ADM

Average Distance Measure

Gianluca Demartini
Università degli Studi di Udine
15/06/2005

Scaletta

- Misure classiche: richiami
- ADM: richiami, esempi, estensioni
- Esperimenti fatti
- Esperimenti da fare
- Strumenti

Misure classiche (1/2)

- Precision, recall

$$Precision = \frac{|relevant \cap retrieved|}{|retrieved|}$$

$$Recall = \frac{|relevant \cap retrieved|}{|relevant|}$$

- Prec@5, Prec@10, …
  - rel[i]=1 se i-esimo doc reperito è relevant
  - rel[i]=0 altrimenti

$$Prec@j = \sum_{k=1..j} rel[k]/j$$

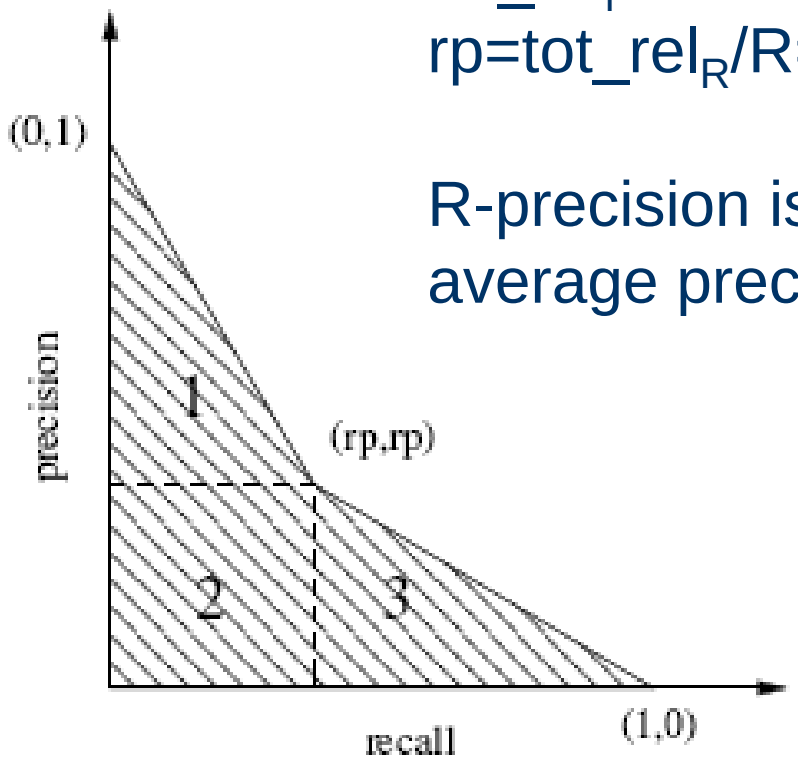
Misure classiche (2/2)

- Avg-Prec:
  - n= num docs reperiti
  - R= num docs relevant

$$AvgPrec = \sum_{j=1..n} (Prec@j * rel[j])/R$$

- R-Prec: prec@R

R-Prec ≈ Avg-Prec



tot_rel_i=# relevant docs reperiti al rank i
$rp=tot\_rel_R/R$=recall al rank R

R-precision is approximately average precision

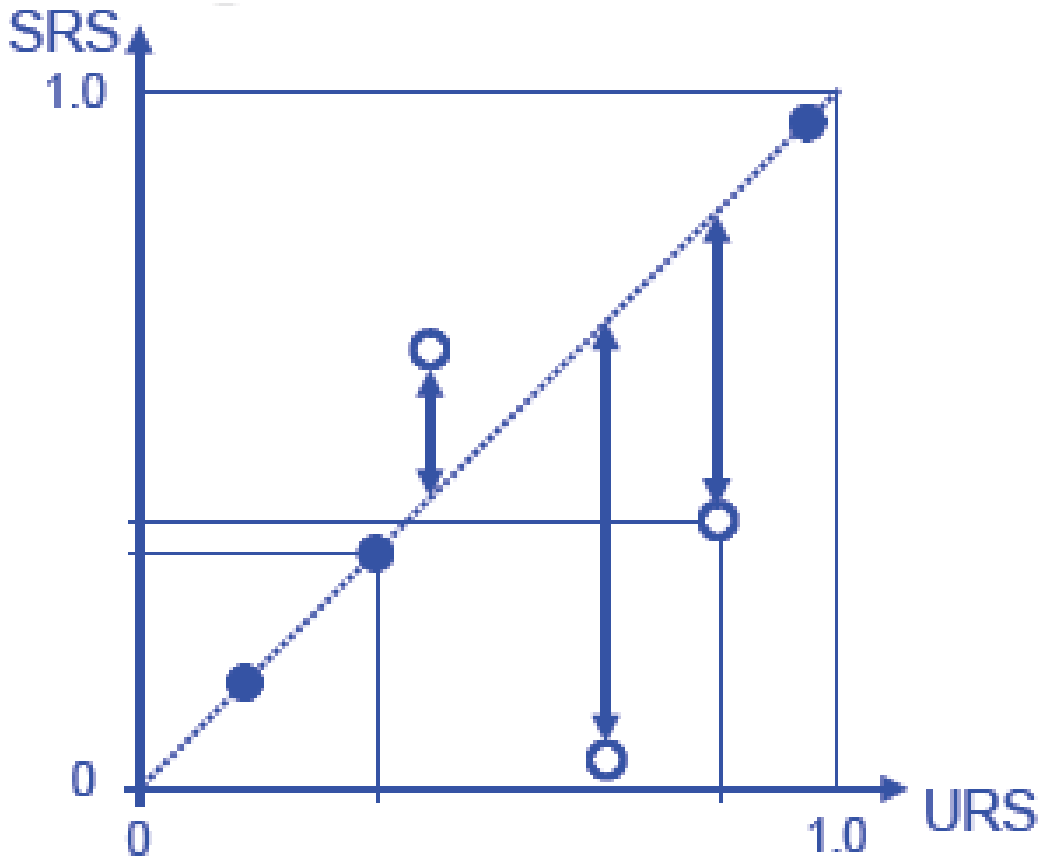
ADM

- Average Distance Measure

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

- Ok per URS e SRS continui (o discreti non binari)

Grafico ADM

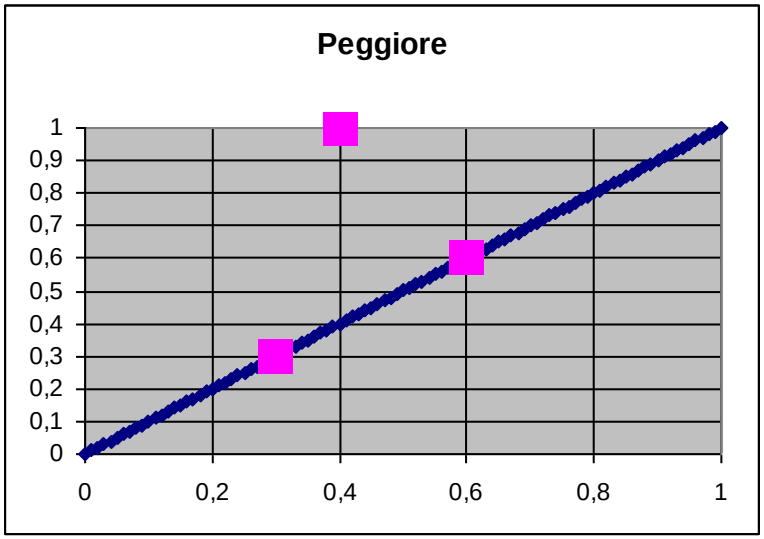


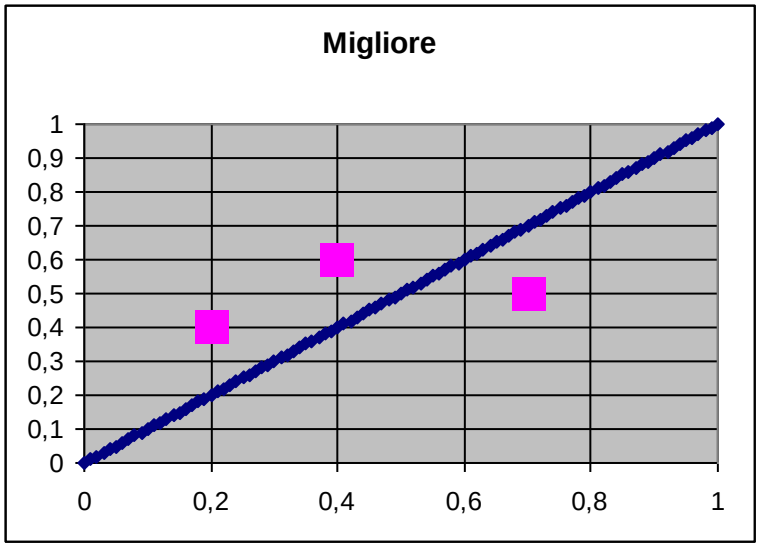
Somma del quadrato delle distanze

- Favorire i sistemi che sono più precisi nel valutare i documenti
- Sfavorire quelli che valutano molto bene certi e peggio altri
- (avendo ADM uguale)

Esempio (1/2)

- Somma del quadrato delle distanze:





Esempio (2/2)

Peggiore			Migliore	
URS	**SRS**		**URS**	**SRS**
0,30	0,30		0,2	0,4
0,40	1,00		0,4	0,6
0,60	0,60		0,7	0,5

ADM(Peggiore)=1-(0+0,6+0)/3=0,8
ADMquad(Peggiore)=1-($0,6^2$)/3=**0,88**

ADM(Migliore)=1-(0,2+0,2+0,2)/3=0,8
ADMquad(Migliore)=1-($0,2^{2+}$ $0,2^2$+ $0,2^2$)/3=**0,96**

Misure legate ad ADM

- ADP: ADM sui “sopravvalutati”
- ADR: ADM sui “sottovalutati”
- ADM = ADP + ADR - 1

- ADM@5, ADM@10: ADM sui primi x docs
- ADM@1

La Relevance

- ADM va bene anche se la relevance non è binaria.
- E’ naturale pensare che un doc può essere più o meno relevant invece che del tutto relevant o per niente relevant

Relevance in INEX

- Two dimensions were employed to define relevance:
  - **Exhaustivity (e-value)** measures the extent to which the given element covers or discusses the topic of request.
  - **Specificity (s-value)** measures the extent to which the given element is focused on the topic of request.
- 4 livelli per ogni dimensione!

Scaletta

- Misure classiche: richiami
- ADM: richiami, esempi, estensioni
- **Esperimenti fatti**
- Esperimenti da fare
- Strumenti

Conferenze di valutazione

- Per gli esperimenti su ADM sono stati utilizzati dati di TREC, INEX, NTCIR
- TREC: conferenza principale
- NTCIR: reperimento su docs in più lingue
- INEX: reperimento sul documenti XML

Esperimenti

- Obiettivi: Scoprire se ADM è più “sensibile” rispetto le altre metriche di valutazione
- Rank dei sistemi: ADM si basa sullo **score** assegnato dai SRI
- Correlazione:
  - *(stat.) Tendenza di due grandezze a variare in modo concomitante*
- Versioni di ADM testate:
  - ADM, ADM@x, ADP, ADR, ADM(rank), ADM(score)

Esperimenti su TREC-8

- Usato il rank indotto dallo score (e non i valori di score)
- Risultati: ADM correla bene con le golden standard

Correlazioni	ADM
AvgPrec	0,876
R-Prec	0,844

Esperimenti su NTCIR

- Score a 4 livelli (S,A,B,C)
- S=7/8, A=5/8, B=3/8, C=1/8
  - (per poter sopravvalutare/sottovalutare)
- Risultati:
  - $ADM_{(4)}^{rank}@N$, $ADM_{(2)}^{rank}@N$
  - $ADM_{(4)}^{score}@N$

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

Esperimenti da fare

- Eero Sormunen ha ri-classificato (su 4 livelli di relevance) alcuni docs di TREC-7/8
- Vedere se con 4 livelli ADM funziona meglio che con 2
- Correlazioni con AvgPrec, R-Prec

Esperimenti da fare

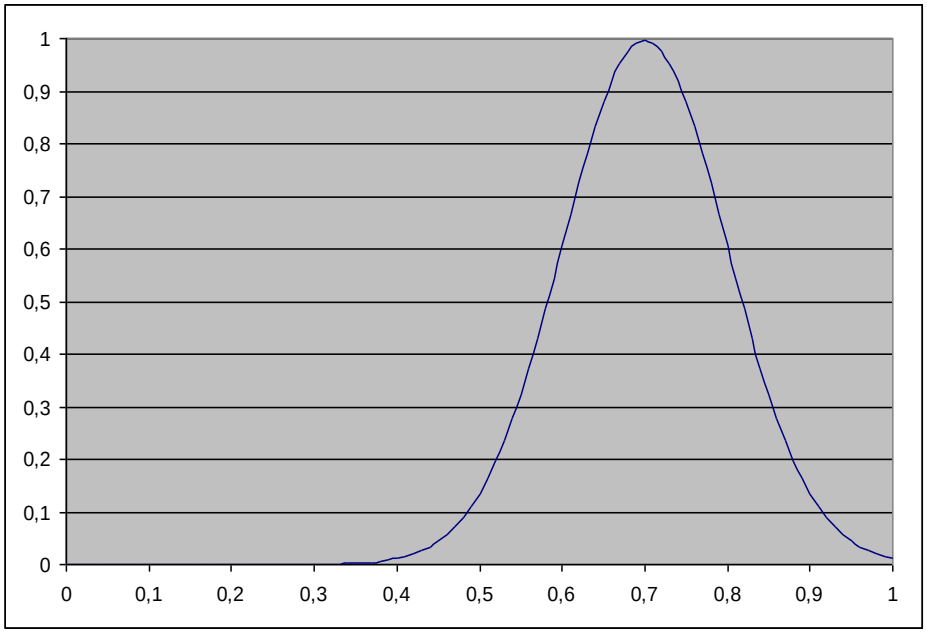
- In TREC-2003:
  - HARD Track: 3 livelli di relevance (hard relevant, soft relevant, no relevant)
  - Robust Retrieval track: pochi docs reperiti
  - TeraByte Collection: su grandi collezioni (tipo il Web) ADM va meglio?

Strumenti usati

- R, correlazione di Kendall, DB, …
- Quello che ho studiato in 5 anni serve!
- Obiettivo: costruire un "banco di prova”
- Quando arrivano i risultati di una conferenza si vede come va ADM

Probabilità di relevance

- Con una relevance non binaria dovrei avere una distribuzione della probabilità sui vari livelli
- Fornire un intervallo di confidenza in cui cade il valore esatto di relevance
- Esempio:
  - Media 0,7
  - Varianza 0,1



Conclusioni

- ADM sembra essere più sensibile (stessi valori con meno docs reperiti)
- Avere una valutazione sistematica che lo provi

Bibliografia (1/3)

- 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. Infotech Oulu, Oulu, Finland, 19-21 September 2001
- 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

Bibliografia (2/3)

- V. Della Mea, L. Di Gaspero, and S. Mizzaro. Evaluating ADM on a four-level relevance scale document set from NTCIR. In Proceedings of NTCIR Workshop 4 Meeting - Supplement Vol. 2, Tokyo, 2-4 giugno 2004. National Institute of Informatics (NII), pages 30-38
- Norbert Fuhr, Saadia Malik, Mounia Lalmas. Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003

Bibliografia (3/3)

- Text REtrieval Conference (TREC) http://trec.nist.gov
- A Geometric Interpretation of Rprecision and Its Correlation with Average Precision (SIGIR05, 2005, in corso di pubblicazione).