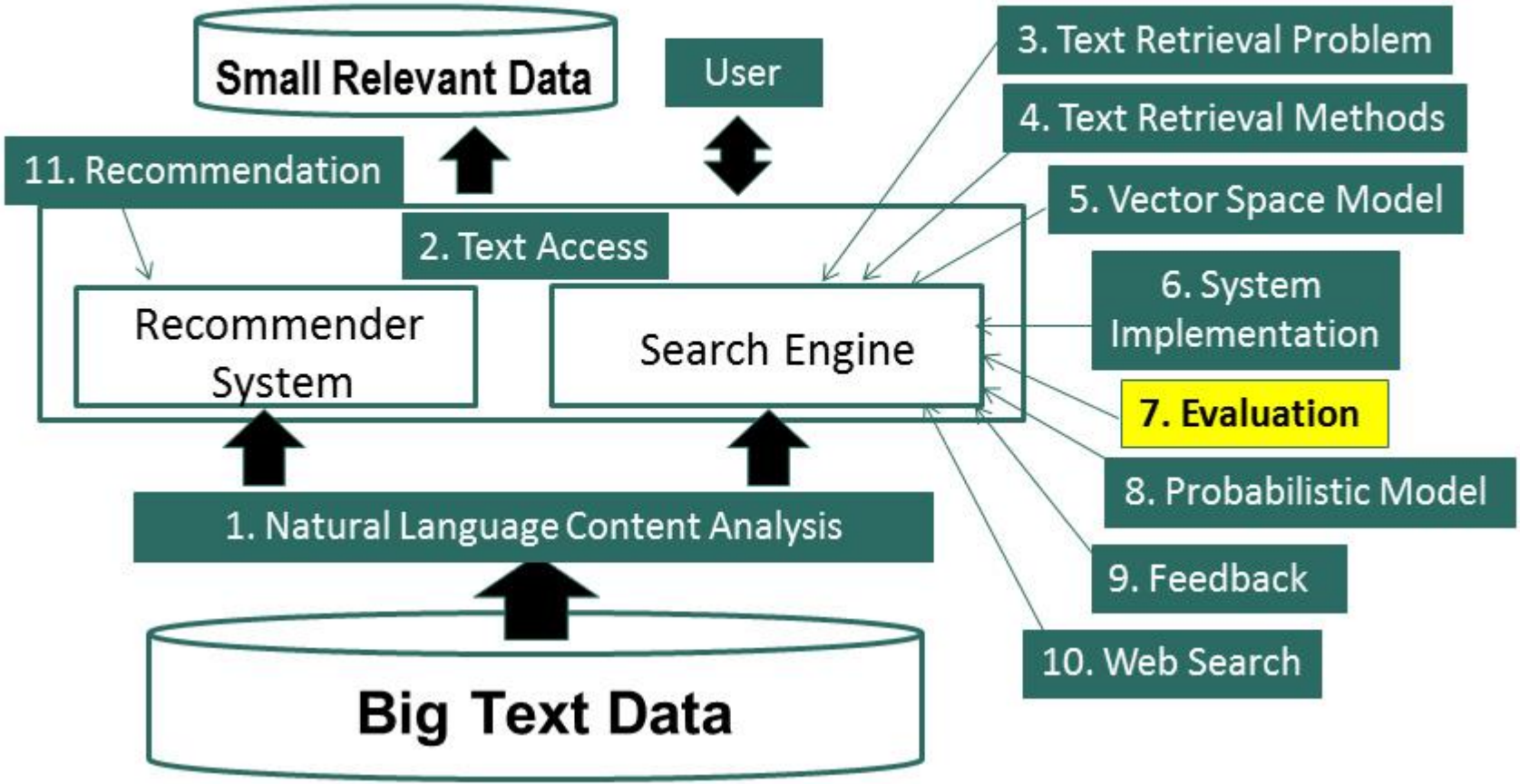
Text Retrieval and Search Engines

Evaluation of TR Systems: Practical Issues

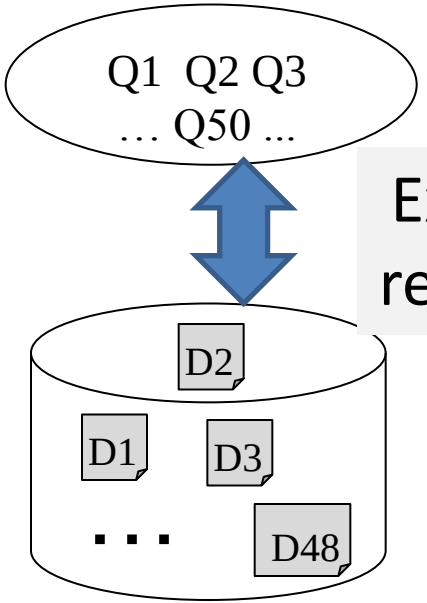
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Course Schedule



Challenges in Creating a Test Collection

Queries: representative & many



Existence of relevant docs

Measures: capture the perceived utility by users

Docs: representative & many

Relevance Judgments

Judgments: completeness vs. minimum human work

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

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

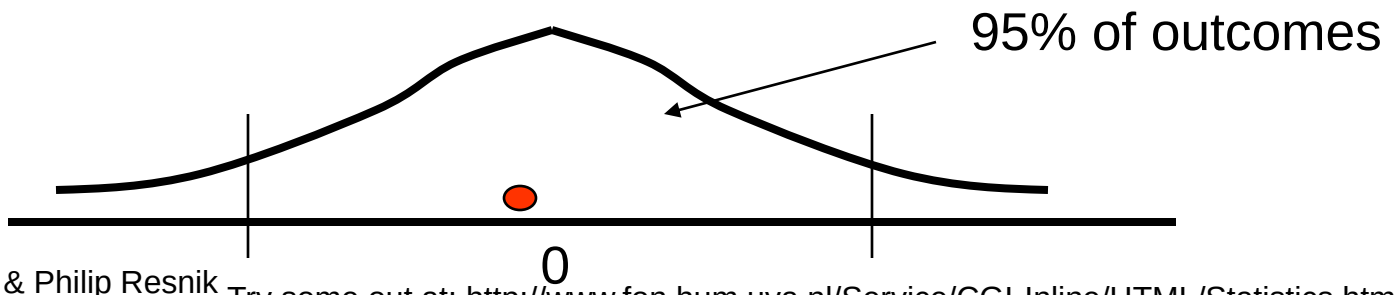
Experiment 2

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

Slide from Doug Oard & Philip Resnik

Statistical Significance Testing

Query	System A	System B	Sign Test	Wilcoxon
1	0.02	0.76	+	+0.74
2	0.39	0.07	-	- 0.32
3	0.16	0.37	+	+0.21
4	0.58	0.21	-	- 0.37
5	0.04	0.02	-	- 0.02
6	0.09	0.91	+	+0.82
7	0.12	0.46	+	+0.34
Average	0.20	0.40	p=1.0	p=0.9375



Slide from Doug Oard & Philip Resnik

Try some out at: http://www.fon.hum.uva.nl/Service/CGI-Inline/HTML/Statistics.html

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]

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)