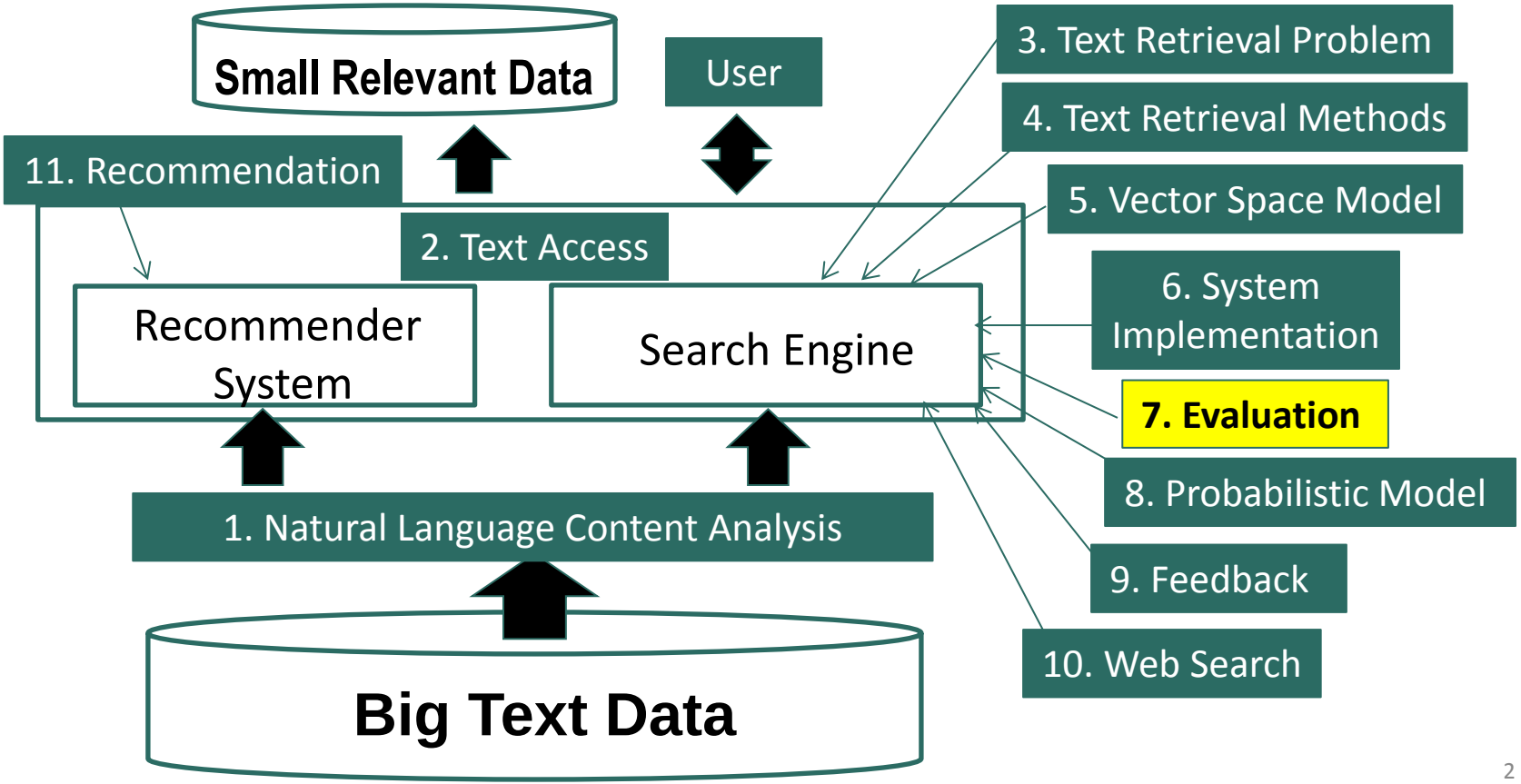
Text Retrieval and Search Engines

Evaluation of TR Systems

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

Evaluation of Text Retrieval Systems



1. Evaluation of TR 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?

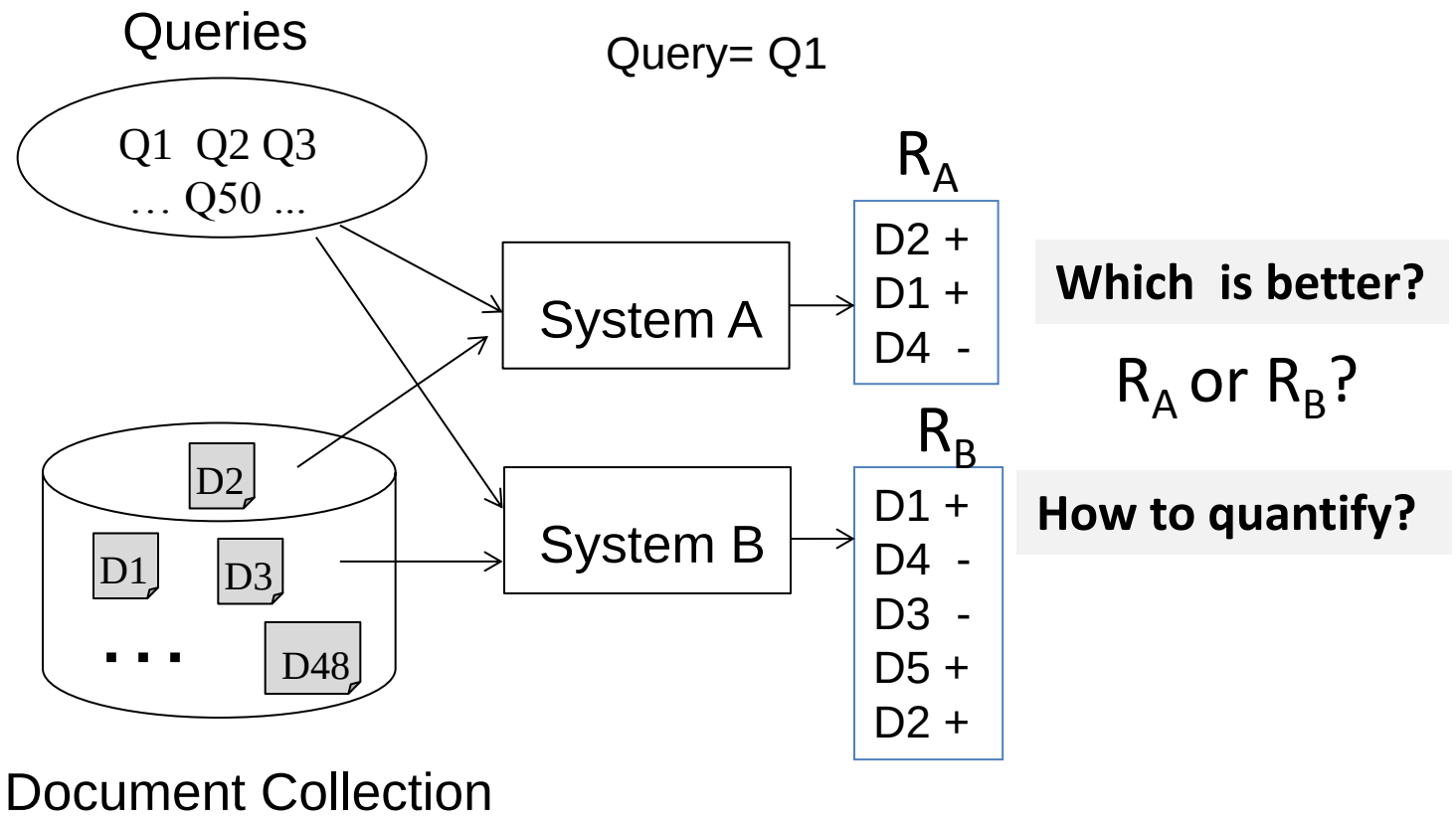
- Effectiveness/Accuracy: how accurate are the search results?
  - Measuring a system's ability of ranking relevant docucments on top of non-relevant ones
- Efficiency: how quickly can a user get the results? How much computing resources are needed to answer a query?
  - Measuring space and time overhead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

2. Evaluation of TR Systems - Basic Measures

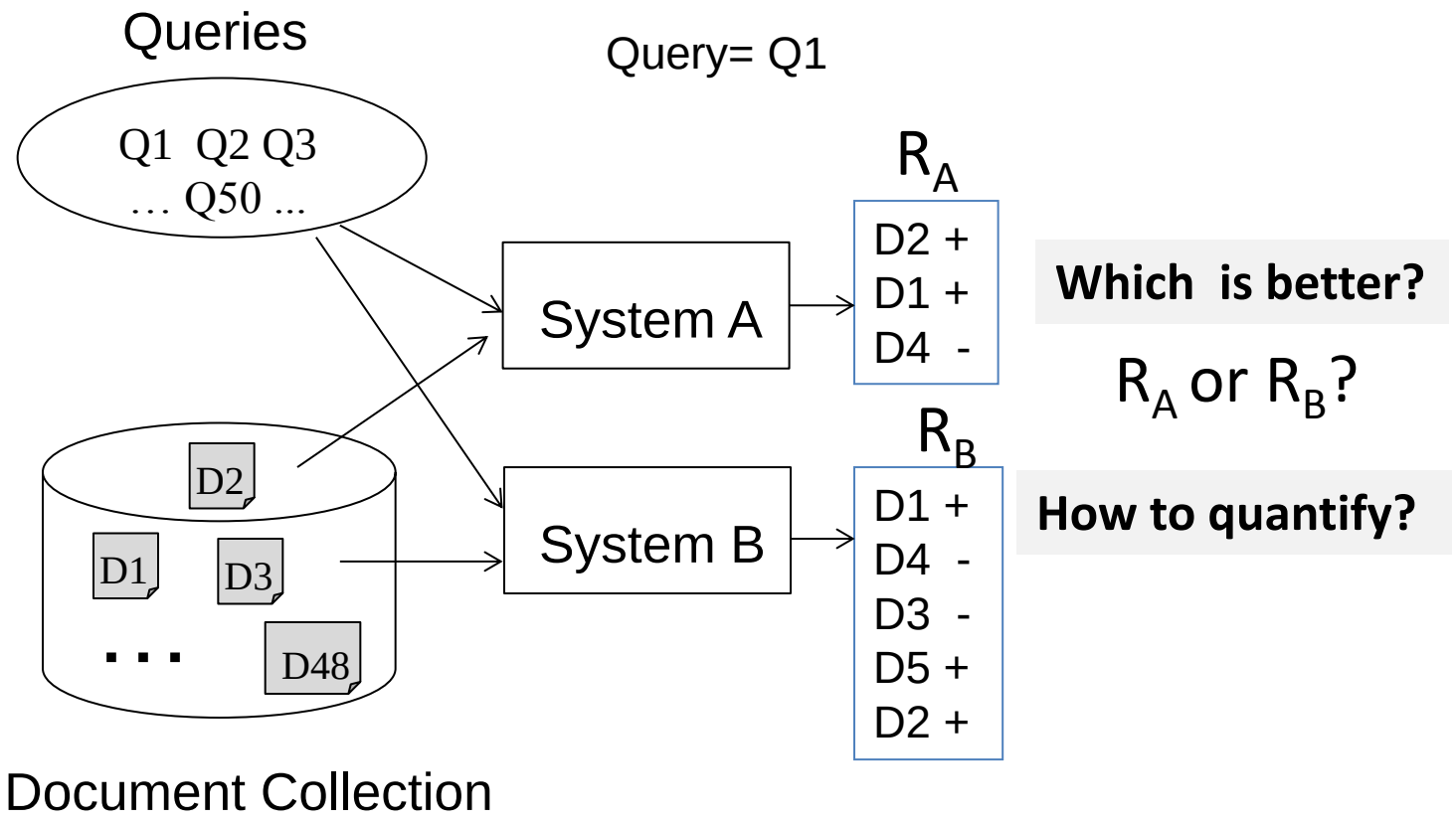
Test Collection Evaluation



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 +

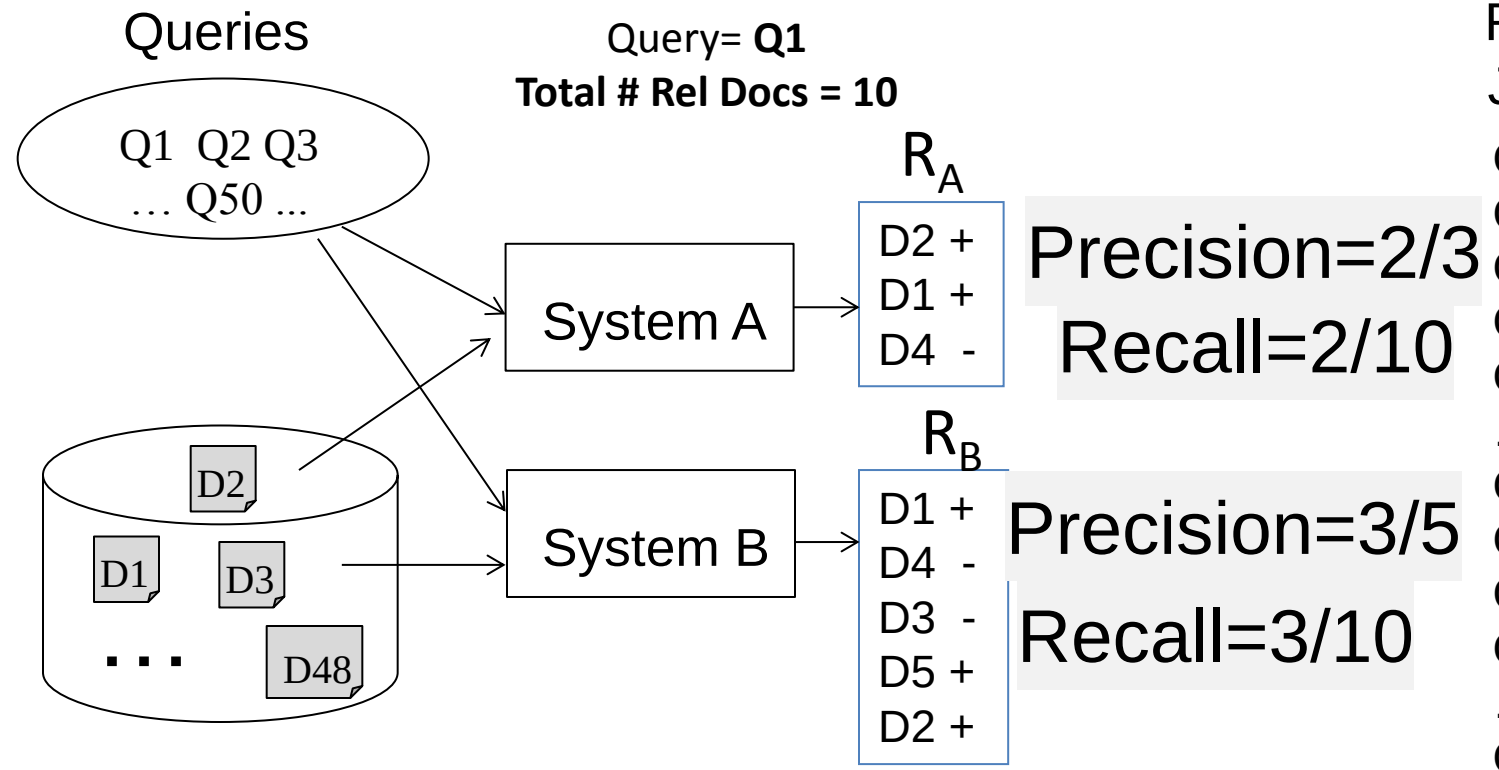
Test Collection Evaluation



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 +

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)

Summary

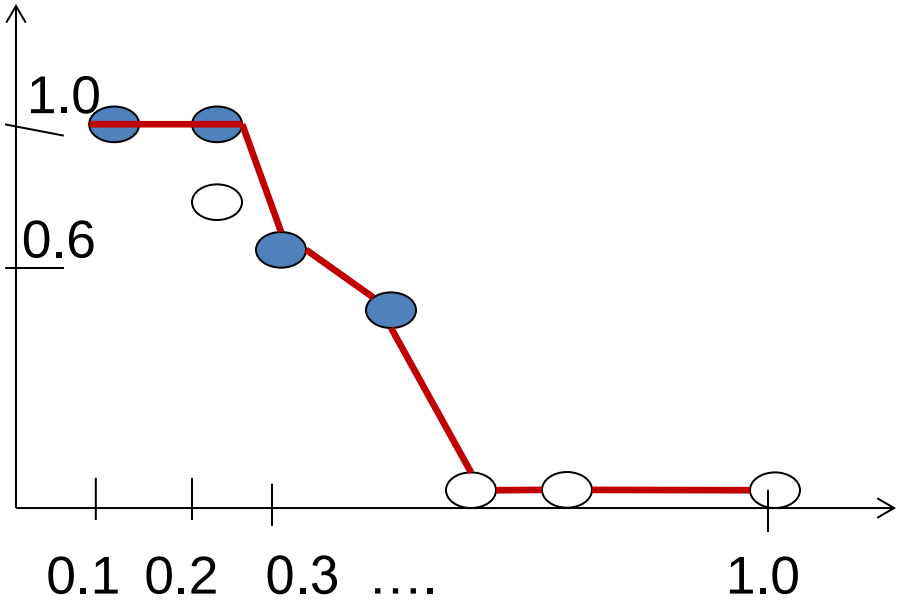
- Precision: are the retrieved results all relevant?
- Recall: have all the relevant documents been retrieved?
- F measure combines Precision and Recall
- Tradeoff between Precision and Recall depends on the user’s search task

3. Evaluation of TR Systems - Evaluating Ranked Lists (1)

Evaluating Ranking: PR Curve

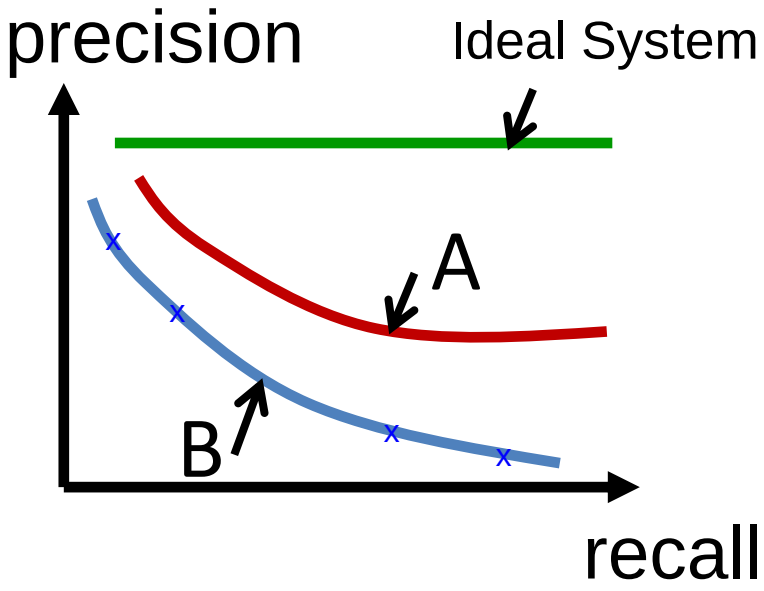
Total number of relevant documents in collection: 10

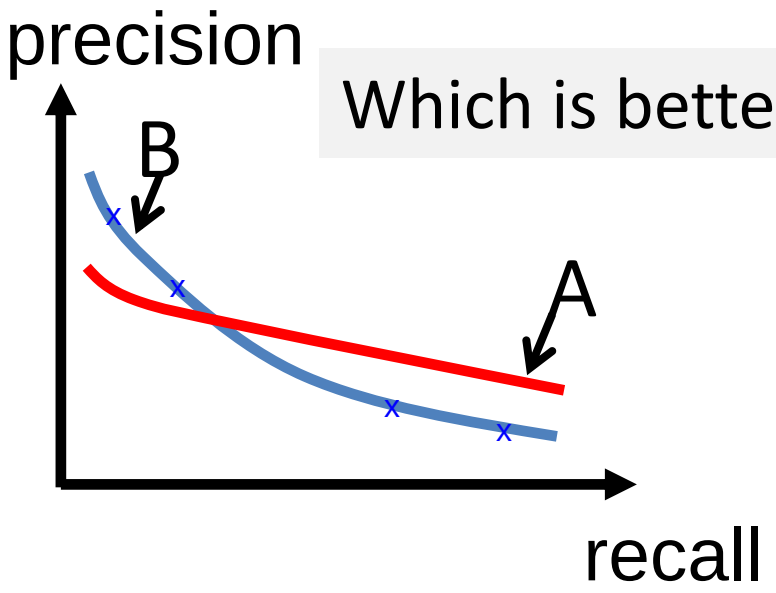
	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	10/10



Assume Precision=0?

Comparing PR Curves





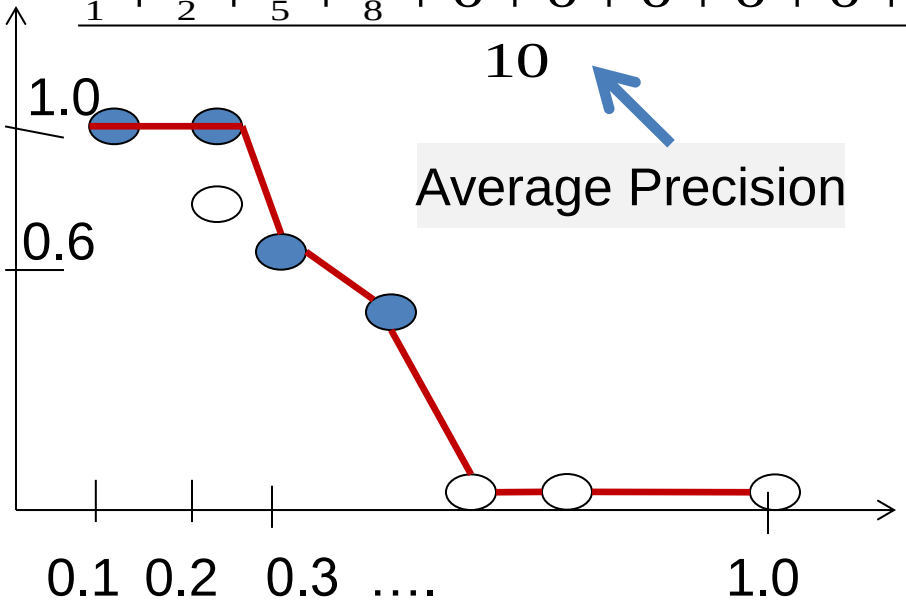
How to Summarize a Ranking

Total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	0	10/10

$$\frac{\frac{1}{1}+\frac{2}{2}+\frac{3}{5}+\frac{4}{8}+0+0+0+0+0+0}{10}$$

Average Precision



3. Evaluation of TR Systems - Evaluating Ranked Lists (2) Mean Average Precision (MAP)

- Average Precision:
  - The average of precision at every cutoff where a new relevant document is retrieved
  - Normalizer = the total # of relevant docs in collection
  - Sensitive to the rank of each relevant document
- Mean Average Precisions (MAP)
  - MAP = arithmetic mean of average precision over a set of queries
  - gMAP = geometric mean of average precision over a set of queries
  - Which is better: MAP or gMAP?

Special Case: Mean Reciprocal Rank

- When there's only one relevant document in the collection (e.g., known item search)
  - Average Precision = Reciprocal Rank = 1/r, where r is the rank position of the single relevant doc
  - Mean Average Precision ➔ Mean Reciprocal Rank
  - Why not simply use r?

Summary

- Precision-Recall curve characterizes the overall accuracy of a ranked list
- The **actual** utility of a ranked list depends on how many top-ranked results a user would examine
- Average Precision is the standard measure for comparing two ranking methods
  - Combines precision and recall
  - Sensitive to the rank of **every** relevant document

What if we have multiple levels of relevance judgments?

5. Evaluation of TR Systems - Multi-Level Judgements

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		

▪ ▪ ▪

▪ ▪ ▪

Normalized DCG=?

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

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

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

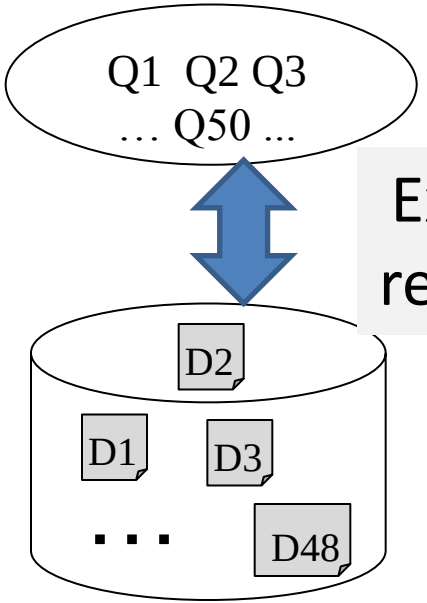
D Assume: there are 9 documents rated “3” in total in the collection

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

6: Evaluation of TR Systems - Practical Issues
Challenges in Creating a Test Collection

Queries: representative & many



Existence of relevant docs

Judgments: completeness vs. minimum human work

Measures: capture the perceived utility by users

Docs: representative & many

Relevance Judgments

…
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

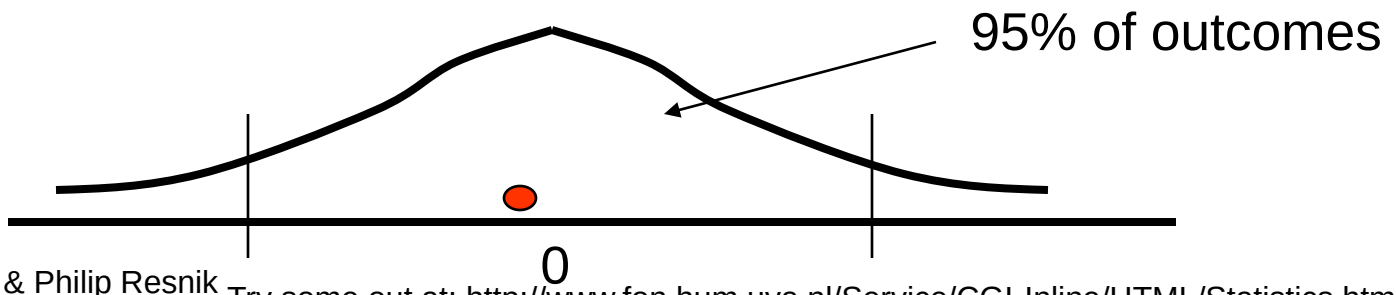
Experiment 2

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)