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

<table>
<thead>
<tr>
<th>Content type</th>
<th>Amount of content produced per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published content</td>
<td>3-4 GB</td>
</tr>
<tr>
<td>Professional web content</td>
<td>~ 2 GB</td>
</tr>
<tr>
<td>User generated content</td>
<td>8-10 GB</td>
</tr>
<tr>
<td>Private text content</td>
<td>~ 3 TB (300x more)</td>
</tr>
<tr>
<td>Upper bound on typed content</td>
<td>~700 TB (~200x more)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Amount of metadata produced per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchortext</td>
<td>100 MB</td>
</tr>
<tr>
<td>Tags</td>
<td>40 MB</td>
</tr>
<tr>
<td>Pageviews</td>
<td>180 GB</td>
</tr>
<tr>
<td>Reviews</td>
<td>Around 10 MB</td>
</tr>
</tbody>
</table>

[Ramakrishnan and Tomkins 2007]

### Examples

**Explicit Metadata**
- Wordnet
- RDF
- Wikipedia
- ODP
- Y! Answers
- Flickr

**Implicit Text**
- Blogs, Groups
- Anchors + links
- Queries + clicks
- Private

**Scale**

**Quality?**
• 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

  “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”.

Aggregating data
The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

- 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

Query → flickr → unsorted photos

- tag graph
  - tag:type
  - tag:type
  - tag:type

WORDNET
Wikipedia
flickr

- lights
- architecture
- sun set
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)

---

### Retrieval Performance

- **Unambiguous topics**

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

- **Ambiguous topics**

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

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

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

Pablo Ruiz Picasso (October 25, 1881 – April 8, 1973), often referred to simply as Picasso, was a Spanish painter and sculptor. His full name is Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso. One of the most recognized figures in 20th century art, he is best known as the co-founder, along with Georges Braque, of cubism.
Pablo Picasso was born in Málaga, Spain.

Extending metadata

Pablo Picasso was born in Málaga, Spain.

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, ....
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, this firm had a long standing relationship with the Field family. Started in 1889, completed in 1931, and built in the same art deco style as the Chicago Board of Trade Building, its cost was reported as both $52 million and $39 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, facade 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 state-of-the-art innovations and facilities for the time.

**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, facade 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 1924 as a 525 foot (156 m) 48-story skyscraper on S. Clark Street in Chicago, USA. 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.
Synthetic page - example

Category: Dinosaurs of South America

W. Bautavator: It was found in Argentina and was described in 2006. The fossilized bones were found in 2005 in sandstone in Patagonia, Argentina - by an excavation led by Richard McDonald, curator of dinosaurs at the Field Museum in Chicago. Bautavator was discovered in the same bone pit that had earlier yielded Giganotosaurus, one of the largest known carnivorous dinosaurs.

W. Herrerasaurus: Herrerasaurus (meaning “Herrera is lizard”) after the name of the rancher who discovered the first bones of the animal was one of the earliest dinosaurs. This View is further supported by ichnological records showing large trackability footprints that can be attributed only to theropod dinosaurs, dating from the Late Triassic (middle Toarcian) to the Late Triassic (Late Precordian) 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.

W. Unaysaurus: It was recovered from the red beds of the Santa María Formation (also known as the Catarín Formation), which is the geologic formation where similar big dinosaurs like Saltasaurus 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.

W. Carnocharus: Carnotaurus (pronounced/karno-tors/KAH-noh-TAWS-us; meaning “meat-ball”, referring to its distinct ball-like bones (Latin caro = flesh – Greek touros = bull) was a large predatory dinosaur with horns vaguely resembling a njj. Carnotaurus lived in Patagonia, Argentina during the Valanginian stage of the Late Cretaceous, and was discovered by José F. Bonaparte, who has discovered many other large South American dinosaurs. Together, these dinosaurs form the subfamily Carnotaurinae in the family Abelisauridae.

W. Eoraptor: Eoraptor was one of the world’s earliest dinosaurs. Early dinosaurs The bones of these primitive dinosaurs were first discovered in 1991 by University of Chicago paleontologist Paul Sereno in the rock formations of Argentina.

W. Argysaurus: Argysaurus (pronounced/arg-zor-es/ARG-ys-ors; meaning “Silver Beast” because it was discovered in Argentina, which is sometimes known as “Silver Land” (Greek argyros meaning silver) and sources meaning beast) was a genus of herbivorous theropod dinosaur that lived about 70 million years ago, during the Late Cretaceous Period in what is now South America (Argentina and Uruguay). It was one of the largest dinosaurs, having a height of 4.8 meters, a length of up to 8.2 meters, and a weight of up to 60 meters. It was a herbivore.

W. Argentinosaurus: Argentinosaurus (meaning “Argentina lizard”) was a titanosaurid sauropod dinosaur genus that was among the largest and heaviest ever lived. Argentinosaurus is featured prominently in the permanent exhibition of the Field Museum of Natural History in Chicago, Illinois. USA. The display depicts a hypothetical portrait of a Argentinosaurus and the largest armored dinosaur Giganotosaurus. At 120 feet long, this skeletal reconstruction represents the largest dinosaur mount ever to be assembled.

W. Neuquensaurus: Neuquensaurus (meaning “Neuquen lizard”) is a sauropod dinosaur that appeared in the Late Cretaceous, 71 million years ago in Argentina and Uruguay. This dinosaur was 10-15 meters (30-45 feet) long, and is believed to have possessed armor-like osteoderms.

W. Guanadocetes: Guanadocetes (Woodward, 1901) is a genus of ceratopsian theropod dinosaur from the Lower Cretaceous of South America. The holotype material (MLP 33-39, Museo de La Plata, La Plata, Argentina) was collected from the Ceballos Group, Departamento Paso de los Indios in the Chubut Province of Argentina and consists of an incomplete skull, including the premaxilla, portions of both maxillae, the right and left dentary, many teeth, a fragment of the infratemporal, and parts of the squamosal bones.

Queries as implicit tags

Click Graph

Session (Query-Flow) Graph

Query-reformulation types

Correct

Specialize

barcelona

brcelona

Generalize

Generalize

barcelona f.c.

barcelona hotels

Parallel move

Specialize

cheap barcelona hotels

luxury barcelona hotels


SearchPad

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

CNET
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

Explicit

Metadata
RDF
Wikipedia
ODP
Y! Answers
Flickr

Implicit

Text
Anchors - links
Queries - clicks

Wordnet

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