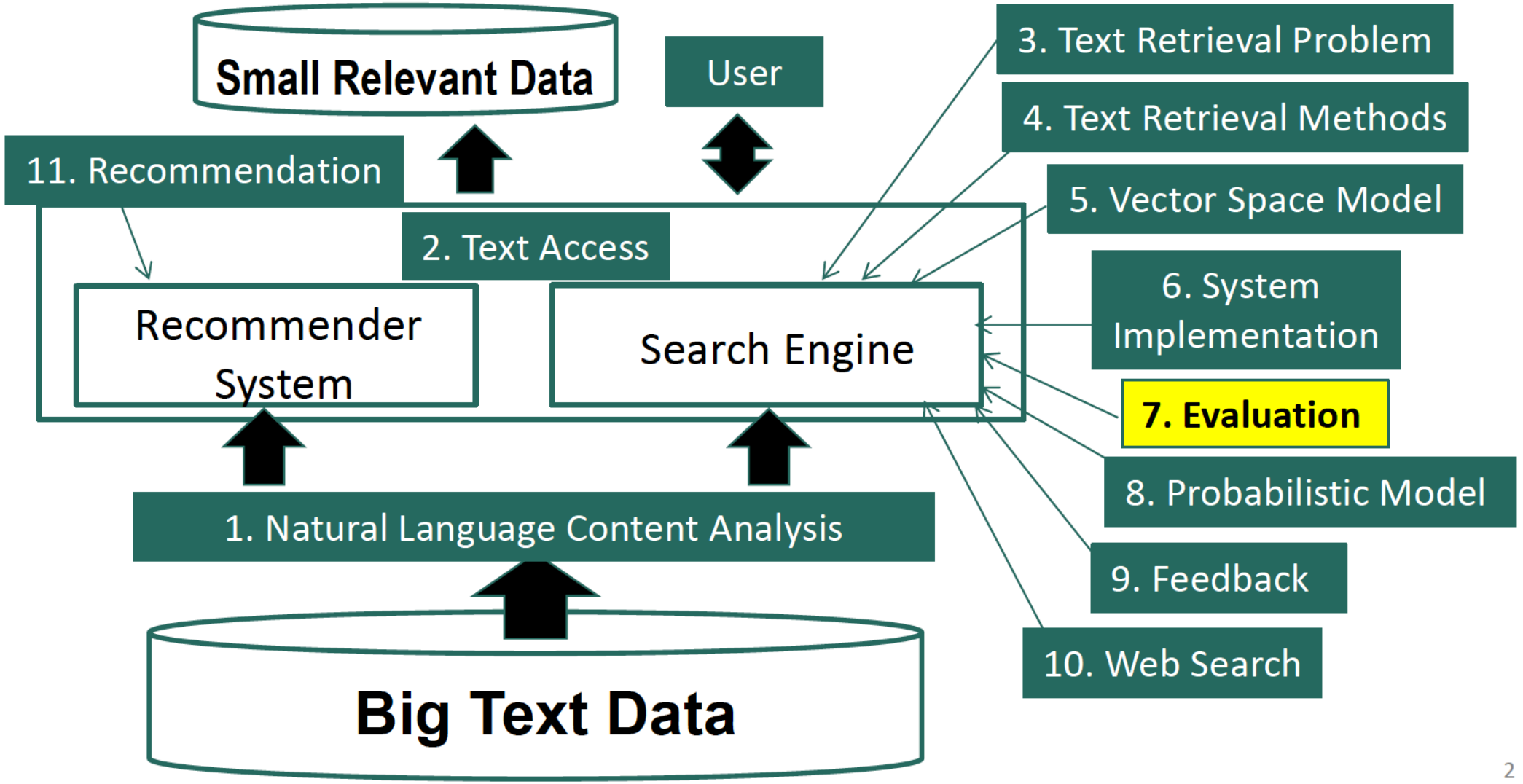
Text Retrieval & Search Engines

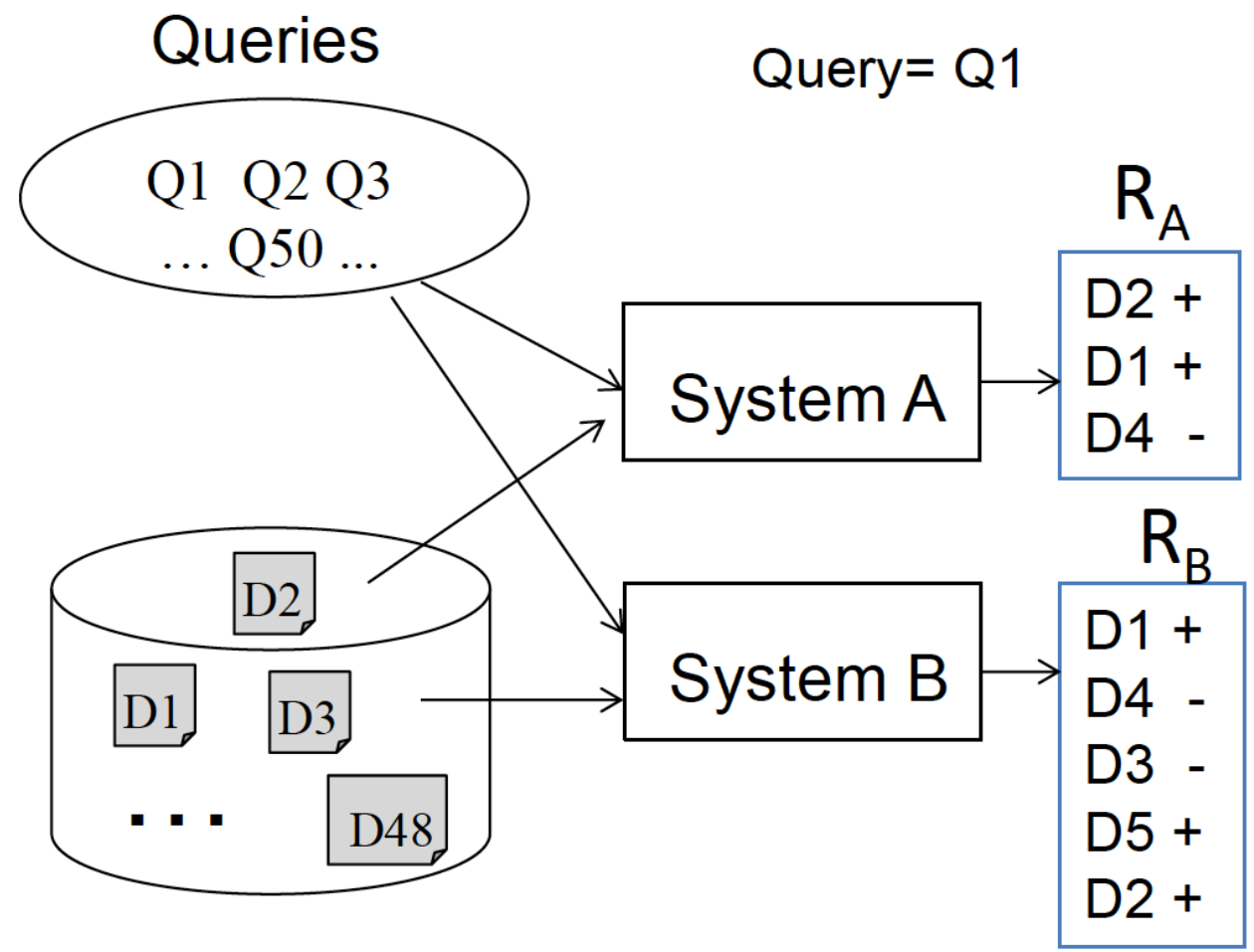
Evaluation of Text Retrieval Systems: Basic Measures

Dr. Iqra Safder

Evaluation of Text Retrieval Systems



Test Collection Evaluation



Which is better?

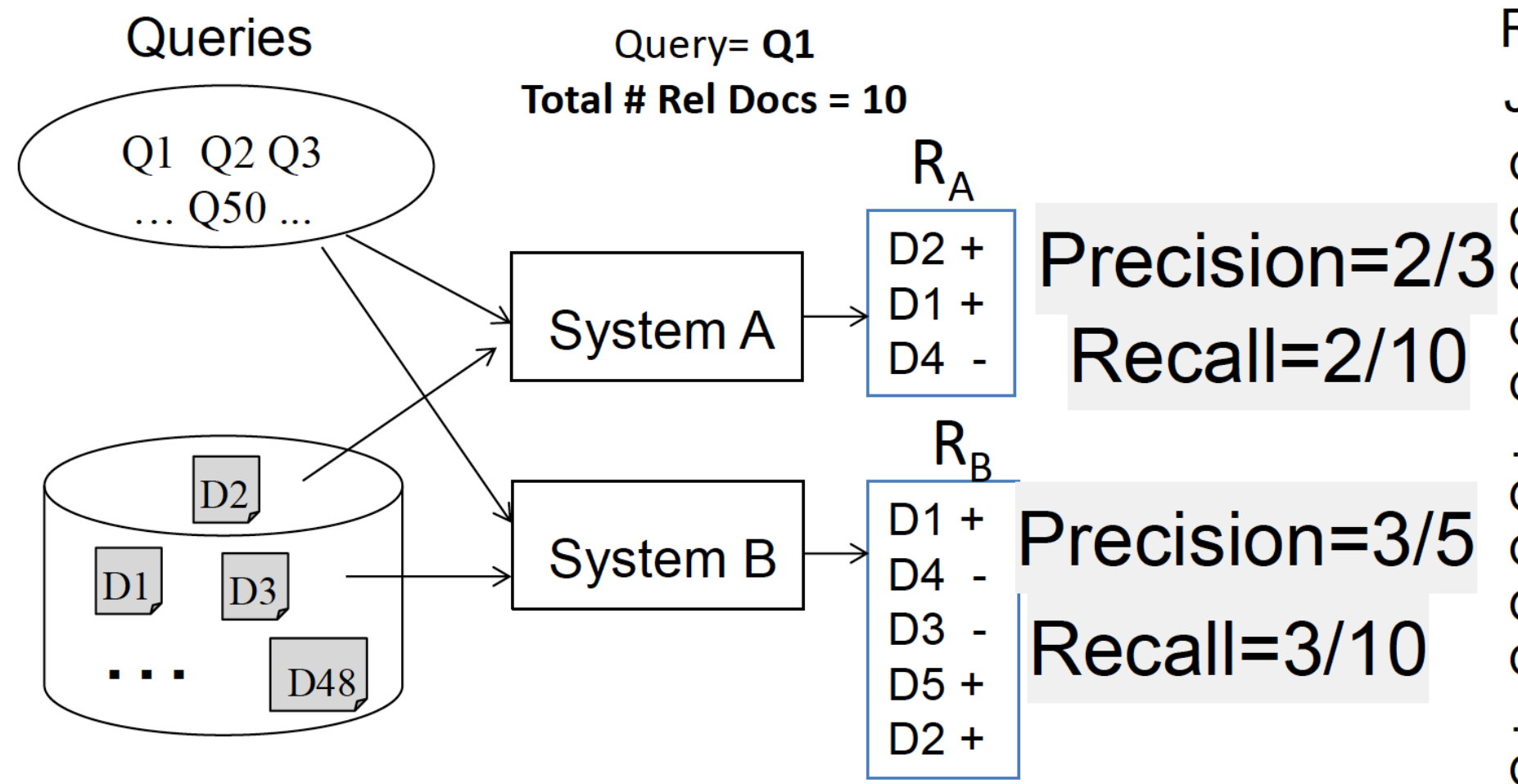
R_A or R_B?

How to quantify?

Relevance Judgments

Q1 D1 +
Q1 D2 +
Q1 D3 –
Q1 D4 –
Q1 D5 +
…
Q2 D1 –
Q2 D2 +
Q2 D3 +
Q2 D4 –
…
Q50 D1 –
Q50 D2 –
Q50 D3 +

Test Collection Evaluation



Relevance Judgments

Q1 D1 +
Q1 D2 +
Q1 D3 –
Q1 D4 –
Q1 D5 +
…
Q2 D1 –
Q2 D2 +
Q2 D3 +
Q2 D4 –
…
Q50 D1 –
Q50 D2 –
Q50 D3 +

Evaluating a Set of Retrieved Docs: Precision and Recall

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved **a**	Relevant Rejected **b**
Not relevant	Irrelevant Retrieved **c**	Irrelevant Rejected **d**

$$\text{Precision} = \frac{a}{a+c}$$

$$\text{Recall} = \frac{a}{a+b}$$

Ideal results: Precision=Recall=1.0

In reality, high recall tends to be associated with low precision

Set can be defined by a cutoff (e.g., precision @ 10 docs)

Combine Precision and Recall: F-Measure

$$F_\beta = \frac{1}{\frac{\beta^2}{\beta^2+1}\frac{1}{R} + \frac{1}{\beta^2+1}\frac{1}{P}} = \frac{(\beta^2+1)P*R}{\beta^2 P + R}$$

$$F_1 = \frac{2PR}{P+R}$$

Why not 0.5*P+0.5*R?

P: precision

R: recall

β: parameter (often set to 1)

Combine Precision and Recall: F-Measure

$$F_\beta = \frac{1}{\frac{\beta^2}{\beta^2+1}\frac{1}{R} + \frac{1}{\beta^2+1}\frac{1}{P}} = \frac{(\beta^2+1)P*R}{\beta^2 P + R}$$

$$F_1 = \frac{2PR}{P+R}$$

Why not 0.5*P+0.5*R?

P: precision

R: recall

β: parameter (often set to 1)

Sum will be dominated by the higher value in arithmatic mean.

Summary

- Precision: are the retrieved results all relevant?
- Recall: have all the relevant documents been retrieved?
- F measure combines Precision and Recall
- Tradeoff between Precision and Recall depends on the user's search task