Eval`.

## Acknowledgements

We warmly thank NTCIR organizers, and especially Noriko Kando, for providing the experimental data.

## References

[1] A. Bookstein. Relevance. *Journal of the American Society for Information Science*, 30(5):269–273, 1979.

[2] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 324–331, Melbourne, Australia, 1998.

[3] H. W. Bruce. A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*, 45(3):142–148, 1994.

[4] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference*, pages 33–40, Athens, Greece, 2000.

[5] C. Buckley and E. Voorhees. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pages 316–323, Tampere, Finland, 2002.

[6] V. Della Mea and S. Mizzaro. 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, 2004.

[7] M. Eisenberg. *Magnitude Estimation and the Measurement of Relevance*. PhD thesis, Syracuse University, Syracuse, NY, 1986.

[8] M. Eisenberg and X. Hu. Dichotomous relevance judgments and the evaluation of information systems. In *Proceedings of the American Society for Information Science*, pages 66–69, Medford, NJ, 1987. Learned Information.

[9] M. B. Eisenberg. Measuring relevance judgments. *Information Processing & Management*, 24(4):373–389, 1988.

[10] H. Frei and P. Schauble. Determining the effectiveness of retrieval algorithms. *Information Processing and Management*, 27(2):153–164, 1991.

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

[12] J. W. Janes. The binary nature of continuous relevance judgments: A case study of users' perceptions. *Journal of the American Society for Information Science*, 42(10):754–756, 1991.

[13] J. W. Janes. Relevance judgments and the incremental presentation of document representations. *Information Processing & Management*, 27(6):629–646, 1991.

[14] J. W. Janes. Other people's judgments: A comparison of user's and other's judgments of document relevance, topicality, and utility. *Journal of the American Society for Information Science*, 45(3):160–171, Apr. 1994.

[15] J. W. Janes and R. McKinney. Relevance judgments of actual users and secondary judges. *Library Quarterly*, 62:150–168, 1992.

[16] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference*, pages 41–48, Athens, 2000.

[17] R. R. Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, 1997.

[18] R. Manmatha, T. Rath, and F. Feng. Modeling score distribution for combining the output of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 267–275, New Orleans, LA, 2001.

[19] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, Sept. 1997. John Wiley & Sons Inc., New York, NY.

[20] S. Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322, June 1998. ISSN: 0953-5438. Elsevier, The Netherlands.

[21] S. Mizzaro. 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)*, pages 43–52, Oulu, Finland, Sept. 2001. Infotech Oulu. ISBN: 951-42-6489-4.

[22] M. E. Rorvig. Psychometric measurement and information retrieval. In *Annual Review of Information Science and Technology*, volume 23, pages 157–189. 1988.

[23] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1984.

[24] L. Schamber. Relevance and information behavior. In *Annual Review of Information Science and Technology*, volume 29, pages 3–48. 1994.

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

[26] K. Spark Jones. Automatic indexing. *Journal of Documentation*, 30:393–432, 1974.

[27] A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management*, 34(5):599–621, 1998.

[28] J. A. Swets. *Effectiveness of Information Retrieval Methods*. Bolt, Beranek and Newman, Cambridge, MA, 1967.

[29] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[30] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 74–82, New Orleans, Louisiana, USA, 2001.

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

[32] Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.