

course:

Searching the Web (NDBI038)

Searching the Web and Multimedia Databases (BI-VWM)

© Tomáš Skopal, 2020

lecture 1:

Web space, search engines, web retrieval modes

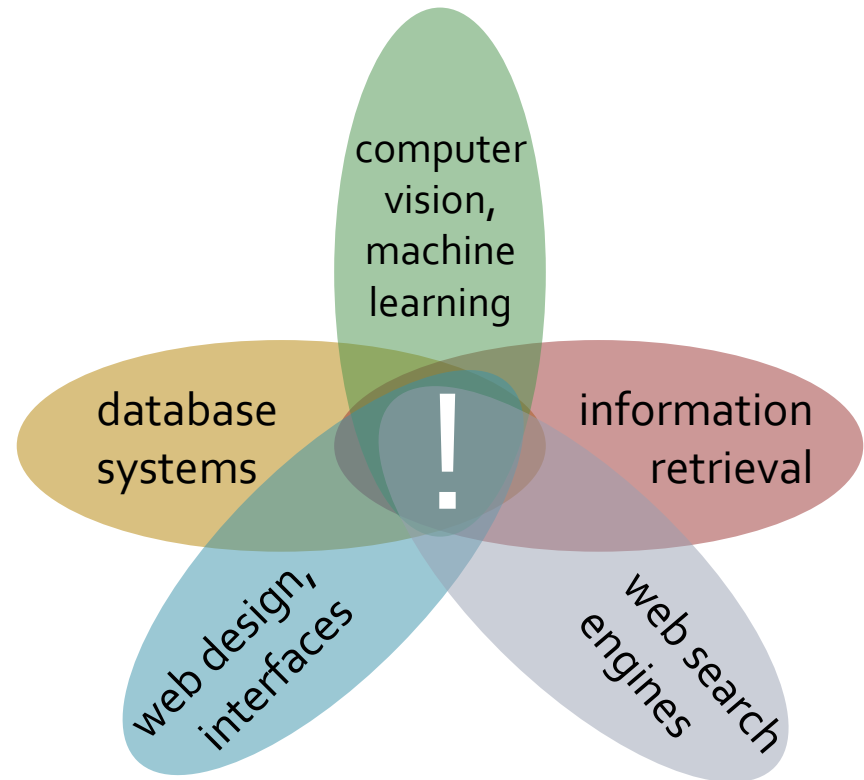
prof. RNDr. Tomáš Skopal, Ph.D.

Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague

Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague

The course focus

Searching the Web



What is the course about?

- it is about
 - retrieving information from web (databases)
 - searching, browsing, querying
 - web pages (text, hyper-text, social networks)
 - multimedia documents/objects (content-based search)
- it is not about
 - general web application architectures
 - user interfaces not related to retrieval
 - networking, protocols, and other low-level infrastructure

Today's lecture outline

- the web space
 - data, multimedia, and communities on the web
- web search engines
 - history
 - web crawling, indexing, searching
 - multimedia retrieval
- web retrieval modes
 - browsing, queries, filtering
- software architectures for multimedia retrieval
 - search engines, hosting servers, enterprise applications

What is the Web space?

- World Wide Web (WWW)
 - founded by Tim Berners-Lee (CERN) in 1989
 - an **internet graph of web pages**
+ other resources hosted on web servers
 - communication over the Internet using HTTP
(**hyper-text transfer protocol**)
 - GUI-based internet space
 - presentation and human-readability (HTML code and page rendering)
 - limited readability by machine (syntactic)

What is the Web space?

- web page
 - hyper-text document written in HTML (hyper-text markup language) or descendants, like XHTML, PHP
 - text including links to resources using URL (uniform resource locator)
 - web page, or resource on the internet (multimedia object, etc.)
 - depending on HTML command, the URL source could be shown as link, or downloaded (and embedded) within the referencing web page
 - node in the Web graph (URL links are the out-links from that node)
- web site
 - web subgraph of related web pages, e.g., personal web, newspapers, etc.
 - pages share domain/subpath in URL

Web 1.0 vs. Web 2.0

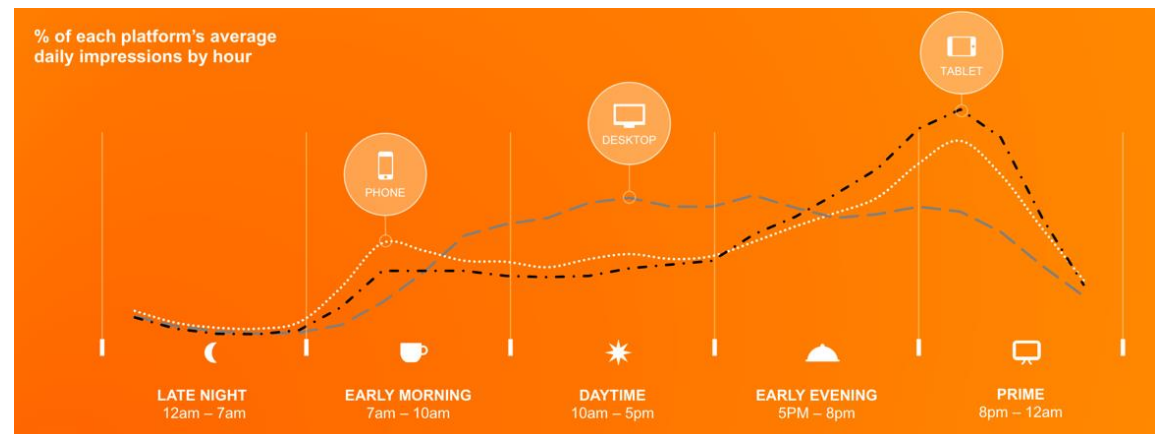
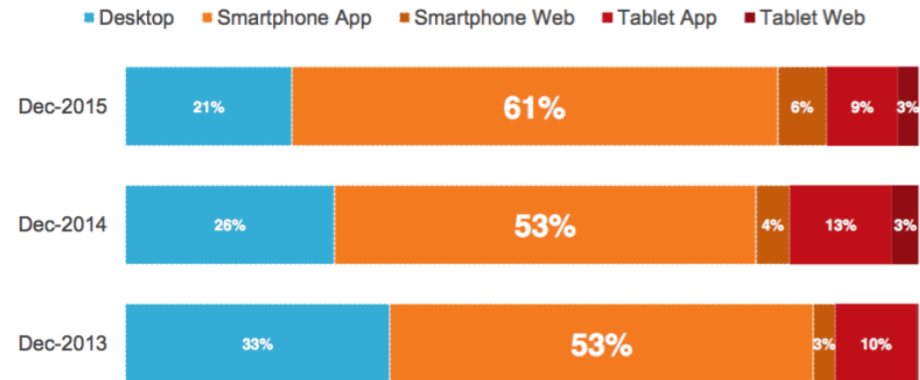
- Web (1.0)
 - the first 15 years of WWW
 - personal websites, publishing, rather static content
 - passive browsing dominant
- Web 2.0
 - the WWW era since 2004
 - available high-speed internet connection
 - social networks, blogs, site aggregations, information society
 - participation instead of passive browsing, interactivity
 - other devices than PCs (smart mobile devices)

Client platforms

- desktop/notebook
 - on decline
- mobile
 - 70% YouTube (2017)
 - 80% social networks (2016)

Share of Time Spent on Social Media Across Different Platforms

Source: comScore Media Metrix Multi-Platform & Mobile Metrix, US, Dec 2015 / Dec 2014 / Dec 2013

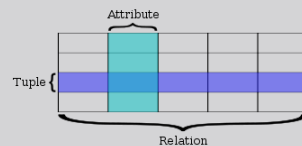


David Chaffey, 2017

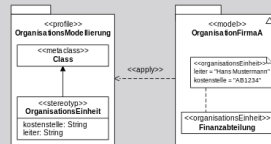
Data on the Web

structured data (with schema)

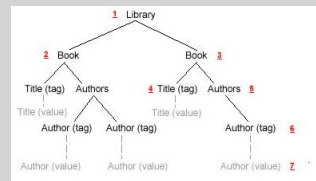
relational (SQL)



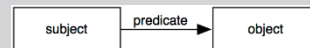
object(-relational)



XML (XPath, XQuery)



RDF (SPARQL)



key-value (NoSQL)

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

NewSQL (ACID + scalability of NoSQL)

time

Big Data era

unstructured data (schemaless)

text

strings

time series

sensory data

multimedia

biometrics

medical/scientific

industrial

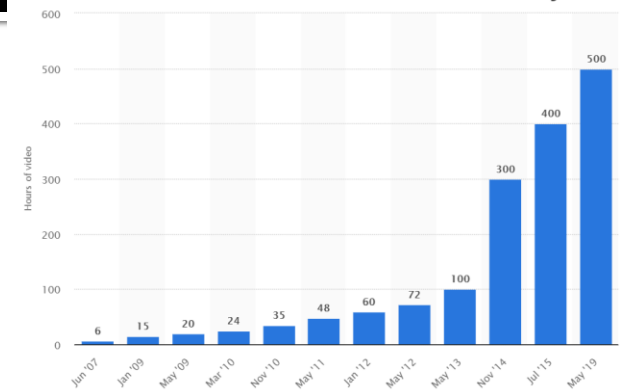
similarity
search

stream/cloud processing

- 400 hrs/min of video uploaded @ YT
- 10 GB/sec from LHC in CERN

Multimedia content on the Web

Hours of video uploaded to YouTube every minute as of May 2019



- > 99% of web space stores multimedia content, mostly at social networks
 - **billions** of photos **per day** uploaded
 - 100M photos+videos/day @ Instagram, **50 billion** in total (2019)
 - 350M photos/day @ Facebook, **250 billion** in total (2019)
 - **500 hrs** of video **per min** uploaded @YT, **5 billion** watched daily (2019)
- factors
 - high-speed internet, increasing computational power, clouds, cheap capturing devices (digital cameras/mobile phones/tablets)
 - everybody is data/information producer (Web 2.0)
 - human activities move to internet (cloud) in a large extent
 - entertainment (social networks), services (e-banking, e-shops, e-gov, ...)

Web information retrieval

- real life moves into the Web at a large scale
 - entertainment – FaceBook, YouTube, Flickr, news
 - work – cloud computing applications (webmail, office, collaborative)
 - shopping – e-Commerce
 - study – e-Learning, student information systems, YouTube tutorials
 - state administration services – e-Government
- in all the mentioned Web activities, the essential is:

retrieval of an information (about entity)

Web search engines

- web search engine
 - the prominent means in the web information retrieval
 - at least keyword-based (full-text) search
 - additional information for ranking acquired from the Web graph (e.g., PageRank, HITS)
- meta-search engine
 - engine that aggregates results returned by several other web search engines
 - could be more effective (e.g., <http://www.copernic.com/>)
 - good for retail business competition (e.g., www.zbozi.cz)

Web search engines – history

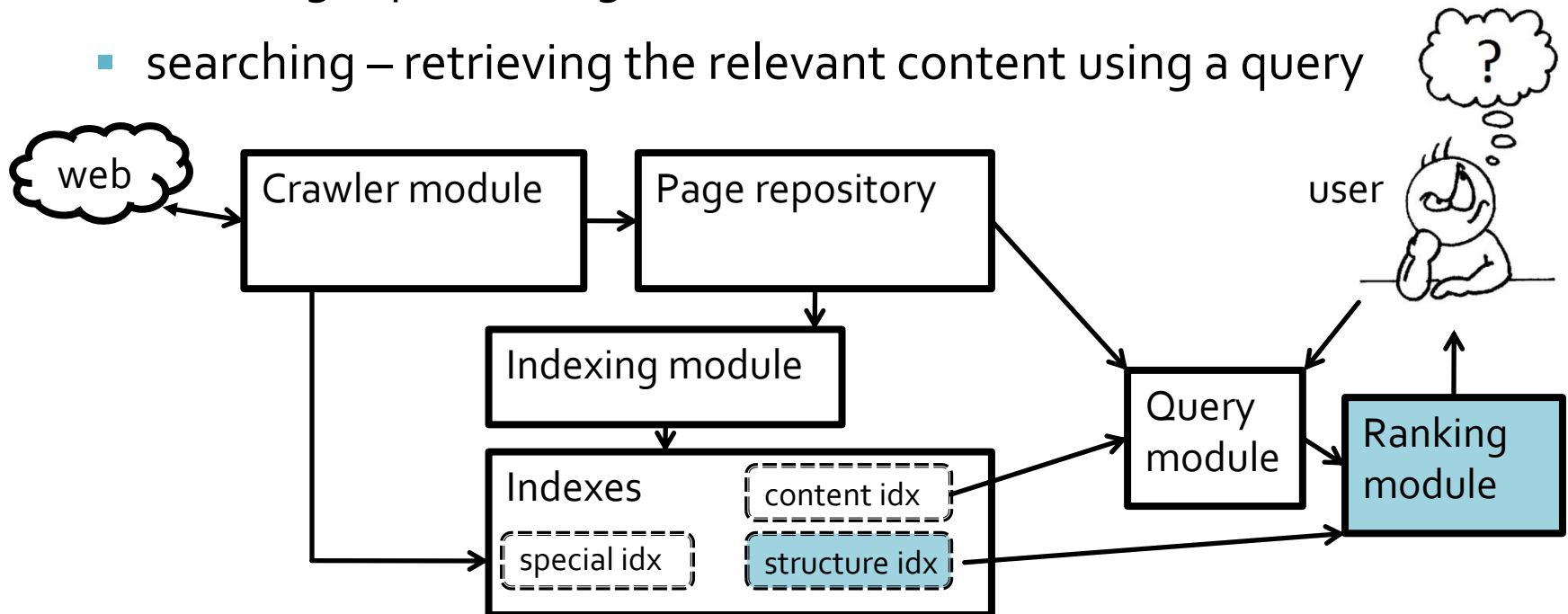
- information retrieval (since 1960s)
 - models and indexing techniques for full-text search
 - older than Web, i.e., not designed for web retrieval, rather digital libraries/archives, collections of full-text documents
 - three basic models – boolean, vector, probabilistic
 - with popularity of Web, it included also the link analysis (1998)
- web retrieval prior to search engines
 - only web directories – edited by humans
 - hard to find any web page not listed!
 - e.g., central list of web servers hosted on a CERN web server

Web search engines – history

- the pioneers (1990-1998)
 - Archie – search in downloaded directory listings of public FTPs (1990)
 - W3Catalog – first search engine in catalogues (manually maintained)
 - World Wide Web Wanderer – first robot (crawler)
 - followed by AltaVista, Inktomi, Yahoo! (though only directory search)
- the Google era (since 1998)
 - PageRank algorithm gave much better results than existing engines based just on the full-text search (classic information retrieval models)
 - success also due to simple GUI, thus fast download in the slow-speed internet age (not downloading Ads and unnecessary portal GUI)
 - followed by Bing, Yahoo! Search...
but Google still has over 90% of the market

Web search engines

- typical elements of a web search process
 - crawling – downloading the content (web pages)
 - indexing – processing the content into a form suitable for search
 - searching – retrieving the relevant content using a query



Web search engines – crawling

- crawler
 - a short program that instructs spiders (or robots/bots/agents) on how and which pages to retrieve
- limited by resources (web size, bandwidth)
 - repeated visits, prioritized visits, optimal policies
- ethic crawling
 - robot exclusion protocol
- other resources on web crawlers
 - book “Spidering Hacks” – tricks and tips for crawler implementation
 - book “Numerical Computing with Matlab” contains an example of crawler written in Matlab, see file surfer.m at www.mathworks.com/moler/ncmfilelist.html

Web search engines – indexing

- crawled web pages need to be organized for searching
- an **index** is built, allowing to process only a small fraction
- many various indexes
 - content index –full-text search
 - structure/citation index – page ranking
 - special purpose index, e.g., for content-based multimedia search
- index design factors
 - merge, storage, size, lookup, maintenance, fault tolerance
- index data structures
 - inverted file, signature file, suffix tree, etc.
- web caches
 - Internet archive project (<http://web.archive.org>)

Web search engines – searching

- traditional web search engines
 - full-text indexing + link analysis
 - keyword or full-text queries
 - keyword query
 - (small) set of terms (words, phrases)
 - full-text query
 - text file (web page), parsed to keyword query
- multimedia web search engines
 - content-based queries (in addition to keyword search)

Multimedia on the Web

- annotation – external description, high-level semantics
 - explicit: keywords, full text, URL, classification, GPS, ...
 - contextual: e.g., relations in social network (comments, likes, shares)
- content-based feature descriptors
 - features extracted from the visual content (image pixels)
 - rather low-level features lacking semantics (semantic gap)
 - using deep learning, visual features become semantic as well
- multi-modal descriptors
 - combination of multiple descriptors
 - using deep learning, the content and annotation fuses together

Multimedia on the Web

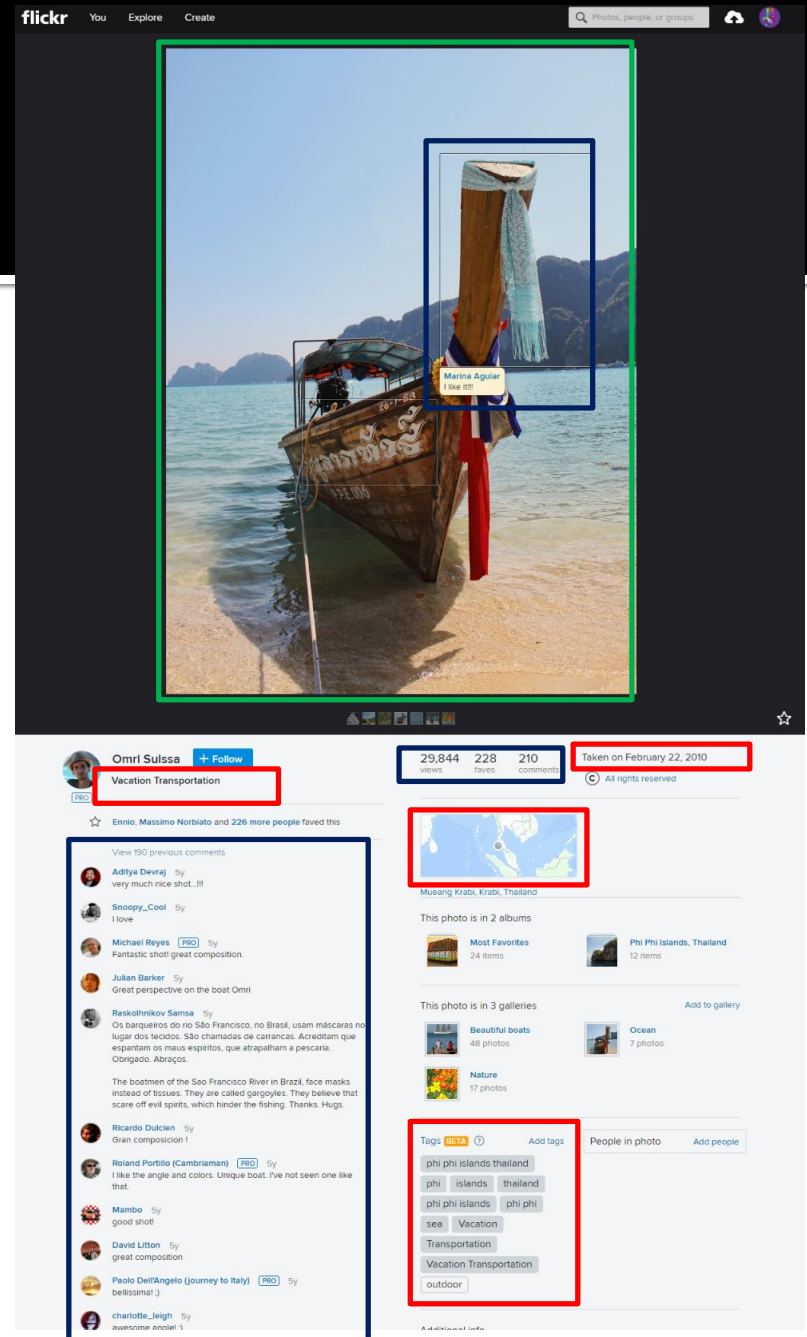
Example:

a photograph hosted at Flickr using keyword “vacation”

visual content (image pixels)

explicit annotation

contextual annotation



Modes of web retrieval

- query
- browsing
- filtering (recommendation)
- combinations
 - query + browsing
 - browsing + query
 - filtering + query
 - etc.

Modes of web retrieval – query

- assumption
 - we are able to specify our search intent
- query = explicit formulation of one-shot search intent
- query models
 - keyword-based (annotation based)
 - content-based
 - file upload, URL, sketch
- query result
 - binary relevance – unordered set of objects
 - multi-value relevance – ranking on database objects
- relevance feedback, re-ranking

Modes of web retrieval – query

- keyword query

- multiple answers

- web pages
- multimedia
- other...

The screenshot shows a Google search interface. The search bar contains the text "vacations in Hawaii". Below the search bar, it says "About 4,160,000 results (0.35 seconds)". To the left of the search bar is the Google logo. Below the search bar is a sidebar with filters: "Everything", "Images", "Videos", "News", and "More". There is also a "Search near..." section with a text input "Enter location" and a "Set" button. Below that are time filters: "Any time", "Latest", "Past 24 hours", "Past week", "Past month", "Past year", and "Custom range...". At the bottom of the sidebar is a link to "More search tools". The main search results are displayed on the right. A red box highlights the first five results, which are all web pages. The results are: "Resorts Hawaii", "Hawaii Vacation Rentals", "Vacation Packages To Hawaii", "Hawaii Vacations | Hawaii Vacation Packages and Deals", and "Hawaii's Official Tourism Site -- Travel Info for Your Hawaii Vacation". Each result includes a title, a brief description, and a URL. A red arrow points from the "multimedia" box in the list on the left to the "Images" filter in the sidebar. Another red arrow points from the "web pages" box in the list on the left to the first result in the search results.

Google

vacations in Hawaii

About 4,160,000 results (0.35 seconds)

Advanced search

Everything
Images
Videos
News
More

Search near...
Enter location Set

Any time
Latest
Past 24 hours
Past week
Past month
Past year
Custom range...
More search tools

Resorts Hawaii
Beautiful Beach Resort Close to Hawaii Attractions. From \$199/Night
www.hiltonhawaiianvillage.com

Hawaii Vacation Rentals
The Finest Hawaii Vacation Rentals From \$600-\$6800/nt. Request Today!
luxuryretreats.com/Hawaii/Rentals

Vacation Packages To Hawaii
Visiting Hawaii? Find Deals & Read Hotel Reviews!
London Hotels - Paris Hotels - New York City Hotels - Rome Hotels
tripadvisor.com/Hawaii

Hawaii Vacations | Hawaii Vacation Packages and Deals ☆ 🔍
29 Jan 2011 ... Planning a Hawaii vacation? Experience the best Hawaii has to offer without spending a fortune on your Hawaii vacation.
Packages and Flights - Hawaiian Airlines Sale - Explore Maui - Explore Oahu
www.hawaii.com/ - Cached - Similar

Hawaii's Official Tourism Site -- Travel Info for Your Hawaii Vacation ☆ 🔍
The People of Hawaii would like to share their Islands with you. The fresh, floral air energizes you. The warm, tranquil waters refresh you.
www.gohawaii.com/ - Cached - Similar

Hawaii Vacations (Discount Hawaiian Vacation Packages) ☆ 🔍
Hawaii Aloha Travel offers the great discount on Hawaii vacations and hawaiian vacation packages. Hawaii Hotels, Airfare, and all inclusive Hawaii vacations ...
www.hawaii-aloha.com/ - Cached - Similar

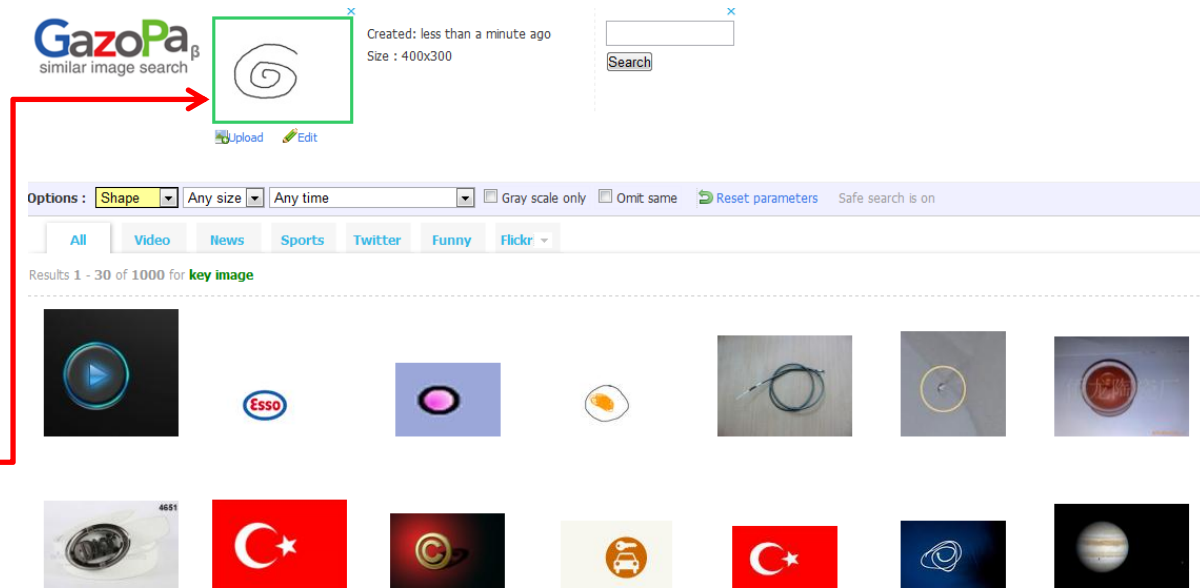
Hawaii Vacation Packages - Cheap Hawaii Vacations - Hawaii Family ... ☆ 🔍
Best Hawaii Vacation Packages on Expedia - Check out Cheap Hawaii Vacations deals and Discount All Inclusive Hawaii Family Vacations at Expedia.
www.expedia.com/.../vacations/hawaii/default.asp - United States - Cached - Similar

Modes of web retrieval – query

- content-based query

- image search, query by

- sketch
- example



Give the url of a Web image:

...or select a file on your computer:

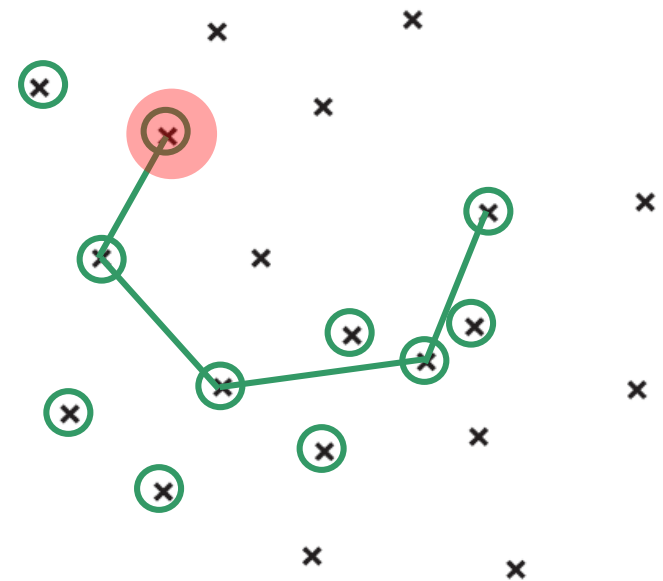
...or search the text associated to images:

Modes of retrieval – browsing

- assumption
 - we are not able to (well) specify our search intent
- browsing = iterative navigation in the database
- browsing models
 - explicit graph (links between entities)
 - virtual graph (series of queries)
- browsing result
 - subset of database objects
 - set of clusters (representatives)
 - hierarchy (ontology)

Modes of retrieval – browsing

- browsing explicit graph
 - static hierarchy of categories
 - dynamic network
- browsing virtual graph
 - series of queries
 - an object in query result is used to specify the new query



Example – browsing

- explicit graph
 - links in user's portfolio
- implicit graph
 - links to similar pages/objects
 - query by keywords obtained from the annotation
 - weighted keywords
 - weight = font size

Keywords (Report | Suggest)

active animal attentive breed canine carnivore command
companion cute daylight dog energetic face fur furry
golden lab labby labrador male mammal outdoor pedigree
pet portrait pose predator protector pure purebred rescue
retriever retrieving side social stand trained watchdog yellow



Add to lightbox

Click for comp image

Share

- click image to zoom -

More similar stock images of "Yellow Labrador Dog 2"



More images

More stock photos from user portfolio



(credit: dreamstime.com)

Modes of retrieval – filtering

- filtering = formulation of fixed search intent
 - explicit = static request
 - query, subscription
 - implicit = recommending
 - based on user preferences, profile, search history, collaborative filtering, etc.
- filtering result
 - dynamic, changing in time
 - like database view

Example – explicit filtering

- RSS channel + reader
 - filtering/recommendation tool
- keyword query

(credit: fotolia.com)

The screenshot shows the Fotolia website interface. At the top, there's a search bar with 'dog labrador' entered. Below the search bar, it says 'Search results: 5063 files'. The main content area displays a grid of image thumbnails. The first thumbnail, showing a close-up of a dog's face, is highlighted with a red rectangle. To the left of the grid, there are filters for 'Category Search', 'Gallery Search', 'Order by' (Relevance, Date, Price, Downloads, Popularity), 'Images' (Standard, Infinite, All), 'Type' (Photo, Illustration, Vector, Video, All), and 'Orientation' (Horizontal, Vertical, All).

The screenshot shows a Windows Internet Explorer browser window displaying search results for 'dog labrador' on the Fotolia website. The address bar shows the URL: [http://rss.fotolia.com/?l=dog+labrador&filters\[content_type%3Aphoto\]=1&filters\[content_type%3Aphoto\]=1&filters\[content_type%3Aphoto\]=1](http://rss.fotolia.com/?l=dog+labrador&filters[content_type%3Aphoto]=1&filters[content_type%3Aphoto]=1&filters[content_type%3Aphoto]=1). The search results are displayed in a list format. The first result is 'Kitten and a pup together.' followed by 'Chien avec un panneau blanc', 'Horse and dog', 'joy', and 'Dog'. The 'Dog' result is highlighted with a red rectangle. The right sidebar shows 'Displaying 32 / 32' items, 'Sort by: Date', and 'Filter by category:' with a list of categories and their counts.

Example – explicit filtering

- subscription to a “channel”
- membership in a group
- etc.

The screenshot shows a YouTube channel page for 'Planes, Cockpits & Flights HD' by neocastillo. The main video player displays a Continental airplane on a tarmac. Below the video, the title 'Aeropuerto Merida 1:1 -HD-' is shown, along with the upload date (September 11, 2010) and view count (1,171 views). The right sidebar features a search bar and a list of recommended videos, including 'Aeropuerto Merida 1:2 -HD-', 'Aeropuerto Merida 1:3 -HD-', 'Cessna 210 Centurion Turbo - Rolls Royce', 'Cockpit - Hawker Siddeley HS-125-1A', 'Estafeta Carga Aerea - Boeing 737-300', and 'Hawker BeechJet 400A Cockpit - HD'.

YouTube

Search Browse Upload

Create Account Sign In

Planes, Cockpits & Flights HD
neocastillo's Channel

Subscribe Uploads

Search

Date Added | Most Viewed | Top Rated

Aeropuerto Merida 1:1 -HD-
1,171 views - 2 weeks ago

Aeropuerto Merida 1:2 -HD-
278 views - 1 week ago

Aeropuerto Merida 1:3 -HD-
171 views - 5 days ago

Cessna 210 Centurion Turbo - Rolls Royce
2,006 views - 1 month ago

Cockpit - Hawker Siddeley HS-125-1A
619 views - 2 months ago

Estafeta Carga Aerea - Boeing 737-300
8,776 views - 8 months ago

Hawker BeechJet 400A Cockpit - HD

www.youtube.com/neocastillo

00:00 / 14:59 360p

Info Favorite Share Playlists Flag

Aeropuerto Merida 1:1 -HD-

From: neocastillo | September 11, 2010 | 1,171 views

Merida Airport First Compilation mixed with
Music: Vonyo Sessions CD1 by Paul van Dyk

I love Planes and Trance Music enjoy!

[View comments, related videos, and more](#)


Like

... (more info)


Example – implicit filtering

- implicit filtering (recommending) as an automatic query based on user's previous searches
 - history of viewing content serves as relevance feedback

Sponzorováno Vytvořit reklamu



zalando.cz
Potřebujete inspiraci pro svůj outfit? Objevte nejnovější trendy na Zalando. 🍌🍌











490 Kč – Objednejte si svůj bezkontaktní ter...
sumup.cz
☒ Žádné další smlouvy. ☒ Žádné měsíční poplatky. ☒ Žádné závazky. Proč? Protože nic z toho...

YouTube

Search Browse Upload

Recommended for You [Learn More](#)

Edit

 Trigger The Bloodshed - The 1 year ago 139,652 views <i>Because you watched Instant Species -...</i>	 The White Stripes - Fell In Love... 4 years ago 6,000,137 views <i>Because you watched Instant Species -...</i>	 OMD Sailing On The Seven Seas 4 years ago 1,102,354 views <i>Because you watched Instant Species -...</i>	 Fell in Love with a Girl Live 3 years ago 92,617 views <i>Because you watched Instant Species -...</i>
 The White Stripes - Fell In Love... 3 years ago 760,678 views <i>Because you watched Instant Species -...</i>	 Walking With A Ghost by Tegan an... 3 years ago 1,381,024 views <i>Because you watched Instant Species -...</i>	 ENTER SHIKARI - JUGGERNAUTS - of... 1 year ago 2,003,711 views <i>Because you watched Instant Species -...</i>	 Tinchy Stryder - Star in the hoo... 2 years ago 10,605 views <i>Because you watched Instant Species -...</i>

Example – combination

The screenshot shows the Amazon website's search results for the query "information retrieval". The search bar at the top is circled in red, with a red line pointing to a box labeled "query". The first search result, "Introduction to Information Retrieval" by Christopher D. Manning and Prabhakar Raghavan, is also circled in red, with a red line pointing to a box labeled "recommendation". The star rating for this book is circled in red, with a red line pointing to a box labeled "re-ranking". The left sidebar contains filters for "Books", "Kindle Store", and "Refine by". The main content area lists three books with their covers, titles, authors, prices, and ratings.

amazon
Shop by Department
Your Amazon.com Today's Deals Gift Cards Sell Help
Hello, Sign in Your Account Try Prime Wish List Cart

1-16 of 58,781 results for "information retrieval" Choose a Department to sort

Show results for

Books >

- Computers & Technology
- Computer Network Administration
- Network Storage & Retrieval Administration
- Databases & Big Data
- Computer Programming
- + See more

Kindle Store >

- Computer Databases
- Computers & Technology
- Computer Programming
- Linguistics

+ See All 26 Departments

Refine by

International Shipping

☐ Ship to Czech Republic

Eligible for Free Shipping

Free Shipping by Amazon

Book Format

- Hardcover
- Paperback
- Kindle Edition

Computer User Books

- Beginners & Seniors
- Advanced & Power Users
- Kids & Teens

Book Language

☐ English

Introduction to Information Retrieval Jul 7, 2008
by Christopher D. Manning and Prabhakar Raghavan

Hardcover

\$25.92 to rent **\$64.79** to buy
Get it by **Thursday, Oct 8**

More Buying Choices
\$31.26 used & new (72 offers)

Kindle Edition
\$50.20
Auto-delivered wirelessly

Other Formats: Paperback

Information Retrieval: Implementing and Evaluating Search Engines Jul 23, 2010
by Stefan Büttcher and Charles L. A. Clarke

Hardcover

\$41.48 ~~\$66.00~~
Only 14 left in stock - order soon.

More Buying Choices
\$41.48 used & new (35 offers)

Kindle Edition
\$47.68
Auto-delivered wirelessly

Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books) Feb 10, 2011
by Ricardo Baeza-Yates and Berthier Ribeiro-Neto

Paperback

\$63.74 ~~\$74.99~~
Get it by **Thursday, Oct 8**

More Buying Choices
\$57.74 used & new (35 offers)

Search Engines: Information Retrieval in Practice Feb 16, 2009
by Bruce Croft and Donald Metzler

Paperback

query

recommendation

re-ranking

Example – combination

- known-item search in video collections
 - mental query or no initial query at all
 - interactive search (browsing + querying)

