Multimedia Information Retrieval and Technology

Lecture 8 Evaluation I

By : Fangyu Wu

Room: SD555

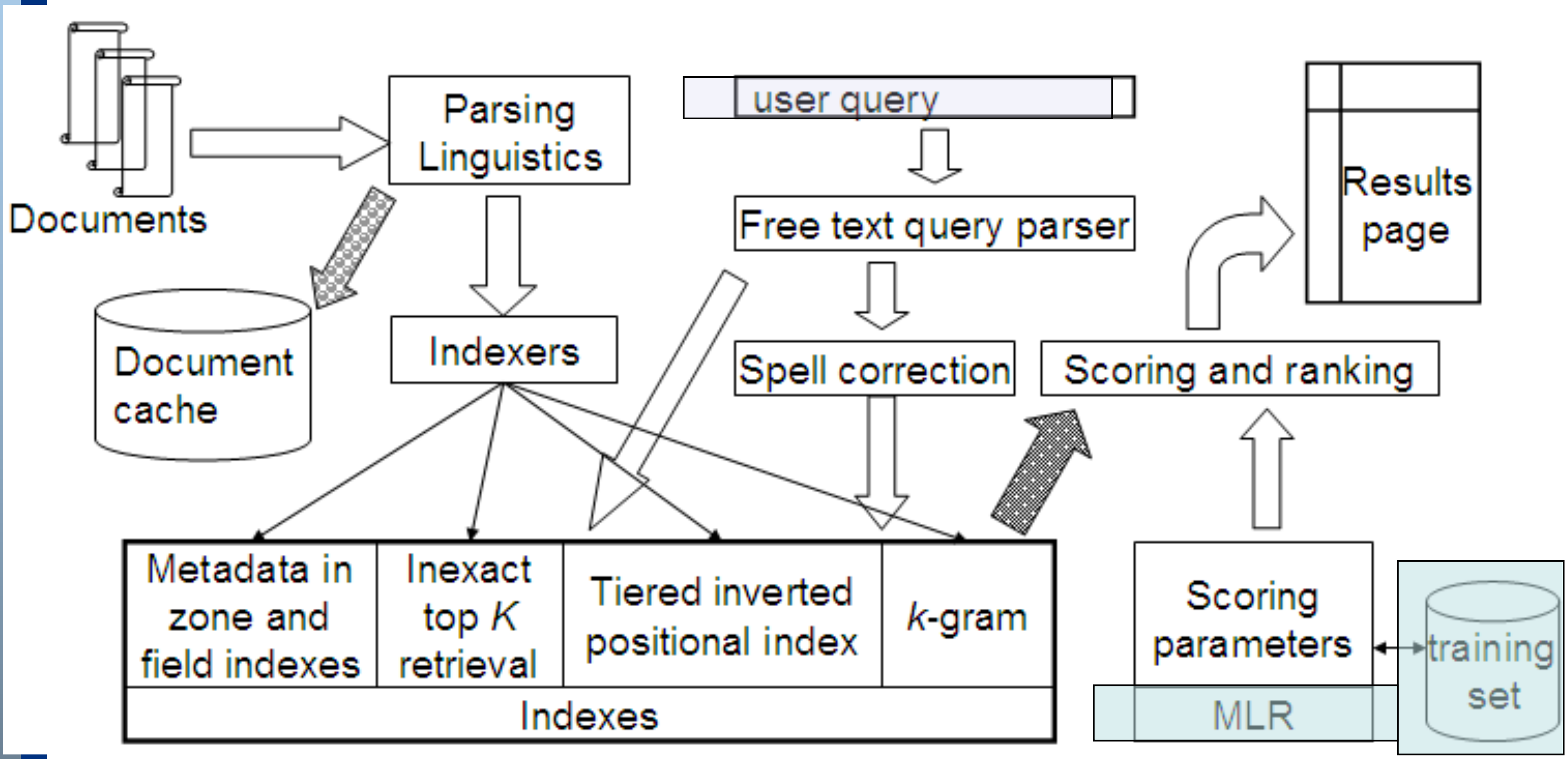
Xi’an Jiaotong-Liverpool University
西交利物浦大学

Exercise

Draw yourself a diagram showing the various function modules in a search engine incorporating all the functionality we have talked about.

Identify some of the key design choices in the search engine pipeline if you can.

Recap


Documents
Parsing
Linguistics
user query
Results
page
Free text query parser
Document
cache
Indexers
Spell correction
Scoring and ranking
Metadata in
zone and
field indexes
Inexact
top K
retrieval
Tiered inverted
positional index
k-gram
Indexes
Scoring
parameters
MLR
training
set

- Evaluation
  - ⍰ Evaluation of unranked retrieval results
  - ⍰ Evaluation of ranked retrieval results
    - ❑ Eleven-point interpolated average precision
    - ❑ MAP
    - ❑ Precision@K
    - ❑ Normalized Discounted Cumulative Gain

Situation

Thanks to your stellar performance, you quickly rise to VP of Search at internet retail giant nozama.com. Your boss brings in her nephew Sergey, who claims to have built a better search engine for nozama.

What do you do?

What could you ask Sergey?

How fast does it index?

- Number of documents/hour
- Incremental indexing – nozama adds 10K products/day

How fast does it search?

- Latency and CPU needs for nozama’s 5 million products

Does it recommend related products?

what is more important?

You want nozama’s users to be happy with the search experience

How do you tell if users are happy?

Search returns products **relevant** to users

- How do you assess this at scale?

Search results get clicked a lot

- Misleading titles/summaries can cause users to click

Users buy after using the search engine

- Or, users spend a lot of $ after using the search engine

Repeat visitors/buyers

- Do users leave soon after searching?
- Do they come back within a week/month/… ?

Evaluation

Factors include:

1. Speed of response
2. Size of index (storage)
3. Uncluttered UI
4. Relevance results

User happiness!

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Searcher finds what she was looking for.
- Web search engine: advertiser. Do searchers click through to my ads?
- Ecommerce: buyer. Buyer buys what she came to the site to buy.
- Ecommerce: seller. Seller is able to sell her wares (because the search function directed buyers to the correct items).
- Enterprise: CEO. Employees are more productive because they find right away what they are looking for.

Measuring relevance

To measure IR effectiveness, we need a test collection consisting of **Three elements**:

1. A benchmark document collection

   Must be representative

2. A benchmark suite of information needs, expressible as queries

   Again, representative

3. An assessment of either Relevant or Nonrelevant judgments for each query-document pair.

   How?

Exercise I

$$P = \frac{TP}{TP+FP} = \frac{8}{18} \approx 0.44$$

$$R = \frac{TP}{TP+FN} = \frac{8}{20} = 0.4$$

An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

	20 Relevant	Non-relevant
Retrieved	8	10
Not retrieved	12	

Solution

Precision = 8/18 = 0.44; Recall = 8/20 =0.4.

- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.
  - Relevance assessments are only usable if they are consistent.
  - How can we measure this consistency or agreement among judges?

Relevance judgments

Binary (relevant vs. non-relevant) in the simplest case, more nuanced (0, 1, 2, 3 …) in others.

This decision is referred to as the ***gold standard*** or ***ground truth*** judgment of relevance.

Evaluating an IR system

Note: **user need** is translated into a **query**

E.g., Information need: *My swimming pool bottom is becoming black and needs to be cleaned.*

Query: ***pool cleaner***

Assess whether the doc addresses the underlying need, not whether it has these words.

Some public test Collections

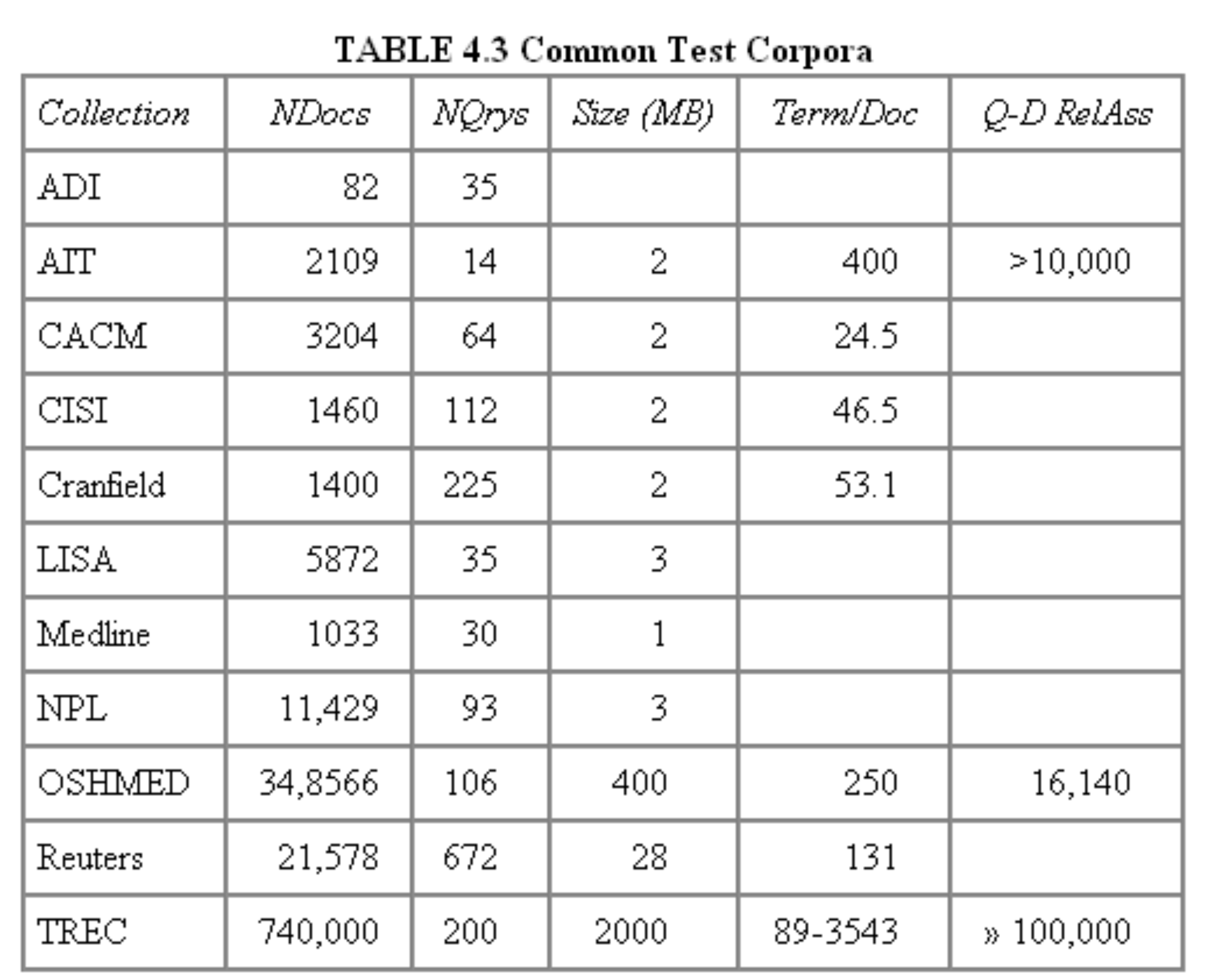


TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Standard relevance benchmark: Cranfield

- Pioneering: first testbed allowing precise quantitative measures of information retrieval effectiveness
- Late 1950s, UK
- 1398 abstracts of aerodynamics journal articles, a set of 225 queries, exhaustive relevance judgments of all query-document-pairs
- Too small, too untypical for serious IR evaluation today

Standard relevance benchmark: TREC

- TREC = Text Retrieval Conference (TREC)
- Organized by the U.S. National Institute of Standards and Technology (NIST)
- TREC is actually a set of several different relevance benchmarks.
- Best known: TREC Ad Hoc, used for first 8 TREC evaluations between 1992 and 1999
- 1.89 million documents, mainly newswire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
- Rather, NIST assessors' relevance judgments are available only for the documents that were among the top *k* returned for some system which was entered in the TREC evaluation for which the information need was developed.

Standard relevance benchmarks: Others

- GOV2
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- NTCIR
  - East Asian language and cross-language information retrieval
- Cross Language Evaluation Forum (CLEF)
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
- Many others

Evaluation

Evaluation of unranked retrieval results

Evaluation of ranked retrieval results

- Eleven-point interpolated average precision
- MAP
- Precision@K
- Normalized Discounted Cumulative Gain

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Recap

- Precision (P) is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

Precision and recall

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$\begin{aligned} P &= TP/(TP+FP) \\ R &= TP/(TP+FN) \end{aligned}$$

Exercise I

An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

Solution

Precision = 8/18 = 0.44; Recall = 8/20 =0.4.

Precision and Recall

Different applications have different concerns

The advantage of having the two numbers for precision and recall :

- Typical web surfers (high precision)
- Paralegals and intelligence analysts (high recall)

Recall is a nondecreasing function of the number of documents retrieved.

In general, we want to get some amount of recall while tolerating only a certain percentage of false positives.

Evaluation

Evaluation of unranked retrieval results

Evaluation of ranked retrieval results

- Eleven-point interpolated average precision
- MAP
- Precision@K
- Normalized Discounted Cumulative Gain

Rank-Based Measures

Precision, recall, and the F measure are set-based measures. They are computed using unordered sets of documents.

Binary relevance

- Precision@K (P@K)
- Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)

Multiple levels of relevance

- Normalized Discounted Cumulative Gain (NDCG)

Binary relevance

100

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN}$$

	Relevant	Irrelevant
ret'd	TP	FP
Not r'd	FN	TN

There are a total of 100 relevant documents in the collection. What is the precision of the system at each step , and what is its recall?

① 1 | 0
 99 | 0

② 2 | 0
 98 | 0

③ 2 | 1
 98 | 0

④ 2 | 2
 98 | 0

⑤ 3 | 2
 97 | 0

⑥ 3 | 3
 97 | 0

⑦ 4 | 3
 96 | 0

		P	R
1	R	1	0.01
2	R	1	0.02
3	N	≃0.67	0.02
4	N	0.5	0.02
5	R	0.6	0.03
6	N	0.5	0.03
7	R	≃0.57	0.04

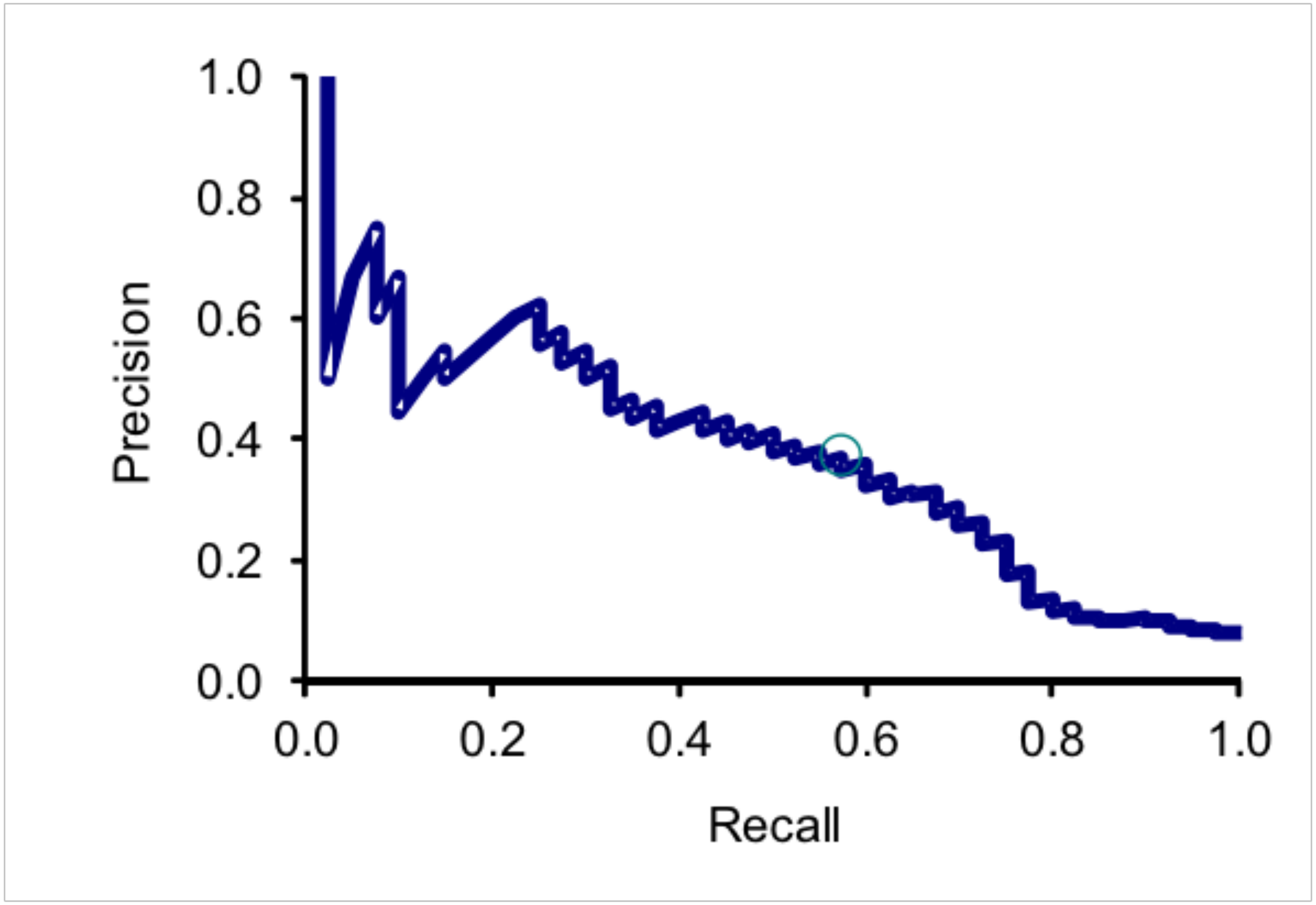
Binary relevance

There are a total of 100 relevant documents in the collection. What is the precision of the system at each step , and what is its recall?

		P	R
1	R	1	0.01
2	R	1	0.02
3	N	2/3	0.02
4	N	2/4	0.02
5	R	3/5	0.03
6	N	3/6	0.03
7	R	4/7	0.04

A precision-recall curve

- The top *k* score documents.
- Precision and recall values can be plotted to give a ***precision-recall curve***,



The precision-recall curves

The precision-recall curves have a **distinctive sawtooth shape**:

- If the (k+1)th document retrieved is nonrelevant, then recall is the same as for the top k documents, but precision has dropped.
- If it is relevant, then both precision and recall increase.

Examining the **entire precision-recall curve** is very informative, but there is often a desire to boil this information down to a few numbers or even a single number.

The precision-recall curves

The *interpolated precision* p_{interp} at a certain recall level *r* is defined as the highest precision found for any recall level $r^{'} \geq r$:

$$p_{interp}(r) = max_{r^{'} \geq r} p(r^{'})$$

take maximum of all future points!

Red line: Interpolated precision.

Xi'an Jiaotong-Liverpool University
西交利物浦大学

A interpolated precision-recall curve？

$p_{interp}(0.03) = ?$ (first 7 steps)

$= \max_{r' \geq r} p(r') = 0.6$

9		P	R
1	R	1	0.01
2	R	1	0.02
3	N	2/3	0.02
4	N	2/4	0.02
5	R	3/5 0.6	0.03
6	N max	3/6 0.5	0.03
→ 7	R	4/7 0.57	0.04
→ 8	R	5/8 = 0.625	0.04
9 ←	N	5/9 ≈ 0.56	0.05

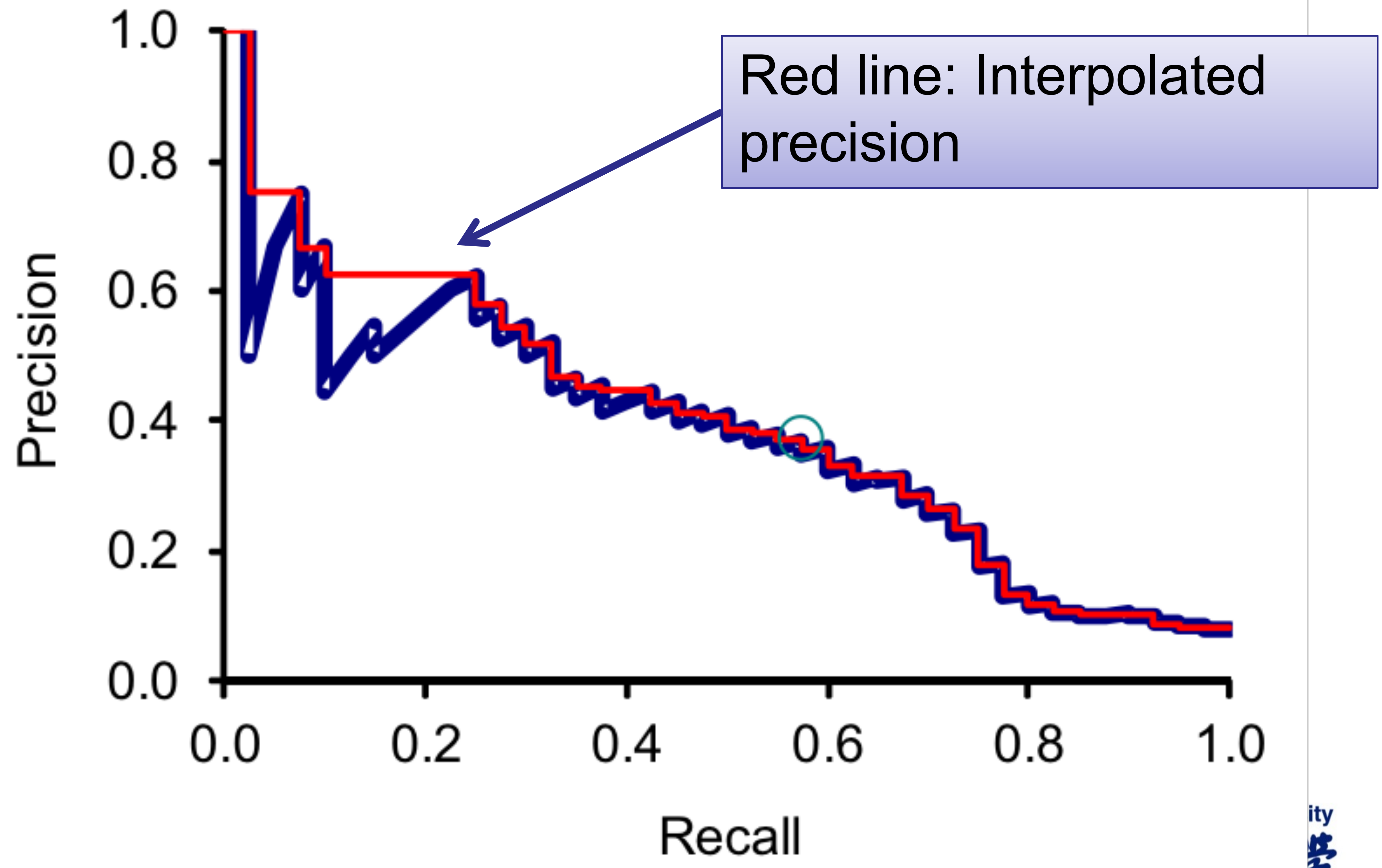
$\begin{array}{c|c} 4 & 3 \\ \hline 96 & 0 \end{array}$

$\begin{array}{c|c} 5 & 3 \\ \hline 95 & 0 \end{array}$

$\begin{array}{c|c} 5 & 4 \\ \hline 95 & \end{array}$

9		P	R	interpolated precision
1	R	1	0.01	1
2	R	1	0.02	1
3	N	2/3	0.02	
4	N	2/4	0.02	
5	R	3/5	0.03	0.625
6	N	3/6	0.03	
7	R	4/7	0.04	0.625
8	R	5/8	0.05	0.625
9	N	5/9	0.05	

A precision-recall curve



Red line: Interpolated precision

Eleven-point interpolated average precision

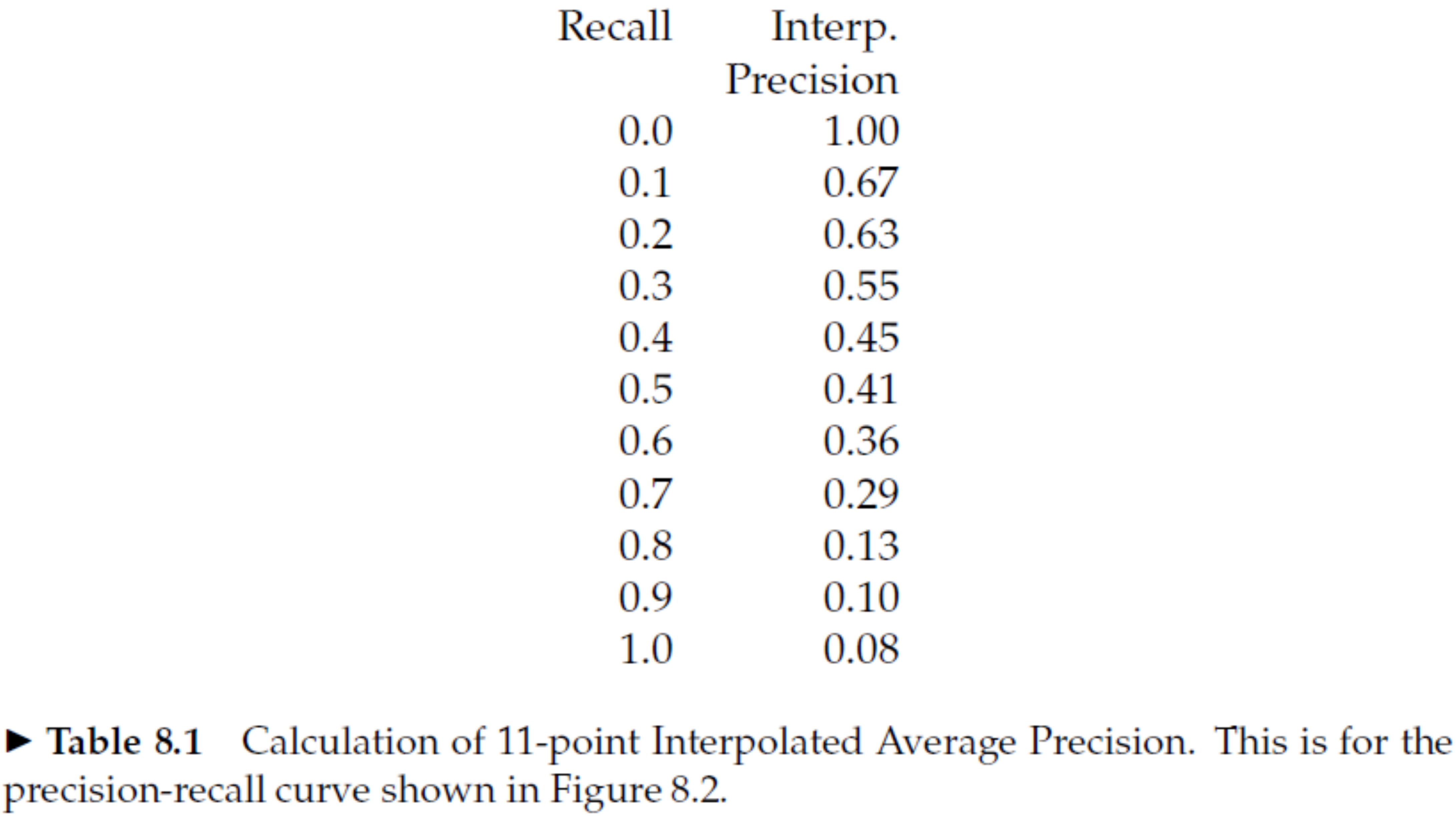
Eleven-point interpolated average precision: for each information need, the interpolated precision is measured at the 11 recall levels of

0.0, 0.1, 0.2, …, 1.0.

For each recall level, we calculate the interpolated precision at that recall level.

Do this for each information need in the test collection, then average over those information needs.(the arithmetic mean)

Eleven-point interpolated average precision

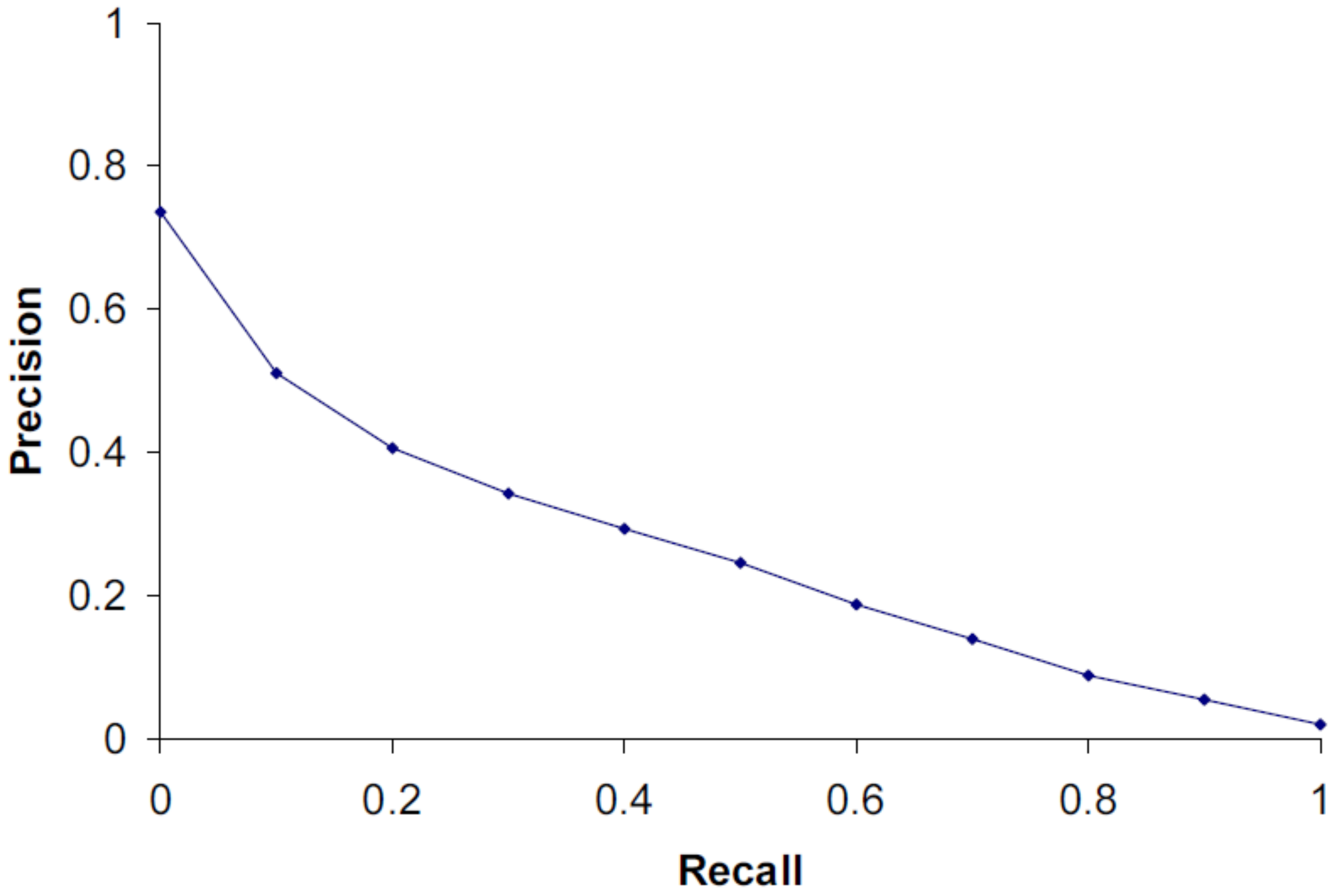


Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

▶ **Table 8.1** Calculation of 11-point Interpolated Average Precision. This is for the precision-recall curve shown in Figure 8.2.

Eleven-point interpolated average precision

A composite precision recall curve showing 11 points can then be graphed.



► **Figure 8.3** Averaged 11-point precision/recall graph across 50 queries for a representative TREC system. The Mean Average Precision for this system is 0.2553.

Eleven-point interpolated average precision

1) Compute interpolated precision at recall levels:0.0, 0.1, 0.2, …,1.0
2) Do this for each of the queries in the evaluation benchmark;
3) Average over queries;

This measures performance at all recall levels.

Exercise

The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N   N N N R N   R N N N R   N N N N R

a. What is the precision of the system on the top 20 results?

b. What is the uninterpolated precision of the system at 37.5% recall?

c. What is the interpolated precision at 37.5% recall?

Solution

a) 0.3

b) 0.33 or 0.3

c) 0.364