A NEW MEASURE OF RETRIEVAL EFFECTIVENESS (OR: WHAT'S WRONG WITH PRECISION AND RECALL)

Stefano Mizzaro
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 — Loc. Rizzi
I33100 Udine, Italy
mizzaro@dimi.uniud.it
http://www.dimi.uniud.it/˜mizzaro

Abstract. Most common effectiveness measures for information retrieval systems 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 are questioned, and a new measure named ADM (Average Distance Measure) is proposed. ADM turns out to be both adequate to measure the effectiveness of information retrieval systems, and useful for revealing some problems about precision and recall.

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 ranking the retrieved documents and, sometimes, showing their weights, do already exist. In this paper, after a brief presentation of most common effectiveness measures (Sect. 2), a new measure is proposed (Sect. 3). It is shown that the measure seems adequate and is useful for understanding some problems related to the classical effectiveness measures precision and recall (Sect. 4). Sect. 5 concludes the paper.

2 Related work: Measures of retrieval effectiveness

Many measures of retrieval effectiveness have been proposed. In this section, some of them are recalled and grouped into categories; most of them are described in, e.g., [20, Ch. 7], [18, Ch. 5], [13, Ch. 8] (or in the references below).

Binary relevance and binary retrieval. The most known measures of retrieval effectiveness are based on binary notions of relevance (either a document is relevant or it is not) and retrieval (either a document is retrieved or it is not). This is probably due to an historical reason: the first IRSs were boolean, i.e., they either retrieved or not retrieved a document. The two dichotomies allow to speak of the sets of retrieved, nonretrieved, relevant, and nonrelevant documents, and to define the well known measures: precision and recall, fallout, generality factor, E-measure, mean average precision, and so on.

Binary relevance and ranking retrieval. The binary relevance, binary retrieval view changed, probably because of some developments in IRS (the coordination level [20, Ch. 5], the vector space model [18, Ch. 4], and the probabilistic models [20, Ch. 6]). After these developments, the documents database is no more divided into two subsets (retrieved and nonretrieved). Rather, the IRS assigns to each document in the database a weight that measures the document-query similarity; the weights allow to *rank* the documents in the database in decreasing order of similarity (i.e., most similar first); and the documents are presented to the user accordingly (some of them, those with the lowest—or zero—similarity, might be not presented). The weight assigned to each document in the database has been named *Retrieval Status Value* (RSV) [1]; I prefer (and use in the following) the term *System Relevance Estimate* (SRE). Actually, in the binary relevance and ranking retrieval view, it is not the SRE that is shown to the user, just the ranking that it induces. Accordingly, some measures that determine the effectiveness of a ranking (on the basis of the binary relevance view) are defined: normalized precision and recall, expected search length.

Binary relevance and continuous retrieval. Instead of simply using the ranking induced by SRE, one can fully use the SRE values potential, defining a measure that evaluates the goodness of the SRE values. Swets's E-measure is such a measure.

Ranking relevance and ranking retrieval. All the above measures assume binary relevance judgments. An obvious step is to go towards ranking relevance. The "relevance equivalent" of SRE can be defined: *User Relevance Estimate* (URE). Similarly to what said for SRE, the continuous values (say, in the $[0..1]$ range), representing the relevance of each document, can be used to derive preference relations between documents. Some measures based on the preference (in terms of relevance and retrieval) of one document on another one have been proposed: ndpm [21], usefulness measure [6].

Continuous relevance and continuous retrieval. In my opinion, the position that the ranking is important, not the SRE and URE, is wrong. Indeed, some IRSs do show the SRE (or a bar representing it); an advanced user interface might use virtual reality techniques to graphically show the space of documents; the user might rely on the set of SRE for deciding if the retrieved (actually, displayed) set of documents is good enough; an "intelligent" IRS might use the statistical distribution of the SREs for suggesting a query reformulation; and so on. The binary relevance assumption can be questioned too. In [10, 11, 12] line-length magnitude estimation was found reliable for exploring the consistency of relevance judgments. Bruce [2] empirically found that magnitude estimation (numeric estimation and hand grip) is appropriate to let the judge express the importance ascribed to various characteristics of documents and information.

If we rely on a continuous relevance judgment, the above mentioned measures are no more adequate. Moreover, the approach of introducing some thresholds to divide the documents into sets of relevant or retrieved (e.g., SRE $\geq$ 0.5 means retrieved; URE $\geq$ 0.5 means relevant) has some limitations, due to the somewhat arbitrary choice of the thresholds and to the experimentally demonstrated difficulty in choosing them: when judges collapse their scaled judgments into dichotomous judgments, the break between relevant and nonrelevant seems below 0.5 [3, 4, 9]. The alternative is to define effectiveness measures that exploit the full potential of SRE and URE, like the sliding ratio.

The path among the various categories of measures can be represented graphically as in Figure 1: from binary relevance and binary retrieval, through binary relevance and ranking

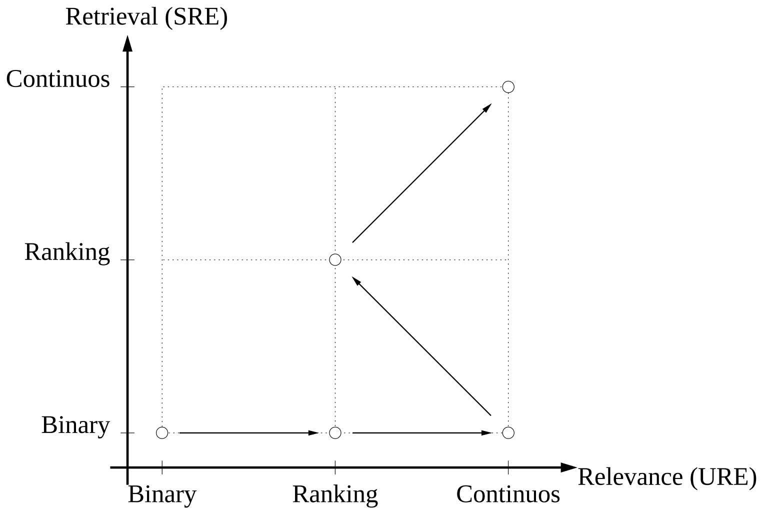


Figure 1: The path among the various combinations of relevance and retrieval categories.

retrieval, binary relevance and continuous retrieval, ranking relevance and ranking retrieval, we have reached the "top" continuous relevance and continuous retrieval. Of course, I have not considered some combinations. It is peculiar that most commonly used measures of IR effectiveness (precision and recall) are at the "bottom".

3 The average distance measure

I propose a new retrieval effectiveness measure, named *Average Distance Measure* (ADM), which simply measures the average distance—or difference—between UREs (the actual relevances of documents) and SREs (their estimates by the IRS). In a more formal way, for a given query q, we can define two relevance weights for each document d_i in the database D: the SRE for d_i with respect to q (I denote it with $\text{SRE}_q(d_i)$), and the URE for d_i with respect to q ($\text{URE}_q(d_i)$). ADM is then defined as the average distance between $\text{SRE}_q(d_i)$ and $\text{URE}_q(d_i)$:

$$\text{ADM}_q \quad = \quad 1 - \frac{\Sigma_{d_i \in D} \, |\text{SRE}_q(d_i) - \text{URE}_q(d_i)|}{|D|} \tag{1}$$

(where the denominator is the number of documents in the database D). ADM is in the $[0..1]$ range, with 0 representing the worst performance. Averaging ADM_q on some queries we obtain a measure of the effectiveness of an IRS.

We can graphically understand ADM in the following way. Let's assign to each document in the database its own SRE and URE values (in the $[0..1]$ range) and plot these values on a standard Cartesian diagram in the $[0..1]^2$ square (see Fig. 2). Each document is therefore a point in the URE–SRE plane; the closer the point to the ideal $\text{SRE} = \text{URE}$ line (the dotted line in the figure), the best the estimate by the IRS (the points on the line are represented by white circles in figure). The last thing we need to define is the distance between a point and the ideal line. Since the URE value is fixed, the distance is not the standard distance between a point and a line (obtained measuring 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. This is the definition used in Eq. (1).

Let's see an example. Tab. 1 shows five hypothetical documents, with their UREs and the corresponding SREs for three different IRSs. The last four columns of the table contain the

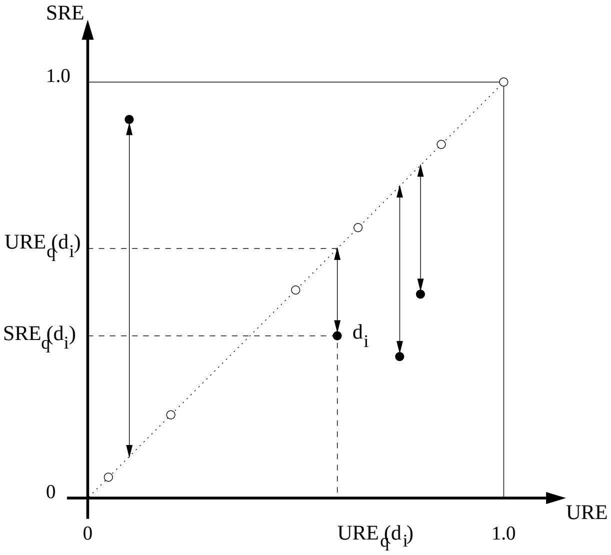


Figure 2: Graphical representation of ADM.

Table 1: An example.

Docs.	d_1	d_2	d_3	d_4	d_5	P	R	E	ADM
URE	**0.8**	**0.6**	0.4	0.2	0.1				
IRS1 (○)	**0.9**	**0.5**	**0.5**	0.1	0.2	0.67	1	0.84	0.9
IRS2 (×)	**1.0**	0.4	**0.6**	0.0	0.3	0.5	0.5	0.5	0.8
IRS3 (□)	**0.8**	**0.6**	0.4	0.2	**1.0**	0.67	1	0.84	0.8

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. 3, where circles are IRS1 points, crosses are IRS2 points, and squares are IRS3 points.

Let's briefly analyze this example (more detailed discussion about ADM follows in the next section). System IRS1 performs constantly better than IRS2 (each circle is closer to the ideal SRE $=$ URE line than the corresponding cross); this is reflected in all the values of the four measures. Systems IRS1 (circles) and IRS3 (squares) are more difficult to compare, since IRS3 performs better than IRS1 an all but one of the documents (d_5), but on d_5 the SRE by IRS3 is really wrong. Precision, recall, and E-measure for IRS1 and IRS3 do not differ, whereas there is a slight difference in the two ADM values.

4 Discussion

4.1 ADM *vs.* classical effectiveness measures

ADM satisfies the four desirable properties introduced by Swets and reported also in [20, Ch. 7]: it measures the effectiveness only, isolating it from efficiency; 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. Besides being adequate for measuring the

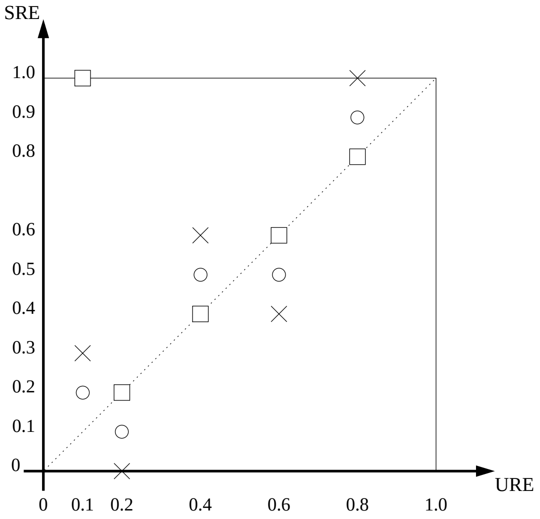
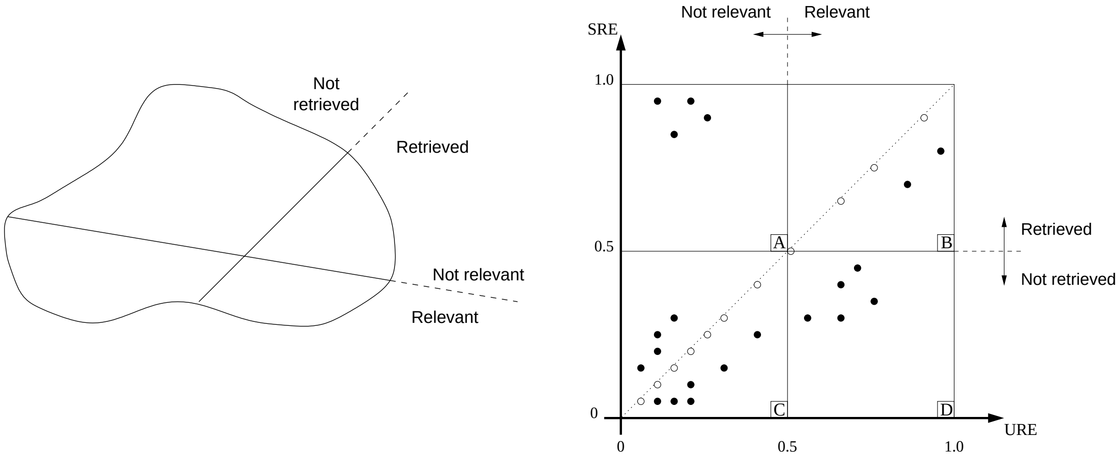


Figure 3: Graphical representation of the example in Tab. 1.



(a) (b)

Figure 4: From binary relevance (a) to continuous relevance (b).

effectiveness of IRSs, ADM leads us to reconsider the effectiveness measures usually adopted in retrieval evaluation. For space limitations, what follows concerns mainly precision and recall, but it can be generalized to other measures as well.

We can now graphically express the generalization from binary relevance and retrieval to continuous relevance and retrieval (Sect. 2): Fig. 4(a) (adapted from [18, Ch. 5]) can be generalized as in Fig. 4(b), where, in place of clear cut divisions between relevant and nonrelevant, and between retrieved and nonretrieved documents, two axes single out continuous values for relevance and retrieval.

Using Fig. 4(b), precision, recall, fallout, and generality factor turn out to be defined respectively as:

$$P = \frac{|B|}{|A|+|B|}, \;\; R = \frac{|B|}{|B|+|D|}, \;\; F = \frac{|A|}{|A|+|C|}, \;\; G = \frac{|B|+|D|}{|A|+|B|+|C|+|D|},$$

where $|A|$, $|B|$, $|C|$, and $|D|$ are the numbers of documents in the A, B, C, and D sectors, respectively.

From Fig. 4(b) one can also understand that 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 URE–SRE plane turn out to be in either $(0, 0)$, $(0, 1)$, $(1, 0)$, or $(1, 1)$, and therefore the distances from the ideal line must be either 0 or 1. The definition of ADM_N^M measures (based on N categories of relevance and M categories of retrieval) is straightforward too.

A comparison of ADM with precision and recall shows how 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. 4(b), some points were added to the C 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 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 (A, B, and D), and the 75% with less documents, since most of them will be in the C sector (in general, given a query, most of the documents are neither relevant not retrieved).

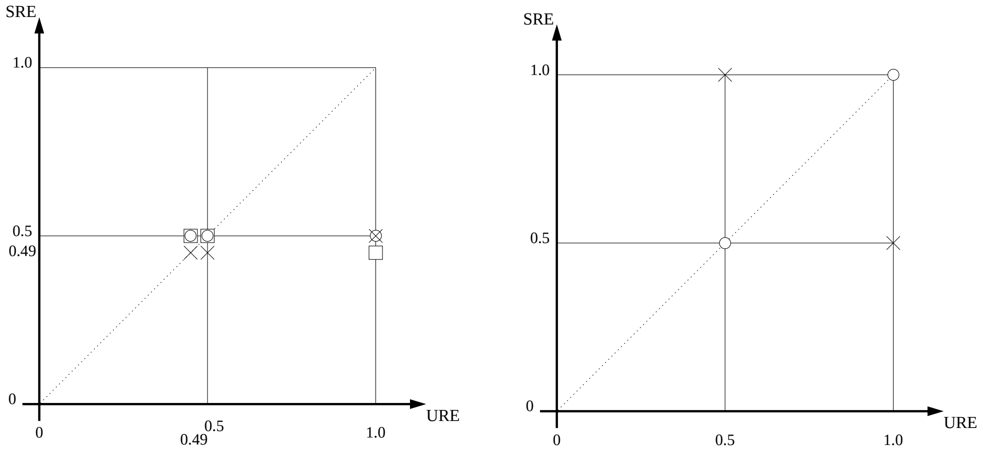
- Precision and recall do not use the full-fledged distance from the ideal line used in Eq. 1, since all the documents within each sector (A, B, C, and D) are considered as equivalent (the distance used is 0 if the document is in sector B or C, 1 if the document is in sector A or D: the same limitation of 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 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.

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 system 1, crosses for system 2). 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 cannot grasp the difference, whereas ADM does.

Therefore, the two problems about precision and recall are: first, small differences in the SRE can lead to very different precision, recall, and E-measure figures, whereas small differences do not affect ADM; second, big differences in SRE 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) [7]. Moreover, looking at it from a different perspective, one might say that using ADM in place of precision and recall, there is no more need for many queries in information retrieval experiments, and the effectiveness for queries with very few relevant documents is measured in a more reliable way.



(a) (b)

Figure 5: Small (a) and big (b) differences in SRE values.

Table 2: Effectiveness measures for Figs. 5(a) and 5(b).

	P	R	E	ADM
IRS1 ($\bigcirc$)	0.67	1	0.84	0.83
IRS2 ($\times$)	1	0.5	0.75	0.83
IRS3 ($\Box$)	0.5	0.5	0.5	0.826

(a)

	P	R	E	ADM
IRS1 ($\bigcirc$)	1	1	1	1
IRS2 ($\times$)	1	1	1	0.5

(b)

Both problems depend on the thresholds on SRE and URE. The second one, however, has a further component: the equal status given to documents within each sector in the calculation of precision and recall. Indeed, it seems not fair to consider all the documents in, say, B simply as “retrieved and relevant”; a fairer categorization might be the one shown in Fig. 6(a), where the documents in the white areas A1 (closer to the ideal line) are considered as correctly evaluated (their distance is < 0.5), whereas the document in the grey areas A2 are not correctly evaluated (distance ≥ 0.5). If the grey and white parts must have the same area, as seems reasonable, with simple calculations we obtain $x = 1 - \frac{\sqrt{2}}{2}$ (see Fig. 6(b)).

On the basis of this new categorization, one might, e.g., define a new version of precision and recall as:

$$P' = \frac{|3| + |4|}{|1| + |2| + |3| + |4| + |5|} \quad R' = \frac{|3| + |8|}{|2| + |3| + |5| + |8| + |9|}$$

(where $|N|$ stands for the number of documents in sector N, see Fig. 6(b)).

4.2 Using ADM in practice

One might wonder how to compute ADM in practice: indeed, there are some issues that need to be dealt with before its practical usefulness is clear.

A first issue is how to get the URE values. I see two ways for doing that. Either we can ask the judges to express the usual dichotomous (or even discrete, using a category rating scale) relevance judgments, and then average them to get the continuous judgments; or we could ask the users to directly express their judgments in a continuous way. Some previous studies explored

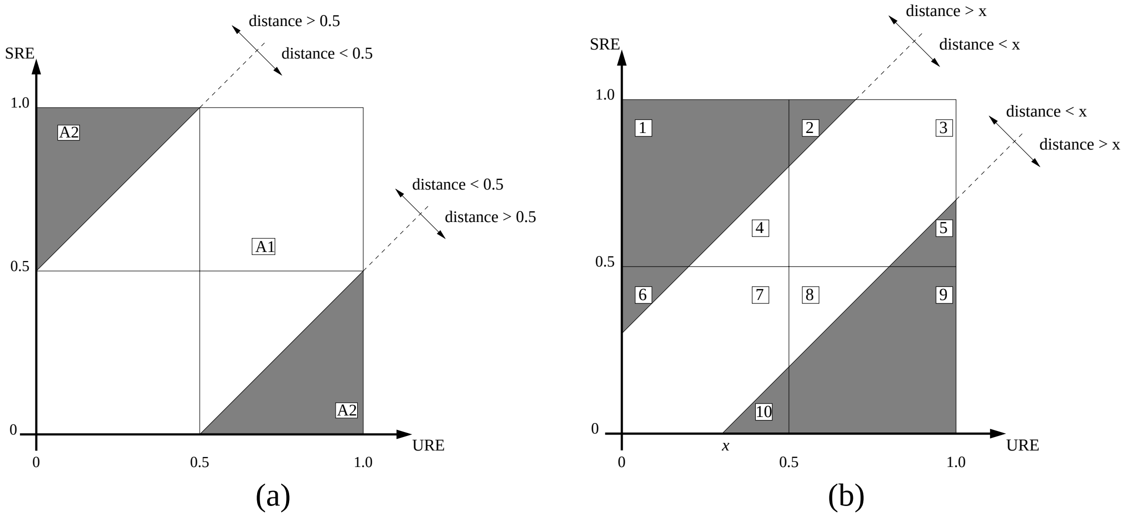


Figure 6: A better categorization than the classical one in Fig. 4(b).

the second approach, showing that the so called magnitude estimation techniques (numeric estimation, hand grip, and line length) are not only reliable to obtain continuous judgments, but also preferable to the classical dichotomous and category rating scales judgments [2, 3, 5, 9, 11, 12]. If more users are available, both methods can be used (and one could even allow some users to express classical dichotomous or discrete judgments, as they prefer). If only one user is available, the only chance is to ask her to express continuous judgments.

A second problem is the following. Accordingly to Eq. (1), to exactly compute the ADM for a given query, URE values for all the documents in the database are needed (as well as SRE values, but these are easily obtained from the IRS being evaluated). Of course, it is not feasible to ask the users to evaluate the relevance of thousands of documents. What can be done instead is to sample in some way the database, asking the users to evaluate the documents in the sample only. The sample can be obtained in various ways: one can use the retrieved documents only, or select documents at random from the database, or select from sets of documents having SRE values in some predefined ranges, and so on.

Let's remark that we have a similar problem even when using precision (in which case the sample consists of the retrieved documents only) and recall (that, as it is well known, can be only estimated in realistic databases of documents). On the basis of the above remarks, perhaps ADM turns out to be more adequate for the evaluation of information filtering (or routing) systems, where the sampling is obtained in a natural way by grouping all the documents received and filtered in the last time period, and it seems feasible to ask the users to express their URE on all these documents.

Moreover, if one has to evaluate many IRSs, as for example in the TREC experiments series (see http://trec.nist.gov), a completely automatic procedure can be devised: the URE values are simply obtained as the average of all the SRE values from all the IRSs, without any human judgment. Therefore, the pooling method used in TREC (that inspired this automatic procedure), even without human assessors, might lead in a natural way to continuous relevance assessment.

5 Conclusions and future developments

In this very preliminary work, I have proposed a new measure of information retrieval effectiveness, ADM, and shown how precision and recall can be better understood on its basis.

After a brief historical analysis that motivates the way we measure IR effectiveness today, I have defined ADM. Then, I have described ADM potential for being an effective measure, and presented some limitations of precision and recall, mainly their hyper-sensitiveness to small variations, and their lack of sensitiveness to big variations. Due to the presence of thresholds, with precision and recall, slight improvements to an IRS might lead to big improvements in effectiveness; with ADM this cannot happen.

Let me emphasize that I have assumed, in this paper, that the "actual" relevance is not a dichotomous yes/no values, but a continuous one. This seems natural to me (also given my previous research on this topic [14, 15]), but some readers might not agree with this position. Let's try to convince them. Let's start by trying to show that some documents can be "more relevant" than others. Given a user that wants to study the issue of relevance in IR, the classical paper by Saracevic [19] is unquestionably more relevant than a paper about, say, retrieval evaluation, that only slightly hints at the issue of relevance (personal experience!). Given a user that wants to understand why the classical "endosystem" view[1] of IR is in some way incomplete, Peter Ingwersen's book about cognitive IR interaction [8] is more relevant than the classical book by Salton and McGill [18].

This, I hope, is enough to convince that yes/no relevance is just a (sometimes convenient, but sometimes misleading) approximation of a more complex phenomenon. But the skeptical readers might still insist that relevance might just be a matter of categories: fine, more than two, but having a continuum is not necessary. To them, I can answer with a question: how many categories? Two is not adequate. Almost all the values from three to eleven categories have been used in the past, but nothing prevents us to use more than that. The asymptotic limit is, of course, to use a continuous range, that, at least, can approximate in an effective way a scale with any number of categories.

Skeptic readers have the the last resort of "preference" judgments, i.e., judgments of preference of one document over another [17]. These can be transformed into continuous relevance judgments by relying on two assumptions: that the most (respectively, least) preferred document corresponds to a 1.0 (respectively, 0.0) relevance score, and that the intermediate documents have a uniform distribution. These two assumptions, I admit, are rather strong; however, nothing better can be done for dichotomous or discrete judgments.

Another point worth to be briefly mentioned is that the disagreement among relevance judges seems to decrease after some discussion among them [16]. As seen under the light of the ADM measure, this phenomenon might not be due to a so high initial disagreement, but simply to the clear cut division between dichotomous judgments of relevance and nonrelevance.

Many points need further work. In Eq. (1), one might use standard deviation in place of the sum of the absolute differences. The sampling needed for calculating ADM in real databases deserves further study, to choose the most reliable approach. Also calculating the ADM value if the documents in the database are assigned a random SRE might be useful. From a practical side, I intend to evaluate some IRSs using ADM. Perhaps the first thing to do is to re-analyze the data of some past experiments (probably from TREC) to verify if the problems with precision and recall do occur in the real world.

Acknowledgements

Many thanks to one of the referees, who gave useful comments that helped to improve the paper.

[1] Korfhage [13] defines the *endosystem* as, approximatively, the search engine *per se*, and the *ectosystem* as the system comprising the user.

References

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

[2] 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.

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

[4] 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.

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

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

[7] 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.

[8] P. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, 1992.

[9] 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.

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

[11] 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.

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

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

[14] 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.

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

[16] J. O'Connor. Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation*, 20(4):311–319, 1969.

[17] M. E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, 1990.

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

[19] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

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

[21] 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.