Mining the Web 2.0
to improve Search

Ricardo Baeza-Yates
VP, Yahoo! Research

Agenda

- The Power of Data
- Examples
  - Improving Image Search (Flickr)
  - Searching the Wikipedia
  - Understanding Queries (SearchPad)
- Impacts not only relevance but also the UI
- Concluding Remarks

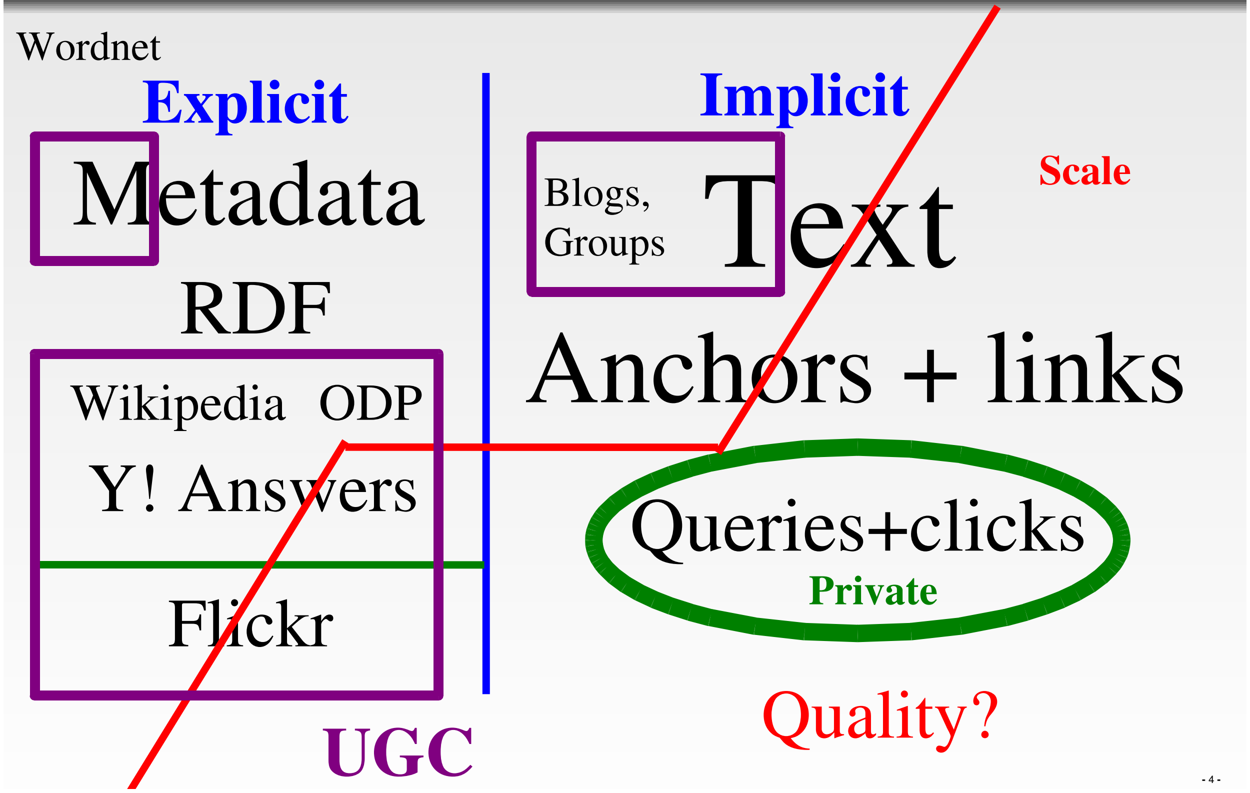
Content and Metadata trends

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

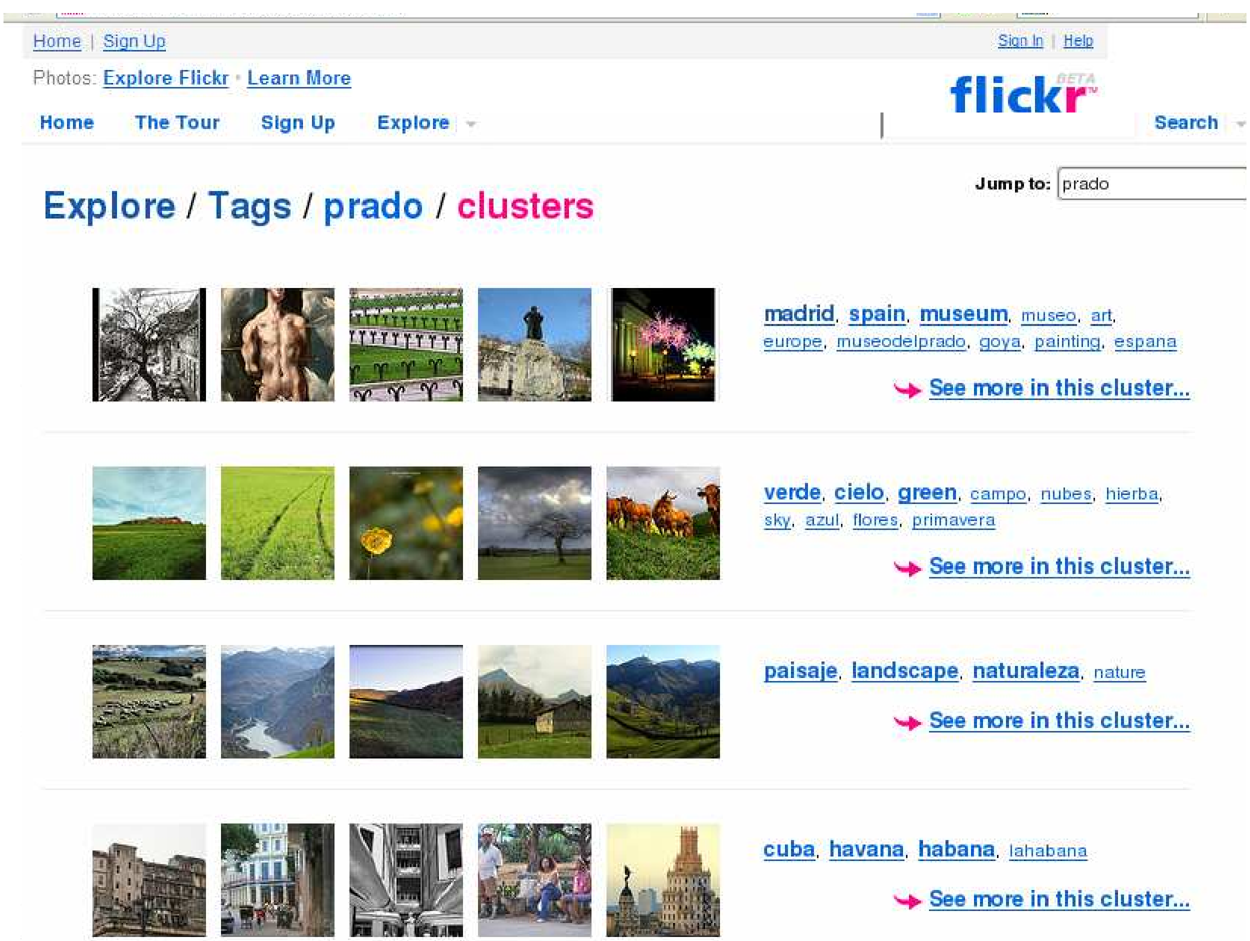


The Wisdom of Crowds

- 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".



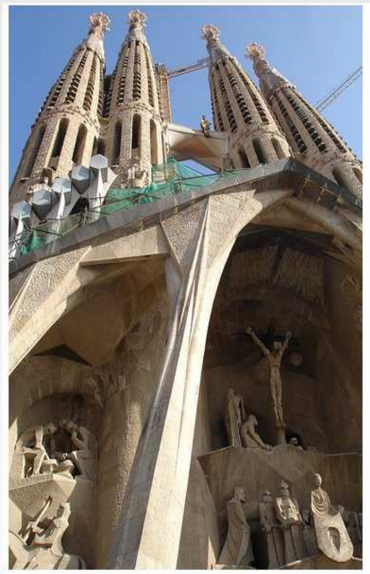
The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

The Wisdom of Crowds

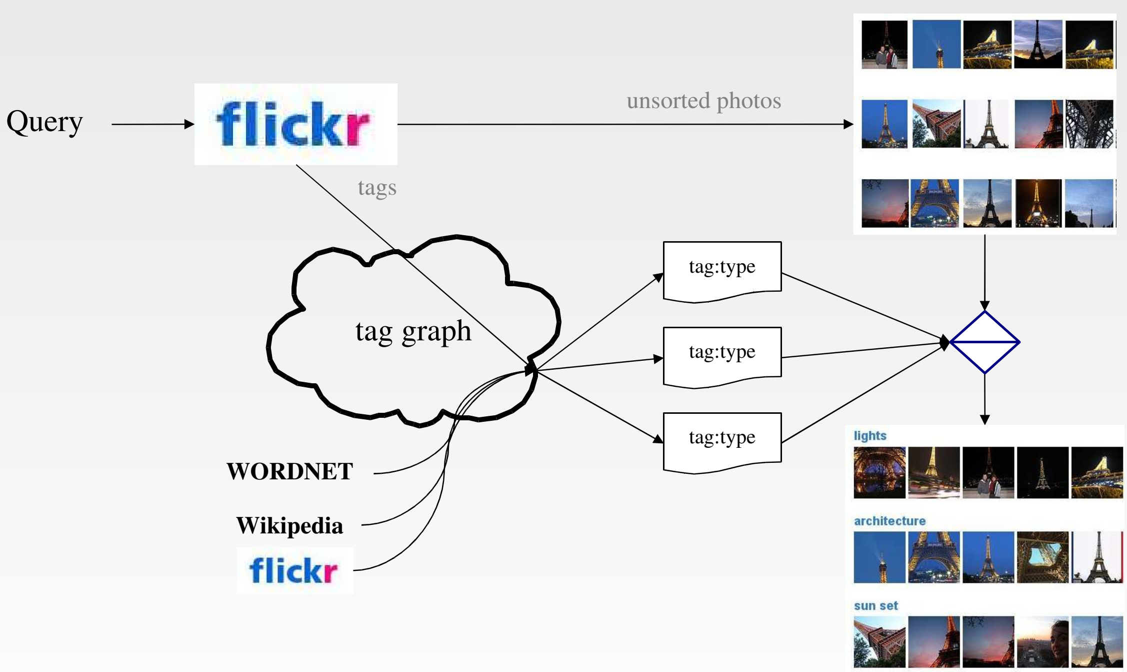
- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)

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

Improving Image Search



TagExplorer

- http://sandbox.yahoo.com/TagExplorer
- A prototype for browsing Flickr photos
- Provides query refinement for …
  - … drilling in to more specific topics
  - … zooming out to more general topics
  - … side-track to a related topic
- Organizes refinement terms …
  - … in a tag-cloud
  - … groups together semantically similar terms

Dynamic Tag Clouds

- For the user query a list of related terms is presented and can be used to refine the query (visualized as a tag-cloud)
- The related terms are derived using tag co-occurrence among 250 million Flickr photos
- The related terms are calculated using a probabilistic framework using different conditional probabilities to get a mixture of general and specific terms

Sigurbjornsson and van Zwol, WWW 2008

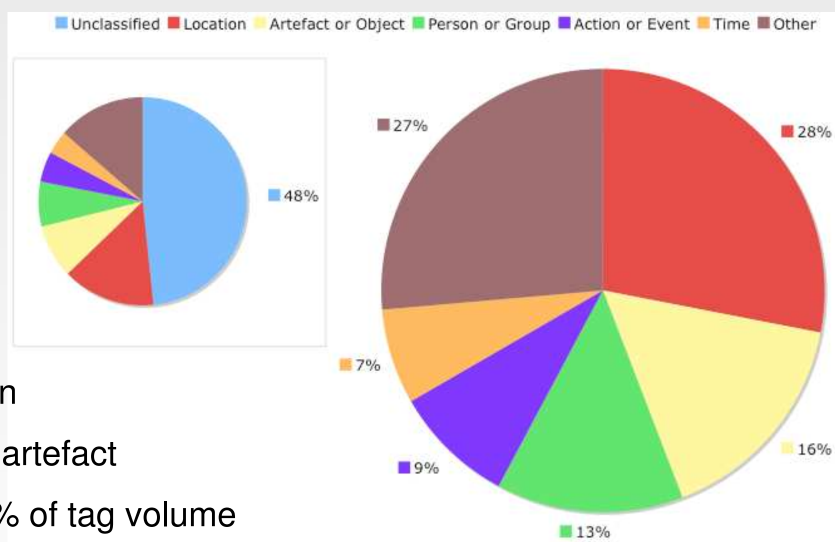
Semantic Breakup of Tag Clouds

- Tag-cloud is organized by grouping together tags that have similar meaning
- The grouping is a two levels
  - Where? What? When?
  - Locations, subjects, names, activities, time
- The classification of tags is derived using a machine learned classification of Wikipedia pages

Overell, Sigurbjornsson and van Zwol, WSDM 2009

Tag Mining - Classification

- Assign tag semantics using WordNet broad categories

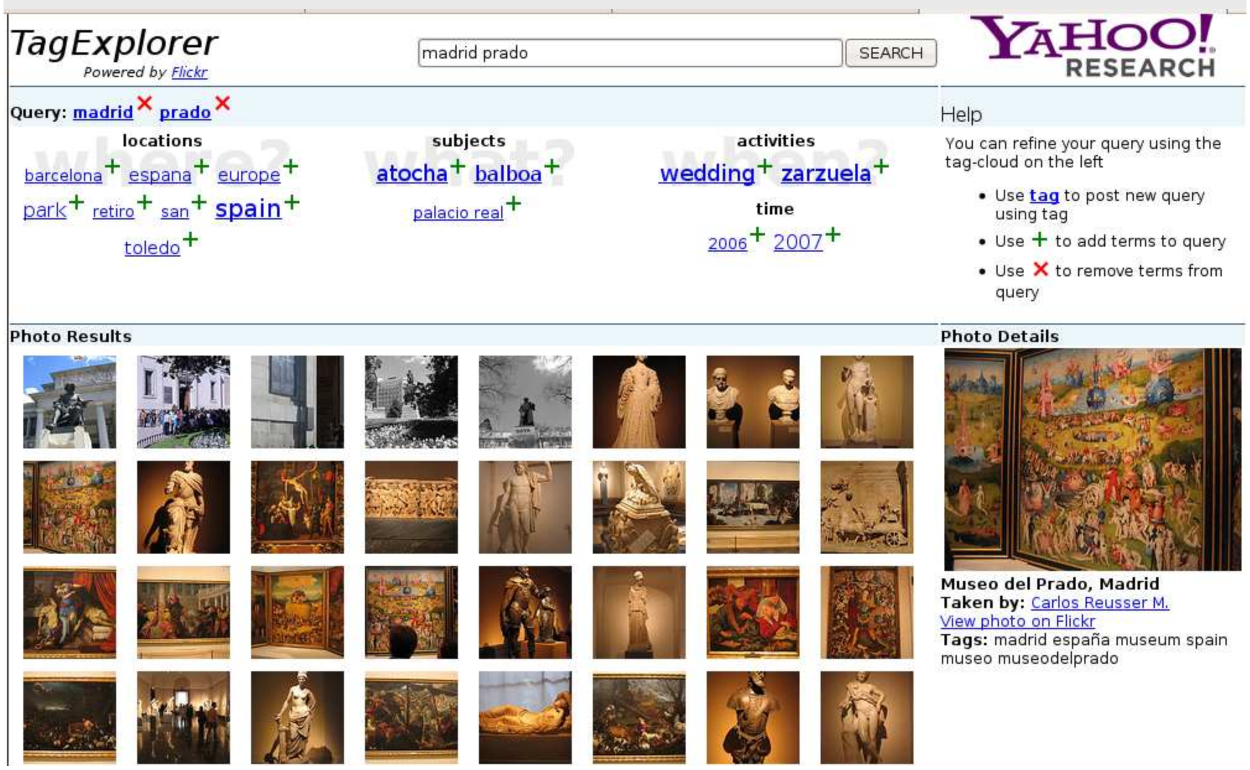


  - Paris :: location
  - Eiffel Tower :: artefact
  - Coverage: 52% of tag volume

Tag Mining – Classification

- Extend this mapping using patterns found in Wikipedia
  - Upper bound for coverage: 78.6% of the tag volume
  - Based on SVM approach
    - Features: Wikipedia templates and categories
    - Training data: Wikipedia entries found in WordNet
  - Extended coverage: 68% of the tag volume
  - Mapping from Wikipedia pages to tags
    - Reduces ambiguity in the classification

TagExplorer - Example



Could suggest tags: nice but ....

Dimensions of Diversity

- **Topical diversity**

  Query: “Jaguar”

- **Visual diversity**

  Query: “Jaguar X-type”

- Other dimensions: spatial, temporal, social

Topical Diversity

- **Diversification as part of the retrieval model**
  - Query Likelihood (full index, tags only)
  - Relevance model (full index, tags only, dual index)
- Topics
  - 95 topics extracted from Flickr search logs
  - 25 ambiguous topics
- Collection:
  - 6M public photos from Flickr (Title, description and tags)

van Zwol, Murdock, Garcia-Pueyo, Ramírez. ACM MIR 2008.

Retrieval Performance

- Unambiguous topics

Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.747	0.733	0.733	0.719	0.709	0.701	0.667
Query Likelihood (Tags Only)	**0.779**	0.749	0.720	0.712	0.703	0.700	0.673
Relevance Model	0.758	0.743	0.720	0.708	0.706	0.699	0.677
Relevance Model (Tags Only)	**0.779**	0.726	0.717	0.719	0.714	0.710	**0.683**
Relevance Model (Dual Index)	0.768	**0.754**	**0.739**	**0.726**	**0.719**	**0.716**	0.680

- Ambiguous topics

Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.680	0.760	0.720	0.725	0.734	0.744	0.734
Query Likelihood (Tags Only)	0.800	0.736	0.732	0.720	0.736	0.736	0.734
Relevance Model	0.720	0.760	**0.768**	**0.784**	**0.788**	**0.792**	**0.778**
Relevance Model (Tags Only)	**0.840**	0.728	0.744	0.741	0.756	0.752	0.735
Relevance Model (Dual Index)	0.720	**0.776**	**0.768**	0.755	0.754	0.760	0.763

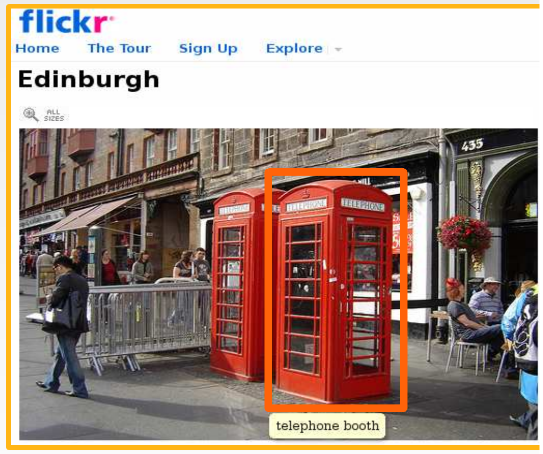
Use Visual Annotations

Flickr allows another kind of annotations (notes)

- Associate text with visual area
- Highly relevant to content

  → ***Visual Annotation***

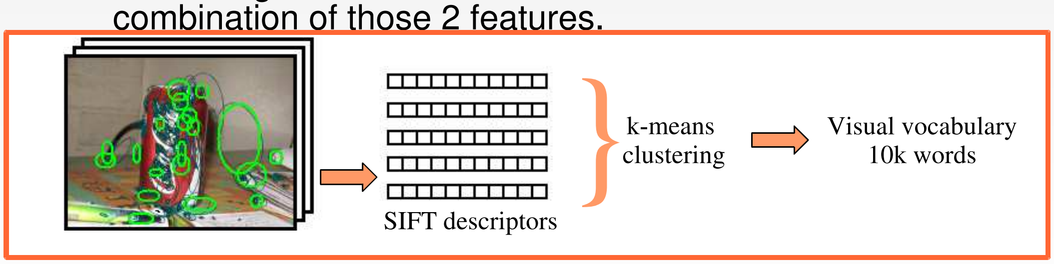
- Valuable to learn different the visual representations of an object



Olivares, Ciaramita, van Zwol. ACM Multimedia 2008

Content-based Image Retrieval

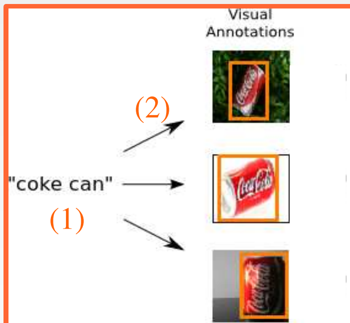
1. Extract visual features and describe them
   - Processed 12,000 images.
   - Computed Harris and Hessian features
   - Described using SIFT
2. Build visual vocabulary
   - Clustered SIFT descriptors to create vocabulary of 10,000 words
   - Implemented an approximate K-means algorithm
   - 3 resulting vocabularies: based on Harris, Hessian and a combination of those 2 features.



High-level search outline

(1) User performs a query (e.g. ”coke can”)

(2) Visual annotations matching the query are selected

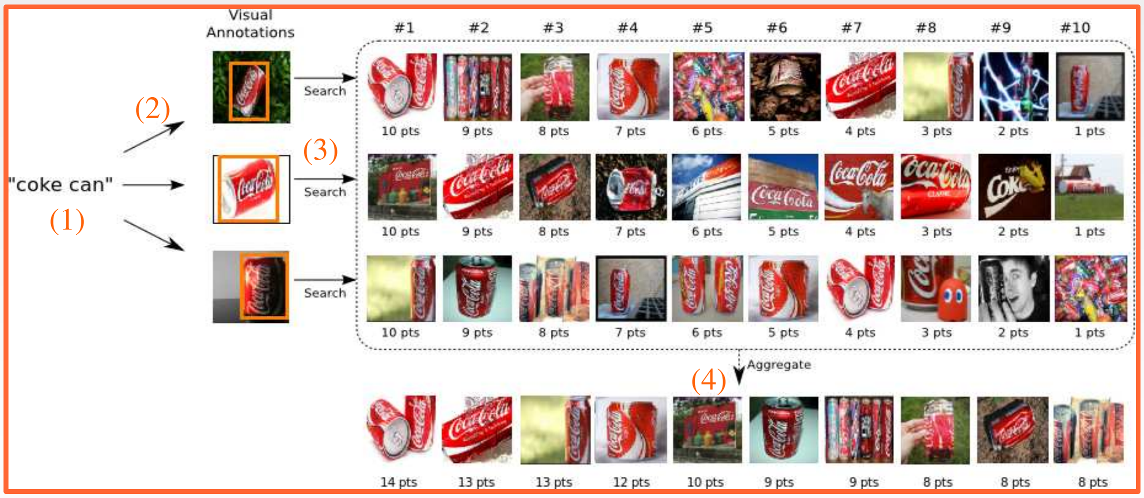


High-level search outline

(3) For each annotation, the top *k* similar images are retrieved, using content-based image retrieval

High-level search outline

(4) The result lists are aggregated to obtain the final result ranking

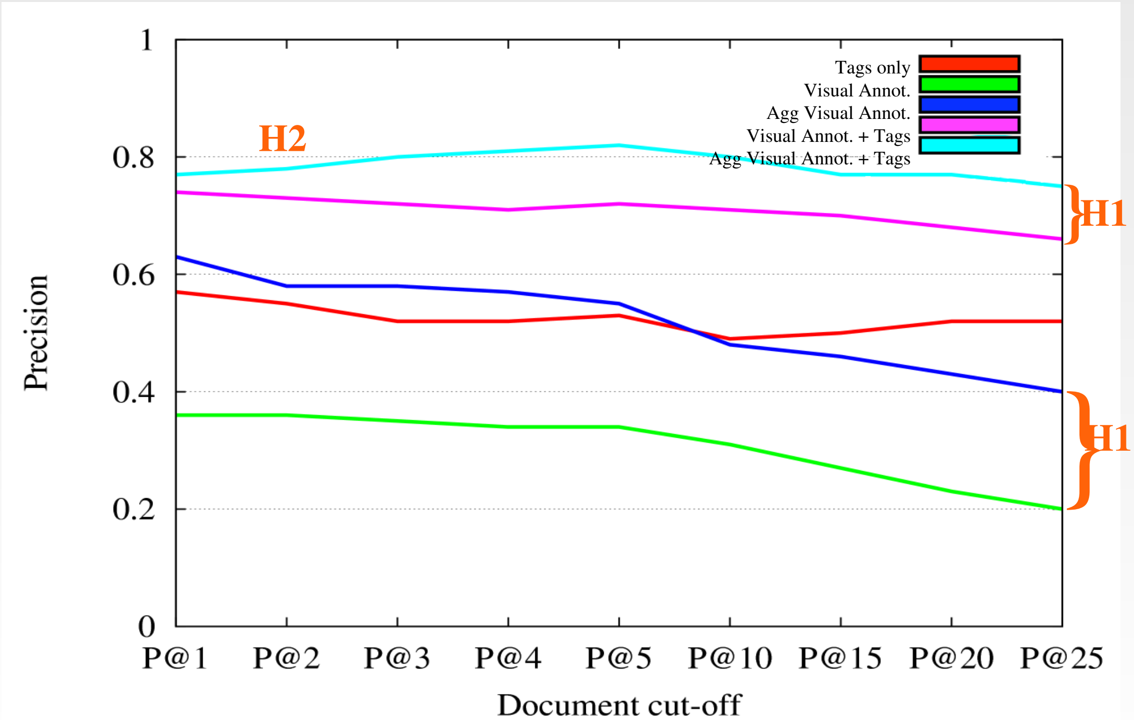


Evaluation

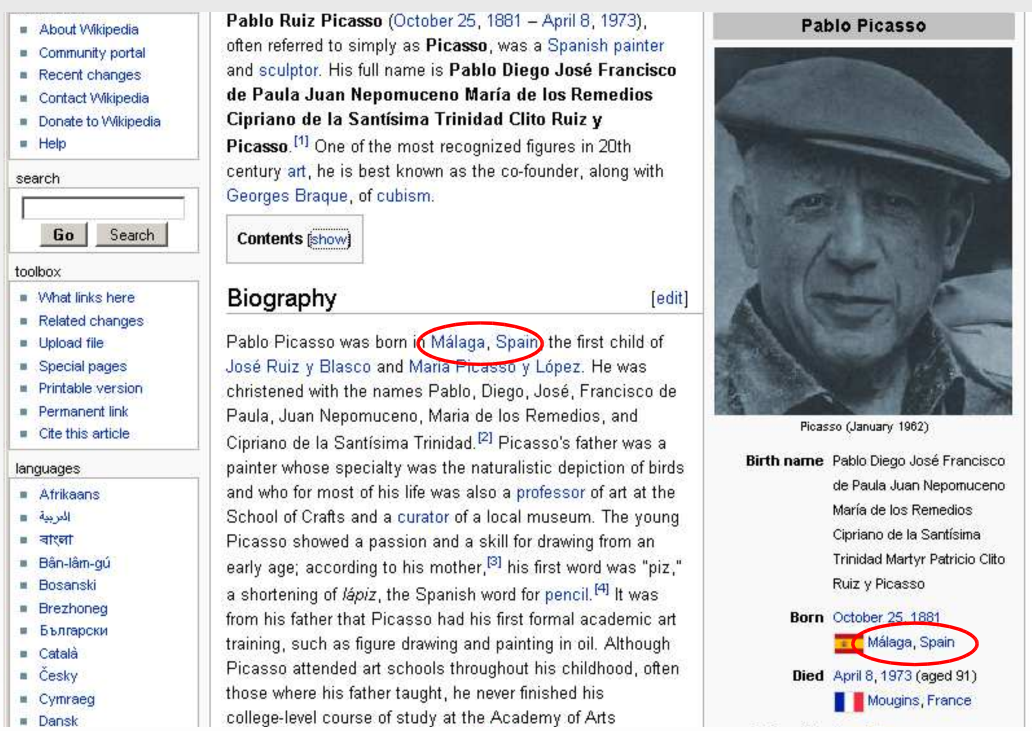
Hypotheses:

- H1: Rank aggregation using visual annotations will significantly improve the retrieval performance in terms of precision
- H2: Tag-based search combined with CBIR using visual annotations will improve retrieval in terms of precision

Results: Systems comparison



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

was born Picasso Málaga Pablo Spain

Text

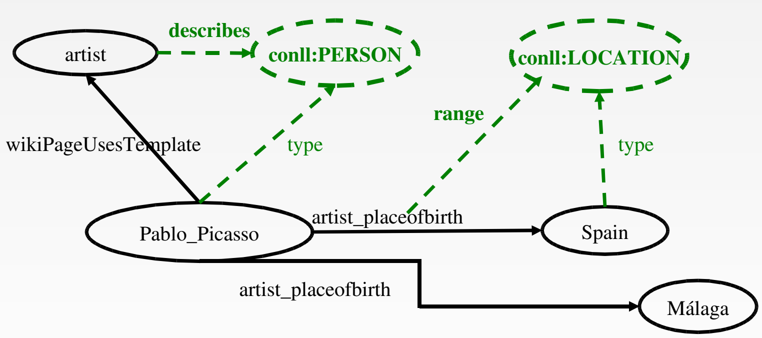
÷£¿≠¥ ÷ŝc£ËËð №£Ë ¿¥r© ŝ© X£≠£g£, Ë÷£ŝ©.
Pablo Picasso was born in Málaga, Spain.

Extending metadata

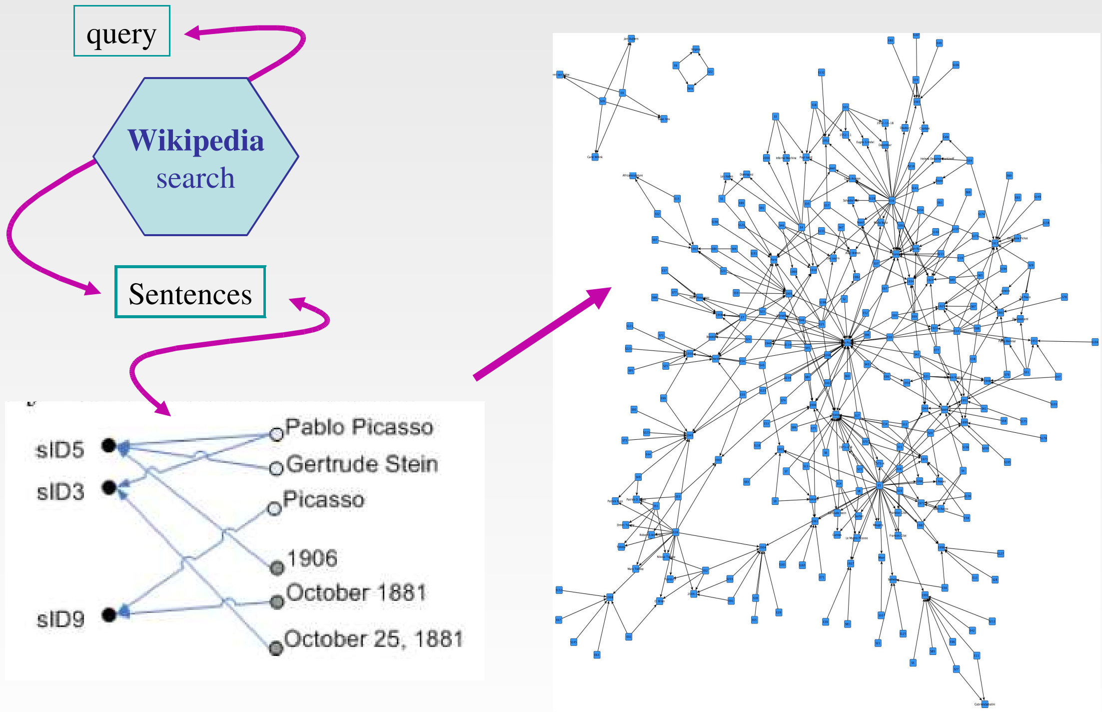
Pablo Picasso was born in Málaga, Spain.

Pablo Picasso	Málaga	Spain
PER	LOC	LOC
E:PERSON	GPE:CITY	GPE:COUNTRY
artist:name	artist:placeofbirth	artist:placeofbirth

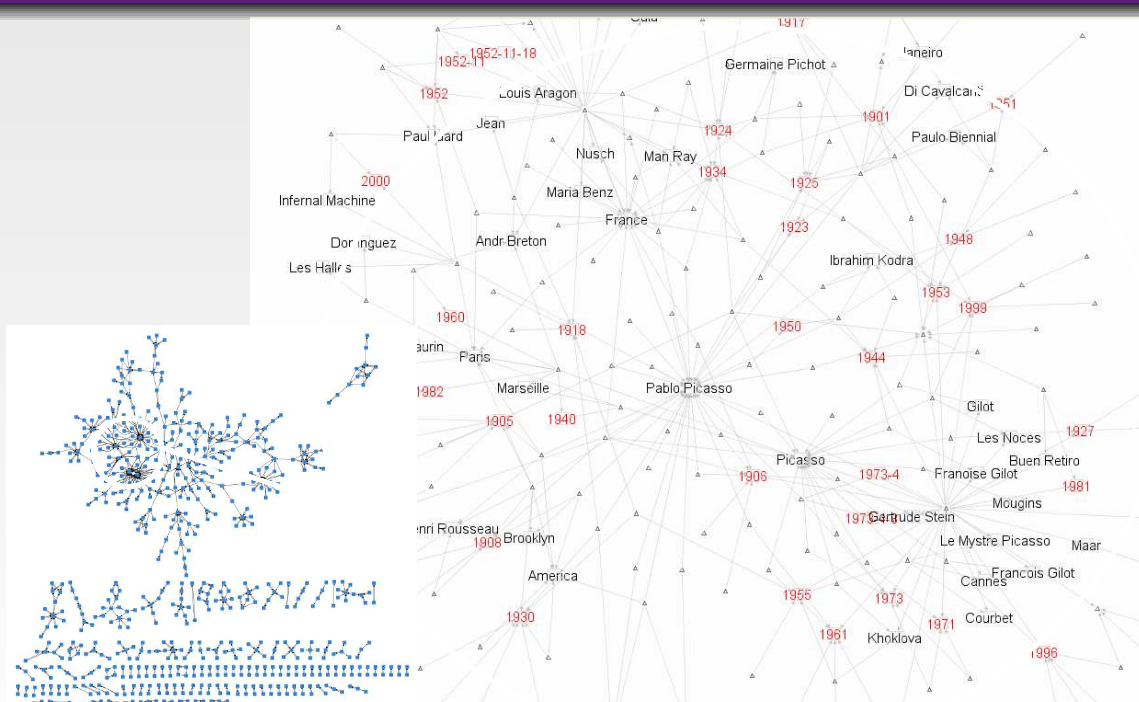
If most artists are persons, than let's assume all artists are persons.
If most places of birth are locations, then let's assume all are.



Entity Containment Graph



Example: Picasso

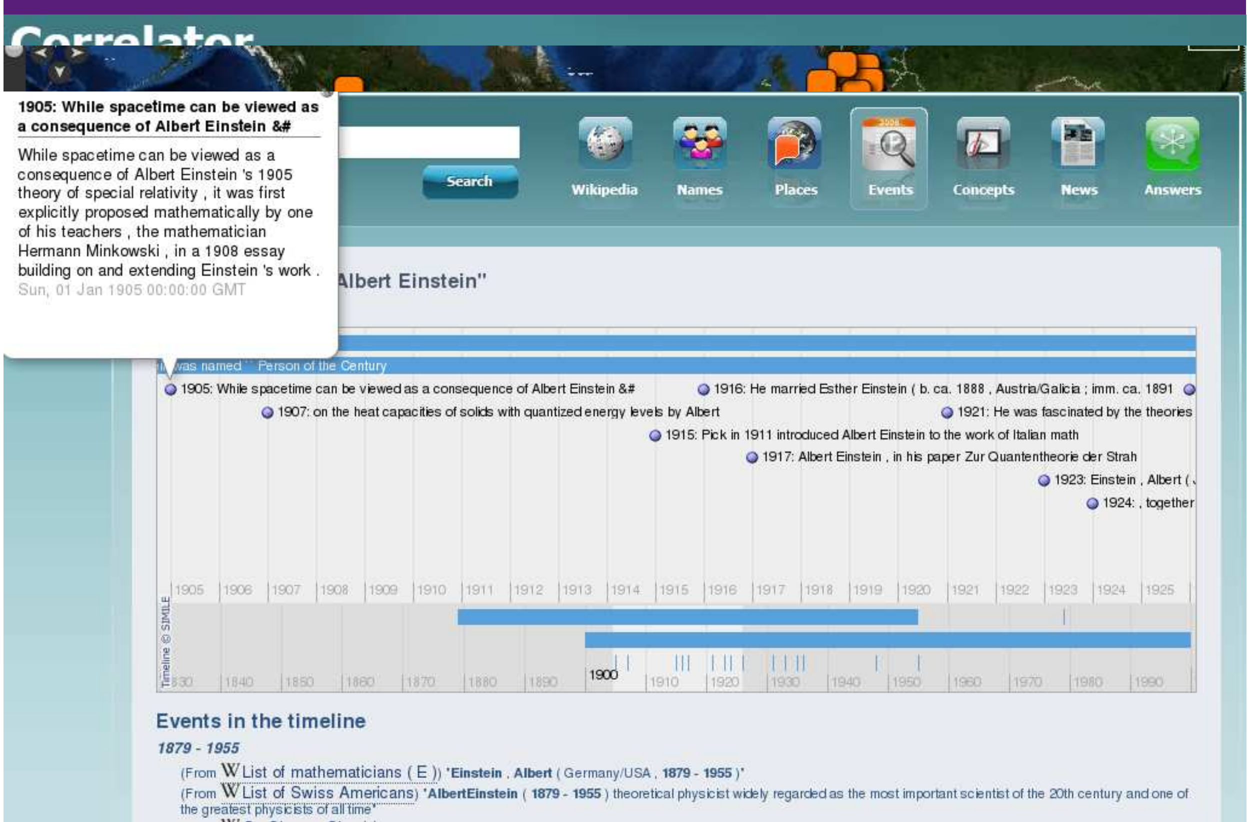


Correlator

- URL: correlator.sandbox.yahoo.com
- Find relations in the Wikipedia
  - Relate entities: names, places, dates
  - Change the result interface
- If the query is not an entry in the wikipedia
  - Synthetic page is created
- Based on linear time natural language parsing and competitive quality

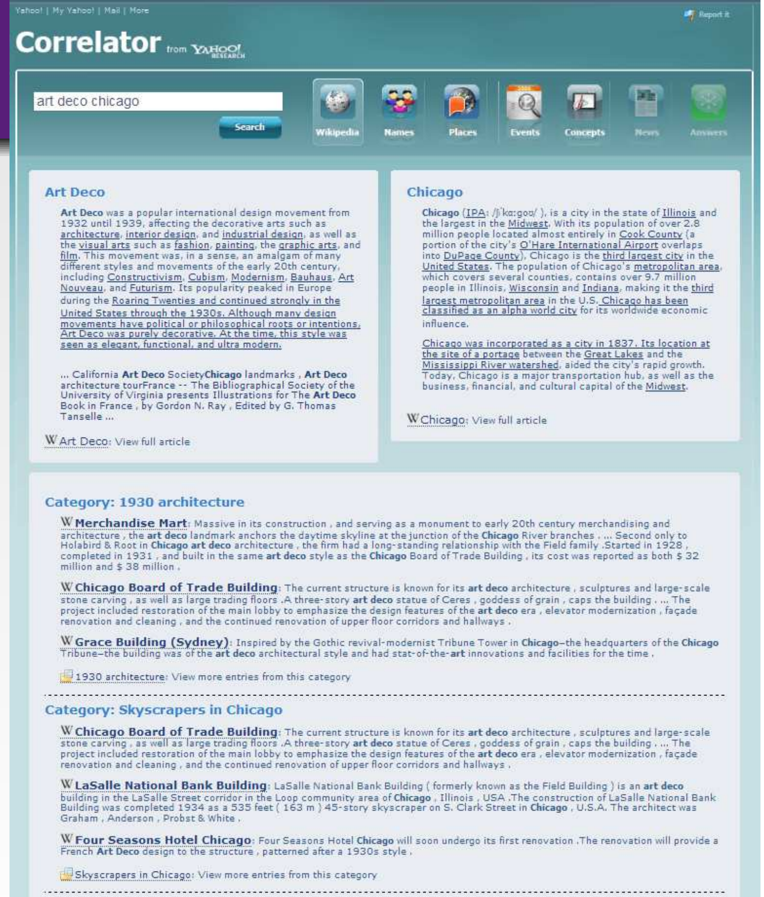
Zaragoza, Attardi, Ciaramita, Atserias, Castillo, Mika, Surdeanu, ...

Correlator - Examples



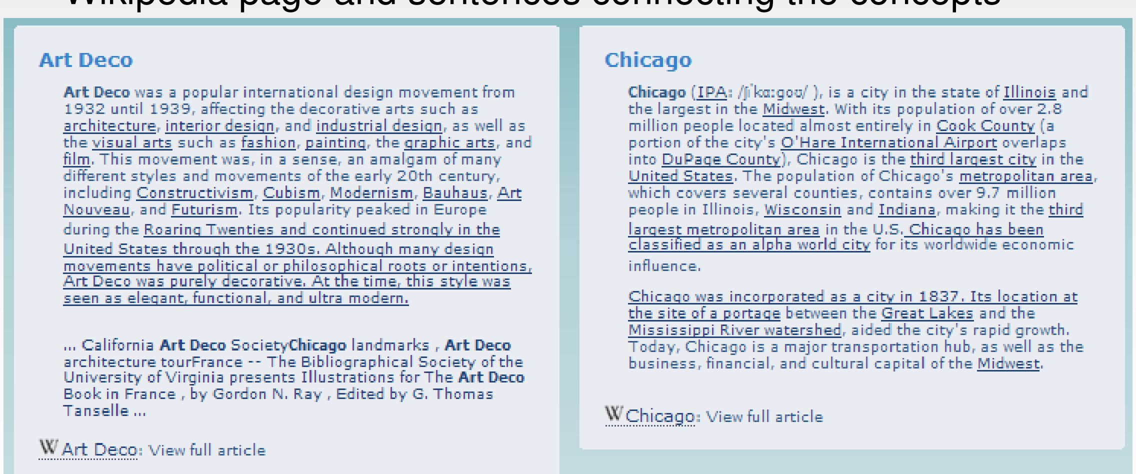
Overview page

- 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



Step 1: Definitions of query concepts

- Parse query using Wikipedia titles and redirects
  - **nyc parks => “New York City” parks**
  - **art deco chicago => “Art Deco” Chicago**
- Display first paragraphs of each from each concept’s Wikipedia page and sentences connecting the concepts



Step 2: Relations between query concepts (1/2)

- Retrieve related sentences
  - **Output: Ranked list of sentences**
- Aggregate sentences over Wikipedia pages
  - **Page score is the sum of the score of its sentences**
  - **Output: Ranked list of pages**
- Aggregate pages over Wikipedia categories
  - **Each relevant page votes for its categories**
  - **Category score is the sum of its votes**
  - **Output: Ranked list of categories containing relevant pages**

Step 2: Relations between query concepts (2/2)

Category: 1930 architecture

W **Merchandise Mart**: Massive in its construction , and serving as a monument to early 20th century merchandising and architecture , the **art deco** landmark anchors the daytime skyline at the junction of the **Chicago** River branches . ... Second only to Holabird & Root in **Chicago art deco** architecture , the firm had a long-standing relationship with the Field family .Started in 1928 , completed in 1931 , and built in the same **art deco** style as the **Chicago** Board of Trade Building , its cost was reported as both $ 32 million and $ 38 million .

W **Chicago Board of Trade Building**: The current structure is known for its **art deco** architecture , sculptures and large-scale stone carving , as well as large trading floors .A three-story **art deco** statue of Ceres , goddess of grain , caps the building . ... The project included restoration of the main lobby to emphasize the design features of the **art deco** era , elevator modernization , façade renovation and cleaning , and the continued renovation of upper floor corridors and hallways .

W **Grace Building (Sydney)**: Inspired by the Gothic revival-modernist Tribune Tower in **Chicago**–the headquarters of the **Chicago** Tribune–the building was of the **art deco** architectural style and had stat-of-the-**art** innovations and facilities for the time .

1930 architecture: View more entries from this category

Category: Skyscrapers in Chicago

W **Chicago Board of Trade Building**: The current structure is known for its **art deco** architecture , sculptures and large-scale stone carving , as well as large trading floors .A three-story **art deco** statue of Ceres , goddess of grain , caps the building . ... The project included restoration of the main lobby to emphasize the design features of the **art deco** era , elevator modernization , façade renovation and cleaning , and the continued renovation of upper floor corridors and hallways .

W **LaSalle National Bank Building**: LaSalle National Bank Building (formerly known as the Field Building) is an **art deco** building in the LaSalle Street corridor in the Loop community area of **Chicago** , Illinois , USA .The construction of LaSalle National Bank Building was completed 1934 as a 535 feet (163 m) 45-story skyscraper on S. Clark Street in **Chicago** , U.S.A. The architect was Graham , Anderson , Probst & White .

W **Four Seasons Hotel Chicago**: Four Seasons Hotel **Chicago** will soon undergo its first renovation .The renovation will provide a French **Art Deco** design to the structure , patterned after a 1930s style .

Skyscrapers in Chicago: View more entries from this category

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 /ˌkɑrnoʊ'tɔrəs/ 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** .

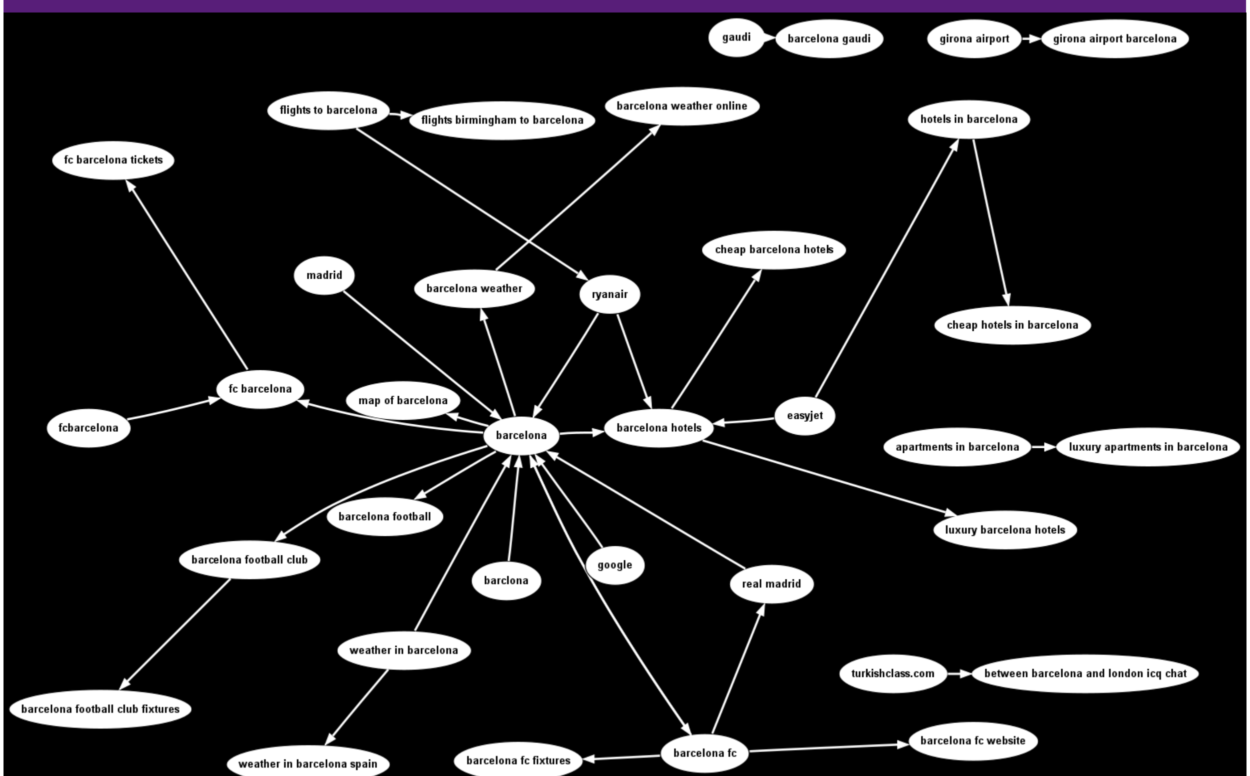
Argyrosaurus: Argyrosaurus (pronounced /ˌɑrdʒɪroʊ'sɔrəs/ AHR-ji-ro-SAWR-us) meaning 'Silver lizard' , because it was discovered in **Argentina** , which is sometimes known as 'Silver land ' (Greek argyros meaning 'silver ' and sauros meaning 'lizard') was a genus of herbivorous titanosaurid **dinosaur** 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 .

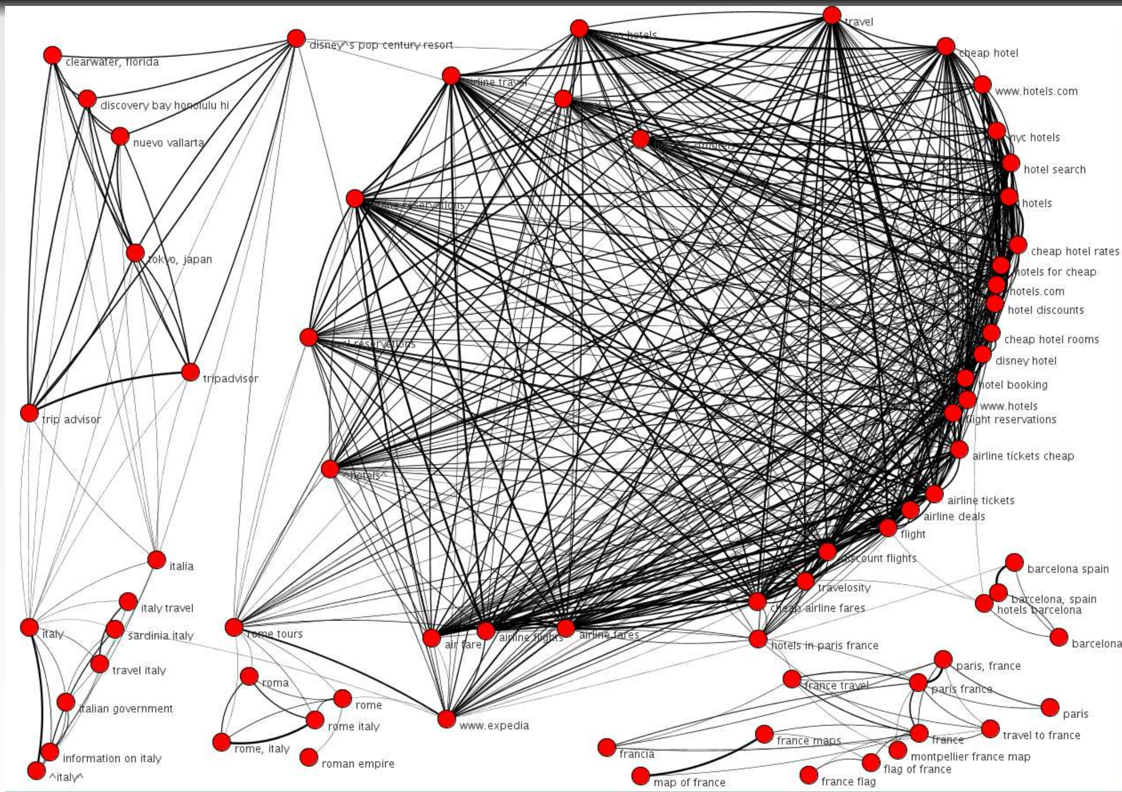
Genyodectes: Genyodectes (Woodward , 1901) is a genus of ceratosaurian theropod **dinosaur** from the Lower Cretaceous of South America .The holotype material (MLP 26-39 , Museo de La Plata , La Plata , **Argentina**) was collected from the Cañadón Grande , Departamento Paso de Indios in the Chubut Province 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 .

Queries as implicit tags

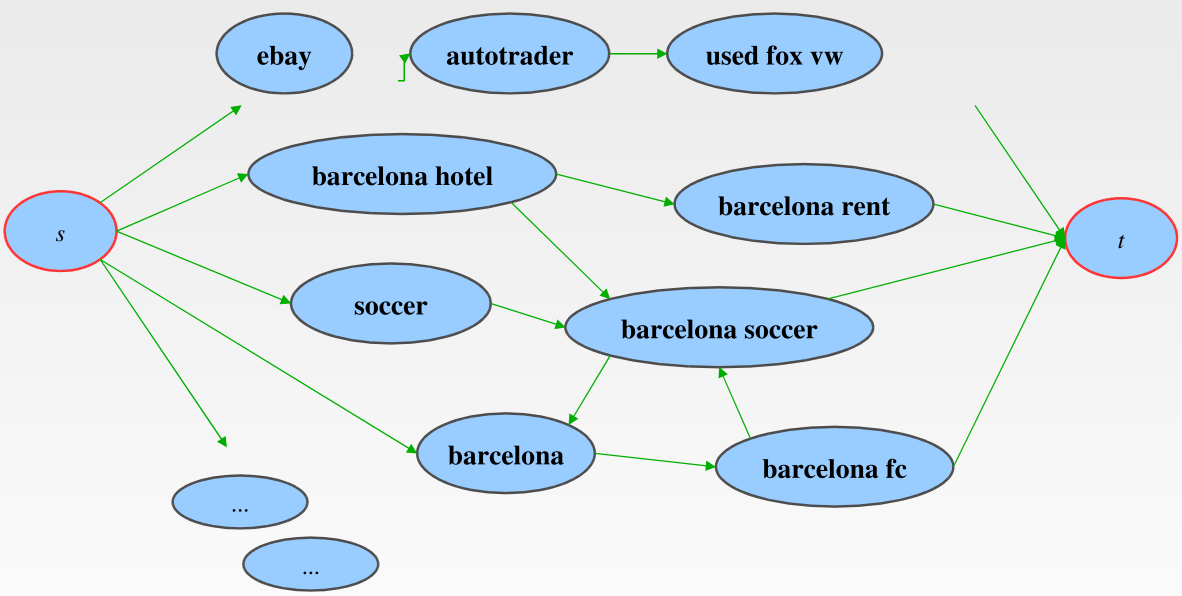


R. Baeza-Yates: “Graphs from search engine queries”. SOFSEM 2007.

Click Graph

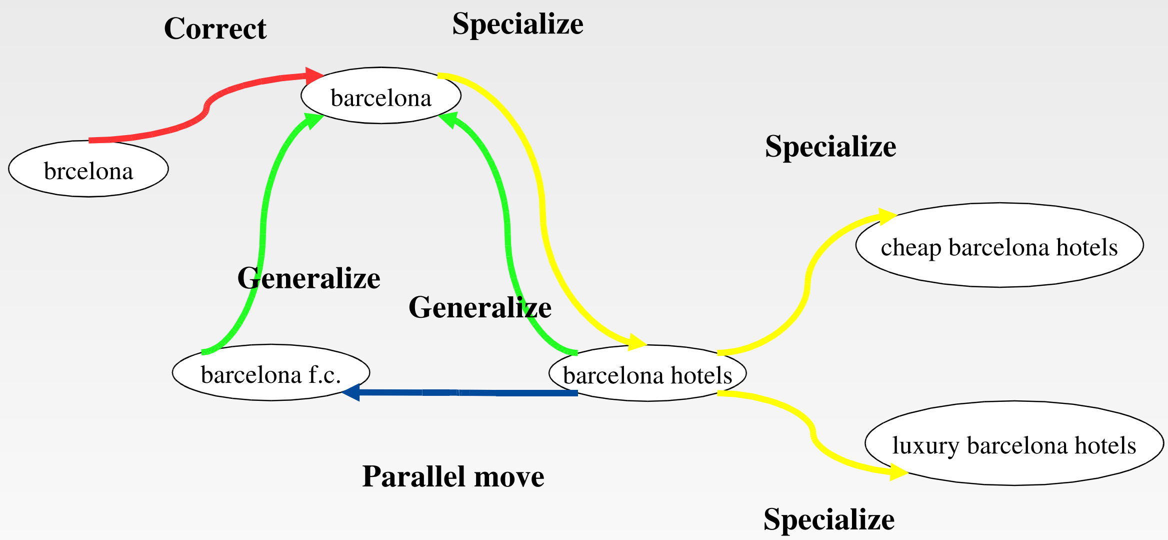


Session (Query-Flow) Graph



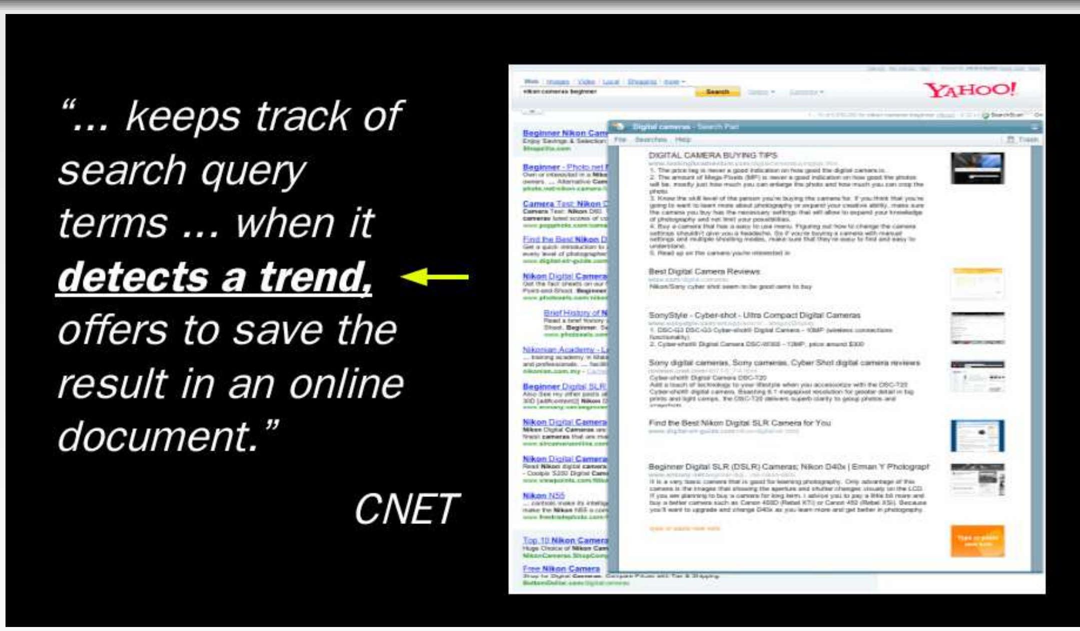
Boldi, Bonchi, Castillo, Donato, Gionis, Vigna. CIKM 2008.

Query-reformulation types



Rieh, S. Y. and Xie, H: "Analysis of multiple query reformulations on the web". IPM 2006

SearchPad



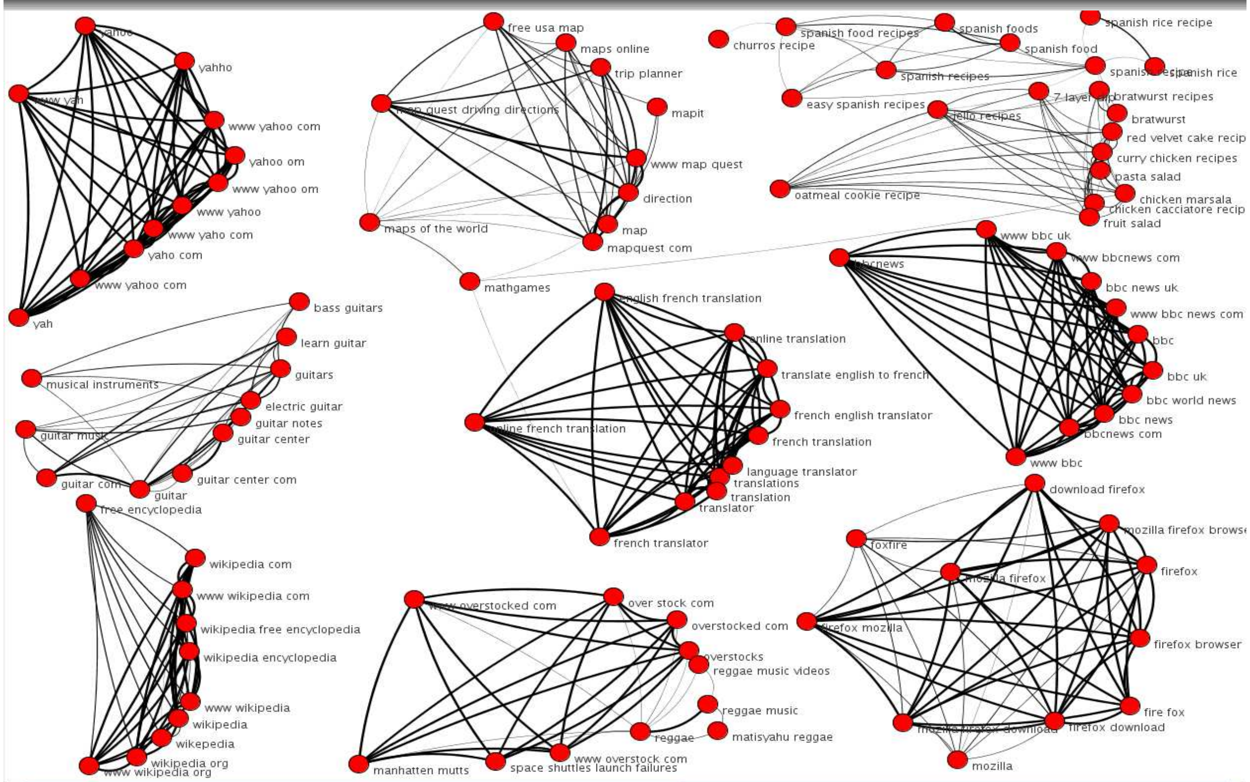
Research Session

- What are the characteristics of a research session?
- Possible scenarios:
  - Buying a house
  - Migraine treatment
  - Piano tuning
  - ....
- Detection using machine learning

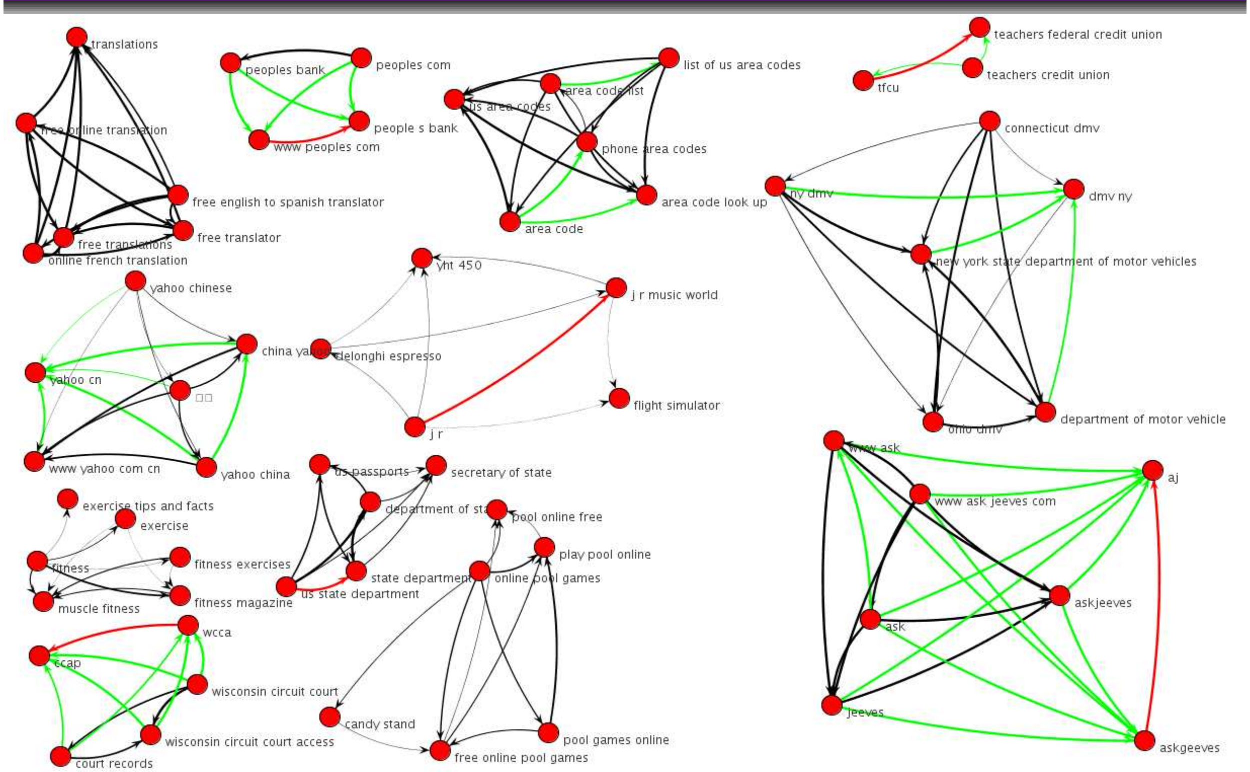
Research sessions

- **Complex activities**
  - Need to compare different sources of information
  - Carefully readings of most of the results returned by the SE
- **Usually performed during many physical sessions (weeks, months)**
  - The average time for buying a flat is 4 months
- **Need of taking notes and remembering past actions or decisions**

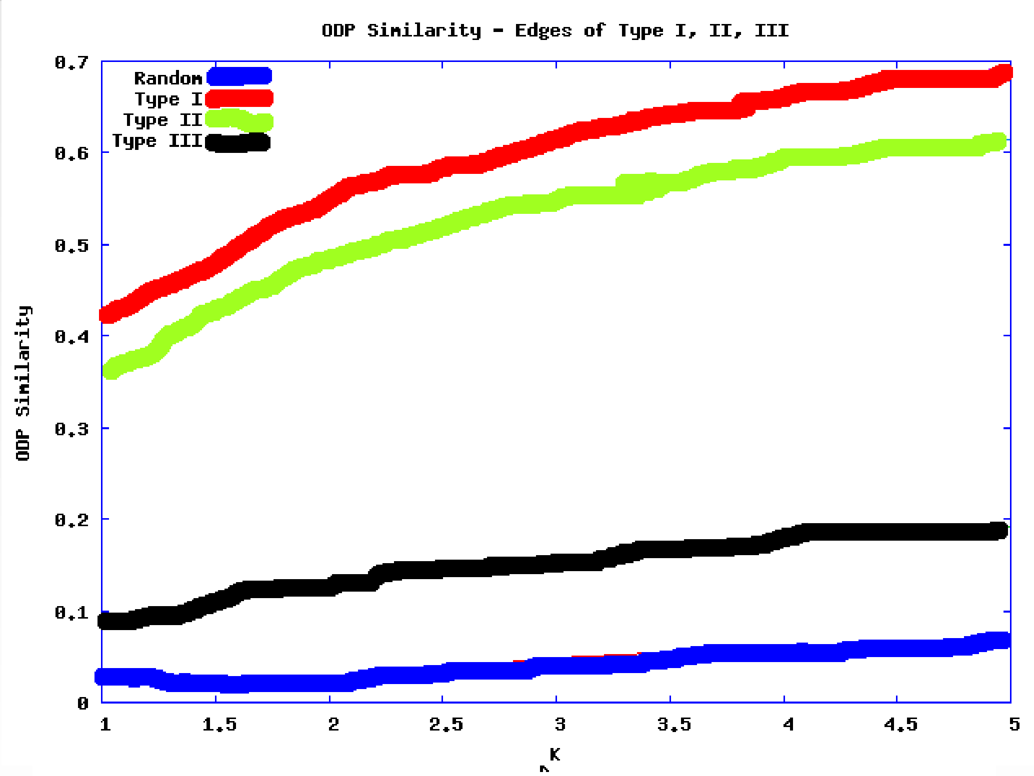
Implicit Folksonomy?



Implicit Knowledge? Web slang!



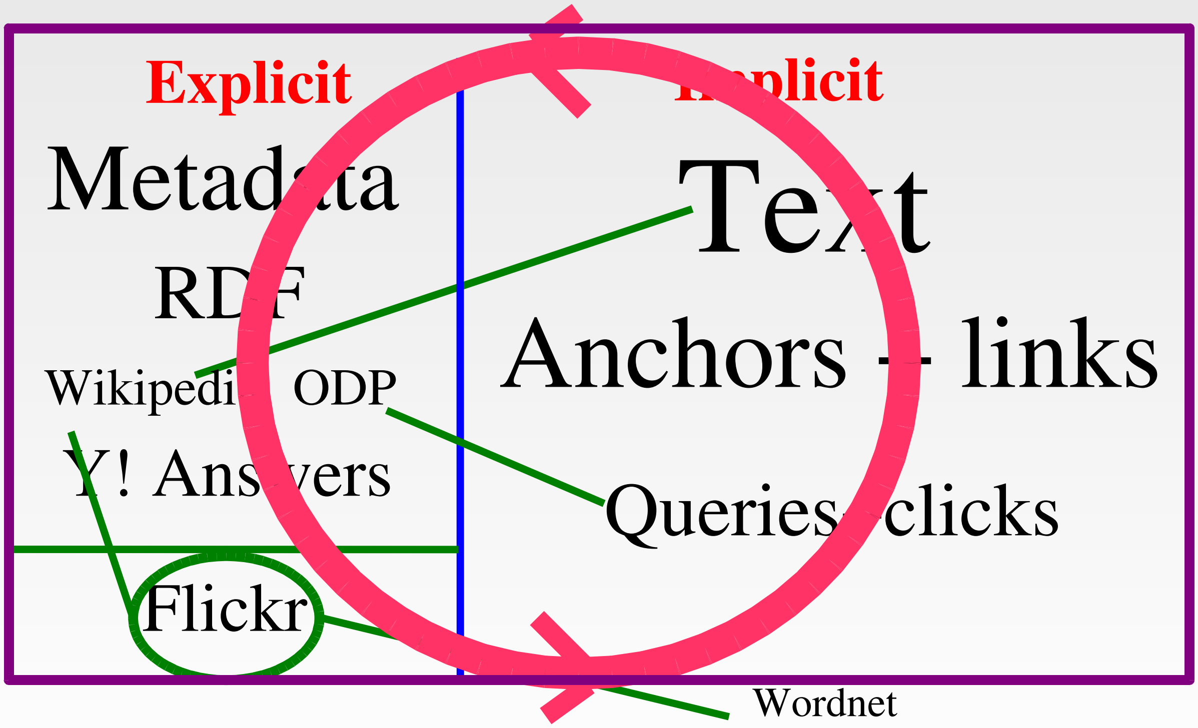
Experimental Evaluation



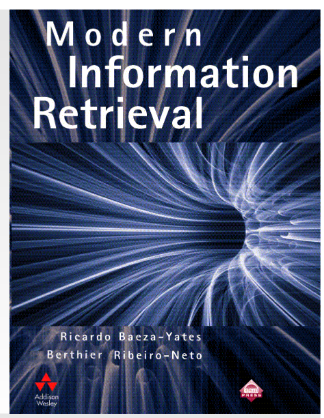
Open Issues

- Data Volume versus Better Algorithms
- Explicit versus implicit social networks
  - Any fundamental similarities?
- How to evaluate with (small) partial knowledge?
  - Data volume amplifies the problem
- User aggregation versus personalization
  - Optimize common tasks
  - Move away from privacy issues

The Virtuous Cycle



Second edition coming soon



Questions?

Contact: rbaeza@acm.org

Thanks to Carlos Castillo, Debora Donato, Aris Gionis, Peter Mika, Borkur Sigurbjornsson, Roelof van Zwol, Hugo Zaragoza