INTRODUCTION TO INFORMATION RETRIEVAL
Organization

- **Timetable**
  - Lectures: Mondays 13:30 – 15:00, room H-E.51
  - Exercises: biweekly, Tuesdays 15:30 – 17:00, room H-E.52

- **Exam**
  - Condition for admission: oral presentation of at least two solutions during the tutorials
  - Form of exam: oral exam at the end of the term
What is this lecture about?

“Know where to find the information and how to use it – That's the secret of success.”

Albert Einstein

Subject

Automated techniques for finding information relevant to an information need from a large corpus of data.

- Content & structure analysis
- Indexing, search, ranking
- Classification, clustering
- Annotation, extraction, summarization, visualization
- Personalization & collaboration
- Interaction with knowledge bases
Application areas of IR

**Web search**
- Web pages (or multimedia) about people, products, locations, events, news, problem solutions, ...
- Main problems: scalability, spam, format diversity, deep (hidden) web, ...

**Enterprise search**
- Intranet documents/pages (or multimedia) about people/experts, products, divisions, problem solutions, ...
- Main problems: low information redundancy, few links, security issues, ...

**Indexing and retrieval for knowledge discovery & data mining**
- Bioinformatics: finding promising medical therapies, retrieving salient interaction paths in biochemical networks, ...
- Business intelligence: risk identification, composition of equity baskets, ...
- Scientific observations: identifying typical or noticeable patterns
- Socio-political observations: analyzing social concerns, developments, political issues, ...
## Differences to database systems

<table>
<thead>
<tr>
<th>Information retrieval systems</th>
<th>Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
<td></td>
</tr>
<tr>
<td>Unstructured, semi-structured information (text, multimedia, etc.)</td>
<td>Structured, semi-structured data, e.g., database table(s), XML-document(s)</td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td></td>
</tr>
<tr>
<td>Imprecise, keyword-based, typically with “and’ish” search semantics</td>
<td>Precise, structured query language, e.g., SQL, Xpath, Datalog, SPARQL …</td>
</tr>
<tr>
<td><strong>Retrieval</strong></td>
<td></td>
</tr>
<tr>
<td>Approximate, frequency-, similarity-based</td>
<td>Exact match of attribute values</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td></td>
</tr>
<tr>
<td>Ranked by relevance</td>
<td>Unranked, or ranked by specific attribute(s) (from the order-by clause of the query)</td>
</tr>
</tbody>
</table>
Historic background

1945: far-reaching ideas about personal & collective information gathering, indexing, aligning, associating, sharing, reasoning, ... mentioned in “As we may think” by Vannevar Bush

Vannevar Bush

Memex extends human memory and reasoning capabilities
Roots in search over library catalogs

- Early work on IR, in the 50’s and 60’s

  - Boolean search (i.e., combine search terms with Boolean operators: **and**, **or**, **not**) unsatisfactory

  - Need for approximate search with “vaguely” Boolean operators needed (e.g., “**and’ish**” search semantics)

- Need for ranked results
Early research

- **Hans Peter Luhn** (IBM researcher)

- **Gerard Salton** (professor at Cornell)
  - 1960s: supervised the development of **SMART (System for the Mechanical Analysis and Retrieval of Text)**
  - 1969: published “**Automatic Information Organization and Retrieval**”
  - Introduces important concepts, e.g., **vector space model**, **relevance feedback**, **Rocchio classification**, ...

- Other influential scientists
  - **Karen Spärck Jones** introduced **inverse-document-frequency** in 1972
  - **Steven Robertson** introduced **Okapi BM25** ranking model together with Karen Spärck Jones in 1976
... and then the web was invented

~ 630 Mio. active sites

~ 315 Mio. active sites

Numbers from: http://news.netcraft.com/archives/2012/

Web 2.0 era

Collaborating (real-time)

Searching

Publishing

1969 ARPANET

1989 WWW “Mesh” proposal by Berners Lee

1998 Google

2000/01 Wikipedia

2004/05 Social networks Web 2.0

2009 Cloud, real-time, mobile

Dr. Gjergji Kasneci | Introduction to Information Retrieval | WS 2012-13
General infrastructure for web IR

Source: Modern Information Retrieval, Baeza-Yates, Ribeiro-Neto
Crawling

Downloader (visiting URLs, breadth-first, multi-threaded)

Scheduler

Web

URL analyzer (Is URL known? Has it been visited? Should it be crawled?)

Indexer (text analysis, cleansing, tokenizing, stemming, annotating, indexing)

Repository of indexed terms, URLs, metadata, thesaurus...

Web graph manager (graph construction, URL management)

Hyperlink extractor (hyperlink extraction, normalization)
**Indexing**

- **Document analyzer**
  (Structure, language, detect spam, duplicated or dynamic content ...)

- **Cleanser**
  (Remove HTML tags and scripts, tokenize, stem, remove stop words ...)

- **Score allocator**
  (Term frequency, inverted document frequency, ...)

- **Thesaurus**
  (with synonyms, hyponyms, hypernyms, meronyms, ...)

- **Inverted Index (B+ tree)**
  (camera, iphone)

- URLs & scores

---

Dr. Gjergji Kasneci | Introduction to Information Retrieval | WS 2012-13
Query answering

query (expressing search need)

User interface

Query processor
(parsing, cleansing,
spell correction,
entity recognition,
query expansion,
reformulation, ...)

Result provider
(top-k retrieval &
ranking by relevance,
authority, freshness, ...)

Thesaurus
(with synonyms,
hyponyms, hypernyms,
meronyms, ...)

Inverted Index
(B+ tree)
camera
iphone

URLs & scores
URLs & scores
Merging results through metasearch

query

response

Metasearch engine

User interface

Query distributor

Result aggregator

Search engine 1

Search engine 2

Search engine 3

Web
Metasearch example

**Free Pussy Riot**
takeaction.amnestyusa.org • Sponsored: Ads by Yahoo!
Free the Russian Punk Band. Tell Putin, Music is not a crime

**Russian punk band found guilty of 'hooliganism,' given two-year jail...**
news.yahoo.com/...d-guilty-hooliganism-115937812.html • Found exclusively on: Google
22 hours ago ... From the blog The Lookout: Three members of Pussy Riot—a Russian punk band and feminist collective that mocked Russian president ...

**Russian Punk Band Is Sentenced to 2 Years - The New York Times**
www.nytimes.com/...sian-punk-band.html?pagewanted=all • Found on: Google, Yahoo! Search
20 hours ago ... Three women who staged a protest against Vladimir V. Putin, Russia's president, in a cathedral and became a cause célèbre of artists around ...

**Guilty: Russian punk band Pussy Riot face three years hard labor...**
Guilty: Russian punk band Pussy Riot face three years hard labor after being convicted of hooliganism

**Category: Russian people convicted of murder - Wikipedia, the free encyclopedia**
en.wikipedia.org/...ussian_people_convicted_of_murder • Found exclusively on: Yandex
See also category: People convicted of murder by Russia. ... Retrieved from "http://en.wikipedia.org/w/index.php?
title=Category:Russian_people_convicted_of_murder&oldid=251697077".

**Russian Punk Band Is Sentenced to 2 Years for Anti-Putin Protest...**
www.nytimes.com/...for-jailed-russian-punk-band.html • Found exclusively on: Google
1 day ago ... A Moscow judge issued stiff prison sentences to three women from the punk band Pussy Riot, who staged a protest against Vladimir V. Putin in ...

**Russian punk band P**** Riot convicted of hooliganism - Politics...**
www.gamefaqs.com/boards/261-politics/63768692 • Found exclusively on: Yahoo! Search
On Politics, a GameFAQs message board topic titled "Russian punk band P**** Riot convicted of hooliganism".
Taxonomy of search on the web

- Typical semantics: **and’ish** search, e.g., “cheap flights Rome”

- Categorization of query types (by A. Broder, *A taxonomy of web search*, 2002) through query log analysis and user study:
  
  - ~50% informational queries: “General relativity”, “Capital of Taiwan”, “Obama campaign”, ...
  
  - ~20% navigational queries: “facebook”, “amazon”, “ebay”, ...
  
  - ~30% transactional/resource-oriented queries: “hotels Seattle”, “flights Berlin Rome”, “eminem mp3”, “skype download”, ...

- Different numbers presented in study by Rose & Levinson in 2004 (~62% informational, ~13% navigational, ~25% transactional queries)
Advanced query types

- Regular expression queries (e.g., on Google):
  -inurl:(htm|html|php) intitle:"index of" +"last modified" +"parent directory"
  +description +size +(wma|mp3) "Nirvana"

- Conversion queries:
  10 Euro in USD

- Calculation queries:
  110*(654/8+3)

- Natural language questions:
  What is the time in Hong Kong
Search results

Google

what is the time in Hong Kong

Ergebnisse für what is the time in Hong Kong
Stattdessen suchen nach: what is the timin Hong Kong

00:35 Samstag (HKT) - Zeit in Hongkong

Current local time in Hong Kong – Hong Kong
www.timeanddate.comROID World Clock - Diese Seite übersetzen
Find out current local time in Hong Kong – Hong Kong. Get Hong Kong’s weather and area codes, time zone and DST. Explore Hong Kong’s sunrise and sunset.

Current local time in Hong Kong
www.worldtimezone.com/current_time_in_HK.. - Diese Seite übersetzen
4 hours ago – Hong Kong - get the current local time and date before placing a telephone call or making travel plans for a cheap flight or discount hotel.

Hong Kong Time: What is the actual current local time now in Hong...
www.greenwichmeantime.com/time.../hong-kon.. - Diese Seite übersetzen
24 Jan 2012 – What’s the time in Hong Kong now? What time zone is Hong Kong in? Does Hong Kong observe daylight savings time / summer time?

Current time in Hong Kong - World Time Clock & Map
24timezones.com/.../current_hong_kong_time.p.. - Diese Seite übersetzen
Hong Kong current time in Hong Kong, China daylight time change date 2012 Hong Kong world clock.

Time in Hong Kong - World Time Clock & Map
24timezones.com/.../time_in_hong_kong.php - Diese Seite übersetzen
Exact time in Hong Kong time zone now. Official Hong Kong timezone and time change dates for year 2012.
Ranking results

➢ **By relevance**
  ➢ Binary (relevant or non-relevant).
  ➢ Query/information-need dependent (i.e., similarity estimation between query terms and each document)

➢ **By importance**
  ➢ Authority/popularity of result page (e.g., likelihood of being useful to a majority of users)
    ➢ Links pointing to that page (PageRank, HITS)
    ➢ Estimation of click-through-rate
    ➢ ...
  ➢ Interestingness & context (e.g., likelihood of being useful for specific user in specific context ➔ personalization)
  ➢ Freshness (e.g., most recent news).

Modern search engines combine relevance and importance criteria for ranking!
Clustering results

Source: http://yippy.com/

Grouping topical categories

Top 163 results of at least 1,640,000,000 retrieved for the query HTML (definition) (details)

- noun - hypertext markup language, hypertext mark-up language, HTML -- (a set of tags and rules (code) for hypertext documents)

Nasal Instillation

A nasal installation is a medicine solution prepared for administration into the nose, either by drops or nasal sprays. Purpose: The purpose of a nasal instillation is to deliver medicine directly into the nose.

HTML.COM
Communicate with your customers using our new HTML newsletter tool! www.html.com
www.html.com - [cache] - Additional Sources, Bing

HTML - Wikipedia, the free encyclopedia

HyperText Markup Language (HTML) is the main markup language for web pages. HTML elements are the basic building-blocks of web pages. HTML is written in the … en.wikipedia.org/wiki/HTML en.wikipedia.org/wiki/HTML - [cache] - Additional Sources, Bing, Gigablast
Visualizing result clusters

Source: [http://search.carrotsearch.com/](http://search.carrotsearch.com/)
Visualizing result clusters
Human edited categories in DMOZ

Top: Regional: Europe: Germany (3,986)

- Localities (1)
- States (3,273)

- Arts and Entertainment (34)
- Business and Economy (227)
- Education (10)
- Employment (5)
- Government (103)
- Guides and Directories (11)
- Health (15)
- Maps and Views (5)
- News and Media (21)
- Real Estate (7)
- Recreation and Sports (40)
- Science and Environment (56)
- Shopping (23)
- Society and Culture (127)
- Transportation (9)
- Travel and Tourism (5)
- Weather (7)

Top: Regional: Europe: Germany: Arts and Entertainment

- Berlin (86)
- Bremen (4)
- Hamburg (2)
- Saarland (4)

- Architecture (3)
- Bodyart Studios (1)
- Clubs and Venues (0)
- Dance (9)
- Film Festivals (11)
- Film Societies (0)
- Galleries (31)
- Literature (10)
- Magicians (5)
- Modeling Agencies (13)
- Museums (0)
- Music (17)
- Theatre (2)
Using knowledge bases: Google’s knowledge graph
Using social networks: Bing’s social search

**Recent searches**
- sea food
- honolulu
- chicago weather

**Search results**

**Seattle restaurants**
- IL Fornaio Restaurant: Fresh Pastas, Sauces and Breads. Private Banquet Rooms Available. 600 Pine Street, Seattle · (206) 430-1556
- Seattle Restaurants | Urbanspoon: www.urbanspoon.com/loc/11/Saattle-restaurants
- May 02, 2012: Seattle restaurant reviews from critics, food blogs and fellow diners.

**Ray's Boathouse**
- Recent reviews:
  - OpenTable: Great location. Excellent food! I'll definitely choose this place again for my future guests! ... Read more
  - OpenTable: Great place to celebrate! Melt in your mouth fish options, dreamy mashed potatoes and ... Read more

**Cuoco**
- bing.com/local
  - 310 Terry Ave N, Seattle · (206) 971-0710
    - Directions · Menu · $30 - $40

**Ray's Boathouse**
- bing.com/local
  - 6049 Seaview Ave NW, Seattle · (206) 789-3770
    - Directions · Menu · $40 - $60

**Serious Pie**
- bing.com/local

**Activity**
- Susan Burk posted “Okay... so I want to go again. Who's with me?”

- Sara Davis posted “We decide to go to St Louis by train, I'm curious how to go from the St Louis Train station to St Louis...”
Semantic search: question answering

Who was president of us when obama was born?

You asked: Who was president of us when obama was born?

John F. Kennedy

John Fitzgerald "Jack" Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK, the 35th President of the United States, serving from 1961 until his assassination in 1963.

Rate this answer:

Available on the App Store

How do we know?
Computational question answering

Population and Family Planning Law

One-Child Policy

Population and Family Planning Law
Community-based question answering

What is the longest river in Germany?

In: Germany [Edit categories]

Moving To-From Germany  www.europemove.com
Get an instant Quote now! Fully Insured, Secure Removals.

Cheap Tickets (original)  www.CheapTickets.de
super billige Flüge online suchen, schnell finden und direkt buchen!

Answer:

The Danube is the longest river of all running through Germany. But in Germany the Rhine is the longest river.
Mobile web search

- Multimodal (videos, images, sounds, ...)
- Personalized
- Context-aware
- Natural language
- Exact (answers)
- Social
- ...
In a nutshell, Goggles lets users search for objects using images rather than words. Simply take a picture with your phone’s camera, and if we recognize the item, Goggles returns relevant search results. Right now Goggles identifies landmarks, works of art, and products (among other things), and in all cases its ability to "see further" is rooted in powerful computing, pervasive connectivity, and the cloud:

- We first send the user’s image to Google’s datacenters
- We then create signatures of objects in the image using computer vision algorithms
- We then compare signatures against all other known items in our image recognition databases; and
- We then figure out how many matches exist; and
- We then return one or more search results, based on available meta data and ranking signals; and
- We do all of this in just a few seconds
Social annotation for image search

- Game with a purpose (exploits “wisdom of the crowds”)
- Project initiated by Luis von Ahn et al. In 2004
- Estimation: 5000 people could tag all images on the web in 4 weeks
Social web (web 2.0)

- **Social collaboration**
  - Tagging, answering, computing, rating, editing, recommending, discussing, ...
  - Examples: Wikipedia, flickr, del.icio.us, librarything, Amazon, Netflix, answers.com, ESP, foldit, ...

Web 2.0 concepts include:

- Folksonomy
- Participation
- Usability
- Simplicity
- Design
- CSS
- AJAX
- Mobility
- RSS
- Remixability
- DataDriven
- Standardization
- Microformats
- REST
- SOAP
- Modularity
- Accessibility
- Microformats
- The Long Tail
- SEO
- XML
- User Centered
- Joy of Use
- widgets
- Browser
Monetization of web search - sponsored search

Mainline slots

Sidebar slots

Organic Search Results

Paid Search Results

1. Investment Opportunities
2. Personal Online Investing
3. Invest in Stocks for $4
4. How to Invest Small Amounts of Money Wisely - wikiHow
5. Investing Money
6. How to Become a Millionaire - The Top 5 Tips
7. 4 Ways To Invest Money Wisely - Financial Web
8. Invest Wisely, Mutual Funds
For Advertisers: **Google AdWords**

Reach people when they are actively looking for information about your products and services online, and send targeted visitors directly to what you are offering. With AdWords cost-per-click pricing, it’s easy to control costs—and you only pay when people click on your ad.

For Web Publishers: **Google AdSense**

Earn more revenue from your website, while providing visitors with a more rewarding online experience. Google AdSense™ delivers ads targeted to your content pages and, when you add Google WebSearch to your site, AdSense delivers targeted ads to your search results pages too. With AdSense you earn more ad revenue with minimal effort—and no additional cost.
Worldwide market share (as of May 2012)

Source: [http://www.karmasnack.com/about/search-engine-market-share/](http://www.karmasnack.com/about/search-engine-market-share/)
Indexed web pages

Source:
http://www.worldwidewebsize.com/
Deep web

- Content of the web that
  - resides in backend databases and
  - is not accessible through hyperlinks
  - can only be surfaced by submitting queries to query forms

- Examples
  - Yellow pages
  - Scientific databases (bio-medical-, geo-, weather-related data, publications, measurements, ...)
  - Census data (i.e., population analysis, statistics, etc.)

- Estimation: several times larger than the visible web
- High-quality & structured information
Deep web: example (1)

Discover over 70,000+ searchable databases and specialty search engines.

Search these Deep Web sources simultaneously with the Deep Query Manager. Learn more...

Note: Your query will search All Topics

Source: http://aip.completeplanet.com
Deep web: example (2)

Source: http://ned.ipac.caltech.edu
Sufacing the deep web

- **Strategy 1** (used by Google, see Madhavan et al. 2006):
  - Probe query boxes in the form with random keywords
  - Use words in the results as new query terms (bootstrapping)

- **Strategy 2** (much more challenging -- semantic analysis):
  - Infer the data type (or semantic concept) behind a query box by analyzing scripts (for required values), exploiting surrounding terms, or probing
  - Surface information by querying with appropriate values (e.g. from other databases / knowledge bases)
  - Analyze associations in surfaced data to infer underlying schema

- **Problems**
  - Map queries to relevant interfaces
  - Integrating schemas of sources from similar domains
  - Merging results into global ranking
Related literature

- **Books**
  - Modern Information Retrieval
  - **Important conferences**
    - SIGIR, TREC, WSDM, WWW, WebSci, CIKM, ECIR, ICML, ECML, KDD, JCDL, ...
  - **IR toolkits**
    - Lucene (http://lucene.apache.org/core/)
    - The Lemur Project (http://www.lemurproject.org/)
What else can we expect from future search?

Source: funnyjunk.com