Search Evaluation

Tao Yang

CS293S

Slides partially based on text book [CMS] [MRS]

Table of Content

- **Search Engine Evaluation**
- **Metrics for relevancy**
  - Precision/recall
  - F-measure
  - MAP
  - NDCG

Difficulties in Evaluating IR Systems

- **Effectiveness is related to the *relevancy* of retrieved items.**
- **Relevancy is not typically binary but continuous. Not easy to judge**
- **Relevancy, from a human standpoint, is:**
  - Subjective/cognitive: Depends upon user's judgment, human perception and behavior
  - Situational and dynamic:
    - Relates to user's current needs. Change over time.
  - E.g.
    - CMU. US Open. Etrade.
    - Red wine or white wine

Measuring user happiness

- **Issue: who is the user we are trying to make happy?**
- **Web engine: user finds what they want and return to the engine**
  - Can measure rate of return users
- **eCommerce site: user finds what they want and make a purchase**
  - Is it the end-user, or the eCommerce site, whose happiness we measure?
  - Measure time to purchase, or fraction of searchers who become buyers?

Aspects of Search Quality

- **Relevancy**
- **Freshness& coverage**
  - Latency from creation of a document to time in the online index. (Speed of discovery and indexing)
  - Size of database in covering data coverage
- **User effort and result presentation**
  - Work required from the user in formulating queries, conducting the search
  - Expressiveness of query language
  - Influence of search output format on the user’s ability to utilize the retrieved materials.

System Aspects of Evaluation

- **Response time:**
  - Time interval between receipt of a user query and the presentation of system responses.
  - Average response time
    - at different traffic levels (queries/second)
    - When # of machines changes
    - When the size of database changes
    - When there is a failure of machines
- **Throughputs**
  - Maximum number of queries/second that can be handled
    - without dropping user queries
    - Or meet Service Level Agreement (SLA)
      - For example, 99% of queries need to be completed within a second.
  - How does it vary when the size of database changes

System Aspects of Evaluation

- **Others**
  - Time from crawling to online serving.
  - Percentage of results served from cache
  - Stability: number of abnormal response spikes per day or per week.
  - Fault tolerance: number of failures that can be handled.
  - Cost: number of machines needed to handle
    - different traffic levels
    - host a DB with different sizes

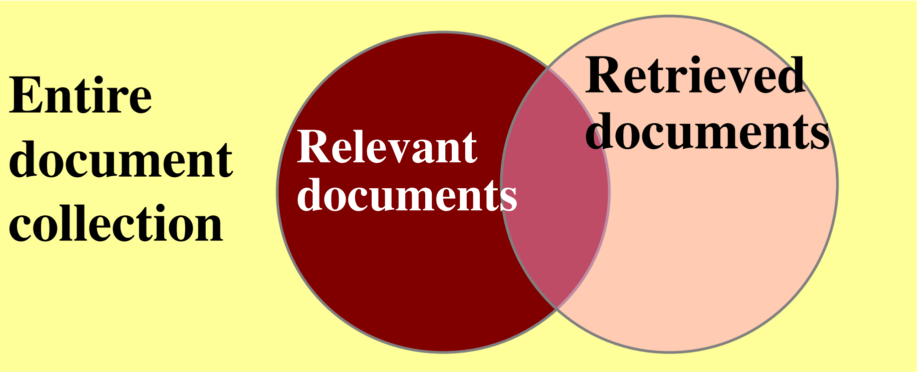
Unranked retrieval evaluation: Precision and Recall

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

	Relevant	Not Relevant
Retrieved	tp (True positive)	fp
Not Retrieved	fn	tn

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

Precision and Recall: Another View



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

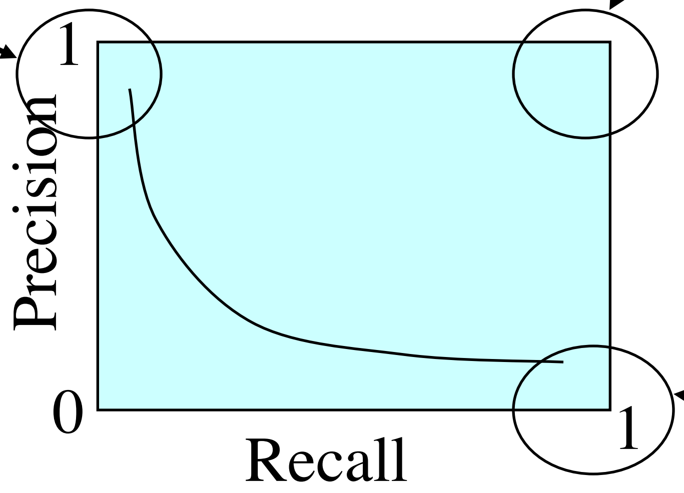
$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Determining Recall is Difficult

- **Total number of relevant items is sometimes not available:**
  - Use queries that only identify few rare documents known to be relevant

Trade-off between Recall and Precision



Returns relevant documents but misses many useful ones too

The ideal

Returns most relevant documents but includes lots of junk

F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

E Measure (parameterized F Measure)

- **A variant of F measure that allows weighting emphasis on precision over recall:**

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

- **Value of β controls trade-off:**
  - β = 1: Equally weight precision and recall (E=F).
  - β > 1: Weight precision more.
  - β < 1: Weight recall more.

Computing Recall/Precision Points for Ranked Results

- For a given query, produce the ranked list of retrievals.
- Mark each document in the ranked list that is relevant according to the gold standard.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.

R- Precision (at Position R)

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

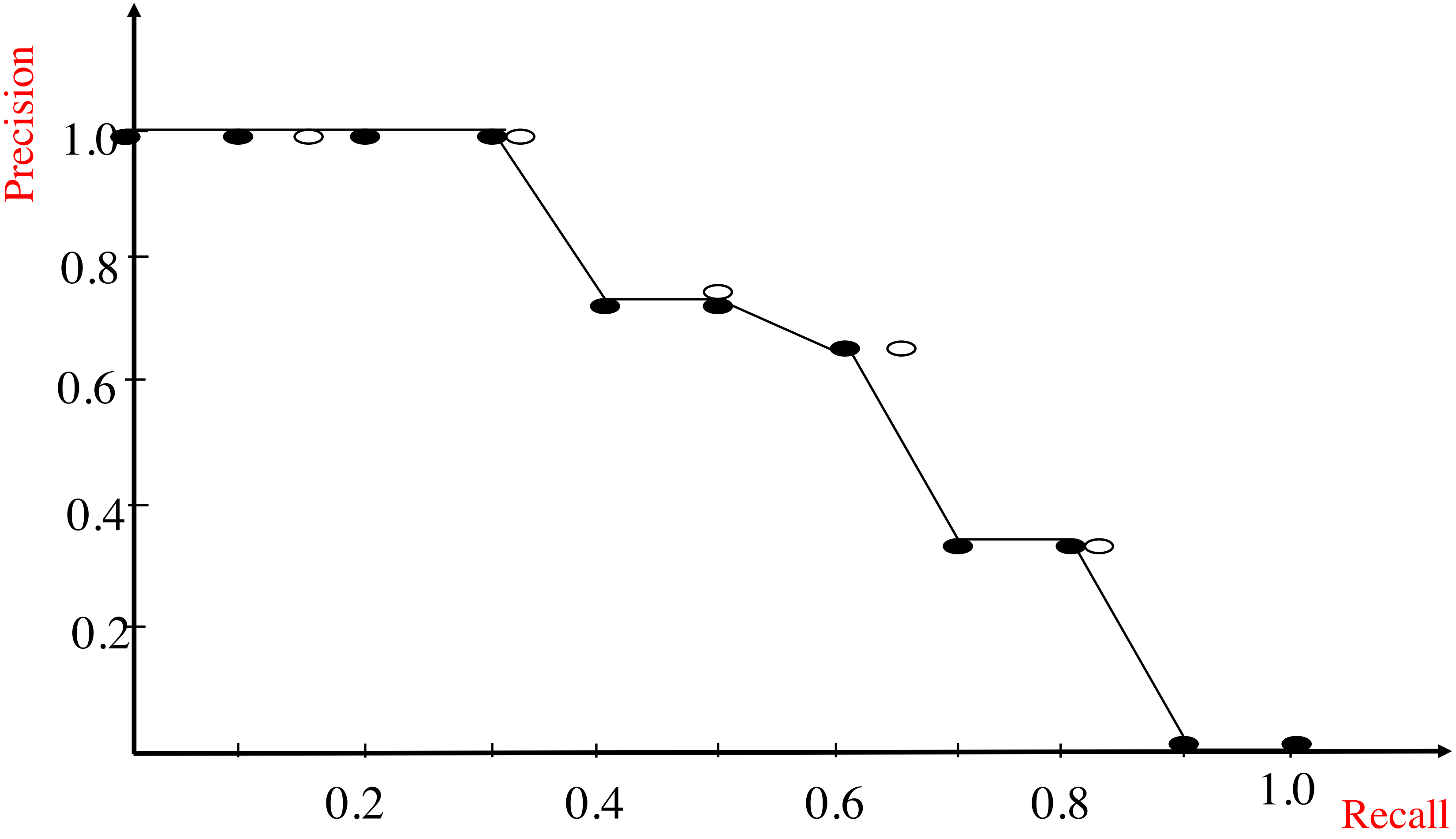
R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document.
Never reach 100% recall

Interpolating a Recall/Precision Curve: An Example



Averaging across Queries: MAP

- ***Mean Average Precision*** **(MAP)**
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers
  - assumes user is interested in finding many relevant documents for each query
  - requires many relevance judgments in text collection

MAP Example:

■ ■ ■ ■ ■ = relevant documents for query 1

Ranking #1	■	□	■	□	□	■	□	□	■	■
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

■ ■ ■ = relevant documents for query 2

Ranking #2	□	■	□	□	■	□	■	□	□	□
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Discounted Cumulative Gain

- **Popular measure for evaluating web search and related tasks**
- **Two assumptions:**
  - Highly relevant documents are more useful than marginally relevant document
    - Support relevancy judgment with multiple levels
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined
- **Gain is *discounted*, at lower ranks, e.g. 1/*log (rank)***
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

- ***DCG*** **is the total gain accumulated at a particular rank *p*:**

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- **Alternative formulation:**

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

DCG Example

- **10 ranked documents judged on 0-3 relevance scale:**

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- **discounted gain:**

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- **DCG@1, @2, etc:**

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Normalized DCG

- **DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking***
  - *Example:*
    - *DCG@5 = 6.89*
    - *Ideal DCG@5=9.75*
    - *NDCG@5=6.89/9.75=0.71*
- **NDCG numbers are averaged across a set of queries at specific rank values**

NDCG Example with Normalization

- **Perfect ranking:**

  3, 3, 3, 2, 2, 2, 1, 0, 0, 0

- **Ideal DCG@1, @2, …:**

  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10

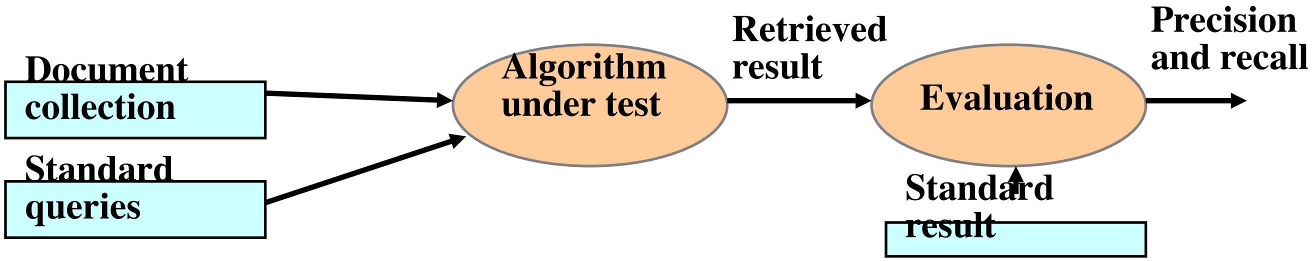
- **NDCG@1, @2, …**
  - normalized values (divide actual by ideal):

  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88
  - NDCG $\leq$ 1 at any rank position

CREATING TEST COLLECTIONS FOR IR EVALUATION

Relevance benchmarks

- **Relevant measurement requires 3 elements:**
  1. A benchmark document collection
  2. A benchmark suite of queries
  3. Editorial assessment of query-doc pairs
     - Relevant vs. non-relevant
     - Multi-level: Perfect, excellent, good, fair, poor, bad



- **Public benchmarks**
  - TREC: http://trec.nist.gov/
  - Microsoft/Yahoo published learning benchmarks

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**

From document collections to test collections

- **Still need**
  - Test queries
  - Relevance assessments
- **Test queries**
  - Must be germane to 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?

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**
  - Relative observed agreement of judges
- **P(E) – what agreement would be by chance**
  - hypothetical probability of chance agreement
- **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(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

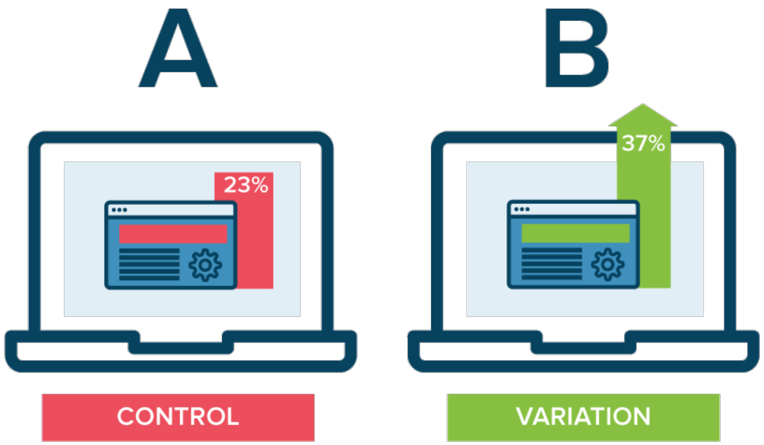
- Kappa > 0.8 = good agreement
- 0.67 < Kappa < 0.8 -> "tentative conclusions" (Carletta '96)
- Depends on purpose of study
- For >2 judges: average pairwise kappas

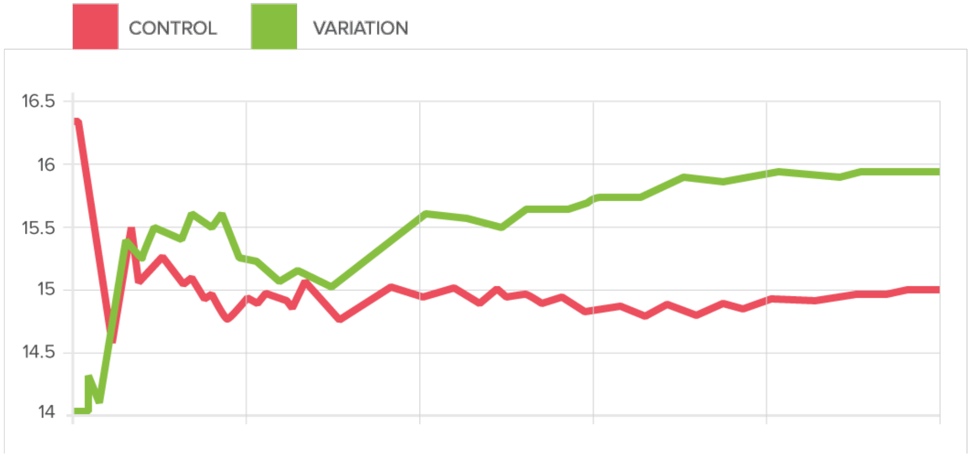
Can we avoid human judgment?

- **No**
  - But once we have test collections, we can reuse them (so long as we don't overtrain too badly)
  - Makes experimental work hard
    - Especially on a large scale
- **In some very specific settings, can use proxies**
  - E.g.: for approximate vector space retrieval, use cosine distance closeness
- **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 (variation)
- Prerequisite: Website with large traffic
- 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
  - Clickthrough.
  - Now we can directly see if the innovation (variation) does improve user happiness.

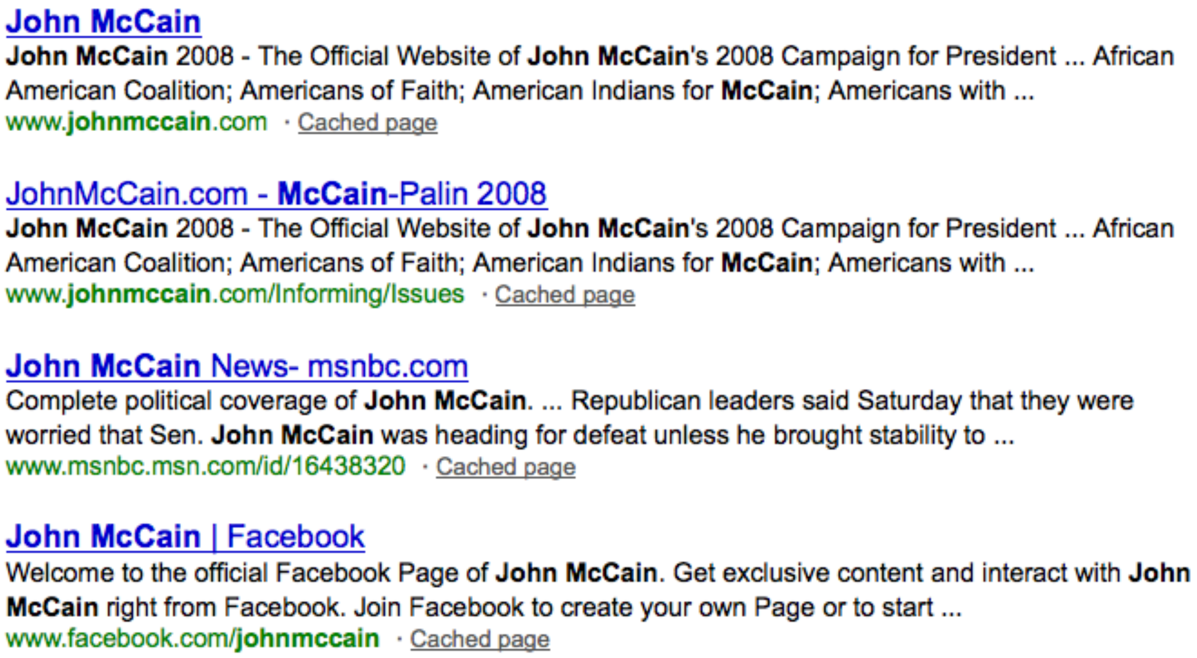




RESULTS PRESENTATION

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

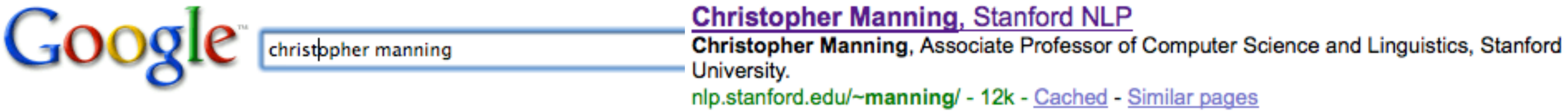
- **The title is often automatically extracted from document metadata. What about the summaries?**
  - This description is crucial.
  - User can identify good/relevant hits based on description.
- **Two basic kinds:**
  - Static
  - Dynamic
- **A static summary of a document is always the same, regardless 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

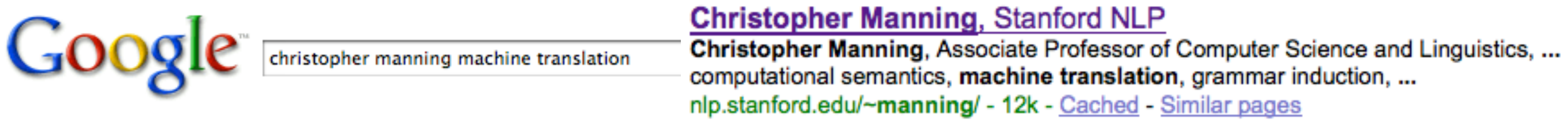
Dynamic summaries

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



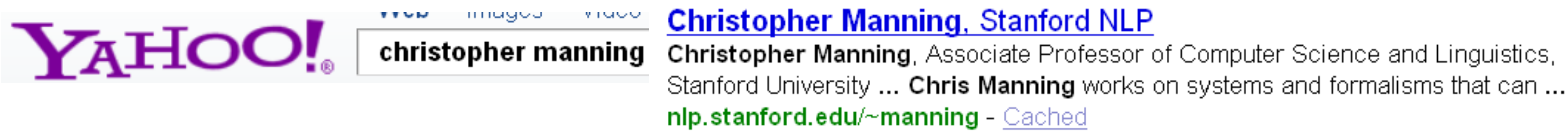
christopher manning

[**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 machine translation

[**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

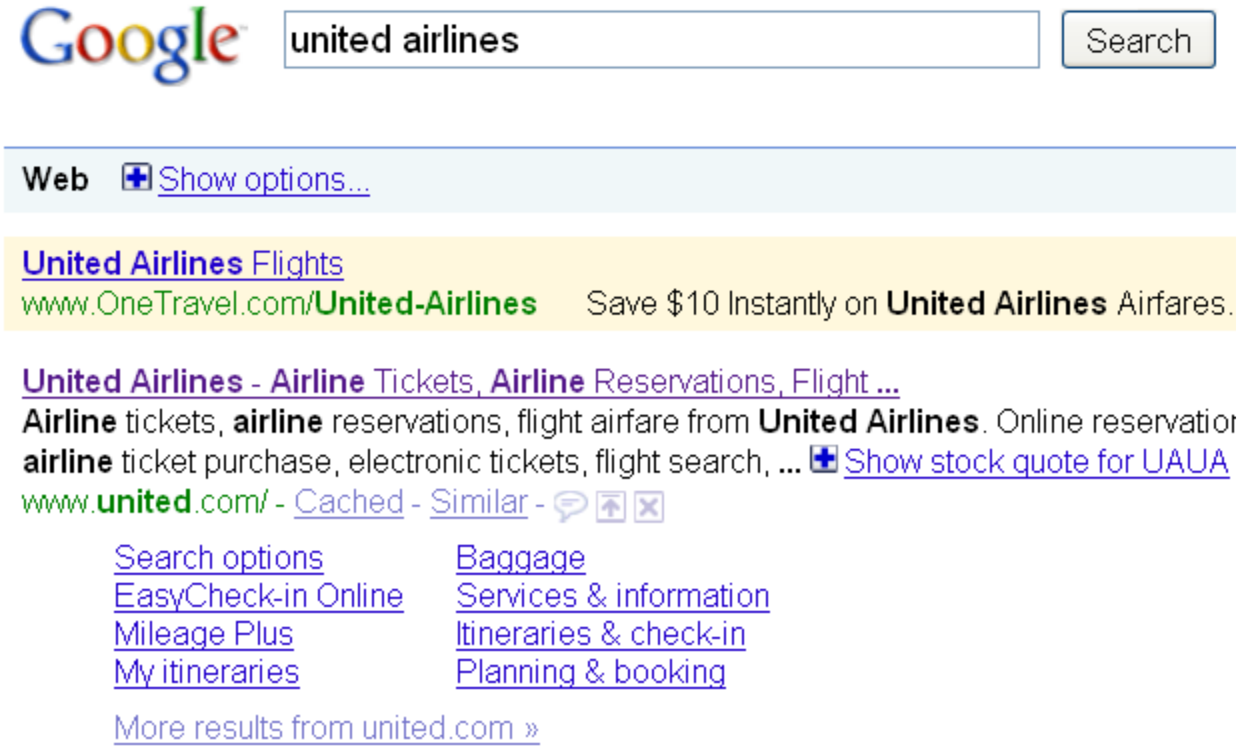
[**Christopher Manning**, Stanford NLP](#)
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University ... **Chris Manning** works on systems and formalisms that can ...
nlp.stanford.edu/~**manning** - Cached

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 –
  - 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

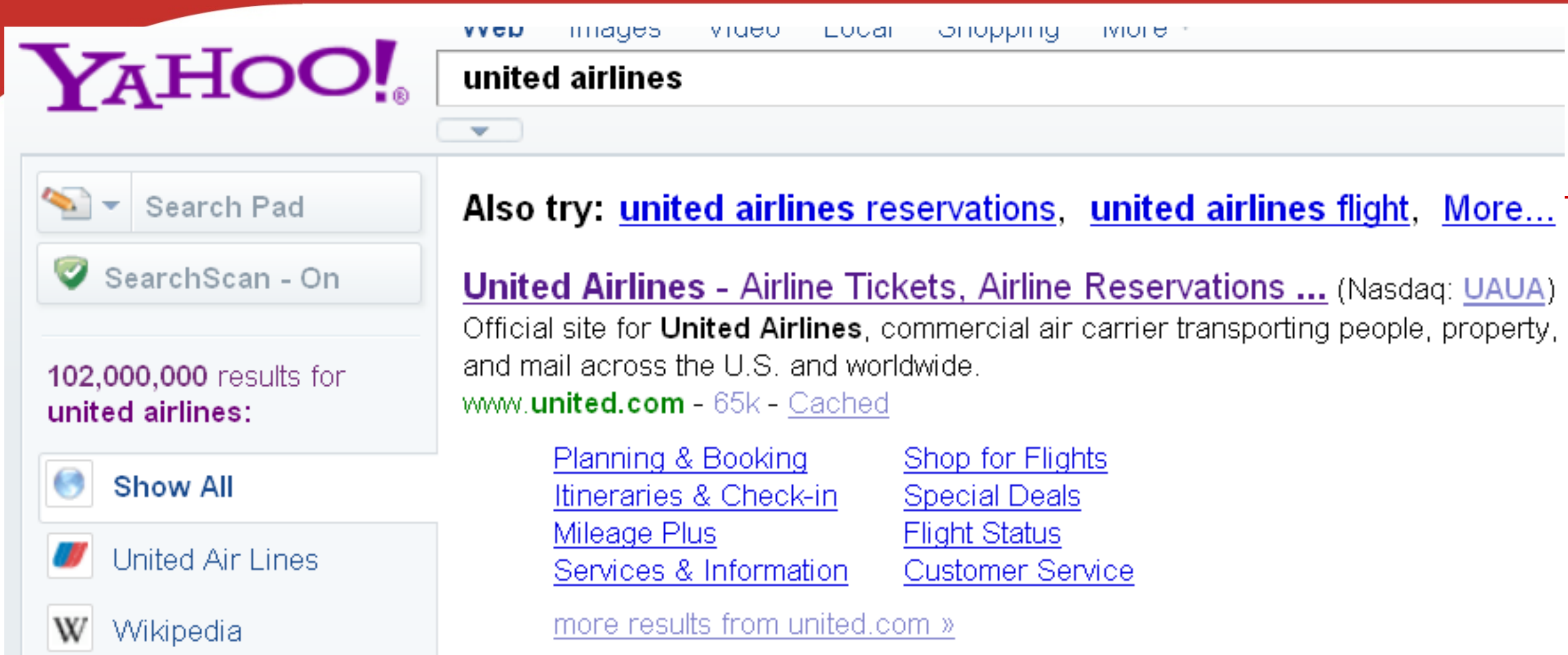


United Airlines Flights
www.OneTravel.com/United-Airlines Save $10 Instantly on United Airlines Airfares.

United Airlines - Airline Tickets, Airline Reservations, Flight ...
Airline tickets, airline reservations, flight airfare from United Airlines. Online reservatio
airline ticket purchase, electronic tickets, flight search, ... Show stock quote for UAUA
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

Also try: united airlines reservations, united airlines flight, More...

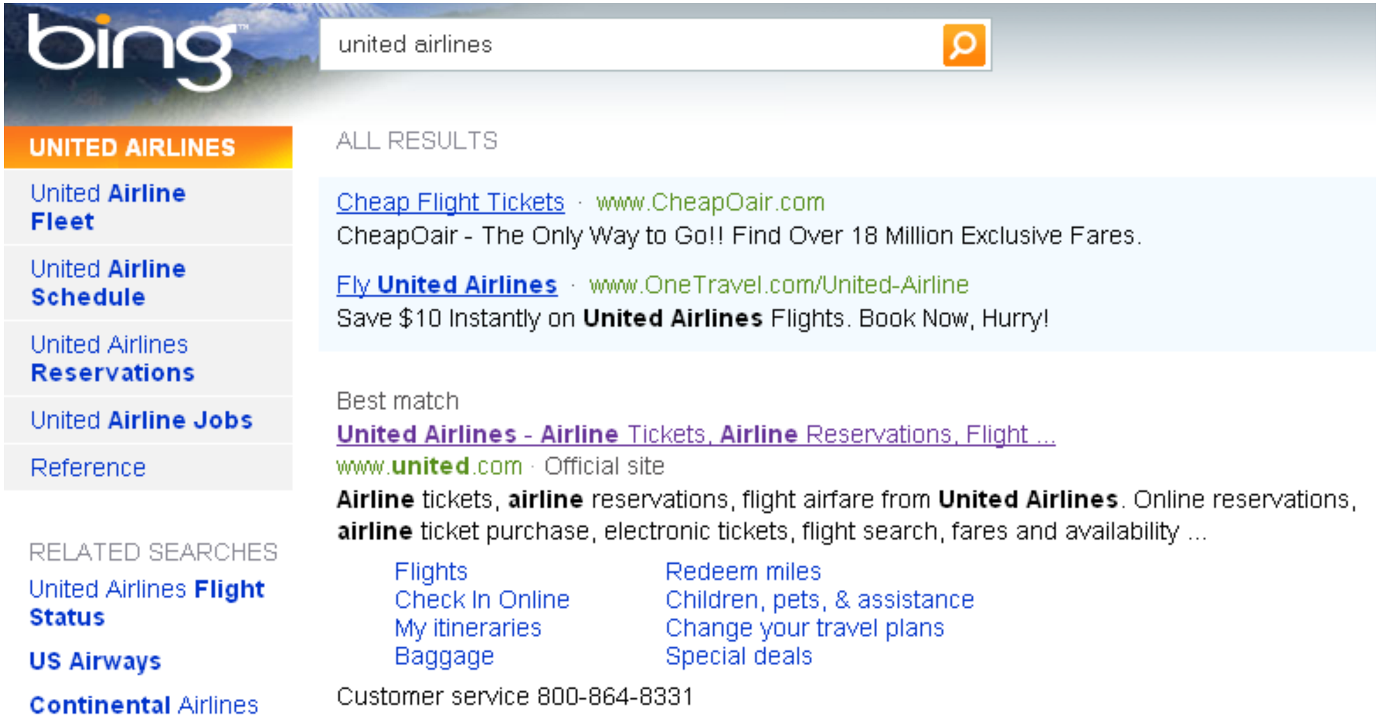
United Airlines - Airline Tickets, Airline Reservations ... (Nasdaq: UAUA)
Official site for **United Airlines**, commercial air carrier transporting people, property, and mail across the U.S. and worldwide.
www.**united.com** - 65k - Cached

Planning & Booking	Shop for Flights
Itineraries & Check-in	Special Deals
Mileage Plus	Flight Status
Services & Information	Customer Service

more results from united.com »

102,000,000 results for **united airlines:**

Show All
United Air Lines
Wikipedia



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.

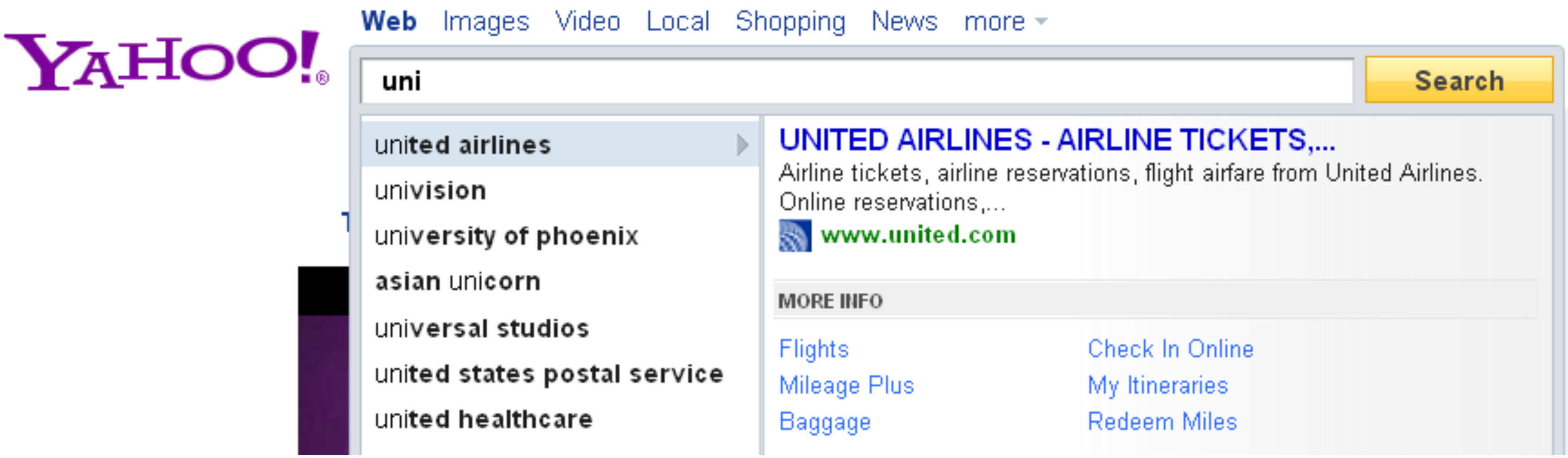
Fly **United Airlines** · www.OneTravel.com/United-Airline
Save $10 Instantly on **United Airlines** Flights. Book Now, Hurry!

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?



uni

united airlines
univision
university of phoenix
asian unicorn
universal studios
united states postal service
united healthcare

UNITED AIRLINES - AIRLINE TICKETS,...
Airline tickets, airline reservations, flight airfare from United Airlines. Online reservations,...
www.united.com

MORE INFO

Flights	Check In Online
Mileage Plus	My Itineraries
Baggage	Redeem Miles