Mining the Web 2.0 to improve Search

Ricardo Baeza-Yates
VP, Yahoo! Research

Agenda

• The Power of Data
• Examples
  – Improving Image Search (Faceted Clusters)
  – Searching the Wikipedia (Correlator)
  – Understanding Queries (Search Pad)
• 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

**Quality?**

**Scale**
• 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”.

---

**The Wisdom of Crowds**

- "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

Semantic Breakup of Tag Clouds

• Tag-cloud is organized by grouping together tags that have similar meaning
• The related terms are calculated using a probabilistic framework using different conditional probabilities to get a mixture of general and specific terms
• 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

Sigurbjörnsson and van Zwol, WWW 2008
Overell, Sigurbjörnsson 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

- 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

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

Update annotation

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><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 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

1. 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]

Application: Faceted Image Search

![Image search results for "Barcelona"]
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. He is one of the most recognized figures in 20th century art, he is best known as the co-founder, alongside Georges Braque, of cubism.

Biography

Pablo Picasso was born in Málaga, Spain, the first child of José Ruiz y Blasco and María 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. His father was a jeweler 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, he first learned to draw with a "piz," a shortening of lápiz, 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.

City-in-Country

{Person, Painter, XXth Century}

born-in

Málaga, Spain

Language, Text, Search & “Semantics”...

Semantics

NLP

IR

Text

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

query

Wikipedia search

Sentences
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
- 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 the Rookwood Kiln in Chicago art deco architecture, the firm had a long-standing relationship with the firm starting 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.
- 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.
- Grace Building (Sydney): Inspired by the Gothic revival modernist Tribune Tower in Chicago, the headquarters of the Chicago Tribune—building just 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, 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.
- 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 (160 m) 42-story skyscraper on S. Clark Street in Chicago, U.S.A. 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.

Category: Dinosaurs of South America
- Butteraptor: It was found in Argentina and was described in 2005. The fossilized bones were found in 2003 in sandstone in Patagonia, Argentina by an excavation led by PHD student in Chicago, curator of dinosaurs of the Field Museum in Chicago. Butteraptor was discovered in the same molasse that had early yielded Gigantoraptor, one of the largest known carnivorous dinosaurs.
- Herrerasaurus: Herrerasaurus ([her-er-uh-sawr-us]), after the name of the wash where the first fossil was found, is one of the earliest dinosaurs... This view is further supported by the phylogeny of the large theropod, which can be attributed only to the theropod dinosaur, dating from the Late Triassic period of the Late Triassic Formation in Argentina. The importance of the Herrerasauridae and Eoraptor is that their remains allow for the first time for the reconstruction of the life 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 María Formation, also known as the Catarinian Formation, which is the geologic formation where similar old dinosaurs like Iguanodon 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 kah-roht-uh-russ) is a large, robust theropod with short arms and large, powerful forelimbs. Carnotaurus lived in Patagonia, Argentina during the Middle Jurassic period of the Late Triassic, and was discovered by José F. Bonaparte, who has uncovered many other popular South American dinosaurs. Together, these dinosaurs form the subfamily Carnotauroidea 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 Colorado Paleontologist Paul Sereno, in the Chubut Province of Argentina.
- Argentinosaurus: Argentinosaurus (Argentino = Argentina and saur = dinosaur) is a genus of herbivorous theropod dinosaur that lived about 90 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 12 meters, a length of up to 30 to 35 meters, and weighed up to 80 tons it was a herbivore.
- Argentinosaurus: Argentinosaurus (Latin: Argen- = Argentina, tino- = dinosaur, saur- = reptile) is a large, herbivorous sauropod dinosaur genus that was among the largest land animals that ever lived... Argentinosaurus is estimated to have weighed about 100 metric tons, which is about the weight of 19 fully grown elephants. It was discovered in 2010 in the province of Entre Ríos in Argentina.
- Neuquensaurus: Neuquensaurus (Neuquen = Neuquen, saur = dinosaur) is a small sauropod dinosaur that lived in the Late Cretaceous, 71 million years ago in Argentina and Uruguay. In South America, this dinosaur was 10 to 15 meters (34 to 51 feet) long, and is believed to have possessed armor-like osteoderms.
- Gonyphodon: Gonyphodon (Greek, gonyphodon = a crocodile) is an extinct crocodyliform that lived from the Lower Cretaceous to the Oligocene of South America. The only known specimen is a single skull from the Lower Cretaceous of Brazil. This reptile is part of the family Coelurosauria, which includes crocodyliforms and other extinct groups of crocodylians.
Queries as implicit tags

Click Graph

Query-reformulation types

Correct  Specialize

correct

barcelona

Generalize  Generalize

Generalize

cheap barcelona hotels

Parallel move

cheap barcelona hotels

Specialize

correct

“... 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
  – ....
• Need for taking notes and remembering past actions and decisions
• Detection using machine learning
Implicit Folksonomy?

Implicit Knowledge? Web slang!
Experimental Evaluation

![Graph showing ODP Similarity - Edges of Type I, II, III](image)

- Cliques & articulation points
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

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

More data via larger communities, makes data less personalized

\textit{wisdom of crowds does not work well on small corpora}

Privacy

Over personalization endangers privacy \text{ → }

Contextualization doesn't

Long-term logs endanger privacy

Contextualize the task: query intent detection

The Virtuous Cycle

Explicit

Metadata

RDF

Wikipedia, ODP

Y! Answers

Flickr

Text

Anchors + links

Queries + clicks

Explicit

Wordnet