



INLS 490-154W: Information Retrieval Systems Design and Implementation

Fall 2009 Web-based course

Evaluation-2

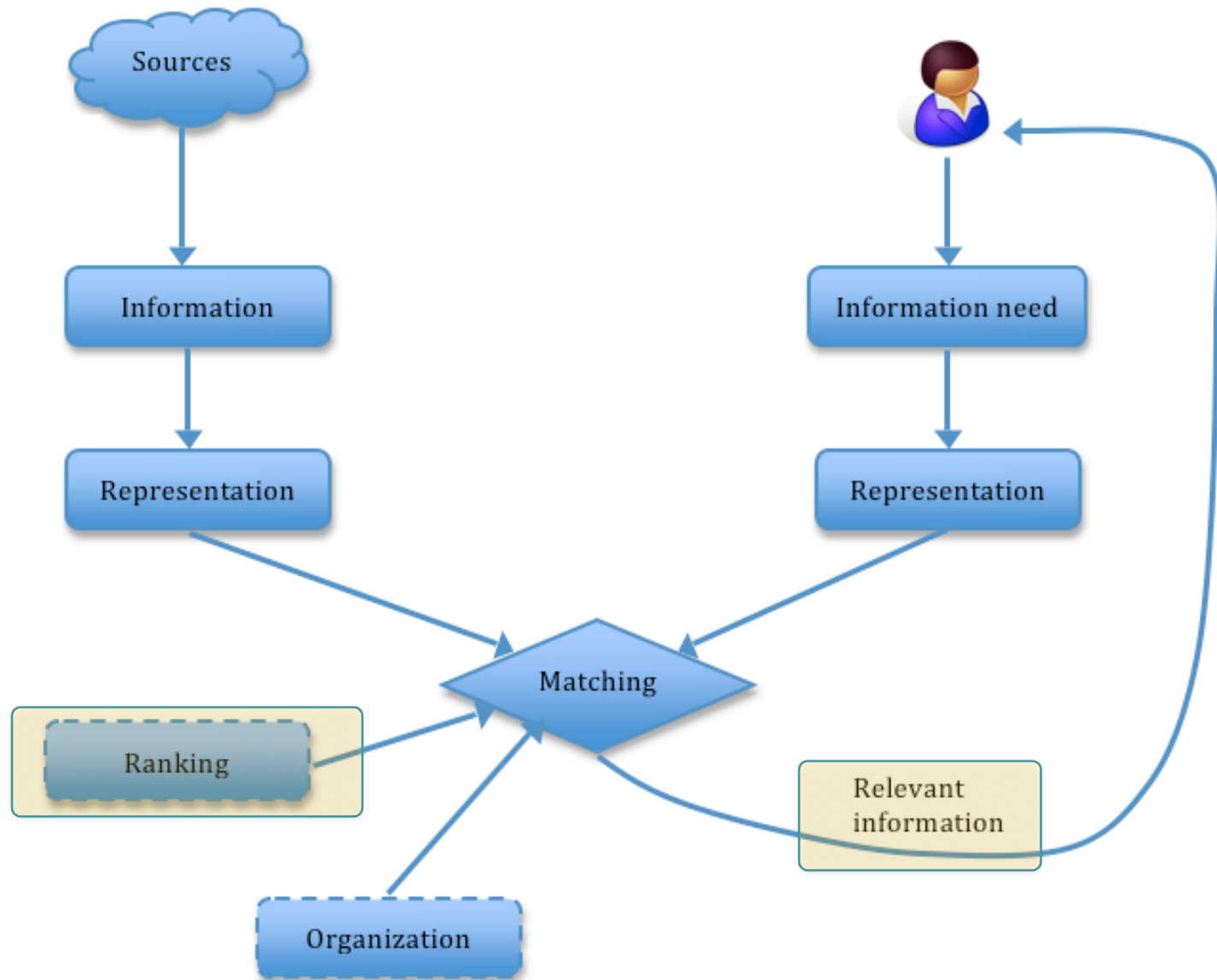
Art of identifying good, bad and ugly

Chirag Shah

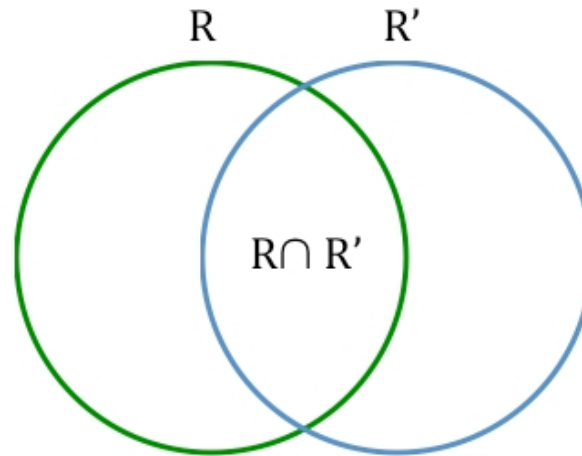
School of Information & Library Science (SILS)

UNC Chapel Hill

Today's lab



Review: recall and precision



R= Relevant, R'=Retrieved

$$\text{Recall} = \frac{R \cap R'}{R}$$

$$\text{Precision} = \frac{R \cap R'}{R'}$$



Review: other measures

- Average precision
- MAP
- R-precision

GMAP

- Geometric mean of per-query average precision, in contrast with MAP, which is the arithmetic mean.
- Designed to highlight improvements for low-performing queries.

$$AP = \frac{1}{m} \sum_{j=1}^m Precision(Recall_j)$$

$$MAP = \frac{1}{|Q|} \sum_{i=1}^Q AP_i$$

$$GMAP = \sqrt[|Q|]{\prod_{i=1}^Q AP_i} = \exp \frac{1}{|Q|} \sum_{i=1}^Q \log AP_i$$

MAP vs. GMAP

- System #1

API = 0.5, AP2 = 0.01

MAP=0.255, GMAP=0.071

- System #2

API = 0.49, AP2 = 0.02

MAP=0.255, GMAP=0.099

Reciprocal rank and MRR

$$RR = \frac{1}{rank}$$

$$MRR = \frac{1}{|Q|} \sum_{i=1}^Q \frac{1}{rank_i}$$

bpref

- Computes a preference of whether judged relevant documents are retrieved ahead of judged non-relevant documents.
- Designed for situations where we do not have enough relevance judgments.

$$bpref = \frac{1}{R} \sum_r \left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)} \right)$$

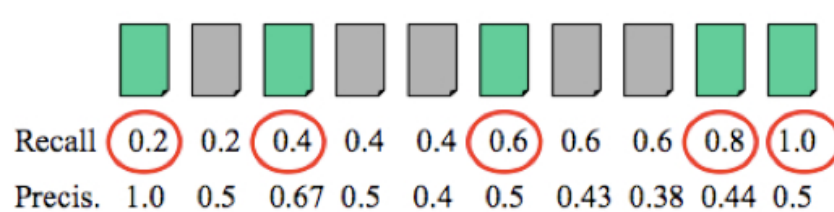
R: total number of judged relevant documents

N: total number of judged non-relevant documents

r: relevant document retrieved

n: member of the first *R* judged non-relevant documents retrieved

Evaluations examples

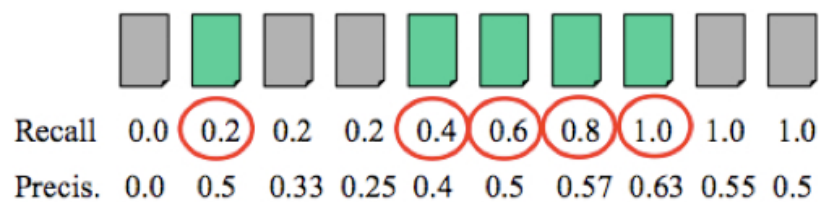


AP = 0.622

R-precision = 0.4

RR = 1

bpref = $\frac{1}{5}((1-0/5) + (1-1/5) + (1-3/5) + (1-5/5) + (1-5/5))$
= 0.44



AP = 0.520

R-precision = 0.4

RR = 0.5

bpref = $\frac{1}{5}((1-1/5) + (1-3/5) + (1-3/5) + (1-3/5) + (1-3/5))$
= 0.48

MAP = $(0.622+0.520)/2 = 0.571$

GMAP = $\sqrt{(0.622*0.520)} = 0.569$

MRR = $(1+0.5)/2 = 0.75$

Source: James Allan, UMass Amherst



Design implications

- General Web search – higher precision
- Legal/medical informatics – higher recall
- Literature review – higher recall
- Homepage finding – higher RR/MRR



Comparing rank lists

- Pearson's covariance
- Spearman's Rho
- Kendall's Tau

Kendall's Tau

- System#1:

- Scores:

a 0.5

b 0.4

c 0.3

d 0.2

- Ranking: a b c d : 1 2 3 4

- Pairs: {(a, b), (a, c), (a, d), (b, c), (b, d), (c, d)}

- System#2:

- Scores:

a 0.4

b 0.1

c 0.25

d 0.05

- Ranking: a c b d : 1 3 2 4

- Pairs: {(a, c), (a, b), (a, d), (c, b), (c, d), (b, d)}

$$\begin{aligned}\tau &= \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \\ &= \frac{5 - 1}{6} = 0.667\end{aligned}$$

Comparing rank lists using R

```
> x<-c(0.5,0.4,0.3,0.2)
> y<-c(0.4,0.1,0.25,0.05)
> order(-x)
[1] 1 2 3 4
> order(-y)
[1] 1 3 2 4
> cor(order(-x), order(-y), method="pearson")
[1] 0.8
> cor(order(-x), order(-y), method="spearman")
[1] 0.8
> cor(order(-x), order(-y), method="kendall")
[1] 0.6666667
```

Summary

- Evaluation metrics that we saw:
 - Recall
 - Precision
 - Average Precision (AP)
 - MAP
 - GMAP
 - R-precision
 - Reciprocal Rank (RR)
 - MRR
 - bpref
- There is no “perfect” evaluation measure.
- Choosing “right” measure(s) to evaluate an IR system depends on the task and requirements.



Next time

- Linking the search back-end to a front-end
- Creating interactive and dynamic user interfaces