Evaluation in Information Retrieval

DAPI . Information Description, Storage and Retrieval Course
MIEIC, 2020/21 Edition

Sérgio Nunes
DEI, FEUP, U.Porto

Based on Chapter 8 from Introduction to Information Retrieval, Manning, C. et al. (2008)

Evaluation in Information Retrieval

Evaluation of Information Retrieval Systems

- ➔ Evaluation is at the heart of Information Retrieval.
- ➔ Evaluation is important to:
  - ➔ Understand the use of a system by its users.
  - ➔ Make decisions on new designs and features to implement.
- ➔ A primary distinction must be made between <u>effectiveness</u> and <u>efficiency</u>.
  - ➔ Effectiveness measures the ability of a search system to find the right information.
  - ➔ Efficiency measures how quickly a search system provides an answer.
- ➔ User satisfaction encapsulates these and other aspects (ux, coverage, effort, etc).

Information Retrieval System Evaluation

- ➔ To measure the effectiveness of a search system in the standard way, we need three things:
  - ➔ A document collection;
  - ➔ A test suite of information needs, expressible as queries;
  - ➔ A set of relevance judgements, standardly a binary assessment of either relevant or non-relevant for each query-document pair.
- ➔ The standard approach to IR system evaluation revolves around the notion of relevant and non-relevant documents.
- ➔ With respect to a user information need, a document in the test collection is given a binary classification as either relevant or non-relevant (gold standard or ground truth).

Information Need

- ➔ Relevance is assessed relative to an information need, not a query.
- ➔ An information need might be:
  - ➔ Information on whether drinking red wine is more effective at reducing your risk of heart attacks than drinking white wine.
- ➔ This might be translated into a query such as:
  - ➔ [wine red white heart attack effective]
- ➔ A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query. This distinction is often misunderstood in practice, because the information need is not clear.

The Cranfield Paradigm

➔ Evaluation of Information Retrieval systems is the result of early experimentation initiated in the 50's by Cyril Cleverdon.

➔ The insights derived from these experiments provide a foundation for the evaluation of IR systems.

➔ These experiments culminated in the metrics of Precision and Recall.

➔ Cyril Cleverdon introduced the notion of test reference collections, composed of documents, queries, and relevance judgements.

➔ Reference collections allows using the same set of documents and queries to evaluate different ranking systems.

Illustration of Cranfield Evaluation Methodology

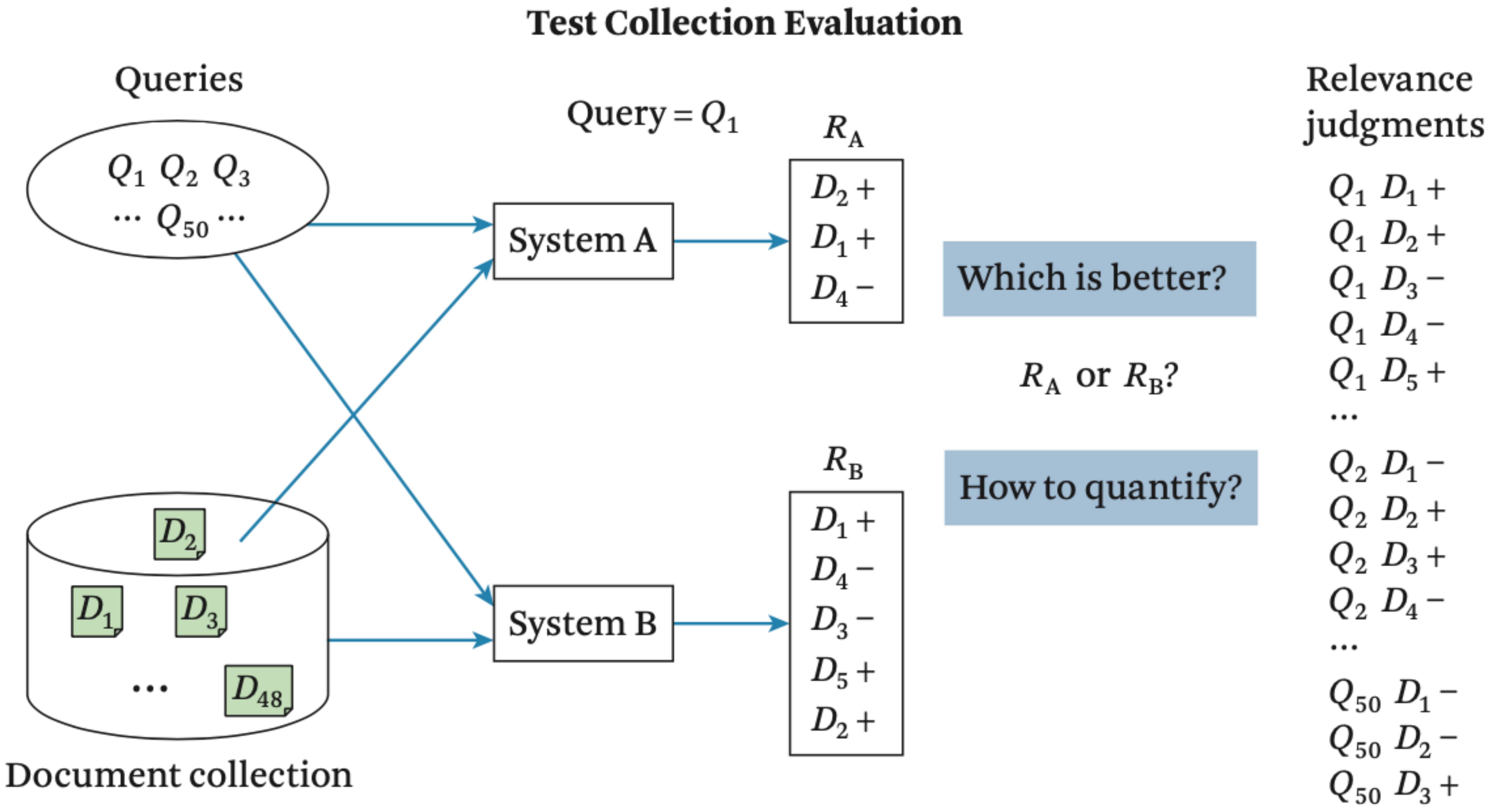


Figure 9.1 **Illustration of Cranfield evaluation methodology.**

From: Zhai, ChengXiang, Sean Massung. Text data management and analysis: a practical introduction to information retrieval and text mining. ACM and Morgan & Claypool, 2016.

TREC Topic Example

<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>

From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Example Test Collections

- ➔ CACM: Titles and abstracts from the Communications of the ACM from 1958–1979. Queries and relevance judgments generated by computer scientists.
- ➔ AP: Associated Press newswire documents from 1988–1990 (from TREC disks 1–3). Queries are the title fields from TREC topics 51–150. Topics and relevance judgments generated by government information analysts.
- ➔ GOV2: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701–850. Topics and relevance judgments generated by government analysts.

From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Example Test Collections

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 MB	64
AP	242,918	0.7 GB	474
GOV2	25,205,179	426 GB	1073

Table 8.1. Statistics for three example text collections. The average number of words per document is calculated without stemming.

From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Example Test Collections

Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

Table 8.2. Statistics for queries from example text collections

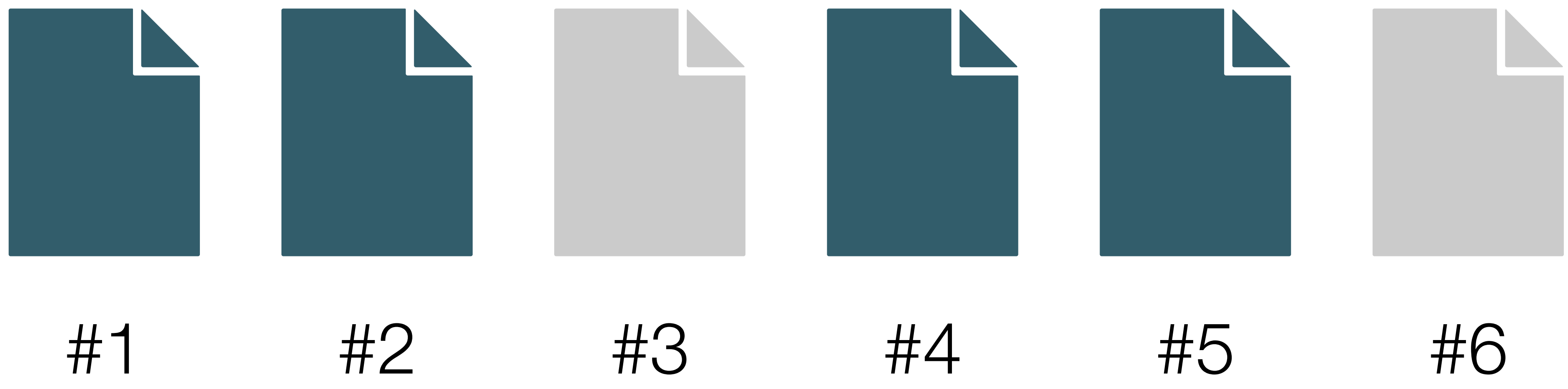
From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Evaluation of Unranked Retrieval

- ➔ The two most frequent and basic measures for information retrieval effectiveness are Precision and Recall.
- ➔ Precision is the fraction of retrieved documents that are relevant.
  - ➔ Precision (P) = #(relevant items retrieved) / #(retrieved items)
- ➔ Recall is the fraction of relevant documents that are retrieved.
  - ➔ Recall (R) = #(relevant items retrieved) / #(relevant items)
- ➔ Precision and Recall are set-based measures.

Precision and Recall

Results for query q1



Precision

P(q1) = 4 relevant documents retrieved / 6 documents retrieved = 0.67

Recall

R(q1) = 4 relevant docs retrieved / 8 existing relevant docs = 0.5

Contingency Table

	relevant	not relevant
retrieved	true positives (tp)	false positives (fp)
not retrieved	false negatives (fn)	true negatives (tn)

➔ Precision = true positives / (true positives + false positives)

➔ Recall = true positives / (true positives + false negatives)

➔ Accuracy, the fraction of classifications that are correct (not useful for IR).

- ➔ Accuracy = #(true positives + true negatives) / #(tp + fp + fn + tn)

F measure

➔ A measure that trades-off Precision versus Recall is the F measure (or F score), which is the weighted harmonic mean of precision and recall.

➔ By default a balanced harmonic mean is used (α = 1/2) resulting in a balanced F measure defined by:

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 \times P + R} \qquad F_{\beta=1} = \frac{2PR}{P + R}$$

➔ It is possible to emphasize Precision (β < 1) or Recall (β > 1).

Evaluation of Ranked Retrieval

- ➔ Precision and Recall are computed over unordered sets of documents.
- ➔ These measures need to be extended to evaluate the ranked lists of results common in search engines.
- ➔ In ranked retrieval contexts, appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each set, Precision and Recall values can be computed.
- ➔ These values can be plotted to obtain a precision-recall curve.

Precision-Recall Curves

- ➔ Consider the ordered set of relevant (R) and non-relevant (N) results from a search system A:
  - ➔ S_a = R R N R N N R N R N
- ➔ In this ranking, the first result is relevant and corresponds to 20% of all (available) relevant documents.
  - ➔ We say that we have 100% precision at 20% recall.
- ➔ At position 4, three documents out of four are relevant, and three documents of a total of five relevant document have been retrieved.
  - ➔ We say that we have 75% precision at 60% recall.
- ➔ Using this data, we can plot a precision-recall curve.

Precision-Recall Curves

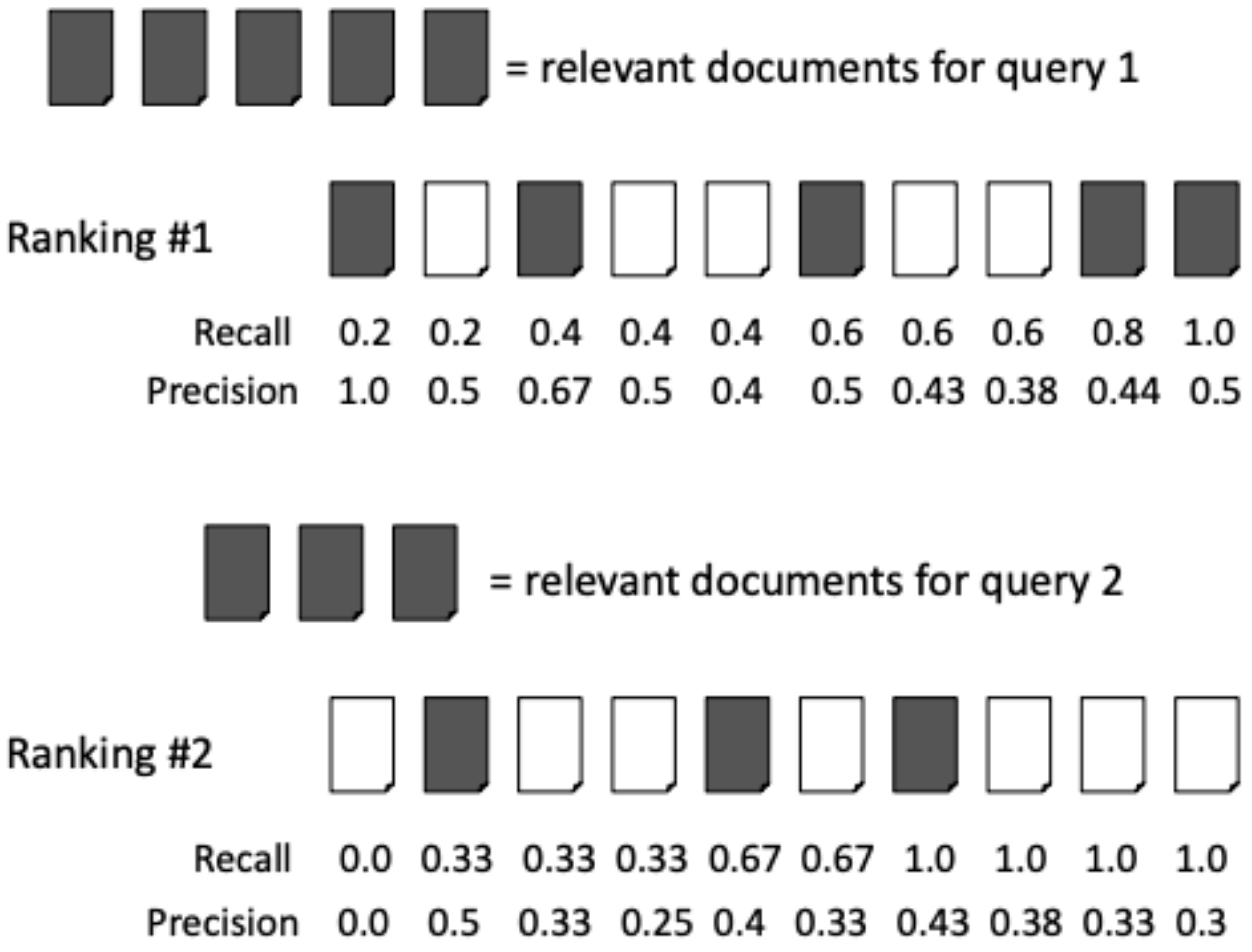


Fig. 8.3. Recall and precision values for rankings from two different queries

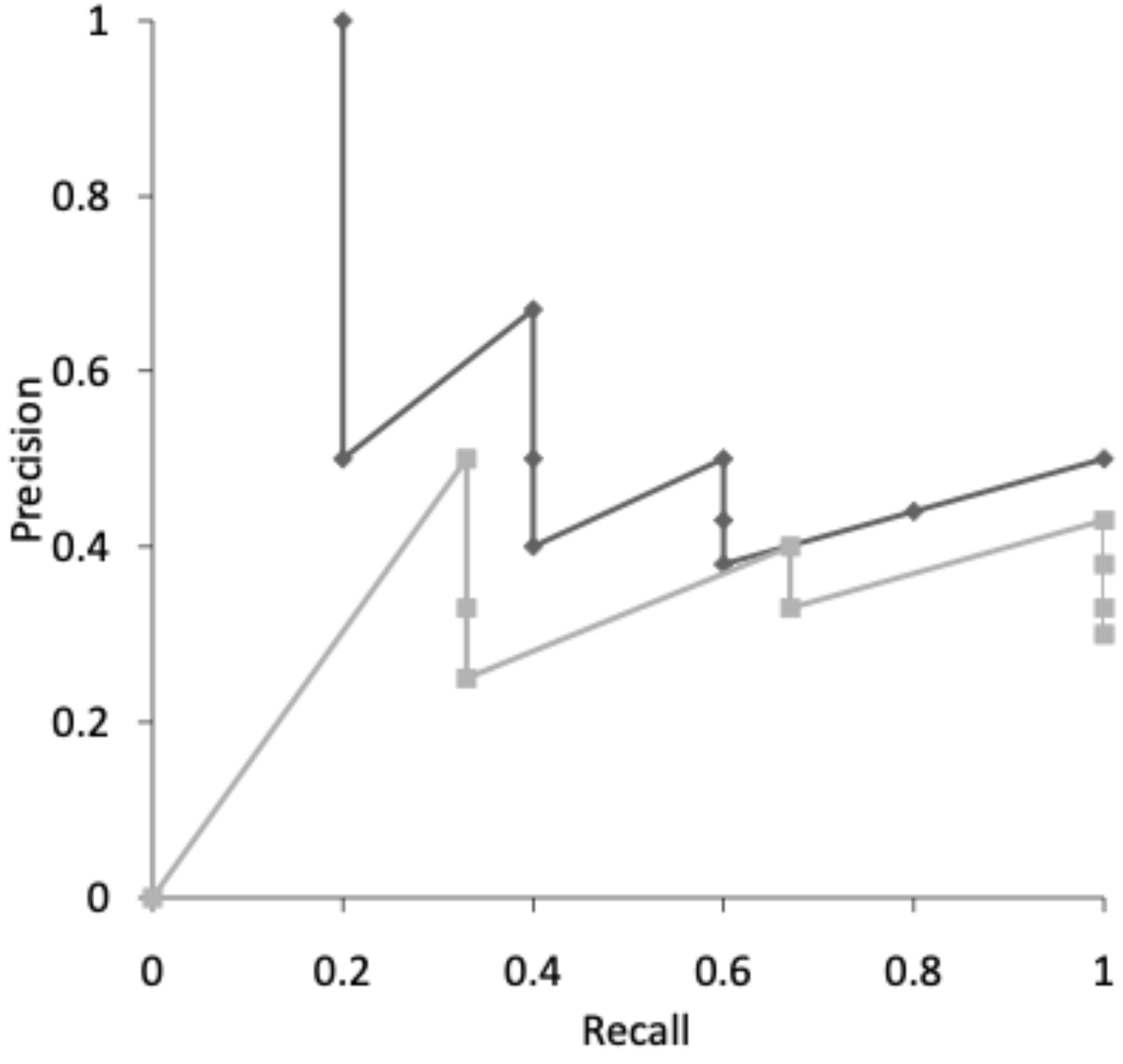


Fig. 8.4. Recall-precision graphs for two queries

From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Precision-Recall Curves (interpolated)

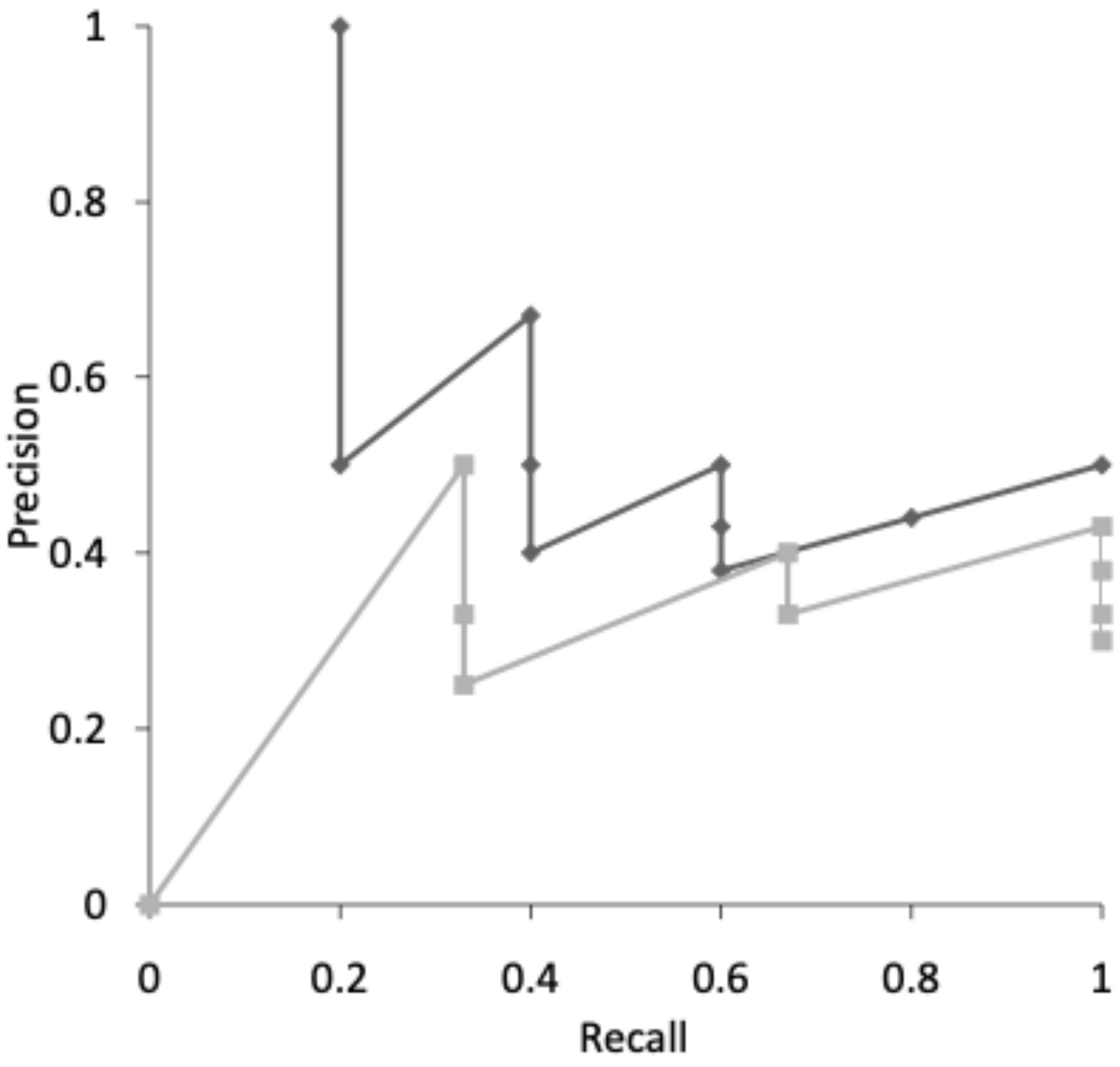


Fig. 8.4. Recall-precision graphs for two queries

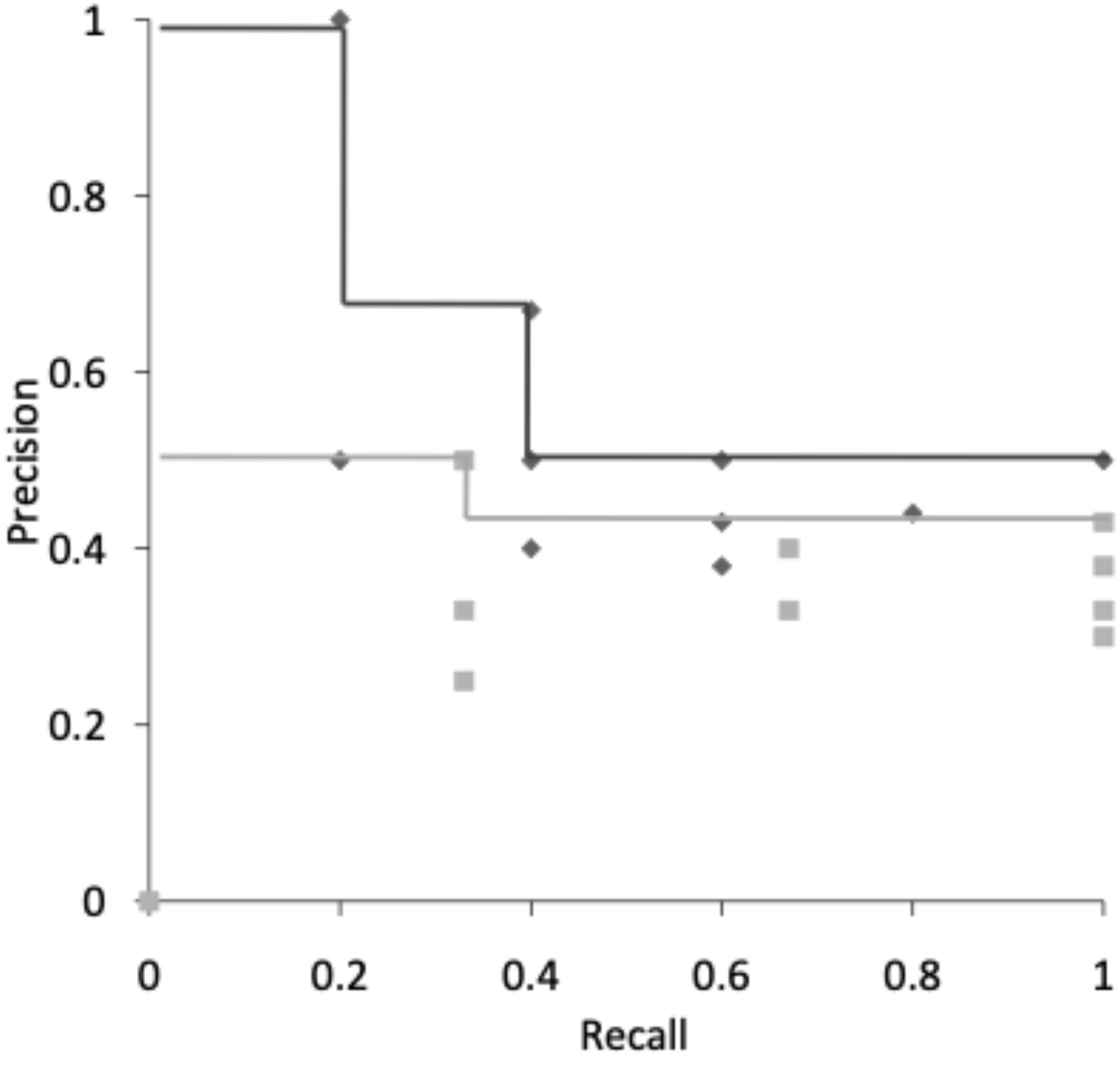


Fig. 8.5. Interpolated recall-precision graphs for two queries

From: W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines: Information Retrieval in Practice, Pearson, 2009.

Comparing Systems

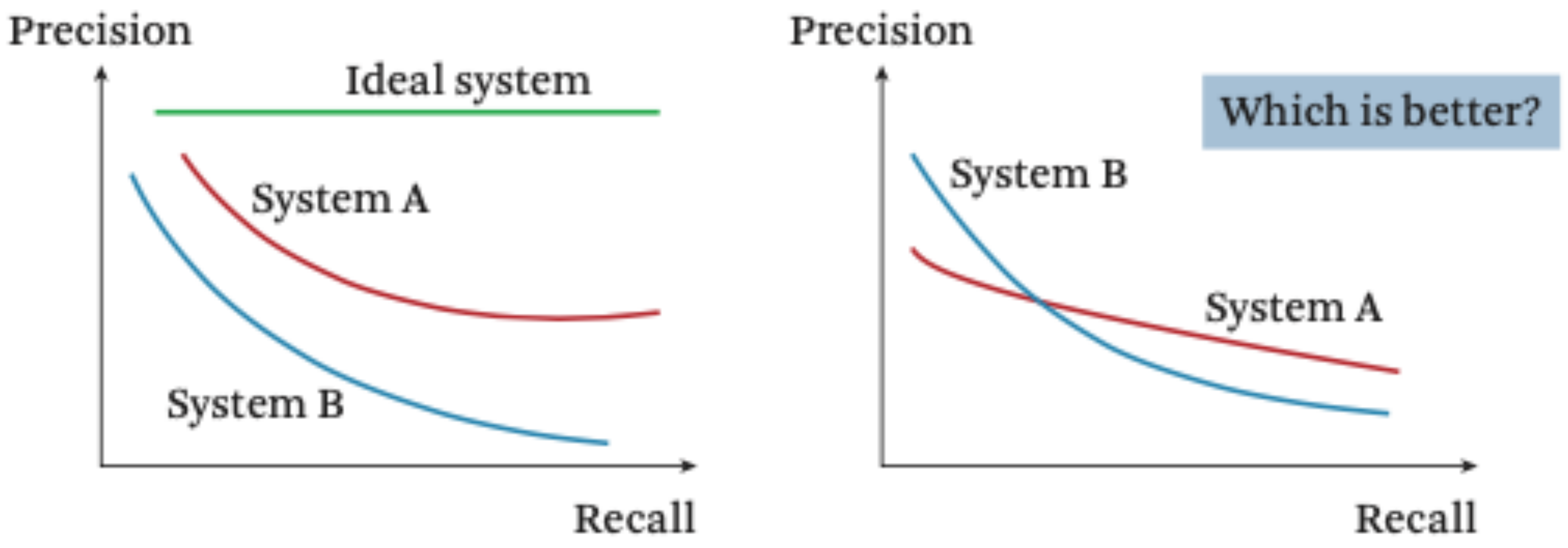


Figure 9.5 **Comparison of two PR curves.** (Courtesy of Marti Hearst)

From: Zhai, ChengXiang, Sean Massung. Text data management and analysis: a practical introduction to information retrieval and text mining. ACM and Morgan & Claypool, 2016.

Precision at k (P@k)

- ➔ In the case of web search, the majority of users do not require high recall.
- ➔ What matters are high quality results on the first page. This leads to measuring precision at fixed low levels of retrieved results.
- ➔ For example, "precision at 5" (P@5) or "precision at 10" (P@10).
- ➔ Considering the following ranking for a given query:
  - ➔ R R N N R N R R R R
- ➔ P@5 = 0.6; P@10 = 0.7

Mean Average Precision

➔ Average Precision (AvP) provides a single-figure measure of quality across recall levels for a single query.

➔ For a single information need, average precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved.

➔ Given a set of queries, the Mean Average Precision (MAP) is the mean over the AvP values. This is one of the most commonly used measures in IR.

Average Precision

➔ Ranking #1

	X		X	X	X	X				X
R	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
P	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

➔ Ranking #2

		X			X	X	X		X	X
R	0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
P	0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

➔ AvP (R#1) = (1 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78

➔ AvP (R#2) = (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.52

➔ MAP = (0.78 + 0.52) / 2 = 0.65

Measuring Efficiency

Example Efficiency Metrics

Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

Table 8.5. Definitions of some important efficiency metrics