Introduction to

Information Retrieval

CS276

Information Retrieval and Web Search

Pandu Nayak and Prabhakar Raghavan

Lecture 8: 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
- 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
- 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 s/he wants and returns to the engine
    - Can measure rate of return users
  - User completes task – search as a means, not end
  - See Russell http://dmrussell.googlepages.com/JCDL-talk-June-2007-short.pdf

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

Evaluating an IR system

- Note: the **information need** is translated into a **query**
- Relevance is assessed relative to the **information need** *not* the **query**
- 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

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)

A combined measure: *F*

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

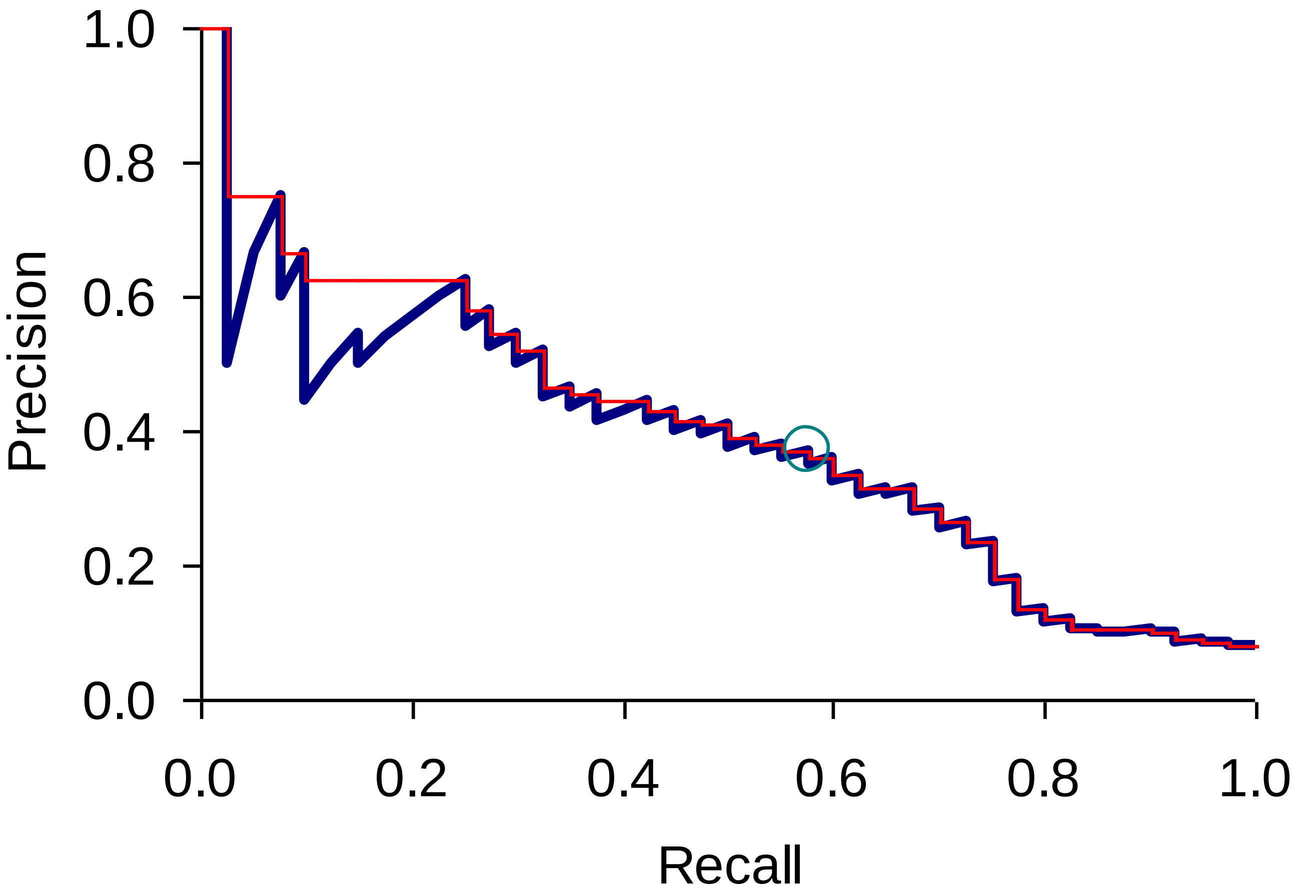
$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ or $\alpha = ½$
- Harmonic mean is a conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

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



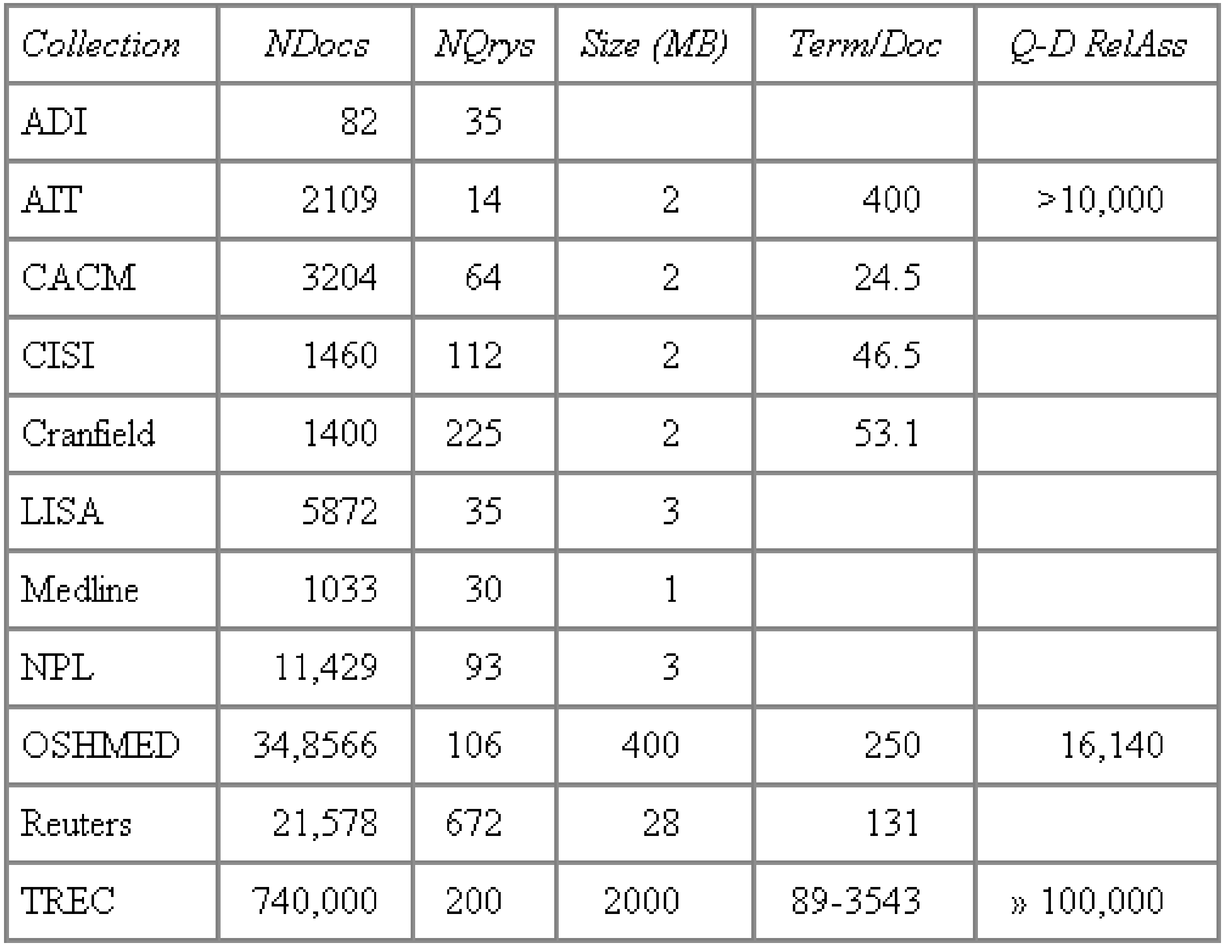
Averaging over queries

- 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.

CREATING TEST COLLECTIONS FOR IR EVALUATION

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 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
- P(E) – what agreement would be by chance

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

- 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

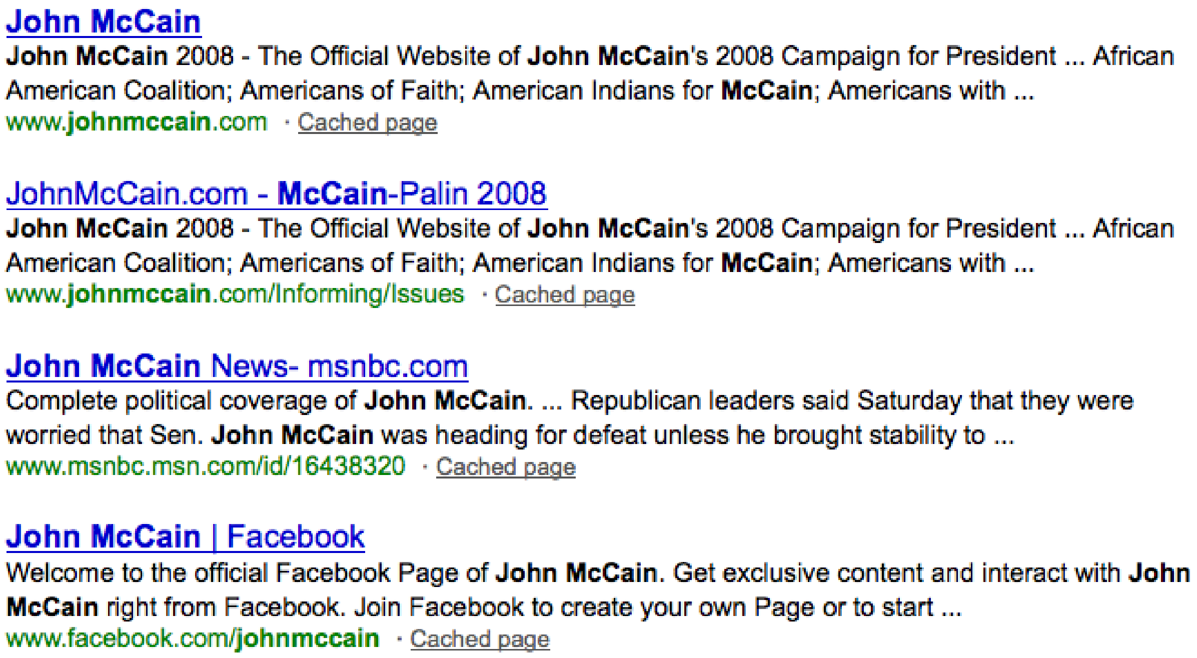
Evaluation at large search engines

- 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)
- 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

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

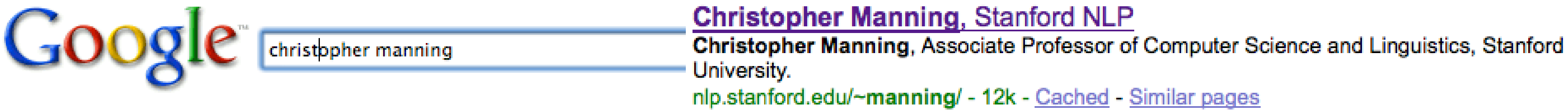
- The title is often automatically extracted from document metadata. What about the summaries?
  - This description is crucial.
  - User can identify good/relevant hits based on description.
- Two basic kinds:
  - Static
  - Dynamic
- A **static summary** of a document is always the same, regardless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

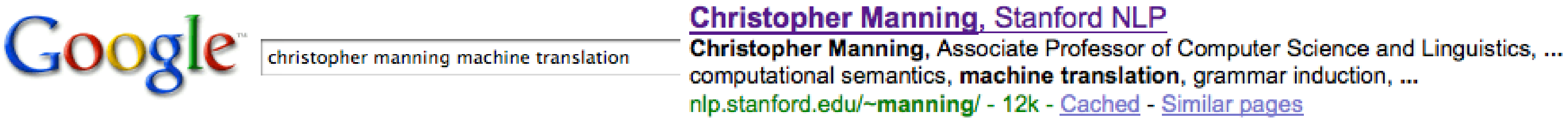
Dynamic summaries

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



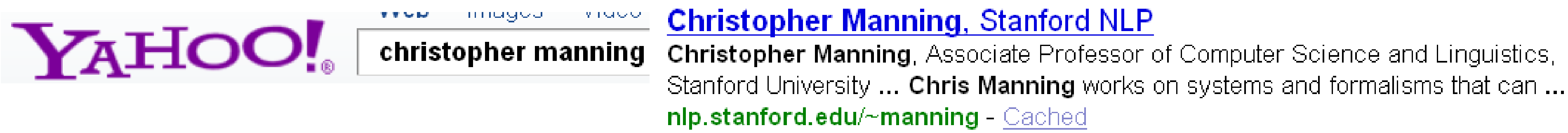
christopher manning

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

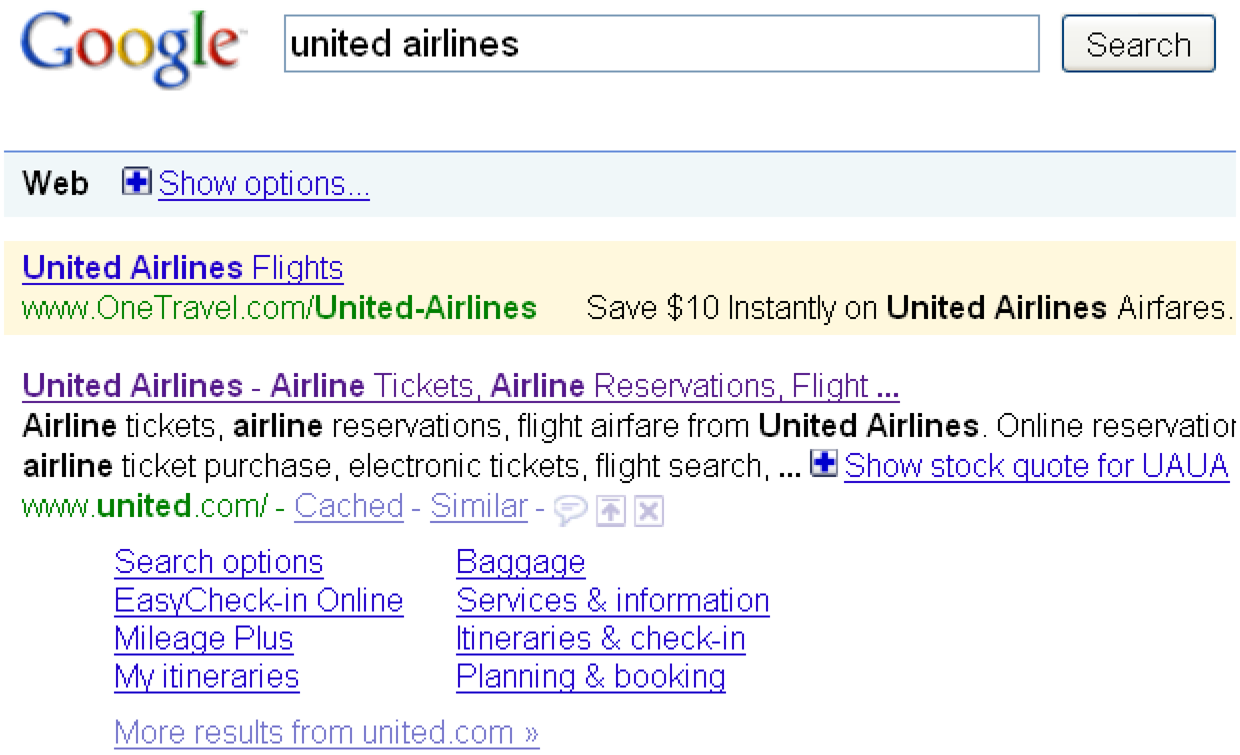
Christopher Manning, Stanford NLP
Christopher Manning, Associate Professor of Computer Science and Linguistics, Stanford University ... **Chris Manning** works on systems and formalisms that can ...
nlp.stanford.edu/~manning - Cached

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 – methodology to be covered Nov 12th
- 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 | united airlines | Search

Web ⊞ Show options...

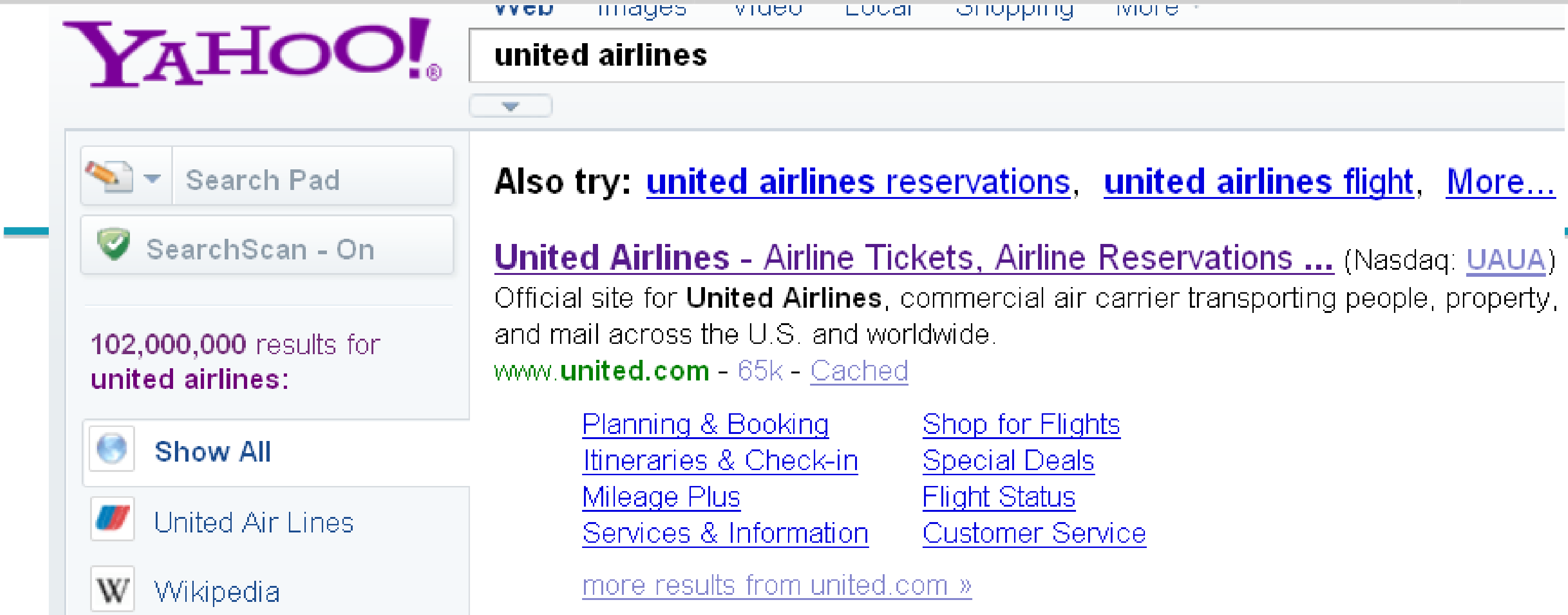
United Airlines Flights
www.OneTravel.com/United-Airlines Save $10 Instantly on **United Airlines** Airfares.

United Airlines - Airline Tickets, Airline Reservations, Flight ...
Airline tickets, **airline** reservations, flight airfare from **United Airlines**. Online reservatio
airline ticket purchase, electronic tickets, flight search, **...** ⊞ Show stock quote for UAUA
www.**united**.com/ - Cached - Similar -

Search options
EasyCheck-in Online
Mileage Plus
My itineraries

Baggage
Services & information
Itineraries & check-in
Planning & booking

More results from united.com »



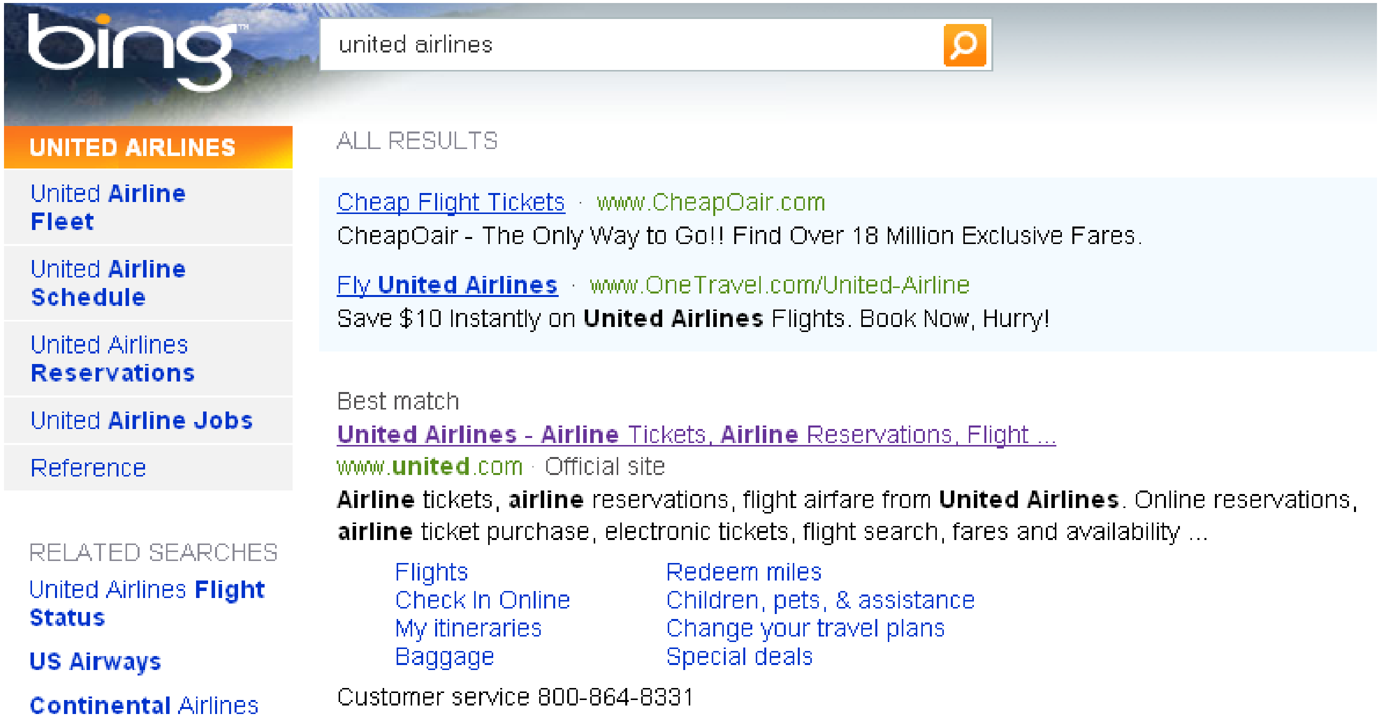
102,000,000 results for **united airlines:**

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 »



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.

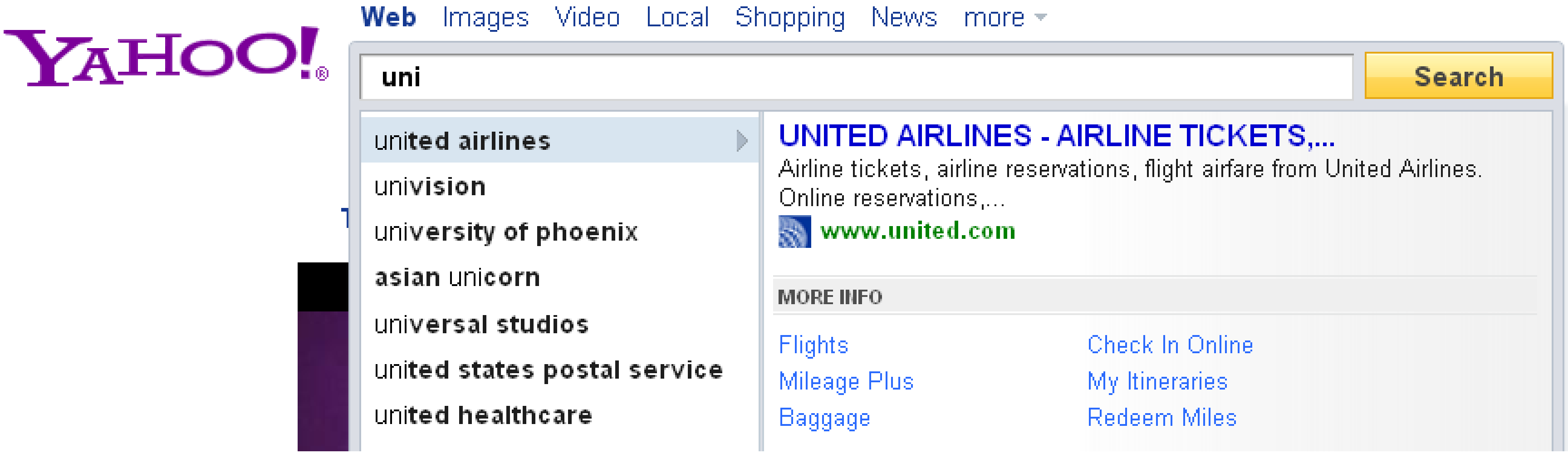
Fly **United Airlines** · www.OneTravel.com/United-Airline
Save $10 Instantly on **United Airlines** Flights. Book Now, Hurry!

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

Resources for this lecture

- IIR 8
- MIR Chapter 3
- MG 4.5
- Carbonell and Goldstein 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR 21.