

First Relevant Score

Stephen Tomlinson
Hummingbird

Abstract

Below is the source material for the Hummingbird poster at TREC 2005 (Nov 16-18, 2005).

First Relevant Score:

A More Accurate Measure of

Primary Recall,

Robustness Across Topics,

Early Precision

and User Reading Saved.

S. Tomlinson
Hummingbird

1 What is First Relevant Score?

FRS is

- 1.08^{1-r} (where r is the rank of the first relevant retrieved), or
- zero (if no desired page was retrieved).

1.1 First Relevant Score vs. Reciprocal Rank

- at rank 1: FRS=1.00, RR=1.00 (same)
- at rank 2: FRS=0.93, RR=0.50
- at rank 4: FRS=0.79, RR=0.25
- at rank 10: FRS=0.50, RR=0.10
- at rank 20: FRS=0.23, RR=0.05
- at rank 50: FRS=0.02, RR=0.02

2 Which is the Larger Difference?

Rank 1 vs. 4 or Rank 4 vs. 100?

- RR says 1 vs. 4 is much bigger
(0.75 vs. 0.24)
- FRS says 4 vs. 100 is much bigger
(0.21 vs. 0.79)

FRS better reflects the potential reading saved.

2.1 What's Special About 1.08?

- it makes FRS a generalization of Success@10
- $1.08^{1-10} = 0.5002$
- if you round FRS to the nearest 1 or 0
it becomes the same as Success@10
- hence 'mean FRS' and 'mean Success@10'
(over a set of topics)
will usually be close to the same

3 Primary vs. Secondary Recall

- “Primary Recall” is recall of the first relevant for a topic.
- “Secondary Recall” is recall of the additional relevants.

Primary Recall Measures:

- Success@n, Reciprocal Rank, FRS

Secondary Recall Measures:

- Precision@n ($n \geq 2$), MAP, gMAP, bpref, Interpolated Precision

Duplicate Filtering

If duplicate information was successfully filtered:

- primary recall measures would likely reflect the benefit
- secondary recall measures would likely penalize the technique.

4 A Litmus Test for Robustness Measures

Blind feedback on top 2 rows.

Rich get richer, poor get poorer:

- likely to find more relevants when relevant in first 2 rows
- likely to push down first relevant if no relevant in first 2 rows.

Opposite of robustness goals.

Experiment: humR05dl vs. humR05dle

- MAP up: 0.16 to 0.20
- gMAP up: 0.09 to 0.11
- FRS down: 0.80 to 0.75

All 3 differences passed t-test at 5% level (statistically significant).

- MAP and gMAP failed litmus test.
- FRS passed litmus test.

5 Which Measures Passed the Robustness Litmus Test?

3 blind feedback experiments:

- T: (Robust Titles: humR05txl vs. humR05txle)
- D: (Robust Descriptions: humR05dl vs. humR05dle)
- TB: (Terabyte Titles: humT05xl vs. humT05xle)

Measures Which Passed Litmus Test (in at least one experiment):

- FRS: $t=2.11$ (D), $t=1.98$ (T), $t=1.81$ (TB)
- RR: $t=2.15$ (T), $t=1.87$ (TB)
- S10: $t=2.06$ (D), $t=1.43$ (TB)
- Rec0: $t=1.69$ (T), $t=1.50$ (TB)

Measures Which Failed Litmus Test (in at least one experiment):

- MAP: $t=4.71$ (D), $t=3.75$ (T), $t=2.35$ (TB)
- R-Prec: $t=4.05$ (D), $t=3.01$ (T), $t=2.08$ (TB)
- gMAP: $t=3.24$ (D), $t=2.46$ (T), $t=1.33$ (TB)
- P20: $t=3.09$ (D), $t=1.56$ (T), $t=1.52$ (TB)
- Rec10: $t=2.74$ (TB), $t=2.65$ (D), $t=1.83$ (T)
- P10: $t=1.84$ (T), $t=1.40$ (D)

$t > 2.0$ implies 95% confidence.

$t > 1.3$ implies 80% confidence.

FRS = First Relevant Score

RR = Reciprocal Rank

Rec0 = Interpolated Precision at 0% Recall

Rec10 = Interpolated Precision at 10% Recall

MAP = Mean Average Precision

gMAP = Geometric MAP

R-Prec = R-Precision

S10 = Success@10

P10 = Precision@10

P20 = Precision@20

6 Rating the Tasks with FRS

Easiest (0.80-1.00):

- Terabyte Title Relevance:
FRS=0.93, S10=47/50 (94%), S1=70%
- QA Answer Document:
FRS=0.88, S10=46/50 (92%), S1=68%
- Robust Title Relevance:
FRS=0.83, S10=44/50 (88%), S1=54%

Medium (0.60-0.80):

- Enterprise Known Email:
FRS=0.72, S10=95/125 (76%), S1=38%
- Terabyte Highly Relevant:
FRS=0.69, S10=33/47 (70%), S1=40%
- Robust Highly Relevant:
FRS=0.66, S10=32/45 (71%), S1=29%

Hardest (0.40-0.60):

- Terabyte Named Page Finding:
FRS=0.41, S10=103/252 (41%), S1=21%

(mean scores of (almost) identical full-text search runs)