Making things findable

Peter Mika
Researcher and Data Architect
Yahoo! Inc.
This was planned to be a presentation (only) about Semantic Search…

But we are celebrating!
Anno 2001: The birth

- Scientific American article
- Semantic Web Working Symposium at Stanford
- Semantic Web standardization begins
- EU funding! DAML funding!
- The Semantic Web starts a career… and so am I…
Anno 2004-2006: Reality sets in

• Agents? Semantic Web Services? Common sense?
• Engineers are not logicians
• Humans will have to do (most of) the work
• Two separate visions
• No one actually invests… other than the EU
• Where are the ontologies?
• Bad reputation for SW research
• The Semantic Web is looking for a job.

and so am I 😊
Anno 2007: A second chance

- Data first, schema second…logic third
- Linked Data…. billions of triples
- A sense of community
- Some ontologies
- We get the standards we need
- Startups, tool vendors appear
- SemTech
- The Semantic Web is in business…

*and so am I!*
Meanwhile, in Search…
Search is really fast, without necessarily being intelligent
Why Semantic Search? Part I

- Improvements in IR are harder and harder to come by
  - Machine learning using hundreds of features
    - Text-based features for matching
    - Graph-based features provide authority
  - Heavy investment in computational power, e.g. real-time indexing and instant search
- Remaining challenges are not computational, but in modeling user cognition
  - Need a deeper understanding of the query, the content and/or the world at large
  - *Could Watson explain why the answer is Toronto?*
Poorly solved information needs

- Multiple interpretations
  - paris hilton
- Long tail queries
  - george bush (and I mean the beer brewer in Arizona)
- Multimedia search
  - paris hilton sexy
- Imprecise or overly precise searches
  - jim hendler
  - pictures of strong adventures people
- Searches for descriptions
  - countries in africa
  - 32 year old computer scientist living in barcelona
  - reliable digital camera under 300 dollars
Poorly solved information needs

- Multiple interpretations
  - paris hilton
- Long tail queries
  - george bush (and I mean the beer brewer in Arizona)
- Multimedia search
  - paris hilton sexy
- Imprecise or overly precise searches
  - jim hendler
  - pictures of strong adventures people
- Searches for descriptions
  - countries in africa
  - 32 year old computer scientist living in barcelona
  - reliable digital camera under 300 dollars

Many of these queries would not be asked by users, who learned over time what search technology can and cannot do.
Dealing with sparse collections
Dealing with sparse collections

Note: don’t solve the sparsity problem where it doesn’t exist.
Contextual Search: content-based recommendations

Hovering over an underlined phrase triggers a search for related news items.
Contextual Search: personalization

Machine Learning based ‘search’ algorithm selects the main story and the three alternate stories based on the users demographics (age, gender etc.) and previous behavior.

Display advertizing is a similar top-1 search problem on the collection of advertisements.
Contextual Search: new devices
Contextual Search: new devices

Yahoo! Connected TV

Who will win the WBO Featherweight Title tonight?

JuanMa" Lopez 59%
Rafael Marquez 39%
Draw

Show related content

Connect to friends watching the same
Aggregation across different dimensions

The best news, deals and things to do around you

Love Lane Mews: Gordon Gekko Meets Mary Poppins
4 hrs ago
The Wall Street Journal recently caught up with Love Lane Mews architect Stephen B. Jacobs. The takeaway — the new condos there will be "supercalafragalisticexpialadoush" meets "greed is good": WSJ: But when it comes to preservation, the...

Dumbo Links Week of 21Nov2010
18 hrs ago
The following are selected links from this past week on blogs and websites with discussion about Dumbo (and its neighboring areas): More Apple Talk in Dumbo?, 24Nov10, Brownstoner 30 Washington St Two Fl Extension?, 22Nov10, Brownstoner Toll Bros...

Posts From Community Writers

Charles Manley wrote an article: Aeropostale Opens on Newly Paved Fulton Street in Downtown Brooklyn

Lucrezia Wise wrote an article: Sing Along at the Jingle Bell Jamboree!
Aggregation across different dimensions

Hyperlocal: showing content from across Yahoo that is relevant to a particular neighbourhood.
Why Semantic Search? Part II

- The Semantic Web is now a reality
  - Billions of triples
  - Thousands of schemas
  - Varying quality

- End users
  - Keyword queries, not SPARQL

- Searching data instead or in addition to searching documents
  - Providing direct answers (possibly with explanations)
  - Support for novel search tasks
Direct answers in search

Search Pad

SearchScan - On

SearchScan - On

SearchPad

17,400,000 results for
vyenna, austria

Related Points of In

Maria am Ge

Kunsthistorifs

Hunderwas

W

CNET Reviews

Hawolka

Wikipedia

Shopping Sites

Spanish Ridi

Sponsored Results

Alsergrund

related Searches

blu ray software

best buy

playstation 3

Suggested search results are “Powered by Bing” since Aug, 2010.
Novel search tasks

• Aggregation of search results
  – e.g. price comparison across websites
• Analysis and prediction
  – e.g. world temperature by 2020
• Semantic profiling
  – recommendations based on particular interests
• Semantic log analysis
  – understanding user behavior in terms of objects
• Support for complex tasks (search apps)
  – e.g. booking a vacation using a combination of services
Why Semantic Search? Part III

• There is a model
  – Publishers are (increasingly) interested in making their content searchable, linkable and easier to aggregate and reuse
  – So that social media sites and search engines can expose their content to the right users, in a rich and attractive form

• This is about creating an ecosystem…
  – More advanced in some domains
  – In others, we still live in the tyranny of rate-limited APIs, licenses etc.
Example: rNews

- RDFa vocabulary for news articles
  - Easier to implement than NewsML
  - Easier to consume for news search and other readers, aggregators

- Under development at the IPTC
  - March: v0.1 approved
  - Final version by Sept
Example: Facebook’s Like and the Open Graph Protocol

• The ‘Like’ button provides publishers with a way to promote their content on Facebook and build communities
  – Shows up in profiles and news feed
  – Site owners can later reach users who have liked an object
  – Facebook Graph API allows 3rd party developers to access the data

• Open Graph Protocol is an RDFa-based format that allows to describe the object that the user ‘Likes’
Example: Facebook’s Open Graph Protocol

- RDF vocabulary to be used in conjunction with RDFa
  - Simplify the work of developers by restricting the freedom in RDFa
- Activities, Businesses, Groups, Organizations, People, Places, Products and Entertainment
- Only HTML <head> accepted
- [http://opengraphprotocol.org/](http://opengraphprotocol.org/)

```html
<html xmlns:og="http://opengraphprotocol.org/schema/">
<head>
  <title>The Rock (1996)</title>
  <meta property="og:title" content="The Rock" />
  <meta property="og:type" content="movie" />
  <meta property="og:url" content="http://www.imdb.com/title/tt0117500/" />
  <meta property="og:image" content="http://ia.media-imdb.com/images/rock.jpg" />
  …
</head>
```

Semantic Search
Semantic Search: a definition

• Semantic search is a retrieval paradigm that
  – Makes use of the structure of the data or explicit schemas to understand user intent and the meaning of content
  – Exploits this understanding at some part of the search process

• Combination of document and data retrieval
  – Documents with metadata
    • Metadata may be embedded inside the document
    • *I’m looking for documents that mention countries in Africa.*
  – Data retrieval
    • Structured data, but searchable text fields
    • *I’m looking for directors, who have directed movies where the synopsis mentions dinosaurs.*
Semantics at every step of the IR process

The IR engine

Query interpretation

q="bla" * 3

θ(q,d)

Indexing

Ranking

The Web

Document processing

Result presentation

What a huge mess!
Data on the Web
Data on the Web

• Data on the Web as a complement to professional data providers

• Most web pages are generated from databases, but the data in not always directly accessible
  – APIs offer limited views over data

• The structure and semantics (meaning) of the data is not directly accessible to search engines

• Two solutions
  – Extraction using Information Extraction (IE) techniques (implicit metadata)
  – Relying on publishers to expose structured data using standard Semantic Web formats (explicit metadata)
Information Extraction methods

• Natural Language Processing
• Extraction of triples
• Filling web forms automatically (form-filling)
• Extraction from HTML tables
• Wrapper induction
  – Kushmerick et al. Wrapper Induction for Information Extraction. IJCAI 2007
Semantic Web

• Sharing data across the Web
  – Publish information in standard formats (RDF, RDFa)
  – Share the meaning using powerful, logic-based languages (OWL, RIF)
  – Query using standard languages and protocols (HTTP, SPARQL)

• Two main forms of publishing
  – Linked Data
    • Data published as RDF documents linked to other RDF documents and/or using SPARQL end-points
    • Community effort to re-publish large public datasets (e.g. Dbpedia, open government data)
  – RDFa
    • Data embedded inside HTML pages
    • Recommended for site owners by Yahoo, Google, Facebook
RDFa: metadata embedded in HTML

Roi's homepage

...<p type="foaf:Person" about="http://example.org/roi">
<span property="foaf:name">Roi Blanco</span>.
<a rel="owl:sameAs" href="http://research.yahoo.com/roi"> Roi Blanco </a>.
You can contact him at <a rel="foaf:mbox" href="mailto:roi@yahoo-inc.com"> via email </a>.
</p>
...
Crawling the Semantic Web

• Linked Data
  – Similar to HTML crawling, but the crawler needs to parse RDF/XML (and others) to extract URIs to be crawled
  – [Semantic Sitemap/VOID](#) descriptions

• RDFa
  – Same as HTML crawling, but data is extracted after crawling
  – [Mika et al. Investigating the Semantic Gap through Query Log Analysis, ISWC 2010.](#)

• SPARQL endpoints
  – Endpoints are not linked, need to be discovered by other means
  – [Semantic Sitemap/VOID](#) descriptions
Data fusion

• Ontology matching
  – Widely studied in Semantic Web research, see e.g. list of publications at ontologymatching.org
    • Unfortunately, not much of it is applicable in a Web context due to the quality of ontologies

• Entity resolution
  – Logic-based approaches in the Semantic Web
  – Studied as record linkage in the database literature
    • Machine learning based approaches, focusing on attributes
  – Graph-based approaches, see e.g. the work of Lisa Getoor are applicable to RDF data
    • Improvements over only attribute based matching

• Blending
  – Merging objects that represent the same real world entity and reconciling information from multiple sources
Data quality assessment and curation

• Heterogeneity, quality of data is an even larger issue
  – Quality ranges from well-curated data sets (e.g. Freebase) to microformats
    • In the worst of cases, the data becomes a graph of words
  – Short amounts of text: prone to mistakes in data entry or extraction
    • Example: mistake in a phone number or state code

• Quality assessment and data curation
  – Quality varies from data created by experts to user-generated content
  – Automated data validation
    • Against known-good data or using triangulation
    • Validation against the ontology or using probabilistic models
  – Data validation by trained professionals or crowdsourcing
    • Sampling data for evaluation
  – Curation based on user feedback
Query Interpretation
Query interpretation in search

• Provide a higher level representation of queries in some conceptual space
  – Ideally, the same space in which documents are represented
  – Limited user involvement in the case document retrieval
    • Examples: search assist, facets
• Interpretation happens before the query is executed
  – Federation: determine where to send the query
    • Example: show business listings from Yahoo! Local for local queries
  – Ranking feature
    • Blend multiple possible interpretations of the same query
  – Deals with the sparseness of query streams
    • 88% of unique queries are singleton queries (Baeza et al)
    • Spell correction (“we have also included results for…”), stemming
Query interpretation in Semantic Search

• “Snap to grid”
  – Using the ontologies, a summary of the data or the whole data we find the most likely structured query matching the user input
    • Example: “starbucks nyc” -> company name:”Starbucks” in location:”New York City”

• Larger user involvement
  – Guiding the user in constructing the query
    • Example: Freebase Suggest
  – Displaying back the interpretation of the query
    • Example: TrueKnowledge
Indexing and Ranking
Indexing

- Search requires matching and ranking
  - Matching selects a subset of the elements to be scored
- The goal of indexing is to speed up matching
  - Retrieval needs to be performed in milliseconds
  - Without an index, retrieval would require streaming through the collection
- The type of index depends on the query language to support
  - SPARQL is a highly-expressive SQL-like query language for experts
    - DB-style indexing
  - End-users are accustomed to keyword queries with very limited structure (see Pound et al. WWW2010)
    - IR-style indexing
IR-style indexing

- Index data as text
  - Create virtual documents from data
  - One virtual document per subgraph, resource or triple
    • typically: resource

- Key differences to Text Retrieval
  - RDF data is structured
  - Minimally, queries on property values are required

- MapReduce is an ideal model for building inverted indices
  - Map creates (term, \{doc1\}) pairs
  - Reduce collects all docs for the same term: (term, \{doc1, doc2…\})
  - Sub-indices are merged separately
    • Term-partitioned indices

Ranking

• Previously, expert applications using specialized datasets
  – Queries given as logical formulas or highly selective DB-style queries
  – Expert users
  – Small, high quality datasets
  – Possible to give a precise answer (question-answering)

• Increasingly, end-user applications using Web data
  – Using keyword queries at least as a starting point
  – Non-expert users
  – Large datasets with potential mistakes
  – Not possible to give precise answers, only to provide relevant answers
Ranking methods

• The unit of retrieval is either an object or a sub-graph
  – The sub-graph induced by matching keywords in the query

• Ranking methods from Information Retrieval
  – TF-IDF, BM25, probabilistic methods (language models)
  – Methods such as BM25F allow weighting by predicate

• Machine learning is used to tune parameters and incorporate query-independent (static) features
  – Example: authority scores of datasets computed using PageRank (Harth et al. ISWC 2009)

• Additional topics
  – Relevance feedback and personalization
  – Click-based ranking
  – De-duplication
  – Diversification
Evaluation

Harry Halpin, Daniel Herzig, Peter Mika, Jeff Pound, Henry Thompson, Roi Blanco, Thanh Tran Duc
Semantic Search evaluation at SemSearch 2010 and 2011

- Evaluation is a critical component in developing IR systems
- Keyword search over RDF data
  - Entity Search
    - Queries where the user is looking for a single real world object
    - *Pound et al. Ad-hoc Object Retrieval in the Web of Data, WWW 2010.*
  - List search (new in 2011)
    - Queries where the user is looking for a class of objects
- Focus on relevance, not efficiency
- Real queries and real data
  - Yahoo! and Microsoft query logs
  - Billion Triples Challenge 2009 dataset
- TREC style evaluation
  - Focusing on ranking, not question answering
  - Evaluated using Amazon’s Mechanical Turk
Be a part of the next wave of Web Search

The Web is rapidly evolving into a vast knowledge base of structured information, thanks to advances in technology and standards for data sharing, and the willingness of businesses and public organizations to publish information in an open, interoperable manner.

This emerging Web of Objects promises to herald a new area of Web Search, where search engines index specific items of information about objects, instead of matching keywords in text. These new types of search engines enable Semantic Search, i.e., giving precise answers to complex queries. In the context of the Web, such search engines help users with the additional challenge of consolidating information from multiple sources.

Organizing committee

Roi Blanco (Yahoo! Research)
Harry Halpin (University of Edinburgh)
Daniel M. Herzig (Karlsruhe Institute of Technology)
Peter Mika (Yahoo! Research)
Henry S. Thompson (University of Edinburgh)
Thanh Tran Duc (Karlsruhe Institute of Technology)
Assessment with Amazon Mechanical Turk

- Evaluation using non-expert judges
  - Paid $0.2 per 12 results
    - Typically done in 1-2 minutes
    - $6-$12 an hour
    - Sponsored by the European SEALS project
  - Each result is evaluated by 5 workers
- Workers are free to choose how many tasks they do

- Makes agreement difficult to compute

Number of tasks completed per worker (2010)
**william penn university**

Assess this search result for the above query:

<table>
<thead>
<tr>
<th>property</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>William_Penn_University</td>
</tr>
<tr>
<td>label</td>
<td>William Penn University</td>
</tr>
<tr>
<td>type</td>
<td>University</td>
</tr>
<tr>
<td>type</td>
<td>University108286163</td>
</tr>
<tr>
<td>type</td>
<td>Thing</td>
</tr>
<tr>
<td>subject</td>
<td>Category:Council_of_Independent_Colleges</td>
</tr>
<tr>
<td>subject</td>
<td>Category:Mahaska_County%2C_Iowa</td>
</tr>
<tr>
<td>subject</td>
<td>Category:Universities_and_colleges_affiliated_with_the_Religious_Society_of_Friends</td>
</tr>
<tr>
<td>comment</td>
<td>William Penn University is a private, liberal arts university in Oskaloosa, Iowa, United States. It was founded by members of the Religious Society of Friends (Quakers) in 1873 as Penn College. In 1933, the name was changed to William Penn College, and finally to William Penn University in 2000. Ath. . .</td>
</tr>
<tr>
<td>sameAs</td>
<td>Mx4rv8bUMpwpEbGdrcN3Y29ycA</td>
</tr>
<tr>
<td>reference</td>
<td><a href="http://www.wmpenn.edu/">http://www.wmpenn.edu/</a></td>
</tr>
<tr>
<td>page</td>
<td>William_Penn_University</td>
</tr>
<tr>
<td>wordnet_type</td>
<td>synset university noun 2</td>
</tr>
</tbody>
</table>

- Excellent - describes the query target specifically and exclusively
- Not bad - mostly about the target
- Poor - not about the target, or mentions it only in passing
Results

• 5-6 teams participated in both 2010 and 2011
  – Web queries and web data… difficult task!
• The Semantic Web is not (necessarily) all DBpedia
  – Though a search engine on Dbpedia only would have done well
• Computing agreement is difficult
  – Each judge evaluates a different number of items
  – Follow up experiments validated the Mechanical Turk approach
    • Blanco et al. Repeatable and Reliable Search System Evaluation using Crowd-Sourcing, SIGIR2011
• Result rendering is important
  – Influences the perception of relevance
  – No ‘canonical’ result rendering in Semantic Web search
• Assessments are made public
  – Measure your own system against the best of the field
Next steps

• New tasks
  – More complex queries (but finding them is difficult)
  – Retrieval on RDFa data
  – Aggregated search
  – Ranking properties of objects

• Achieve repeatability
  – Simplify our process and publish our tools
  – Automate as much as possible… except the Turks ;)

• Positioning compared to other evaluation campaigns
  – TREC Entity Track
  – Question Answering over Linked Data
  – SEALS campaigns

• Join the discussion at semsearcheval@yahooogroups.com
Search interface
Search Interface

• Semantic Search brings improvements in
  – Snippet generation
  – Adaptive and interactive presentation
    • Presentation adapts to the kind of query and results presented
    • Object results can be actionable, e.g. buy this product
    • User can provide feedback on particular objects or data sources
  – Aggregated search
    • Grouping similar items, summarizing results in various ways
    • Filtering (facets), possibly across different dimensions
  – Query and task templates
    • Help the user to fulfill the task by placing the query in a task context
Snippet generation using metadata

• Yahoo displays enriched search results for pages that contain microformat or RDFa markup using recognized ontologies
  – Displaying data, images, video
  – Example: GoodRelations for products
  – Enhanced results also appear for sites from which we extract information ourselves

• Also used for generating facets that can be used to restrict search results by object type
  – Example: “Shopping sites” facet for products

• Documentation and validator for developers
  – http://developer.search.yahoo.com

• Formerly: SearchMonkey allowed developers to customize the result presentation and create new ones for any object type

• Haas et al. Enhanced Results in Web Search. SIGIR 2011
Example: Yahoo! Enhanced Results

Enhanced result with deep links, rating, address.
Example: Yahoo! Vertical Intent Search

Related actors and movies
**Example: Time Explorer**

**Congressional Office Forecasts Drop in Deficit, With Possibility of a Later Surplus**

If the United States reduced the number of troops in Iraq to 30,000 in 2010, from more than 130,000 today, the agency estimated, the spending on Iraq would be about $300 billion less than it assumes.

**The Cost of Invading Iraq: Imponderables Meet Uncertainties**

Another study of Iraq war costs, by Linda J. Bilmes of Harvard and Joseph E. Stiglitz of Columbia, comes up with an eye-catching estimate of $2.2 trillion, assuming the United States is no longer... in Iraq in 2015.

**Iraq Is Strategic Issue for Oil Giants, Too**

Mr. Chalabi, in his paper, wrote, "Iraq has the potential to produce 4.7 million barrels a day more oil from discovered fields that are ready to be developed." With the right investments, Iraq
Future work in Semantic Web Search

• (Semi-)automated ways of metadata creation
  – How do we go from 5% to 95%?
• Data quality
  – How do we assess the quality of data?
• Reasoning
  – Automated ontology mapping, instance mapping and blending
  – To what extent is reasoning useful in general?
• Scalability
  – What is between databases and IR engines?
• Ontology reuse
  – How do we get people to reuse ontologies?
• Display
  – How do we automatically generate effective displays for data we don’t understand or only partially understand?
In 2011

• The Semantic Web is still evolving
  – Finally, a JSON syntax for RDF!

• Leaner and meaner
  – Bottom-up approaches: graphs, statistics, mining, retrieval, visualization…

• We get some credit in other fields
  – RDF data management is now a topic at VLDB, others
  – CIKM 2011 has 67 submitted abstracts on Semantic Search!

• The Semantic Web is 10 years wiser… and so am I
The End