The Web as a Semantic Source

Ricardo Baeza-Yates
VP, Yahoo! Research

and Applications

Outline

• Internet moved from a curiosity to a substrate for life activity
• Content growing, changing, diversifying, fragmenting
• Semantics of content unlock the value of the data ecosystem
• Explicit and implicit semantic sources
• Applications
• A virtuous feedback cycle for enhancing semantics
Content Growth and Trends

Content 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>

[Ramakrishnan and Tomkins 2007]
Metadata trends

<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]

Content ownership

- Content consumption is fragmenting – nobody owns more than 10% of the Web page views
- No single place will own all the content
- Best of breed processing will operate on the web version (?)
- Value transitions to ecosystem
Content Consumption is fragmenting

By age...

By topic

Content access is fragmenting
Time of Openness

Semantic Sources
Semantic Sources in the Web

- Intrinsically Collaborative
- Explicit or Implicit
- Taxonomies vs. Folksonomies
- Different Size and Growth
- Diversity of Quality
- Public or Private

Examples

Explicit
Metadata
  RDF
  Wikipedia ODP
  Answers

Implicit
Text
  Anchors + links

Scale
  Queries+clicks
    Private

UGC
Wordnet
Social properties had 115M unique visitors worldwide, 56M “under 35”.

- **Yahoo! Groups**: 8 million groups, 1 in 10 Internet users
- **Del.icio.us**: 2 million users
- **Flickr**: 1 million photos uploaded daily
- **Yahoo! Answers**: 90M unique users, 250M answers
- **Messenger**: 85M unique users

(2007 data)
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!)

Metadata: Microformats
Microformats

- microformats.org
- Originated by Tantek Celik and others
- Agreements on the way to encode certain kinds metadata in HTML
  - Reuse of semantic-bearing HTML elements
  - Based on existing standards
  - Community process
  - Persons, events, listings etc. but also syntactic metadata: licenses, tags
- Microformats have no shared syntax
  - Each microformat has a separate syntax tailored to the vocabulary
- Microformats are not ontologies
  - No formal descriptions of schema, only text
  - Limited reuse, extensibility of schemas
  - No datatypes
- No namespaces, unique identifiers (URIs)
  - no interlinking
  - mapping between instances is required
- Relationship to page context is unclear
- Widely used in millions of documents
  - User-generated as well as automatically generated

Example: tags and machine tags

James Morrison - (El Rey Theatre, 14th June 2007)

Tags:
- James Morrison
- Live music
- concert
- Los Angeles
- The El Rey
- Upcoming event=183351

Additional Information:
- All rights reserved
- Taken at James Morrison, Charlotte Sometimes
- Taken in Los Angeles, California (impr)
- Taken on June 14, 2007
- Viewed once
Metadata is out there

• Question:
  – Just how much data is out there?
  – What is the quality?

• Idea: bring metadata to the surface of search

• How does it work?
  – User enters query
  – Metadata is extracted dynamically
  – Entity reconciliation
  – Metadata is used to display
    • rich abstracts,
    • related pages
    • spatial, temporal visualization

• Microsearch prototype
  • Play at http://www.yr-bcn.es/demos/microsearch/

Example: ivan herman

- Rich abstract
- Related pages based on metadata
- Geolocation
- Events from personal calendar, Conferences, and bio from LinkedIn
Example: peter site:flickr.com

Flickr users named “Peter” by geography

Example: san francisco conference

Conferences in San Francisco by date
Lessons

- More metadata than we expected
  - 53% of unique queries have at least one metadata-enabled page in top 10 (n=7848)
- Performance is poor
  - Metadata needs to come from the index for performance
- Metacrap does exist
  - Users have to see metadata to spot mistakes in their markup, warn others
- RDF templating is hard
  - Adds extra complexity
- Scalability
Exploiting Metadata: SearchMonkey

- Creating an ecosystem of publishers, developers and end-users
  - Motivating and helping publishers to implement semantic annotation
  - Providing tools for developers to create compelling applications
  - Focusing on end-user experience
- Rich abstracts as a first application
- Addressing the long tail of query and content production
- Standard Semantic Web technology
  - dataRSS = Atom + RDFa
  - Industry standard vocabularies
What is SearchMonkey?

an open platform for using structured data to build more useful and relevant search results

**Before**

- **Topics for Getting Pregnant - BabyCenter**
  Find out how to boost your chances of getting pregnant, what you can do if you're having a problem conceiving, and more... do before you try to get pregnant...
  www.babycenter.com/getting-pregnant - 91k - Cached

- **WebMD Allergies Health Center - Find allergy information and latest...**
  Information and articles on the diagnosis, symptoms, treatment, and the prevention of allergies.
  www.webmd.com/diseases_and_conditions/allergies.htm - 130k - Cached

- **Italy Travel Guide and Travel Information - LonelyPlanet**
  Lonely Planet Italy includes information on events, attractions, activities, and transportation for the independent traveller.
  www.lonelyplanet.com/travelguides/destinations/europe/italy - 67k - Cached

**After**

- **Topics for Getting Pregnant - BabyCenter**
  BabyCenter's getting pregnant tips and information can boost your chances of conception by helping you chart your cycle, read your chemical messengers, and pinpoint ovulation.
  www.babycenter.com/getting-pregnant - 79k - Cached

- **WebMD Allergies Health Center - Find allergy information and latest...**
  Allergies are an abnormal response of the immune system. People who have allergies have an immune system that reacts to usually harmless substances in the environment.
  www.webmd.com/allergies - 111k - Cached

- **Italy Travel Guide, Overview - Lonely Planet WorldGuide**
  When to go: Italy is at its best in spring (April-May) and autumn (September-October). During these seasons, the scenery is beautiful, the temperatures are pleasant, and air pollution is low.
  www.lonelyplanet.com/travelguides/destinations/europe/italy - 44k - Cached

**Enhanced Result**

- deep links
- image
- name/value pairs or abstract
UGC: Exploiting Flickr Tags

Dr. Seuss's beloved story about an elephant that discovers a tiny society existing on a spot of dust springs to life in this vibrant animated adaptation from Academy Award-winning animator Chuck Jones. Agreeing to protect Who-ville from harm, Horton the elephant inspires the Whos to make their presence known among all the inhabitants of the jungle and champion the idea that "a person is a person, no matter how small.

Enjoyed By Members Who Enjoyed

- The Muppet Show: Season 2 (4-Disc Series)
- 101 Dalmatians (Animated)
- It's the Great Pumpkin, Charlie Brown
Tag Mining

- **Objective:**
  - Deploy collective knowledge that exists within Social Media services (Flickr and Delicious)

- **Approach:**
  1. Use tag co-occurrence statistics for media annotation and retrieval
  2. Semantic analysis of large tag-spaces

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
Tag Mining - Semantics

• Assign tag semantics using WordNet broad categories
  – Paris :: location
  – Eiffel Tower :: artifact
  – Coverage: 52% of tag volume

Tag Mining - Semantics

• Extend this mapping using patterns found in Wikipedia
  – Upperbound 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
Understanding tags

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

Semantic Photo Search

Query ➔ flickr ➔ tags ➔ tag graph ➔ tag:type ➔ unsorted photos ➔...
Understanding Text

Document Understanding Cartoon

Complexity of Document Understanding

grep

search engines

semantic web?

Q & A

domain expert

our work!
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.
Example: Picasso

emergency landing

Events related to "emergency landing"

Events in the timeline

1943
- (From [WRAF_Woodbridge]) "The airfield was constructed as an Emergency Landing Ground and was operational from 1943."
- (From [W_Fort_Hartz]) "In 1943 and 1944 the primary purpose of Fort Hartz was to gather intelligence and to cover an airstrip which served as an emergency landing ground for planes flying from India to China over the eastern end of the Himalayan Mountains."
- (From [WRAF_Foulsham]) "Many aircraft made emergency landings at Foulsham, including USAAF B-17 F "Ruthie II", which made an emergency landing there in 1943 after an epic return flight for which co-pilot John C. Morgan was awarded the highest U.S. medal, the Medal of Honor."

June 2, 1983
- (From [Cincinnati/Northwest_Kentucky_International_Airport]) "On June 2, 1983, Air Canada Flight 797, DC-9 flying on Houston-Dallas-Toronto route, made an emergency landing at Cincinnati due to cabin fire."
- (From [W_Air_Canada]) "June 2, 1983: Flight 797 McDonnell Douglas DC-9 had an electrical fire in the aft lavatory during flight, resulting in an emergency landing at Cincinnati/Northwest Kentucky International Airport."
- (From [W_Dallas-Fort_Worth_International_Airport]) "On June 2, 1983, Air Canada Flight 797, which was flying on Houston-Dallas/Fort Worth route, made an emergency landing at DFW due to a small fire in the engine."

2000
- (From [W_RA] "Emergency landing"
- (From [W_Air_Canada]) "June 2, 2000: Air Canada Flight 797, which was flying on the Houston-Dallas route, made an emergency landing at DFW due to a fire in the cabin."
- (From [W_Dallas-Fort_Worth_International_Airport]) "On June 2, 2000, Air Canada Flight 797, which was flying on the Houston-Dallas route, made an emergency landing at DFW due to a fire in the cabin."

2001
- (From [W_RA] "Emergency landing"
Synthetic Document

Query

{ Syntactic matches: extract snippets }

Other: group by relevant categories

Synthetic Document

Climbing

Climbing is the activity of using one's hands and feet to move up the surface of a steep object. It is pursued both recreationally, either to get to a destination otherwise inaccessible or for its own enjoyment, and also professionally, as part of activities such as maintenance of a structure, or military operations.

Rock climbing, the scaling of steep rocky surfaces, is perhaps the most familiar sort of climbing; other types of climbing include ice climbing, tree climbing, buildings (climbing on the outside of buildings), and pole climbing.

Mountaineering, the general activity of ascending mountains, often requires the use of climbing techniques.

Climbing may be divided into two broad categories: aid climbing and free climbing.

Climbing communities in many countries, as well as individual regions, have developed their own climbing rating systems. Ratings, Bouldering, Surfing, Free climbing, and Aid climbing all share these factors to one degree or another.

more about Climbing ...

Climbing areas

Rock climbing in the Head District: Generally the climbing style is free climbing (as opposed to aid climbing) and the rock is either granite or limestone. ... There is a long-standing tradition of climbing routes in the traditional climbing style at this site.

Bluffs is slowly gaining recognition as a destination climbing area due to its mild weather, easy access and high number of rock climbing routes.

For Head: The Del Rios Rock Climbing Club maintains a climbing pad in the area.

Climbing clubs are listed in the Directory of Climbing clubs.

Golden Horsehead geography

Climbing techniques

Lead climbing: Lead climbing is a method of climbing where the leader climbs a route, and the follower (or 'follower') follows behind. The leader is responsible for setting the route and ensuring the safety of the climbers.

Climbing commands: A climbing command is a short phrase used by a leader to give instructions to the follower(s). The commands are used to ensure the smooth operation of the climbing system.

Auto-belay: A device that involves using a fixed point around the climbing rope, and may be combined with other climbing equipment.
Related Entities Search

Sentences

Query

Sentence
Sentence
Sentence
Sentence
Sentence

Entity graph

Possibly filtered by type

Top entities

Entity
Sentence
Entity
Sentence
Entity
Sentence

1 relevant sentence per entity

Locations related to 'emergency landing'

Town
Okinawa
Naha
Ragigad
Goose Bay
Los Angeles

Country
Cuba

Island
Naruto Island

- From W/Harta International Airport - The airliner was able to make an emergency landing in Okinawa.
- From W/Philippine Airlines Flight 434 - The Boeing 747-223B, tail number ELBWF, made an emergency landing in Naha Airport, Okinawa, one hour after the bomb exploded.
- From W/Oyon Island - The Boeing 747-200 safely made an emergency landing in Naha, Okinawa.
Syntactic/Semantic Tagging

- Goals: Identify multiword expressions and entities, support generalization, coarse disambiguation
- Tagger: Average perceptron HMM (Collins, 2002) general purpose tagger: efficient (millions of features, hundreds of classes), fast (thousands of sentences/sec)
- State of the art: 3rd NIST Automatic Content Extraction Evaluation (ACE) 2007 (Surdeanu & Ciaramita, 2007)
- Tasks: PoS Tagging, supersense tagging (Ciaramita & Altun, 2006), named entity detection (CoNLL, BBN-WSJ, ACE, etc.)
- Research problems: robustness on Web data (domain adaptation), learning/evaluating from user-generated data (Mika et al., forthcoming)

Parsing/SRL

- Goal: extract structured information at sentence level (beyond the bag of words/document-centric models)
- Dependency parsing:
  - Parser: Fast discriminative multilingual Shift/Reduce parsing (hundreds of sentences/sec) (Attardi, 2006); 2nd in Adaptation Task of CoNLL 2007 (Attardi et al., 2007)
- Semantic role labeling:
  - Joint parsing and SRL: 3rd best system at CoNLL 2008 (Forthcoming)
- Research problems: how can structured linguistic representations be used to improve search/ranking problems? (Surdeanu, Ciaramita & Zaragoza, ACL 2008)
Web Usage: Extracting Semantics from Queries

Relating Queries (Baeza-Yates, 2007)
## Qualitative Analysis

<table>
<thead>
<tr>
<th></th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td><strong>Session</strong></td>
<td>Medium</td>
<td>High</td>
<td>Physical</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sessions</td>
</tr>
<tr>
<td><strong>Click</strong></td>
<td>High</td>
<td>Medium</td>
<td><strong>Multitopic pages</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Click spam</strong></td>
</tr>
<tr>
<td><strong>Link</strong></td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td><strong>Term</strong></td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>

### Click Graph

![Click Graph Image]
Implicit Folksonomy?

Set Relations and Graph Mining

- Identical sets: **synonyms**
- Subsets: **specificity**
  - directed edges
- Non empty intersections (with threshold)
  - degree of relation
- Dual graph: URLs related by queries
  - High degree: multi-topical URLs

Baeza-Yates & Tiberi
ACM KDD 2007
Implicit Knowledge? Webslang!

Evaluation: ODP Similarity

- A simple measure of similarity among queries using ODP categories
  - Define the similarity between two categories as the length of the longest shared path over the length of the longest path
  - Let $c_1,.., c_k$ and $c'_1,.., c'_k$ be the top $k$ categories for two queries. Define the similarity (@$k$) between the two queries as $\max\{ \text{sim}(c_i,c'_j) \mid i,j=1,..,K \}$
ODP Similarity

- Suppose you submit the queries “Spain” and “Barcelona” to ODP.

- The first category matches you get are:
  
  – Regional/ Europe/ Spain
  
  – Regional/ Europe/ Spain/ Autonomous Communities/ Catalonia/ Barcelona

- Similarity @1 is 1/2 because the longest shared path is “Regional/ Europe/ Spain” and the length of the longest is 6

Experimental Evaluation

- We evaluated a sample of 1,000 thousand edges for each kind of relation

- We also evaluated a sample of 1,000 random pairs of not adjacent queries (baseline)

- We studied the similarity as a function of $k$ (the number of categories used)
Experimental Evaluation

Final Remarks
Open Issues

• Explicit vs. implicit social networks
  – Any fundamental similarities?

• How to evaluate with (small) partial knowledge?
  – Data volume amplifies the problem

• User aggregation vs. personalization
  – Optimize common tasks
  – Move away from privacy issues

The Virtuous Cycle

Explicit
Metadata
RDF
Wikipedia
ODP
Answers
Flickr

Implicit
Text
Anchors + links
Queries + clicks
Wordnet
The Future: Web 3.0?

- We are at Web 2.0 beta
- People wants to get tasks done
  - Where I do go for a original holiday with 1,000 US$?
- Take in account the context of the task

Yahoo! Experience

Second edition coming soon

Thanks to Massi Ciaramita, Peter Mika, Borkur Sigurbjornsson, Mihai Surdeanu, Andrew Tomkins, Roelof van Zwol, Hugo Zaragoza

No Questions?
Contact: rbaeza@acm.org