Information Retrieval

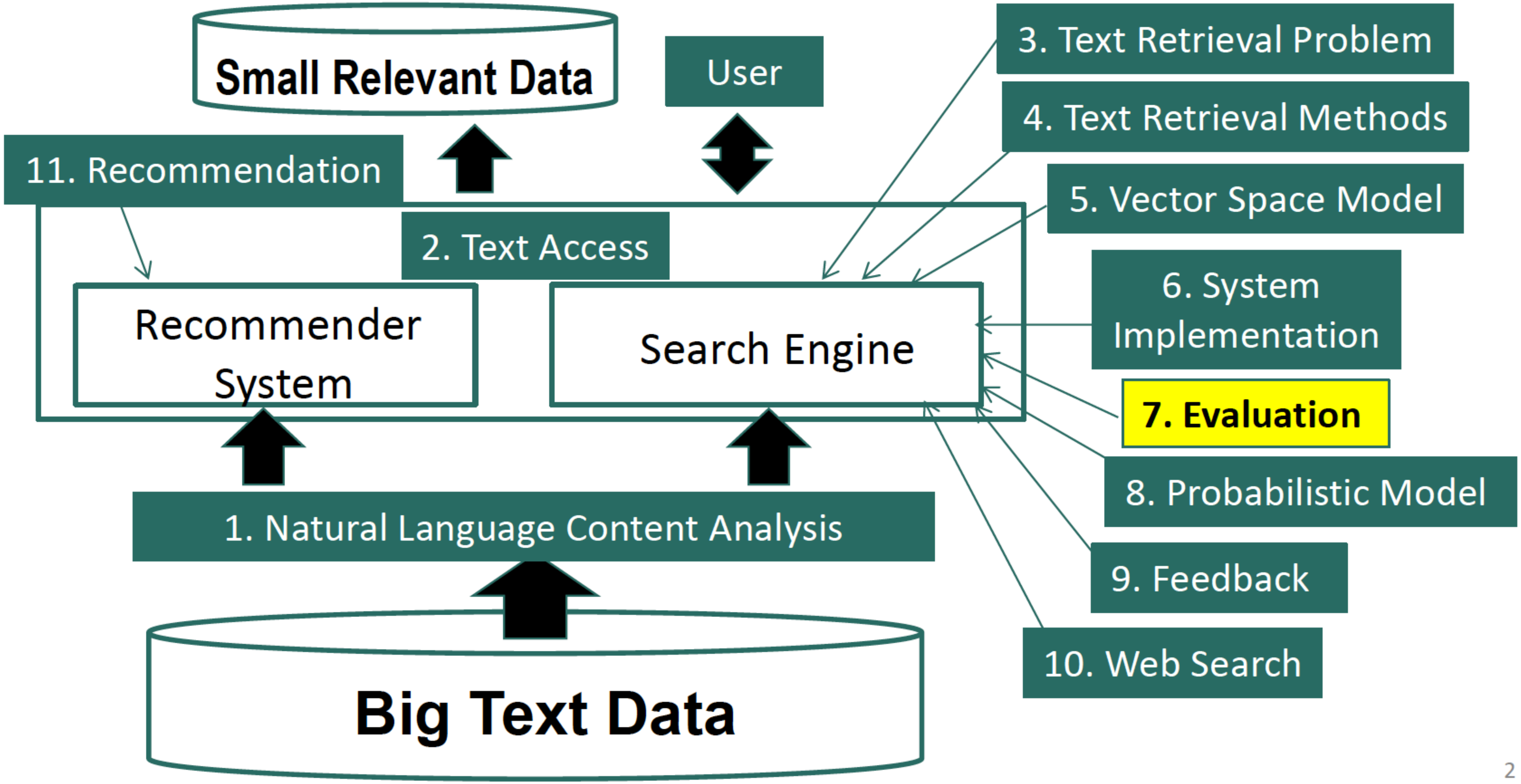
Evaluation of Text Retrieval Systems:

Multi level Ranked list Evaluations

Practical Issues

Dr. Iqra Safder

Evaluation of Text Retrieval Systems



Quiz 3 & 4

- 22nd April (11-14)
- 23rd April (14-17)

What If We Have Multi-level Relevance Judgments?

Relevance level: r=1 (non-relevant) , 2 (marginally relevant), 3 (very relevant)

	Gain	Cumulative Gain	Discounted Cumulative Gain
D1	3	3	3
D2	2	3+2	3+2/log 2
D3	1	3+2+1	3+2/log 2+1/log 3
D4	1	3+2+1+1	
D5	3		
D6	1		
D7	1	. . .	
D8	2		
D9	1		
D10	1		

- **DCG@10 = 3+2/log 2+1/log 3 +…+ 1/log 10**

Normalized DCG=?

$$\frac{DCG@10}{IdealDCG@10}$$

IdealDCG@10 = 3+3/log 2+3/log 3 +…+ 3/log 9+ 2/log 10

Assume: there are 9 documents rated “3” in total in the collection

	Gain	Cumulative gain	Discounted cumulative gain
D_1	3	3	3
D_2	2	$3 + 2$	$3 + 2/\log 2$
D_3	1	$3 + 2 + 1$	$3 + 2/\log 2 + 1/\log 3$
D_4	1	$3 + 2 + 1 + 1$	$\cdots$
D_5	3	$\cdots$	
D_6	1		
D_7	1		
D_8	2		
D_9	1		
D_{10}	1		

$$\text{Normalized DCG} = \frac{\text{DCG@10}}{\text{IdealDCG@10}}$$

$$\text{DCG@10} = 3 + 2/\log 2 + 1/\log 3 + \cdots + 1/\log 10$$

$$\text{IdealDCG@10} = 3 + 3/\log 2 + 3/\log 3 + \cdots + 3/\log 9 + 2/\log 10$$

Relevance level: $r = 1$ (non-relevant), 2 (marginally relevant), 3 (very relevant)

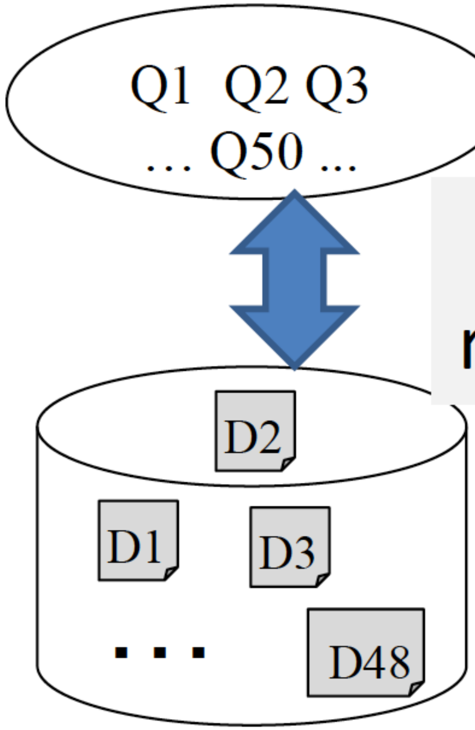
Assume: there are 9 documents rated “3” in total in the collection

Normalized Discounted Cumulative Gain (nDCG)

- Applicable to multi-level judgments in a scale of [1,r], r>2
- Main idea of nDCG@k documents
  - Measure the total utility of the top k documents to a user
  - Utility of a lowly ranked document is discounted
  - Normalized to ensure comparability across queries

Challenges in Creating a Test Collection

Queries: representative & many



Existence of relevant docs

Docs: representative & many

Relevance Judgments

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

Judgments: completeness vs. minimum human work

Measures: capture the perceived utility by users

Statistical Significance Tests

- How sure can you be that an observed difference doesn't simply result from the particular queries you chose?

Experiment 1

Query	System A	System B
1	0.20	0.40
2	0.21	0.41
3	0.22	0.42
4	0.19	0.39
5	0.17	0.37
6	0.20	0.40
7	0.21	0.41
Average	0.20	0.40

Experiment 2

Query	System A	System B
1	0.02	0.76
2	0.39	0.07
3	0.16	0.37
4	0.58	0.21
5	0.04	0.02
6	0.09	0.91
7	0.12	0.46
Average	0.20	0.40

How reliable is our conclusion by simply looking at the mean average precision? YES, Intutively we can say E1 is better, but how can we quantitatively answer this question…that's the reason we need statistical significance test.

Pooling: Avoid Judging all Documents

- If we can’t afford judging all the documents in the collection, which subset should we judge?
- Pooling strategy
  - Choose a diverse set of ranking methods (TR systems)
  - Have each to return top-K documents
  - Combine all the top-K sets to form a pool for human assessors to judge
  - Other (unjudged) documents are usually assumed to be non-relevant (though they don’t have to)
  - Okay for comparing systems that contributed to the pool, but problematic for evaluating new systems

Summary of TR Evaluation

- Extremely important!
  - TR is an empirically defined problem
  - Inappropriate experiment design misguides research and applications
  - Make sure to get it right for your research or application
- Cranfield evaluation methodology is the main paradigm
  - MAP and nDCG: appropriate for comparing ranking algorithms
  - Precision@10docs is easier to interpret from a user's perspective
- Not covered
  - A-B Test [Sanderson 10]
  - User studies [Kelly 09]

Sign Test
Wilcoxon Signed Rank   Test

Additional Readings

- Donna Harman, Information Retrieval Evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers 2011

- Mark Sanderson, Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval 4(4): 247-375 (2010)

- Diane Kelly, Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3(1-2): 1-224 (2009)