Web search: From the noun to the verb

Prabhakar Raghavan, Google
Simple view of a Web search engine
Simple view of a Web search engine
Simple view of a Web search engine
Simple view of a Web search engine
Simple view of a Web search engine

The Web

Spider

Indexer

Web

Indexes

Search

Retrieval and Ranking
Simple view of a Web search engine

The Web Spider

Indexer

Indexes

Retrieval and Ranking

Search
Simple view of a Web search engine

The Web

Web Spider

Indexer

Indexes

Retrieval and Ranking

Search
A brief history of web search

● First crawls of the Web ~1994
  ○ Simple indexes, keyword matches
  ○ Yahoo’s first search of its directory was a grep

● By 1995, several engines at scale
  ○ Altavista - proof point for DEC’s Alpha servers
    ■ Introduced Mercator - the first distributed and continuous web crawler
    ■ Trading freshness, quality, reach

Pinkerton 1994
Heydon/Najork 1999
Early emphasis on Recall

- **Recall** = the fraction of relevant results returned on a search
  - Say the Web contains 100 pages relevant to the query [ISWC 2014]
  - If the engine retrieves 70 of these for this query, **Recall is 0.7**
- Recall is a function of comprehensiveness of the index, and also *query phrasing* vs the user’s *intent*
- Altavista, Infoseek, Lycos … 10’s of millions of pages
User task: Get rid of mice in a politically correct way

Info need: Info about removing mice without killing them

Query: how trap mice alive

Misconception?

Misformulation?
Basic structure - inverted index

- For each word that occurs on any web page, record every document it occurs in
  - And every position it occurs in
  - Building it is a massive sort operation

- What’s a “word”? 
  - Basically, every string in any language
  - Billions of distinct “words”
Inverted index - example

- **2014**: 81090, 92793, 112793, 112883
- **iswc**: 13256, 92793
- **italy**: 89, 431, 967, 2145, 3279
- **semantic**: 5294, 21349
- **web**: 1, 2, 3, 5, 7, 8, 11
Inverted index - example

Sorted by docID
Inverted index - example

Sorted by docID
Early ranking

● Various early attempts at ranking from classic Information Retrieval
  ○ Derived from Library Science
    ▪ Scored documents by number of query terms, rarity of query terms
  ○ Presumes high quality authorship
  ○ Presumes no spam
Why did early Web ranking fail?

- Presumption of “safe” corpus
- Uniform authorship
- Aligned motives in content creation
- Ideas like term density easily hijacked
- *GOTO* - a commercial search engine
  - Websites bid for ranking on keywords
Precision became important

- *Precision* = fraction of retrieved results that are relevant
- Simple Boolean notion - is a web page relevant or not, for each query
- Thus if 4 of the 10 results for a query are relevant, precision = 0.4
- Precision can be measured at 1, 3, 10 results ...
Late 1990s - a new generation

- **inktomi** - perhaps the first “cloud-based” engine
- Large data centers with workstations, early notions of hot failure/replacement
- Hand tuned scoring function
  - Scoring every doc for every query became expensive
Index refinements

- semantic
- web
- semantic web
Index refinements

- semantic
  - Bigrams: 5294 ➔ 21349

- web
  - Bigrams: 1 ➔ 2 ➔ 3 ➔ 5 ➔ 8 ➔ 13 ➔ 21

- semantic web
  - Bigrams: 5294 ➔ 56374
Index refinements

- semantic
- web
- semantic web
- web
- riva del garda

Bigrams

Index tiers

5294 21349
1 2 3 5 8 13 21
5294 56374
4 6 9 12 17 5294
9793

Link analysis and off-page signals

- Google and others introduced link analysis
  - Measures of page quality independent of query
  - Link patterns between pages are harder to spam for a single website owner
  - Led to a new industry of website owners colluding to spam link analysis ... the battle rages on!

Brin/Page, Kleinberg 1998, Kumar et al 1999
Main ideas in link analysis 1

- A link is an endorsement of the page it links to
- Pages with many links coming in are better
- Pages with many links from other good pages are even better
- Lead to fixed-point iterative eigenvalue computations $\rightarrow$ pagerank
Main ideas in link analysis 2

- A link is a **description** of the page it links to
  - When indexing a document $D$, include the anchor text from links pointing to $D$

  www.ibm.com
Main ideas in link analysis 2  

- A link is a **description** of the page it links to
  - When indexing a document $D$, include the anchor text from links pointing to $D$

```
Armonk, NY-based computer giant IBM announced today ...

Big Blue today announced record profits for the quarter
```

www.ibm.com
But why do people query?

● In the beginning - like browsing - curiosity
  ○ What’s out there?
  ○ [to be or not to be]

● Very quickly (ca. 2001) turned into: the need behind a query
Lufthansa | lufthansa.com
www.lufthansa.com
106,600+ followers on Twitter
Sccopra le offerte: Tantissime destinazioni a prezzi imbattibili!
- Offerte A380
- Prenota online
- Tutte le offerte

Tutti i Voli Low Cost | Jetcost.it
www.Jetcost.it/Voli-LowCost
Confronta voli tra più di 600 compagnie di Linea e Low Cost!
Voli Milano A/R da 19€ · Voli Roma A/R da 19€ · Voli Palermo A/R da 29€

Lufthansa - Biglietti Economici.
Offerte-Voli.FindWith.me
Sccopri Ora Lufthansa!
Voli Internazionali · Voli Last Minute · Voli Last Second · Voli Low Cost

Lufthansa ® Italia - Prenotazione Voli Online... Translate this page
www.lufthansa.com/it/it/Homepage · Official site
Prenoti ora i voli alle tariffe più convenienti. Voli in Italia ed Europa da 99 €, andata e ritorno. - Lufthansa Italia

Check-in
Check-in online, check-in mobile, check-in via SMS, check-in al ...

Booking
Flying with Lufthansa. Booking, Top Offers, Miles & More and more

Lufthansa
La Deutsche Lufthansa AG, o più semplicemente Lufthansa, è la principale compagnia aerea tedesca.

www.lufthansa.com
Founded: 1953
Fleet size: 282
CEO: Carsten Spohr
Headquarters: Colonia

People also search for
The Best Gazpacho Recipes | Food & Wine
www.foodandwine.com/articles/the-best-gazpacho-recipes
Food & Wine offers the best gazpacho recipes along with spectacular variations of the chilled soup - ranging from rustic and chunky to smooth and elegant.
Gazpacho Recipe - Trey Foshee - Rustic Gazpacho - Raichlen's Gazpacho on Fire

Gazpacho Recipe : Ina Garten : Food Network
www.foodnetwork.com/recipes/gazpacho-r... Food Network
Rating: ★★★★★ Rating: 5 - 210 reviews - 20 mins
Get this all-star, easy-to-follow Gazpacho recipe from Ina Garten.

Gazpacho Recipe : Alton Brown : Food Network
www.foodnetwork.com/recipes/gazpacho-r... Food Network
Rating: ★★★★★ Rating: 5 - 87 reviews - 2 hrs 45 mins
Get this all-star, easy-to-follow Gazpacho recipe from Alton Brown.

Gazpacho Recipes - Allrecipes.com
allrecipes.com/recipes/soups-soups-chili/gazpacho/ Allrecipes.com
Looking for gazpacho recipes? Allrecipes has more than 30 trusted gazpacho recipes complete with ratings, reviews and cooking tips.

Gazpacho Recipe | Epicurious.com
www.epicurious.com/recipes & menus Epicurious
Rating: ★★★★☆ Rating: 4/4 - 7 reviews - 58 cal
umbrella in German

- **English**: umbrella
- **German**: Regenschirm

Translation of umbrella in German - Vocabulix

Translate the word umbrella to German. The dictionary languages are English-German: umbrella stand. umbrella.

German Translation of “umbrella” | Collins English German...

German Translation of “umbrella” | The official Collins English to German Dictionary online. Over 100000 German translations of English words and phrases.

German Word of the Day: Regenschirm - umbrella
Broder 2002 taxonomy

- **Navigational**
  - [lufthansa]; [facebook]
- **Informational**
  - [gazpacho recipe]; [eiffel tower height]
- **Transactional**
  - [nikon camera]; [plumber]

- Results depend on the type of query
- Grey areas [iswc]. Exceptions (news topics)
Stanford University - U.S. News & World Report
colleges.usnews.rankingsandreviews.com

Stanford, CA 94305
Tuition: $44,757
Student: 7,061 (53% male / 47% female)
Student-Faculty Ratio: 5:1
National Universities: 1888
Type of School: private, coed college
Rate of Acceptance:

Stanford Online
online.stanford.edu
Stanford Online courses from all parts of campus are reaching millions of learners globally.

Stanford University - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Stanford_University
Leland Stanford Junior University, or more commonly Stanford University, is a private research university in Stanford, California, and one of the world's most ...
Metrics: ranked list of results

*Discounted Cumulative Gain (DCG)* of a results page

\[ \sum_i d_i r(i) \]

- \( d_i \) is a (decreasing) discount for the \( i \)th result
- \( r(i) \) is the relevance of the \( i \)th result
- not necessarily Boolean!
- Query-doc relevance judged by panels
Heat maps
Heat maps

User attention diminishes top to bottom and left to right: Eye tracks, clicks ...
Heat maps

User attention diminishes top to bottom and left to right: Eye tracks, clicks ...

computational vs traditional social sciences
Refining query taxonomies

[plumber] is local and transactional
signals not in the query - IP location
Nuanced parsing of queries e.g.,
[jfk rome] vs [kennedy rome]
Fly Rome - New York - Aer Lingus
www.aerlingus.com/
Fly from Rome via Dublin with Aer Lingus from €249

Flights from New York, USA (JFK) to Rome, Italy (all airports)
www.google.com/flights

<table>
<thead>
<tr>
<th>Depart</th>
<th>Return</th>
<th>Nonstop</th>
<th>All flights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat, November 1</td>
<td>Wed, November 5</td>
<td>Delta</td>
<td>Aeroflot</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9h 10m</td>
<td>18h 15m+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>from $1,599</td>
<td>from $1,298</td>
</tr>
</tbody>
</table>

New York to Rome Flights: Book Flights from JFK to FCO ...
flights.expedia.com / North America / USA / New York / JFK / Expedia
Book your JFK to FCO flights with Expedia and find last-minute New York to Rome airfare. Expedia offers discount airfare on multiple airline carriers that fly direct ...

(JFK) - Rome - Kayak
www.kayak.com/Cheap-Flights-New-York-to-Rome-JFK-ROM.co ... Kayak
Search cheap non-stop flights New York JFK Rome FCO (JFK) Rome (FCO) from $1,298 with Kayak
jfk rome

About 68,500,000 results (0.92 seconds)

Fly Rome - New York - Aer Lingus.com
www.aerlingus.com/
Fly from Rome via Dublin with Aer Lingus from €249

Flights from New York, USA (JFK) to Rome, Italy (all airports)
www.google.com/flights

<table>
<thead>
<tr>
<th>Depart</th>
<th>Sat, November 1</th>
<th>Return</th>
<th>Wed, November 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonstop</td>
<td>Delta</td>
<td>9h 10m</td>
<td>from €1,599</td>
</tr>
<tr>
<td></td>
<td>Alitalia</td>
<td>9h 10m</td>
<td>from €1,759</td>
</tr>
<tr>
<td>All flights</td>
<td>Aeroflot</td>
<td>18h 15m+</td>
<td>from €1,298</td>
</tr>
<tr>
<td></td>
<td>Royal Air Maroc</td>
<td>22h 30m+</td>
<td>from €1,365</td>
</tr>
<tr>
<td></td>
<td>Other airlines</td>
<td>10h 5m+</td>
<td>from €1,570</td>
</tr>
</tbody>
</table>

More Google flight results

New York to Rome Flights: Book Flights from JFK to FCO ...
flights.expedia.com | North America | USA | New York | JFK | Expedia
Book your JFK to FCO flights with Expedia and find last-minute New York to Rome airfare. Expedia offers discount airfare on multiple airline carriers that fly direct ...

(jfk) - Rome - Kayak
www.kayak.com/Cheap-Flights-New-York-to-Rome JFK-ROM.co ...
Kayak

kennedy rome

About 33,600,000 results (0.62 seconds)

Hotel Kennedy - Rome, Italy - Booking.com
www.booking.com
4.0 ★★★★★ rating for booking.com
Kennedy Rome No reservation costs. Great rates.
Hotel Photos • Hotel Reviews • View Hotel on a Map • Book Now
Most Popular Online Booking Website in 2013 – CNN Travel

Hotels in Dubai Hotels in Singapore
Hotels in Melbourne Hotels in Sydney

Kennedy Hotel (Rome, Italy) - Hotel Reviews - TripAdvisor
www.tripadvisor.co.uk | Lazio | Rome | Rome hotels | TripAdvisor
★★★☆☆☆ Rating: 3.5 - 457 reviews - Price range: £
Kennedy Hotel, Rome: See 457 traveller reviews, 195 candid photos, and great deals for Kennedy Hotel, ranked #699 of 1264 hotels in Rome and rated 3.5 of 5 ...

Kennedy Hotel in Rome - Book a Charming hotel in Rome ...
www.hotelkennedy.net/en
Welcome to the website of the Kennedy Hotel, a Charming 3 star hotel in Rome. Center. Book your hotel in Rome at the best price on the internet.

The Kennedy Hotel, Rome, Accommodation, 3 star hotel
www.kennedyhotelrome.com

(JFK) - Rome - Kayak
www.kayak.com/Cheap-Flights-New-York-to-Rome JFK-ROM.co ... Kayak
Scoring/ranking functions

- Combine 100s of factors such as term densities, link counts, user clicks, ...
- Combination can be manually tuned, or machine learned (MLR)
- For MLR, train on editorial judgments, clicks or both
Some nuances

- Already, examples show more than simple ranking - maps, flights …
  - Queries must be classified into Nav, Info, etc.
- Impractical to score all docs on all queries
  - Only docs whose text/anchors contain query terms
  - Only authoritative documents
  - Fallback strategies
Revised view of a Web search engine
Revised view of a Web search engine
Revised view of a Web search engine
Revised view of a Web search engine

Indexer

Words

Links

Indexes

Query Classification

Retrieval and ranking

Maps

Flights
Revised view of a Web search engine

Indexer

Words Links

Indexes

Query Classification

Retrieval and ranking

Search

Maps Flights
Lots of cleverness in assembling pages - massive AB experiments

Revised view of a Web search engine
Entities

● Over 99% of queries have noun (phrases)
  ○ People, places, companies, airports …
● When the entity occurs in a query, insert a “card” about the entity
San Francisco International Airport - Official Site
www.flysfo.com

Flight Info | www.flysfo.com
www.flysfo.com/flight-info
San Francisco International Airport (SFO) offers non-stop flights to more than 31 international cities on 30 international carriers. The Bay Area's largest airport ...

www.sfoitalia.com
www.sfoitalia.com/home.php
Non è possibile visualizzare una descrizione perché il sito non lo consente.

Related searches for sfo
SF DRS Meteo San Francisco Airport Departures
SF Fernsehen Sanitaria Viale Delle Milizie Roma
PSP File Extension PBP SFO Roma

San Francisco International Airport - Wikipedia, the free ... en.wikipedia.org/wiki/San_Francisco_International_Airport
San Francisco International Airport (IATA: SFO, ICAO: KSFO, FAA LID: SFO) is 13 miles (21 km) south of downtown San Francisco, California, near Millbrae and San Bruno ...

Aeroporto Internazionale di San Francisco
L'Aeroporto Internazionale di San Francisco è aeroporto situato circa 21 km dal centro di San Francisco (CA)

Airport code: SFO
Address: 780 S. Airport Blvd - 94128 San Francisco (CA)

People also search for
Aeroporto Internazionale di San Francisco ...
American Airlines Flight 24
On-time - departs in 5 hours 27 mins

SFO → JFK

Departs San Francisco, Friday, October 17
Arrives New York City, Friday, October 17

Time  Terminal  Gate  Time  Terminal  Gate
7:00 AM  2  58B  3:40 PM  8  C46

American Airlines (AA) #24 FlightAware
flightaware.com/live/flight/AAL24   FlightAware

American Airlines (AA) #24 Flight Tracker (AAL24) Flight Tracker (en route flights, arrivals, departures, history) with live maps and aircraft photos.

Thought for the Day -- Hazelden
www.hazelden.org/web/public/thought.view?catid...
Hazelden Foundation

You are reading from the book Twenty Four Hours a Day Hardcover (24 Hours) ... Am I glad to be a part of the great work that A.A. is doing and do I feel a deep ...

Touchstones - Hazelden -- Thought for the ... - 24 Hours - Sign up
Entities

● Can we synthesize a compendium of all entities?
● How? Wikipedia, metaweb, feeds
  ○ Keeping it up to date
  ○ The long tail problem
Palo Alto - Il Fornaio
www.illianforno.com/paloalto
Il Fornaio Palo Alto offers elegant private and semi-private dining space along with a wide range of specialized menus and services to match our authentic, ...

3.6 ★★★★★ 217 Google reviews · Write a review

520 Cowper St, Palo Alto, CA 94301, United States
(650) 853-3888
PASTICCERIA - FRUTTA FRESCA - SPREMUTE

Il Fornaio
www.illianforno.com
Founded in Italy in 1972, Il Fornaio offers lunch and dinner guests the most authentic Italian dining experience available at 21 restaurants across the US.

Il Fornaio - Palo Alto, CA | Yelp
www.yelp.com › Restaurants › Italian › Yelp
★★★★★ Rating: 3 - 394 reviews - Price range: $$
394 Reviews of Il Fornaio "In short.. Friendly waiter, good food and a nice patio like seating. Try their capellini pomodoro ( for vegetarians ). Their butternut ...

Il Fornaio - Palo Alto Restaurant - Palo Alto, CA | OpenTable
www.opentable.com › Palo Alto restaurants › OpenTable
★★★★★ Rating: 3.9 - 159 reviews - Price range: $30 and under
Book now at Il Fornaio - Palo Alto in Palo Alto, explore menu, see photos and read 159 reviews: "The food was so delicious! The service was spot on!"
Metrics

- As more information is surfaced directly
  - List of links draws fewer clicks
  - Need behind the query often answered directly

- How can we tell if the user is happy?

- Some rudimentary ideas
  - Shake phone to report a bug
  - MSR - correlation to eye tracking

Huang et al. 2011 No clicks, no problem …
The verb behind the noun

- **Select** a restaurant (list of restaurants)
  - by Geo, genre, open hours ...
- **Book** a restaurant (booking service)
- **Go** to the restaurant (map/directions)
How can we tell which verb?

- Conversational assistants e.g., Siri
  - Explicit vs implicit specification
- Consumer behavior: *decision funnel*
- Signals - time of day, location etc
  - [roger federer]
  - [french laundry]
Spoken queries and the party factoid

Google search result for "what is the height of the Eiffel Tower":
- **986' (301 m)**
  - Eiffel Tower, Height
- **Burj Khalifa**
  - 2,717'
- **Empire State Building**
  - 1,250'
- **Statue of Liberty**
  - 305'

Google Maps showing the location of the Eiffel Tower in Paris, France.
Proactive contextual experiences

- Anticipating the next query
  - Proactively show you the answer *before you query*
- Numerous startups, Microsoft Cortana, Google Now
- Simplest example - calendar alert
Now examples
Now examples
Query lengths

- For over 15 yrs, query lengths ~2.5 words
  - Modulo high traffic periods in early engines (!)
- Bimodal query lengths
  - Proactive notifications: the “zero word” query
  - Spoken queries: the 8-word query
Summary

- Web search remains work in progress
- Research questions span
  - User behavior
  - New devices - mobile++
  - New infrastructure - be computationally bound!
- Still early for deep semantics in search
  - Utility trumps technical virtuosity
  - Conditioning users’ expectations - infallibility
Summary 2

- **User needs direct solutions**
  - Computational social sciences grow
- **Scientific methods, engineering results**
- **Teaching web search engines**
  - 6 x 1 hr lectures
  - After 3 x 1 hr lectures, competent CS undergrads can build a search engine for their dept website
  - Final 3 x 1 hr lectures - more advanced material, plus basics of advertising
Thank you!