Evaluation in Information Retrieval

PRI 22/23 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

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

PRI Status

- ➔ Week Plan
  - ➔ Lecture: Information Retrieval Evaluation
  - ➔ Labs: Solr tutorial; Collection indexing;
- ➔ Start of Milestone #2
  - ➔ Define documents
  - ➔ Index documents
  - ➔ Retrieval testing

Outline for Today

- ➔ Information Retrieval Evaluation
  - ➔ Review of main IR concepts
  - ➔ Test collections
  - ➔ Offline evaluation metrics
  - ➔ Online evaluation strategies
- ➔ Solr tutorial

Notes on PRI Project

Milestone #1

Notes on Milestone #1 Reports

- ➔ Scientific writing is expected
  - ➔ Scientific writing is technical writing for communication in science
  - ➔ Factual, concise, evidence-based
  - ➔ Avoid creative, verbose prose
  - ➔ Attention to expressions with rigorous meanings in CS, Math or Engineering, e.g. "significant" (statistics), "exponential growth" (math, CS).
  - ➔ Do not use "decorative" figures
  - ➔ Abstracts are not an introduction to the document, they should act as 'mini-documents' with all the main elements: why, what, how, results.

Notes on Milestone #1 Reports

- ➔ Other aspects
  - ➔ Missing information needs — necessary for milestone 2 evaluation
  - ➔ Missing analysis of textual fields — important to understand the data

- ➔ Follow-up deliveries build upon the previous
  - ➔ correct the main problems reported in the evaluation.
  - ➔ use \parts (Latex) for each milestone;

Milestone #2

Milestone #2

- ➔ The second milestone is achieved with the implementation and use of an information retrieval tool on the project datasets and its exploration with free-text queries.
  - ➔ choose the information retrieval tool (Solr, Elasticsearch, Lucene, Terrier, …);
  - ➔ analyze the documents and identify their indexable components;
  - ➔ use the selected tool to build the indexes;
  - ➔ use the selected tool to configure and execute the queries;
  - ➔ demonstrate the indexing and retrieval processes;
  - ➔ manually evaluate the returned results;
  - ➔ evaluate the results obtained for the defined information needs.

Milestone #3

Milestone #3

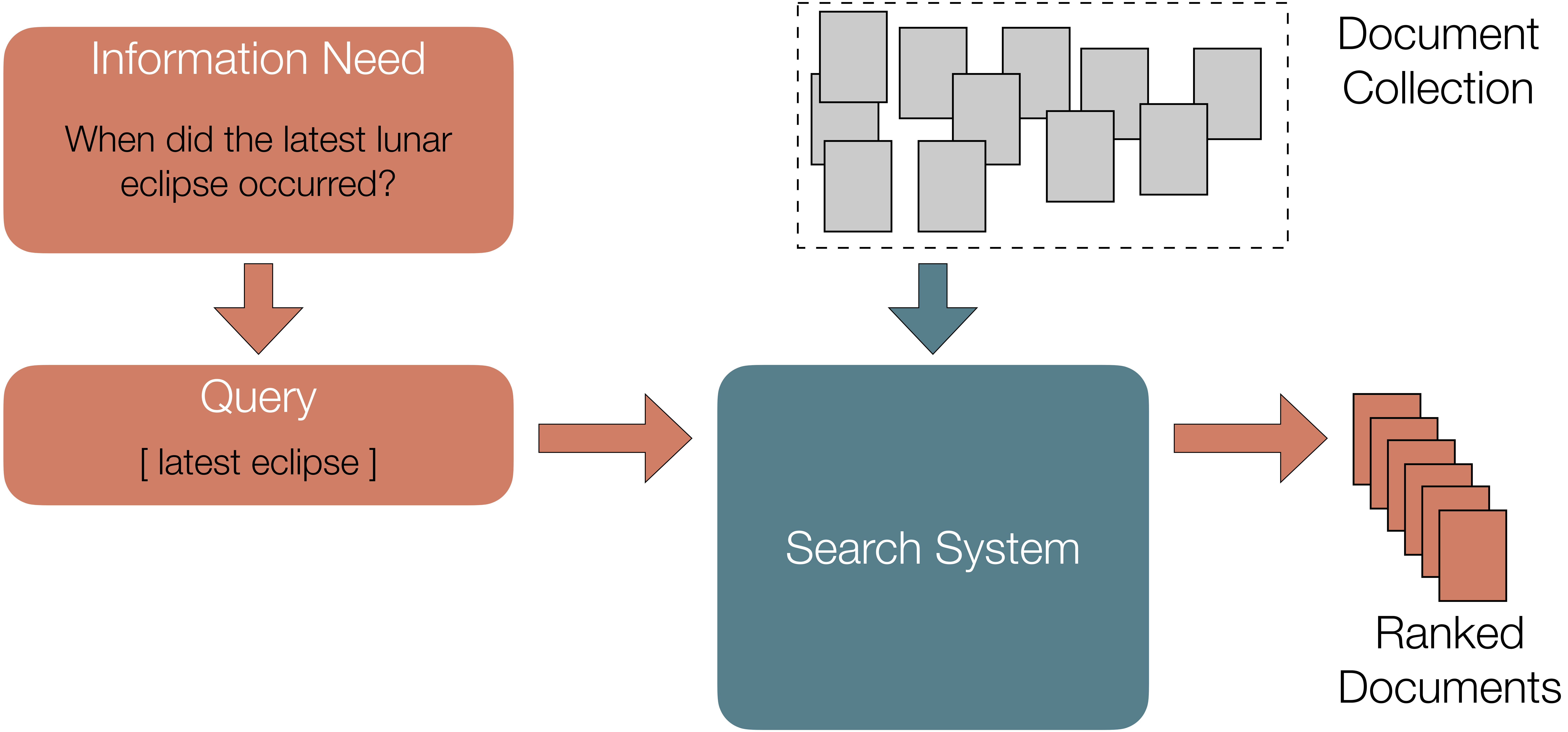
- ➔ The 3rd milestone is achieved with the development of the final version of the search system.
- ➔ This version is an **improvement over the previous milestone**, making use of features and techniques with the goal of improving the quality of the search results.
- ➔ For this milestone, each group is expected to explore innovative approaches and ideas, and will heavily depend on the context and data of each group.
- ➔ An **extended evaluation of the results** and **a comparison with the previous version** of the search system is expected.
- ➔ As examples of topics to explore, groups **may**:
  - ➔ Incorporate new information retrieval algorithms;
  - ➔ Expand the information available for each document by adding and linking new datasets;
  - ➔ Work on user interfaces by developing a frontend for the search system (**not sufficient**);

Information Retrieval Recap

Information Retrieval Recap

- ➔ Information Retrieval dates back to 1940s.
- ➔ Well established field in Computer Science.
- ➔ Deals with a wide range of information access problems.
- ➔ Organized in tasks: web search, filtering, question answering, podcast search, conversational, enterprise / intranet search, entity-oriented search
- ➔ Strong links with information science, NLP, machine learning, databases, etc.
- ➔ Core concepts: information need, index, query, relevance.

Classic Information Retrieval



Architecture of an Information Retrieval System

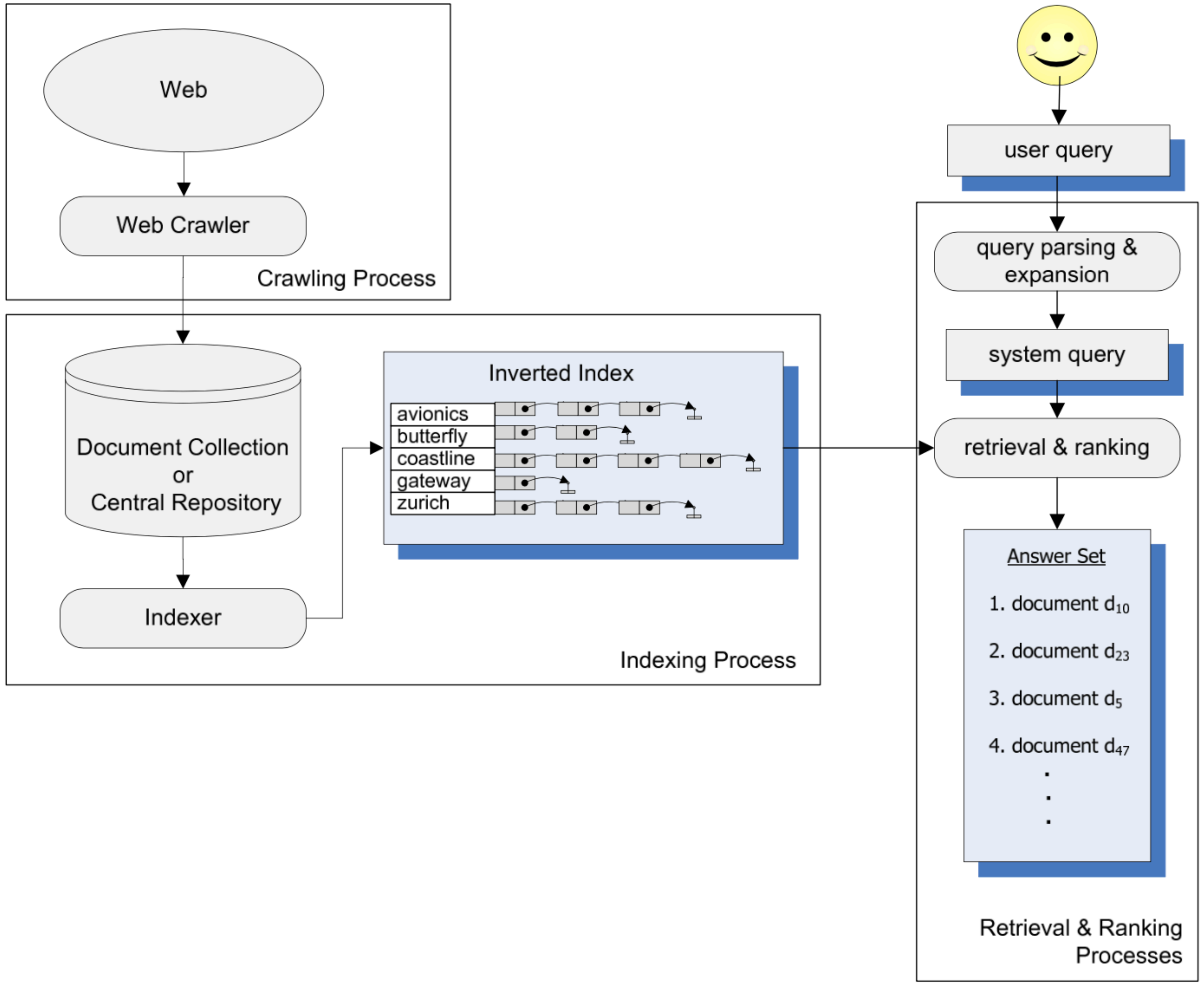
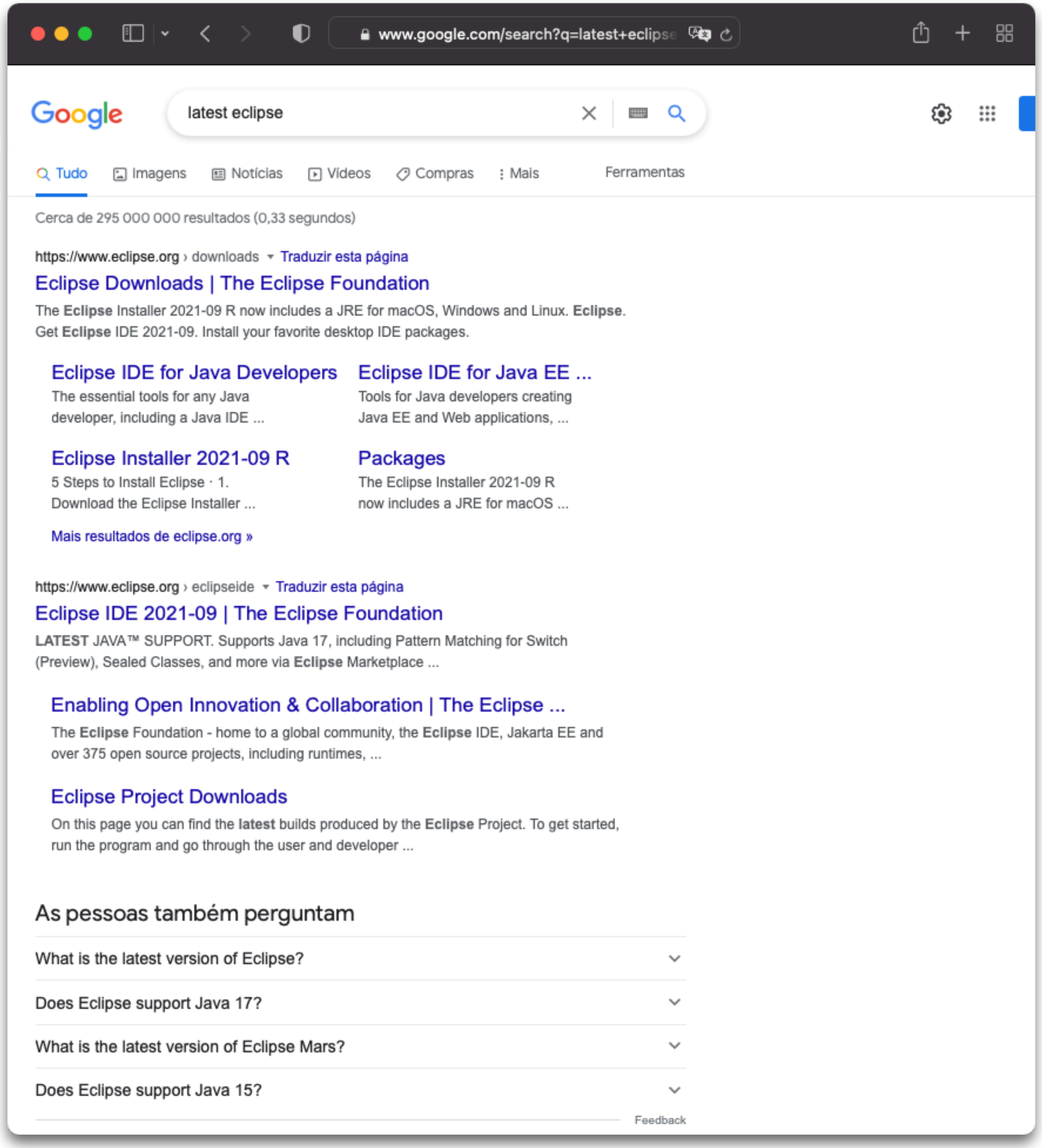
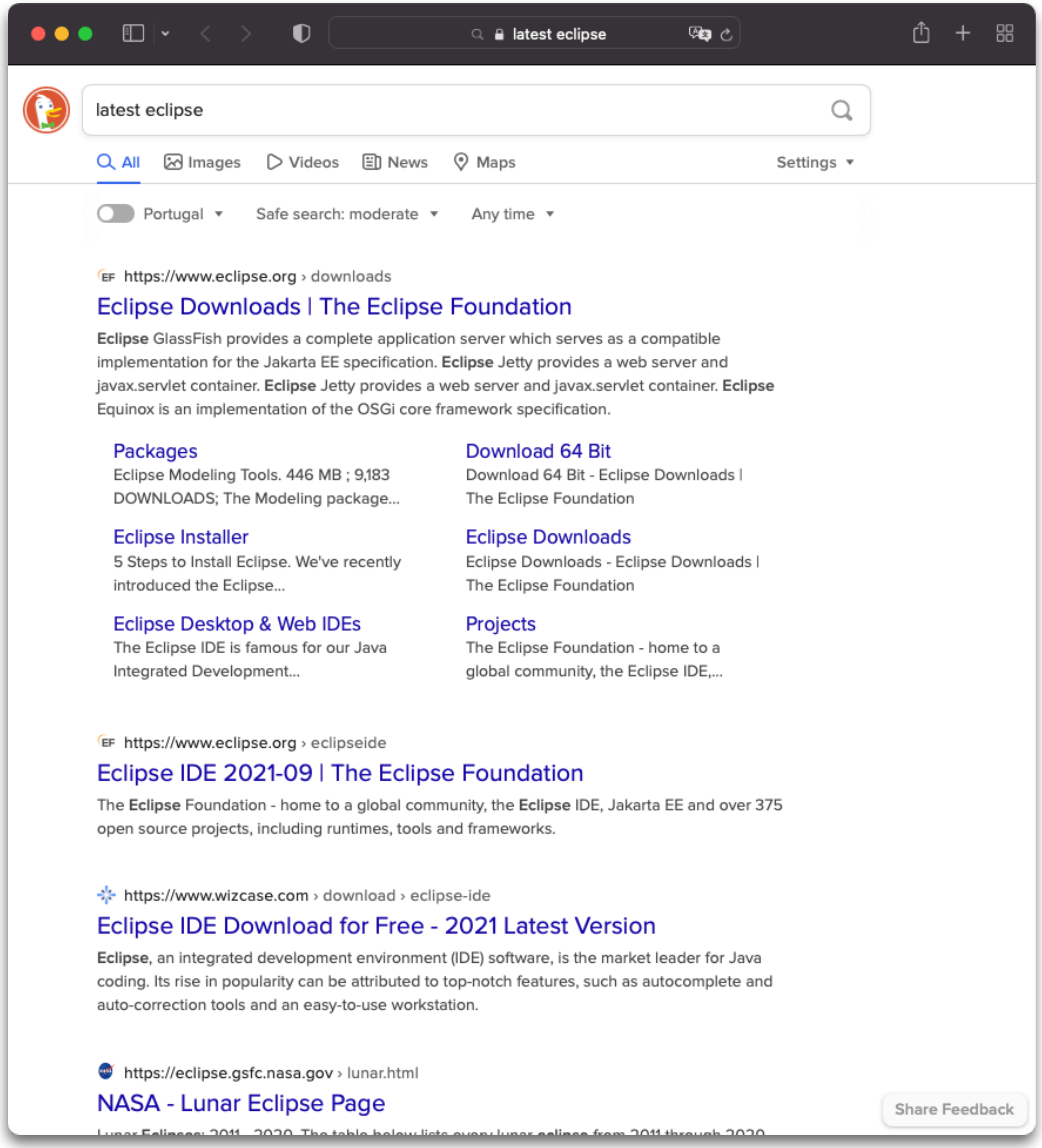


Image from Modern Information Retrieval, Baeza-Yates et al. (2011)

Evaluation in Information Retrieval


www.google.com/search?q=latest+eclipse
Google
latest eclipse
Tudo
Imagens
Notícias
Vídeos
Compras
Mais
Ferramentas
Cerca de 295 000 000 resultados (0,33 segundos)
https://www.eclipse.org › downloads
Traduzir esta página
Eclipse Downloads | The Eclipse Foundation
The Eclipse Installer 2021-09 R now includes a JRE for macOS, Windows and Linux. Eclipse. Get Eclipse IDE 2021-09. Install your favorite desktop IDE packages.
Eclipse IDE for Java Developers
The essential tools for any Java developer, including a Java IDE ...
Eclipse IDE for Java EE ...
Tools for Java developers creating Java EE and Web applications, ...
Eclipse Installer 2021-09 R
5 Steps to Install Eclipse · 1. Download the Eclipse Installer ...
Packages
The Eclipse Installer 2021-09 R now includes a JRE for macOS ...
Mais resultados de eclipse.org »
https://www.eclipse.org › eclipseide
Traduzir esta página
Eclipse IDE 2021-09 | The Eclipse Foundation
LATEST JAVA™ SUPPORT. Supports Java 17, including Pattern Matching for Switch (Preview), Sealed Classes, and more via Eclipse Marketplace ...
Enabling Open Innovation & Collaboration | The Eclipse ...
The Eclipse Foundation - home to a global community, the Eclipse IDE, Jakarta EE and over 375 open source projects, including runtimes, ...
Eclipse Project Downloads
On this page you can find the latest builds produced by the Eclipse Project. To get started, run the program and go through the user and developer ...
As pessoas também perguntam
What is the latest version of Eclipse?
Does Eclipse support Java 17?
What is the latest version of Eclipse Mars?
Does Eclipse support Java 15?
Feedback


latest eclipse
latest eclipse
All
Images
Videos
News
Maps
Settings
Portugal
Safe search: moderate
Any time
https://www.eclipse.org › downloads
Eclipse Downloads | The Eclipse Foundation
Eclipse GlassFish provides a complete application server which serves as a compatible implementation for the Jakarta EE specification. Eclipse Jetty provides a web server and javax.servlet container. Eclipse Jetty provides a web server and javax.servlet container. Eclipse Equinox is an implementation of the OSGi core framework specification.
Packages
Eclipse Modeling Tools. 446 MB ; 9,183 DOWNLOADS; The Modeling package...
Download 64 Bit
Download 64 Bit - Eclipse Downloads | The Eclipse Foundation
Eclipse Installer
5 Steps to Install Eclipse. We've recently introduced the Eclipse...
Eclipse Downloads
Eclipse Downloads - Eclipse Downloads | The Eclipse Foundation
Eclipse Desktop & Web IDEs
The Eclipse IDE is famous for our Java Integrated Development...
Projects
The Eclipse Foundation - home to a global community, the Eclipse IDE,...
https://www.eclipse.org › eclipseide
Eclipse IDE 2021-09 | The Eclipse Foundation
The Eclipse Foundation - home to a global community, the Eclipse IDE, Jakarta EE and over 375 open source projects, including runtimes, tools and frameworks.
https://www.wizcase.com › download › eclipse-ide
Eclipse IDE Download for Free - 2021 Latest Version
Eclipse, an integrated development environment (IDE) software, is the market leader for Java coding. Its rise in popularity can be attributed to top-notch features, such as autocomplete and auto-correction tools and an easy-to-use workstation.
https://eclipse.gsfc.nasa.gov › lunar.html
NASA - Lunar Eclipse Page
Share Feedback

Evaluation of Information Retrieval Systems

- ➔ Evaluation depends on the task, collection, information need type.
- ➔ Collections: news articles, web pages, scientific articles, etc.
- ➔ Different information needs:
  - ➔ What is the homepage of the Dept. of Informatics Engineering at FEUP?
  - ➔ Who are the teachers of information systems courses at FEUP?
  - ➔ Open source information retrieval frameworks?
- ➔ Informational vs. navigational; Number of relevant documents; Answers needed.

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 effectiveness (*eficácia*) and efficiency (*eficiência*).
  - ➔ 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, typically 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

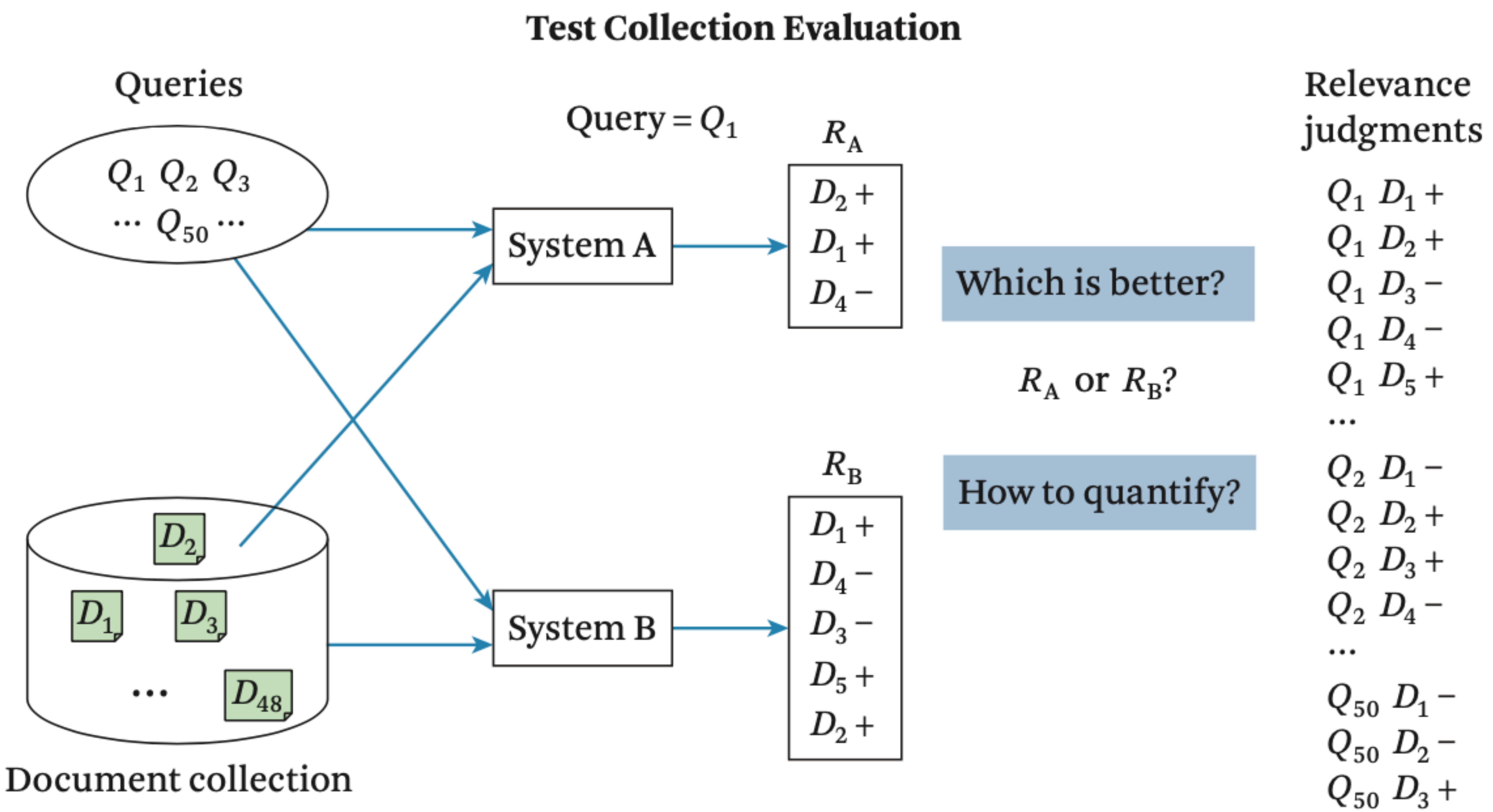


Figure 9.1 **Illustration of Cranfield evaluation methodology.**

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 produced 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 produced 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 produced 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.

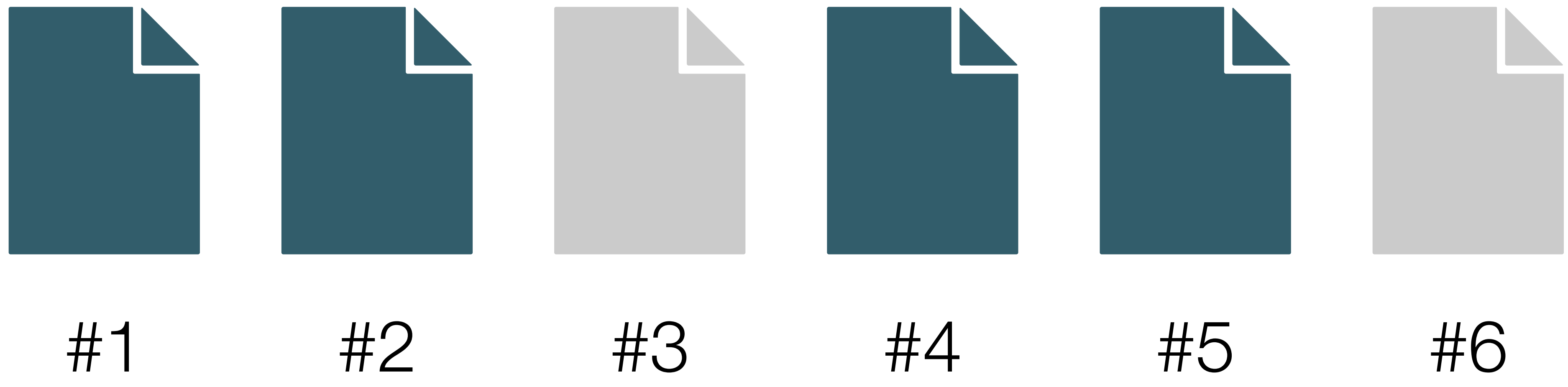
Evaluation of Unranked Retrieval

- 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

Not retrieved for q1

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 ($\beta < 1$) or Recall ($\beta > 1$).

- Often criticized because it is opaque to interpretation.

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

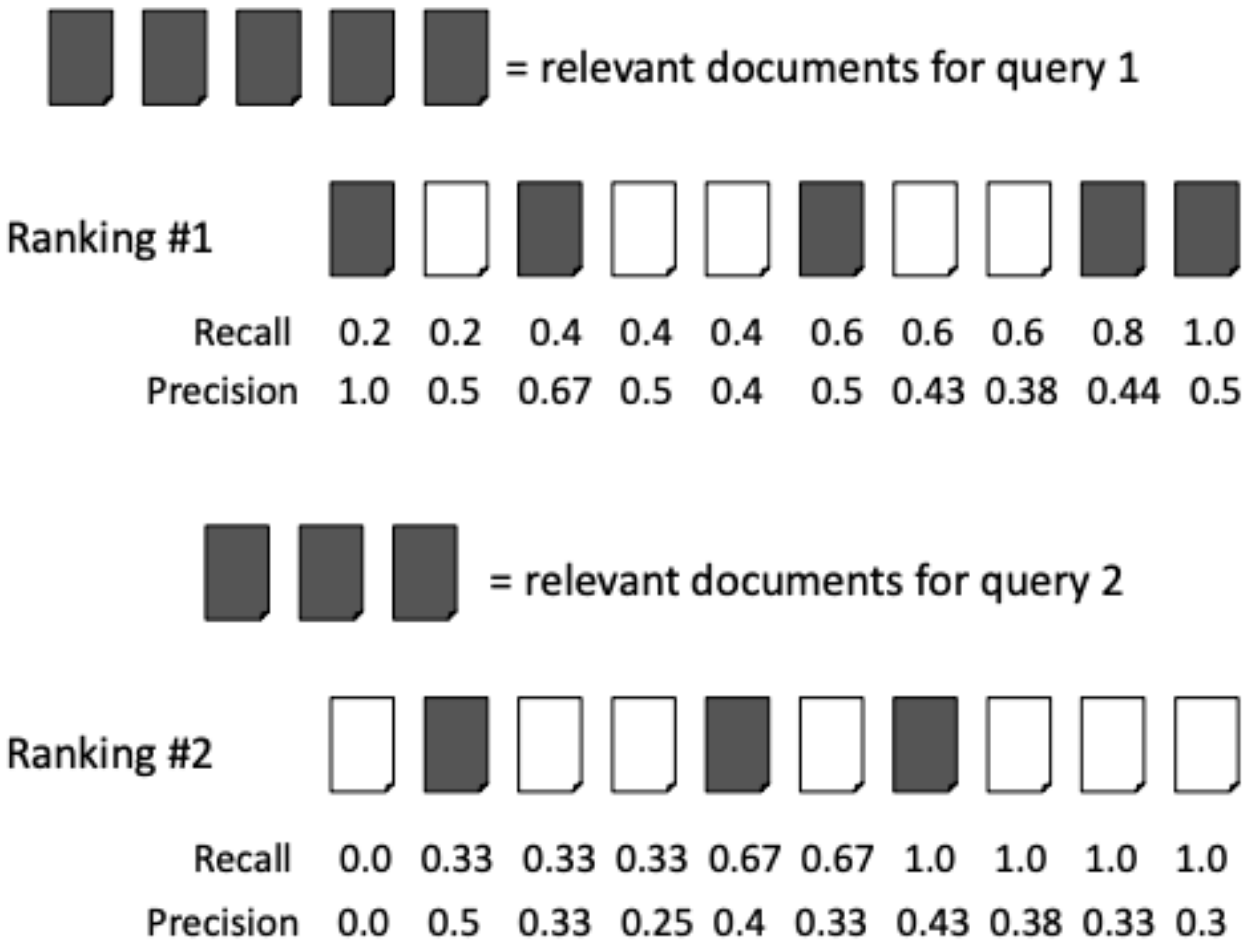


Fig. 8.3. Recall and precision values for rankings from two different queries

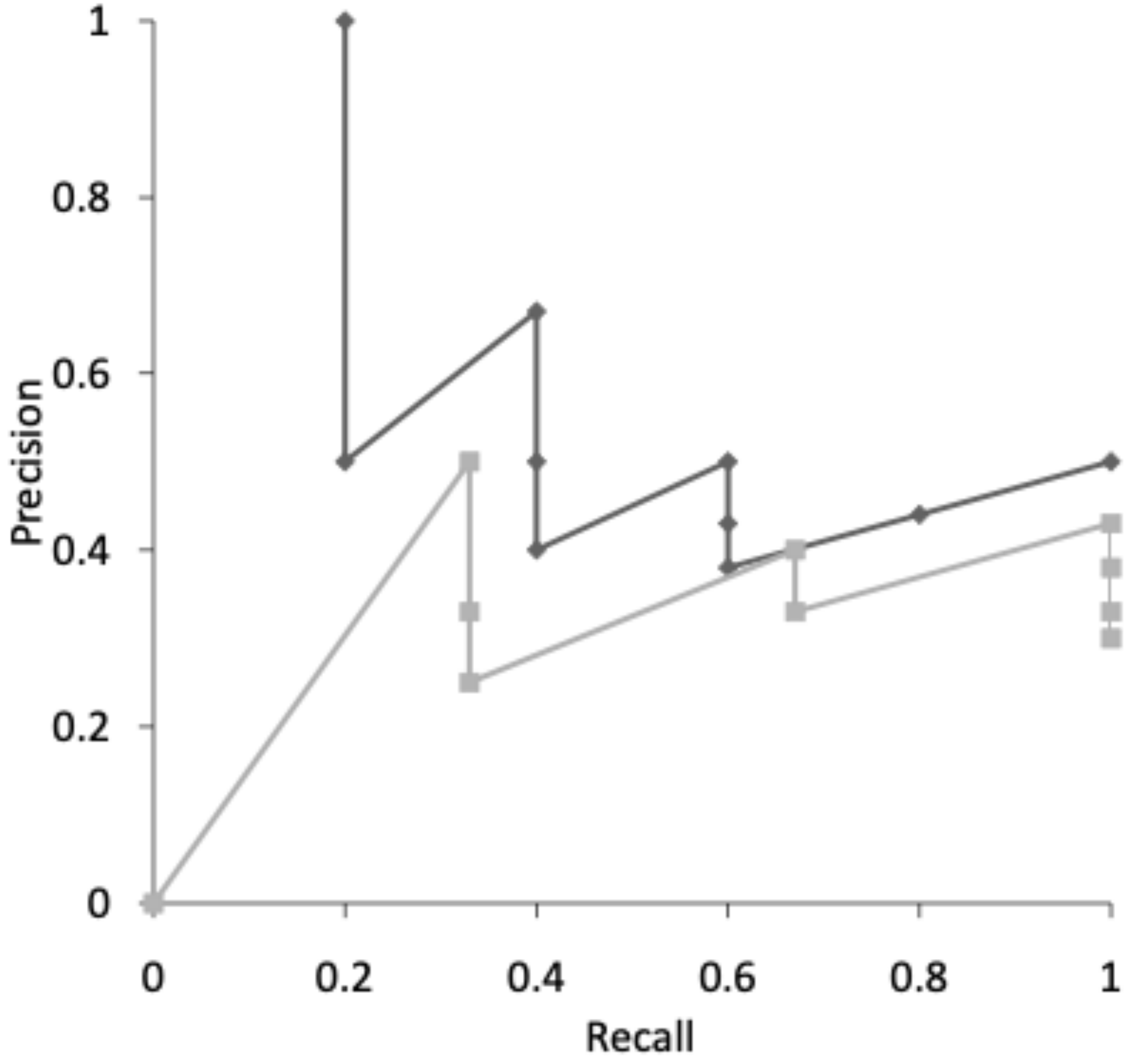


Fig. 8.4. Recall-precision graphs for two queries

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

Precision-Recall Curves (interpolated)

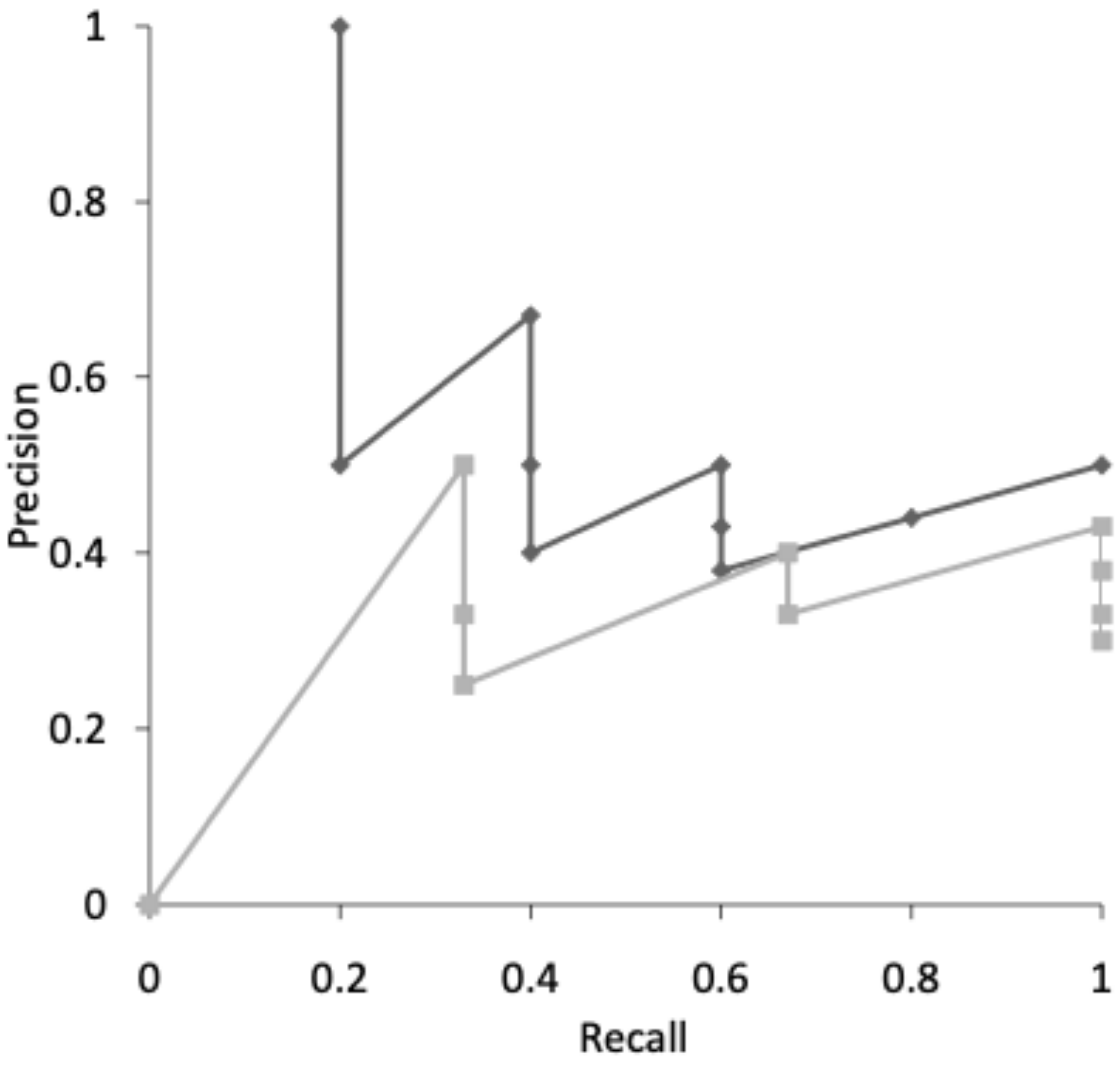


Fig. 8.4. Recall-precision graphs for two queries

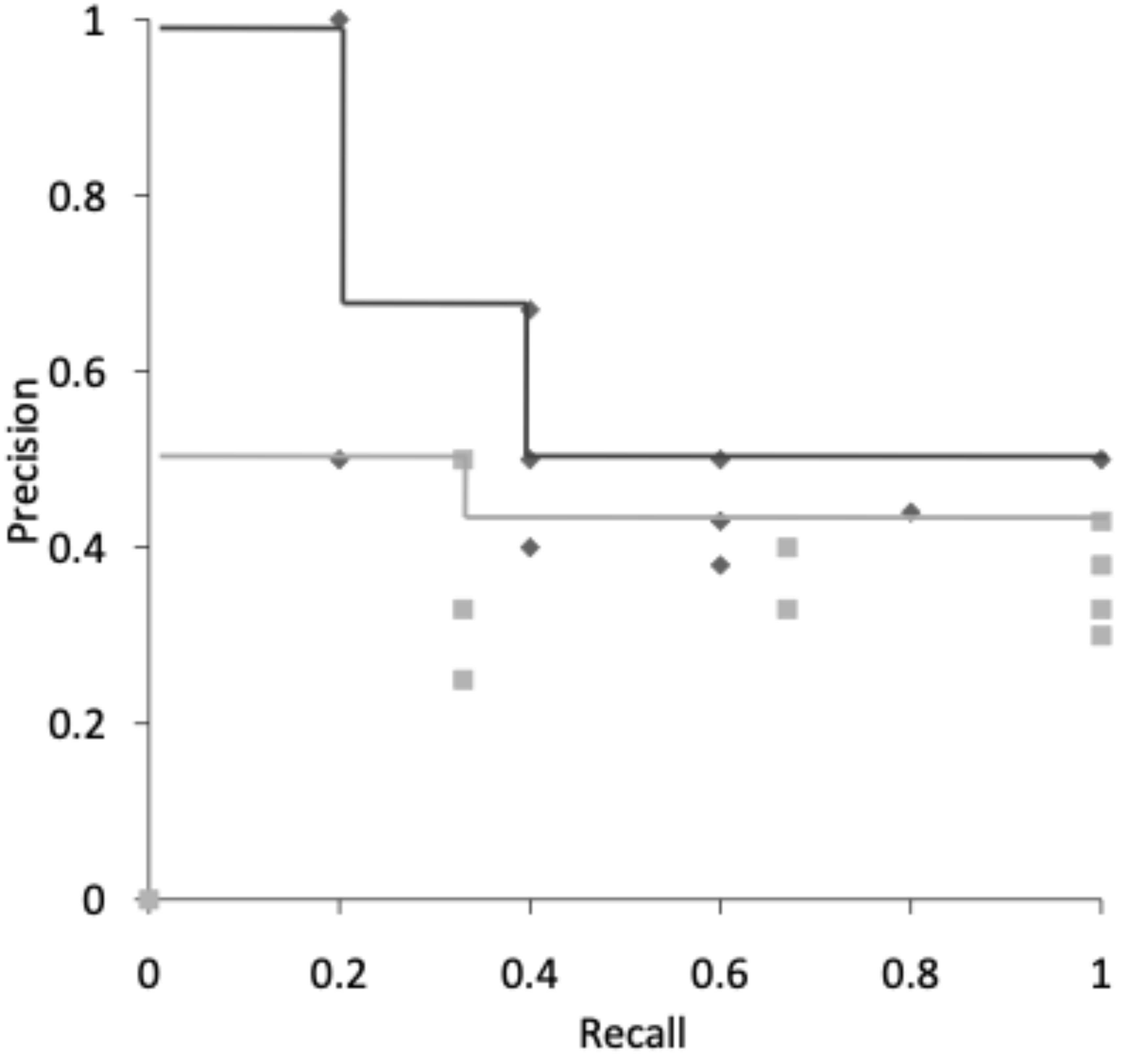


Fig. 8.5. Interpolated recall-precision graphs for two queries

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

Comparing Systems

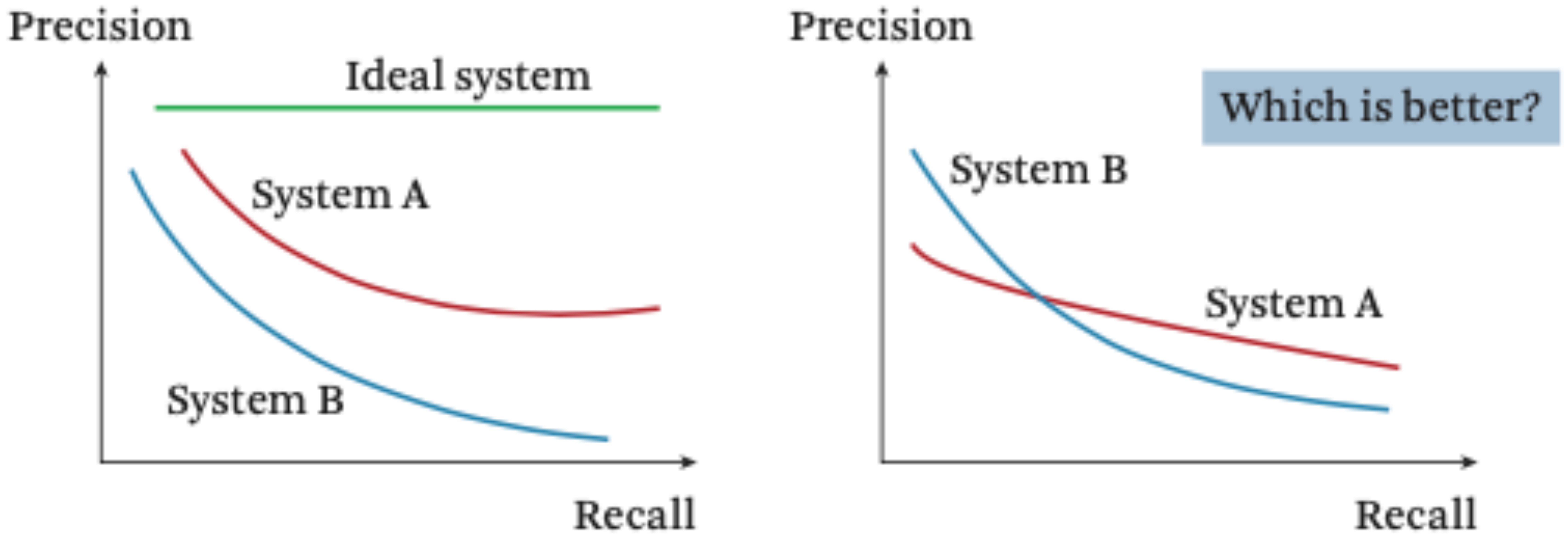


Figure 9.5 Comparison of two PR curves. (Courtesy of Marti Hearst)

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

User-Centered and Online Evaluation

- Instead of using standard collection, systems may be evaluated using real user interaction data. This requires a working system or prototype with a user base.
- In user-centered evaluation, different techniques may be used
  - Observational studies, observe the user in controlled sessions;
  - Observe server-side query logs, to measure performance;
  - Use client-side logging tools, to study interaction patterns;
- Online, large-scale evaluation experiments can be made with A/B tests. Common practice in industry - run multiple experiments with different versions of the system.

Measuring Efficiency

Example Efficiency Metrics

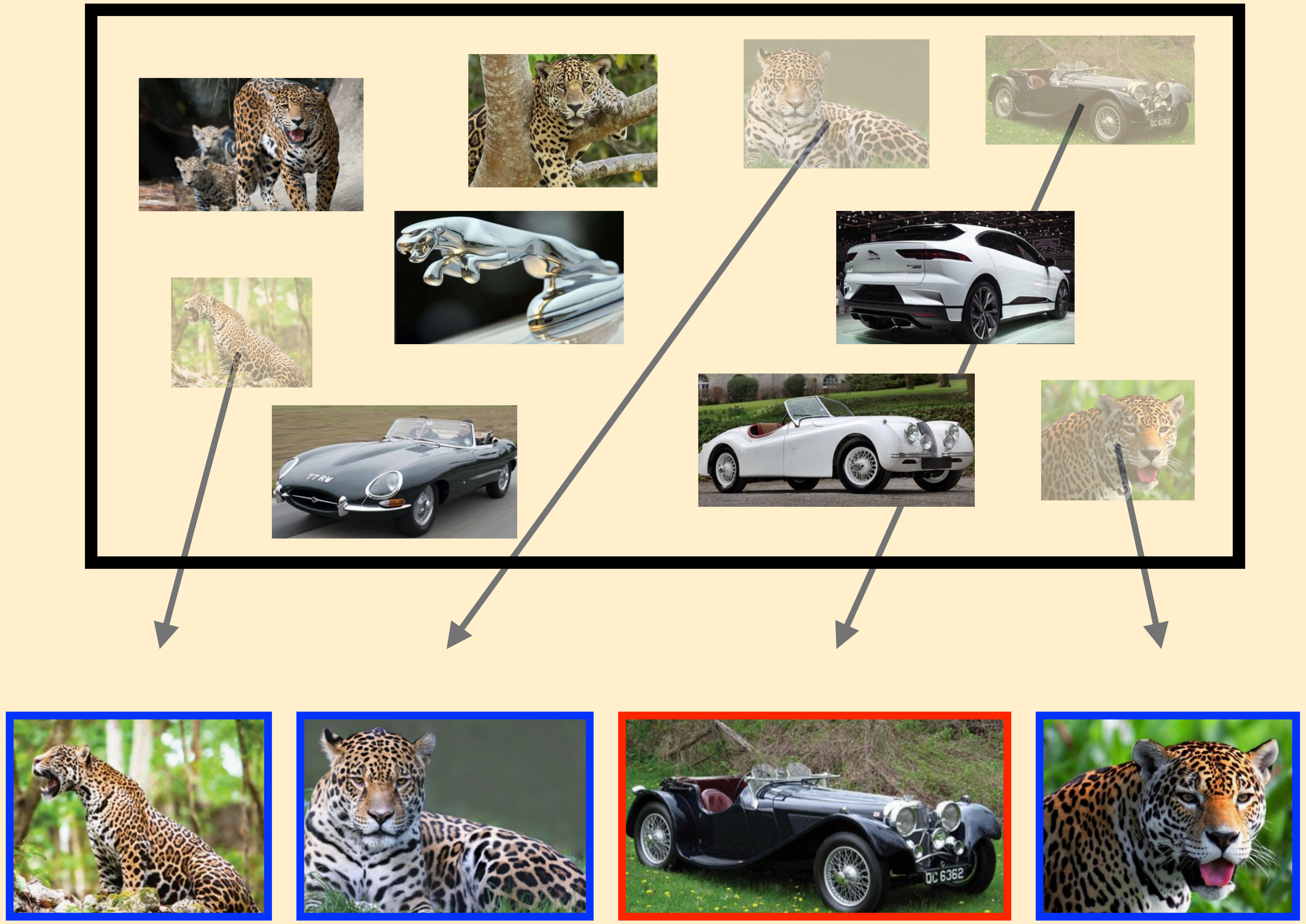
Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains 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

Exercises

Calculate Precision and Recall

➔ Consider the following scenario and a search for [jaguar] (the animal).
What is the precision and recall of this system?



Retrieved documents

	relevant	not
retrieved	3	1
not	2	4

Calculate Precision and Recall

	relevant	not
retrieved	3	1
not	2	4

➔ Precision = #(relevant items retrieved) / #(retrieved items) = 3 / 4 = 0.75

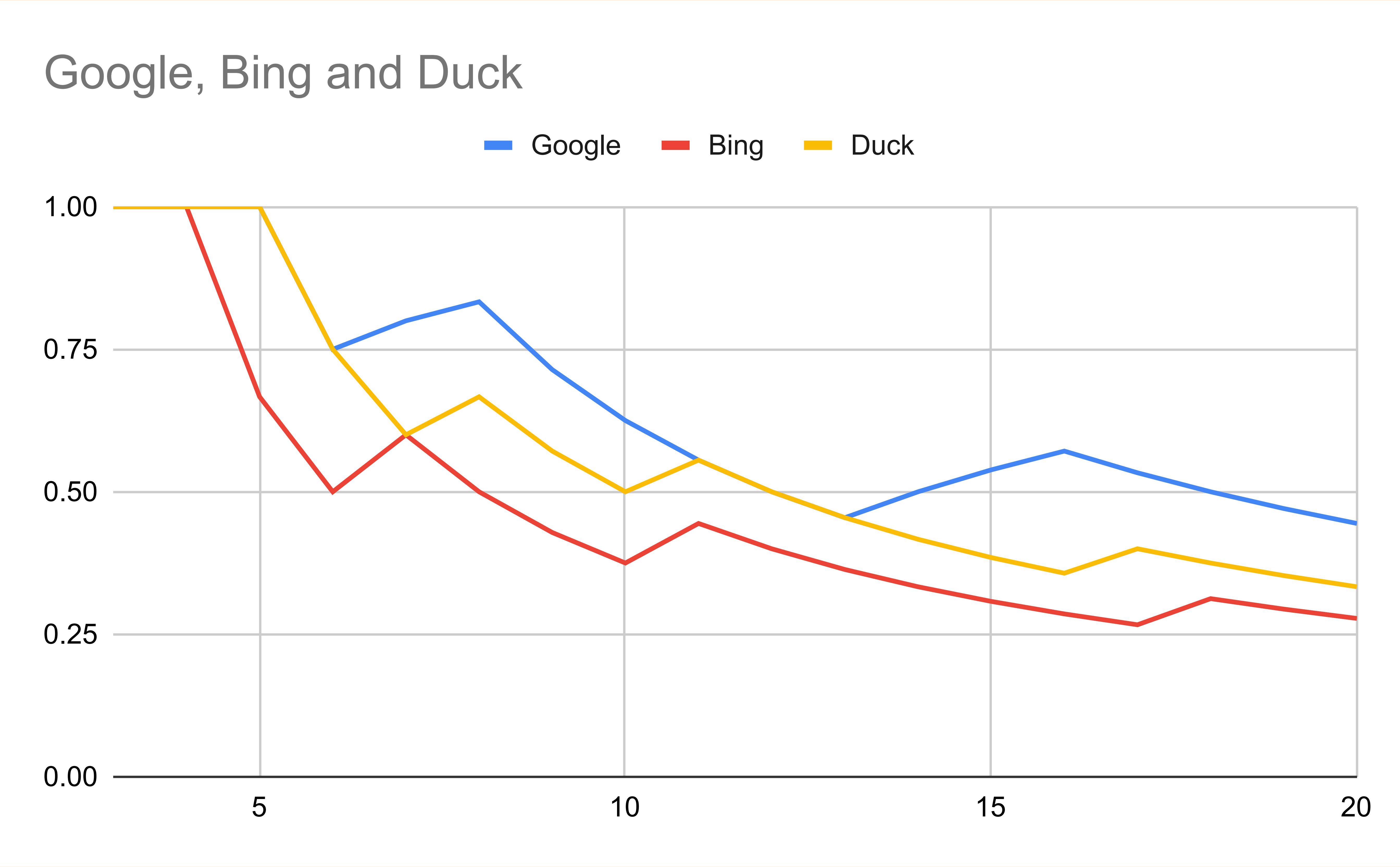
➔ Recall = #(relevant items retrieved) / #(relevant items) = 3 / 5 = 0.6

Calculate Precision and Recall

- ➔ Compare Google, Bing, DuckDuckGo with the following queries:
  - ➔ [eclipse]
  - ➔ [eclipse -lunar]
- ➔ For these two queries, for each search engine,
  and considering "eclipse the editor" the underlying information need:
  - ➔ Plot P@ values (x-axis) for each search engine for the first 20 results returned. Use a spreadsheet.
  - ➔ Calculate AvP for each query for each search engine.
  - ➔ Calculate MAP values for each search engine.

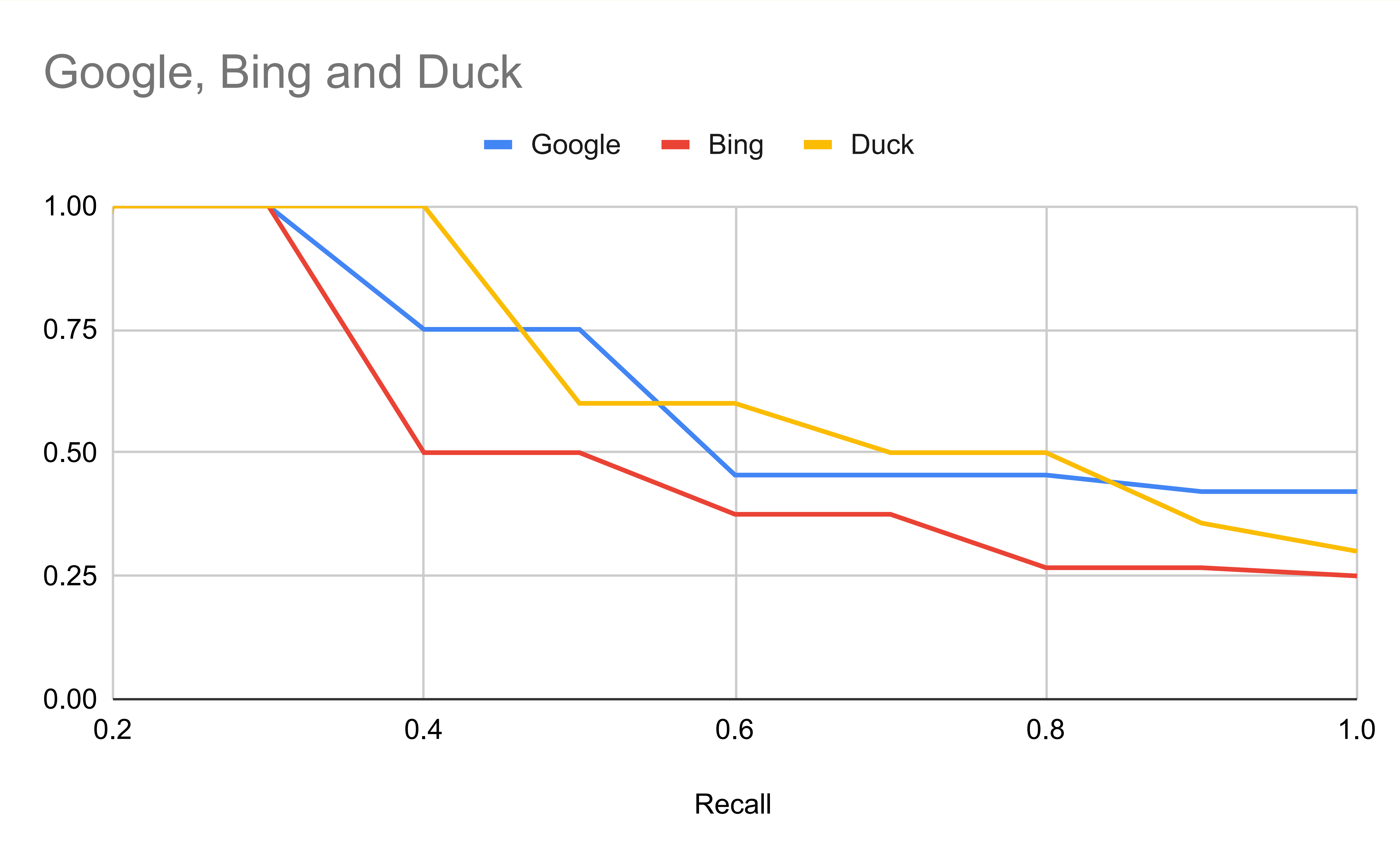
P@ Plot

➔ Q = [eclipse]



Precision-Recall Graph

➔ Q = [eclipse]



Average Precision Values

➔ Q = [eclipse]

Rank	Google		Bing		Duck	
1	x	1/1=1	x	1	x	1
2	x	2/2=1			x	1
3					x	1
4			x	2/4=0.5		
5			x	3/5=0.6		
6					x	4/6=0.67
7	x	3/7=0.43				
8						
9			x	4/9=0.44	x	5/9=0.56
10						
AP		0.81		0.635		0.846

Next Steps

- ➔ Work on Milestone #2
  - ➔ Define information needs
  - ➔ Index documents
  - ➔ Retrieve documents — explore different rankings