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

<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**
- Wordnet
- RDF
- Wikipedia
- ODP
- Y! Answers

**Implicit**
- Blogs, Groups
- Anchor text + links
- Queries + clicks

Quality?
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

  “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

Explore / Tags / prado / clusters

```
madrid, spain, museum, museo, art, europe, museodelprado, goya, painting, espana

See more in this cluster...
```

```
verde, cielo, green, campo, nubes, hierba, sky, azul, flores, primavera

See more in this cluster...
```

```
paisaje, landscape, naturaleza, nature

See more in this cluster...
```

```
cuba, havana, habana, lahabana

See more in this cluster...
```
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!)
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

![Diagram showing tag graph and tag types](image)
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

Sigurbjörnsson 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 ....

London Eye

London Eye and Golden Jubilee Bridge seen from Westminster Bridge.

Tag list

london eye, thames,

Suggested tags

- London
- England
- UK
- River
- Eye
- South Bank
- Big Ben
- Night
- Bridge
- 2006

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)


<table>
<thead>
<tr>
<th>Retrieval Performance</th>
</tr>
</thead>
</table>

• 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><strong>0.779</strong></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><strong>0.779</strong></td>
<td>0.726</td>
<td>0.717</td>
<td>0.719</td>
<td>0.714</td>
<td>0.710</td>
<td><strong>0.683</strong></td>
</tr>
<tr>
<td>Relevance Model (Dual Index)</td>
<td>0.768</td>
<td><strong>0.754</strong></td>
<td><strong>0.739</strong></td>
<td><strong>0.726</strong></td>
<td><strong>0.719</strong></td>
<td><strong>0.716</strong></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><strong>0.768</strong></td>
<td><strong>0.784</strong></td>
<td><strong>0.788</strong></td>
<td><strong>0.792</strong></td>
<td><strong>0.778</strong></td>
</tr>
<tr>
<td>Relevance Model (Tags Only)</td>
<td><strong>0.840</strong></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><strong>0.776</strong></td>
<td><strong>0.768</strong></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
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**

![Graph showing precision vs. document cut-off for different systems comparison]

**Bridging implicit and explicit metadata**

Pablo 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.

**Biography**

Pablo Picasso was born in Málaga, Spain, the first child of José Ruiz y Blasco and Maria Picasso y López. He was christened with the names Pablo, Diego, José, Francisco de Paula, Juan Nepomuceno, María de los Remedios, and Cipriano de la Santísima Trinidad. Picasso's father was a painter whose specialty was the naturalistic depiction of birds and who for most of his life was also a professor of art at the School of Crafts and a curator of a local museum. The young Picasso showed a passion and a skill for drawing from an early age, according to his mother; his first word was "piz," a shortening of 'pizarrón,' the Spanish word for pencil. It was from his father that Picasso had his first formal academic art training, such as figure drawing and painting in oil. Although Picasso attended art schools throughout his childhood, often those where his father taught, he never finished his college-level course of study at the Academy of Arts.
Pablo Picasso was born in Málaga, Spain.

Extending metadata

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

Art Deco

Art Deco was a popular international design movement from 1925 until 1939, affecting the decorative arts such as architecture, interior design, and industrial design, as well as the visual arts such as fashion, painting, the graphic arts, and film. This movement was, in essence, an amalgamation of many different styles and movements of the early 20th century including Cubism, Futurism, Futurism, Art Nouveau, and Esherism. Its popularity peaked in Europe during the Roaring Twenties and continued strongly in the United States through the 1930s. Although many design practitioners had political or philosophical goals or intentions, Art Deco was mainly decorative. At this time, this style was seen as “neat, functional, and world-shaping.”

[View full article]

Chicago

Chicago (Chicago) is a city in the state of Illinois and the largest city in the Midwest. With a population of over 2.8 million people located almost entirely in Cook County (a portion of the city’s O’Hare International Airport overlaps into Dupage County), Chicago is the third largest city in the United States. The population of Chicago’s metropolitan area, which covers several counties, contained over 9 million people in Illinois, Wisconsin, and Indiana, making it the third largest metropolitan area in the U.S. Chicago has been classified as an Alpha World City for its worldwide economic influence.

Chicago was incorporated as a city in 1837, its location at the point where the Chicago River empties into Lake Michigan. Today, Chicago is a global transportation hub, as well as the business, financial, and cultural capital of the Midwest.

[View full article]
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**

- **Merchandise Mart:** Massive in its construction, and serving as a monument to early 20th century merchandising and architecture, this Art Deco landmark anchors the daytime skyline at the junction of the Chicago River branches. Second only to Holabird & Roche's Art Deco architecture, the firm had a long-standing relationship with the Field family. Started in 1929, 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 $39 million.

- **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, tops 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.

- **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**

- **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, tops 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.

- **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 1936 as a 523 feet (159 m) 45-story skyscraper on S. Clark Street in Chicago, USA. The architect was Graham, Anderson, Probst & White.

- **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

W. Buitraperator. It was found in Argentina and was described in 2005. The fossilized bones were found in 2006 in sandstone in Patagonia, Argentina - by an excavation led by Paul McGowan, curator of dinosaurs at the Field Museum in Chicago. Buitraperator was discovered in the same fossil site that had earlier yielded Giganotosaurus - one of the largest known carnivorous dinosaurs.

W. Herrerasaurus. Herrerasaurus (meaning "Herreras is lizard") is the name of the (now extinct) animal that discovered the first fossils of a reptile, one of the earliest dinosaurs. This view is further supported by a new fossil find, which shows that herrerasaurids are the only known dinosaurs, and dates from the Late Triassic (about 230 million years ago). The view that herrerasaurids are the closest relatives to the dinosaurs has led to the idea that dinosaurs and crocodiles are a monophyletic group, i.e., a lineage of organisms that share a common ancestor.

W. Utanysaurus: Utanysaurus was recovered from the red beds of the Ceballos Formation (also known as the Ceballos Formation), which is the geologic formation where similar old dinosaurs like Utanysaurus have been found. The oldest dinosaurs in the world were from here and nearby in Argentina (like the Eoraptor), which suggests that the first dinosaurs may have originated in this area.

W. Carnottaurus. Carnottaurus (pronounced kar-NOH-tawr-us) is a genus of theropod dinosaur discovered in the Late Triassic of Argentina and named for its knobby spines on its back and tail. This dinosaur was a predator, and its fossilized remains have been found in the same area as other Carnottaurus fossils.

W. Eoraptor. Eoraptor was one of the world’s first dinosaurs. Early dinosaurs were first discovered in 1991 at a quarry in Argentina. The quarry contained a fossilized dinosaur, and these fossils are among the oldest known dinosaurs. The quarry contains a fossilized dinosaur, and these fossils are among the oldest known dinosaurs.

W. Argentinosaurus. Argentinosaurus (pronounced AR-gent-I-noh-SAWR-us) is a genus of sauropod dinosaur discovered in the Late Cretaceous Period, which is the most recent period of the Cretaceous. Argentinosaurus was one of the largest land animals that ever lived, and it was discovered in Argentina. It is one of the largest land animals that ever lived, and it was discovered in Argentina.

W. Neuquensaurus. Neuquensaurus is a genus of sauropod dinosaur discovered in the mid-Cretaceous Period. Its fossils were first discovered in the Neuquén Province of Argentina, and it was one of the largest land animals that ever lived. However, it has not been confirmed as a sauropod, and it is not among the largest land animals that ever lived.

W. Gephyrostegid. Gephyrostegids (Gephyrostegus) is a genus of aquatic amphibian discovered in the Late Triassic Period. Its fossils were first discovered in Argentina, and it is not among the largest land animals that ever lived.
Click Graph

Session (Query-Flow) Graph

Query-reformulation types

Correct

Specialize

Generalize

Generalize

Parallel move

Specialize

Specialize


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

Text
Anchors + links
Queries + clicks

Explicit
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