course:
Searching the Web (NDBI038)
Searching the Web and Multimedia Databases (BI-VWM)

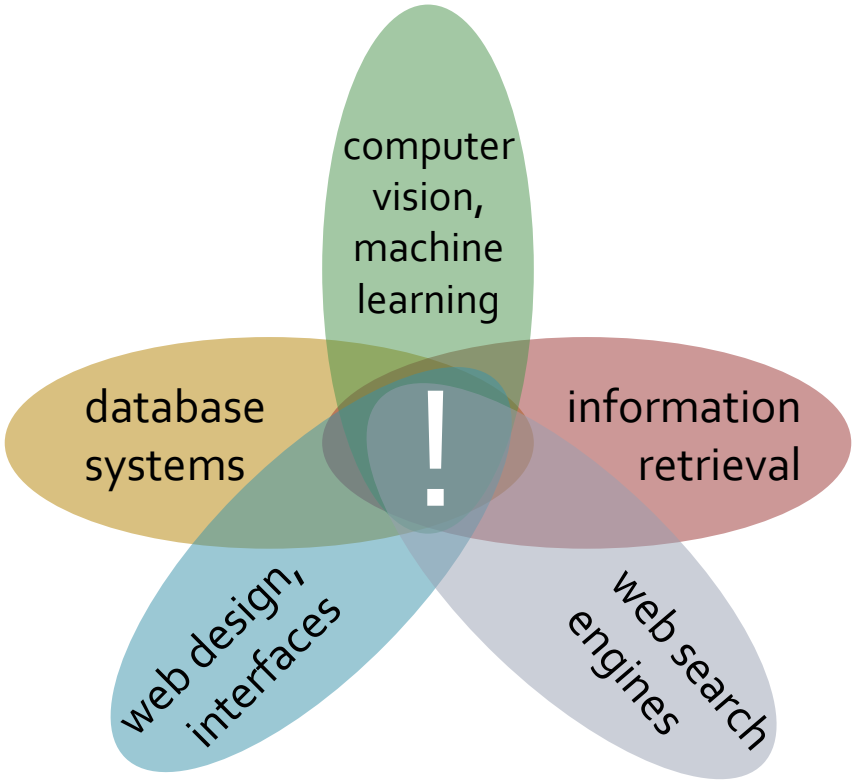
lecture 1:

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
    (**h**yper-**t**ext **t**ransfer **p**rotocol)
  - GUI-based internet space
    - presentation and human-readability (HTML code and page rendering)
    - limited readability by machine (syntactic)

What is the Web space?

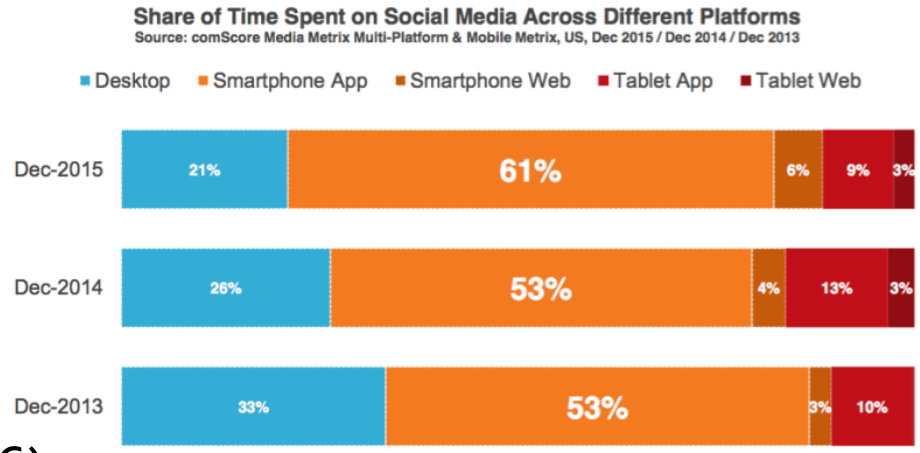
- web page
  - hyper-text document written in HTML (**h**yper-**t**ext **m**arkup **l**anguage) or descendants, like XHTML, PHP
  - text including links to resources using URL (uniform resource locator)
    - web page, or resource on the internet (multimedia object, etc.)
    - depending on HTML command, 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

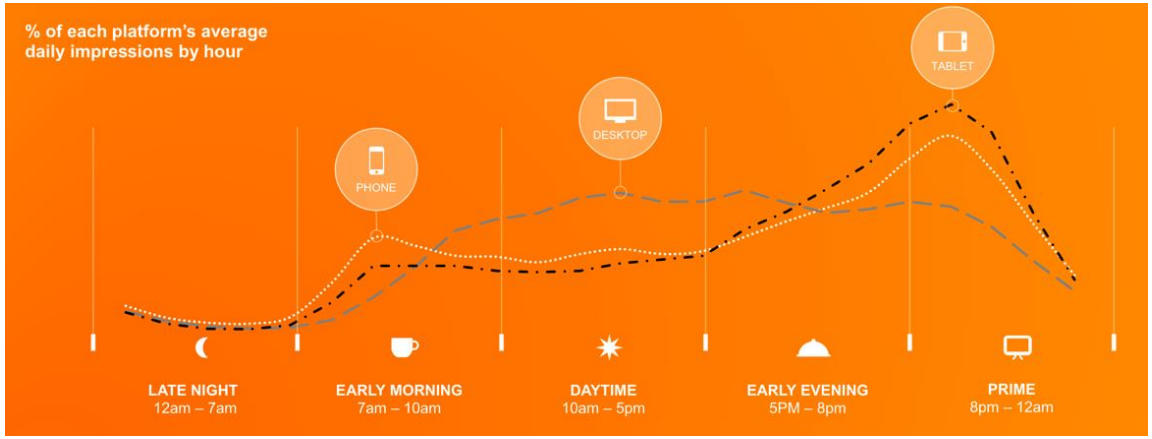
Web 1.0 vs. Web 2.0

- 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)



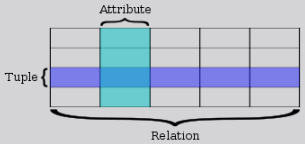


David Chaffey, 2017

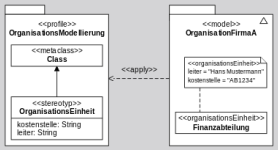
Data on the Web

structured data
(with schema)

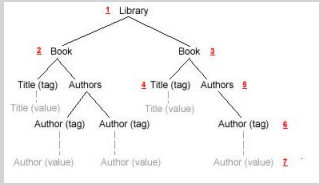
relational (SQL)



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

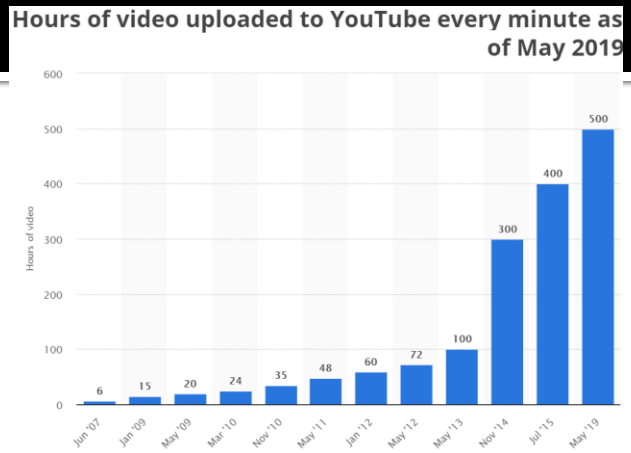
time series

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



- > 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)

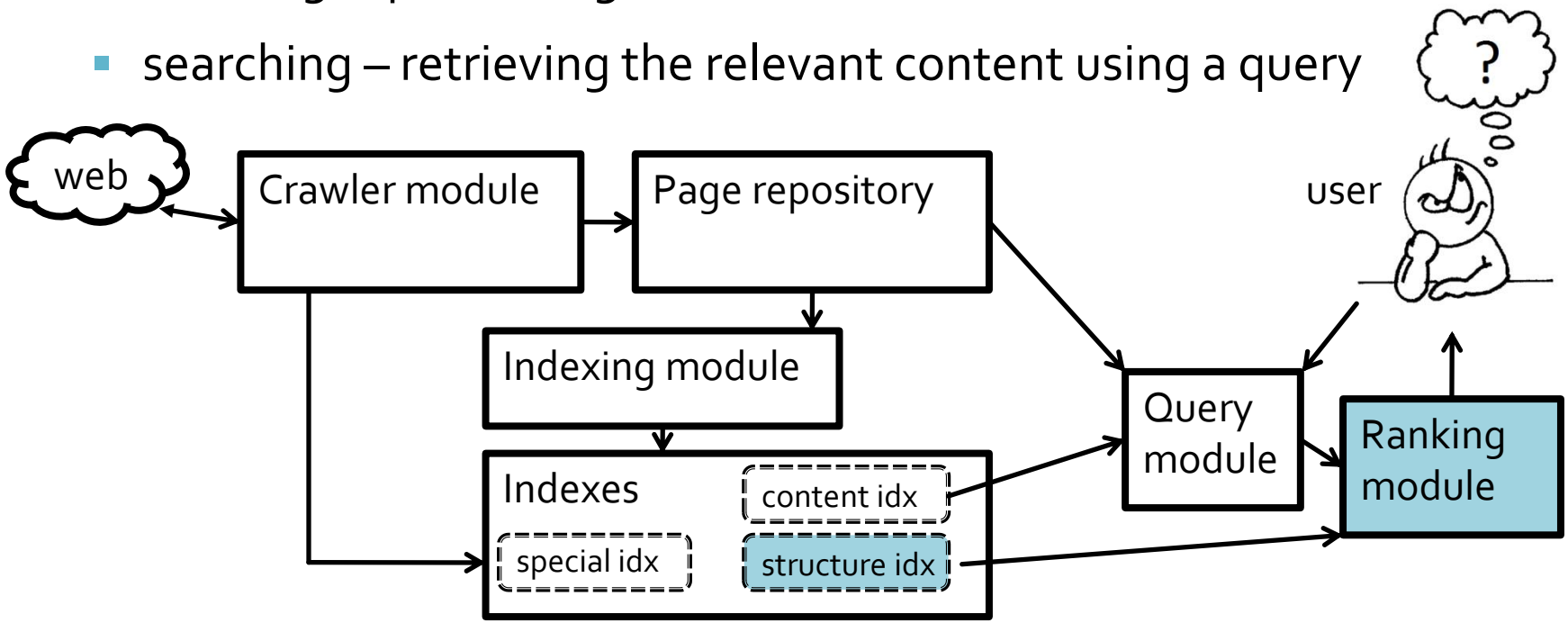
- web retrieval prior to search engines
  - only web directories – edited by humans
  - hard to find any web page not listed!
  - e.g., central list of web servers hosted on a CERN web server

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

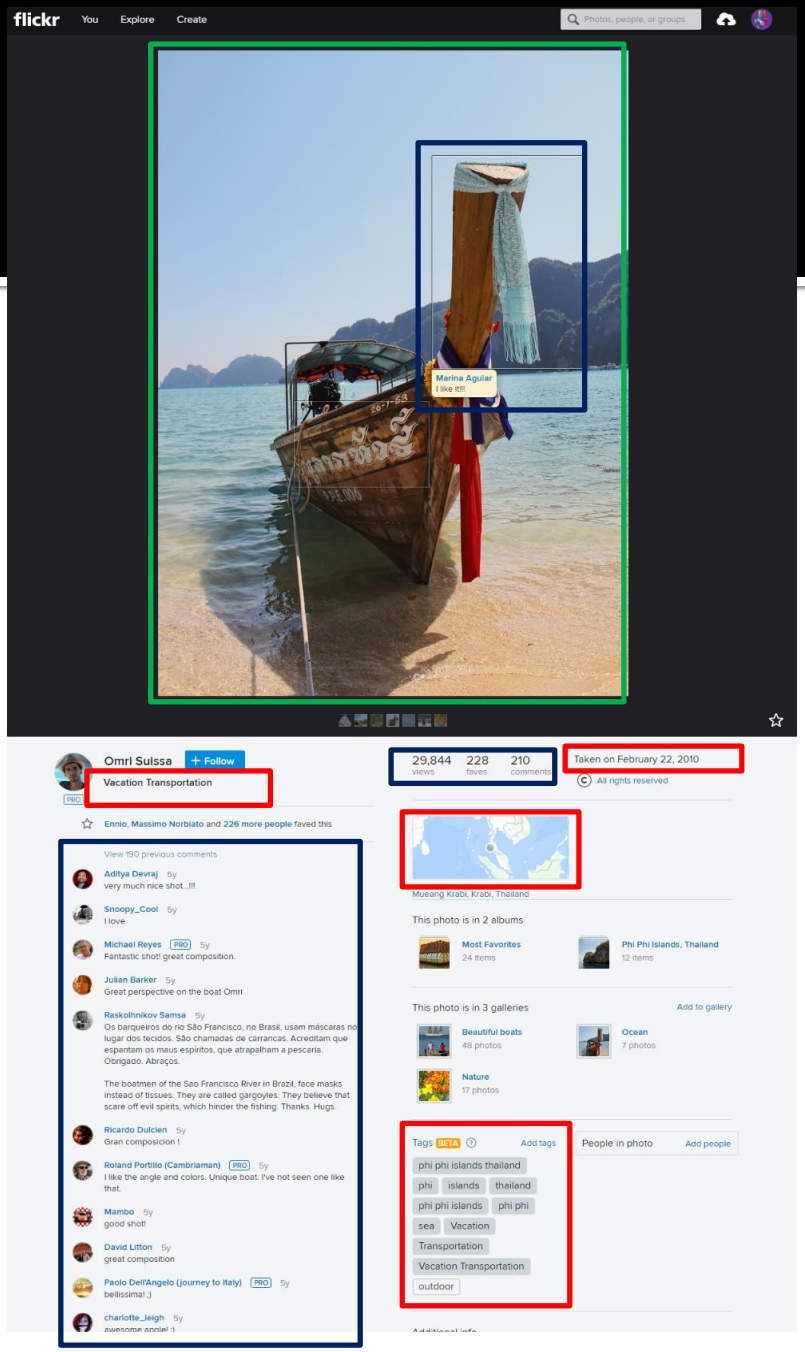
Multimedia on the Web

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)

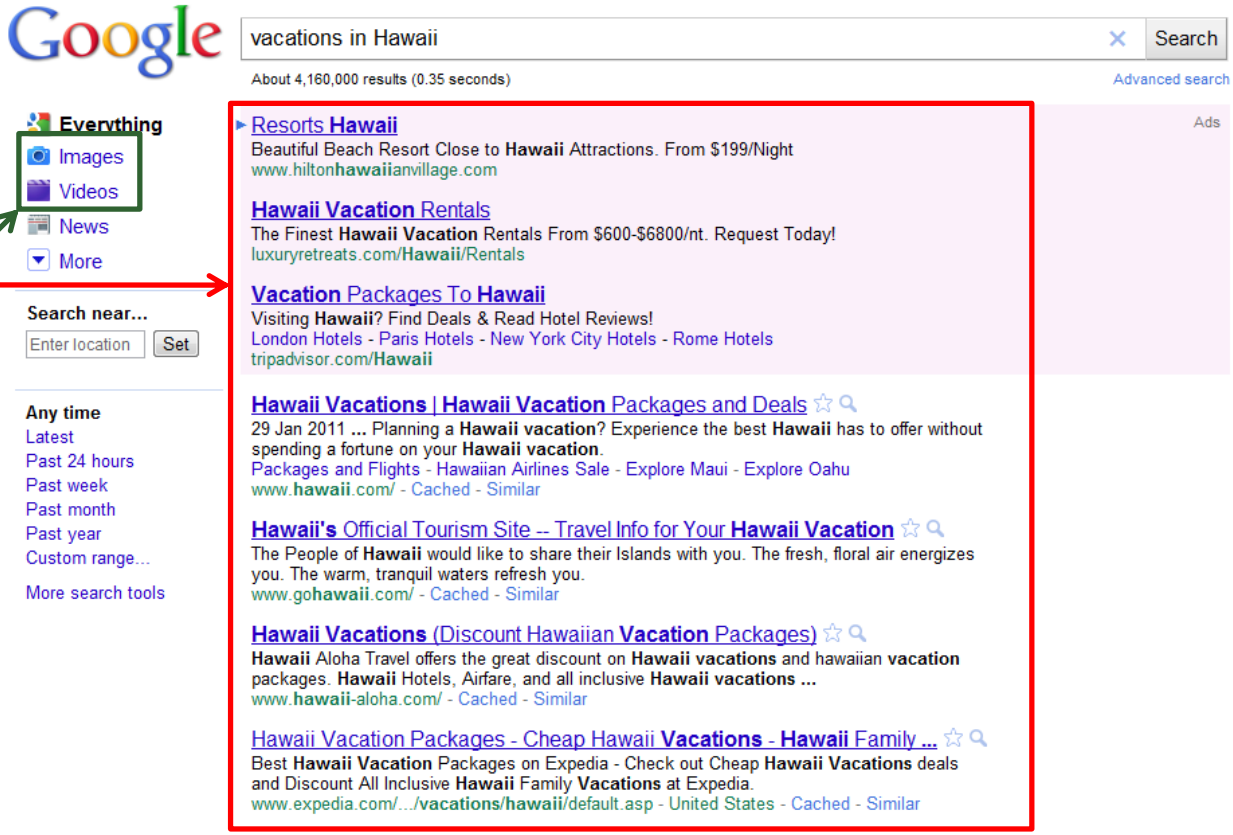
- combinations
  - query + browsing
  - browsing + query
  - filtering + query
  - etc.

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

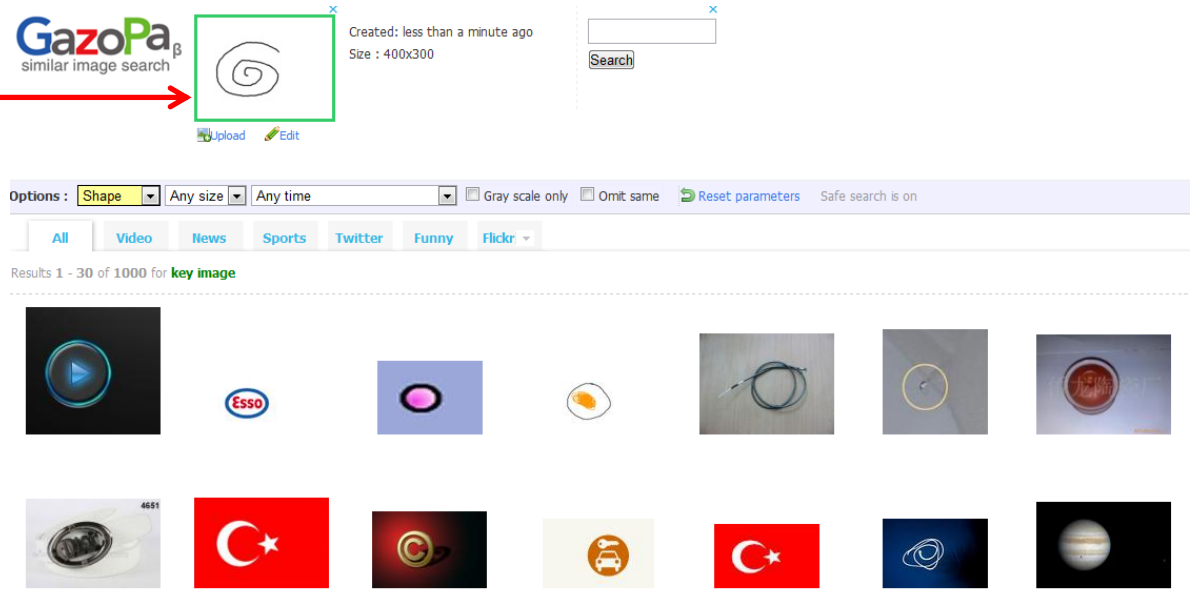
Modes of web retrieval – query

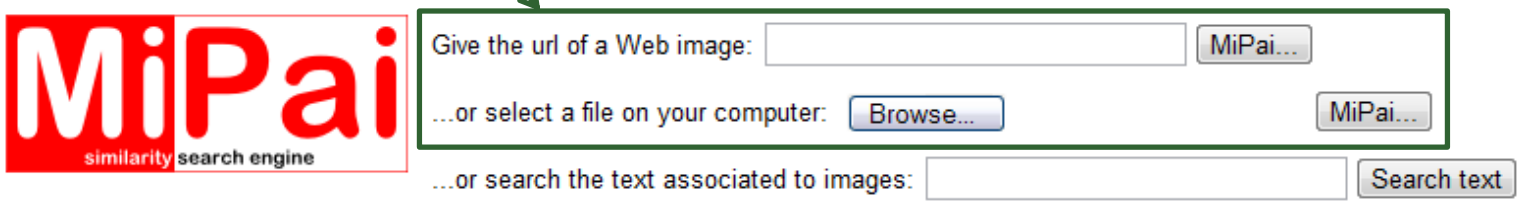
- keyword query
- multiple answers
  - web pages
  - multimedia
  - other…



Modes of web retrieval – query

- content-based query
- image search, query by
  - sketch
  - example



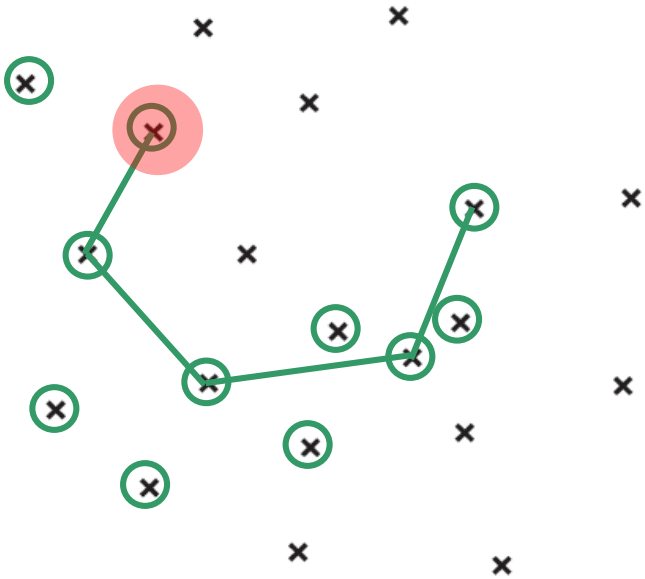


Modes of retrieval – browsing

- 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

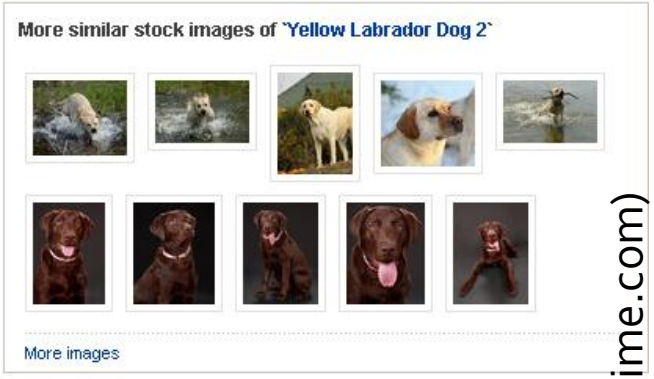
- browsing explicit graph
  - static hierarchy of categories
  - dynamic network
- browsing virtual graph
  - series of queries
  - an object in query result is used to specify the new query



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 energentic 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



(credit: dreamstime.com)

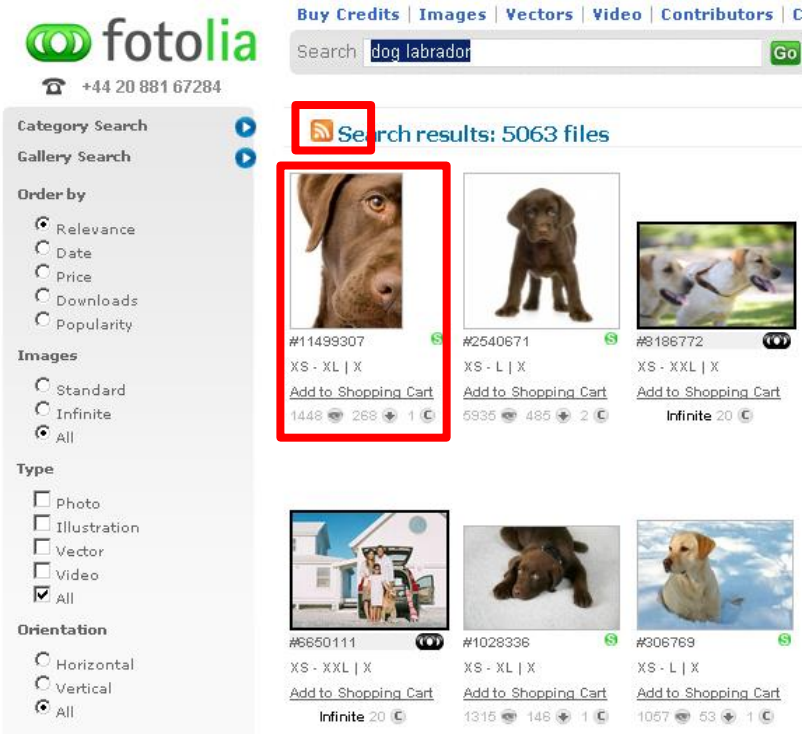
Modes of retrieval – filtering

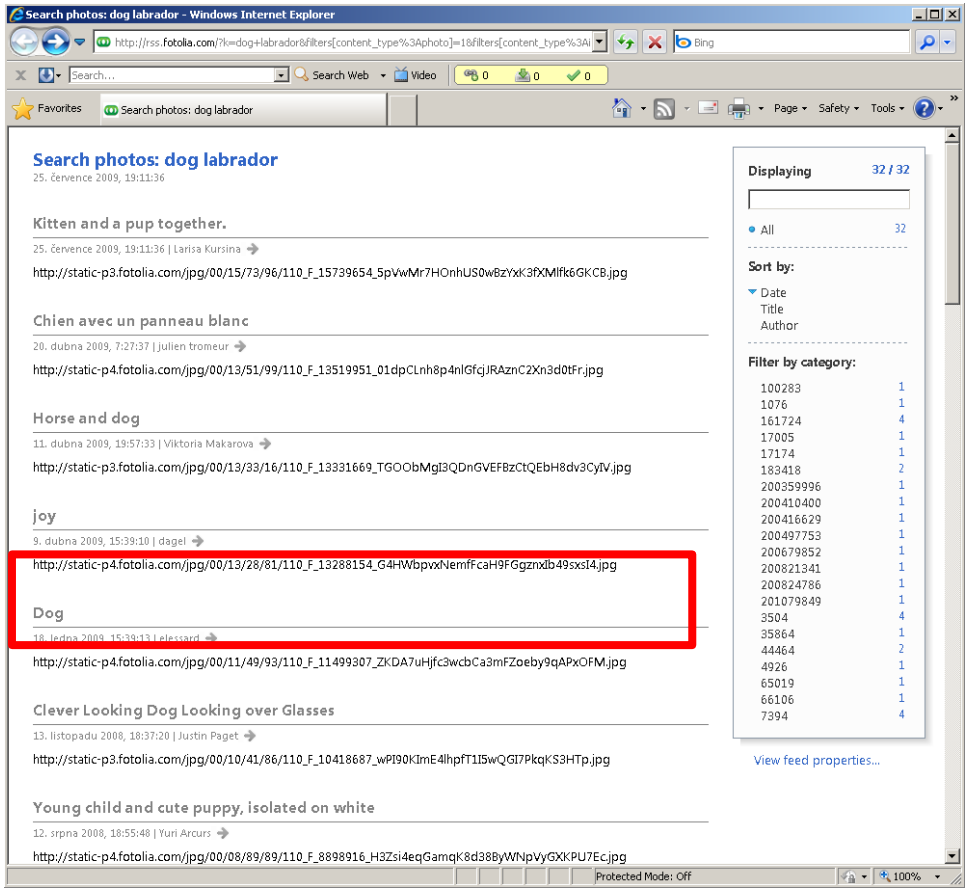
- filtering = formulation of fixed search intent
  - explicit = static request
    - query, subscription
  - implicit = recommending
    - based on user preferences, profile, search history, collaborative filtering, etc.
- filtering result
  - dynamic, changing in time
  - like database view

Example – explicit filtering

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

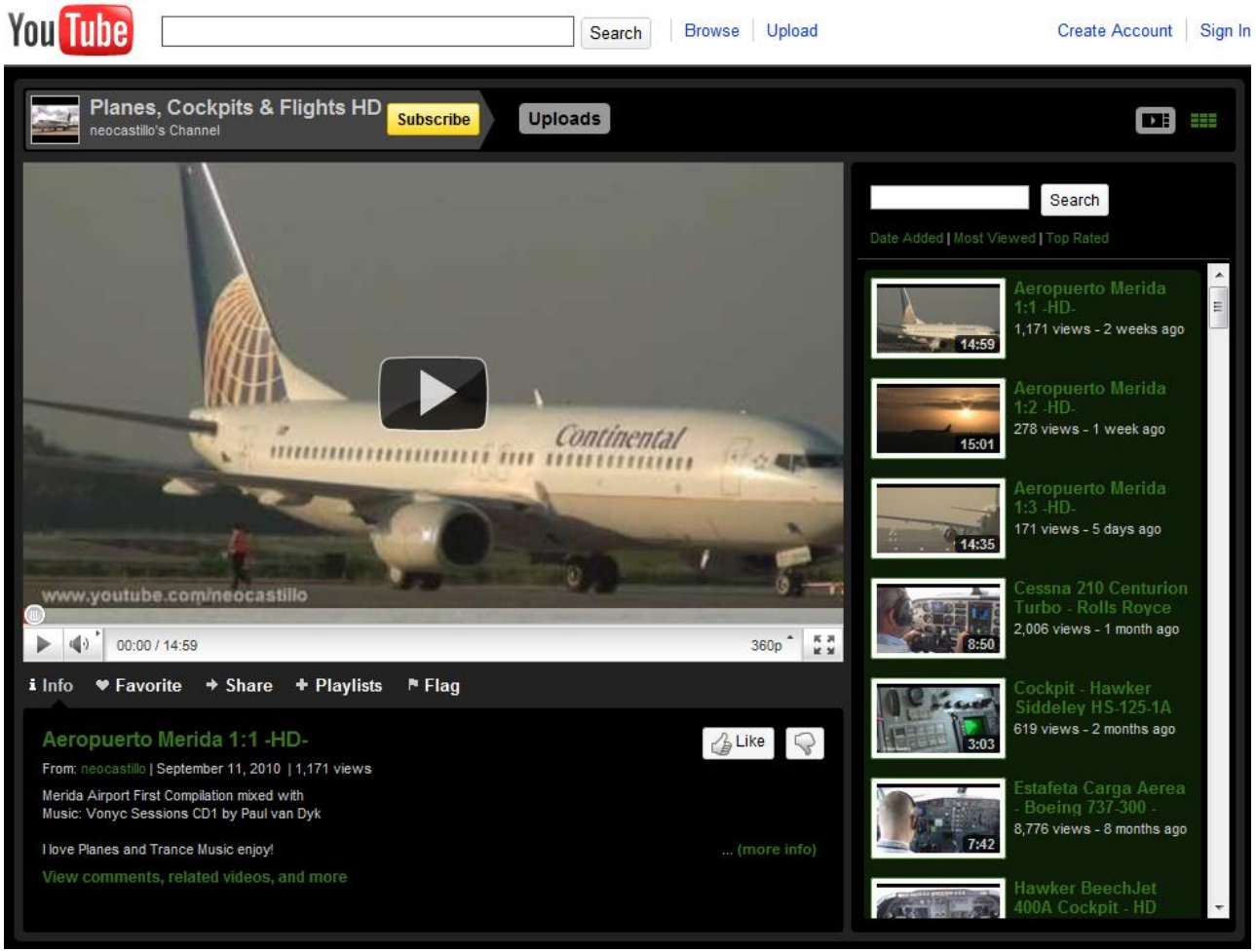
(credit: fotolia.com)





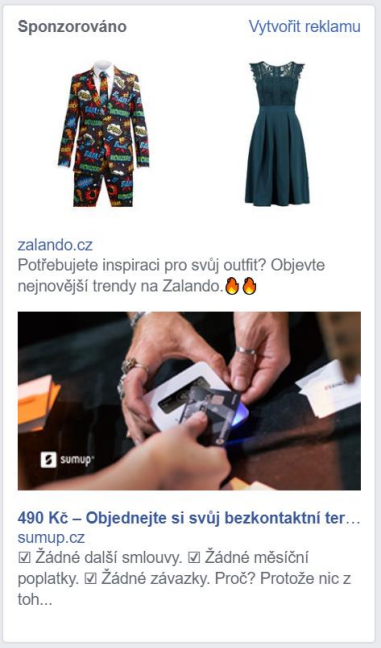
Example – explicit filtering

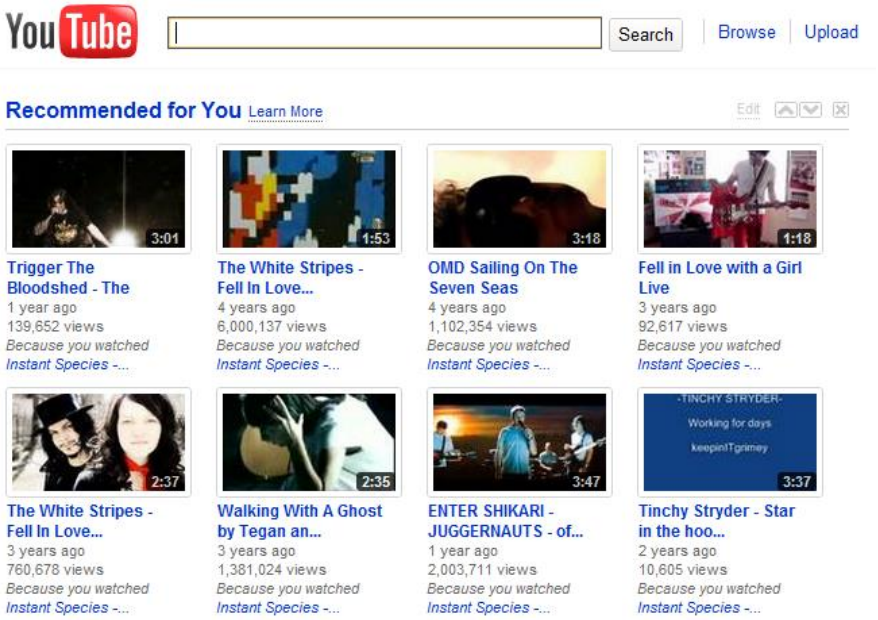
- subscription to a "channel"
- membership in a group
- etc.



Example – implicit filtering

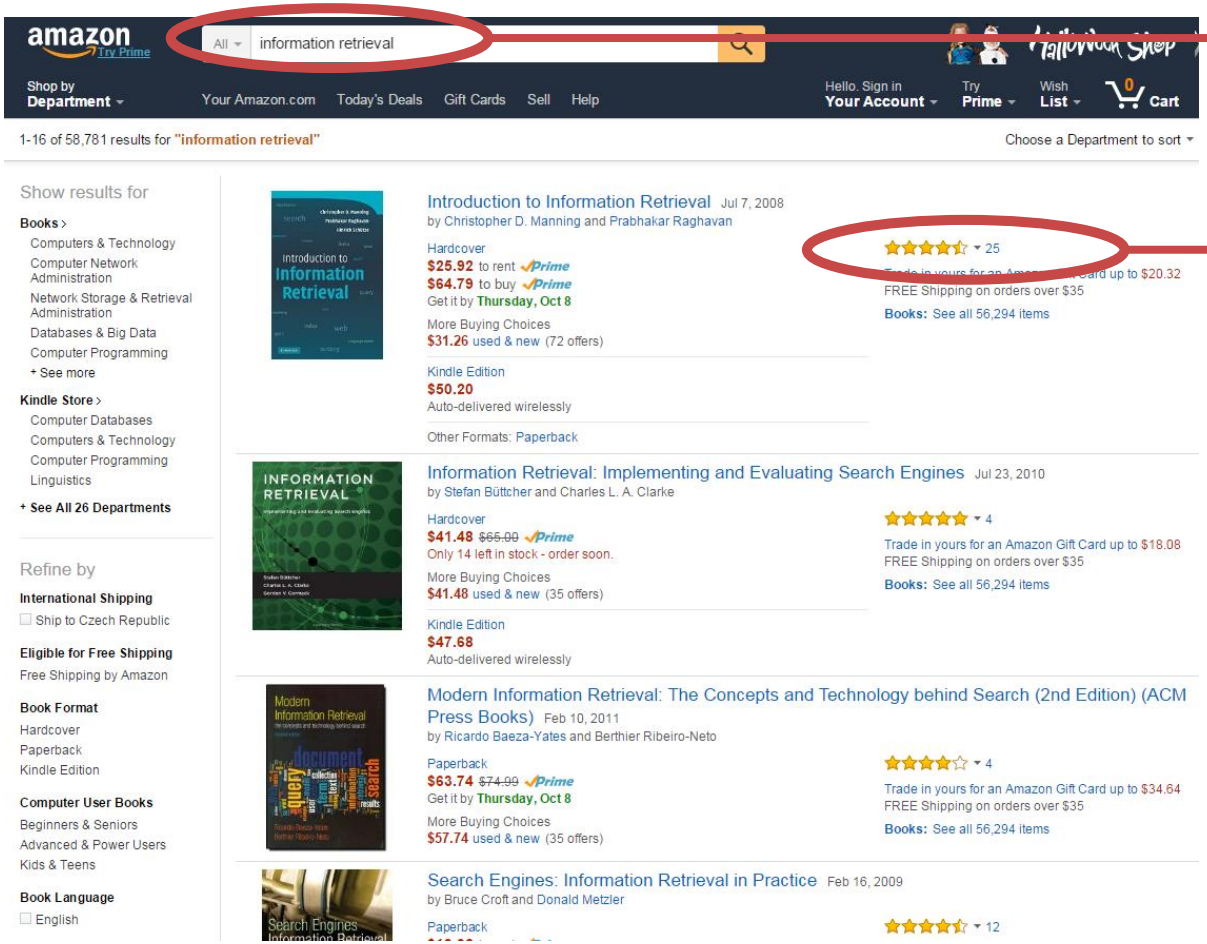
- implicit filtering (recommending) as an automatic query based on user's previous searches
  - history of viewing content serves as relevance feedback





(credit: youtube.com)

Example – combination



query

recommendation

re-ranking

Example – combination

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