



Semantic Search

Peter Mika

Researcher, Data Architect

Yahoo! Research / YST

Yahoo! by numbers (April, 2007)

- There are approximately **500 million users** of Yahoo! branded services, meaning we reach 50 percent – or **1 out of every 2 users** – online, the largest audience on the Internet (Yahoo! Internal Data).
- Yahoo! is the most visited site online with nearly **4 billion visits** and **an average of 30 visits per user per month in the U.S.** and leads all competitors in audience reach, frequency and engagement (comScore Media Metrix, US, Feb. 2007).
- Yahoo! accounts for the largest share of time Americans spend on the Internet with 12 percent (comScore Media Metrix, US, Feb. 2007) and **approximately 8 percent of the world's online time** (comScore WorldMetrix, Feb. 2007).
- **Yahoo! is the #1 home page** with 85 million average daily visitors on Yahoo! homepages around the world, an increase of nearly 5 million visitors in a month (comScore WorldMetrix, Feb. 2007).
- Yahoo!'s social media properties (Flickr, delicious, Answers, 360, Video, MyBlogLog, Jumpcut and Bix) have **115 million unique visitors worldwide** (comScore WorldMetrix, Feb. 2007).
- Yahoo! Answers is the largest collection of human knowledge on the Web with more than 90 million unique users and **250 million answers** worldwide (Yahoo! Internal Data).
- There are more than **450 million photos** in Flickr in total and **1 million photos** are uploaded daily. 80 percent of the photos are public (Yahoo! Internal Data).
- **Yahoo! Mail is the #1 Web mail provider in the world** with 243 million users (comScore WorldMetrix, Feb. 2007) and nearly 80 million users in the U.S. (comScore Media Metrix, US, Feb. 2007)
- Interoperability between Yahoo! Messenger and Windows Live Messenger has formed the largest IM community approaching 350 million user accounts (Yahoo! Internal Data).
- **Yahoo! Messenger is the most popular in time spent** with an average of 50 minutes per user, per day (comScore WorldMetrix, Feb. 2007).
- Nearly 1 in 10 Internet users is a member of a **Yahoo! Groups** (Yahoo! Internal Data).
- Yahoo! is one of only 26 companies to be on both the Fortune 500 list and the Fortune's "Best Place to Work" List (2006).



SANDBOX



A place to play with innovations from Yahoo! Research



BROWSE YAHOO! RESEARCH

[About Yahoo! Research](#)

[Academic Relations](#)

[Events](#)

[Job Opportunities](#)

[News](#)

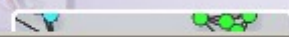
[People](#)

Yahoo! Research is the central advanced research organization of Yahoo! Inc., a leading global Internet brand and one of the most trafficked Internet destinations worldwide.

We're responsible for big inventions - our goals are nothing short of inventing the future of the Internet and creating the next generation of businesses for Yahoo!



Featured Project



Graph Partitioning



Events

ACM Seventeenth

Yahoo! Research Barcelona

- Established January, 2006
- Led by Ricardo Baeza-Yates
- Research areas
 - _ Web Mining
 - content, structure, usage
 - _ Distributed Web retrieval
 - _ Multimedia retrieval
 - _ NLP and Semantics



<http://www.flickr.com/photos/borbis/>



Barcelona, Spain



Semantic Search

State of search

- “We are at the beginning of search.” (Marissa Mayer)
- Old battles are won
 - Marginal returns on investments in crawling, indexing, ranking
 - Solved large classes of queries (e.g. navigational)
 - Lots of tempting, but high hanging fruits
- Currently, the biggest bottlenecks in IR not computational, but in modeling user cognition (Prabhakar Raghavan)
 - If only we could find a computationally expensive way to solve the problem
 - then we should be able to make it go faster
- In particular, solving queries that require a deeper understanding of the query, the content and/or the world at large

Some examples...

- Ambiguous searches
 - Paris Hilton
- Multimedia search
 - Images of Paris Hilton
- Imprecise or overly precise searches
 - Publications by Jim Hendler
 - Find images of strong and adventurous people (Lenat)
- Searches for descriptions
 - Search for yourself without using your name
 - Product search (ads!)
- Searches that require aggregation
 - Size of the Eiffer tower (Lenat)
 - Public opinion on Britney Spears
 - World temperature by 2020

Not just search...

Web | Images | Video | Local | Shopping | more ▾

harrison ford

Search

Options ▾

Customize ▾

YAHOO!

1 - 10 of 41,700,000 for **harrison ford** (About) - 0.04 s | SearchScan

Also try: [harrison ford runner](#), [harrison ford movies](#), [More...](#)

Harrison Ford Web

Fan site includes a large collection of photos, articles, biography, and more on **Harrison Ford**.

www.harrisonfordweb.com

 (3 reviews) - Stumb... ▾ × |  14 bookmarks - delici... ▾ ×

Harrison Ford (I)

Actor: Raiders of the Lost Ark. His father was Irish, his mother Russian-Jewish. ... Discuss this name with other users on IMDb message board for **Harrison Ford (I)** ...

www.imdb.com/name/nm0000148 - 82k - [Cached](#)

 (1 fan) - StumbleU... ▾ × |  5 bookmarks - delici... ▾ ×

Harrison Ford (II)

Actor: Janice Meredith. Silent screen leading man in films from 1915-1932. ... Discuss this name with other users on IMDb message board for **Harrison Ford (II)** ...

www.imdb.com/name/nm0001230 - [Cached](#)

 (Review this site) - StumbleUpon ▾ ×

Harrison Ford - Wikipedia, the free encyclopedia

[Early life](#) | [Career](#) | [Personal life](#) | [Filmography](#)

Harrison Ford (born July 13, 1942) is an award-winning American actor. **Ford** is best known for his performances as the title character in the Indiana Jones film series and as Han Solo in the original Star Wars trilogy. He is...

en.wikipedia.org/wiki/Harrison_Ford - 158k - [Cached](#)



SPONSOR RE

Planning a Night out?

Any Orange customer. Any mate. Any Wednesday. Text film to 241.

www.orange.co.uk/orangewednesdays

Ford Posters – up to 75% Less

Incredible prices. Find **ford** posters & save up to 75%.

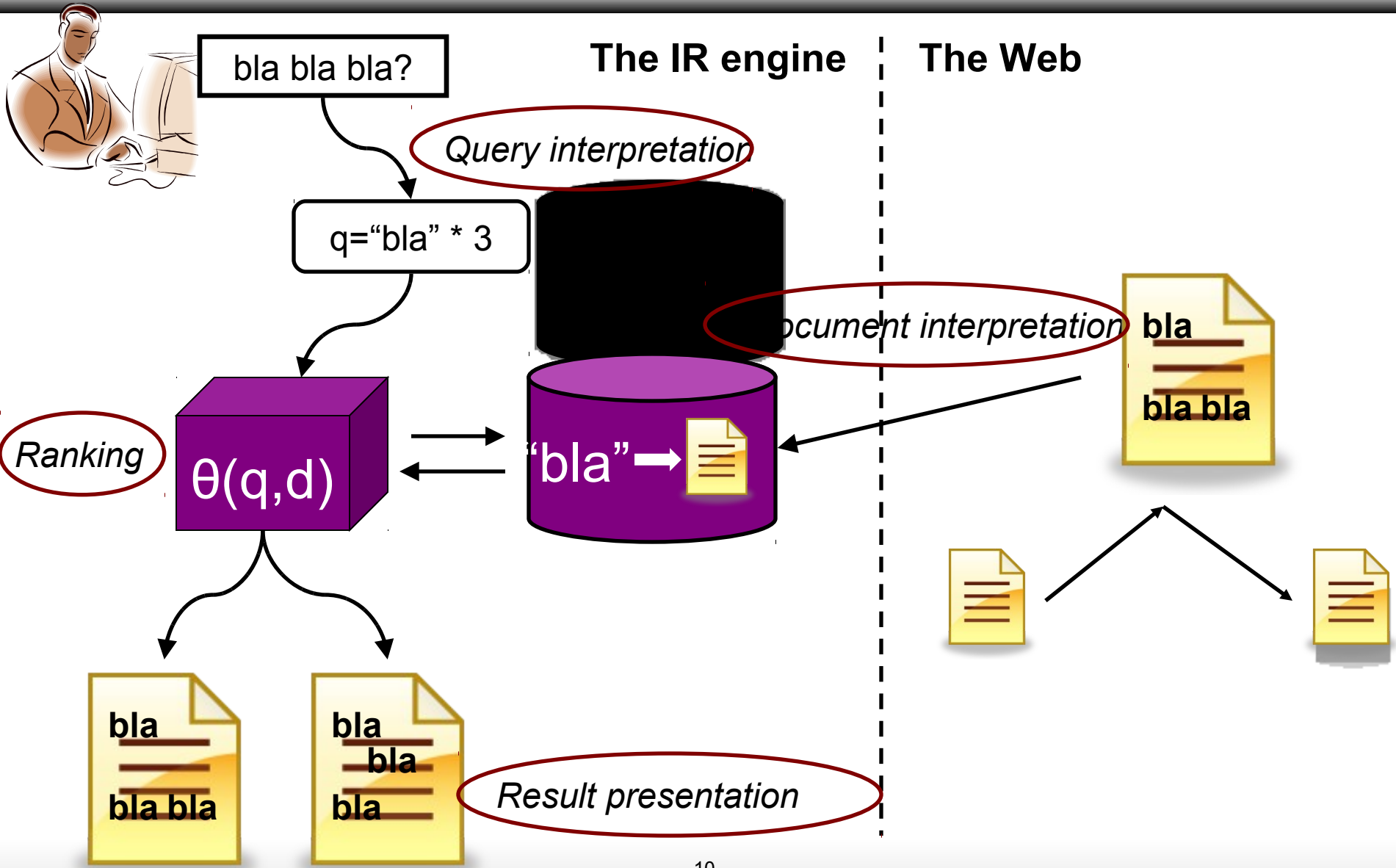
www.uk.Best-Price.com

[See your message here...](#)

Semantic (Web) Search

- **Def.** matching the user's query with the Web's content at a conceptual level, often with the help of world knowledge
 - R. Guha, R. McCool: Semantic Search, WWW2003
- Moving from document search toward search over structured data
 - Without dividing the search space into verticals
- Related disciplines
 - Semantic Web, IR, Databases, NLP, IE
- As a field
 - ISWC/ESWC/ASWC, WWW, SIGIR
 - Exploring Semantic Annotations in Information Retrieval (ECIR08, WSDM09)
 - Semantic Search Workshop (ESWC08, WWW09)
 - Future of Web Search: Semantic Search (FoWS09)

Semantics at every step of the IR process





Document Interpretation

Document processing

- Goal
 - Provide a higher level representation of information in some conceptual space
 - Conceptual space is different for Semantic Web and NLP based search engines
- Limited document understanding in traditional search
 - Page structure such as fields, templates
 - Understanding of anchors, other HTML elements
 - Limited NLP
- In Semantic Search, more advanced text processing and/or reliance on explicit metadata
 - Information sources are not only text but also databases and web services

Automated approaches

- Trade-off between precision, recall and effort
 - Low cost, broad coverage, but error-prone
 - Expensive, targeted but precise
- Variety of approaches
 - NLP
 - Named entity extraction with or without disambiguation
 - From text to triples
 - Linguistic patterns
 - Deep parsing
 - Information Extraction
 - Form filling combined with list extraction (Halevy et al.)
 - Wrapper induction
- (Public) examples:
 - NLP: Zemanta, OpenCalais,
 - Wrapper induction: Dapper, MashMaker

Example: Zemanta

- A personal writing assistant for bloggers
 - Plugin for popular blogging platforms and web mail clients
- Analyzes text as you type and suggests hyperlinks, tags, categories, images and related articles
- API available with the same functionality

Your content enhanced!

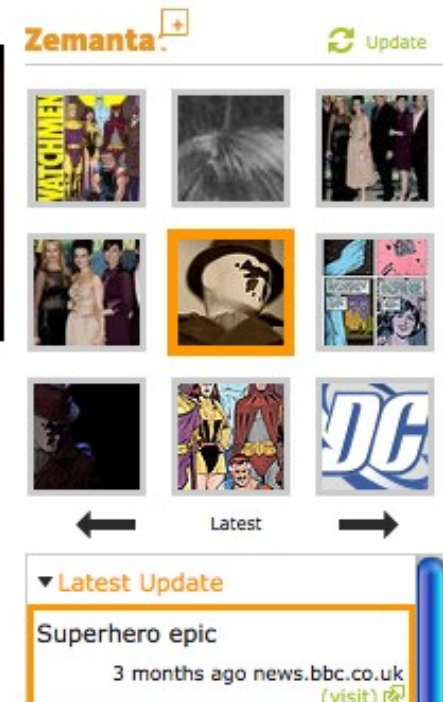
Branded "unfilmable", **Watchmen** - the cult graphic novel about a group of retired, flawed superheroes - has finally made it to the big screen. From the second the opening credits roll, it is clear Watchmen is not your typical superhero movie.



Image by [TCM Hitchhiker](#) via Flickr

An ageing vigilante, The Comedian, is attacked in his high-rise apartment before being hurled 10 storeys to his death... in graphic slow motion. What follows is a two-and-three-quarter hour epic that centres on an outlawed group of deeply flawed former heroes as a Cold War Doomsday clock inches ever closer to midnight and nuclear apocalypse.

First published in 12 parts by DC Comics in 1986, Watchmen was



The screenshot shows the Zemanta plugin interface. At the top left is the 'Zemanta' logo with a plus sign icon. To its right is an 'Update' button with a refresh icon. Below these is a grid of nine image suggestions. The second image in the second row is highlighted with a yellow border. Below the grid are navigation arrows and the word 'Latest'. At the bottom, there is a 'Latest Update' section with a dropdown arrow, containing the text 'Superhero epic' and '3 months ago news.bbc.co.uk (visit)'. A blue vertical bar is on the right side of the interface.

Semantic Web

- Making the Web searchable through explicit semantics
 - Embedded metadata
 - microformats
 - RDFa
 - Microdata (HTML5)
 - Publisher feeds
 - DataRSS
 - Wrappers around websites and web services
 - SearchMonkey, YQL
 - SPARQL endpoints or Linked Data wrappers around databases

Example: microformats and RDFa

Microformat (hCard)

```
<div class="vcard">  
  <a class="email fn" href="mailto:jfriday@host.com">Joe Friday </a>  
  <div class="tel"> +1-919-555-7878 </div>  
  <div class="title"> Area Administrator, Assistant </div>  
</div>
```

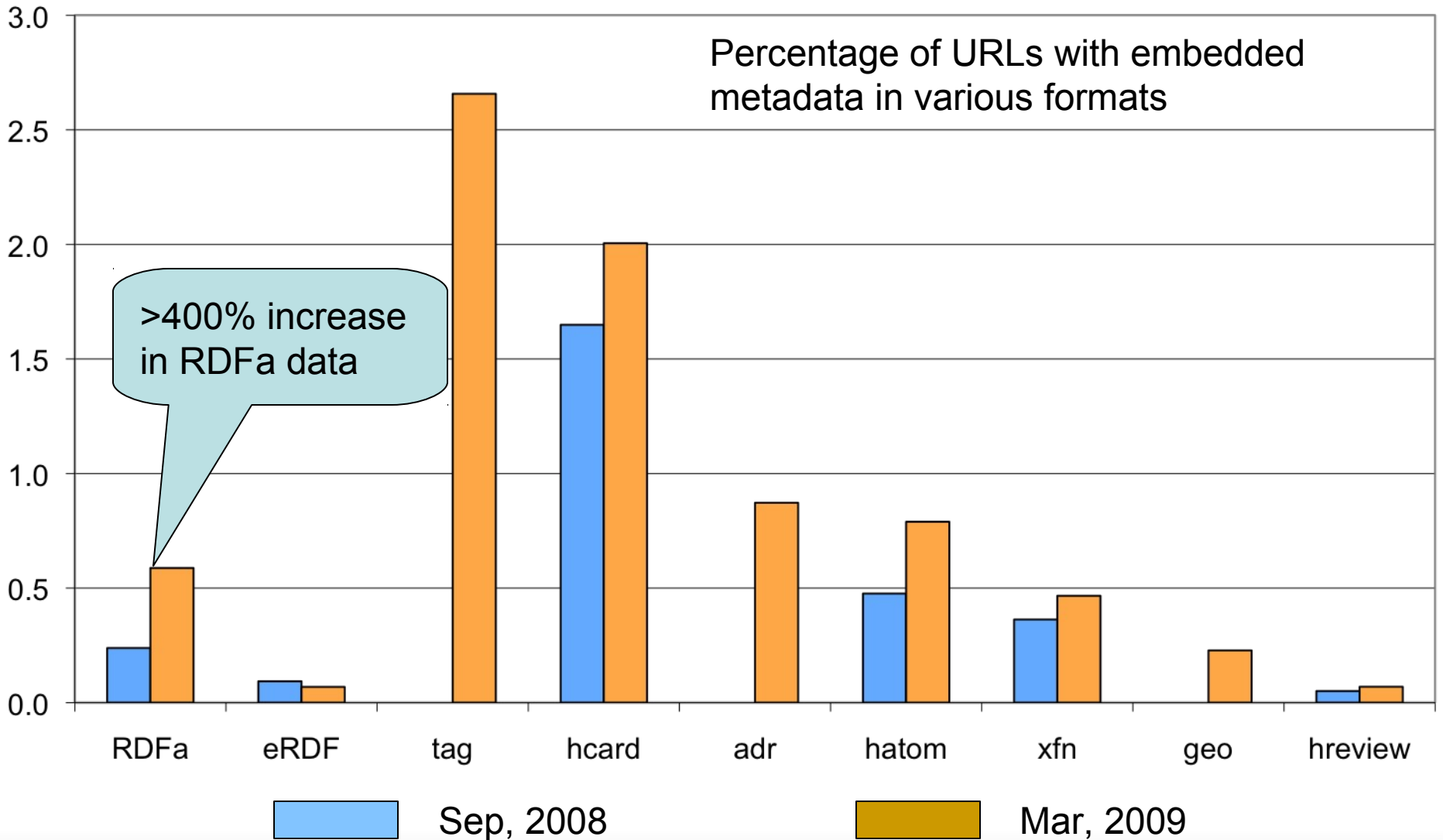
RDFa

```
<p typeof="contact:Info" about="http://example.org/staff/jo">  
  <span property="contact:fn">Jo Smith</span>.  
  <span property="contact:title">Web hacker</span> at  
  <a rel="contact:org" href="http://example.org"> Example.org </a>.  
  You can contact me <a rel="contact:email"  
    href="mailto:jo@example.org">  
  via email </a>.  
</p> ...
```

Coming soon: Microdata in HTML5

How far are we? Lot's of data or very little?

Percentage of URLs with embedded metadata in various formats



Investigating the data gap through query logs

- How big a Semantic Web do we need?
- Just big enough... to answer all the questions that users may want to ask
 - Query logs are a record of what users want to know as a whole
- Research questions:
 - How much of this data would ever surface through search?
 - What categories of queries can be answered?
 - What's the role of large sites?

Method

- Simulate the average search behavior of users by replaying query logs
 - Reproducible experiments (given query log data)
 - BOSS web search API returns RDF/XML metadata for search result URLs
- Caveats
 - For us, and for the time being, search = document search
 - For this experiment, we assume current bag-of-words document retrieval is a reasonable approximation of semantic search
 - For us, search = web search
 - We are dealing with the average search user
 - There are many queries the users have learned not to ask
 - Volume is a rough approximation of value
 - There are rare information needs with high pay-offs, e.g. patent search, financial data, biomedical data...

Data

- Microformats, eRDF, RDFa data
- Query log data
 - US query log
 - Random sample of 7k queries
 - Recent query log covering over a month period
- Query classification data
 - US query log
 - 1000 queries classified into various categories

Number of queries with a given number of results with particular formats (N=7081)

	1	2	3	4	5	6	7	8	9	10	Impressions	Average impressions per query
ANY	2127	1164	492	244	85	24	10	5	3	1	7623	1.08
hcard	1457	370	93	11	3	0	0	0	0	0	291	0.36
rel-tag	1317	350	95	44	14	8	6	3	1	1	291	0.38
adr	456	77	21	6	1	0	0	0	0	0	702	0.10
hatom	450	52	8	1	0	0	0	0	0	0	582	0.08
license	359	21	1	1	0	0	0	0	0	0	106	0.06
xfn	339	26	1	1	0	0	0	1	0	0	106	0.06

On average, a query has at least one result with metadata.

Are tags as useful as hCard?

That's only 1 in every 16 queries.

Notes:

- Queries with 0 results with metadata not shown
- You cannot add numbers in columns: a query may return documents with different formats
- Assume queries return more than 10 results

 Impressions

 Average impressions per query

The influence of head sites (N=7081)

	1	2	3	4	5	6	7	8	9	10		
ANY	2127	1164	492	244	85	24	10	5	3	1	7623	1.08
hcard	1457	370	93	11	3	0	0	0	0	0	2535	0.36
rel-tag	1317	350	95	44	14	8	6	3	1	1	2681	0.38
wikipedia.org	1676	1	0	0								0.24
adr	456	77	21	6								0.10
hatom	450	52	8	1	0	0	0	0	0		582	0.08
youtube.com	475	1	0	0	0	0	0	2	0	0	495	0.07
license	359	21	1	1	0	0	0	0	0	0	408	0.06
xfn	339	26	1	1	0	0	0	1	0	0	406	0.06
amazon.com	345	3	0	0	0	0	1	0	0	0	358	0.05

If YouTube came up with a microformat, it would be the fifth most important.

 Impressions

 Average impressions per query

Restricted by category: local queries (N=129)

	1	2	3	4	5	6	7	8	9	10		
ANY	36	16	10	0	4	1	0	0	0	0	124	0.96
hcard	0	0	0	0	0	0	0	0	0	0	64	0.50
adr	0	0	0	0	0	0	0	0	0	0	41	0.32
local.yahoo.com	24	0	0	0	0	0	0	0	0	0	24	0.19
en.wikipedia.org	24	0	0	0	0	0	0	0	0	0	24	0.19
rel-tag	19	2	0	0	0	0	0	0	0	0	23	0.18
geo	16	5	0	0	0	0	0	0	0	0	26	0.20
www.yelp.com	16	0	0	0	0	0	0	0	0	0	16	0.12
www.yellowpages.com	14	0	0	0	0	0	0	0	0	0	14	0.11

The query category largely determines which sites are important.

 Impressions

 Average impressions per query

Summary

- Time to start looking at the demand side of semantic search
 - Size is not a measure of usefulness
- For us, and for now, it's a matter of who is looking for it
 - “We would trade a gene bank for fajita recipes any day”
 - Reality of web monetization: pay per eyeball
- Measure different aspects of usefulness
 - Usefulness for improving presentation but also usefulness for ranking, reasoning, disambiguation...
- Site-based analysis
- Linked Data will need to be studied separately



Query Interpretation

Query Interpretation

- Provide a higher level representation of queries in some conceptual space
 - Ideally, the same space in which documents are represented
- Interpretation treated as a separate step from ranking
 - Required for federation, i.e. determine where to send the query
 - Used in ranking
 - Due to performance requirements
 - You cannot execute the query to determine what it means and then query again
 - Automated process
 - Limited user involvement, e.g. search assist, facets
- Important also for query normalization in analysis
 - Query log data is extremely sparse: 88% of unique queries are singleton queries

Query Interpretation in Semantic Search

- Queries may be keywords, questions, semi-structured, structured etc.
- For now, the world is using keyword queries. But what kind of structures could be extracted from keyword queries?

Query type	Percent	Example
Entity query	40.60%	starbucks palo alto
Type query	12.13%	plumbers barcelona
Attribute query	4.63%	zip code waterville maine
Relation query	0.68%	chris brown rihanna
Other keyword query	36.10%	nightlife barcelona
Uninterpretable	5.89%	YÖÿ#ÿ±YÖ

Investigating the ontology gap through query logs

- Does the language of users match the ontology of the data?
 - Initial step: what is the language of users?
- Observation: the same type of objects often have the same query context
 - Users asking for the same aspect of the type

Query	Entity	Context	Class
aspirin side effects	ASPIRIN	+side effects	Anti-inflammatory drugs
ibuprofen side effects	IBUPROFEN	+side effects	Anti-inflammatory drugs
how to take aspirin	ASPIRIN	-how to take	Anti-inflammatory drugs
britney spears video	BRITNEY SPEARS	+video	American film actors
britney spears shaves her head	BRITNEY SPEARS	+shaves her head	American film actors

- These are potential attributes or relationships

Models

- Desirable properties:
 - P1: Fix is frequent within type
 - P2: Fix has frequencies well-distributed across entities
 - P3: Fix is infrequent outside of the type

- Models:

M1: Most likely fix (highest mean) [P1]:

$$P(f|T) = \left(\frac{n_{Tf}}{n_{T\bullet}} \right) \propto_T n_{Tf}$$

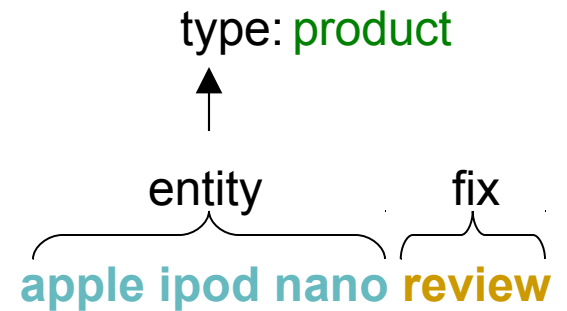
M2: Most discriminant fix [P1,P2]:

$$\frac{P(f|T)}{P(f)} = \left(\frac{n_{Tf}}{n_{T\bullet}} \right) \left(\frac{n_{\bullet f}}{n_{\bullet\bullet}} \right)^{-1} \propto_T \frac{n_{Tf}}{n_{\bullet f}}$$

M3: Most disuniform outside type [P3]:

$$\beta_f = (P(T|f), 1 - P(T|f))$$

$$-H(\beta_f) = P(f|T) \log_2 P(f|T) + [1 - P(f|T)] \log_2 [1 - P(f|T)]$$



Models cont.

M4: Most typical fix (geometric mean) [P1,P2]:

$$G(f|T) = \left(\prod_{e \in T} n_{ef} \right)^{1/|T|}$$

M4' taking outside type into account [P1,P2,P3]:

$$G(f|T)/\lambda G(f)$$

$$G(f|T) - \lambda G(f)$$

M5: Most uniform fix (entropy) [P2]:

$$\theta_{f|T} = (P(e|f, T))_{e \in T}$$

$$H(\theta_{f|T}) = - \sum_{e \in T} P(e|f, T) \log_2 P(e|f, T)$$

Demo

Pre- and postfix search

Start typing for auto completion:

ru^ssia

Then select a type for that entity:

Russian-speaking_countries_and_territories

Currently, this demo shows Wikipedia **categories** as types. The models below are described [here](#). Some examples to try:

[Madonna](#)

[Russia](#)

[Singapore](#)

[Aspirin](#)

[Transformers](#)

Entity results Type results for: *Russian-speaking countries and territories*

Most frequent for *ru^ssia*

Fix
free
yahoo
from
nude
sexy
map of
learn
young
with love
forward
naked

Most discriminant (M2)

Fix
moenchengladbach
china,korea
british trekker
and ukraine
with love cheats
dortmund football kit
in 2002 and 2003
with love ps2
hot teen photo
must aspired to claim world leadership

Weighed geometric

Fix
wwwplaneta
forbes
photos of
travel to
gay
private
mi-14
flight from
in
looking for
embassy in london
immigration
largest
visit
latest
music
future of
extreme
south
lady
silver
airline
what

Lambda: 0.5

Highest entropy

Fix
embassy
airlines
embassy london
girls
map
airline
visa
in london
flights
news

Qualitative evaluation

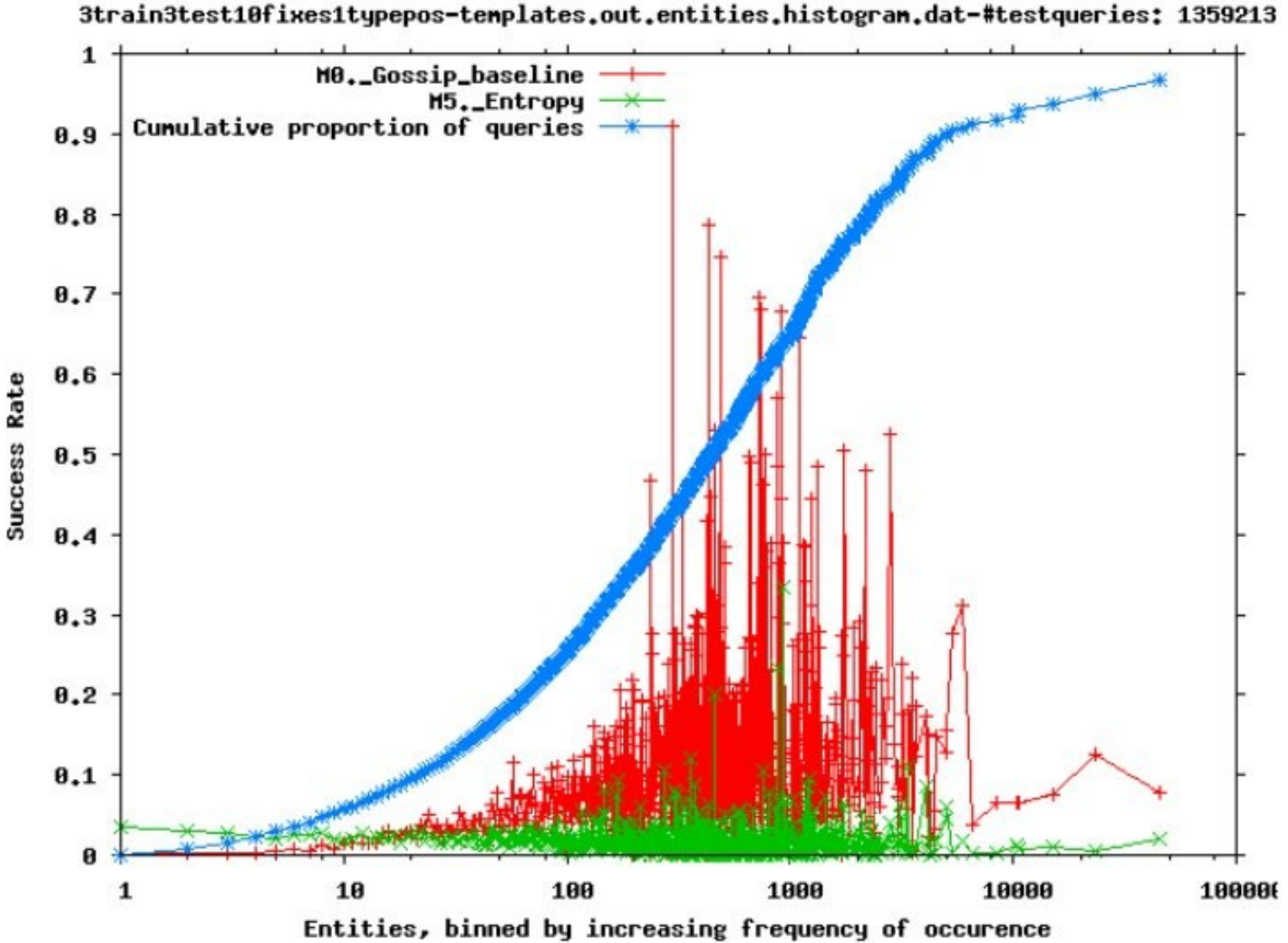
- Four wikipedia templates of different sizes
- **Bold** are the information needs that would be actually fulfilled by infobox data

Settlement	Musical artist	Drug	Football club
hotels	lyrics	buy	forum
map	buy	what is	news
map of	pictures of	tablets	website
weather	what is	what is	homepage
weather in	video	side effects of	tickets
flights to	download	hydrochloride	official website
weather	hotel	online	badge
hotel	dvd	overdose	fixtures
property in	mp3	capsules	free
cheap flights to	best	addiction	logo

Evaluation by query prediction

- Idea: use this method for type-based query completion
 - Expectation is that it improves infrequent queries
- Three days of UK query log for training, three days of testing
- Entity-based frequency as baseline (~current search suggest)
- Measures
 - Recall at K, MRR (also per type)
- Variables
 - models (M1-M6)
 - number of fixes (1, 5, 10)
 - mapping (templates vs. categories)
 - type to use for a given entity
 - Random
 - Most frequent type
 - Best type
 - Combination
 - To do: number of days of training

Results: success rate (binned)



Summary

- Some improvement on query completion task
 - Win for rare queries
 - Raw frequency wins quickly because of entity-specific completions
- Potentially highly valuable resource for other applications
 - Facets
 - Automated or semi-automated construction of query-trigger patterns
 - Query classification
 - Characterizing and measuring the similarity of websites based on the entities, fixes and types that lead to the site
- Further work needed to turn this into a vocabulary engineering method



Indexing and Ranking

Indexing and Ranking

- Goal:
 - Precise matching of the query representation to content representation over as large a base as possible
 - Efficiency in both indexing (offline) and ranking (online)
- Indexing and ranking are the IR core
 - Ranking features (TF-IDF, PageRank, clicks) are studied in great detail
 - Machine Learning used to build the model (formula) of how to combine features
 - Lot's of engineering: specialized data structures, encodings, distributed architectures, etc.
- We don't discuss crawling: largely unchanged

Indexing and Ranking in Semantic Search

- Ranking has not been an issue for the Semantic Web until recently
 - Small volumes of data
 - Logical framework (relevance is binary)
- More recently:
 - Need for searching large volumes of data
 - Using keyword queries at least as a starting point
- Databases can execute structured queries and web search scales... but what's in between?
 - Approaches both from the database and IR worlds

In the best of cases...

- Matching the query intent with the document metadata can be trivial
 - Example: queries composed with Freebase Suggest

Query:

<http://rdf.freebase.com/ns/en.Madonna>

Interpretation:

<http://rdf.freebase.com/ns/en.Madonna>

Data

<http://rdf.freebase.com/ns/en.Madonna>

Query interpretation is a source of uncertainty

Query:

madonna

Interpretation:

```
<adjunct id="com.yahoo.query.intent" version="0.5">
  <type typeof="fb:music.artist foaf:Person">
    <meta property="foaf:name">Madonna</meta>
  </type>
</adjunct>
```

Document metadata:

```
<adjunct id="com.yahoo.page.hcard" version="0.5">
  <type typeof="foaf:Person">
    <meta property="foaf:name">Madonna</meta>
  </type>
</adjunct>
```


Text interpretation is a source of uncertainty

Query:

madonna

Interpretation:

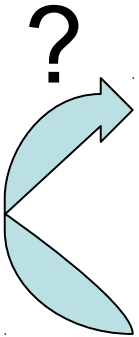
```
<adjunct id="com.yahoo.query.intent" version="0.5">
  <type typeof="fb:music.artist foaf:Person">
    <meta property="foaf:name">Madonna</meta>
  </type>
</adjunct>
```

Document metadata:

```
<adjunct id="com.yahoo.page.hcard" version="0.5">
  <type typeof="foaf:Person">
    <meta property="foaf:name">Madonna</meta>
  </type>
</adjunct>
```

Text:

Madonna, along with Michael Jackson one of the most successful singers of the 1980s...



Matching is a source of uncertainty

Query:

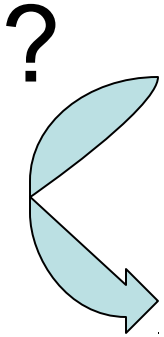
madonna

Interpretation:

```
<adjunct id="com.yahoo.query.intent" version="0.5">
  <type typeof="fb:music.artist foaf:Person">
    <meta property="foaf:name">Madonna</meta>
  </type>
</adjunct>
```

Document metadata:

```
<adjunct id="com.yahoo.page.hreview" version="0.5">
  <type typeof="review:Review">
    <meta property="review:text">The latest single by superstar
    Madonna has topped the charts for seven consecutive weeks...</meta>
  </type>
</adjunct>
```



Indexing and Ranking in Semantic Search

- Data integration, ontology mapping and other forms of reasoning could be used offline or at query time
 - Open question precisely what kinds of reasoning are useful for retrieval
 - Ontology matching?
 - Entity resolution?
- Shared ontologies are still crucial
 - Entity resolution, ranking, presentation can be tailored to the type of resource
- Heterogeneity, quality of data is still an issue
 - Very heterogeneous: from well-curated data sets to microformats

Ontology matching and entity resolution

- 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
- Often combined
 - Ontology is also part of the data in Semantic Web!

Examples of current semantic search engines

- Structured data and hybrid search engines
 - Semantic Web search engines
 - Sindice, SWSE (VisiNav), Watson, Swoogle, Falcon-S
 - Information extraction based
 - Google Squared
 - Searching curated closed world datasets
 - Wolfram Alpha
 - Research demos
 - Semplore, The Information Workbench
- Document (web) search engines with semantic features
 - Yahoo's SearchMonkey
 - Google's Rich Snippets

Future work: semantic search evaluation

- Problem: lack of rigorous evaluation of results
- A typology of web queries based on structure
 - Entity, entity-attribute, entity + context entity, entity type, entity relationship, other
 - Tool for annotating web queries
- Relevance evaluation
 - Focusing on entity and entity type queries
 - Keyword queries, results are lists of entities
- Goal: INEX style evaluation campaign in 2010
 - Linked Data or embedded metadata corpus
 - A selected set of web queries
- Join the discussion group
 - <http://groups.yahoo.com/group/semsearcheval/>



Result presentation

Search Interface

- Goal is to facilitate the interaction between the user and the system
 - helping the user to formulate queries
 - present the results in an intelligent manner
- Improvements in
 - Snippet generation
 - Adaptive presentation
 - presentation adapts to the kind of query and results presented
 - Aggregated search
 - Grouping similar items, summarizing results in various ways
 - Possibilities for filtering, possibly across different dimensions
 - Task completion
 - Help the user to fulfill the task by placing the query in a task context



- 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
- <http://developer.yahoo.com/searchmonkey/>



Enhanced Result

[Web](#) | [Images](#) | [Video](#) | [Local](#) | [Shopping](#) | [more](#) ▾

[Options](#) ▾

[Yahoo!](#) | [My Yahoo!](#) | [Mail](#) | Welcome, [Guest](#) ([Sign In](#)) | [Help](#)

YAHOO!

1 - 10 of about 8,140,000 for **art of pizza chicago** ([About this page](#)) - 0.23 sec.

deep links

image

name/value pairs or abstract

[The Art of Pizza - Chicago, IL : Citysearch.com](#)

Get details on The **Art of Pizza - Chicago, IL**, at Citysearch - over 1 million user reviews & editorials about local businesses.
www.citysearch.com/profile/3735139

[The Art of Pizza - Chicago, IL, 60657-3035 - Citysearch](#)

... information, directions, and reviews on The **Art of Pizza** and other 5-Star Restaurants in Chicago ... The **Art of Pizza**. 3033 N Ashland Ave. Chicago, IL 60657 ...
chicago.citysearch.com/profile/.../chicago_il/the_art_of_pizza.html - 76k - [Cached](#)

[The Art of Pizza - Lakeview - Chicago, IL 60657](#)



[Reviews](#)
[Photos](#)
[Send to a friend](#)
[Send to Phone](#)

Ratings: ★★★★★ (173)
Address: 3033 N Ashland Ave, Chicago, IL
Phone: (773) 327-5600
Price Range: \$

www.acmreviews.com/biz/the-art-of-pizza-chicago - 145k - [Cached](#)

[The Art of Pizza | Metromix Chicago](#)

This modest storefront eatery's deep-dish pizza was voted Chicago's best by the Tribune. ... Rate The Art of Pizza "its not worth it" circuit from logan ...
chicago.metromix.com/restaurants/italian/.../140369/content - [Cached](#)

[Art of Pizza: Chicago Reader Restaurants: Restaurant Raters' Comments](#)

Art of Pizza. 3033 N. Ashland Ave. Chicago ... Chicagoan, the **Art of Pizza** is one of the pizza greats in Chicago, and a relatively hidden gem. ...
www.chicagoreader.com/cgi-bin/rrr/comments.cgi?numb=2736 - [Cached](#)

[Art of Pizza on Centerstage Chicago - Art of Pizza : 3033 N. Ashland ...](#)

Art of Pizza on Centerstage, Publisher of Honest Info By Chicago, For Chicago ... Eateries like Lakeview's **Art of Pizza** prove that a food lover's paradise can be ...
centerstagechicago.com/restaurants/art-of-pizza.html

SPONSOR RESULTS

[Chicago Pizza Chicago](#)

Pizza Now - Call Today Chicago Style, World Class Taste.
dexonline.com/ledos-pizza

[Nick & Bruno's Pizzeria](#)

Chicago area pizza restaurant Home of the Monster Pizza.
Nick-Brunos.com

[Pizza Place in Chicago](#)

Find & review local restaurants and cafes. Free search.
Chicago.Citysearch.com

[Pizza Art](#)

Find **Pizza Art** and Compare prices at Smarter.com.
www.smarter.com

[See your message here...](#)



Yahoo! My Yahoo! Mail Welcome, **Guest** [Sign In] Help

Web | Images | Video | Local | Shopping | more ▾

Dr. Seuss Horton Hears a Who Movie Options ▾

1 - 10 of about 17,800,000 for **Dr. Seuss' Horton Hears a Who Movie** (About this page) - 0.33 sec.

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

Dr. Seuss' Horton Hears a Who (2008) Movie - Acme Movies

	Movie Details	Critics Review: ★★★★★ (173)
	Showtimes & Tickets	MPAA Rating: G
	Trailers & Clips	Running Time: 1 hr. 28 min.
	Critics Reviews	Release Date: March 14th, 2008 (wide)

acmemovies.com/hortonhearsawho - Cached

F Reviews ★★★★★ (77) | IMDb Cast & Crew | Netflix ▲

Horton Hears a Who! (2008) - Netflix **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."

★★★★☆

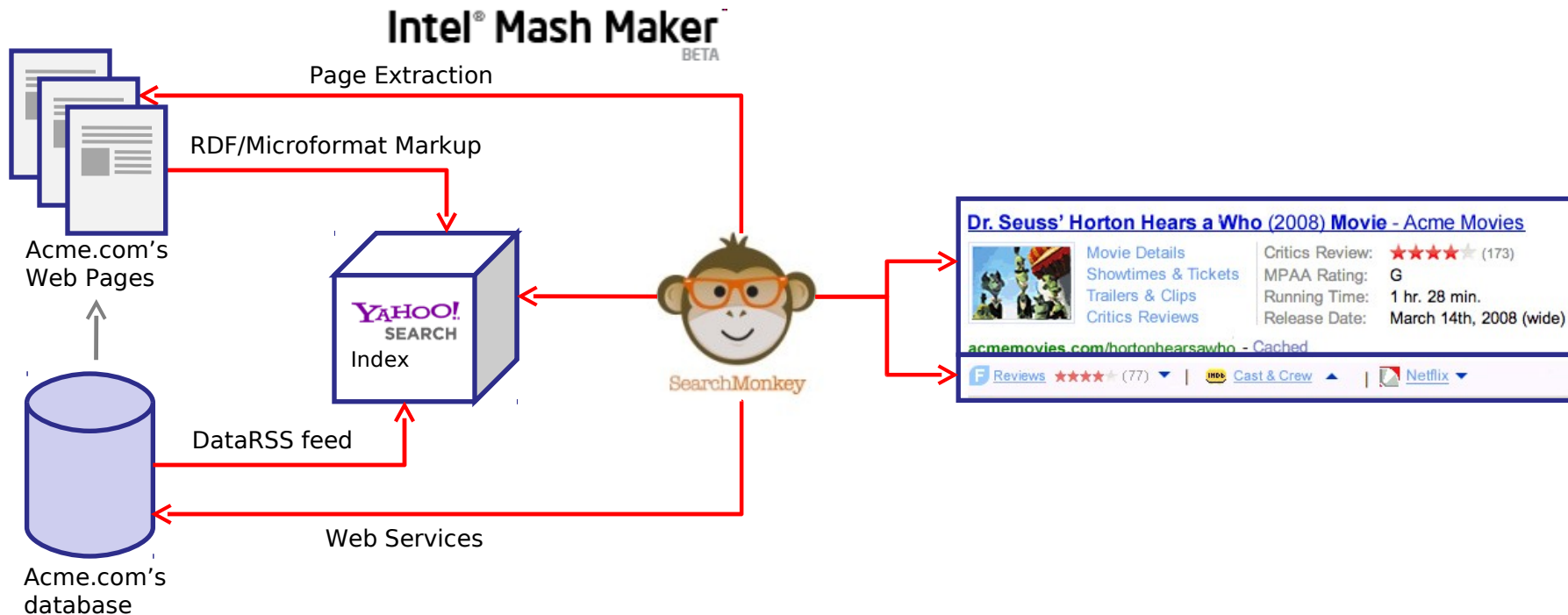
Enjoyed By Members Who Enjoyed

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

SearchMonkey



- 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



Standard enhanced results

Embed markup in your page, get an enhanced results without any programming

STANDARD ENHANCED RESULTS:

If you have any of the following types of data embedded on your site, create an enhanced result by adding a few lines of code to your page.



PRODUCT

Highlight prices, images, and product info directly in search results.

[Show Me How](#)



LOCAL

Display phone numbers, addresses, and other local info.

[Show Me How](#)



NEWS

Show the publication date and a photo for news articles.

[Show Me How](#)



VIDEO

Embed Flash video from your site directly in search results.

[Show Me How](#)



EVENT

Show the address and date for festivals and other activities.

[Show Me How](#)



DOCUMENTS

Display presentations and other documents.

[Show Me How](#)



DISCUSSION

Show the number of comments and thread date for discussions.

[Show Me How](#)



GAMES

Preview games right on the search results page.

[Show Me How](#)

Documentation

Simple and advanced, examples, copy-paste code, validator

SearchMonkey > Start Overview > Video



Video

A clip or other video

EXAMPLE

[YOUTUBE - IRISH DANCING MONKEYS](#)

Monkeys can also dance. And the music they prefer is Irish Step!

www.youtube.com/watch?v=44Y-_JAJAwE - 101k



DESCRIPTION

Play videos directly in Yahoo! Search. Add code to indicate to Yahoo! that a video exists on your page, and when we next crawl your site, we'll take care of the rest. From your markup, we'll be able to extract structured data from your site and render it automatically in an enhanced result in search results.

RDFa

Facebook Share

Feed

MINIMUM

Click any item to toggle description on or off

[xmlns:media](#) (url)

[media:thumbnail](#) (gif, jpg, png)

[media:video](#) (url)

EXAMPLE 1

```
<object width="512" height="296" rel="media:video"
  resource="http://example.com/video_object.swf"
  xmlns:media="http://search.yahoo.com/searchmonkey/media/">
  <a rel="media:thumbnail" href="http://example.com/thumbnail_preview.jpg"
    <param name="movie" value="http://example.com/video_object.swf" />
  <embed src="http://example.com/video_object.swf" type="application/x-shoc
  </embed>
</object>
```

NOTE: Your video player's embed code might vary.



Yahoo! | My Yahoo! | Mail | More ▾ Welcome , serendipity588 | Sign out | Help

YAHOO! SEARCH GALLERY BETA Y Search [Web Search](#)

[Search](#) Sort [Most Popular](#) ▾ [View Options](#) [Developer Tool](#) [Help](#)

My Enhancements

All Enhancements

Business & Finance (3)

Education (1)

Entertainment (16)

Health (4)

Home & Living (6)

Information (15)


Local (7)

News & Blogs (3)

Shopping (4)

Social (4)


Sports (7)



Y Yahoo! Search Video Player ✕ [Remove](#)

Developed by: [Yahoo! Search](#)


Discover new content for your media-hungry mind! Watch music videos, news, sports, comedy, movie previews and more - right from your search page.



in LinkedIn Public Profile ✕ [Remove](#)

Developed by: [Yahoo! user](#)


Render LinkedIn Public Profiles in a richer and more compelling format within Yahoo! Search results. Currently, this plug-in is compatible only with the public profiles of LinkedIn members who have claimed a custom public profile URL.



*** Yelp - Local Business Ratings, Reviews & Info** ✕ [Remove](#)

Developed by: [Yahoo! user](#)

Useful info about local businesses from Yelp.com. Get a preview of photos, reviews, rating, address, phone number & more while you search.



Y Yahoo! Local Enriched Businesses ✕ [Remove](#)

Developed by: [Yahoo! Local Team](#)

The Yahoo! Local Enriched Businesses application makes searching for local businesses faster, easier, and more fun by pulling in key info and reviews.

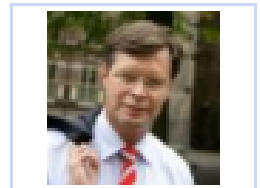
Example apps

- LinkedIn
 - hCard plus feed data

[Jan Peter Balkenende - Prime Minister at Netherlands](#)

- Location: The Hague Area, Netherlands
- Current: Prime Minister, Netherlands

www.linkedin.com/in/janpeterbalkenende - [Cached](#)




- Cr
 - CC in RDFa

[Ben Adida](#)

Ben Adida. ben@adida.net. Projects Papers Talks CV/bio B(en)log ... © **Ben Adida**, distributed under a Creative Commons license. ...

ben.adida.net - [Cached](#)

 Attribution - Creative Commons 

 Attribution - Creative Commons



See the [detailed CC Attribution License](#)

Example apps. II.

- Other me by Dan Brickley
 - Google Social Graph API wrapped using a Web Service



[Alexandre Passant](#)

Alexandre Passant. Home. About me. Blog. testtt. RSS. SDoW2008 tagcloud ... Copyright © 2007 **Alexandre Passant** • Powered by WordPress • Using Silhouette ...
[apassant.net/blog](#) - 63k - [Cached](#)

 [apassant.net myopenid.com](#) - otherme ▾ 

[About me : Alexandre Passant](#)

I am **Alexandre Passant**, a Ph.D. student at LaLIC, Université Paris-Sorbonne. ... Copyright © 2007 **Alexandre Passant** • Powered by WordPress • Using Silhouette ...
[apassant.net](#) - [Cached](#)

 [apassant.net](#) - otherme ▾ 

[Flickr: Alexandre Passant](#)

Flickr is almost certainly the best online photo management and sharing ... **Alexandre Passant's** public groups. American Punk/Hardcore Archive (78-91) Dead Computers ...
[www.flickr.com/people/terraces](#) - [Cached](#)

 [apassant.net apassant.net](#) - otherme ▾ 

Google's Rich Snippets

- Shares a subset of the features of SearchMonkey
 - Encourages publishers to embed certain microformats and RDFa into webpages
 - Currently reviews, people, products, business & organizations
 - These are used to generate richer search results

[Drooling Dog Bar B Q - Colfax, CA](#)

★★★★☆ 15 reviews - Price range: \$\$

Drooling Dog has some really good BBQ. I had the pulled pork sandwich, Drooling Dog BBQ is a great place to stop at on your way up the hill to Tahoe ...

www.yelp.com/biz/drooling-dog-bar-b-q-colfax - 75k - [Cached](#) - [Similar pages](#)

- SearchMonkey is customizable
 - Developers can develop applications themselves
- SearchMonkey is open
 - Wide support for standard vocabularies
 - API access

Yahoo BOSS: Build your Own Search Service

- Ability to re-order results and blend-in additional content
- No restrictions on presentation
- No branding or attribution
- Access to multiple verticals (web search, image, news)
- 40+ supported language and region pairs
- Pricing (BOSS)
 - Pay-by-usage
 - 10,000 queries a day still free
 - Serve any ads you want
- For more info, <http://developer.yahoo.com/search/boss/>

BOSS for structured data

- Simple HTTP GET calls, no authentication
 - You need an Application ID: register at developer.yahoo.com/search/boss/
- `http://boss.yahooapis.com/ysearch/web/v1/{query}?appid={appid}&format=xml&view=searchmonkey_feed`
- Restrict your query using special words
 - `searchmonkey:com.yahoo.page.uf.{format}`
 - {format} is one of hcard, hcalendar, tag, adr, hresume etc.
 - `searchmonkey:com.yahoo.page.rdf.rdfa`

Demo: resume search

- Search pages with resume data and given keywords
 {keyword} searchmonkey:com.yahoo.page.uf.hresume
- Parse the results as DataRSS (XML)
- Extract information and display using YUI

Demo: resume search

This is an example of what a developer can build with SearchMonkey's semantic data and the BOSS APIs. It took about four hours with over half of the time spent wondering why I chose PHP as the rendering engine. The data used in this demo is data publicly available through the BOSS APIs and based off of [hResume](#). Some of the sites using hResume today include LinkedIn.com and plaxo.com.



Resumes

Name	Company	Title	StartDate	Education	Graduation	Photo
Bill Gates	Bill & Melinda Gates Foundation	Co-chair	2000		2008-07-01	
bill gates	microsoft	Owner		California University of Pennsylvania	1981-12-31	
Les Bill Gates	Editaddition (Self-employed)	Independent Professional	2007-11-01	University of Exeter	1973-12-31	
0.Bill Gates						
Gal Horvitz	pnmssoft	CEO	1994	Technion-Machon Technologi Le' Israel	1996-12-31	
Gates Bill						
Michael Morgenstern, MD	Medwiser	Founder & CEO	2005-06-01	Drexel University College of Medicine	2009-12-31	
Michael Bacino	Thought Equity Motion	Content Sales Manager	2007-05-01	Indiana University Bloomington	1995-12-31	
Bill Yerkes	Solaix	Owner and Chief Technology Officer	2003	Stanford University	1957-12-31	
George Rajna	IBM Global Services	Microsoft .NET Solutions Architect	1999-08-01	EÄ¶tvÄ¶s LorÄ¶nd TudomÄ¶nyegyetem	1999-08-01	

[Show me the BOSS query](#)

Match case

Done

 FoxyProxy: Patterns 

Yahoo Correlator

- Named entity recognition applied to Wikipedia
- Sentence-level indexing
- Rich visualization
 - Places on the map
 - Dates on timeline
 - Names of people, organizations in a graph
- <http://sandbox.yahoo.com/Correlator>

Demo: Correlator

Yahoo | My Yahoo! | Mail | More

Report it

Correlator from YAHOO! RESEARCH

Search



Wikipedia



Names



Places



Events



Concepts



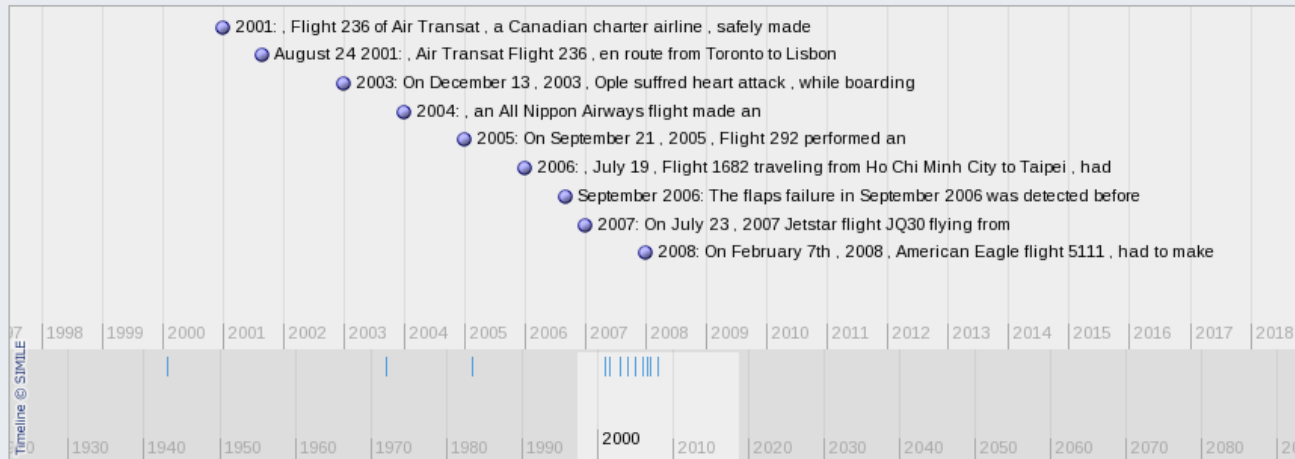
News



Answers

Events related to "emergency landing"

Timeline



Events in the timeline

1943

(From [WRAF Foulsham](#)) "Many aircraft made emergency landings at Foulsham, including USAAF B17 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."

(From [WFort Hertz](#)) "In **1943** and 1944 the primary purpose of Fort Hertz was to gather intelligence and to cover an airstrip which served as an emergency landing ground for planes flying The Hump from India to China over the eastern end of the Himalayas."

(From [WRAF Woodbridge](#)) "The airfield was constructed as an Emergency Landing Ground and was operational from **1943**."

1972

(From [WCondensation trap](#)) "The **1972** Movie Family Flight features a family forced to make an emergency landing in the Mexican desert."



(From [WLinda Kelsey](#)) "Kelsey's professional career began with stage appearances in her home of Minneapolis, Minnesota, with her good looks and striking mane of red hair winning her success that ultimately landed her in Los Angeles in **1972**, with appearances in small roles on television shows like Emergency! and The Rookies, and the television movie The Picture of Dorian Gray (1973)."

(From [WLand and Property laws in Israel](#)) "Lands so acquired would often be sold to the JNF. These regulations remained in place until **1972**."


Demo: Yahoo Quest




- Helping users find the right question to ask
- Navigate possible questions based on nouns and verbs frequently found in questions related to your keywords
- NLP parsing adapted to questions, inverted index and forward for quick counting
- Based on the Yahoo Answers collection of question/answer pairs
- <http://sandbox.yahoo.com/Quest>

Demo: Quest

SANDBOX Presented by Yahoo! Research [More Demos](#)  [More Experiments](#) 

YAHOO! QUEST

 Find questions about

Sort	Keywords	Questions
	More Specific	
<input type="checkbox"/>	old dalmatian	4
<input type="checkbox"/>	dalmatian puppy	2
<input type="checkbox"/>	dalmatian female	2
	Nouns	
<input type="checkbox"/>	dalmatians	19
<input type="checkbox"/>	dog	7
<input type="checkbox"/>	puppy	3
<input type="checkbox"/>	female	2
	Verbs	
<input type="checkbox"/>	breed	2
<input type="checkbox"/>	bite	2
<input type="checkbox"/>	train	2
<input type="checkbox"/>	feed	2
<input type="checkbox"/>	own	2
<input type="checkbox"/>	come	2
<input type="checkbox"/>	know	4
<input type="checkbox"/>	need	2
<input type="checkbox"/>

57 results

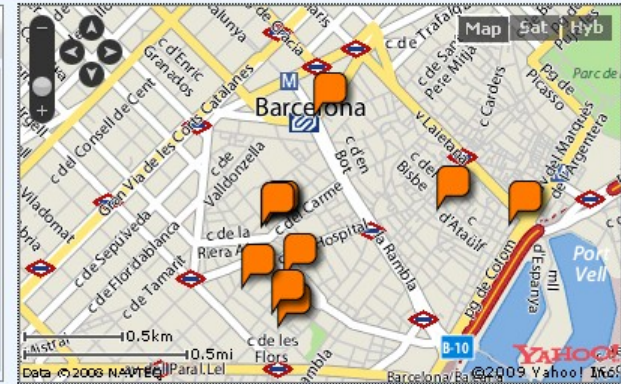
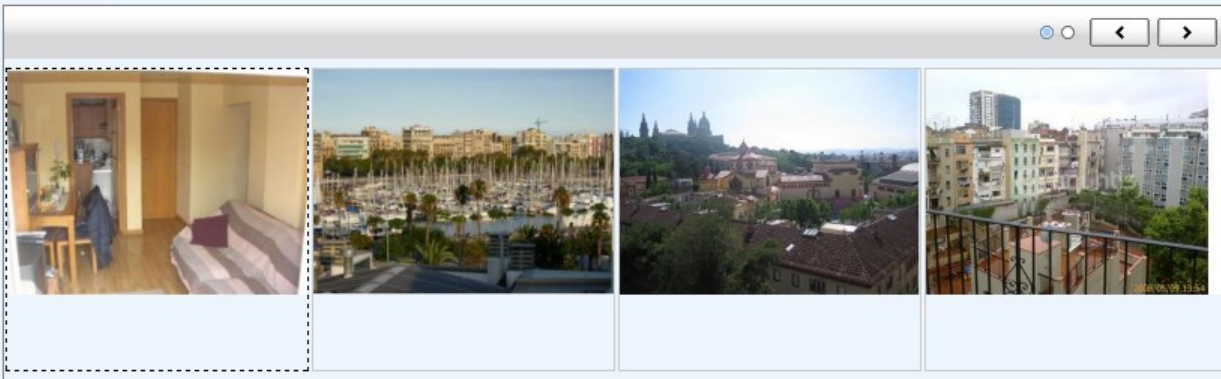
- [+ Why do firehouses always have dalmatians?](#)
In Dogs
- [+ Which breed of dogs is better german shepherd or dalmatian?](#)
In Dogs
- [+ how many cups of dogfood am I to give a 3 year old dalmatian?](#)
In Dogs
- [+ How can you make a Dalmatian mix stop biting you?](#)
In Dogs
- [+ how do you get your dalmatian to be nice to neighbors?](#)
In Dogs
- [+ where can i find a dalmatian puppy in reno nv?](#)
In Other - Business & Finance
- [+ where should i post my ad that says"dalmatian for sale" where Filipinos will surely buy it?](#)
In Dogs
- [+ How do i get mv 6 year old dalmatian to listen?](#)

Semantic bookmarking

- Extracting metadata from the user's Delicious profile
 - Metadata pulled from index and SearchMonkey data services
- Rich representation of bookmarks based on structured data
 - Tabular display
 - Sorting on attributes
 - Map
- Tracking changes in data
 - Alert me when the price drops below...
- Prototype: house search application

Demo: semantic bookmarking

Housing About



Data Table

title	price	address	size	link	previous
ático en venta en c/ hort de la bomba, 6, barcelona	590,000	c/ hort de la bomba, 6,08001,barcelona,Spai	290	www.idealista.com	0
Ático en Venta en Calle Aurora de Raval, Barcelona	145,000	Calle Aurora,Barcelona,Spain	60	www.fotocasa.es	0
Ático en Venta en Calle Padilla de Sagrada Familia - Fort Pienc, Barcelo	399,000	Calle Padilla,Barcelona,Spain	95	www.fotocasa.es	0
Ático en Venta en Calle San Isidre 2 de Poble Sec - Font de la Guatlla, Ba	350,000	Calle San Isidre 2 de Poble Sec - Font,Barce	78	www.fotocasa.es	0
penthouse for sale in st. carretes, 50, barcelona	325,000	st. carretes, 50,08001,barcelona,Spain	82	www.idealista.com	0
Ático en Venta en Calle Peu de la Creu 21 de Raval, Barcelona	330,000	Calle Peu de la Creu 21,Barcelona,Spain	70	www.fotocasa.es	0
Ático en Venta en Calle Nápoles de Sagrada Familia - Fort Pienc, Barcelo	324,000	Calle Nápoles,Barcelona,Spain	75	www.fotocasa.es	0
Ático en Venta en Calle Sant Miquel de Barceloneta - Born - Sta. Caterina	319,000	Calle Sant Miquel,Barcelona,Spain	67	www.fotocasa.es	0
ático en venta en c/ carretes, 8, barcelona	189,000	c/ carretes, 8,08001,barcelona,Spain	50	www.idealista.com	196,000
Ático en Venta en Calle Merce de Gòtic, Barcelona	420,000	Calle Merce,Barcelona,Spain	84	www.fotocasa.es	0
ático en venta en c/ joaquin costa, 3, barcelona	345,000	c/ joaquin costa, 3,08001,barcelona,Spain	80	www.idealista.com	0
penthouse for sale in st. peu de la creu, 21, barcelona	330,000	st. peu de la creu, 21,08001,barcelona,Spain	70	www.idealista.com	0

idealista.com  fotocasa.es

Summary

- Semantic Search impacts every step of the process
 - Document processing
 - Human effort: developers, site owners
 - Machine effort: NLP, IE
 - Query intent analysis
 - Indexing and ranking
 - Interface
 - Both query input and result presentation
- This presentation focused on the Semantic Web story
 - Did not discuss pure NLP search engines such as PowerSet, Hokia, TrueKnowledge

Key areas of future work and emerging topics

- (Semi-)automated ways of metadata creation
 - How do we go from 5% to 95%?
- Data quality
 - Can we trust statements people make about each other's data?
- Reasoning
 - To what extent is reasoning useful?
- Scale
 - What is between databases and IR engines?
- Solving the ontology problem
 - How do we get people to reuse vocabularies?
- Semantic ads
- Personalization
- Mobile

Contact

- Peter Mika
 - pmika@yahoo-inc.com
- SearchMonkey
 - developer.yahoo.com/searchmonkey/
 - mailing lists
 - searchmonkey-developers@yahoogroups.com
 - searchmonkey-siteowners@yahoogroups.com
 - forums
 - <http://suggestions.yahoo.com/searchmonkey>



SearchMonkey