Introduction to Information Retrieval (36662)

Dr. David Elsweiler
Chair for Information Science
Faculty of Language, Literature and Cultural Sciences
David.Elsweiler@sprachlit.uni-regensburg.de

Universität Regensburg

Outline

① Intro and Motivation

② A Framework for Evaluation

③ Metrics

④ Building Test Collections

⑤ Discussion

Outline

① Intro and Motivation

② A Framework for Evaluation

③ Metrics

④ Building Test Collections

⑤ Discussion

„My Search Engine is better than yours!“

- These kind of statements are often sought-after in IR research
- If it is not opinion then it is the result of IR Evaluation
- IR has a very strong empirical tradition
- No such statements have merit unless shown, by rigorous evaluation, to be well founded
- IR evaluation is itself a HUGE area of research

Goals of today's lecture

- Understand how Information Retrieval Systems are evaluated (at least from a system's perspective)
- Look at the components / design of Systems IR evaluation
- Look at different metrics, the thinking behind them and how to calculate them
- Give you a feel for the strengths and limitations of systems evaluation in IR

Outline

What is Evaluation?

- Systematic determination of merit and significance of something using criteria against a set of standards [Järvelin, 2009 in Ruthven and Kelly p.113]
- So, we need:
- Some object to be evaluated (e.g. IR system)

  What do we mean by an IR System?
- Some goal that should be achieved (e.g. quality of retrieved result)

  What is the quality of the retrieved result and how can we measure it?

  As we can answer these questions in different ways, there are lots of different kinds of IR Evaluation

In Practical Life …

- People engage with IR when they need support to find information in order to complete a current task
  - Subjective or objective task benefit
  - How much did the system help them?
  - Difficult and expensive to investigate (control and measure)
- We need surrogates and goals that are easier to measure, to control and to repeat.
- Typically in Systems IR our object is not the human user completing a task, but the (ranked) list of documents returned

Types of IR Evaluation

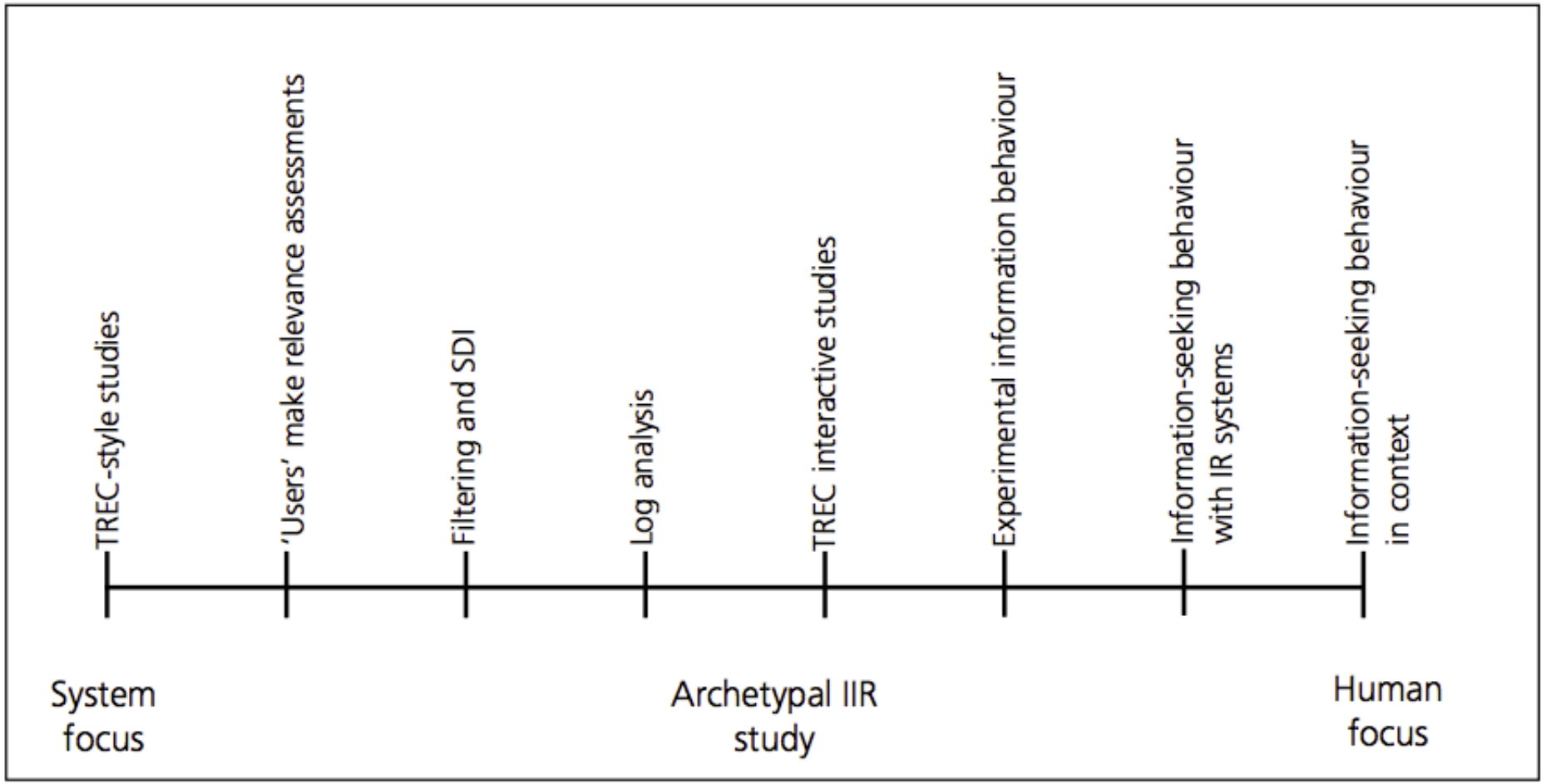


Figure 7.2 *The continuum of information retrieval evaluation studies (Kelly, 2009)*

Design an experiment:

- Starting point for any experiment is to design a hypothesis pair:

  H1: The quality of the ranking produced by System A (our new system) is better than that produced by a baseline system

  H0: There is no difference between the rankings produced

- Dependent variables (ranking produced – operationalised by some metric)
- Independent variables (Systems i.e. retrieval models)
- Concomitant variables – fixed to prevent uncontrolled variation (e.g. test collections and topics)
- Other variables remain hidden / confounded (motivation,knowledge, expertise etc.)

Test Collections

- Test collections consist of:
  - A document database used for retrieval
  - A set of test requests (topics) which represent information needs
  - A set of relevance judgements indicating , for each request, which documents are relevant (should be retrieved by a system)
- Numerous collections exist for many domains
- We will look later on at how these are created

What topics look like

- Consist of an **information need** and a **query** representative of a query that might be generated in this scenario
- Relevance is assessed relative to the **information need** *not* the **query**
- E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- Query: ***wine red white heart attack effective***
- Evaluate whether the doc addresses the information need, not whether it has these words

Design an experiment:

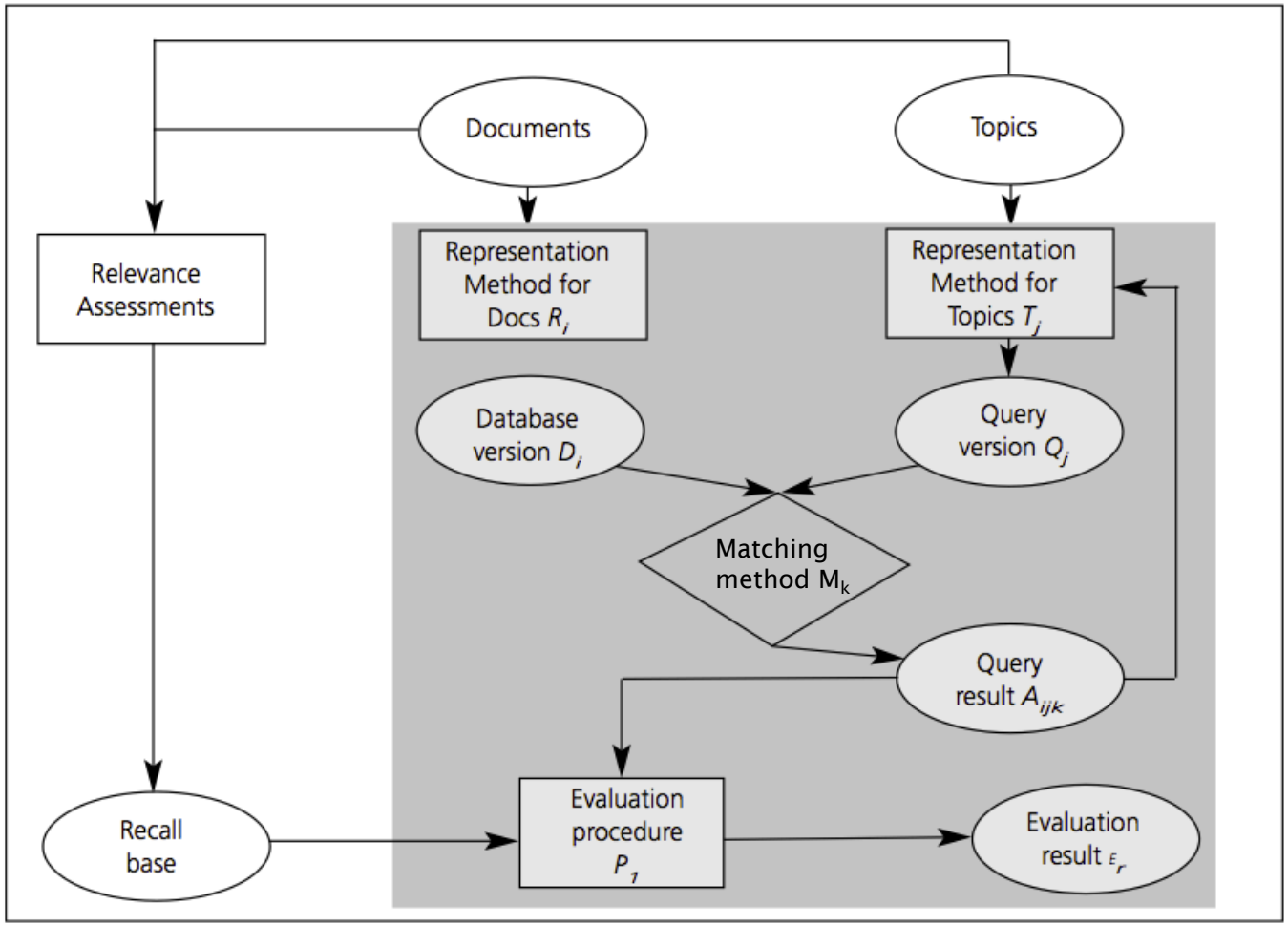


Figure 7.3 *Test-collection-based information retrieval evaluation setting (based on Järvelin, 2007)*

Concomitant variables

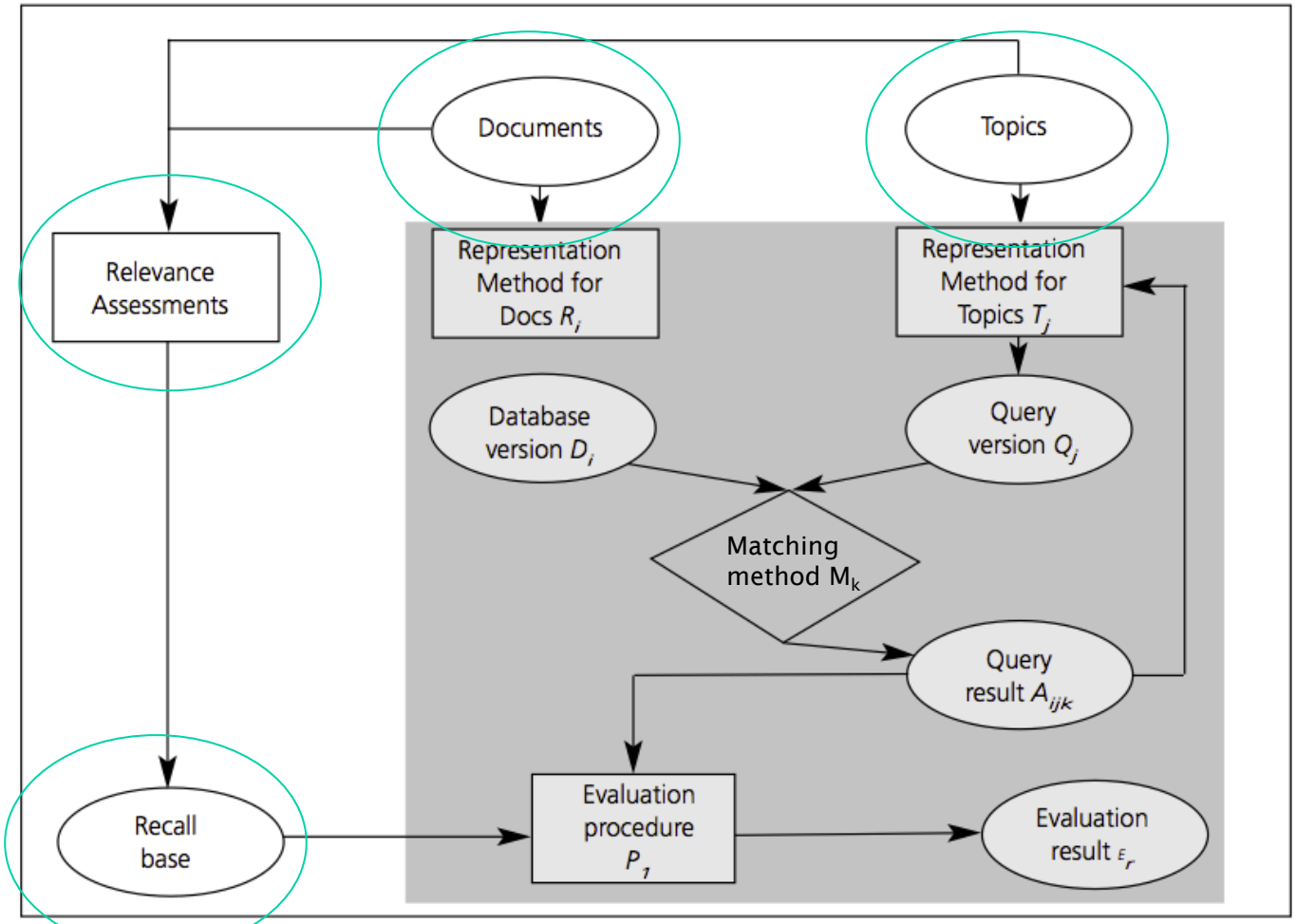


Figure 7.3 *Test-collection-based information retrieval evaluation setting (based on Järvelin, 2007)*

Dependent variables

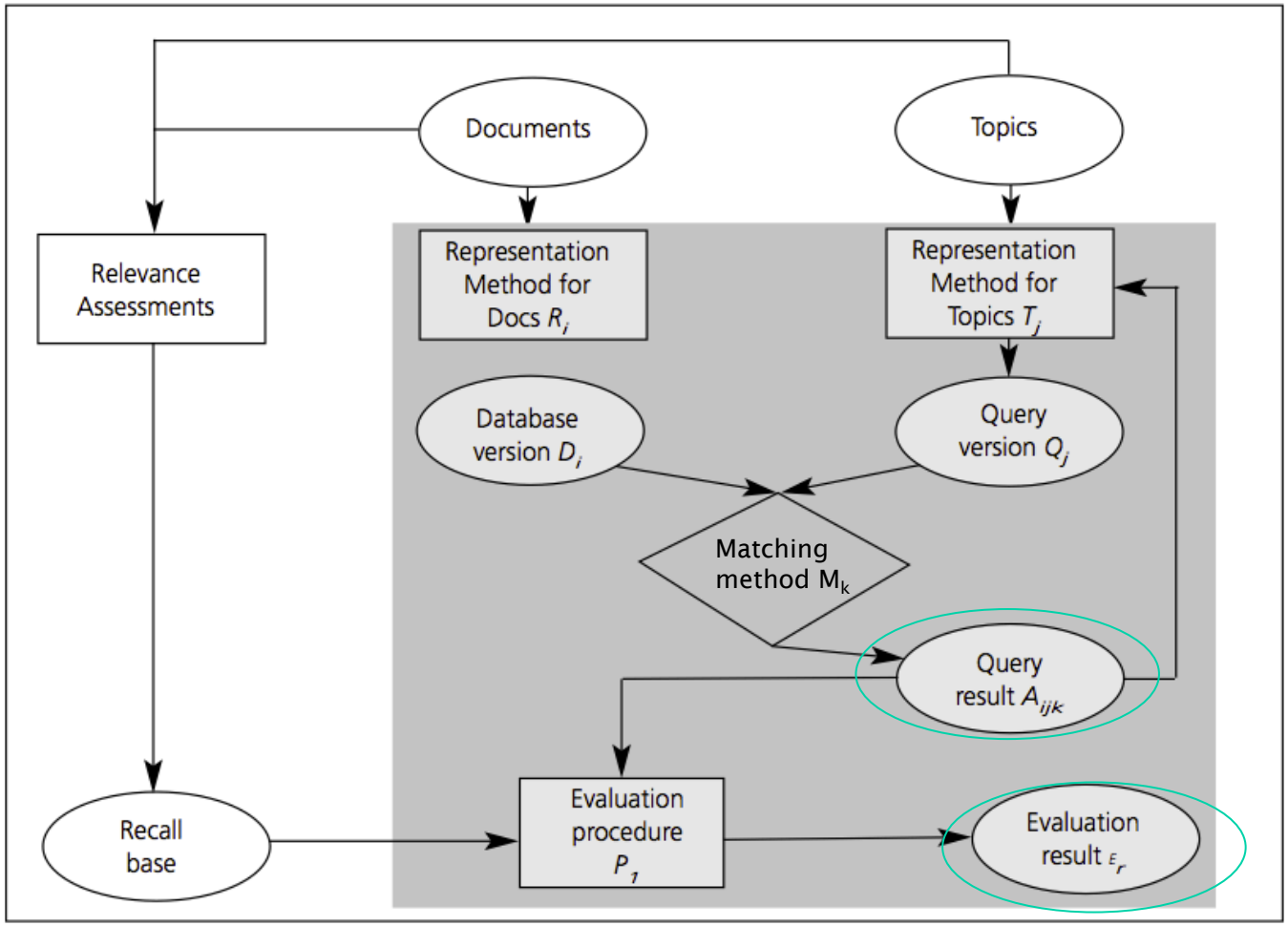


Figure 7.3 *Test-collection-based information retrieval evaluation setting (based on Järvelin, 2007)*

Everything else we can manipulate to suit our experimental needs

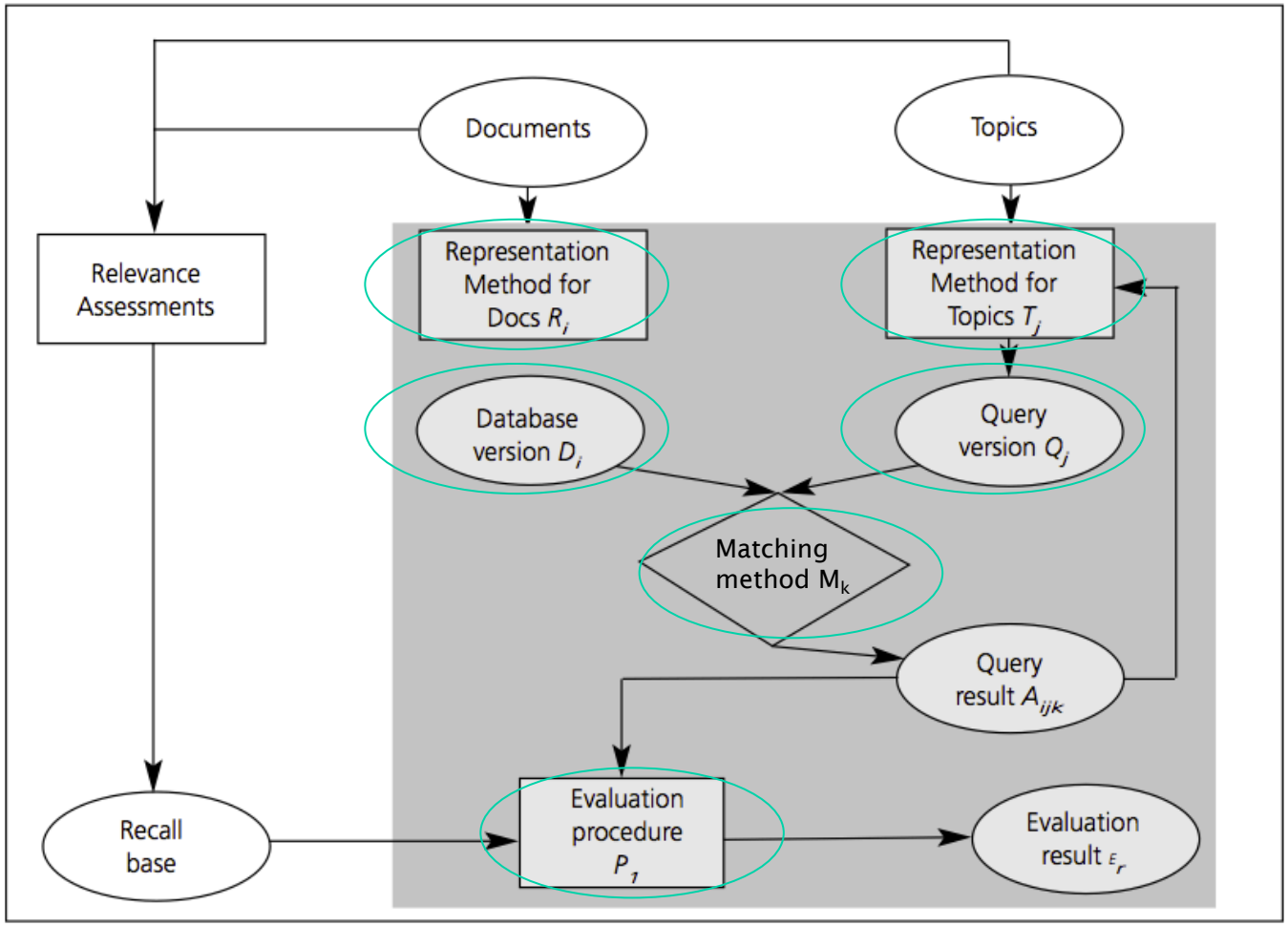


Figure 7.3 *Test-collection-based information retrieval evaluation setting (based on Järvelin, 2007)*

Standard relevance benchmarks

- TREC - National Institute of Standards and Technology (NIST) has run a large IR test bed for many years
- Reuters and other benchmark doc collections used
- "Retrieval tasks" specified
  - sometimes as queries
- Human experts mark, for each query and for each doc, Relevant or Nonrelevant
  - or at least for subset of docs that some system returned for that query

Outline

Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved)
- **Recall**: fraction of relevant docs that are retrieved = P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Should we instead use the accuracy measure for evaluation?

- Given a query, an engine classifies each doc as "Relevant" or "Nonrelevant"
- The **accuracy** of an engine: the fraction of these classifications that are correct
  - (tp + tn) / (tp + fp + fn + tn)
- **Accuracy** is a commonly used evaluation measure in machine learning classification work
- Why is this not a very useful evaluation measure in IR?

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

Precision/Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved

- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation

Difficulties in using precision/recall

- Should average over large document collection/ query ensembles
- Need human relevance assessments
  - People aren't reliable assessors
- Assessments have to be binary
  - Nuanced assessments?
- Heavily skewed by collection/authorship
  - Results may not translate from one domain to another

A combined measure: *F*

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ or $\alpha = ½$

$$\frac{2PR}{P+R}$$

- Harmonic mean is a conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

Calculating the F-measure

- What is the F_1 when P = 0.4 and R = 0.4

a) 0.2
b) 0.4
c) 0.8
d) 2.5

Calculating the F-measure (2)

- What is the F_1 when P = 0.75 and R = 0.25

a) 0.125

b) 0.25

c) 0.375

d) 0.5

To think about …

- Why not just use the arithmetic mean (averaging) to combine precision and recall?
- Hint: what do we need to do to get 100% recall?
- Do that and the you automatically get 50% F-score, which doesn't reflect that the system is not doing its job at all.
- However, because the harmonic mean tends to be much closer to the minimum of two values, in this case the score of 2% is much more appropriate

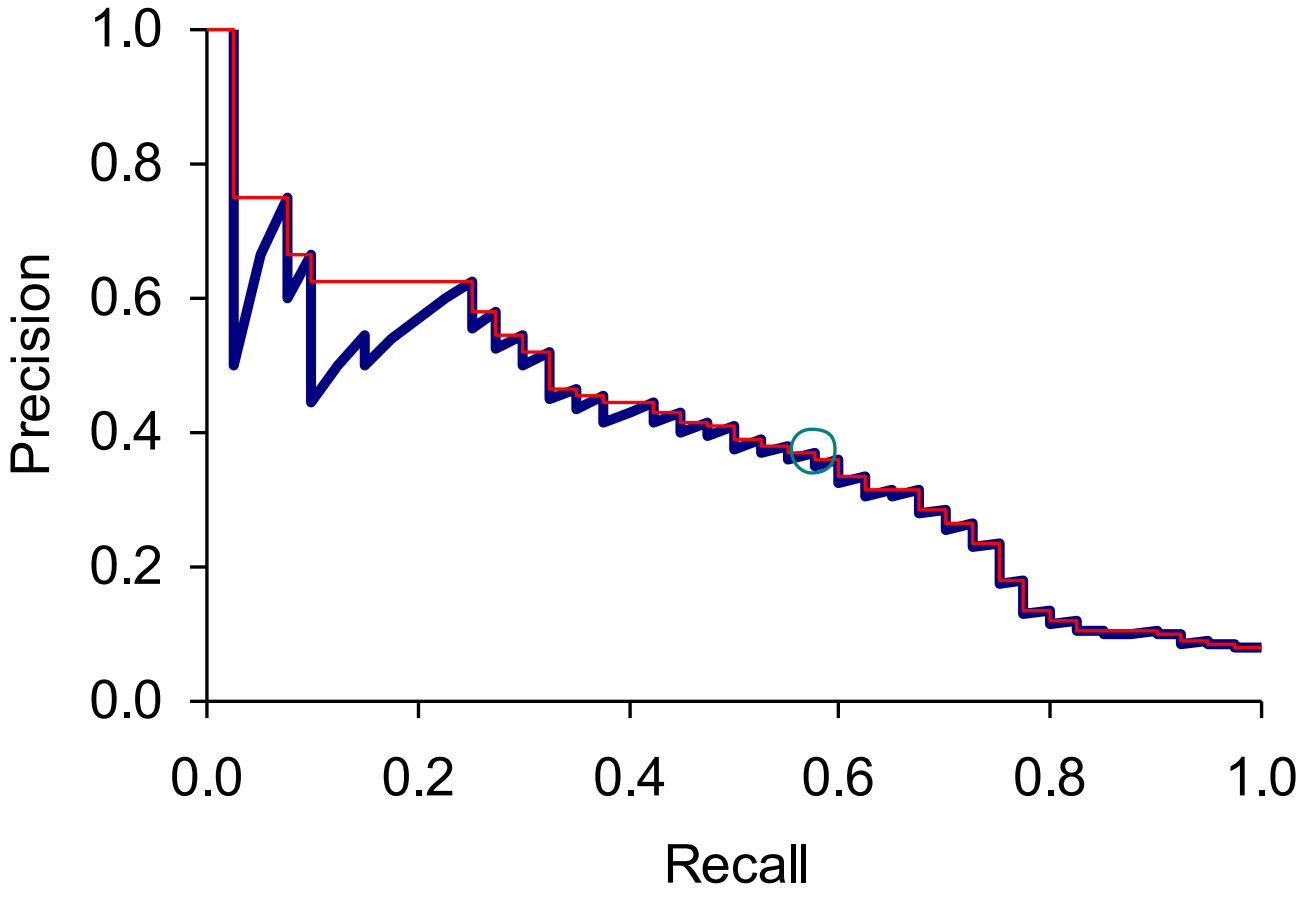
Evaluating ranked results

- Evaluation of ranked results:
  - The system can return any number of results
  - Precision, Recall and the F-Measure are set-based metrics
    - We need to extend these to deal with ranking
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

Precision and Recall for Ranked Lists

		R	P	10 relevant docs in total
1	R			
2	N			
3	N			
4	R			
5	R			
6	N			
7	N			
8	N			
9	N			
10	N			

A precision-recall curve

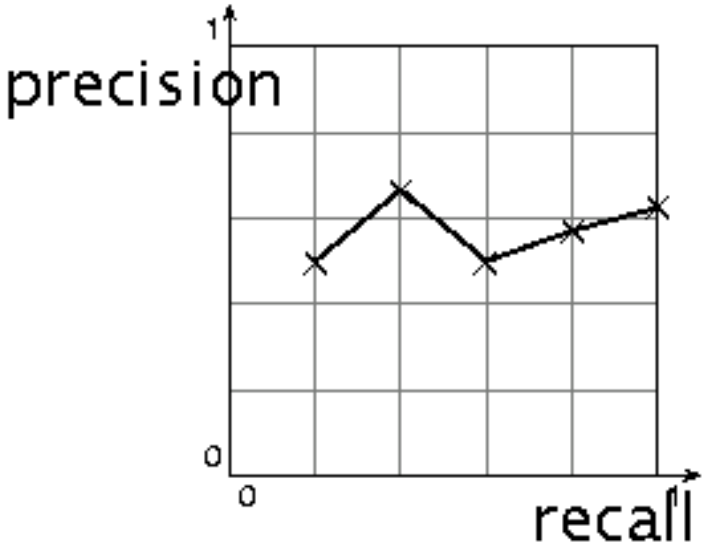


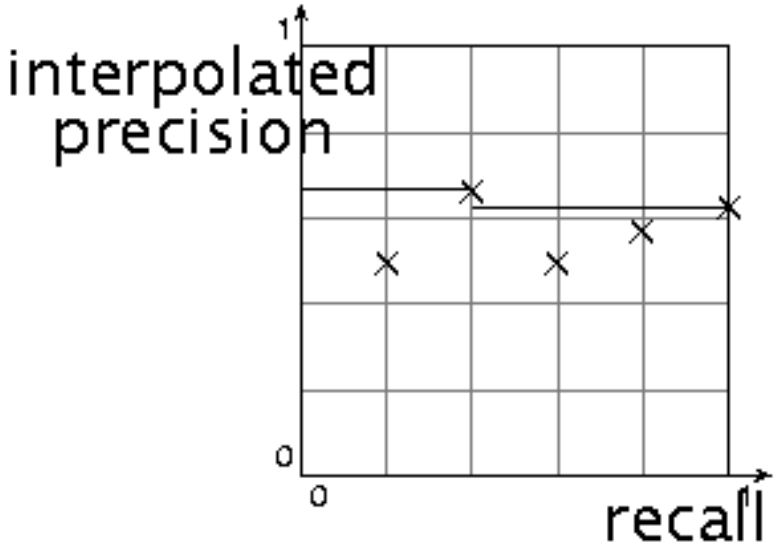
Averaging over queries

- A precision-recall graph for one query isn't a very sensible thing to look at
- You need to average performance over a whole bunch of queries.
- But there's a technical issue:
  - Precision-recall calculations place some points on the graph
  - How do you determine a value (interpolate) between the points?

Interpolated precision

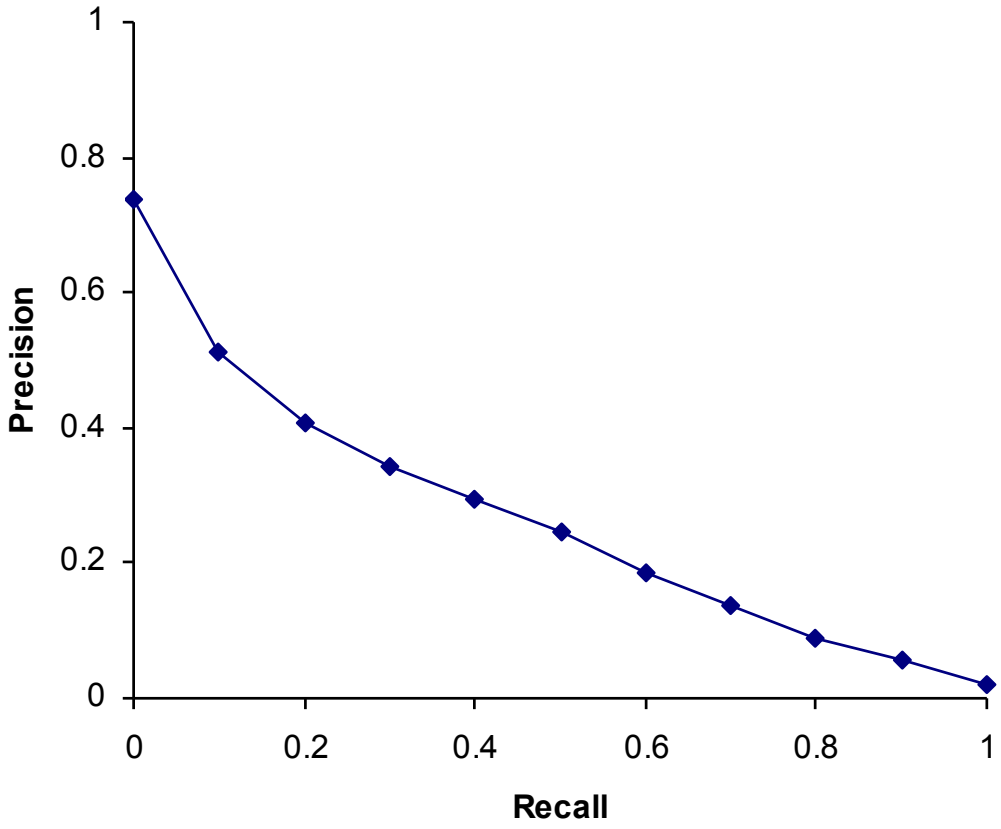
- Idea: If locally precision increases with increasing recall, then you should get to count that…
- So you take the max of precisions to right of value





Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



Evaluation

- Graphs are good, but people want summary measures!
  - Precision at fixed retrieval level
    - Precision-at-*k*: Precision of top *k* results
    - Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages
    - But: averages badly and has an arbitrary parameter of *k*

Mean Average Precision

- Mean average precision (MAP)
  - Average of the precision value obtained for the top *k* documents, each time a relevant doc is retrieved
  - Avoids interpolation, use of fixed recall levels
  - MAP for query collection is arithmetic ave.
    - Macro-averaging: each query counts equally

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

Variance

- For a test collection, it is usual that a system does poorly on some information needs (e.g., MAP = 0.1) and excellently on others (e.g., MAP = 0.7)
- Indeed, it is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query.

- That is, there are easy information needs and hard ones!

Outline

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

From document collections to test collections

- Still need
  - Test queries
  - Relevance assessments
- Test queries
  - Must be suited to docs available
  - Best designed by domain experts
  - Random query terms generally not a good idea
- Relevance assessments
  - Human judges, time-consuming, how is this possible with large document collections (1 Million plus?)
  - Are human panels perfect? Is this a problem?

Kappa measure for inter-judge (dis)agreement

- Kappa measure
  - Agreement measure among judges
  - Designed for categorical judgments
  - Corrects for chance agreement
- Kappa = [P(A) – P(E)] / [1 – P(E)]
- P(A) – proportion of time judges agree
- P(E) – what agreement would be by chance
- Kappa = 0 for chance agreement, 1 for total agreement.

P(A)? P(E)?

Kappa Measure: Example

Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

Kappa Example

- P(A) = 370/400 = 0.925
- P(nonrelevant) = (10+20+70+70)/800 = 0.2125
- P(relevant) = (10+20+300+300)/800 = 0.7878
- P(E) = 0.2125^2 + 0.7878^2 = 0.665
- Kappa = (0.925 – 0.665)/(1-0.665) = 0.776

Interpreting the kappa

κ	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

- Depends on purpose of study
- For >2 judges: average pairwise kappas

TREC

- TREC Ad Hoc task from first 8 TRECs is standard IR task
  - 50 detailed information needs a year
  - Human evaluation of pooled results returned
  - More recently other related things: Web track, HARD
- A TREC query (TREC 5)

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>

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

Impact of Inter-judge Agreement

- Impact on absolute performance measure can be significant (0.32 vs 0.39)
- Little impact on ranking of different systems or relative performance
- Suppose we want to know if algorithm A is better than algorithm B
- A standard information retrieval experiment will give us a reliable answer to this question.

Outline

① Intro and Motivation

② A Framework for Evaluation

③ Metrics

④ Building Test Collections

⑤ Discussion

Discussion

- Trec-style evaluation gives us a means to perform controlled experiments
  - Repeatable
  - Fair comparison across models
  - Low cost (once collections are created)
- A few issues:
  - Difficult to deal with user interaction (does this accurately model the way people search)
  - Relevance vs Marginal Relevance (A document can be redundant even if it is highly relevant- think duplicates)

Summary

- Looked at how IR Systems can be evaluated using the TREC approach
- Components, how the experiment is setup