Information Retrieval & Text Mining

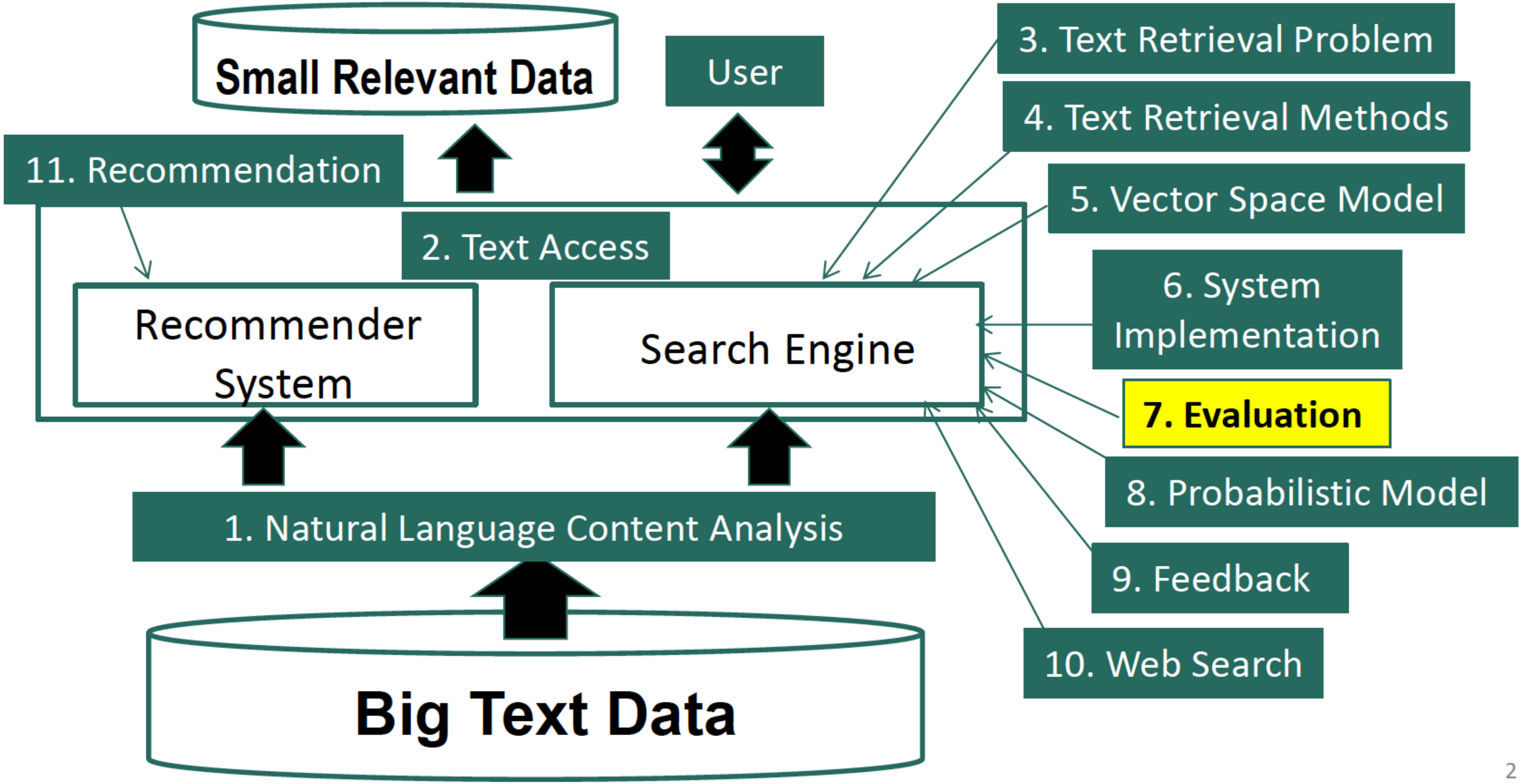
Evaluation of Text Retrieval Systems:

Basic Measure

Dr. Saeed Ul Hassan

Information Technology University

Evaluation of Text Retrieval Systems



Why Evaluation?

- Reason 1: Assess the actual utility of a TR system
    - Measures should reflect the utility to users in a real application
    - Usually done through user studies (interactive IR evaluation)
- Reason 2: Compare different systems and methods
    - Measures only need to be correlated with the utility to actual users, thus don't have to accurately reflect the exact utility to users
    - Usually done through test collections (test set IR evaluation)

What to Measure?

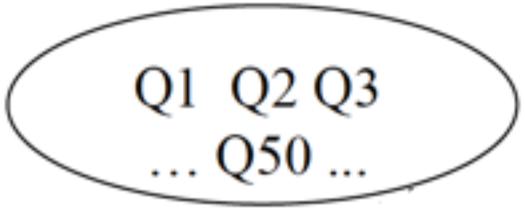
- Effectiveness/Accuracy: how accurate are the search results?
  - Measuring a system's ability of ranking relevant docucments on top of non-relevant ones
- Efficiency: how quickly can a user get the results? How much computing resources are needed to answer a query?
  - Measuring space and time overhead
- Usability: How useful is the system for real user tasks?
  - Doing user studies

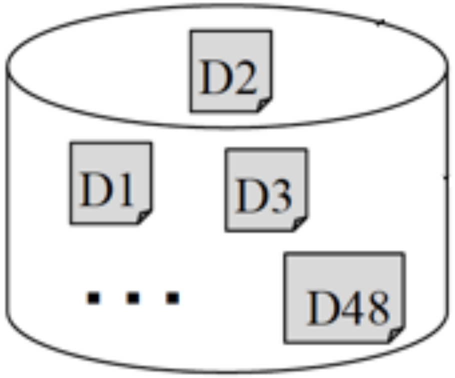
The Cranfield Evaluation Methodology

- A methodology for laboratory testing of system components developed in 1960s
- Idea: Build reusable test collections & define measures
  - A sample collection of documents (simulate real document collection)
  - A sample set of queries/topics (simulate user queries)
  - Relevance judgments (ideally made by users who formulated the queries) ➔ Ideal ranked list
  - Measures to quantify how well a system’s result matches the ideal ranked list
- A test collection can then be reused many times to compare different systems

Test Collection Evaluation

Queries



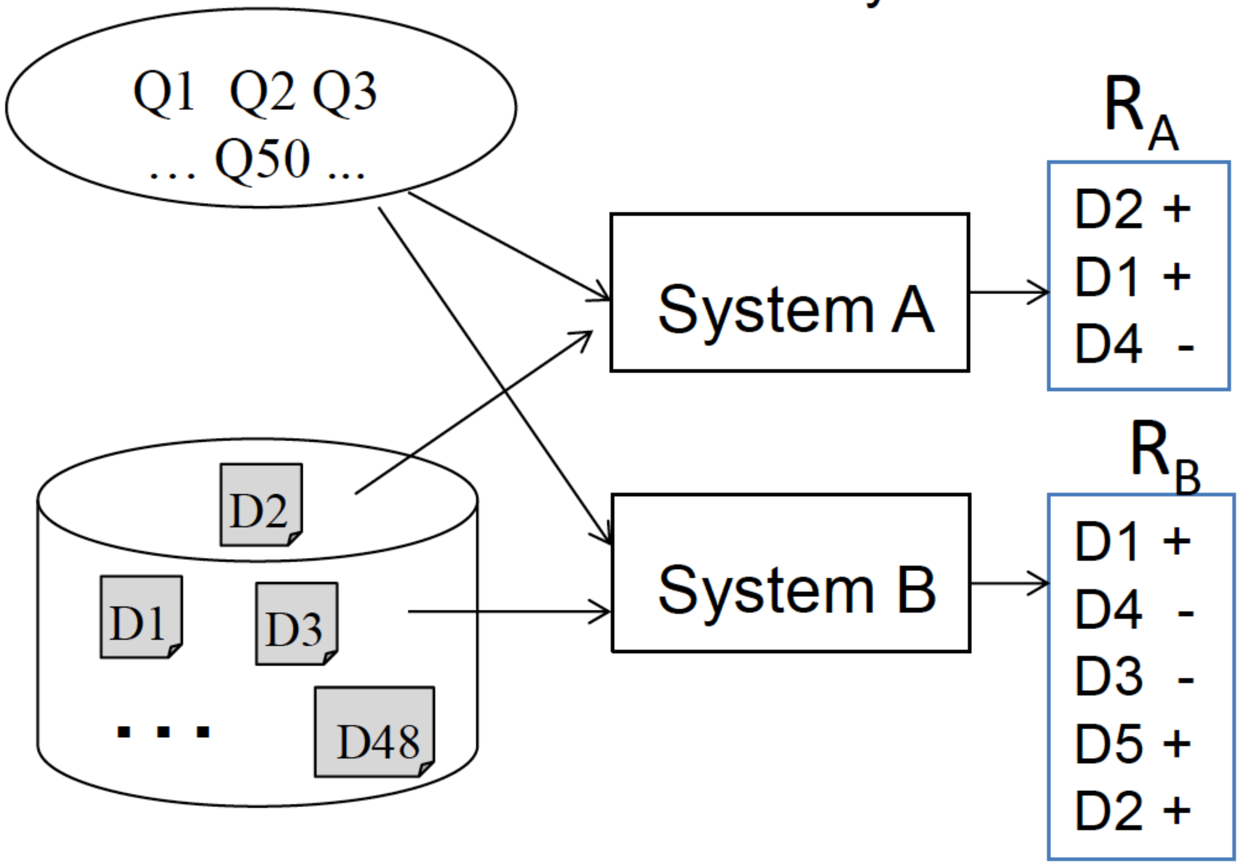


Document Collection

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



Queries

Query= Q1

Q1 Q2 Q3
… Q50 …

System A

System B

R_A

D2 +
D1 +
D4 -

R_B

D1 +
D4 -
D3 -
D5 +
D2 +

D2
D1
D3
. . .
D48

Document Collection

Which is better?

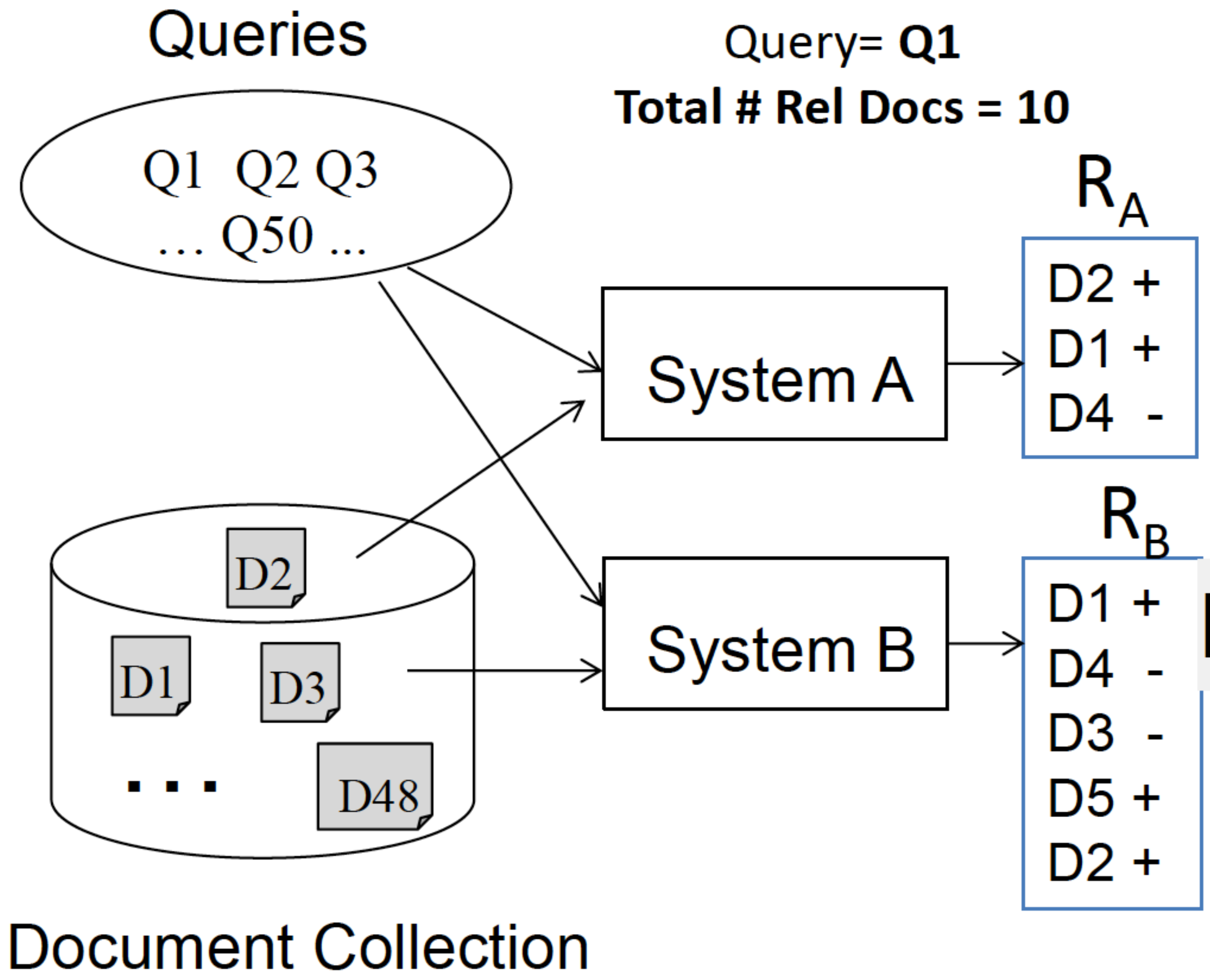
R_A or R_B?

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



Queries

Q1 Q2 Q3
… Q50 …

Document Collection

Query= **Q1**

Total # Rel Docs = 10

R_A

D2 +
D1 +
D4 -

Precision=2/3

Recall=2/10

R_B

D1 +
D4 -
D3 -
D5 +
D2 +

Precision=3/5

Recall=3/10

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

Action Doc	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)

How to combine Precision and Recall ?

How about

$$0.5*P + 0.5*R = ?$$

$$\frac{P + R}{2} = ?$$

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)

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