UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

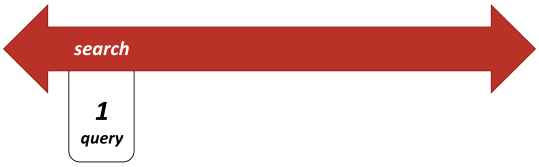
Introduction

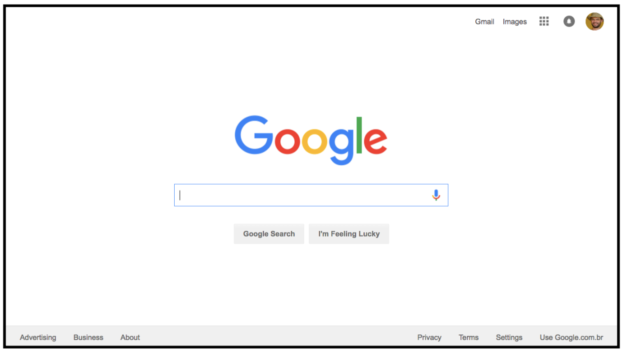
Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Information retrieval

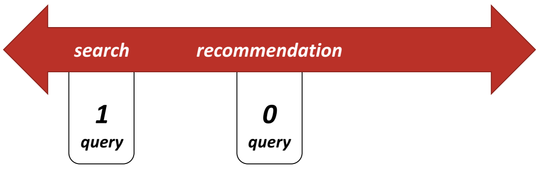
> *Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.*
>
> Gerard Salton, 1968

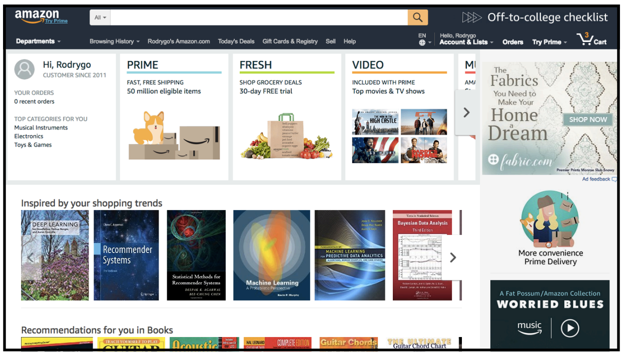
Retrieval tasks



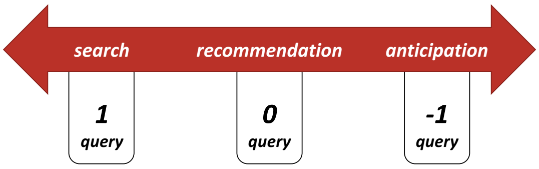


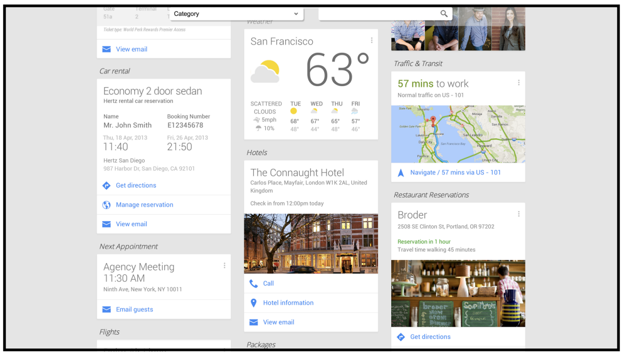
Retrieval tasks





Retrieval tasks




Gate 51a Terminal 2

Ticket type: World Perk Rewards Premier Access

View email

Category

Weather

Car rental

Economy 2 door sedan
Hertz rental car reservation

Name: Mr. John Smith
Booking Number: E12345678

Thu, 18 Apr, 2013 11:40
Fri, 26 Apr, 2013 21:50

Hertz San Diego
987 Harbor Dr, San Diego, CA 92101

Get directions

Manage reservation

View email

Next Appointment

Agency Meeting
11:30 AM
Ninth Ave, New York, NY 10011

Email guests

Flights

San Francisco
63°
SCATTERED CLOUDS
5mph
10%

TUE	WED	THU	FRI
68°	67°	65°	57°
48°	44°	48°	46°

Hotels

The Connaught Hotel
Carlos Place, Mayfair, London W1K 2AL, United Kingdom

Check in from 12:00pm today

Call

Hotel information

View email

Packages

Traffic & Transit

57 mins to work
Normal traffic on US - 101

Navigate / 57 mins via US - 101

Restaurant Reservations

Broder
2508 SE Clinton St, Portland, OR 97202

Reservation in 1 hour
Travel time walking 45 minutes

Get directions

Our focus: search

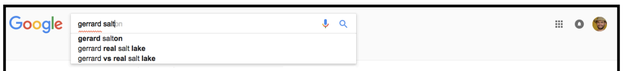
A lot of people ………………………. *60k queries per second*

From a lot of places ………. *whole planet (and beyond?)*

Using a lot of a devices ………………… *smart-you-name-it*

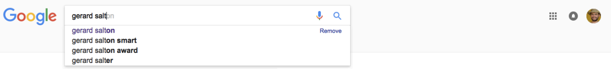
Looking for a lot of info ……………………. *10^{11} documents*

What does a search engine do?


gerrard salton

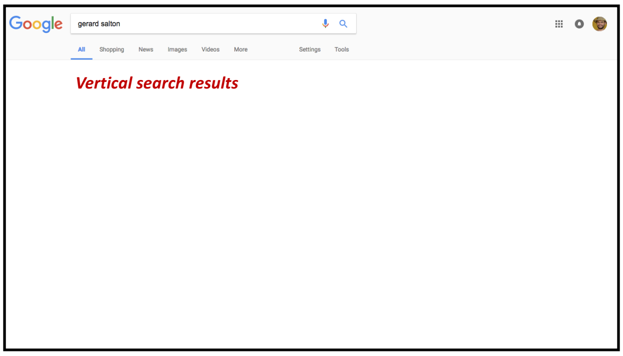
gerard salton
gerrard real salt lake
gerrard vs real salt lake

Spelling correction

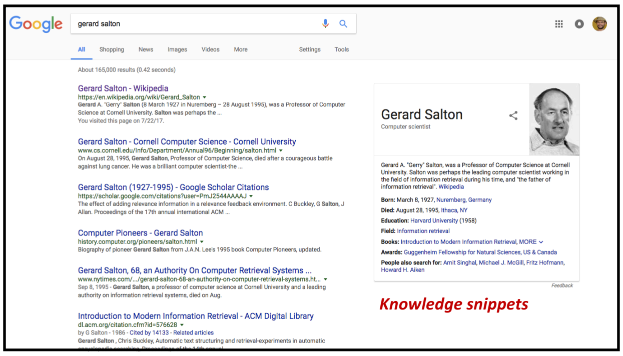

gerard salton

gerard salton — Remove
gerard salton smart
gerard salton award
gerard salter

Query autocompletion
Query suggestions



Vertical search results



About 165,000 results (0.42 seconds)

Gerard Salton - Wikipedia
https://en.wikipedia.org/wiki/Gerard_Salton ▾
Gerard A. "Gerry" **Salton** (8 March 1927 in Nuremberg – 28 August 1995), was a Professor of Computer Science at Cornell University. **Salton** was perhaps the ...
You visited this page on 7/22/17.

Gerard Salton - Cornell Computer Science - Cornell University
www.cs.cornell.edu/Info/Department/Annual96/Beginning/salton.html ▾
On August 28, 1995, **Gerard Salton**, Professor of Computer Science, died after a courageous battle against lung cancer. He was a brilliant computer scientist-the ...

Gerard Salton (1927-1995) - Google Scholar Citations
https://scholar.google.com/citations?user=PmJ2544AAAAJ ▾
The effect of adding relevance information in a relevance feedback environment. C Buckley, G **Salton**, J Allan. Proceedings of the 17th annual international ACM ...

Computer Pioneers - Gerard Salton
history.computer.org/pioneers/salton.html ▾
Biography of pioneer **Gerard Salton** from J.A.N. Lee's 1995 book Computer Pioneers, updated.

Gerard Salton, 68, an Authority On Computer Retrieval Systems ...
www.nytimes.com/.../gerard-salton-68-an-authority-on-computer-retrieval-systems.ht... ▾
Sep 8, 1995 - **Gerard Salton**, a professor of computer science at Cornell University and a leading authority on information retrieval systems, died on Aug.

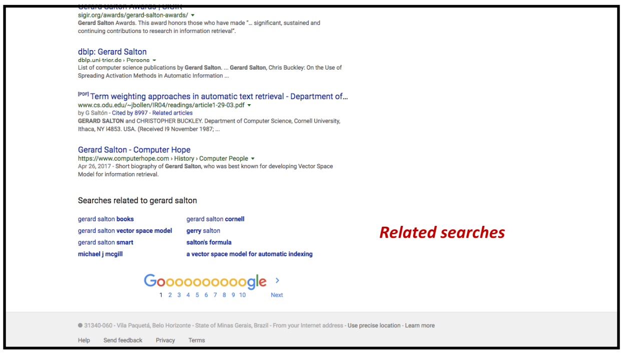
Introduction to Modern Information Retrieval - ACM Digital Library
dl.acm.org/citation.cfm?id=576628 ▾
by G Salton - 1986 - Cited by 14133 - Related articles
Gerard Salton , Chris Buckley, Automatic text structuring and retrieval-experiments in automatic

Gerard Salton
Computer scientist

Gerard A. "Gerry" Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time, and "the father of information retrieval". Wikipedia

Born: March 8, 1927, Nuremberg, Germany
Died: August 28, 1995, Ithaca, NY
Education: Harvard University (1958)
Field: Information retrieval
Books: Introduction to Modern Information Retrieval, MORE ⌄
Awards: Guggenheim Fellowship for Natural Sciences, US & Canada
People also search for: Amit Singhal, Michael J. McGill, Fritz Hofmann, Howard H. Aiken

Knowledge snippets


sigir.org/awards/gerard-salton-awards/ ▾
Gerard Salton Awards. This award honors those who have made "... significant, sustained and continuing contributions to research in information retrieval".

dblp: Gerard Salton
dblp.uni-trier.de › Persons ▾
List of computer science publications by **Gerard Salton**. ... **Gerard Salton**, Chris Buckley: On the Use of Spreading Activation Methods in Automatic Information ...

[PDF] Term weighting approaches in automatic text retrieval - Department of...
www.cs.odu.edu/~jbollen/IR04/readings/article1-29-03.pdf ▾
by G Saltón - Cited by 8997 - Related articles
GERARD SALTON and CHRISTOPHER BUCKLEY. Department of Computer Science, Cornell University, Ithaca, NY l4853. USA. (Received l9 November 1987; ...

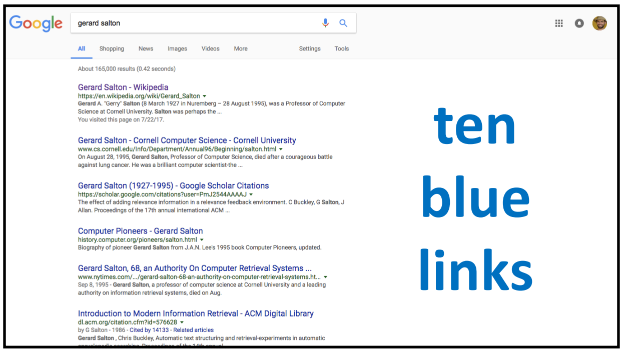
Gerard Salton - Computer Hope
https://www.computerhope.com › History › Computer People ▾
Apr 26, 2017 - Short biography of **Gerard Salton**, who was best known for developing Vector Space Model for information retrieval.

Searches related to gerard salton

gerard salton **books**	gerard salton **cornell**
gerard salton **vector space model**	**gerry** salton
gerard salton **smart**	**salton's formula**
michael j mcgill	**a vector space model for automatic indexing**

● 31340-060 - Vila Paquetá, Belo Horizonte - State of Minas Gerais, Brazil - From your Internet address - Use precise location - Learn more

Related searches



About 165,000 results (0.42 seconds)

Gerard Salton - Wikipedia
https://en.wikipedia.org/wiki/Gerard_Salton ▾
Gerard A. "Gerry" **Salton** (8 March 1927 in Nuremberg – 28 August 1995), was a Professor of Computer Science at Cornell University. **Salton** was perhaps the ...
You visited this page on 7/22/17.

Gerard Salton - Cornell Computer Science - Cornell University
www.cs.cornell.edu/Info/Department/Annual96/Beginning/salton.html ▾
On August 28, 1995, **Gerard Salton**, Professor of Computer Science, died after a courageous battle against lung cancer. He was a brilliant computer scientist-the ...

Gerard Salton (1927-1995) - Google Scholar Citations
https://scholar.google.com/citations?user=PmJ2544AAAAJ ▾
The effect of adding relevance information in a relevance feedback environment. C Buckley, G **Salton**, J Allan. Proceedings of the 17th annual international ACM ...

Computer Pioneers - Gerard Salton
history.computer.org/pioneers/salton.html ▾
Biography of pioneer **Gerard Salton** from J.A.N. Lee's 1995 book Computer Pioneers, updated.

Gerard Salton, 68, an Authority On Computer Retrieval Systems ...
www.nytimes.com/.../gerard-salton-68-an-authority-on-computer-retrieval-systems.ht... ▾
Sep 8, 1995 - **Gerard Salton**, a professor of computer science at Cornell University and a leading authority on information retrieval systems, died on Aug.

Introduction to Modern Information Retrieval - ACM Digital Library
dl.acm.org/citation.cfm?id=576628 ▾
by G Salton - 1986 - Cited by 14133 - Related articles
Gerard Salton , Chris Buckley, Automatic text structuring and retrieval-experiments in automatic

ten
blue
links

The ranking problem

Given

- Some evidence of the user's need

Produce

- A list of matching information items
- In decreasing order of relevance

The ranking problem

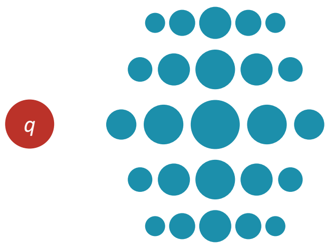
Given

- Some ~~evidence of the user's need~~ ***query***

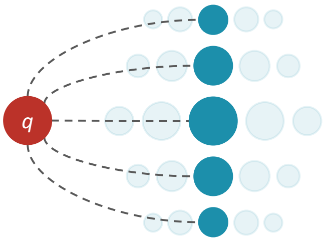
Produce

- A list of matching ~~information items~~ ***documents***
- In decreasing order of relevance

1) What documents do we show?



2) What order do we show them in?



2) What order do we show them in?

$$f(q,d)$$

Isn't it a solved problem?

Credit: Anna Jumped

Efficiency

Efficiency is about doing something (good or bad) in an optimal way (i.e., faster or with fewer resources)

Key performance indicators

- *Query latency:* searching billions of documents
- *Query throughput:* serving thousands of users
- *Document latency:* serving freshly published content

Effectiveness

Effectiveness is about doing the right thing; it's about finding documents that are relevant to the user

Relevance is influenced by many factors

- Topical relevance vs. user relevance
- Task, context, novelty, style

Ranking models define ***a view of*** relevance

Pursuing relevance

Exact matching of words is not enough

- Queries are ambiguous; documents are ambiguous
- Many different ways to write the same thing

Even perfect matching is not enough

- Must counter adversarial content
- Must infer quality beyond content

Assessing relevance

Relevance is a user's prerogative

- We can observe changes in user behavior
- Or directly ask the user how we're doing

Evaluation is an empirical science

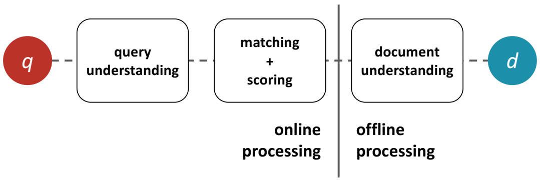
- It must be scientifically rigorous
- It must be economically viable

What do search engineers do?

The ranking problem

$$f(q,d)$$

Ranking pipeline


q

query understanding

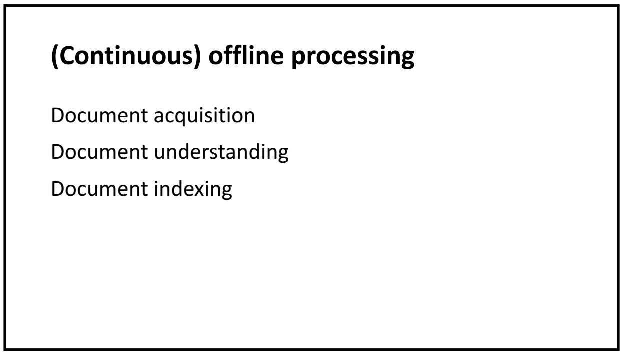
matching + scoring

document understanding

d

online processing

offline processing

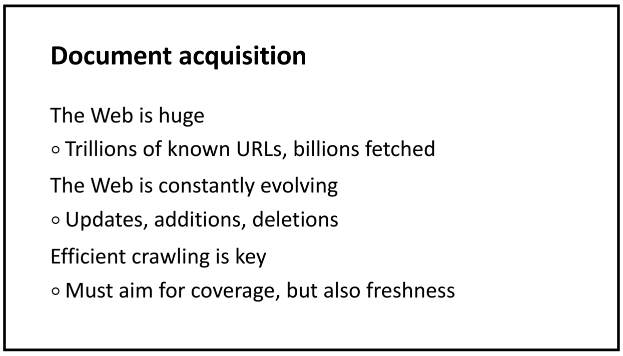


(Continuous) offline processing

Document acquisition

Document understanding

Document indexing



Document acquisition

The Web is huge

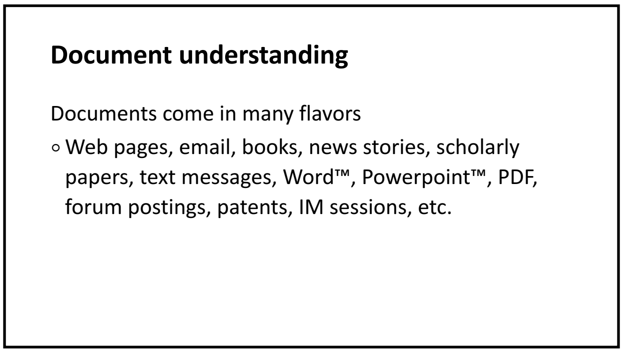
- ○ Trillions of known URLs, billions fetched

The Web is constantly evolving

- ○ Updates, additions, deletions

Efficient crawling is key

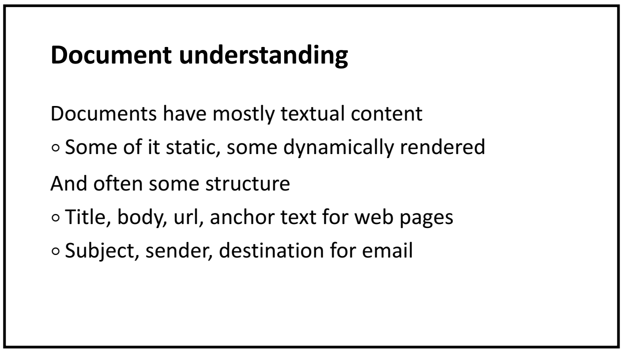
- ○ Must aim for coverage, but also freshness



Document understanding

Documents come in many flavors

- ○ Web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.



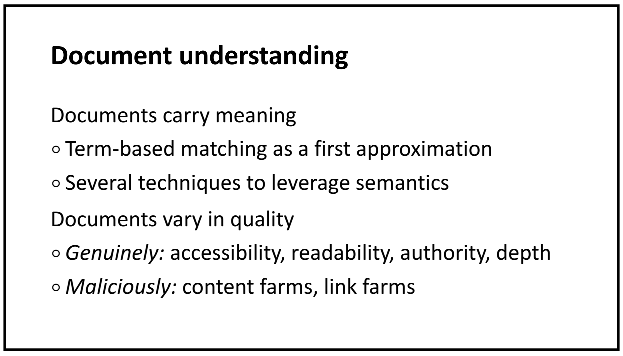
Document understanding

Documents have mostly textual content

- ○ Some of it static, some dynamically rendered

And often some structure

- ○ Title, body, url, anchor text for web pages
- ○ Subject, sender, destination for email



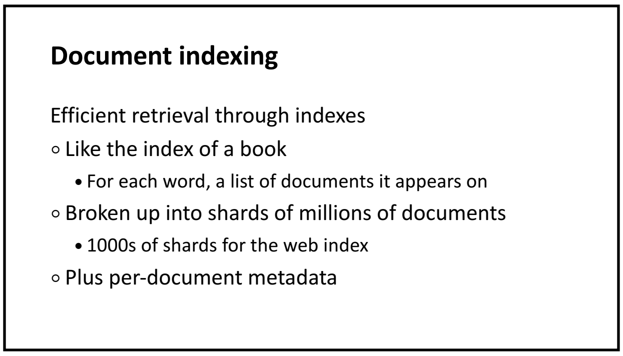
Document understanding

Documents carry meaning

- ○ Term-based matching as a first approximation
- ○ Several techniques to leverage semantics

Documents vary in quality

- ○ *Genuinely:* accessibility, readability, authority, depth
- ○ *Maliciously:* content farms, link farms



Document indexing

Efficient retrieval through indexes

- ○ Like the index of a book
  - • For each word, a list of documents it appears on
- ○ Broken up into shards of millions of documents
  - • 1000s of shards for the web index
- ○ Plus per-document metadata

Online processing

Query understanding

Matching and scoring

Post-processing

Query understanding

Keyword queries are often poor descriptions of the user’s actual information need

- Interaction and context also matter

Query understanding techniques can help

- Query segmentation, query scoping
- Query relaxation, query expansion

Query understanding

Query scoping through semantic annotation

- [**san jose** convention center]
- [**matt cutts**]

Query expansion through acronym expansion

- [**gm** trucks] → [**general motors** trucks]
- [**gm** corn] → [**genetically modified** corn]

Matching and scoring

Send the query to all the shards

Each shard

- Finds matching documents
- Scores each query-document pair
- Sends back the top *n* documents

Combine all the top documents and sort by score

Post-retrieval adjustments

Host clustering, sitelinks

Near-duplicate removal, diversification

Spam demotions, copyright takedowns

Course goals

Provide an in-depth account of ranking models and evaluation methods for information retrieval

Provide an exploration of recent advances and current research directions in the field

Course scope

Focus on ranking

- Query-dependent ranking
- Query-independent ranking
- Machine-learned ranking

Focus on evaluation

- Offline, online, counterfactual

Out-of-scope

We have dedicated courses for:

- Information retrieval
- Recommender systems
- Machine learning
- Data mining

Course materials: textbooks

Search Engines: Information Retrieval in Practice
by B. Croft, D. Metzler, and T. Strohman

Introduction to Information Retrieval
by C. Manning, P. Raghavan, and H. Schütze

Modern Information Retrieval
by R. Baeza-Yates and B. Ribeiro-Neto

Course materials: textbooks

Information Retrieval: Implementing and Evaluating Search Engines
by S. Büttcher, C. Clarke, and G. Cormack

Text Data Management: A Practical Introduction to Information Retrieval and Text Mining
by C. Zhai and S. Massung

Course materials: surveys

Foundations and Trends in Information Retrieval
by several authors

Morgan & Claypool Synthesis Lectures on Information Concepts, Retrieval, and Services
by several authors

Other relevant material

General background

- Algorithms and data structures
- Basic statistics
- Basic linear algebra

Advanced readings

- Google Scholar is your friend

Course grading (tentative)

Exams: 40%

Project: 40%

Assignments: 20%

Course website

http://homepages.dcc.ufmg.br/~rodrygo

References

Google Search Statistics
Internet Live Stats, 2017

How Google Works: A Google Ranking Engineer's Story
Haahr, SMX West 2016

Ten blue links on Mars
Clarke et al., WWW 2017

Writing assignment #0

Fill in a short questionnaire describing your past experience and expectations related to the course

- **Due Tue, Aug 15 @ 23:55 via Moodle**

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Search Architecture

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br