Web Data Management: Powering the New Web

Raghu Ramakrishnan
Chief Scientist for Audience, Yahoo!
Research Fellow, Yahoo! Research
(On leave, Univ. of Wisconsin-Madison)
Outline

• **Trends in Search and Information Discovery**
  – Move towards task-centricity
  – Need to interpret content

• **Evolution of Online Communities**
  – Social Search
  – PeopleWeb
  – Community Information Management

• **Web Data Infrastructure**
  – Massively distributed computing
  – Hosted services
  – Heterogeneous content
Further Reading

- Content, Metadata, and Behavioral Information: Directions for Yahoo! Research, *The Yahoo! Research Team*, IEEE Data Engineering Bulletin, Dec 2006 (Special Issue on Web-Scale Data, Systems, and Semantics)


- Towards a PeopleWeb, R. Ramakrishnan and A. Tomkins, IEEE Computer, August 2007 (Special Issue on Web Search)
Trends in Search
Search and Content Supply

• Premise:
  – People don’t want to search
  – People want to get tasks done

I want to book a vacation in Tuscany.

Broder 2002, A Taxonomy of web search
Y! Shortcuts

Putnam Toyota: Toyota Corolla
www.putnamtoyota.com | Local Info  The Putnam Automotive Group has more than 37 years experience serving the public, and we invite you to become yet another satisfied customer.

Buy a New Toyota Corolla below Invoice
cbi.carsbelowinvoice.com  Why pay sticker? You can buy a new car below invoice.

Used Toyota Corolla Results near Palo Alto, CA - Map All Results
1999 Toyota Corolla - $10,000 - 81,571 miles
2004 Toyota Corolla - $12,991 - 28,158 miles
2002 Toyota Corolla - $9,994 - 54,860 miles

Find Used TOYOTA COROLLA Cars in Palo Alto, California on Yahoo!
Find used TOYOTA COROLLA cars in Palo Alto, California on Yahoo! Autos. Review used TOYOTA COROLLA pictures, price, options and features. Find the best deal and respond to the listings online immediately.
autos.yahoo.com/used-cars/ca-palo_alto-toyota-corolla.html - More from this site - Save

Magnussen's Toyota & Scion of Palo Alto
New and used cars, parts and service.
Category: California > Palo Alto > Automotive Dealers > Toyota
“seafood san francisco”

Category: restaurant
Location: San Francisco

Reserve a table for two tonight at SF’s best Sushi Bar and get a free sake, compliments of OpenTable!

Category: restaurant
Location: San Francisco

Alamo Square Seafood Grill – (415) 440-2828
803 Fillmore St, San Francisco, CA – 0.93mi – map

Category: restaurant
Location: San Francisco
Steps to Task-Centricity

- **Information integration**
  - Information extraction
  - Schema normalization

- **Structure**
  - Extract and exploit

---

hotel near leicester square

Search

Welcome to The Savoy
Located on The Strand in the heart of the West End theatre district,
Semantic Structure – Not Easy

- Colorado, Texas
- Oregon, Alaska
- Peru, Bolivia
- Peru, Argentina
- Washington, Nevada County, California
- Bush, Cheney & Rice, WA state
How Do We Circumvent?

• Unleash community computing
  – Social structure
  – Incentive mechanisms

• More in a moment
Evolution of Online Communities
Rate of content creation

• Estimated growth of content
  – Published content from traditional sources: 3-4 Gb/day
  – Professional web content: ~2 Gb/day
  – User-generated content: 8-10 Gb/day
  – Private text content: ~3 Tb/day (200x more)
  – Upper bound on typed content: ~700 Tb/day
Metadata

• Estimated growth of metadata
  – Anchortext: 100Mb/day
  – Tags: 40Mb/day
  – Pageviews: 100-200Gb/day
  – Reviews: Around 10Mb/day
  – Ratings: <small>

  Drove most advances in search from 1996-present

  Increasingly rich and available, but not yet useful in search

This is in spite of the fact that interactions on the web are currently limited by the fact that each site is essentially a silo
PeopleWeb: Site-Centric → People-Centric

- **Common web-wide id for objects (incl. users)**
  - Even common attributes? (e.g., *pixels* for camera objects)
- As users move across sites, their personas and social networks will be carried along
- **Increased semantics on the web through community activity**
  (another path to the goals of the Semantic Web)

*(Towards a PeopleWeb, Ramakrishnan & Tomkins, IEEE Computer, August 2007)*
Social Search

• **Improve web search by**
  – Learning from shared community interactions, and leveraging community interactions to create and refine content
    • Enhance and amplify user interactions
  – Expanding search results to include sources of information (e.g., experts, sub-communities of shared interest)

Reputation, Quality, Trust, Privacy
Web Search Results for “Lisa”

- Latest news results for “Lisa”. Mostly about people because Lisa is a popular name
- 41 results from My Web!
- Web search results are very diversified, covering pages about organizations, projects, people, events, etc.
My Web 2.0 Search Results for “Lisa”

Related tags: lisa, 2005, usenix, blog, More...

My Community’s Results (41)

1. **SAGEwire LISA 2005 Blog**
   - system administrator n.a. system administrator is one who, as a primary job function, manages computer and network systems on behalf of another, such as an employer or client. LISA 2005 Blog. David N.
   - Saved by: Zhichen Xu and 1 other 10:05 AM PST - Details
   - Tags: blog, lisa, usenix
   - http://sagewire.sage.org/article.pl?sid=05/08/16/0237249 - Save

2. **The LISA ‘05 Conference Blog**
   - Behind the Scenes at the Largest System Administration Conference in the World. November (there’s another burning-hot news nugget that will be posted on the official conference site today)
   - Saved by: Jianchang (JC) Mao 10:09 AM PST - Details
   - Note: Official Lisa conference blog. A lot of good stuff behind the scene.
   - Tags: 2005, blog, lisa, usenix
   - http://blog.lisaconference.org/ - Save

3. **19th Large Installation System Administration Conference**
   - Administrators of all specialties and levels of expertise meet at LISA to exchange ideas, sharpen old skills, learn new techniques, debate current issues, and meet colleagues and friends.
   - Saved by: Jianchang (JC) Mao on November 26, 2005 - Details
   - Note: Qi is giving a keynote speech. Excellent system conference. Dates: Dec 4-9, 2005
   - Tags: conference, lisa, systems
   - http://www.usenix.org/events/lisa05 - Save

Excellent set of search results from my community because a couple of people in my community are interested in Usenix Lisa-related topics.
Google Co-Op

Query-based direct-display, programmed by Contributor

This query matches a pattern provided by Contributor...

...so SERP displays (query-specific) links programmed by Contributor.

Users “opts-in” by “subscribing” to them

Subscribed Link
Challenges in Tag-Based Search

• How do we use these tags better?
• How do you cope with spam?
• What’s the ratings and reputation system?

• The bigger challenge: where else can you exploit the power of the people?
• What are the incentive mechanisms?
  – Luis von Ahn (CMU): The ESP Game
Timely Answers

77% of answers provided within 24h

- No effort to answer each question
- No added experts
- No monetary incentives for enthusiasts
Interesting Problems

• Question categorization
• Detecting undesirable questions & answers
• Identifying “trolls”
• Ranking results in Answers search
• Finding related questions
• Estimating question & answer quality

(Byron Dom: SIGIR talk)
Supplying Search Content

• As information discovery and search become task-oriented, we need to find ways to create semantically rich summaries that address the user’s information needs. Three ways to do this:
  – Editorial, Extraction, UGC
  • Opportunity to focus creation of structured UGC feeds directly into this growing need!

Challenge: Design social interactions that lead to creation and maintenance of high-quality structured content
Better Search via Information Extraction

- Extract, then exploit, structured data from raw text:

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Select Name
From PEOPLE
Where Organization = 'Microsoft'

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>Founder</td>
<td>Free Soft..</td>
</tr>
</tbody>
</table>

Bill Gates
Bill Veghte

(from Cohen's IE tutorial, 2003)
DBLife

- Integrated information about a (focused) real-world community
- Collaboratively built and maintained by the community
- Semantic web via extraction & community
The DBLife Portal

- Faculty: AnHai Doan & Raghu Ramakrishnan
- Students: P. DeRose, W. Shen, F. Chen, R. McCann, Y. Lee, M. Sayyadian
- Prototype system up and running since early 2005
- Plan to release a public version of the system in Spring 2007
- 1164 sources, crawled daily, 11000+ pages / day
- 160+ MB, 121400+ people mentions, 5600+ persons
- See DE overview article, CIDR 2007 demo
DBLife

Data Sources

- Click on the + to show details, and the - to hide them again
- Colloquia
- Conference homepage
- Database group page
- Dbworld
- Department homepage
- Event

Crawled daily, 11000+ pages = 160+ MB / day
Entity Resolution

Selected Publications

- Managing Map versus Local Data in Distributed Geographical Information Systems. P. McCann, I. Hlava, W. Pugh, and J. Wolf. VLDB-05.
- VLDB-05.
- VLDB-05.
- Semantic Integration Research in the Database Community.

co-authors = A. Doan, Divesh Srivastava,

Raghu Ramakrishnan
Resulting ER Graph
Challenges

• Extraction
  – Domain-level vs. site-level
  – Blending extraction with other sources (feeds, wiki-style user edits)
  – Compositional, customizable approach to extraction planning
    • Cannot afford to implement extraction afresh in each application!

• Maintenance of extracted information
  – Managing information Extraction
  – Incremental maintenance of “extracted views” at large scales
  – Mass Collaboration—community-based maintenance

• Exploitation
  – Search/query over extracted structures in a community
  – Search across communities—semantic web through the back door!
  – Detect interesting events and changes

Right in the sweet spot of the “Data to Knowledge” thrust of NSF’s CDI initiative!
Web Data Management: Massively Distributed Hosted Systems
Two Key Subsystems

- **Serving system**
  - Takes queries and returns results
- **Content system**
  - Gathers input of various kinds (including crawling)
  - Generates the data sets used by serving system
- Both highly parallel

Goal: scaleup.
Hardware increments support larger loads.

Serving System

Goal: speedup.
Hardware increments speed computations.

Data sets

Logs

Users

Web sites

Data updates

(Courtesy: Raymie Stata)
A Case for Hosted Infrastructure

• What does it take to get the Next Great Thing off the ground?

• **Now:**
  – Set up multiple replicas of a clustered data store
  – Set up a system for indexing
  – Set up a system for caching
  – Set up auxiliary DBMS instances for reporting, etc.
  – Set up the feeds and messaging between them
  – Write the application logic
  – Fairly complex system at first line of new code

• **Our vision:**
  – Write the application logic
  – Use a hosted infrastructure to store and query your data

  – Or, as Joshua Shachter puts it: “The next cool thing shouldn’t take a team of 30, it should be three guys, PHP and a long weekend”
The PNUTS Project

CREATE TABLE Parts (ID VARCHAR, StockNumber INT, Status VARCHAR ...)

A 42342 E
B 42521 W
C 66354 W
D 12352 E
E 75656 C
F 15677 E

A 42342 E
B 42521 W
C 66354 W
D 12352 E
E 75656 C
F 15677 E
Asynchronous replication
Basic consistency model

- Record lifecycle
  1. Record inserted with a given primary key
  2. Record’s non-primary key attributes updated
     - Primary key cannot be updated
  3. Record deleted
  4. Another record with the same primary key may subsequently be inserted

- What happens to a record with primary key “Brian”?
The Big Picture: Sherpa Data Services

Applications

**YCA: Authorization**

**PNUTS Services**
- Query planning and execution
- Index maintenance

**Distributed infrastructure for tabular data**
- Data partitioning
- Update consistency
- Replication

**YDHT FS**
- Hash tables

**YDHT FS**
- Ordered tables

**YMB**
- Pub/sub messaging

**Zookeeper**
- Consistency service
Data Analysis Platforms

• Understanding online communities, and provisioning their data needs
  – Exploratory analysis over massive data sets
    • Challenges: Analyze shared, evolving social networks of users, content, and interactions to learn models of individual preferences and characteristics; community structure and dynamics; and to develop robust frameworks for evolution of authority and trust; extracting and exploiting structure from web content …

• Examples:
  – Bigtable, Map-Reduce, Hadoop, PIG
The Bigger Picture

- Software-as-a-service
  - E.g., Salesforce.com
- Hosted data systems
  - Amazon’s S3 and EC2
- Web application development
  - Ning, Ruby-on-rails
- Change tracking
  - Stream management
Grand Challenge

• How to maintain and leverage structured, integrated views of web content
  – Web meets DB … and neither is ready!
    • Interpreting and integrating information
      – Result pages that combine information from many sites
    • Scalable serving of data/relationships
      – Multi-tenancy, QoS, auto-admin, performance
  – Beyond search—web as app-delivery channel
    • Data-driven services, not DBMS software
      – Customizable hosted apps!
    • Desktop ➔ Web-top