Searching the Web of Objects

Ricardo Baeza-Yates

VP of Yahoo! Research for EMEA & LatAm
Barcelona, Spain

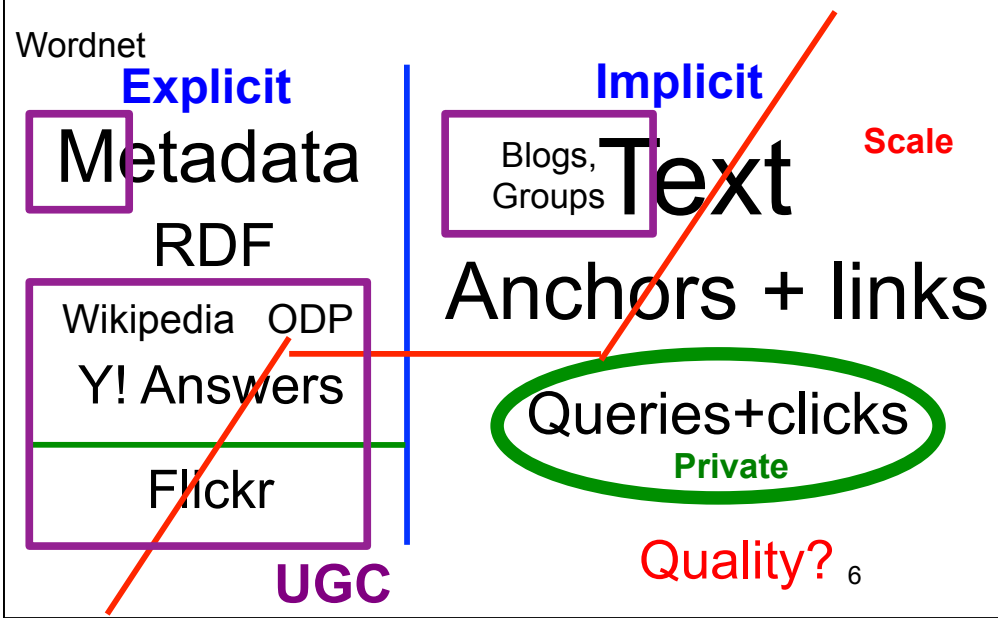
Content and Metadata trends

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	~ 2 GB
User generated content	8-10 GB
Private text content	~ 3 TB (300x more)
Upper bound on typed content	~700 TB (~200x more)

Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

[Ramakrishnan and Tomkins 2007]

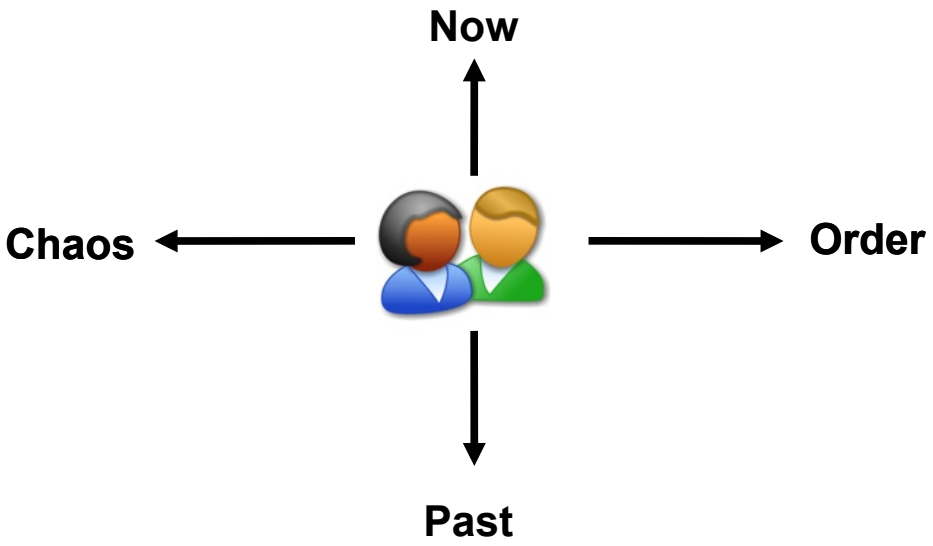
Examples



Trends

- User Generated Content
  - Massive (quality vs. quantity)
  - Social Networks
  - Real time (people + physical sensors)
- Impact
  - Fragmentation of ownership
  - Fragmentation of access (longer tail)
  - Fragmentation of right to access
- Viability
  - Business model based in advertising (?)

What we really want?



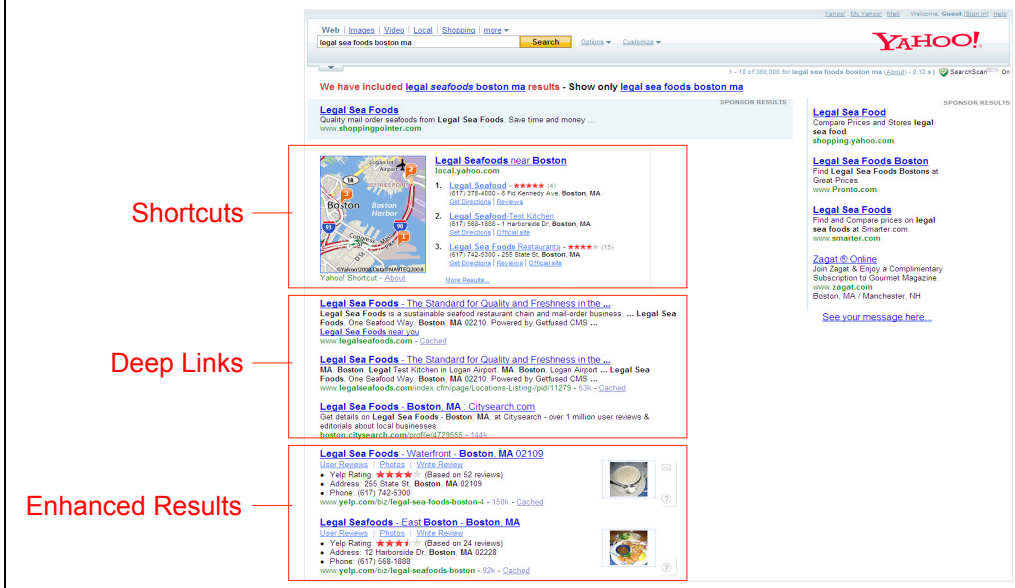
Search is Evolving

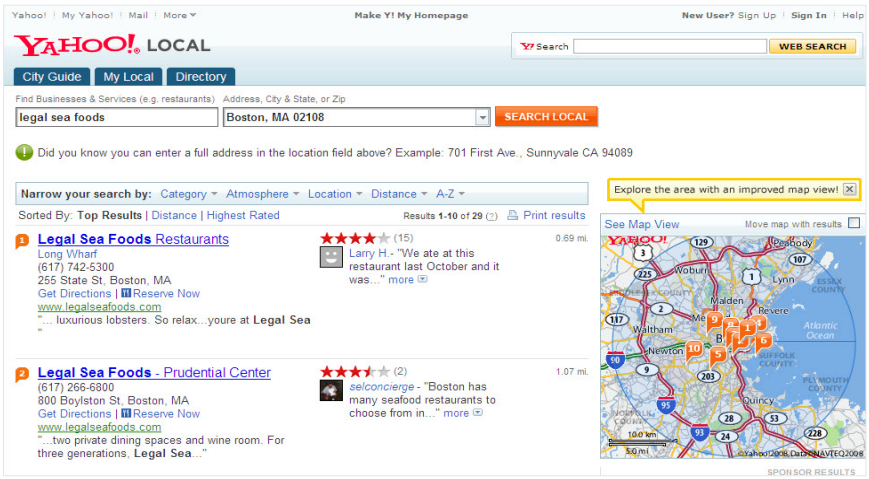
- Already, more than a list of docs
- Moving towards identifying a user's task
- Enabling means for task completion

- New experiences based on the Web 2.0

- Challenges: on-line, scalability

More complete information in one search

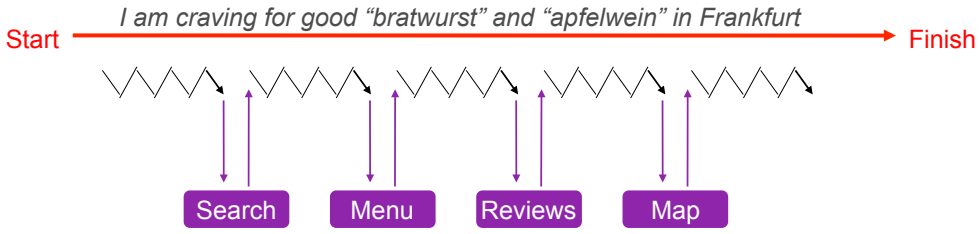




Search: Content vs. Intent

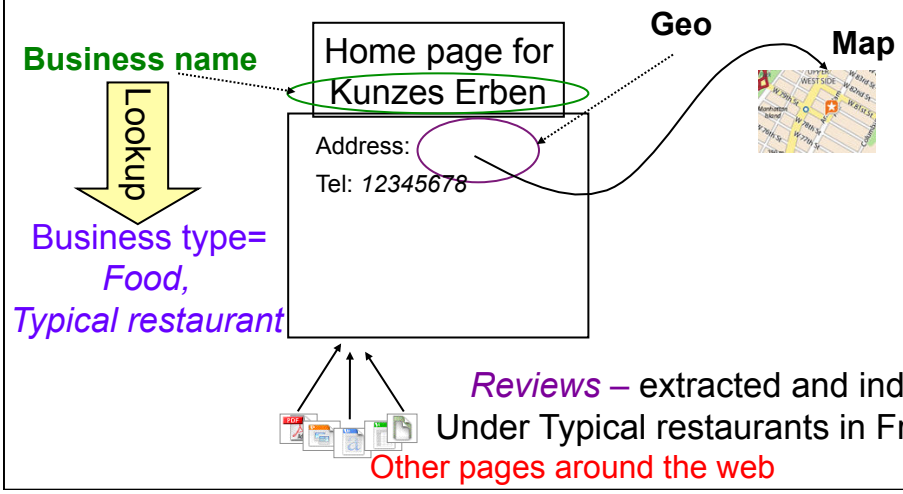
Premise:

- People don't want to search
- People want to get tasks done and get straight to their answers



How this might work – I

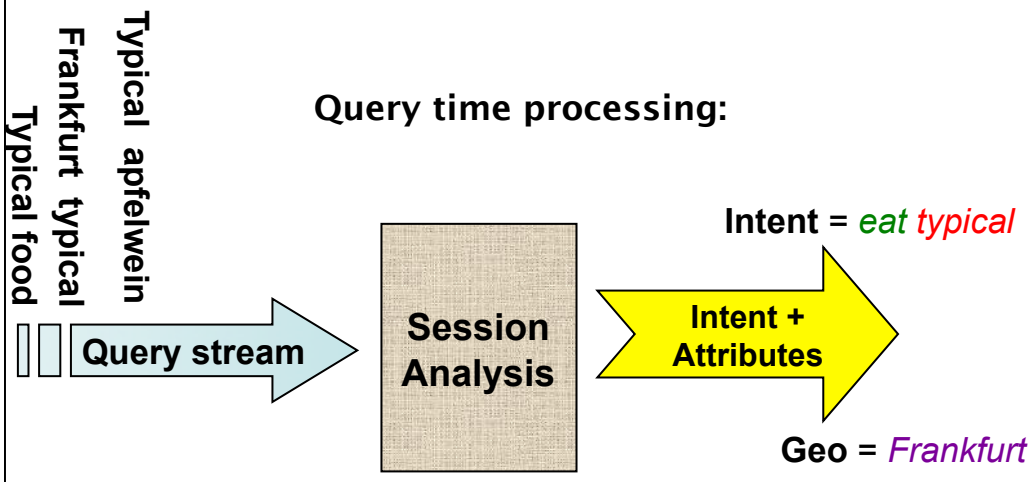
Index time processing:



Reviews – extracted and indexed
Under Typical restaurants in Frankfurt

Other pages around the web

How this might work – II



Net

- We move from a web of pages to a Web of Objects
- Objects are people, places, businesses, restaurants … (named entities)
- Objects have attributes
  - Missing, noisy, etc.
- Intents are satisfied by presenting objects and attributes
- Attributes define faceted search

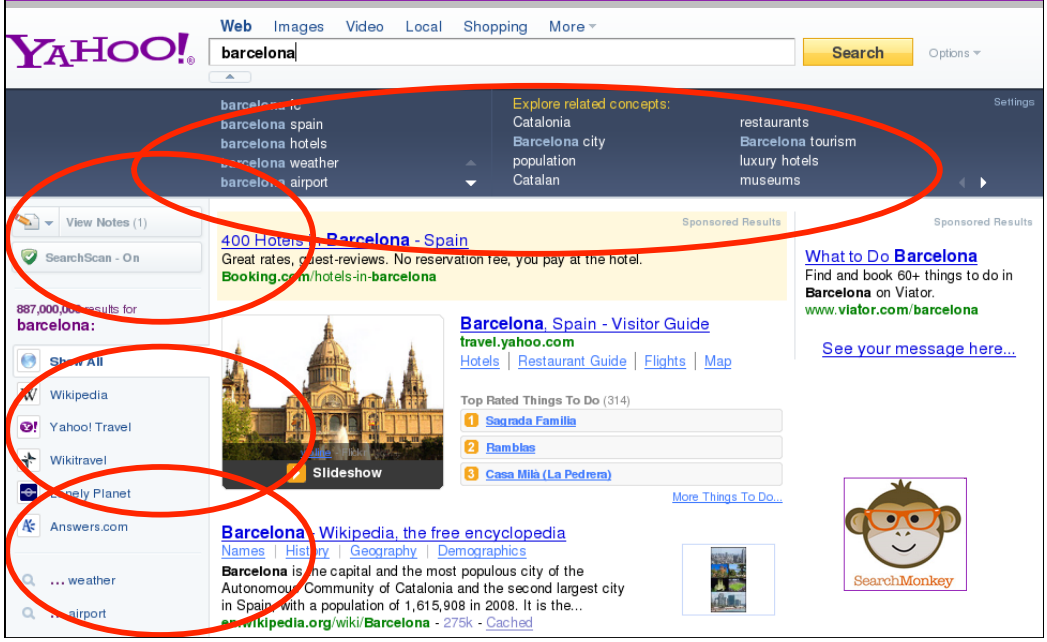
Research Challenges

- Crawling objects
- Object extraction (entities)
- Object disambiguation
- Object consolidation
- Object normalization
- Object indexing
- Object ranking
- Object visualization

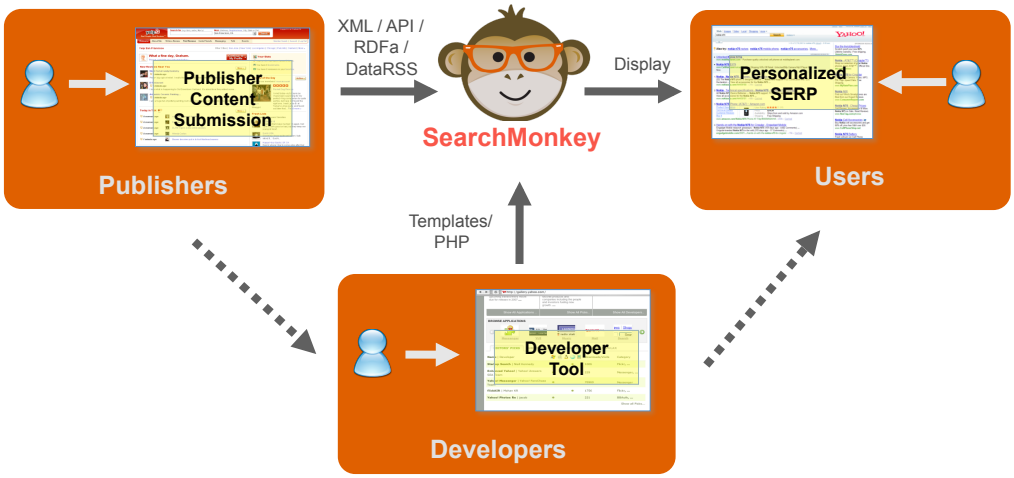
How do we get structured objects/attributes?

- Web Content
  - Metadata/Taxonomies/Folksonomies
  - Machine learning techniques
  - Classification/Extraction/Semantic Web
- Web Usage
  - Implicit relations
- Building out an open ecosystem
  - Publishers have incentives to contribute

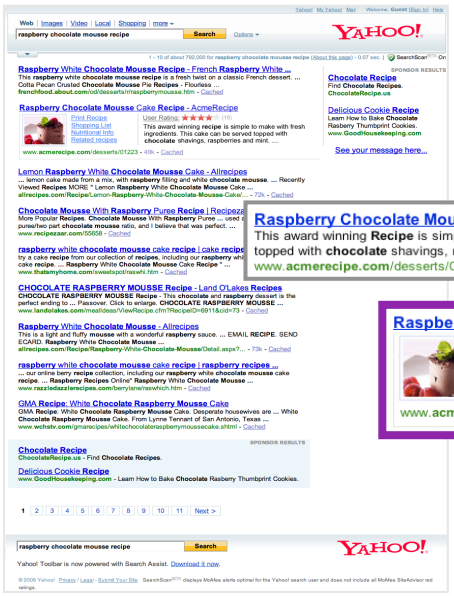
Search Engine Result Page



The SearchMonkey Ecosystem



Site Owners



- Straight to the answers
- Increase in quality of site traffic
- Fosters loyalty and engagement

BEFORE

Raspberry Chocolate Mousse Cake Recipe - AcmeRecipe
This award winning **Recipe** is simple to make with fresh ingredients. This cake can be served topped with **chocolate** shavings, raspberries and mint....
www.**acmerecipe.com**/desserts/01223 - 49k - Cached

AFTER

Raspberry Chocolate Mousse Cake Recipe - AcmeRecipe
Print Recipe
Shopping List
Nutritional Info
Related recipes
User Rating: ★★★★☆ (18)
This award winning **recipe** is simple to make with fresh ingredients. This cake can be served topped with **chocolate** shavings, raspberries and mint. ....
www.**acmerecipe.com**/desserts/01223 - 49k - Cached

Enhanced Results and Infobars:
Ratings and **reviews**, **images**, **deep links**, and other **name-value pairs**.

The Wisdom of Crowds

- James Surowiecki, a ***New Yorker*** columnist, published this book in 2004
  - **"Under the right circumstances, groups are remarkably intelligent"**
- Importance of diversity, independence and decentralization

Aggregating data

"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".

The Wisdom of Crowds

–Crucial for Search Ranking

–Text: Web Writers & Editors

- not only for the Web!

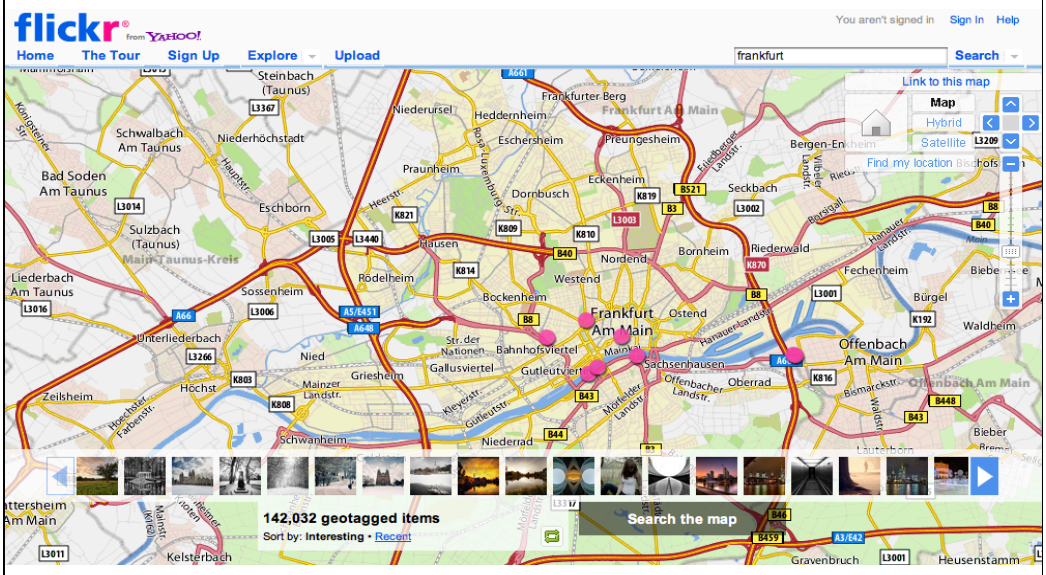
–Links: Web Publishers

–Tags: Web Taggers

–Queries: All Web Users!

- Queries and actions (or no action!)

Geo-tagged Photos in Flickr



The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

Long tail

The Long Tail

Explore Flickr through tags

architecture art australia beach birthday blue bw california canada canon china christmas city concert england europe family festival flower flowers food france friends fun germany green italy japan london music nature new newyork night nikon nyc paris park party people portrait red sanfrancisco sky snow spain street summer sunset taiwan travel trip uk usa vacation water wedding white winter

Heavy tail of user interests

Many queries, each asked very few times, make up a large fraction of all queries

Movies watched, blogs read, words used …

One explanation

Personal distribution has a heavy tail

Many queries, each asked very few times, make up a large fraction of all queries

Applies to word usage, web page access …

We are all partially eclectic

The reality

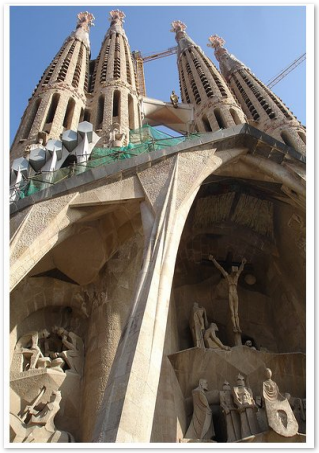
Broder, Gabrilovich, Goel, Pang; WSDM 2009

Why the heavy tail matters

Not because the worst-sellers make a lot of money

But because they matter to a lot of people

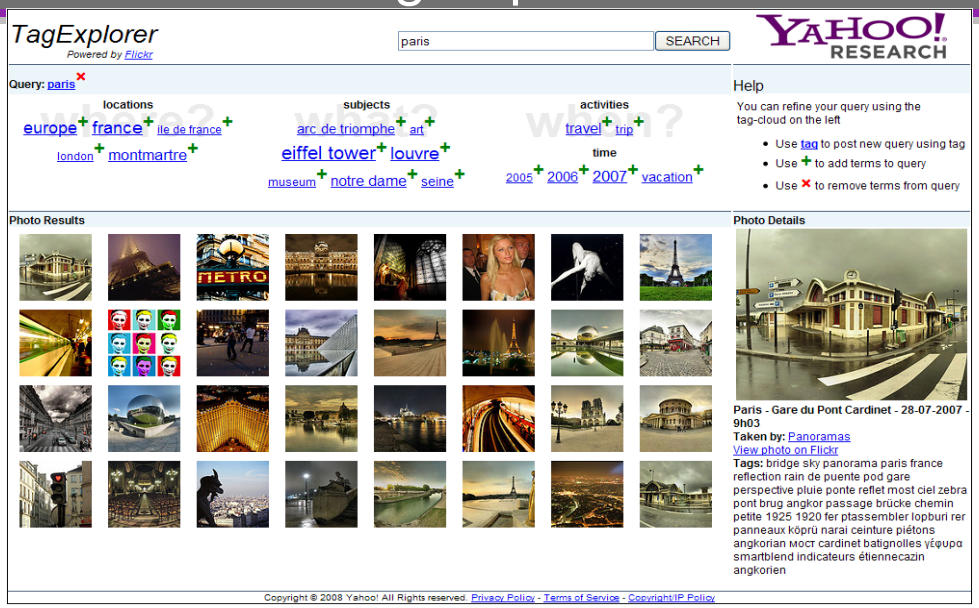
Tag Mining - Collective Knowledge



- Many users annotate photos of “La Sagrada Familia”:
  - Sagrada Familia, Barcelona
  - Sagrada Familia, Gaudi, architecture, church
  - church, Sagrada Familia
  - Sagrada Familia, Barcelona, Spain
- Derived collective knowledge:
  - Barcelona, Gaudi, church, architecture

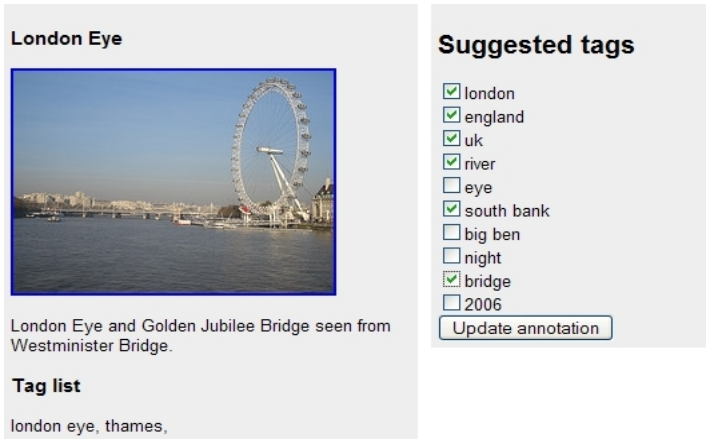
Van Zwol et al, 2008 - 2010

Tag Explorer

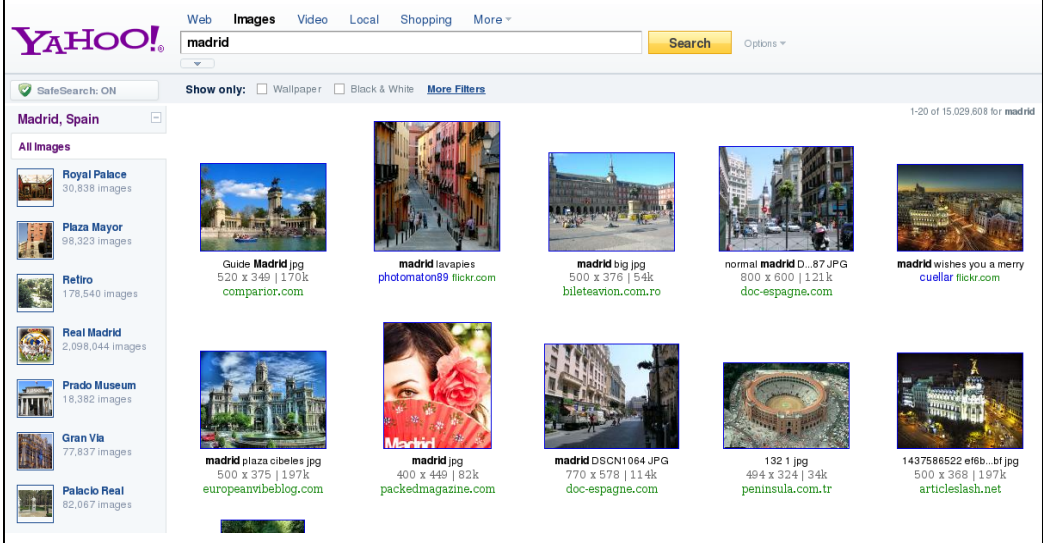


Endless image browsing

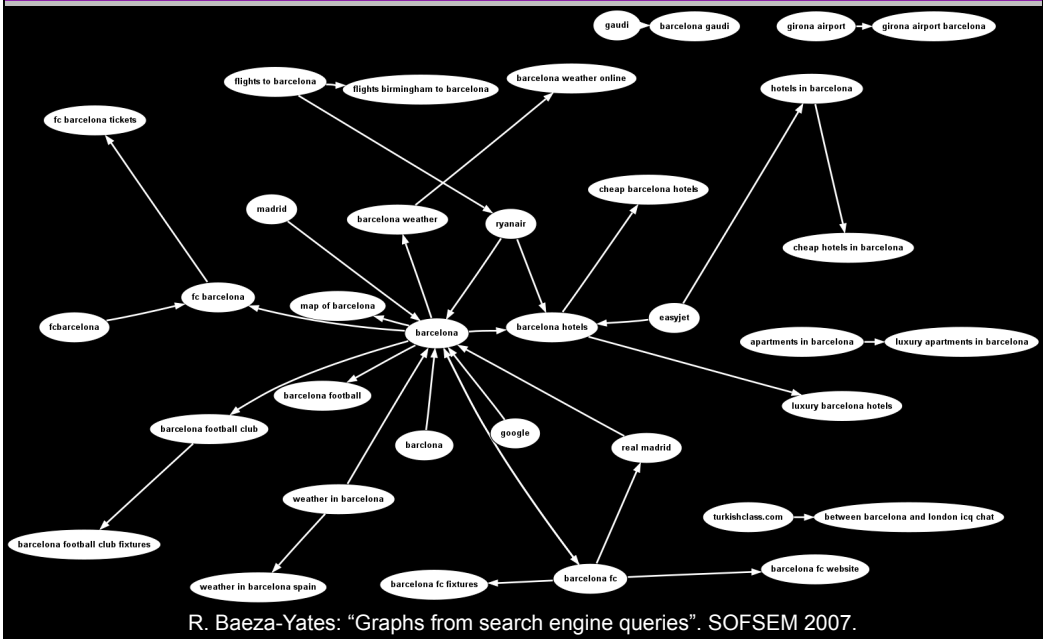
Could suggest tags: nice but ....



Faceted Image Search

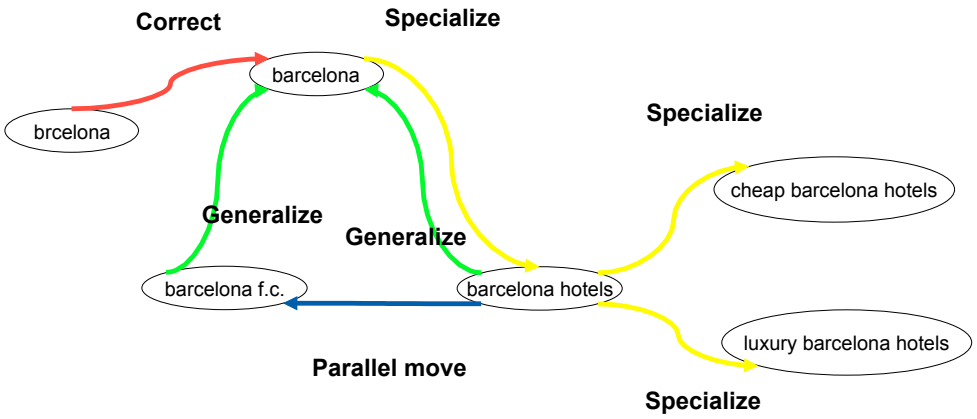


Queries as implicit tags



R. Baeza-Yates: “Graphs from search engine queries”. SOFSEM 2007.

Query-reformulation types



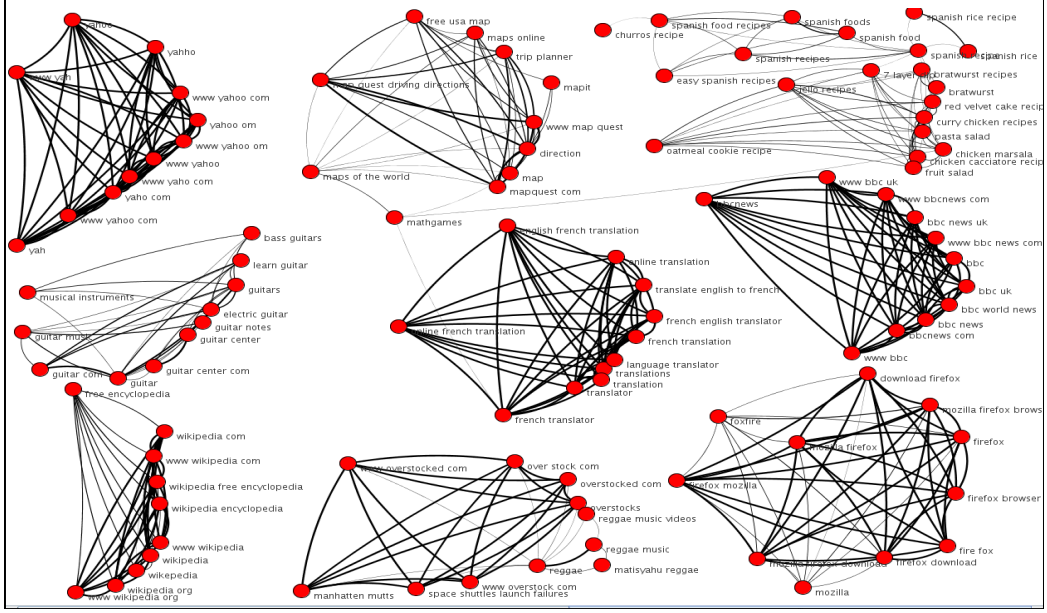
Bonchi, Donato, Castill, etc. Query flow model, 2009

Search Pad

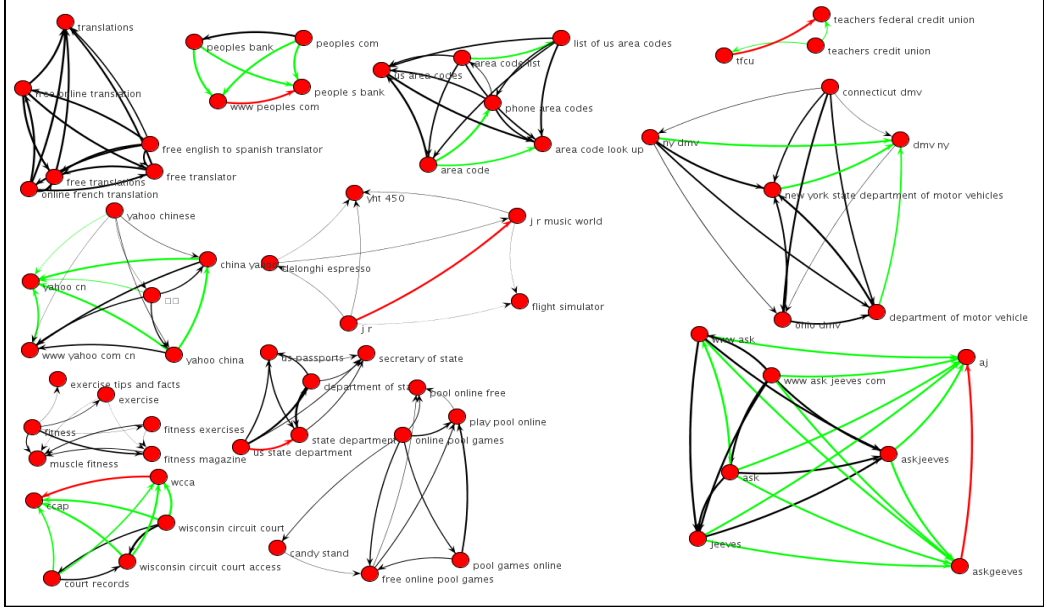
"... keeps track of search query terms ... when it ***detects a trend,*** *offers to save the result in an online document."*

CNET

Implicit Folksonomy?

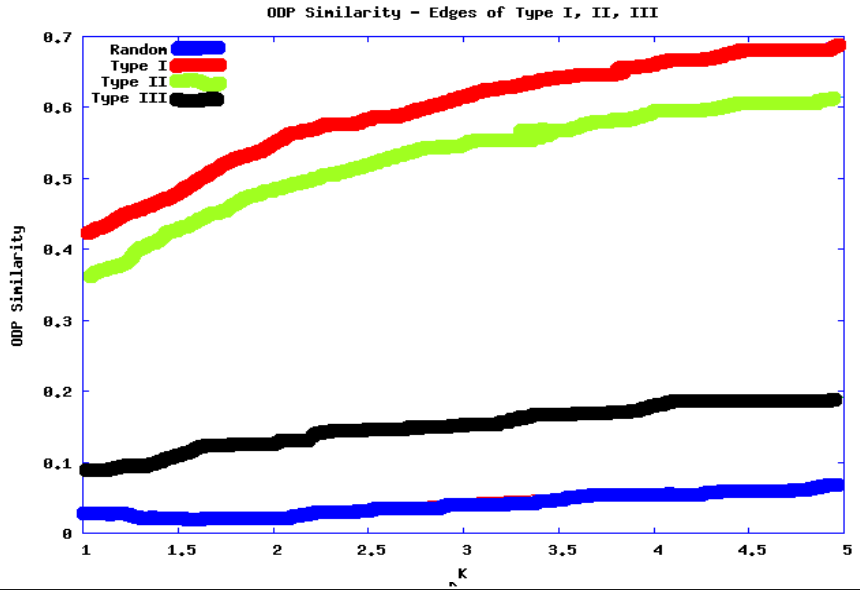


Implicit Knowledge? Web slang!

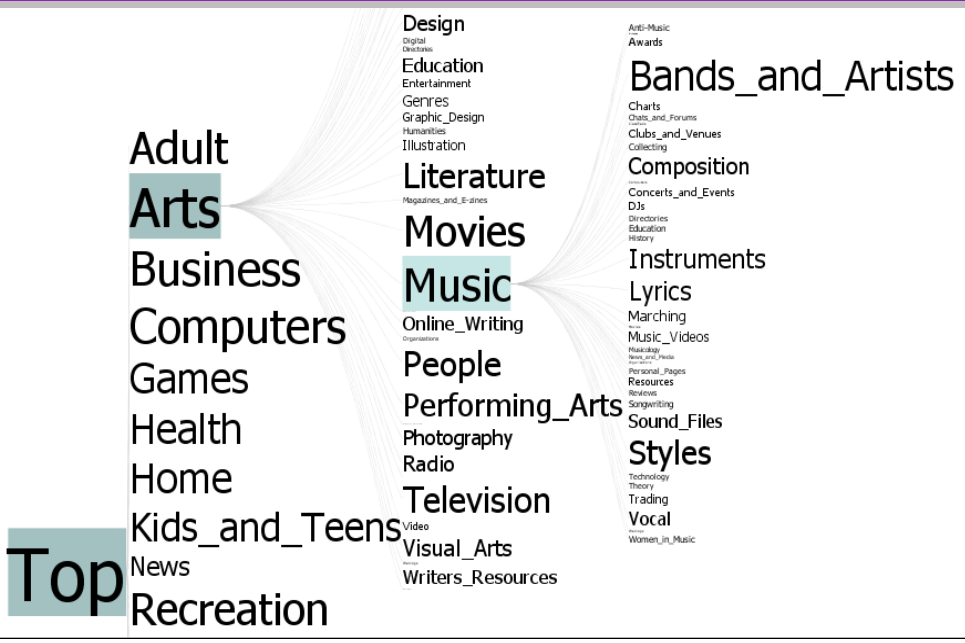


Experimental Evaluation

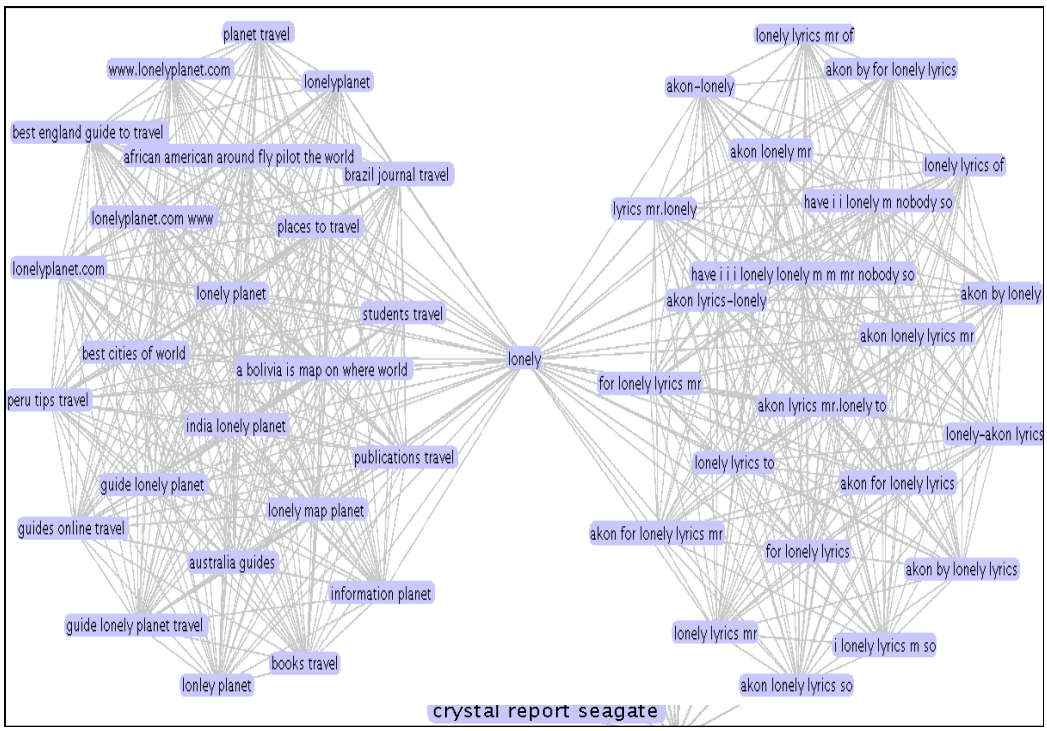
Baeza-Yates & Tiberi, KDD 2007



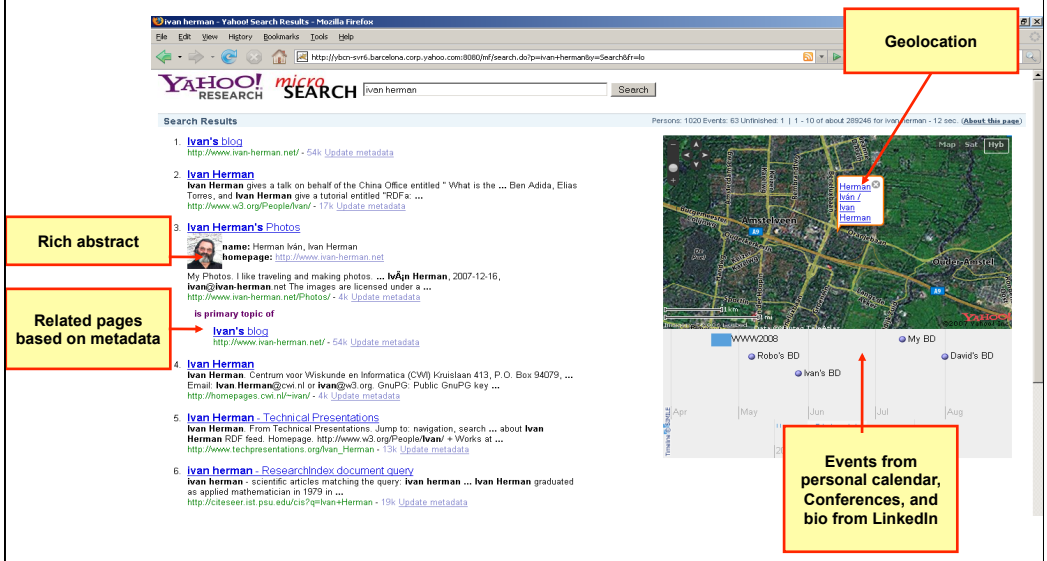
Hierarchical Clustering




Top
free
free
free
free_games_yahoo
candy_jelly_hershey
surveys_paid_jobs
fights_street_video
furniture_ikea_dining
jc_penney_penny
wine_baskets_sattui
preteen_models_lolita
sheet_music_piano
free_cocks_black
fonts_font_free
tax_peas_state
model_modeling_models
cancer_breast_women
ebay_motors_keyless
malaysia_of_kementerian
bed_billet_truck
constitution_amendment_amendments
pay_military_mypay
final_fantasy_advent
furniture_blinds_rugs
lyrics_mymp_moves
potter_harry_goblet
briggs_parts_stratton
navy_old_baby
lagu_pan_peter
gps_garmin_magellan
clothing_bealls_alloy
dade_miami_clerk
tommy_armour_jeff
r_us_toys
kementerian_malaysia_pendidikan
urban_hoax_legends
fax_internet_efax
mvp_ducks_mcmurray
vids_guerra_phone
scuba_diving_snorkel
philippines_philippine_news
ask_jeeves_www.ask
clit_bang_banglbus
philadelphia_wpvi_6
roma_dynamite_napoleon
gangs_gang_street
fry_frys_electronics
hobby_quarters_lobby
lolita_pedo_lolitas
ncaa_basketball_championships
paris_france_hotels
estate_real_banker
sunriver_palau_cbs
palm_zire_palmr
orleans_louisiana_new
friends_happy_tree
trisha_nude_actress
puppies_sale_for
plant_plants_light
auto_parts_autozone
cat_cats_punk
books_half_powell
boot_camps_camp
diet_recipes_exercise
battery_adapter_camcorder
guitar_tab_chords
amityville_horror_story
draft_nfl_mock
hair_prom_styles
encarta_terms_computer
county_harris_cook
pimp_choppers_ride
typing_free_test
south_park:_southpark
snoop_dogg_girls
books_michael_book
penguin_yeti_cubs
naruto_torrent_manga
wedding_patterns_shoes
creative_sound_blaster
court_supreme_cases
pool_tanning_swimming
test_permit_driving
vitamin_eas_liver
bluegreen_williamsburg_marriott
war_world_2
warcraft_world_wow
atlanta_baymont_braves
jazeera_al_aljazeera
mortgage_homes_home
blue_book:_value
inmate_county_orange
credit_report_free
card_credit_citi
sony_panasonic_motorola
states_directions_map
games_nick:_nickelodeon
hair_hsn_qvc
karaoke_music_songs
dictionary_english_spanish
nike_vuitton_louis
sony_camera_canon
college_university_community
code_area_codes
hotels_marcus_neiman
inn_holiday_marriott
illinois_state_of
spyware_adware_lavasoft
free_chart_software
verizon_phone_telephone
victoria_secret_victorias
records_visa_state
exchange_currency_rate
download_music_p2p
homes_sale_houses
estate_real_insurance
airlines_york:_west
pet_dog_foster
california_tax_state
movie_wars_star
chicago_sun_news
weather_newspaper_news
school_schools_high
screensavers_wallpaper_free
metro_bus_mta
lane_oreilly_o
lyrics_song_lyric
apartments_apartment_rent
music_mp3_download
music_top_billboard
iraq_navy_casualties
apprentice_news_nbc
free_sex_porn
pest_roach_tear
trails_forest_hiking
dota_warcraft_allstar
drinks_drink_bartending
aa_route_planner
□□□□_□□□_□□□
cena_john_wwe
lotto_lottery_super
massachusetts_boston_ma
outlet_springs_golf
bcb_spokane_bumiputra
poems_scholastic_children


planet travel
www.lonelyplanet.com
lonelyplanet
best england guide to travel
african american around fly pilot the world
brazil journal travel
lonelyplanet.com www
places to travel
lonelyplanet.com
lonely planet
students travel
best cities of world
a bolivia is map on where world
peru tips travel
india lonely planet
publications travel
guide lonely planet
lonely map planet
guides online travel
australia guides
information planet
guide lonely planet travel
books travel
lonley planet
lonely
lonely lyrics mr of
akon by for lonely lyrics
akon-lonely
akon lonely mr
lonely lyrics of
lyrics mr.lonely
have i i lonely m nobody so
have i i i lonely lonely m m mr nobody so
akon by lonely
akon lyrics-lonely
akon lonely lyrics mr
for lonely lyrics mr
akon lyrics mr.lonely to
lonely-akon lyrics
lonely lyrics to
akon for lonely lyrics
akon for lonely lyrics mr
for lonely lyrics
akon by lonely lyrics
lonely lyrics mr
i lonely lyrics m so
akon lonely lyrics so
crystal report seagate

Exploiting Microformats



Bridging implicit and explicit metadata

Language, Text, Search & "Semantics"

Semantics

{Person, Painter, XXth Century}

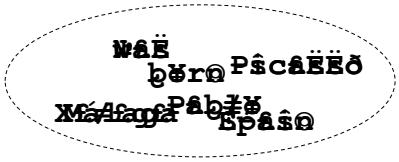
born-in

City-in-Country
{etc., etc.…}

NLP

÷£¿≠¥ ÷ŝc£ËËð №£Ë ¿¥r© ŝ© X£≠£g£, Ë÷£ŝ©.
PER LOC LOC

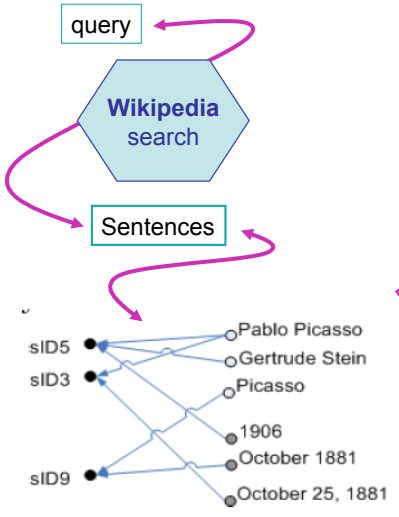
IR

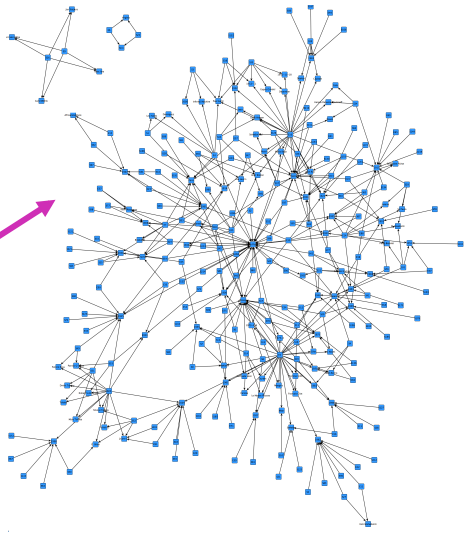


Text

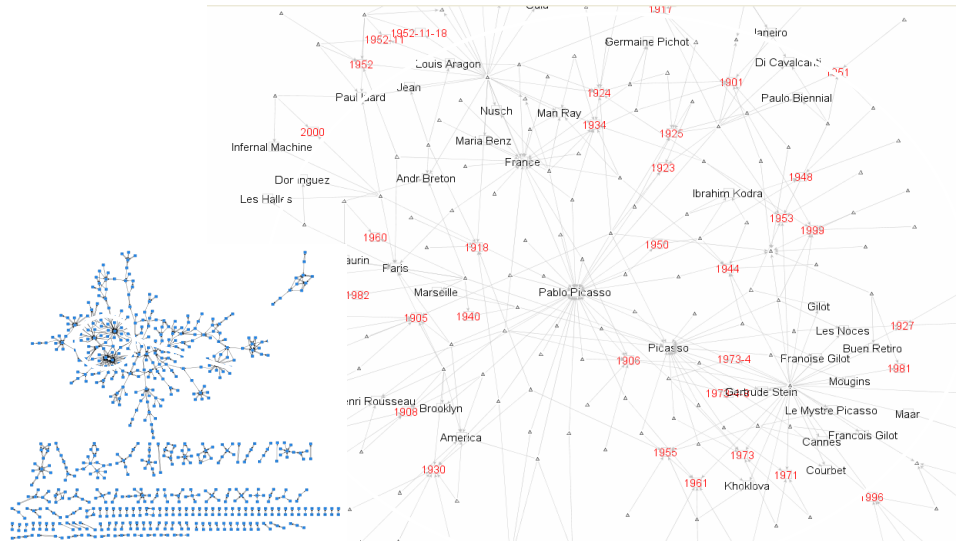
÷£¿≠¥ ÷ŝc£ËËð №£Ë ¿¥r© ŝ© X£≠£g£, Ë÷£ŝ©.
Pablo Picasso was born in Málaga, Spain.

Entity Graph (linear time entity extraction)

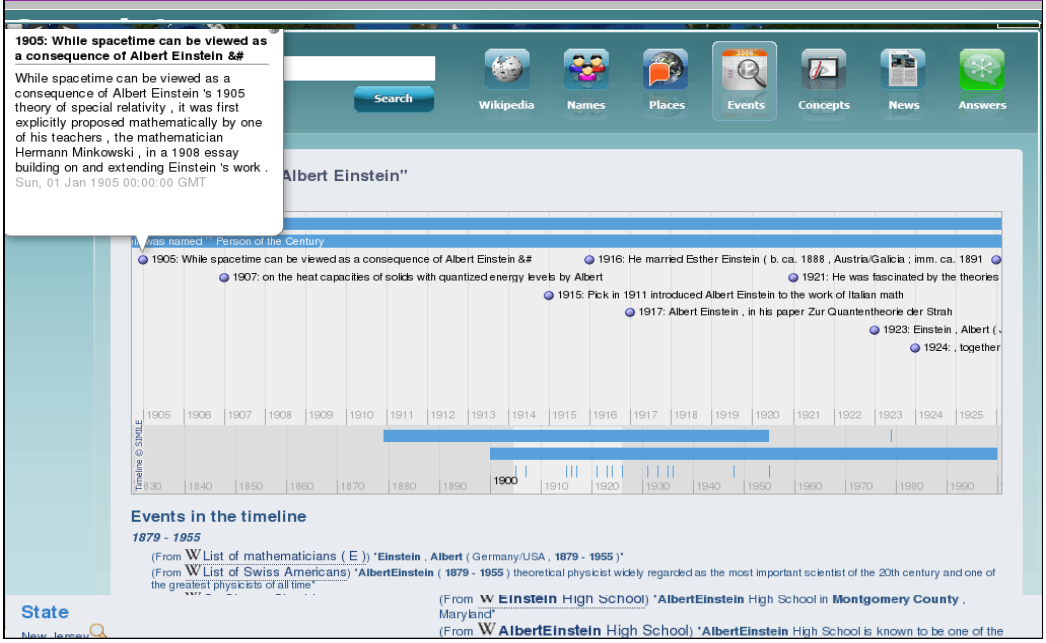




Example: Picasso



Correlator: Relations in Wikipedia



New Pages

- **For topics without a Wikipedia page, Correlator creates a "synthetic page" with an overview of the topic**
- **Query:**
  - **art deco chicago**
- **Synthetic page:**
  - **Defines Art Deco**
  - **Defines Chicago**
  - **Shows relations between Art Deco and Chicago**

Synthetic page - example

Category: Dinosaurs of South America

Buitreraptor: It was found in **Argentina** and was described in 2005 .The fossilised bones were found in 2005 in sandstone in Patagonia , **Argentina** - by an excavation lead by Peter Makovicky , curator of **dinosaurs** at the Field Museum in Chicago) .Buitreraptor was discovered in the same fossil site that had earlier yielded Giganotosaurus , one of the largest known carnivorous **dinosaurs** .

Herrerasaurus: Herrerasaurus (meaning `` Herrera 's lizard , `` after the name of the rancher who discovered the first fossil of the animal) was one of the earliest **dinosaurs** . ... This view is further supported by ichnological records showing large tridactyl footprints that can be attributed only to a theropod **dinosaur** , dating from the Ladinian (Middle Triassic) of the Los Rastros Formation in **Argentina** and predating Herrerasaurus by 3 to 5 million years .The importance of Herrerasaurus and Eoraptor lies in the fact that their remains allow for directly testing the idea of **dinosaurs** being a monophyletic group , i.e. all **dinosaurs** have a common ancestor .

Unaysaurus: It was recovered from the red beds of the Santa Maria Formation (also known as the Caturrita Formation) , which is the geologic formation where similarly old **dinosaurs** like Saturnalia have been found .The oldest **dinosaurs** in the world are from here and nearby in **Argentina** (like the Eoraptor) , which suggests that the first **dinosaurs** may have originated in the area .

Carnotaurus: Carnotaurus (pronounced /ˌkɑrnoʊˈtɔrəs/ KAHR-noh-TAWR-us ; meaning `` meat-bull '' , referring to its distinct bull-like horns (Latin carne = flesh + Greek tauros = bull) was a large predatory **dinosaur** , with horns vaguely resembling a bull's .Carnotaurus lived in Patagonia , **Argentina** during the Maastrichtian stage of the Late Cretaceous , and was discovered by José F. Bonaparte , who has uncovered many other bizarre South American **dinosaurs** . ... Together , these **dinosaurs** form the subfamily Carnotaurinae in the family Abelisauridae .

Eoraptor: Eoraptor was one of the world 's earliest **dinosaurs** . ... Early **dinosaur**The bones of this primitive **dinosaur** were first discovered in 1991 , by University of Chicago paleontologist Paul Sereno , in the Ischigualasto Basin of **Argentina** .

Argyrosaurus: Argyrosaurus (pronounced /ˌɑrdʒɪroʊˈsɔrəs/ AHR-ji-ro-SAWR-us) meaning 'Silver lizard' , because it was discovered in **Argentina** , which is sometimes known as 'Silver land ' (Greek argyros meaning 'silver ' and sauros meaning 'lizard') was a genus of herbivorous titanosaurid **dinosaur** that lived about 70 million years ago , during the Late Cretaceous Period of what is now South America (**Argentina** and Uruguay) .It was one of the largest **dinosaurs** , having a height of 8 metres , a length of up to 20-30 metres and a weight of up to 80 tonnes.It was a herbivore .

Argentinosaurus: Argentinosaurus (meaning `` **Argentina** lizard '') was a herbivorous sauropod **dinosaur** genus that was among the largest land animals that ever lived . ... Argentinosaurus is featured prominently in the permanent exhibition Giants of the Mesozoic at Fernbank Museum of Natural History in Atlanta , Georgia , USA. This display depicts a hypothetical encounter between Argentinosaurus and the carnivorous theropod **dinosaur** Giganotosaurus . ... At 123 feet long , this skeletal reconstruction represents the largest **dinosaur** mount ever to be assembled .

Neuquensaurus: Neuquensaurus (meaning `` Neuquén lizard '') was a titanosaur sauropod **dinosaur** that appeared in the Late Cretaceous , 71 million years ago in **Argentina** and Uruguay in South America .This **dinosaur** was 10-15 meters (34-51 feet) long , and is believed to have possessed armor-like osteoderms .

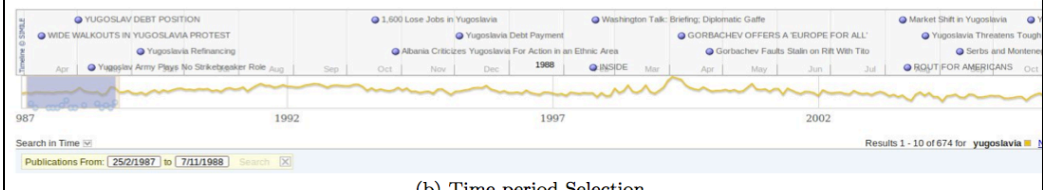
Genyodectes: Genyodectes (Woodward , 1901) is a genus of ceratosaurian theropod **dinosaur** from the Lower Cretaceous of South America .The holotype material (MLP 26-39 , Museo de La Plata , La Plata , **Argentina**) was collected from the Cañadón Grande , Departamento Paso de Indios in the Chubut Province of **Argentina** and consists of an incomplete snout , including the premaxillae , portions of both maxillas , the right and left dentary , many teeth , a fragment of the left splenial , and parts of the supradentaries .

Time Explorer

- **Finding Relations among Entities in News**
  - Past, present or future!
    - Baeza-Yates, Searching the Future, 2005.
  - The clue is the interface
  - Part of the Living Knowledge EU project
- **Winner of the HCIR 2010 Challenge**
  - New York Times collection (1987-2007)
  - Found many interesting examples
  - Generates new NLP research problems

Time Explorer

(a) Basic Timeline



(b) Time period Selection

(c) Timelijne with entity trends

Time Explorer

Figure 4: Searching the Future

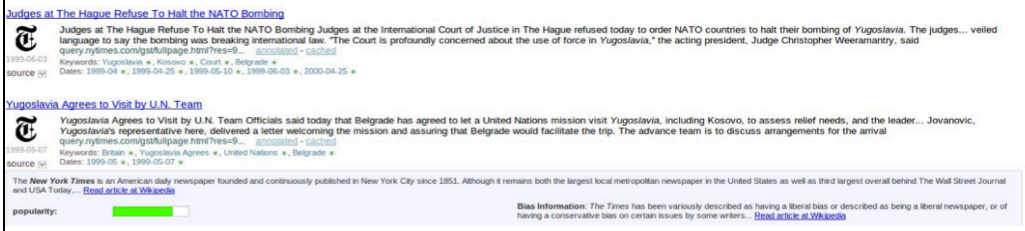


Figure 5: Results

Time Explorer

Implicit Search

Search as a back end process

- Trigger the right search depending on the context
  - Writing email
  - Browsing news
- Research challenges
  - More on query intent prediction
  - Whole page layout optimization
  - Exploit the social layer
  - Interactive manipulation of the answer
  - Measure user engagement in any context
  - Compare user satisfaction across Web sites

So what's next?

We are far from being done with innovation in search engines

- Possible future
  - The new frontiers: front-end and user experience
    - The most probable reason for users to switch between quasi-equivalent engines is a better user experience
  - We still don't understand well information needs (will we ever? brain electrodes don't work☺)
  - New search: contextual content delivery
- Large scale usage data is key to getting there BUT

Three major conflicting factors

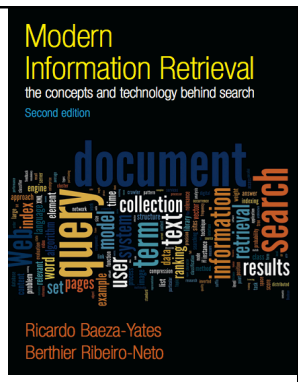
Usage data at a very large scale	
More data over longer periods of time brings more insights	
Contextualization	**Privacy**
More data via larger communities, makes data less personalized *wisdom of crowds does not work well on small corpora*	Over contextualization endangers privacy Long-term logs endanger privacy

Contextualize the task: query intent detection

Conclusions

- Web search is no longer about document retrieval
  - Means for web-mediated goals
  - –New breed of search experiences
  - –Demands search ecosystem combining content with intent
  - –Exploiting the Wisdom of Crowds behind the Web 2.0
  - –User aggregation versus personalization
    - Optimize common tasks
    - Move away from privacy issues

Second edition coming next month



Questions?

rbaeza@acm.org

http://search.yahoo.com

http://labs.yahoo.com

http://sandbox.yahoo.com