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

Vincenzo Della Mea, Luca Di Gaspero, Stefano Mizzaro*

Department of Mathematics and Computer Science
University of Udine
http://www.dimi.uniud.it/~mizzaro
mizzaro@dimi.uniud.it

NTCIR-4, Tokyo, 2 June 2004

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

Vincenzo Della Mea, Luca Di Gaspero, Stefano Mizzaro*

Department of Mathematics and Computer Science
University of Udine
http://www.dimi.uniud.it/~mizzaro
mizzaro@dimi.uniud.it

NTCIR-4, Tokyo, 2 June 2004

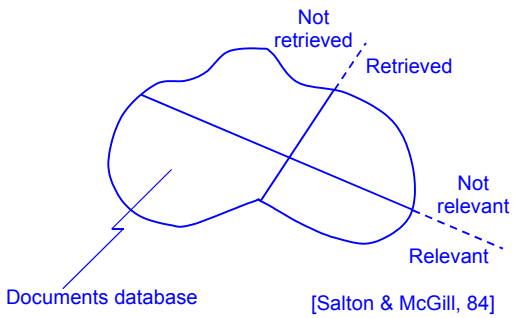
The idea

- ADM: an IR effectiveness measure based on continuous relevance
- Relevance
  - Binary {0,1}
  - Categories {low, medium, high}
  - Continuous [0..1]
- Retrieval: too (boolean, vector space, …)
- V. Della Mea, S. Mizzaro (2004). Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation, *JASIST*, 55(6):530-543
- Draft, p. 30, supplement v. 2

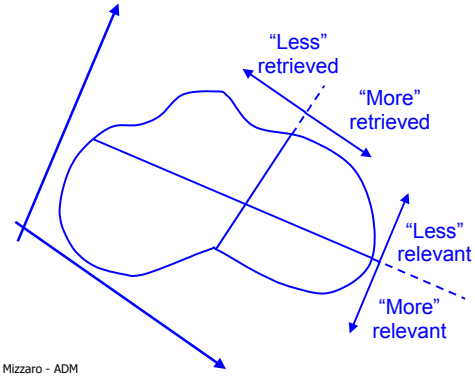
Outline

- Definition
  - The URS/SRS plane
  - ADM (Average Distance Measure)
  - Examples
- Conceptual analysis
  - Problems with precision and recall
- Experimental analysis
  - TREC data
    - ADM is as good as TREC measures
    - ADM is effective with less data than TREC measures
  - NTCIR data: preliminary results

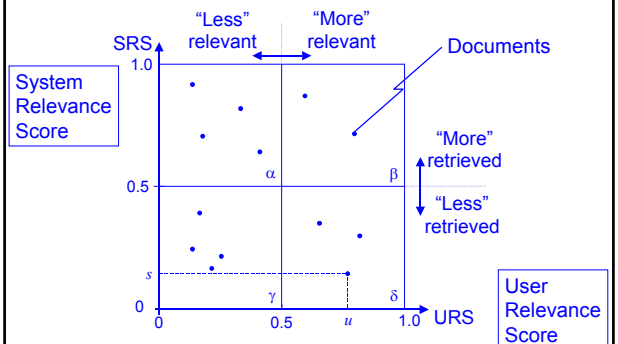
From binary relevance…



… to continuous relevance



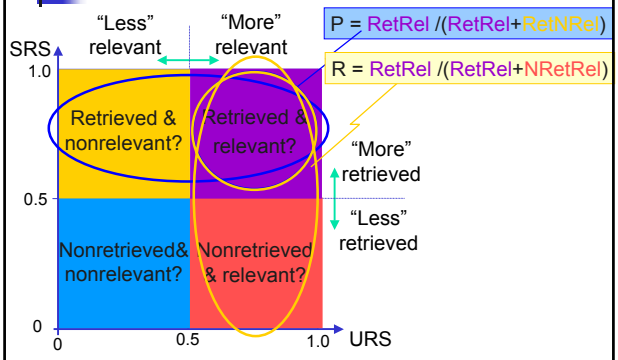
The URS/SRS plane



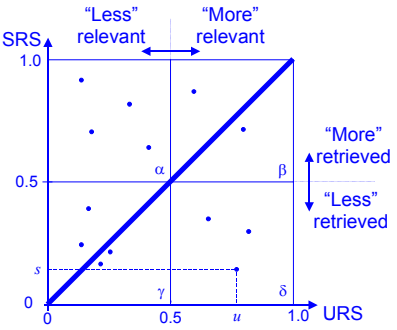
SRS and URS

- SRS (System Relevance Score)
  - Relevance value given by the IRS
- URS (User Relevance Score)
  - Relevance value given by the user
- Real numbers, in the [0..1] range
- Different from
  - RSV (Retrieval Status Value), insensible to rank-preserving transformations
  - Estimate of the probability of relevance

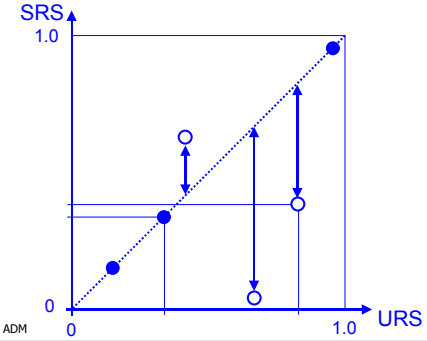
A step backward: P & R



The "right" places…



ADM: Average Distance Measure

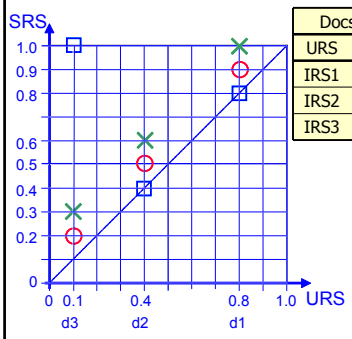


ADM: Average Distance Measure

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

- ADM for one query: 1 - average distance between SRS and URS over all (?) the documents
- ADM for one IRS: average over some queries

An example



Docs.	d1	d2	d3	ADM
URS	0.8	0.4	0.1	
IRS1 ○	0.9	0.5	0.2	0.9
IRS2 ×	1.0	0.6	0.3	0.8
IRS3 □	0.8	0.4	1.0	0.7

Outline

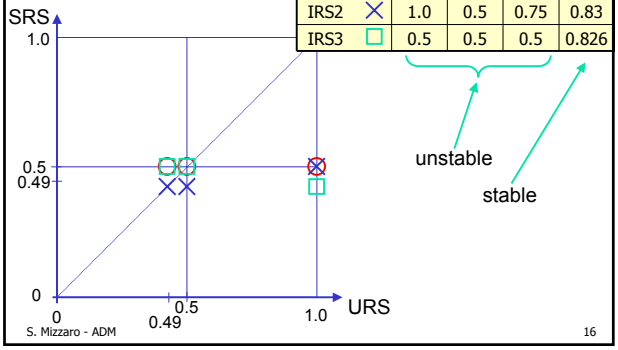
- Definition
  - The URS/SRS plane
  - ADM (Average Distance Measure)
  - Examples
- Conceptual analysis
  - Problems with precision and recall
- Experimental analysis
  - TREC data
    - ADM is as good as TREC measures
    - ADM is effective with less data than TREC measures
  - NTCIR data: preliminary results

ADM vs. P & R

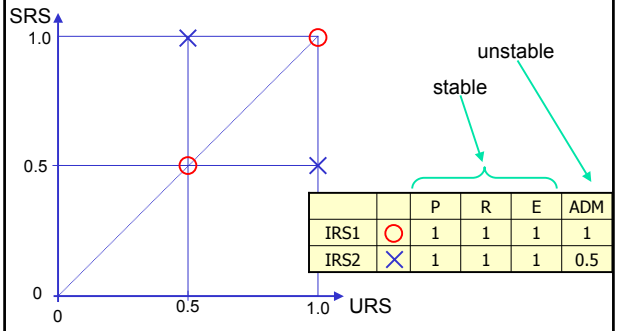
- Precision and recall are:
  - Hyper-sensitive
    - to relevant/nonrelevant and retrieved/nonretrieved thresholds
    - (i.e., 0.49 and 0.51 are two very similar values, but the outcome is very different…)
  - Insensitive
    - to variations within particular areas (0.99 and 0.51 are very different, but the outcome is the same…)

Hyper-sensitiveness: 3 similar IRS

	P	R	E	ADM
IRS1 ○	0.67	1.0	0.84	0.83
IRS2 ×	1.0	0.5	0.75	0.83
IRS3 □	0.5	0.5	0.5	0.826

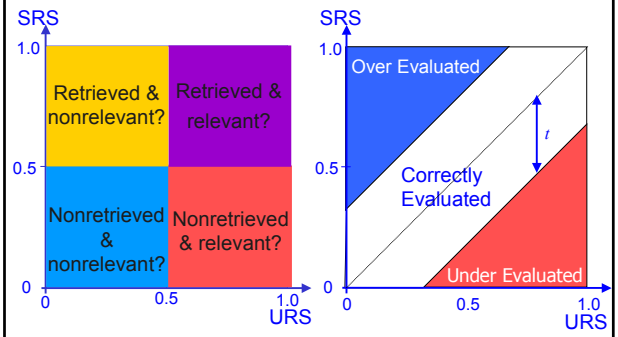


Insensitiveness: 2 different IRS



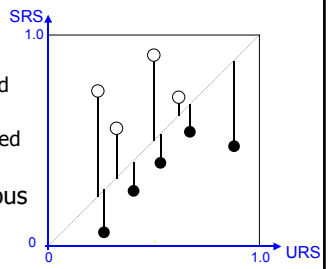
	P	R	E	ADM
IRS1 ○	1	1	1	1
IRS2 ×	1	1	1	0.5

Problem: arbitrary & wrong thresholds



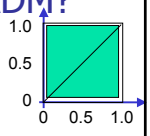
ADM variants

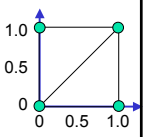
- ADM for precision and recall
  - R: on the over-evaluated documents only
  - P: on the under-evaluated documents only
- ADM with non continuous SRSs and URSs
- …

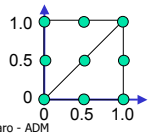


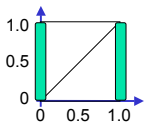
What do we need for ADM?

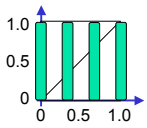
- Ideal situation: Continuous SRS & URS
- Worst situation: "binarized" ADM
  - All the documents in (0,0),(0,1),(1,0),(1,1)
  - Docs in (0,1) e (1,1) only: R
  - Docs in (1,0) e (1,1) only: P
- Intermediate situations: "discrete" ADM
  - Categories, combinations, …











Outline

- Definition
  - The URS/SRS plane
  - ADM (Average Distance Measure)
  - Examples
- Conceptual analysis
  - Problems with precision and recall
- Experimental analysis
  - TREC data
    - ADM is as good as TREC measures
    - ADM is effective with less data than TREC measures
  - NTCIR data: preliminary results

ADM on TREC data

- ADM Variants:
  - (simplifying…)
  - URSs are binary (either relevant or nonrelevant)
  - SRSs are not reliable → We used the ranking

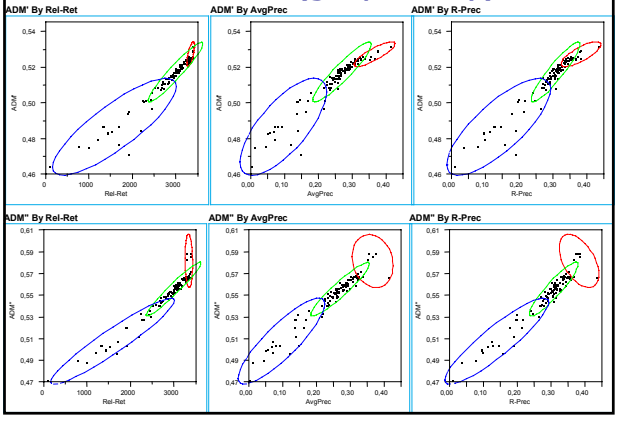
Rank	1st	2nd	3rd	4th	…	999th	1000th	1001st	…
SRS	1.0	0.999	0.998	0.997	…	0.002	0.001	0.000	…

ADM is as good as TREC measures

	ADM	Rel-Ret	AvgPrec	R-Prec
ADM	1			
Rel-Ret	**0.891**	1		
AvgPrec	**0.876**	0.824	1	
R-Prec	**0.844**	0.807	0.902	1

- Kendall Correlations

Correlations (graphically)



ADM is effective with less data than TREC measures

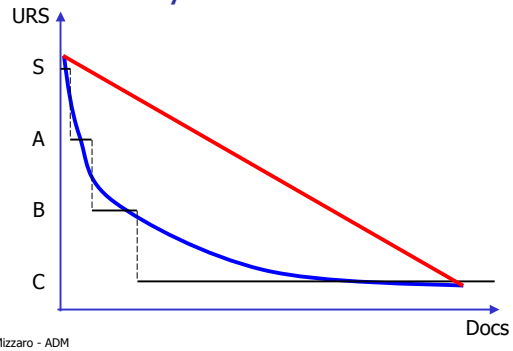
- Correlations between "global" ADM (on the TREC pool docs.) and ADM on subsets:

Set (Ret, Rel, topics)	N. docs (approx.)	ADM
(100%, 100%, 100%)	53000	1
(100%, 100%, 50%)	26000	0.852
(50%, 50%, 100%)	26000	0.910
(10% 10% 100%)	5000	0.802
(50% 50% 50%)	13000	0.807
(100% 0% 100%)	50000	0.935

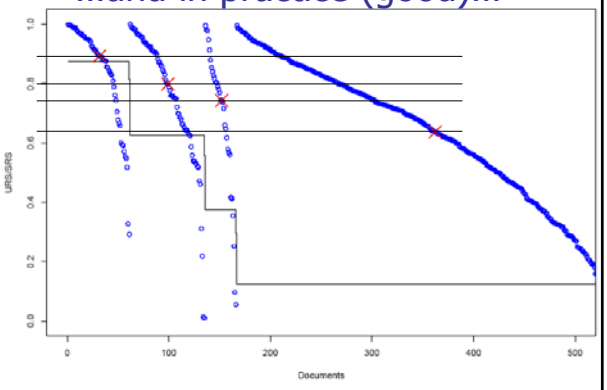
ADM on NTCIR-4 data

- PRELIMINARY RESULTS!
- URS:
  - 4 categories → 4 values (…)
- SRS:
  - Continuous scores → Linear normalization into SRSs
  - Rank, as in TREC

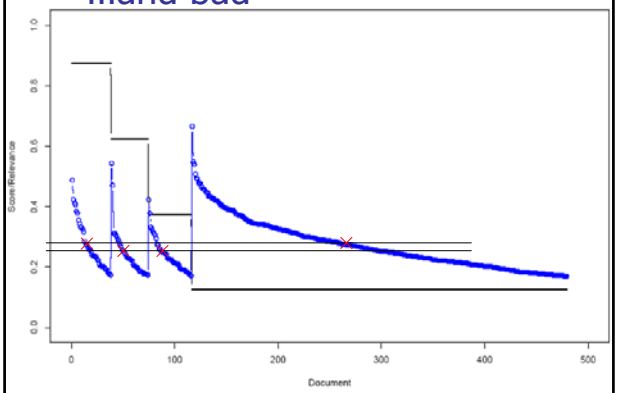
URS and SRS distributions: in theory…



…and in practice (good)…



…and bad



Some results: low correlations…

- No correlation between ADM and standard measures
  - Standard measures are not sensible to how well an IRS approximates the URS distribution
- Good IRS according to standard measures = Good rank
- Good IRS according to ADM = Good approximation of the URS distribution shape

…and some high correlations

- Rank-based ADM
- On the first N retrieved documents

N	*5*	*10*	*20*	*50*
AvgPrec	0.747	0.792	0.8	0.788
R-Prec	0.755	0.802	0.816	0.799

Summary

- Definition
  - The URS/SRS plane
  - ADM (Average Distance Measure)
  - Examples
- Conceptual analysis
  - Problems with precision and recall
- Experimental analysis
  - TREC data
    - ADM is as good as TREC measures
    - ADM is effective with less data than TREC measures
  - NTCIR data: preliminary results

Future work

- Carefully analyze NTCIR-4 data
- A proposal
  - IRSs participating in next NTCIR-5 could be evaluated by ADM too
    - SRSs normalized in [0..1]
    - Carefully decide how to compute the SRSs
    - Try to better approximate the URS distribution
- Continuous URS?
- Distributed IR, data fusion, meta-search, …