Trends in Web search
Content growing, changing, diversifying, fragmenting
Semantics of content unlock the value of the data ecosystem
Applications of the semantic sources in the Web
Impact on semantic search
A virtuous feedback cycle for enhancing semantic resources
Final Goal: Getting things done

Trends

• Web search is no more document retrieval
  – Means for web-mediated goals
  – Computation (merely) the basis for rational and behavioral interactions
• Already, more than a list of docs
• Moving towards identifying a user’s task
  – (As expressed in the query box)
• Enabling means for task completion
papa johns - Yahoo! Search Results - Mozilla Firefox

1. Did you mean papa johns?

1. Papa John's
Official site for the pizza delivery and carry-out chain Papa John's. Includes store locator, nutritional info, and franchise and employment opportunities.
www.papajohns.com - 14k - Cached - More from this site

2. Welcome to Papa John's Online Ordering
Order Papa John's Pizza online today. It's easy to register and order online. Welcome to Papa John's Online Ordering. New online customer? Click Here to get ...
www.papajohnsonline.com - 9k - Cached - More from this site

3. Papa John's Menu
Get a glimpse of Papa John's menu for both our restaurants and online ordering. Papa John's does not add monosodium glutamate to any of our pizza toppings...
www.papajohns.com/menu/index.htm - 49k - Cached - More from this site

4. Papa John's Restaurant Locator
Find the neighborhood Papa John's location near you. You can order online or by phone...
www.papajohnsonline.com/restaurantlocator.jsp - 18k - Cached - More from this site

5. Papa John's Pizza UK - America's Number 1
Better Ingredients, Better Pizza - UK Delivery Operator Of The Year 2004 and 2005 ...
www.papajohns.co.uk - 10k - Cached - More from this site

---

papa john's in Saratoga, CA 95070 on Yahoo Local - Mozilla Firefox

1. Papa John's Pizza
(408) 376-7272, 1834 Saratoga Ave, Saratoga, CA 95070, CA
www.amica.com Map | Directions | Send to Phone
near the Burrito Factory about 3 blocks away, their pizza is much better.
Papa... more
See all Carry Out & Take Out - Pizza - Restaurants
pizzabringe.com/

2. Papa John's
(408) 873-0202, 1920 Cambrian Ave, San Jose, CA 95128, CA
www.amica.com Map | Directions | Send to Phone
Papa John is a take out and delivery restaurant with a red and white counter. They...
See all Pizza - Restaurants
www.papajohns.com/

3. Papa John's
(408) 615-7272, 136 Homestead Rd, Santa Clara, CA 95054, CA
www.amica.com Map | Directions | Send to Phone
Papa John is a take out and delivery restaurant with a red and white counter. They...
See all Pizza - Restaurants
www.papajohns.com/
Rich media and search assistance
**Structured aggregation**

**Search: Content vs. Intent**

- **Premise:**
  - People don’t want to search
  - People want to get tasks done

I want to book a vacation in Tuscany.
How this might work – I

Index time processing:

Business name

Lookup

Business type= *Food, Fast food, pizza*

Home page for Pizza Hut London

Address: Tel: 12345678

Geo

Map

Reviews – extracted and indexed under Pizza Hut London

Other pages around the web

How this might work – II

Query time processing:

Query stream

Session analysis

Intent = *buy pizza*

Geo = *London*
Net result

• We move from a web of pages to a web of objects
• Objects are people, places, business, universities …
• Objects have attributes
  – Missing, noisy, etc.
• Intents are satisfied by juxtapositions of objects

Info assembly for user

• No longer a list of 10 documents

<table>
<thead>
<tr>
<th>Pizza Hut Pisa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address:</td>
</tr>
<tr>
<td>Tel: 12345678</td>
</tr>
<tr>
<td>[Click to call]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Blah blah … ★★★</td>
</tr>
<tr>
<td>3. …                           ★★★</td>
</tr>
</tbody>
</table>
Hard problems to be solved

• How do you model intent?
  – What is the right abstraction?
  – What is the right granularity?
  – What are the top intents?
• How do you measure user happiness?
  – Transaction?
  – User studies? Interviews?

Hard problems to be solved

• What is the right data model for the web of objects?
  – What are the kinds of relationships?
  – How do you infer trust/reliability?
  – What logic/algebra governs scoring them?
• What is the automated framework for relevance?
Ergo: Web search is no longer “just” search!

• Search is no longer about finding documents
  – It is an interface for web-mediated goals
• Old challenges
  – Main driver of user perception of search quality was: precision of navigational queries
  – The technical prowess was about crawling and spam
  – The hard core was indexing and retrieval

New challenges

• (Navigational query precision – taken for granted)
• (Spam remains an issue)
• Classic search quality: just the admission ticket
• The battleground: facilitating the user’s goal
• The technical prowess: data extraction & integration, from documents and queries
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

<table>
<thead>
<tr>
<th>1 to 3</th>
<th>0.5</th>
<th>treats, catnap, daddy, mommy, nursing, mice, playing, napping, scratching, milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 to 15</td>
<td>3.5</td>
<td>etwebdesigning, Jeremy Sumper, Chris Wilson, Emma Watson, T. V., Tom Felton, FUSE, Adam Carson, Guzz, Pac Sun, mall, going online</td>
</tr>
<tr>
<td>16 to 18</td>
<td>25.2</td>
<td>1986(7,9), class of 2004(4,5), dream street, drama club, band trips, 16, Brave New Girl, drum major, talking on the phone, high school, JROTC</td>
</tr>
<tr>
<td>18 to 21</td>
<td>32.8</td>
<td>1996(3,5), class of 2003, dorm life, frat parties, college life, my tattoo, pre-med</td>
</tr>
<tr>
<td>22 to 24</td>
<td>18.7</td>
<td>199(1,2), Dumbledore's army, Midori sour, Long Island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra</td>
</tr>
<tr>
<td>25 to 27</td>
<td>8.4</td>
<td>1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio</td>
</tr>
<tr>
<td>28 to 30</td>
<td>4.4</td>
<td>Hal Hartley, soap opera, Camerilla, Amorcord, Taxi, Concrete Blonde, motherhood, SQL, TRON</td>
</tr>
<tr>
<td>31 to 33</td>
<td>2.4</td>
<td>my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, soap opera, the prisoner, good eats, herbalism</td>
</tr>
<tr>
<td>34 to 36</td>
<td>1.5</td>
<td>Cross Stitch, Thelema, Taxi, parenting, cub, role-playing games, bicycling, shamanism, Burning Man</td>
</tr>
<tr>
<td>37 to 45</td>
<td>1.6</td>
<td>SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot</td>
</tr>
<tr>
<td>46 to 57</td>
<td>0.5</td>
<td>science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking</td>
</tr>
<tr>
<td>&gt; 57</td>
<td>0.2</td>
<td>death, cheese, photography, cats, poetry</td>
</tr>
</tbody>
</table>

Content access is fragmenting
Semantic Sources in the Web

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

Examples

<table>
<thead>
<tr>
<th>Explicit</th>
<th>Implicit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td>Text</td>
</tr>
<tr>
<td>RDF</td>
<td>Anchors + links</td>
</tr>
<tr>
<td>Wikipedia ODP</td>
<td>Queries+clicks</td>
</tr>
<tr>
<td>Y! Answers</td>
<td>UGC Private</td>
</tr>
</tbody>
</table>

Wordnet

UGC

Scale
User Generated Content at Yahoo!

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)
• 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

- 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

Metadata is out there: Microsearch (Mika, 2007)

- 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

Example: peter site:flickr.com

Flickr users named “Peter” by geography
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
Adding 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?

Before

1. **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. [Link](http://www.babycenter.com/getting-pregnant)

2. **WebMD Allergies Health Center - Find allergy information and latest...**
   - Information and articles on the diagnosis, symptoms, treatment, and the prevention of allergies. [Link](http://www.webmd.com/allergies)

3. **Italy Travel Guide and Travel Information - Lonely Planet**
   - Lonely Planet Ita...
UGC: Exploiting Flickr Tags
Tag Mining (Sigurbjörnsson and van Zwol, 2007)

- **Objective:**
  - Deploy collective knowledge that exists within Social Media services (Flicker 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

TagExplorer - Approach

• Combines:
  – Tag semantics
    • Dual level
      – Where?, When?, What?
      – Nouns in WordNet broad categories
        » Location, artifact, activities, event, person, group, etc.
    • Other schemas can be applied.
  – Tag co-occurrence analysis
    • For a given set of tags - a keyword based query - a set of related tags is derived.
tagexplorer - approach

locations
australia barcelone berlin bw california canada chicago china city england europa florida france germany hawaii india italy japan london mexico new york nyc paris park san francisco scotland seattle spain taiwan thailand tokyo uk usa vancouver

activities/events
concert festival fun party travel trip wedding

time
2002 2003 2004 2005 2006 2007 april birthday christmas day holiday july june may night spring summer vacation winter

artifacts/objects
architecture art beach cameraphone canon clouds garden home house lake music nature new nikon portrait sea sky street sun set zoo

people/groups
baby family friends kids me people

plants/animals
cat dog flower flowers tree

food/substance
food water
1. **Search assistants:**
   - TagExplorer, Query Ambiguity Detection, Feedback search

2. **In-context exploration of search results**
   - Connect search assistants with search results
   - Provide **instant feedback** to the user, which of the current results correspond to the context provided by the search assistant.
Media Search - Example

Showing results for **washington**, 36 results found. 1.24 seconds.

Tag Explorer  Feedback  Ambiguity

Search assist has detected that you have provided an ambiguous query: **washington**.
You can disambiguate the results for this query by extending your search with **dc** or **seattle**.

Title: Tulips in Washington

Add the term 'dc' to the query

Tags: seattle monument taj mahal miami dc kids tag5
tulips washington monument ian leal wexler wexler

By "Toakho" on Flickr

Understanding Text
**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.
Related Entities Search (Zaragoza et al, 2007)

Example: Picasso
Synthetic Document (Zaragoza et al, 2008)

**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 own enjoyment, and also professionally, as part of activities such as maintenance of a structure, or military operations.

Rock climbing: the activity of moving up rocky surfaces. It is perhaps the most familiar sort of climbing, other types of climbing include ice climbing, tree climbing, bouldering (climbing on the outside of buildings), and via ferrata (climbing).

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

Climbing may be applied into two basic categories: aid climbing and free climbing.

Climbing communities in many countries, as well as individual regions, have developed their own climbing rating systems.

*Linking* (linking) - related to free climbing, and aid climbing will share these factors in one degree or another.

---

**Climbing areas**

Climbing in the Peak District. Usually the climbing style is free climbing (as opposed to aid climbing) and the rock is either gritstone or limestone. There is a long-standing passion of climbing routes in the traditional style.

Stanage Boulders: The area is mostly a sport climbing area, though many traditional climbing opportunities also exist. Stanage Boulders is highly regarded as a destination climbing area due to its mild weather, easy access and high number of sport climbing routes.

Fair Head: The Ballintoy Climbing Club maintains a climbing base in the area. Categories: Northern Ireland geography.

Mount. Horear Conservation Area: Categories: Climbing areas | Ballet Region | Ontario | Parks in Ontario | Climbing routes

---

**Climbing techniques**

Here's a point in climbing, leading. Diagram of aLead climbing mode essentially means "leading the climb after rope has run out." Categories: climbers; climbers in practice, climbers in practice, climbers in practice. Climbing command. A belaying command is a short standard phrase used in climbing to ensure the smooth operation of the climbing system.

Autobelay: It involves using a friction hitch around the climbing rope, and may be combined with other climbing equipment.
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>Graph</th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td>Session</td>
<td>Medium</td>
<td>High</td>
<td>Physical sessions</td>
</tr>
<tr>
<td>Click</td>
<td>High</td>
<td>Medium</td>
<td>Multitopic pages</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Click spam</td>
</tr>
<tr>
<td>Link</td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td>Term</td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>

### Click Graph

[Image of a complex network graph showing connections between various nodes labeled with terms such as "France", "Barcelona", "cheap hotel", and "travel Italy". The graph illustrates the relationships and click patterns among these terms.]
Connected Components

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

Implicit Knowledge? Webslang!

Baeza-Yates & Tiberi
ACM KDD 2007
• 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 \( c_1, \ldots, c_k \) and \( c'_1, \ldots, c'_k \) be the top \( k \) categories for two queries. Define the similarity \( (@k) \) between the two queries as

\[
\max \{ \text{sim}(c_i, c'_j) \mid i, j = 1, \ldots, K \}
\]

• 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).
Open Issues

• Data Volume vs. Better Algorithms

• 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
To take home

• Syntactic models work in many cases (IR: low hanging fruit)
• Ambiguity many times can be solved with one more keyword
• Ambiguous searches that are not satisfied with the most popular case are a small percentage of Web queries
• However semantic search can solve more complex tasks if we predict well the intention of the query
• Solutions are then constrained by:
  – Must improve precision more than recall
  – Must be scalable: focus on efficiency
  – Must work in many languages
• How: Combination of formal sources plus Web data and usage
  – To achieve this we need powerful tools for graph mining
Questions?

Contact: rbaeza@acm.org

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