Improving Search with Semantic Technologies: Current Research Directions

Hugo Zaragoza
Yahoo! Research (Barcelona)
Yahoo! by numbers (April, 2007)

• **500 million users** of Yahoo! branded services **(1/2 users online)**.

• Most visited site online
  – most visits (nearly 4 billion visits)
  – largests online time share: 12% in America and 8% worldwide.
  – #1 Web mail provider in the world (243 million users)
  – #1 online news destination
  – (leads all competitors in audience reach, frequency and engagement)

• Social media properties: **(Flickr, delicious, Answers, 360, Video, MyBlogLog, Jumpcut and Bix)**
  – 115 million unique visitors worldwide.
  – Flickr: more than 450 million photo, 1 million uploaded daily.
  – Nearly 1 in 10 Internet users is a member of a Yahoo! Groups.

(comScore WorldMetrix Feb.07 & Yahoo! Internal Data )

Hugo Zaragoza, ESTC 2008. (2)
The mission of Yahoo! Research is to develop the world-class science that will deliver the next generation of businesses to the company.
Why should you care about search?
Plan

• Search Monkey (today)

• Web Search: where are we going? (tomorrow)

• Current Research Directions (the future…)
**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 ...

**WebMD Allergies Health Center - Find allergy information and latest...**
Information and articles on the diagnosis, symptoms, treatment, and the prevention of allergies.

**Italy Travel Guide and Travel Information - Lonely Planet**
Lonely Planet Italy includes information on events, attractions, activities, and transportation for the independent traveler.

**After**

**Topics for Getting Pregnant - BabyCenter**
BabyCenter's getting pregnant tools and information can boost your chances of conception by helping you chart your cycle, read your cervical mucus, and pinpoint ovulation.

**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. This substance (pollen, mold, dander, etc.) is called an allergen.

**Italy Travel Guide - Overview - Lonely Planet WorldGuide**
Where to go: Italy is at its best in spring (April-May) and autumn (October-November). During these seasons, the scenery is beautiful, the temperatures are pleasant and are relatively few crowds.
Enhanced Result: what might a developer build?

- deep links
- images
- name/value pairs or abstract
How does it work?

1. site owners/publishers share structured data with Yahoo!.
2. site owners & third-party developers build SearchMonkey apps.
3. consumers customize their search experience with Enhanced Results or Infobars.

Process:
- **Acme.com's database** feeds into Yahoo!'s **SEARCH Index**
- **Page Extraction** and **RDF/Microformat Markup** are applied.
- **DataRSS feed** and **Web Services** connect to Yahoo!'s services.
- **SearchMonkey** processes the data, allowing for customized search results.
Infobar: what might a developer build?

Pull in data from any web service
Infobar: what might a developer build?

Open canvas for HTML

<table>
<thead>
<tr>
<th>Cast &amp; Crew: Horton Hears a Who! (2008) - Internet Movie Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jim Carrey</strong></td>
</tr>
<tr>
<td><strong>Steve Carell</strong></td>
</tr>
<tr>
<td><strong>Carol Bumett</strong></td>
</tr>
<tr>
<td><strong>Dan Fogler</strong></td>
</tr>
<tr>
<td><strong>Isla Fisher</strong></td>
</tr>
</tbody>
</table>

Search Monkey | 10 | http://developer.yahoo.com/searchmonkey
Infobar: what might a developer build?

Yahoo! Search

Dr. Seuss Horton Hears a Who Movie

Dr. Seuss’ Horton Hears A Who
www.hortonmovie.com/splash.html - Cached


Horton Hears a Who! (2008) - Netflix

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-Part Series)
101 Dalmatians (Animated)
It's the Great Pumpkin, Charlie Brown
Everything SearchMonkey:
developer.yahoo.com/searchmonkey

mailing lists
- searchmonkey-developers@yahoogroups.com
- searchmonkey-siteowners@yahoogroups.com

forums
- http://suggestions.yahoo.com/searchmonkey
- http://suggestions.yahoo.com/datarss

data/partners

...and all sites with supported microformats, or eRDF markup!
Plan

• Search Monkey (today)

• Web Search: where are we going? (tomorrow)

• Current Research Directions (the future…)}
1. **Papa John's** (Nasdaq: PZZA)
   Official site for the pizza delivery and carry-out chain Papa John's. Includes store locator, nutritional info, and franchise and employment opportunities.
   [Papa John's near you](http://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 ... 
   [Papa Johnsonline.com](http://www.papajohnsonline.com) - 9k - Cached - More from this site

3. **Papa John's Menu**
   Get a glimpse of Papa John's basic 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](http://www.papajohns.com/menu/index.htm) - 49k - Cached - More from this site

4. **Papa John's Restaurant Locator**
   Find the neighborhood Papa Johns location in your area. You can order online or by phone. ... 
   ©Copyright 2007 Papa John's International Inc. All Rights Reserved. 
   [www.papajohnsonline.com/restlocator.jsp](http://www.papajohnsonline.com/restlocator.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 ... Privacy Statement | Terms and Condition | [Papa John's (GB) Ltd 2006](http://www.papajohns.co.uk) ... 
   [www.papajohns.co.uk](http://www.papajohns.co.uk) - 10k - Cached - More from this site

**Did you mean:** **papa john's**
Flights from San Francisco, CA to Boston, MA

cheapTickets - Expedia - Hotwire - Orbitz - Priceline - Travelocity

FoRK Archive: Re: Oh, the trials of LaMacchia
UA 04-APR-97 0033 BOS SFO 2704 2704 5488 UA 01-APR-97 0160 LAX BOS 2611 2611
5222 UA 28-MAR-97 0159 BOS LAX 2611 2611 5222 UA 25-MAR-97 0160 LAX BOS 2611 ...
www.xent.com/FoRK-archive/may97/0739.html - 10k - Cached - Similar pages - Note this

San Francisco (SFO) to Boston (BOS) departing Apr 2007
San Francisco (SFO) to Boston (BOS) departing Apr 2007 Get this by email Get this ... Top
25 Airport Idea. add another. World Cities, United States, Europe ...
www pinpointtravel.com/h/buzz/view?dt=A&code=SFO&dest=BOS&tm=200704 - 66k -
Cached - Similar pages - Note this

ESPN.com - MLB - Box Score - Twins at Royals
MLB Scoreboard - April 25, 2006 - American League - BOS, 8, F ... Batters faced: S Elarton
25; L Hudson 3; A Sisco 3; A Burgos 4 ...
sports.espn.go.com/mlb/boxscore?gameId=260425107 - 130k -
Cached - Similar pages - Note this

ESPN.com - MLB - Box Score - Devil Rays at Yankees
MLB Scoreboard - April 25, 2006 - American League - BOS, 8, F. CLE, 6. TAM, 1. NYY, 9,
F. BAL, 3 ... STL, 6, F. ARI, 7, F. SDG, 0. NYM, 4, F. SFO, 1 ...
Beyond traditional “Web Search”

• We move from a web of pages to a web of objects

• Objects are people, places, business, flights… and their relationships.

• User Intents are satisfied by juxtapositions of objects
Beyond traditional “Web Search”

• Search is no longer about finding documents
  – It is an interface for web-mediated goals

• Old battles
  – Precision of navigational queries (solved)
  – Document crawling and spam (!)
  – Indexing and retrieval (!)
  – Result relevance (!)
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 is the automated framework for relevance?

• What are the ranking models that can attain it?
Plan

• Search Monkey  
  (today)

• Web Search: where are we going?  
  (tomorrow)

• Current Research Directions  
  (the future…)
  
  – MicroSearch
  – Learning tags
  – Searching Objects
Micro Search, example: ivan herman

Related pages based on metadata

Conferences he plans to attend and other events from homepage plus bio events from LinkedIn

Geolocation

query: ivan herman

Peter Mika. http://www.yr-bcn.es/demos/microsearch
MicroSearch, example: san francisco conference
MicroSearch, example:
greater st. peter
MicroSearch, towards intent: Paris Hilton is a person!
MicroSearch, towards intent: hugo zaragoza

Hugo Zaragoza, ESTC 2008. (26)
Plan

• Search Monkey

• Web Search: where are we going?

• Current Research Directions
  – MicroSearch
  – Learning to tag, tagging to learn.
  – Ranking Objects
Example

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 Clítu Ruiz y Picasso. [1] 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 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. [2] 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, [3] his first word was "piz," a shortening of lápiz, the Spanish word for pencil [4] 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.
Statistical Taggers @ Y!Research Barcelona

• Name Entity Recogniser
  – SuperSense Tagger.

  Massimiliano Ciaramita

• Dependency Parsing
  – 2nd best at CONL 2007

  Giuseppe Attardi (U. Pisa)

• Semantic Role Labeling
  – 2nd best at CONL 2008

  Mihai Surdeanu.
Extending annotations

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.

[Mika, Zaragoza, Ciaramita, Atserias. IEEE AI 2008]
Hugo Zaragoza, ESTC 2008. (30)
Adaptation and improved tagging

- Generate training data based on what is learned
  - i.e. if Paris is also a place of birth, it is also a place

- Dealing with sparse data
  Paris is the capital city of France.

- Improvement
  - On news: +6.1%
  - On Wikipedia: +5.5%
## (property typing examples)

<table>
<thead>
<tr>
<th>Wikipedia property</th>
<th>CONLL</th>
<th>Examples of property values</th>
</tr>
</thead>
<tbody>
<tr>
<td>infoboxNbaPlayer_name</td>
<td>PER</td>
<td>Alex Groza, Elgin Baylor, Jerry West, David Thompson, Glen Rice, Christian Laettner, Richard Hamilton, Juan Dixon, Sean May, Joakim Noah, ...</td>
</tr>
<tr>
<td>infoboxSerialKiller_alias</td>
<td>ORG</td>
<td>Gray Man, the Werewolf of Wysteria, Brooklyn Vampire, Sister, Brian Stewart, Bloody Benders, The Prostishooter, The Rest Stop Killer, The Truck Stop Killer, The Sunset Strip Killer, Cincinnati Strangler, Son of Sam, Plainfield Ghoul, Ed &quot;Psycho&quot; Gein, The Co-ed Killer, ...</td>
</tr>
<tr>
<td>highlanderCharacter_born</td>
<td>LOC</td>
<td>Unknown, unknown, 1659, 1945, 802, 1887, 1950, , Glenfinnan, Scotland, (original birth date unknown)(&quot;Highlander II&quot;), 896 BC, Ancient Egypt (original birth date unknown) (&quot;Highlander II&quot;), California, ...</td>
</tr>
<tr>
<td>infoboxSuperbowl_stadium</td>
<td>LOC</td>
<td>Sun Devil Stadium, Georgia Dome, Miami Orange Bowl, Hubert H. Humphrey Metrodome, Dolphin Stadium, Raymond James Stadium, Louisiana Superdome, Joe Robbie Stadium, Ford Field, Los Angeles Memorial Coliseum, ...</td>
</tr>
<tr>
<td>infoboxWeapon_usedBy</td>
<td>LOC</td>
<td>USA, None, One, none, Italy, United States, Mexico, UK, Russia, Under development, ...</td>
</tr>
<tr>
<td>minorLeagueTeam_league</td>
<td>MISC</td>
<td>Eastern League (1923-37, 1940-63, 1967-68, 1992- ), Pacific Coast League, Arizona League, Texas League, South Atlantic League, California League, Midwest League, Northwest League, International League, Carolina League, ...</td>
</tr>
<tr>
<td>infoboxTea_teaOrigin</td>
<td>LOC</td>
<td>Nuwara Eliya, Sri Lanka near Adam's Peak between 2200 - 2500 metres, Japan, India, Vietnam, Taiwan, Turkey, China, Anhui, Guangdong, Jiangxi, ...</td>
</tr>
<tr>
<td>infoboxPrimeMinister_name</td>
<td>PER</td>
<td>Abdallah El-Yafi, Umar al-Muntasir, Dr. Abdellatif Filali, Abderrahmane Youssoufi, Abdessalam Jalloud, Abdul Ati al-Obeidi, Abdul Hamid al-Bakkoush, Abdul Majid Kubar, Abdul Majid al-Qaud, Abd al-Qadir al-Badri, ...</td>
</tr>
</tbody>
</table>
Beyond annotations, into text

“In a 1945 interview with Jerome Seckler, Picasso declared: ‘I am a communist …’
Plan

• Search Monkey (today)

• Web Search: where are we going? (tomorrow)

• Current Research Directions (the future…)
  – MicroSearch
  – Learning to tag, tagging to learn.
  – Ranking Objects
Why should you care about search?
Why is it hard?

- Information needs are intrinsically vague & ambiguous.
- People like to search in natural language.
- Search (Information Retrieval) technology greatly surpassed Boolean querying in the 80s.
- Attempts to improve search technology with semantic knowledge have repeatedly failed since.
  - (except in very narrow domains).
- Effective query expansion is very difficult.
But new directions arise…

- Entity Ranking  (sort entities by relevance to a query)
  - Expert Finding
  - Temporal Search
  - Geo-search, …

- Example:
  - “Effect of sub-prime crisis in Italy”,
    “Pablo Picasso and the Second World War”
    - Find relevant dates
    - Find relevant people & organisations
    - Find relevant geo-locations
Colored (Typed) Indices

Doc #3: The last time Peter exercised was in the XXth century.

Doc #5: Hope claims that in 1994 she run to Peter Town.

Possible Queries:

“Peter ^ run”
“Peter ^ WNS:N_DATE”
“(WSJ:CITY:*) ^ run”
“(WSJ:PERSON:Hope) ^ run”

Bracketing can also be dealt with
Entity containment graphs

Doc #3: The last time Peter exercised was in the XXth century.

Doc #5: Hope claims that in 1994 she run to Peter Town.

[Zaragoza et. al. CIKM’08]
Entity Containment Graph

Pablo Picasso and the Second World War

Search Engine

Sentences

Sentence to Entity Map
(Correlator demo)
Conclusion

• Web Search is moving changing from Query→Page to: Information Need → Information Objects

• A virtuous cycle needs to emerge between Web Search and Semantic Web technologies.

• Difficult scientific problems remain to be solved before we can effectively search Resources:
  – Level of Representation
  – Automatic Extraction and Adaptation
  – Ranking
  – User Interaction