IR System Evaluation

Why do we perform evaluation?

Example for problems of evaluation

$\Delta = (\substack{+\\-} \mid \substack{-\\-}\ \substack{-\\-}\ \substack{-\\-}\ \substack{-\\-}\ )$

$\Delta = (\ +_{-}+_{-}+ \mid \substack{-\\+}\ \substack{-\\+}\ \ {}^{-})$

Above notation succinctly represents how a system ranks documents and what user likes / dislikes

---

Relevance

\+ user would like to retrieve

$^{-}$ user would not like to retrieve

vertical lines separate different ranks generated by the retrieval system (by means of RSVs)

Evaluation of systems by evaluation of tasks

2* 2 Table

	rel.	nonrel.
retr.	a	b
not retr.	c	d

Recall = a / (a+c) = R

Precision = a / (a+b) = P

Fallout = b / (b+d) = F

Generality = (a+c) / (a+ b + c +d) = G

$$GR/[GR+(1-G)F] = [(a+c)/(a+b+c+d)]*[a/(a+c)] / \{[(a+c)/(a+b+c+d)]*[a/(a+c)] + [(b+d)/(a+b+c+d)]*[b/(b+d)]\}$$

$$= a / (a+b)$$

$$= P$$

Ranked output (according to retrieval status values)

1	2	3			P
…	…	…		…	…
…	…	…		…	…
…	…	…	…	…	…

$\longrightarrow$

user Rank 1,…, P

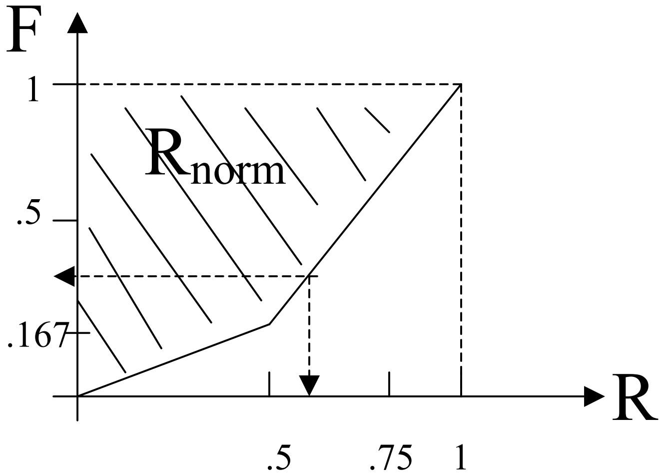
Assumption: User retrieves full ranks

R_v recall after having retrieved v ranks
P_v precision after having retrieved v ranks
F_v fallout after having retrieved v ranks

Commonly used measure in this case: recall_precision (R-P) graph

First consider R-F graph

Example:
(+_+ | _+_ | _ _+ _)



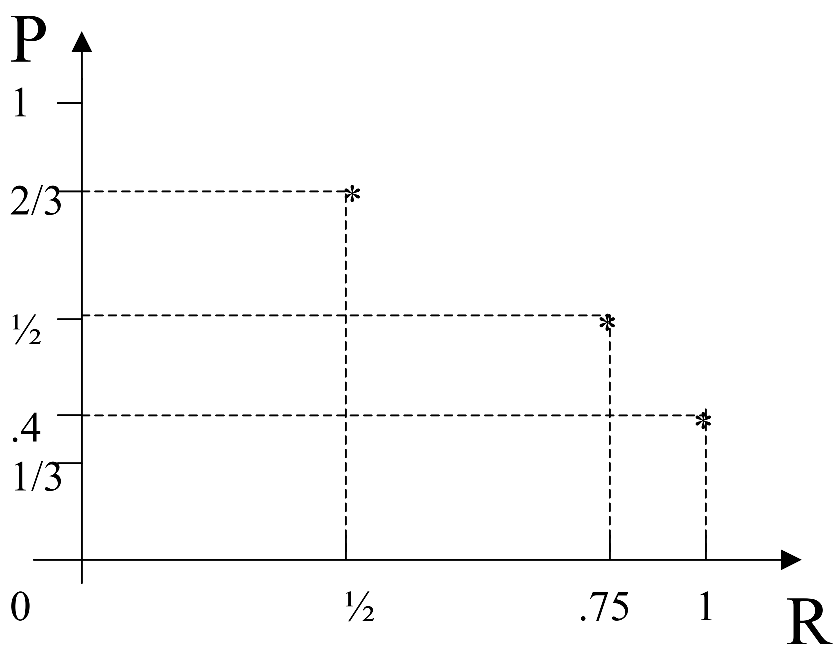
R_{norm} Interpretation:
How much better is the system compared to worst case (Fallout is 1 for all search depths)

R_{norm} = Rocchio 1965

Interpretation of points on R-F graph:

The pair (expected recall after retrieving n documents, expected fallout after retrieving n documents) lies on the Recall_Fallout Graph (n is not necessarily an integer).

Next, we consider R-P graph



How to interpolate?
Straight line interpolation is not meaningful

Interpolation:
Map every point of the Recall_Fallout graph on the Recall_Precision graph using the following formula:

$$(R, F) \rightarrow (R, GR/[GR+(1-G)F])$$

Properties of R-P-graph:
- Every pair (expected Recall after retrieving n documents, expected precision after retrieving n documents) is on the graph.
- Multiplication (creating certain number of copies of every document) of collection gives the same graph
- If considered as graph of a representative sample of very large document collection,

the graph can be interpreted as:

i) expected precision for a given Recall
ii) P(rel/retr.) for a given Recall

<u>Multilevel preference (relevance)</u>

$R_{norm} = ½ [1 + (I^+ - I^-)/ I^+_{max}]$

I^+ number of pairs where a better document precedes a worse document.
I^- number of pairs that are inverted (worse document precedes a better one)
I^+_{max} maximum number of correct pairs

Example:

r relevant
m medium relevant
n non-relevant

3 rel.
4 medrel.
6 nonrel.

r r	r	m
m	m m	n n n
n	n n	

$I^{+} = 2*8+5+4+6 = 31$

$I^{-} = 2+2+3=7$

$I^{+}_{max} =30+24=54$

$R_{norm} =1/2[1+(31-7)/54]$

$=1/2[(54+31-7)/54]$
$=78/108 = .72$

Special case of rel and nonrel (i.e. two-level relevance)

N collection size

n number of relevant

$I^{+}_{max} = n(N-n)$

Example: (+_+ | _+_ | _ ± _)

$I^{+} = 2*5+3 = 13$

$I^{-} = 1+3 = 4$

$I^{+}_{max} = n(N-n) = 4*6 = 24$

$R_{norm} = 1/2[1+(13-4)/24] = ½ [(24+13-4)/24] = 33/48 = 0.69$

Theorem: For binary (two-level) relevance the two definitions of R_{norm} coincide.

Practical Problems:
n number of relevant documents may not be known
→ no R – P graph
no R_{norm}

other evaluation options:
- Just use precision, since it is known
- If more relevant documents are retrieved by a system for the same collection → Recall is higher for that system
- Other measures like expected search length, denoted, esl_k, can be used. It indicates the number of nonrelevant documents that can be expected to be retrieved in order to retrieve k relevant documents.

- “disgust rule” kraft et al. (apprx.1982) how many relevant document can we expect after having retrieved a certain number, k, of nonrelevant documents.