

Part 7: Evaluation



Francesco Ricci

Most of these slides comes from the
course:

Information Retrieval and Web Search,
Christopher Manning and Prabhakar
Raghavan

This lecture

- How do we know if our results are any good?
 - Evaluating a search engine
 - Benchmarks
 - Precision and recall
 - Accuracy
 - Inter judges disagreement
 - Normalized discounted cumulative gain
 - A/B testing
- Results summaries:
 - Making our good results usable to a user.

Measures for a search engine

- How **fast** does it **index**
 - Number of documents/hour
 - (Average document size)
- How **fast** does it **search**
 - Latency as a function of index size
- **Expressiveness** of query language
 - Ability to express complex information needs
 - Speed on complex queries
- **Uncluttered UI**
- Is it free? 😊

Measures for a search engine

- All of the preceding criteria are ***measurable***: we can quantify speed/size
 - we can make expressiveness precise
- But the key measure: **user happiness**
 - What is this?
 - Speed of response/size of index are factors
 - But blindingly fast, useless answers won't make a user happy
- Need a way of quantifying user happiness.

Measuring user happiness

- Issue: who is the user we are trying to make happy?
 - Depends on the setting
- **Web engine:**
 - User finds what they want and return to the engine
 - Can measure rate of return users
 - User completes their task – search as a means, not end
 - See Russell
<http://dmrussell.googlepages.com/JCDL-talk-June-2007-short.pdf>
- **eCommerce site:** user finds what they want and buy
 - Is it the end-user, or the eCommerce site, whose happiness we measure?
 - Measure time to purchase, or fraction of searchers who become buyers?

Measuring user happiness

- **Enterprise** (company/govt/academic): Care about “user productivity”
 - How much time do my users save when looking for information?
 - Many other criteria having to do with breadth of access, secure access, etc.

Reviews

- Bolzano
- Bolzano Tourism
- Bolzano Hotels
 - Parkhotel Laurin
- Flights to Bolzano
- Bolzano Deals
- More On Bolzano**
- Restaurants
- Things to Do
- Travel Forum
- Travel Guide
- Photos
- Map
- Bolzano Deals**
- All Travel Offers

Free Newsletter

Interested in **Parkhotel Laurin** and **Bolzano**?

We'll send you updates with **the latest deals, reviews and articles for Parkhotel Laurin and Bolzano** each week.

Home → Europe → Italy → Trentino-Alto Adige → South Tyrol → Bolzano → Bolzano Hotels

Parkhotel Laurin

[Compare Bolzano business hotels](#)



- [Hotel photos](#)
- [Map this hotel](#)
- [Hotel amenities](#)



Hotel class **★★★★★**
Via Laurin 4, I-39100 Bolzano, Italy

Check Rates and Availability

Check-in: Adults:

mm/dd/yyyy mm/dd/yyyy

- Venere.com
- Booking.com
- it.octopustravel.com
- Expedia.it
- Hotels.com

Opens one window for each offer. Please disable pop-up blockers.

Reviews you can trust

91% Recommend

69 reviews

Excellent		42
Very good		21
Average		3
Poor		2
Terrible		1

Filter reviews by trip type

- All (69)**
- Business (0)
- Couples (19)
- Family (6)
- Friends getaway (2)
- Solo travel (6)

1-10 of 69

« 1 2 ... 7 »

TripAdvisor Popularity Index

1 of 53 hotels in Bolzano

Rating Details Photos (25) Map

TripAdvisor Traveler Rating

69 Reviews

91% | [Write a review](#)

“peculiar hotel!!”
Feb 6, 2010 - fabmc

“Great location”
Nov 27, 2009 - drewantha

Chiama ora per prenotare: da Hotels.com **800 924 633**

[Click here for best prices for Parkhotel Laurin](#)

- Parkhotel Laurin:** [Confronta i nostri prezzi!](#) **Venere.com** Nessun anticipo, pagamento in Hotel. Prenota ora e risparmi!
- Parkhotel Laurin:** [Risparmia fino al 75%](#) **Booking.com** Prezzi bassi e zero commissioni! Prenota ora online, paga in hotel
- Parkhotel Laurin:** [Octopus Travel](#) **it.octopustravel.com** 55.000 hotel in ogni parte del mondo accettati fino al 70%

A Review

“ peculiar hotel!! ”

Parkhotel Laurin





fabmc  27 contributions
Venice, Italy

Feb 6, 2010

Save Review

Very peculiar hotel. I had a few problems with their reservation system, but everything else worked fine.

My ratings for this hotel

 Value  Service
 Rooms
 Location
 Cleanliness

Multiple Criteria

Date of stay February 2010

Visit was for Leisure

Traveled with Other

Member since June 20, 2008

Would you recommend this hotel to a friend? No

Context of the search

This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC.

Was this review helpful? [Yes](#)

[View profile](#) | [Send message](#) | [Compliment reviewer](#)

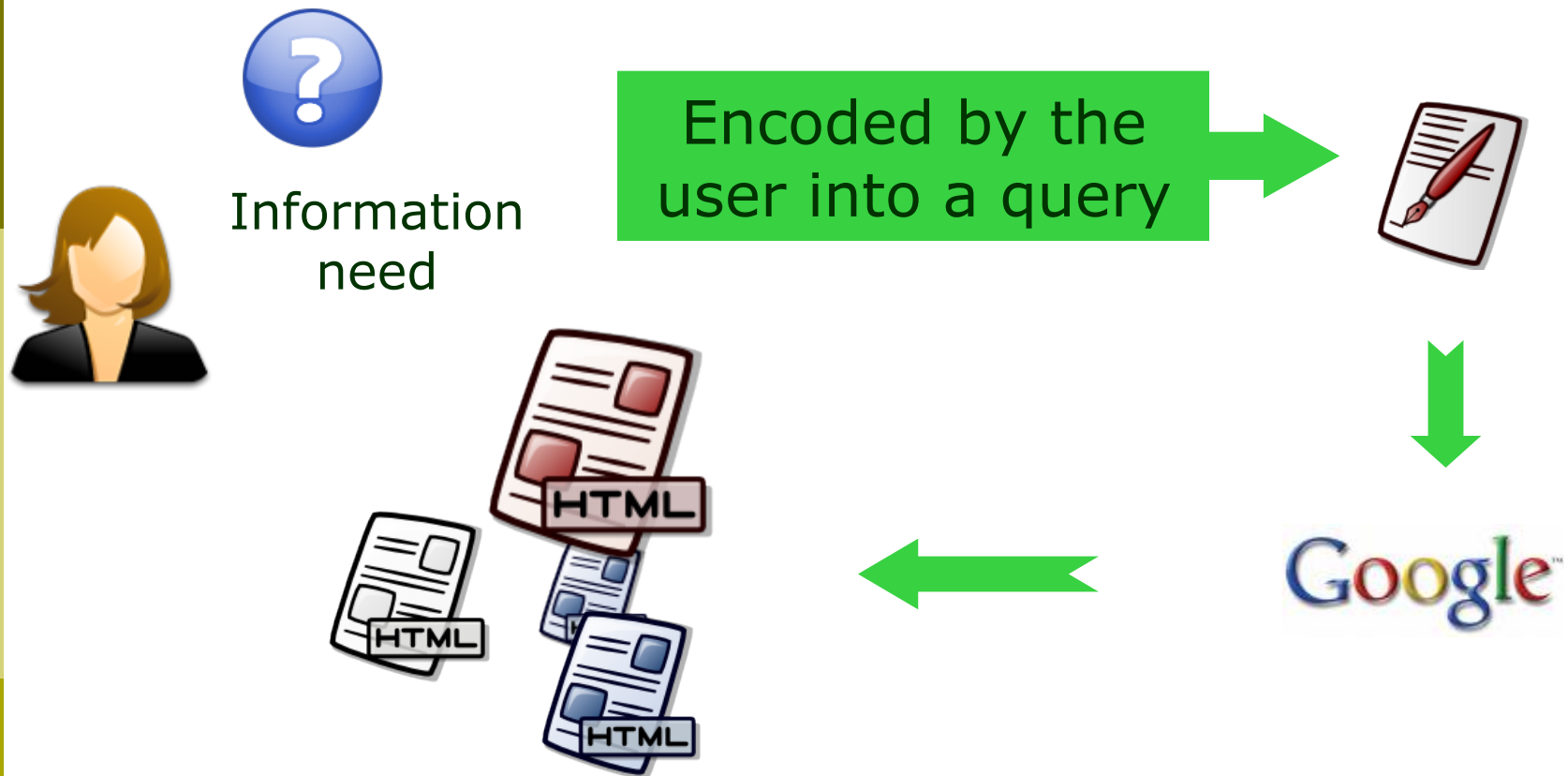
[Report Inappropriate Content](#)

Evaluation and assessment
of the evaluation

Happiness: elusive to measure

- ❑ Most common proxy: **relevance** of search results
- ❑ *But how do you measure relevance?*
- ❑ We will detail a methodology here, then examine its issues
- ❑ Relevance measurement requires 3 elements:
 1. A benchmark document collection
 2. A benchmark suite of queries
 3. A usually binary assessment of either Relevant or Nonrelevant for each query and each document
 - ❑ Some work on more-than-binary, but not the standard.

From needs to queries



- Information need -> query -> search engine -> results -> browse OR query -> ...

Evaluating an IR system

- Note: the **information need** is translated into a **query**
- 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**
- You evaluate whether the doc addresses the information need, not whether it has these words.

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.*

Relevance and Retrieved documents

Information need

relevant

not relevant

TP

FP

retrieved

FN

TN

not retrieved

Documents

Query and system



Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = $P(\text{relevant}|\text{retrieved})$
- **Recall**: fraction of relevant docs that are retrieved = $P(\text{retrieved}|\text{relevant})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision $P = tp/(tp + fp) = tp/\text{retrieved}$
- Recall $R = tp/(tp + fn) = tp/\text{relevant}$

Accuracy

- Given a query, an engine (**classifier**) classifies each doc as "Relevant" or "Nonrelevant"
 - What is retrieved is classified by the engine as "relevant" and what is not retrieved is classified as "nonrelevant"
- The **accuracy** of the 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?

A screenshot of a search engine interface. The text 'snoogle.com' is displayed in a stylized font with blue and pink colors. Below it, the text 'Search for:' is followed by an empty rectangular input box. At the bottom, the text '0 matching results found.' is displayed in a yellow-green color.

snoogle.com

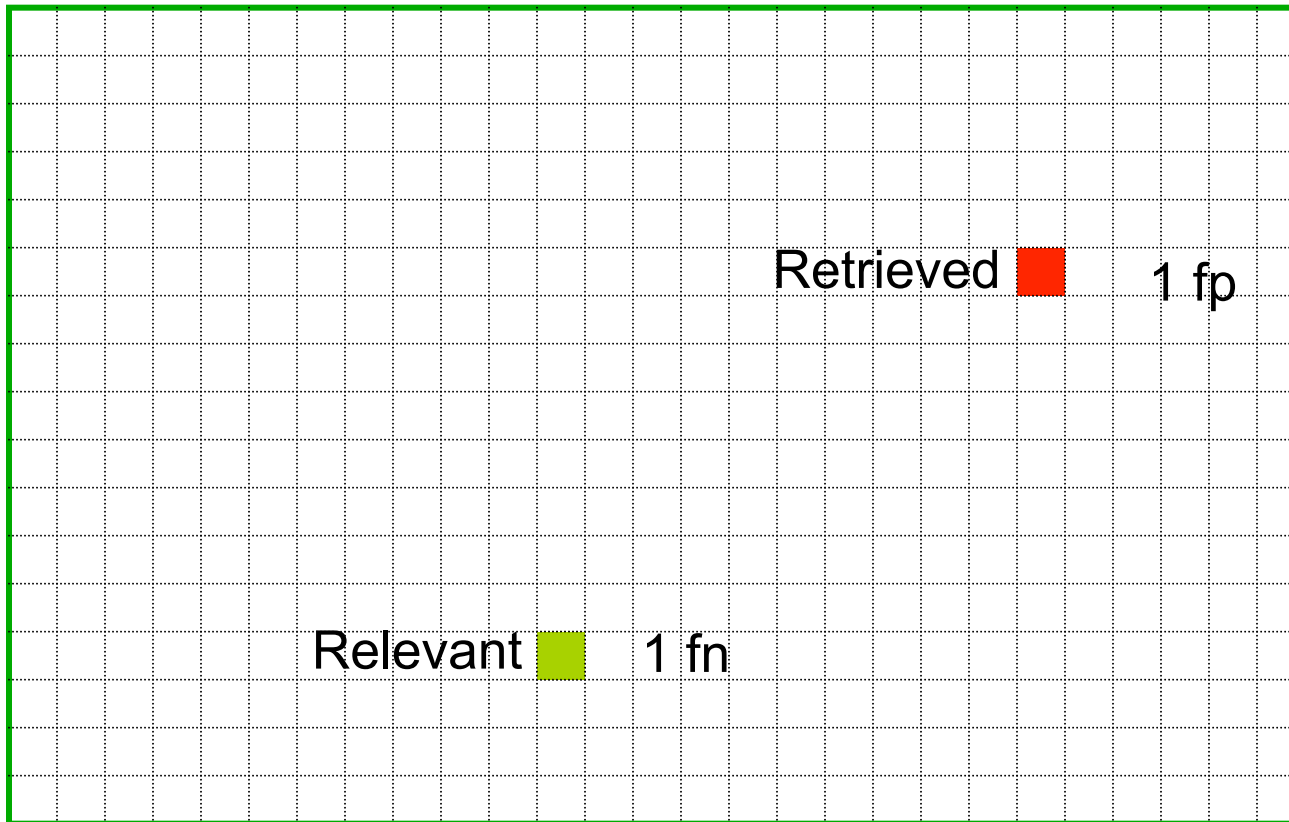
Search for:

0 matching results found.

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

Precision, Recall and Accuracy

Very low precision, very low recall, high accuracy



$$p = 0$$

$$r = 0$$

positive = retrieved
negative = not retrieved

$$a = (tp + tn) / (tp + fp + fn + tn)$$
$$= (0 + (27 * 17 - 2)) / (0 + 1 + 1 + (27 * 17 - 2)) = 0.996$$

Precision/Recall

- What is the recall of a query if you retrieve all the documents?
- 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**
 - *Why?*
- In a good system, **precision decreases as either the number of docs retrieved or recall increases**
 - This is not a theorem (why?), 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

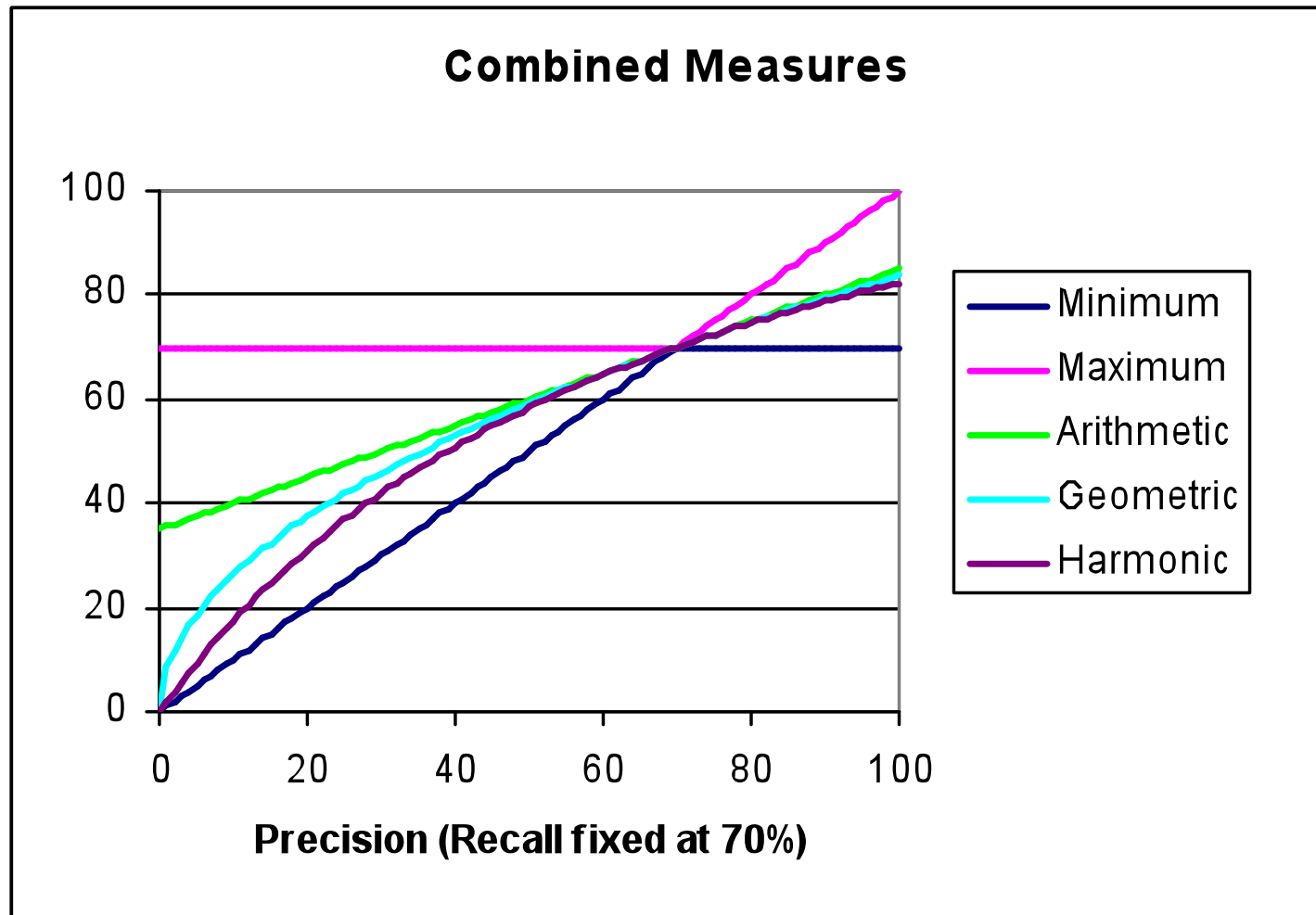
- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

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

$$\beta^2 = \frac{1 - \alpha}{\alpha}$$

- People usually use balanced F_1 measure
 - i.e., with $\beta = 1$ or $\alpha = 1/2$
- Harmonic mean is a **conservative** average
 - See CJ van Rijsbergen, *Information Retrieval*

F_1 and other averages



Geometric mean of a and b is $(a*b)^{1/2}$

Evaluating ranked results

- Evaluation of ranked results:
 - The system can return any number of results – by varying its behavior or
 - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*.

Precision-Recall



Web [+ Show options...](#)

[Cop Land \(1997\)](#)

Do you know that he was paid only \$60,000 for his acting in **Cop Land**, ... To me **Cop land** is the kind of movie Stallone should have made after First Blood. ...
[www.imdb.com/title/tt0118887/](#) - 13 hours ago - [Cached](#) - [Similar](#)

[Aaron Copland - Wikipedia, the free encyclopedia](#)

Before emigrating from Scotland to the United States, **Copland's** father, Travels to Italy, Austria, and Germany rounded out **Copland's** musical education. ...
[Biography](#) - [Composer](#) - [Film composer](#) - [Critic, writer, and teacher](#)
[en.wikipedia.org/wiki/Aaron_Copland](#) - [Cached](#) - [Similar](#)

[Copland - Wikipedia, the free encyclopedia](#)

From Wikipedia, the free encyclopedia. Jump to: navigation, search. **Copland** can mean: [ec Surname. Aaron **Copland** (1900–1990), American composer ...
[en.wikipedia.org/wiki/Copland](#) - [Cached](#) - [Similar](#)

[+ Show more results from en.wikipedia.org](#)

[Books by Aaron Copland](#)

[What to Listen for in Music](#) - 2002 - 308 pages

[Music and Imagination](#) - 1980 - 134 pages

[Aaron Copland: A Reader Selected Writings 1923 ...](#) - 2004 - 416 pages

[books.google.it](#) - [More book results »](#)

[COPLAND](#)

Maker and one line of products: stereo and multi-channel valve amplifier, stereo and multi-channel power amplifier and cd player.
[www.copland.dk/](#) - [Cached](#) - [Similar](#)

[Aaron Copland | American Composer](#)

4 Jan 2010 ... Lucidcafé's profile noting life, works, and style with photograph and links.
[www.lucidcafe.com/library/95nov/copland.html](#) - [Cached](#) - [Similar](#)

[Classical Net - Basic Repertoire List - Copland](#)

As much as anyone, Aaron **Copland** established American concert music through his

What is
1 000?

$P=0/1, R=0/1000$

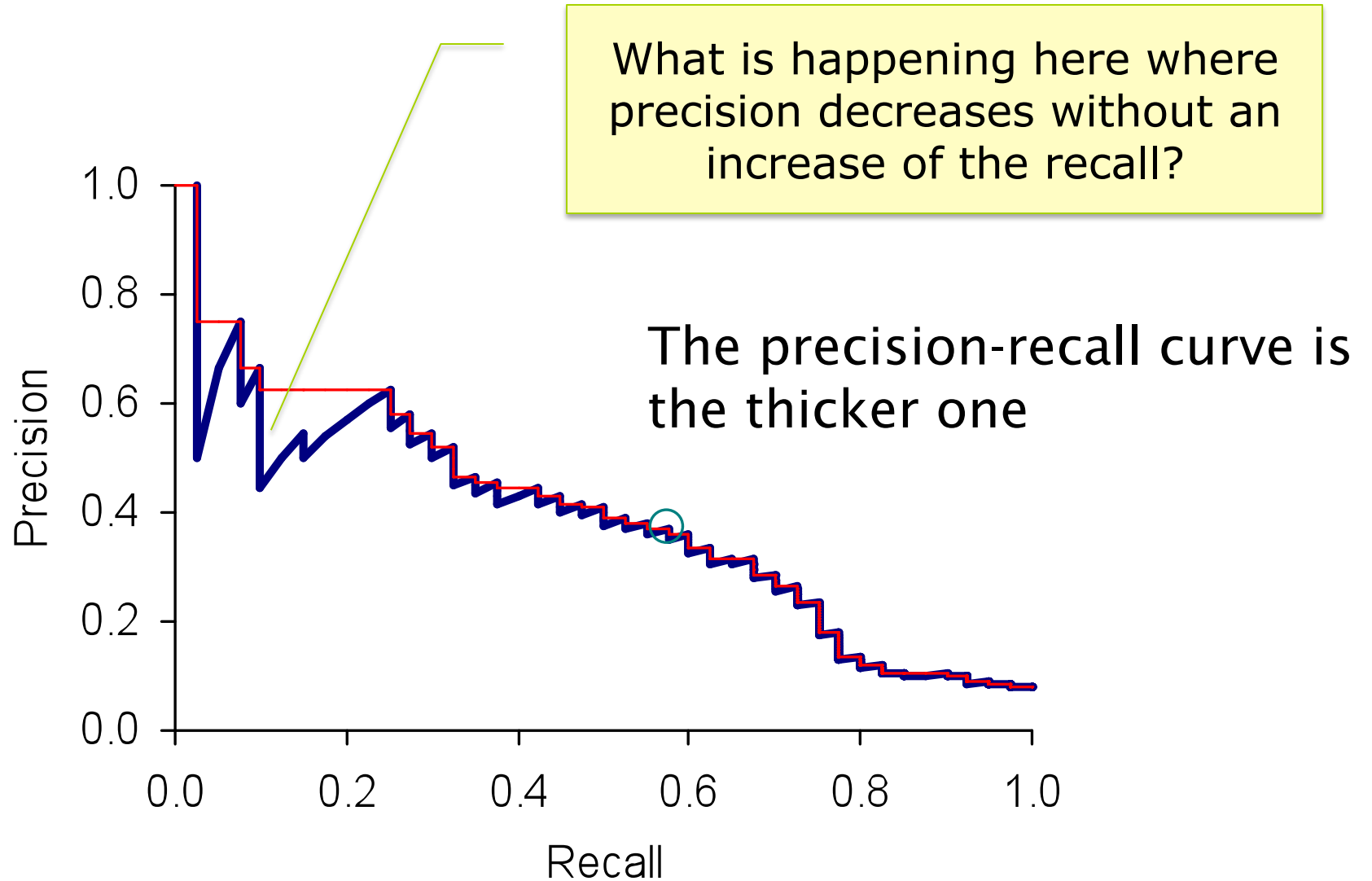
$P=1/2, R=1/1000$

$P=2/3, R=2/1000$

$P=2/4, R=2/1000$

$P=3/5, R=3/1000$

A precision-recall curve

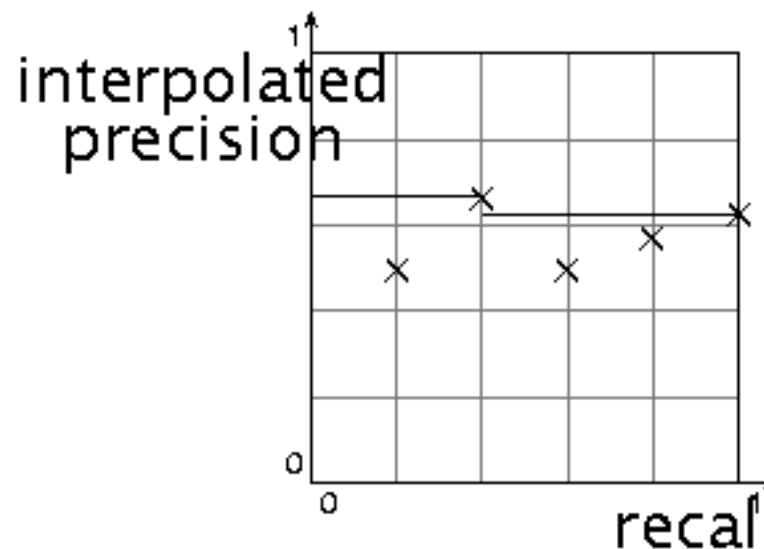
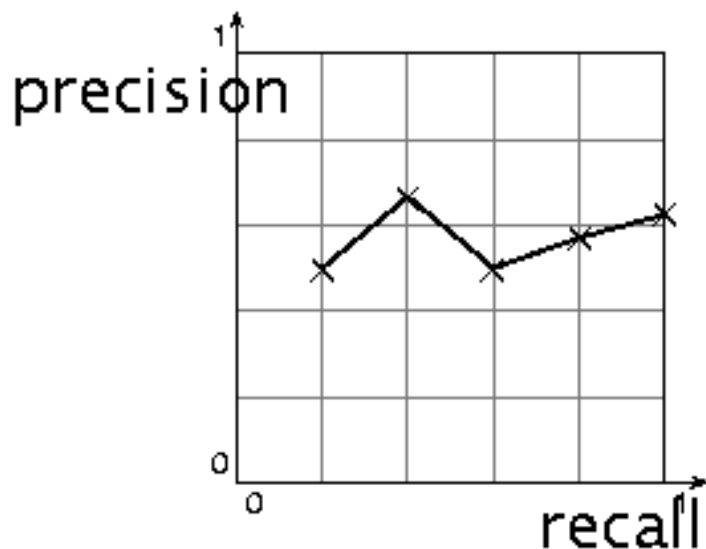


Averaging over queries

- 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 the precisions for all the greater values of recall



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

Definition of
interpolated precision

Evaluation: Precision at k

- 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 .

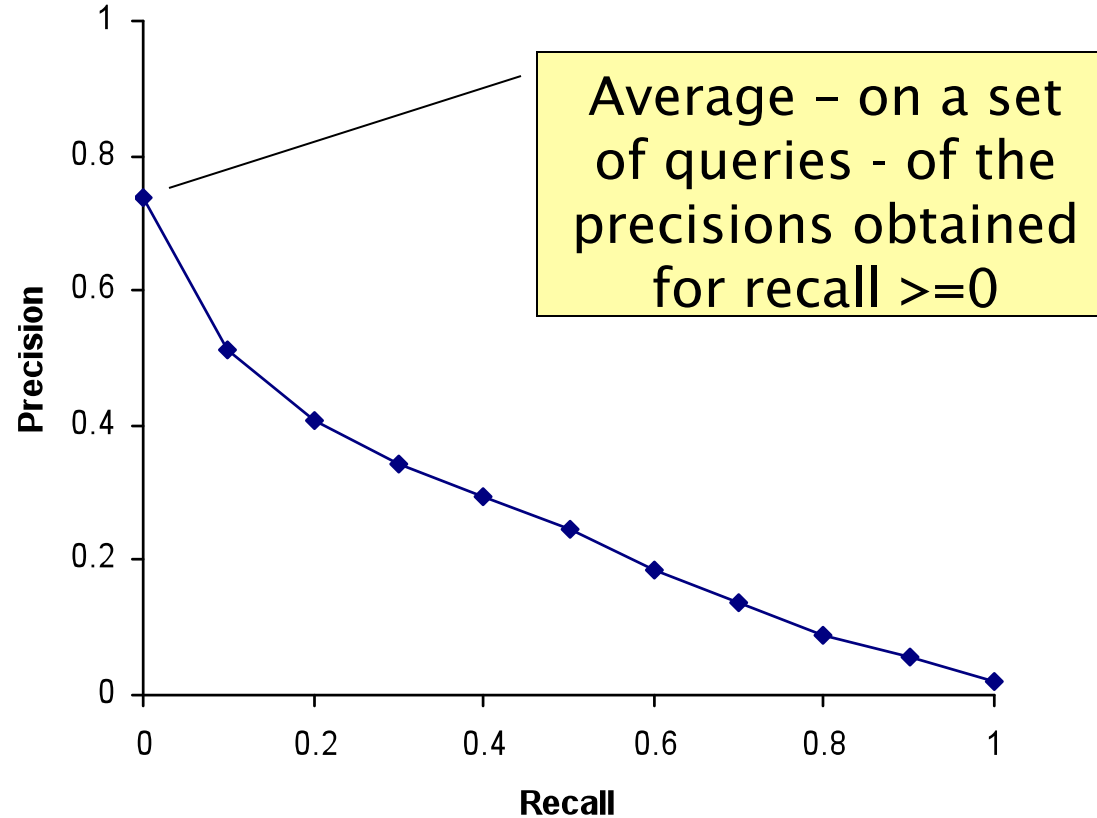
Evaluation: 11-point interpolated prec.

□ 11-point interpolated average precision

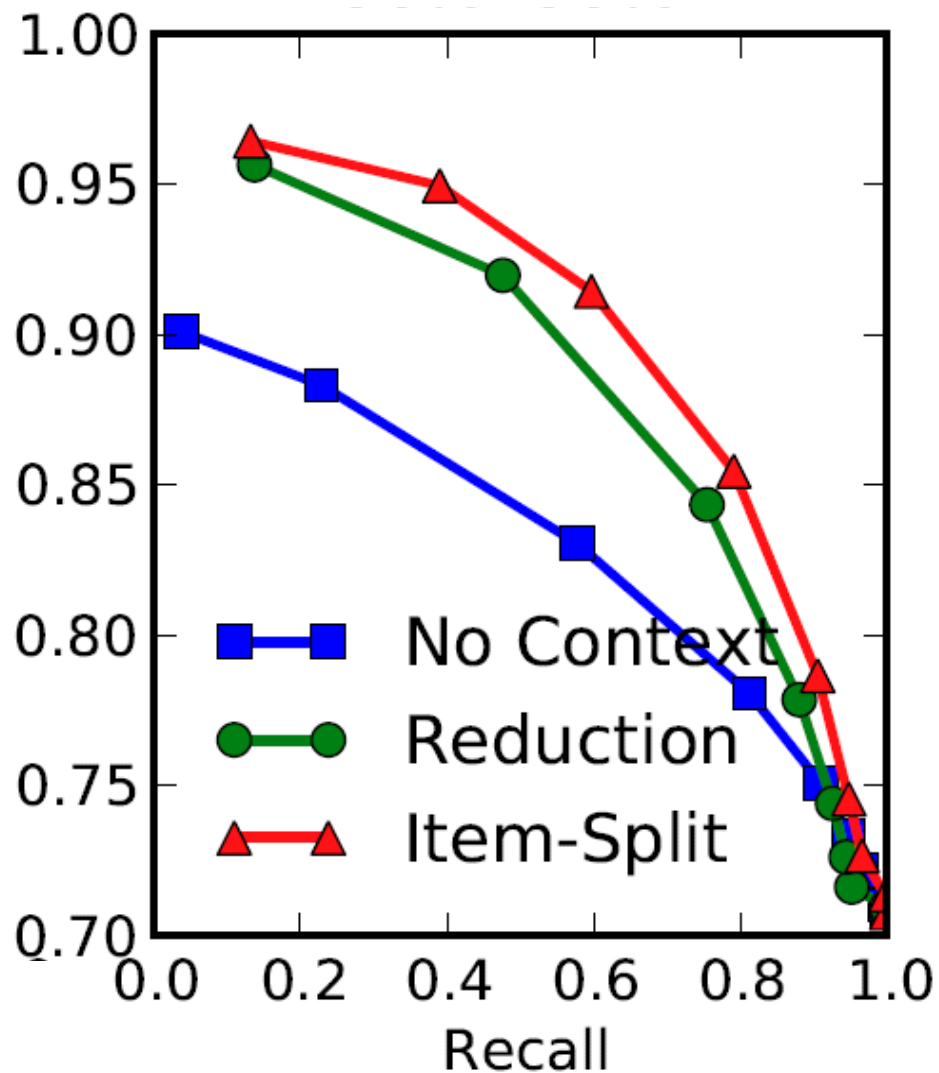
- The standard measure in the early TREC competitions
- Take the interpolated precision at 11 levels of recall varying from 0 to 1 by tenths
- The value for 0 is always interpolated!
- Then **average them**
- Evaluates performance at all recall levels.

Typical (good) 11 point precisions

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



Precision recall for recommenders



- Relevant if the true rating is ≥ 4
- Retrieve all the items whose predicted rating is $\geq x$ ($x=5, 4.5, 4, 3.5, \dots 0$)
- You get 10 points to plot
- Why precision is not going to 0? Exercise.
- What the 0.7 value represents? I.e. the precision a recall = 1.

Mean average precision (MAP)

- Average of the precision values obtained for increasing values of k , for the top k documents, each time a new relevant doc is retrieved
- Avoids interpolation, use of fixed recall levels
- MAP for a query collection is arithmetic average
 - Macro-averaging: each query counts equally
- **Definition:** if the set of relevant documents for an information need q_j is $\{d_1, \dots, d_{m_j}\}$ and R_{jk} is the set of documents retrieved until you get d_k , then:

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

Example

Q1



1/1



2/3



3/7



⋮



Q2



1/1



2/2



3/6



4/7



⋮



$$(1 + 2/3 + 3/7) / 3 = 0.69$$

$$(1 + 1 + 3/6 + 4/7) / 4 = 0.76$$

Average precision =
 $(0.69 + 0.76) / 2 = 0.72$

 nonrelevant

 relevant

R-precision

- If I know the set of relevant documents Rel , then calculate the precision of top $|Rel|$ docs returned
- Perfect system could score 1.0.
- If there are $|Rel|$ relevant documents for a query, we examine the top $|Rel|$ results of a system, and find that r are relevant then
 - $P = r/|Rel|$
 - $R = r/|Rel|$
- R-precision turns out to be identical to the *break-even point*, i.e., where precision is equal to recall.

Variance

- For a test collection, it is usual that a system does very bad 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!



CREATING TEST COLLECTIONS FOR IR EVALUATION

Test Collections

TABLE 4.3 Common Test Corpora

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
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 appropriate for docs available
 - Best designed by domain experts
 - Random query terms generally not a good idea
- Relevance assessments
 - Human judges, time-consuming
 - Are human panels perfect?

TREC (Text REtrieval Conference)

- 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)
 - a **topic id** or number;
 - a **short title**, which could be viewed as the type of **query** that might be submitted to a search engine;
 - a description of the **information need** written in no more than one sentence; and
 - a **narrative** that provided a more complete description of what documents the searcher would consider as relevant.

<http://trec.nist.gov/>

Example TREC ad hoc topic

<top>

<num> Number: 200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description: Document must report on how the importation of foreign textiles or textile products has influenced or impacted on the U.S. textile industry.

<narr> Narrative: The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. "Textile industry" includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.

</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

Unit of Evaluation

- We can compute precision, recall, and F curve for different units
- Possible units (i.e., *what content is retrieved*):
 - Documents (most common)
 - Facts (used in some TREC evaluations)
 - Entities (e.g., car companies)
- May produce different results. Why?

Kappa measure for inter-judge (dis) agreement

- Kappa measure
 - Agreement measure among judges
 - Designed for categorical judgments
 - Corrects for chance agreement
- $\text{Kappa} = [P(A) - P(E)] / [1 - P(E)]$
- $P(A)$ – proportion of time judges agree
- $P(E)$ – what agreement would be by chance – but using the probability to output relevant/nonrelevant as observed in the panel of the judges
- Kappa = 0 for chance agreement, 1 for total agreement.

Kappa Measure: Example

P(A)? P(E)?

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

Kappa Example

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

- $\text{Kappa} > 0.8 = \text{good agreement}$
- $0.67 < \text{Kappa} < 0.8 \rightarrow \text{"tentative conclusions" (Carletta '96)}$
- Depends on purpose of study
- *For >2 judges: average pairwise kappas*

Interjudge Agreement: TREC 3

nonrelevant

information need	number of docs judged	disagreements	NR	R
51	211	6	4	2
62	400	157	149	8
67	400	68	37	31
95	400	110	108	2
127	400	106	12	94

Impact of Inter-judge Agreement

- Impact on **absolute** performance measure can be significant (e.g., 0.32 using a judge vs 0.39 using the other judge)
- 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.

Critique of pure relevance

- Relevance vs **Marginal Relevance**
 - A document can be **redundant** even if it is highly relevant
 - Duplicates
 - The same information from different sources
 - **Marginal relevance is a better measure of utility for the user**
- Using facts/entities as evaluation units more directly measures true relevance
- But harder to create evaluation set.

Can we avoid human judgment?

- No
- Makes experimental work hard
 - Especially on a large scale
- In some very specific settings, can use proxies
- E.g.: for testing an approximate vector space retrieval:
 - compare the cosine distance closeness of the **true closest docs** to those found by the approximate retrieval algorithm
- But once we have test collections, we can reuse them (so long as we don't overtrain too badly).

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web (why?)
- Search engines often use top k precision, e.g., $k=10$
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right: **NDCG (Normalized Cumulative Discounted Gain)**
- Search engines also use non-relevance-based measures:
 - **Clickthrough on first result:** Not very reliable if you look at a single clickthrough ... but pretty reliable in the aggregate
 - Studies of **user behavior in the lab**
 - **A/B testing.**

Normalised Discounted Cumulative Gain

- Like precision at k , it is evaluated over some number k of top search results
- For a set of queries Q , let $R(j, d)$ be the relevance score that **human assessors** gave to document **at rank index d** for query j

$$\text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1 + m)}$$

- where Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at k for query j is 1
- For queries for which $k' < k$ documents are retrieved, the last summation is done up to k' .

A/B testing

- **Purpose:** Test a single innovation
- **Prerequisite:** You have a large search engine up and running.
- Have **most users use old system**
- **Divert a small proportion of traffic** (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness
- Probably the evaluation methodology that large search engines trust most (true also for RecSys).

RESULTS PRESENTATION

Result Summaries

- Having ranked the documents matching a query, we wish to present a results list
- Most commonly, a list of the document titles plus a short summary, aka “10 blue links”

[John McCain](#)

John McCain 2008 - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.johnmccain.com · [Cached page](#)

[JohnMcCain.com - McCain-Palin 2008](#)

John McCain 2008 - The Official Website of **John McCain's** 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.johnmccain.com/Informing/Issues · [Cached page](#)

[John McCain News- msnbc.com](#)

Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...
www.msnbc.msn.com/id/16438320 · [Cached page](#)

[John McCain | Facebook](#)

Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...
www.facebook.com/johnmccain · [Cached page](#)

Summaries

- The title is often automatically extracted from document metadata. What about the summaries?
 - This description is crucial
 - User can identify good/relevant hits based on description
- Two basic kinds:
 - Static
 - Dynamic
- A **static summary** of a document is always the same, regardless of the query that hit the doc
- A **dynamic summary** is a *query-dependent* attempt to explain why the document was retrieved for the query at hand.

Static summaries

- In typical systems, the static summary **is a subset of the document**
- **Simplest heuristic:** the first 50 (or so – this can be varied) words of the document
 - Summary cached at indexing time
- **More sophisticated:** extract from each document a set of “key” sentences
 - Simple NLP heuristics to score each sentence
 - Summary is made up of top-scoring sentences
- **Most sophisticated:** NLP used to synthesize a summary
 - Seldom used in IR; cf. text summarization work.

Dynamic summaries

- Present one or more “windows” within the document that contain several of the query terms
 - “KWIC” snippets: Keyword in Context presentation



[Christopher Manning, Stanford NLP](#)

Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University.

nlp.stanford.edu/~manning/ - 12k - [Cached](#) - [Similar pages](#)



[Christopher Manning, Stanford NLP](#)

Christopher Manning, Associate Professor of Computer Science and Linguistics, ... computational semantics, **machine translation**, grammar induction, ...

nlp.stanford.edu/~manning/ - 12k - [Cached](#) - [Similar pages](#)



[Christopher Manning, Stanford NLP](#)

Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University ... **Chris Manning** works on systems and formalisms that can ...

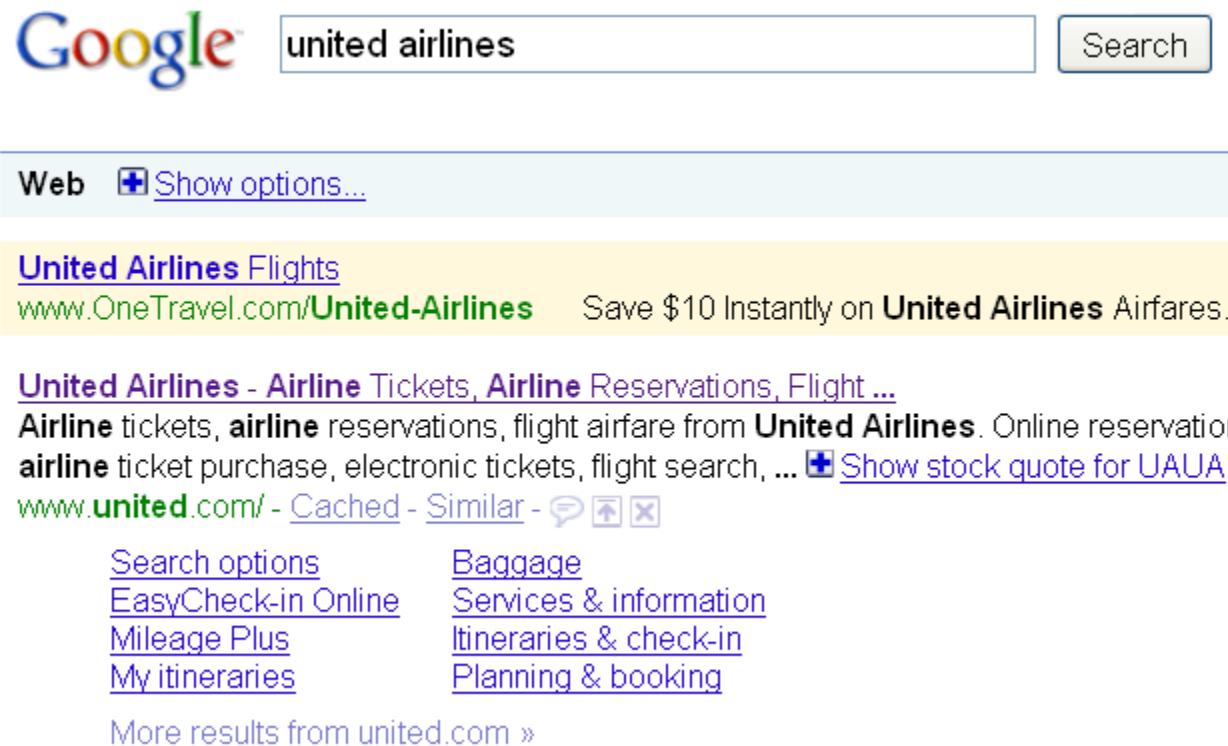
nlp.stanford.edu/~manning - [Cached](#)

Techniques for dynamic summaries

- **Find small windows in doc that contain query terms**
 - Requires fast window lookup in a document cache
- **Score each window wrt query**
 - Use various features such as window width, position in document, etc.
 - Combine features through a scoring function – methodology to be covered later in this course
- **Challenges in evaluation: judging summaries**
 - Easier to do pairwise comparisons rather than binary relevance assessments.

Quicklinks

- For a *navigational query* such as **united airlines** user's need likely satisfied on www.united.com
- Quicklinks provide navigational cues on that home page



Google

Web [+ Show options...](#)

[United Airlines Flights](#)
www.OneTravel.com/United-Airlines Save \$10 Instantly on **United Airlines** Airfares.

[United Airlines - Airline Tickets, Airline Reservations, Flight ...](#)
Airline tickets, **airline** reservations, flight airfare from **United Airlines**. Online reservation **airline** ticket purchase, electronic tickets, flight search, ... [+ Show stock quote for UUA](#)
www.united.com/ - [Cached](#) - [Similar](#) - [🗨](#) [📄](#) [✕](#)

Search options	Baggage
EasyCheck-in Online	Services & information
Mileage Plus	Itineraries & check-in
My itineraries	Planning & booking

[More results from united.com »](#)

united airlines

Search Pad

SearchScan - On

102,000,000 results for united airlines:

Show All

United Air Lines

Wikipedia

Also try: [united airlines reservations](#), [united airlines flight](#), [More...](#)

United Airlines - Airline Tickets, Airline Reservations ... (Nasdaq: [UAUA](#))

Official site for **United Airlines**, commercial air carrier transporting people, property, and mail across the U.S. and worldwide.

[www.united.com](#) - 65k - [Cached](#)

[Planning & Booking](#)

[Shop for Flights](#)

[Itineraries & Check-in](#)

[Special Deals](#)

[Mileage Plus](#)

[Flight Status](#)

[Services & Information](#)

[Customer Service](#)

[more results from united.com »](#)

united airlines



UNITED AIRLINES

[United Airline Fleet](#)

[United Airline Schedule](#)

[United Airlines Reservations](#)

[United Airline Jobs](#)

[Reference](#)

RELATED SEARCHES

[United Airlines Flight Status](#)

[US Airways](#)

[Continental Airlines](#)

ALL RESULTS

[Cheap Flight Tickets](#) · [www.CheapOair.com](#)

CheapOair - The Only Way to Go!! Find Over 18 Million Exclusive Fares.

[Fly United Airlines](#) · [www.OneTravel.com/United-Airline](#)

Save \$10 Instantly on **United Airlines** Flights. Book Now, Hurry!

Best match

[United Airlines - Airline Tickets, Airline Reservations, Flight ...](#)

[www.united.com](#) · Official site

Airline tickets, **airline** reservations, flight airfare from **United Airlines**. Online reservations, **airline** ticket purchase, electronic tickets, flight search, fares and availability ...

[Flights](#)

[Redeem miles](#)

[Check In Online](#)

[Children, pets, & assistance](#)

[My itineraries](#)

[Change your travel plans](#)

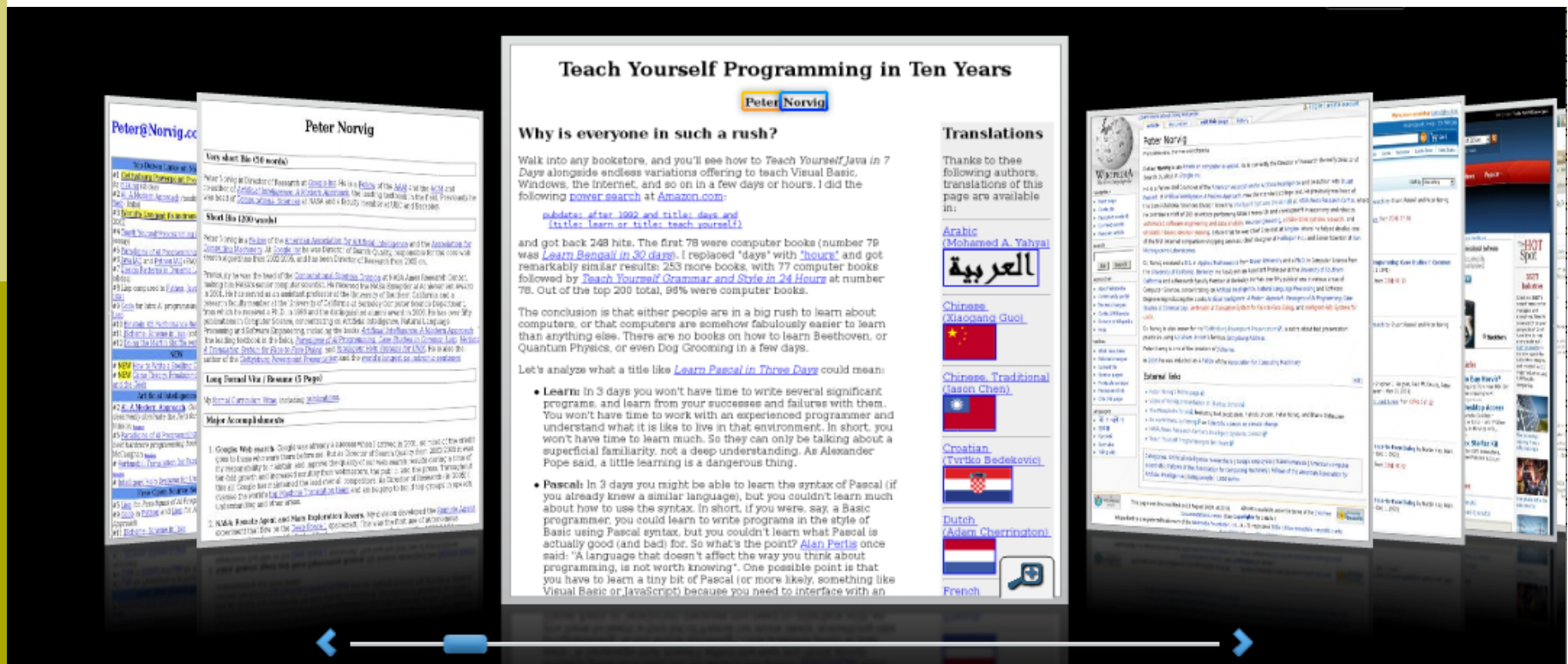
[Baggage](#)

[Special deals](#)

Customer service 800-864-8331

Alternative results presentations?

- An active area of HCI research
- An alternative: <http://www.searchme.com> copies the idea of Apple's Cover Flow for search results
 - (searchme recently went out of business)





Resources for this lecture

- IIR 8