Introduction to Information Retrieval

Evaluation

This lecture

- How do we know if our results are any good?
  - Evaluating a search engine
    - Benchmarks
    - Precision and recall
- Results summaries:
  - Making our good results usable to a user

Evaluating Search Engines

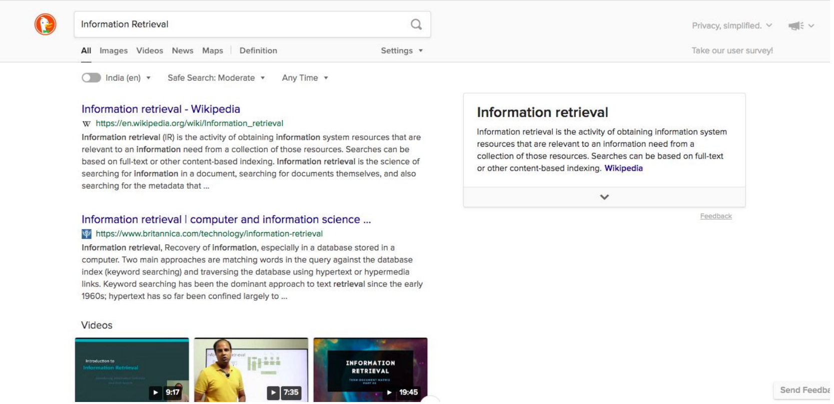
Measures for a search engine

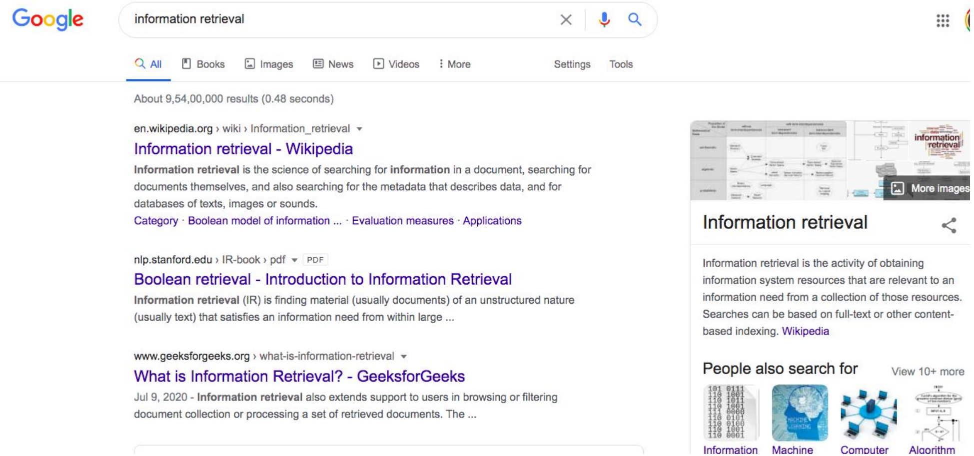
- **How fast does it index**
    - Number of documents/hour
    - (Average document size)
- **How fast does it search**
    - Latency as a function of index size

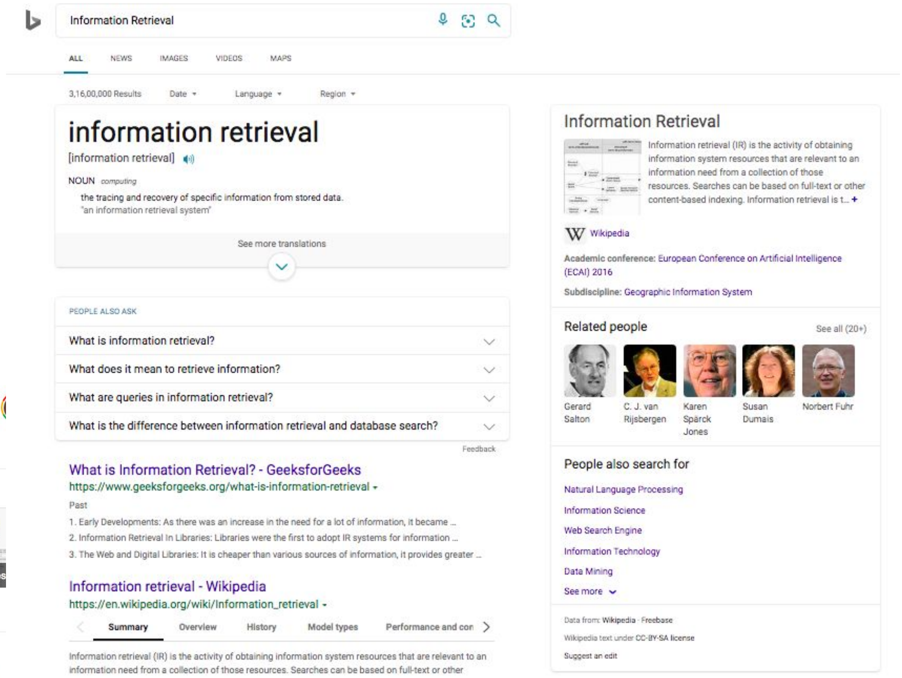
Measures for a search engine

- Expressiveness of query language
    - Ability to express complex information needs
    - Speed on complex queries
- Uncluttered UI
- Is it free?

Uncluttered UI







Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed/size
  - we can make expressiveness precise
- 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

Measuring user happiness

- 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.

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

Happiness: elusive to measure

- Relevance measurement requires 3 elements:
  - A benchmark document collection
  - A benchmark suite of queries
  - 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

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.*

Evaluating an IR system

- 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

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?

Accuracy

- (tp + tn) / (tp + fp + fn + tn)
- 5 +5000/ (5+5 +5 + 5000)

	Relevant	Nonrelevant
Retrieved	tp 5	fp 5
Not Retrieved	fn 5	tn 5000

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 (low inter annotator agreement)
- Assessments have to be binary
- Heavily skewed by collection/authorship
    - Results may not translate from one domain to another

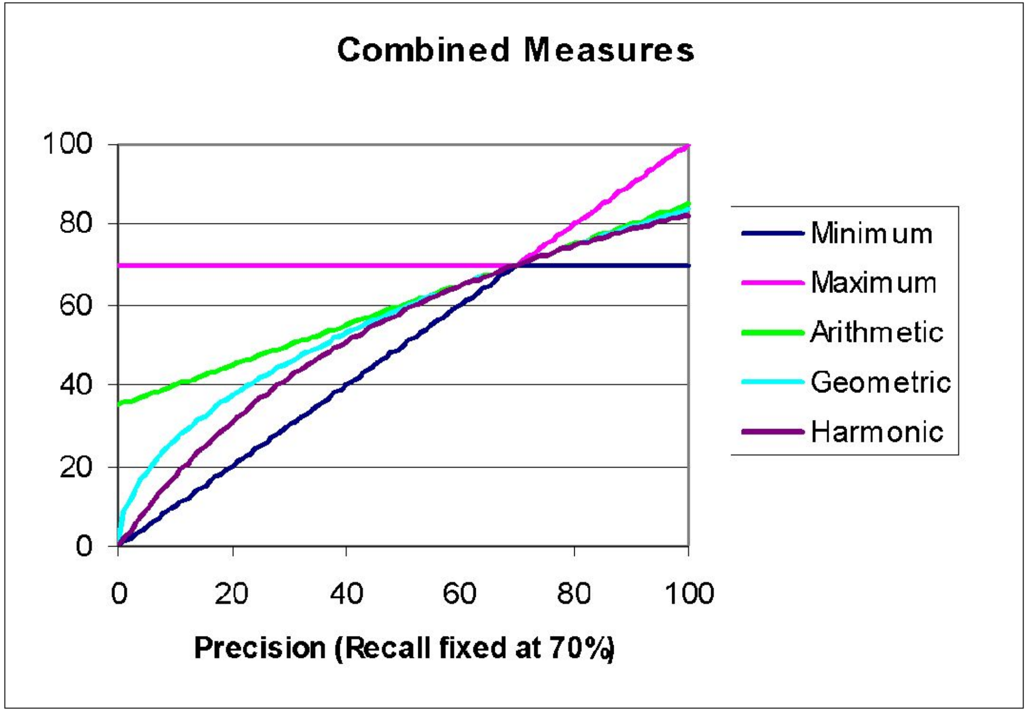
A combined measure: F

- 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 F1 measure
- i.e., with β = 1 or α = ½
- Harmonic mean is a conservative average

F1 and other averages



Evaluating ranked results

- Evaluation of ranked results:
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a precision-recall curve

A precision-recall curve

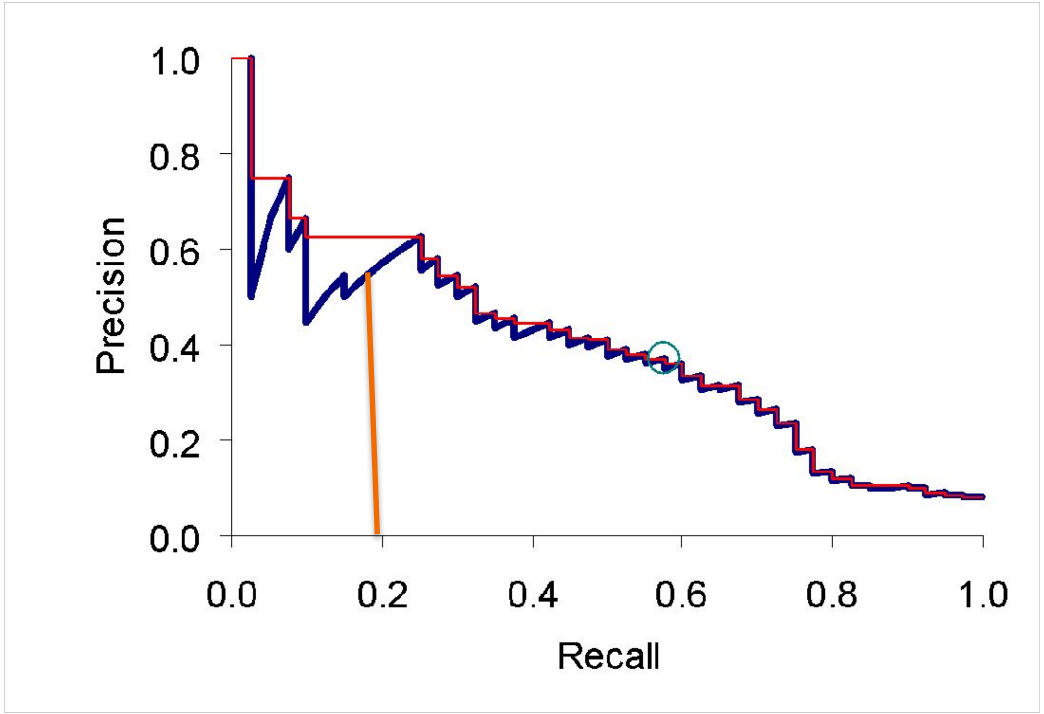
Relevance

1 0 1 1 0 1 0 0 0 1 0 1 1

Precision

100,50,66,75,60,66,58,50,44,

50, 45, 50, (recall = 0.2) 53

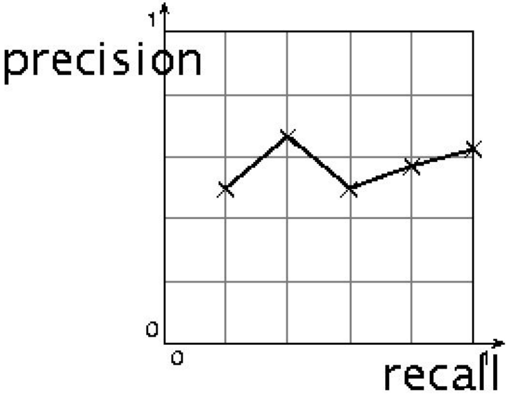


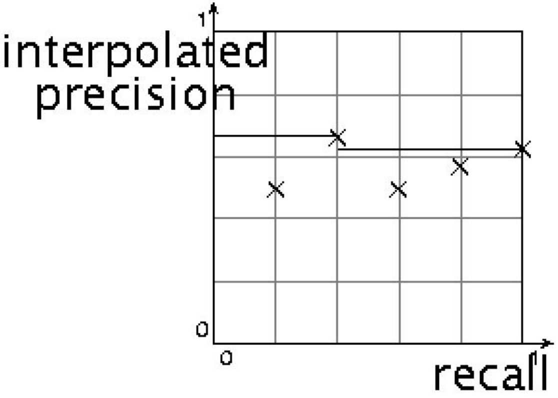
Averaging over queries

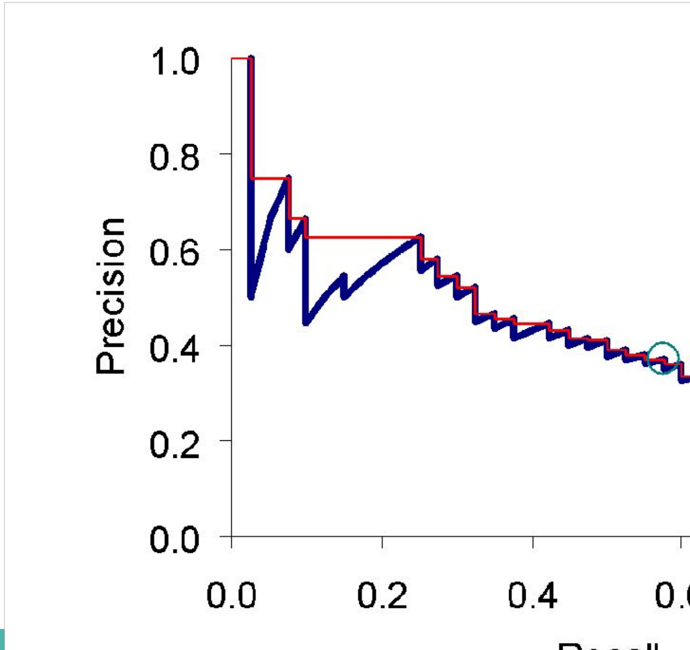
- A precision-recall graph for one query isn't a very sensible thing to look at
- You need to average performance over a whole bunch of queries.
- But there's a technical issue:
  - Precision-recall calculations place some points on the graph
  - How do you determine a value (interpolate) between the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

Interpolated precision

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







Evaluation

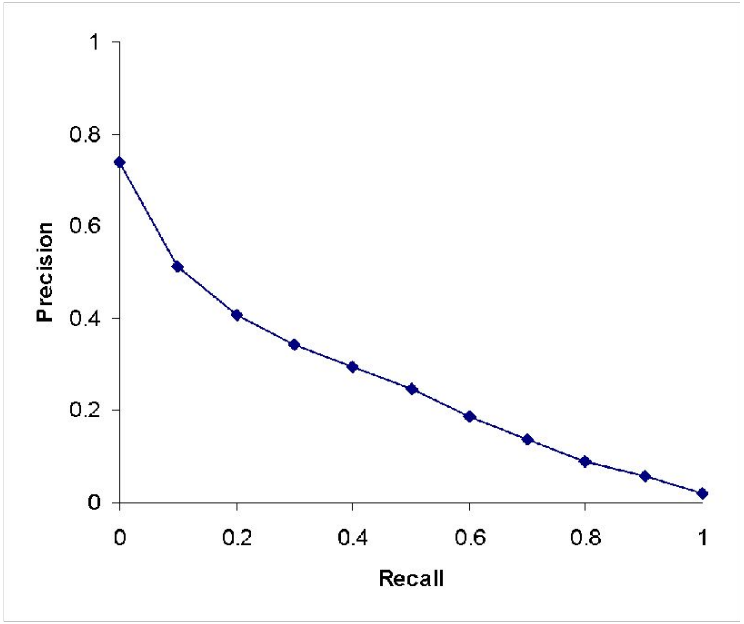
- Graphs are good, but people want summary measures!
    - Precision at fixed retrieval level
        - Precision@k: Precision of top k results
        - Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages
        - But: averages badly and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

Evaluation

- 11-point interpolated average precision
    - The standard measure in the early TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (the value for 0 is always interpolated!), and average them
    - Evaluates performance at all recall levels

Typical (good) 11 point precisions

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



Yet more evaluation measures...

- 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

MAP

the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \ldots d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then

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

MAP

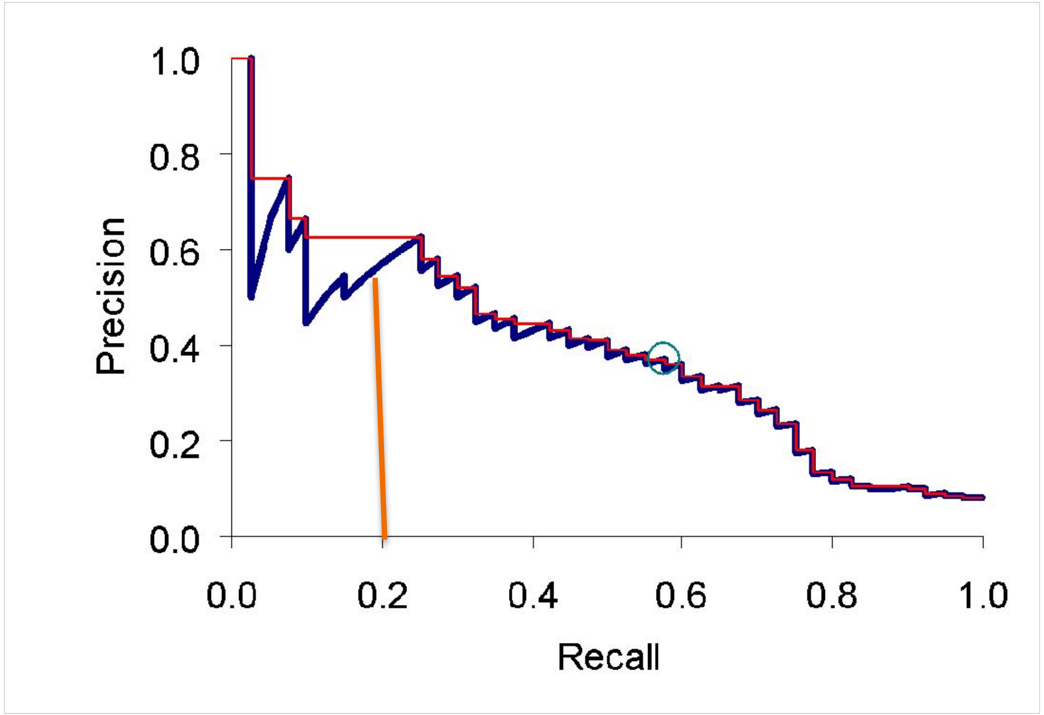
Relevance

1 0 1 1 0 1 0 0 0 1 0 1 1

Precision

100,50,66,75,60,66,58,50,44,

50, 45, 50, (recall = 0.2) 53



MAP = (100 + 66 + 75 + 66 + 50 + 50)/6=67.8

Yet more evaluation measures...

- Precision@k, Unstable, not possible to average, as one query may have more relevant document than other.
- Precision at K
  - Measures factor in precision at all recall level, that is measuring precision at all low level of retrieved level
  - Advantage : Does not require any estimate of the size of the set of relevant documents
  - Disadvantage: It is the least stable of the commonly used evaluation measures and that it does not average well

Yet more evaluation measures...

R-precision

- If have known (though perhaps incomplete) set of relevant documents of size |Rel|, then calculate precision of top |Rel| docs returned
- precision = r/|Rel| where r is the number of relevant documents returned. Recall = Precision,
- Perfect system could score 1.0.

- Break-Point: When precision and recall is same. R-precision is same as break-point.

NDCG (NORMALIZED DISCOUNTED CUMULATIVE GAIN)

lative gain (*NDCG*). NDCG is designed for situations of non-binary notions of relevance (cf. Section 8.5.1). 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 assessors gave to document d for query j. Then,

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

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'.

NDCG

$$\sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)},$$

Relevance : 1 0 0 1

Cumulative Gain : 1 + 0 + 0 + 1 = 2

Discounted CG (DCG) : $1/\log_2(1+1) + 0 + 0 + 1/\log_2(1+4) = 1.43$

Ideal Relevance : 1 1 0 0

Ideal DCG : $1/\log_2(1+1) + 1/\log_2(1+2) + 0 + 0 = 1.63$

NDCG = DCG/ideal DCG = 1.43 / 1.63 = **0.877**

1 1 0 0 : Is equivalent to the ideal relevance, so its DCG = 1.63

NDCG = 1.63 / 1.63 = **1.0**

Variance

- For a test collection, it is usual that a system does 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 Evaluations

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 germane (relevant) to docs available
  - Best designed by domain experts
  - Random query terms generally not a good idea

From document collections to test collections

- Relevance assessments
  - Human judges, time-consuming
  - Are human panels perfect?

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.

Kappa Measure: Example

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

P(A)? P(E)?

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

Kappa Example

- Kappa > 0.8 = good agreement
- 0.67 < Kappa < 0.8 -> “tentative conclusions” (Carletta ’96)
- Depends on purpose of study
- For >2 judges: average pairwise kappas

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.

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.

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 approximate vector space retrieval, we can compare the cosine distance closeness of the closest docs to those found by an 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
- Search engines often use precision at top k, 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)

Evaluation at large search engines

- 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

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

A/B testing

- 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
- In principle less powerful than doing a multivariate regression analysis, but easier to understand

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"

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

Summaries

- 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

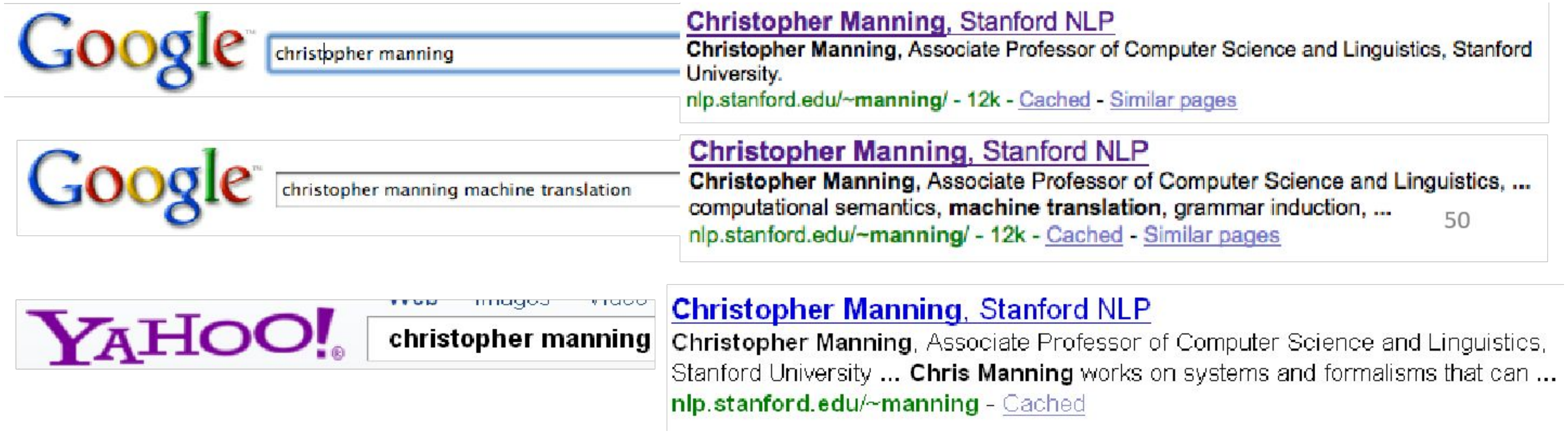
- In typical systems, the static summary is a subset of the document
- Simplest heuristic: the first 50 (or so – this can be varied) words of the document
  - Summary cached at indexing time

Static summaries

- More sophisticated: extract from each document a set of “key” sentences
  - Simple NLP heuristics to score each sentence
  - A sentence score can be sum of idfs of its individual words.
  - Summary is made up of top-scoring sentences. (5 lines)
- 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

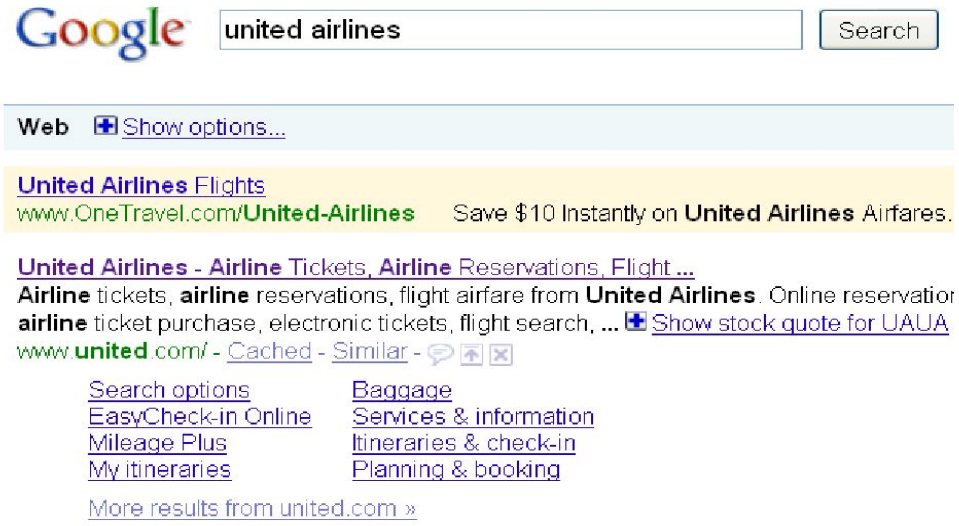


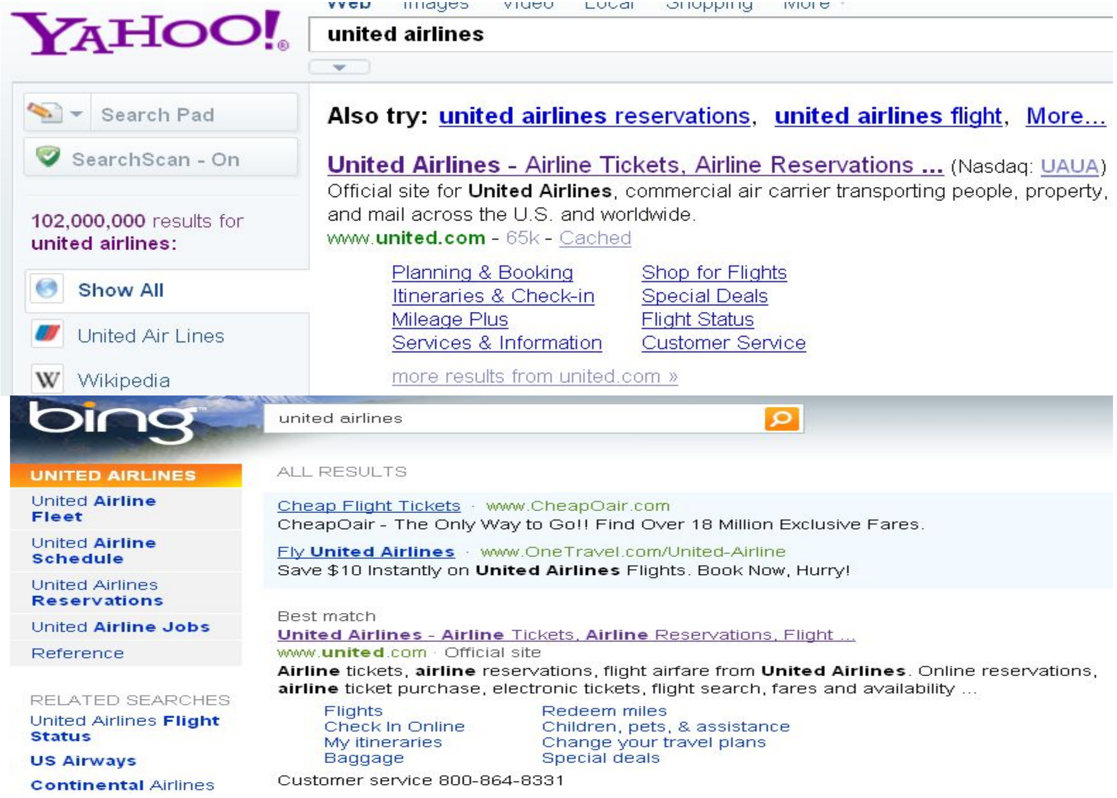
Techniques for dynamic summaries

- 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

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





Web Images Video Local Shopping More
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
Itineraries & Check-in
Mileage Plus
Services & Information
Shop for Flights
Special Deals
Flight Status
Customer Service
more results from united.com »
bing
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
Check In Online
My itineraries
Baggage
Redeem miles
Children, pets, & assistance
Change your travel plans
Special deals
Customer service 800-864-8331

Alternative results presentations?

- An active area of HCI research