Web Search:
Techniques, algorithms and Aplications

Basic Techniques
for Web Search

German Rigau <german.rigau@ehu.es>

[Based on slides by Eneko Agirre …
and Christopher Manning and Prabhakar Raghavan]

Basic Techniques for Web Search

- Review of applications
- Basic Techniques in detail:
  - Boolean search
  - Vocabularies, dictionaries, index
  - **Scoring, evaluation, complete system**
  - Web search
- Semantic search

Complete system (Chap. 7)

- Putting together a complete search system
  - Will require learning about a number of miscellaneous topics and heuristics

Static quality scores

- We want top-ranking documents to be both *relevant* and *authoritative*
- *Relevance* is being modeled by cosine scores
- *Authority* is typically a query-independent property of a document
- Examples of authority signals
  - Wikipedia among websites
  - Articles in certain newspapers
  - A paper with many citations
  - Many diggs, Y!buzzes or del.icio.us marks
  - (Pagerank)

Quantitative

Modeling authority

- Assign to each document a *query-independent* quality score in [0,1] to each document *d*
  - Denote this by *g(d)*
- Thus, a quantity like the number of citations is scaled into [0,1]

Net score

- Consider a simple total score combining cosine relevance and authority
- net-score(q,d) = $g(d)$ + cosine(q,d)
  - Can use some other linear combination than an equal weighting
  - Indeed, any function of the two “signals” of user happiness
- Now we seek the top K docs by <u>net score</u>

Top *K* by net score – fast methods

- First idea: Order all postings by *g(d)*
- Key: this is a common ordering for all postings
- Thus, can concurrently traverse query terms' postings for
  - Postings intersection
  - Cosine score computation

Why order postings by *g(d)*?

- Under *g(d)*-ordering, top-scoring docs likely to appear early in postings traversal
- In time-bound applications (say, we have to return whatever search results we can in 50 ms), this allows us to stop postings traversal early
  - Short of computing scores for all docs in postings

Champion lists in *g(d)*-ordering

- Can combine champion lists with *g(d)*-ordering
- Maintain for each term a champion list of the *r* docs with highest *g(d)* + tf-idf$_{td}$
- Seek top-*K* results from only the docs in these champion lists

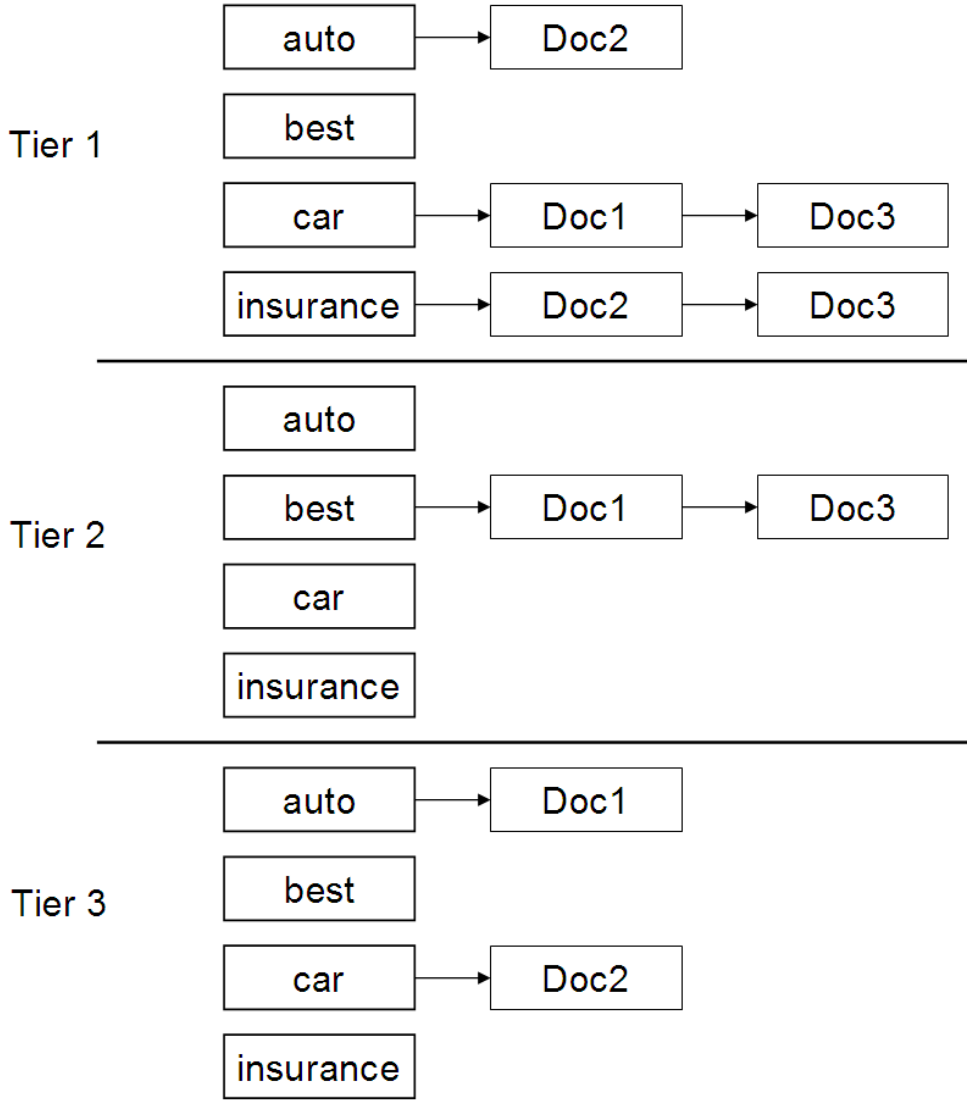
High and low lists

- For each term, we maintain two postings lists called *high* and *low*
  - Think of *high* as the champion list
- When traversing postings on a query, only traverse *high* lists first
  - If we get more than *K* docs, select the top *K* and stop
  - Else proceed to get docs from the *low* lists
- Can be used even for simple cosine scores, without global quality *g(d)*
- A means for segmenting index into two <u>tiers</u>

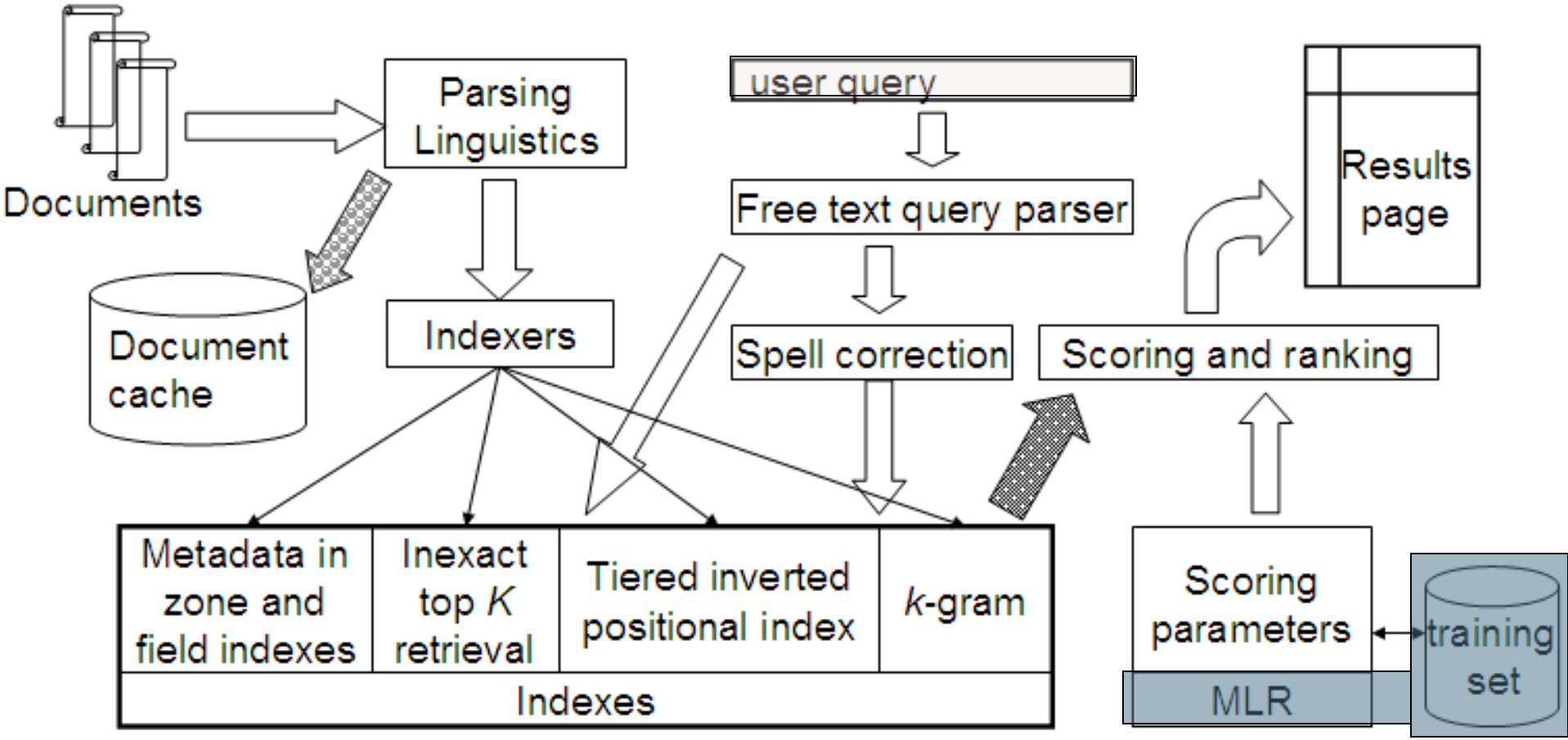
Tiered indexes

- Break postings up into a hierarchy of lists
  - Most important
  - …
  - Least important
- Can be done by *g(d)* or another measure
- Inverted index thus broken up into tiers of decreasing importance
- At query time use top tier unless it fails to yield *K* docs
  - If so drop to lower tiers

Example tiered index



Putting it all together



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

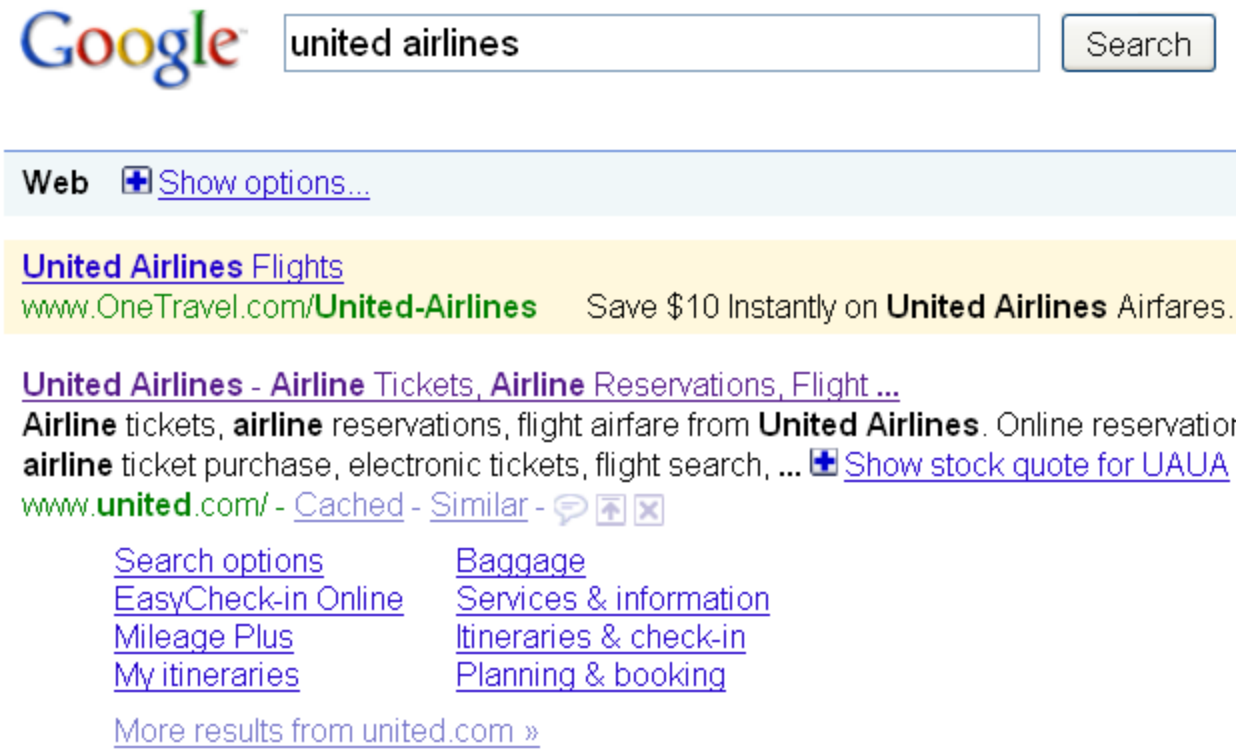
- Present one or more “windows” within the document that contain several of the query terms
  - “KWIC” snippets: Keyword in Context presentation

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



YAHOO!

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

Web Search:
Techniques, algorithms and Aplications

Basic Techniques
for Web Search

German Rigau <german.rigau@ehu.es>

[Based on slides by Eneko Agirre …
and Christopher Manning and Prabhakar Raghavan]