Yahoo! Labs

Silicon Valley

Barcelona, Spain

Bangalore, India

New York, NY

Burbank, CA

Haifa, Israel

Santiago, Chile
• Search Technologies
• Community Systems
• Machine Learning
• Computational Advertising
• Microeconomics and Social Systems
Open platforms

• Open source cloud computing infrastructure
  – Hadoop
  – Pig – Web-scale data management

• BOSS – Build your Own Search Service

• Y!OS: Build your Own Social Application (to be released soon)
  – OpenSocial initiative to provide access to Y!’s social network
Pig

- Chris Olston, Ben Reed, Utkarsh Srivastava
- Map-reduce
  - scalable, fault-tolerant shared-nothing query processor
  - limited data flow structure -> lots of hacking (e.g., join, chain N operators)
- Pig: aim for best of both worlds (map-reduce + SQL-like)
  - high-level dataflow programming model (algebra)
  - compiles into map-reduce sequences
- Sawzall (Google), SCOPE (Microsoft), JAQL (IBM), HIVE (Facebook)
PSOX: Information Extraction

• Philip Bohannon, Raghu Ramakrishnan, Nilesh Dalvi, Srujana Merugu, Ashwin Majchanawalla, Sathiya Keerthi Selvaraj, Cong Yu

• Cost of data is low but cost of supervision, especially large, high-quality training sets, is high

• Challenges
  – Rapid training iteration
  – Tolerance for low-quantity/quality supervision

• Solutions
  – Efficient and effective extraction management system (algebra-based framework)
  – Scalable machine learning extraction prioritization algorithms

• Application to Y!Local
COKE: Content Optimization

- Deepak Agrawal, Wei Chu, Anirban Dasgupta, Raghu Ramakrishnan, Bee-Chung Chen, Pradheep Elango, Seung-Taek Park

- Optimize Y!FrontPage
  - Maximize CTR

- Choose best 4 of 16 editorially selected news items

- Online item scoring per user segment
  - Online statistical model and multi-armed bandit streaming algorithms to learn best items

- 30+% CTR lift compared to current editorial serving
Computational Advertising


- Ad Selection and Ranking for Online Advertising
  - Publishers, Advertisers, Users in an ad exchange network
  - Find the "best ad" between a given user in a given context: contextual event
  - 3 advertising methods
    - Sponsored Search
    - Contextual advertising
    - Display advertising (banner ads)
  - Algorithms
    - guaranteed delivery
    - supply/demand forecasting
• Sihem, David Pennock, Sebastien Lahaie

• An open standard for advertisers (& publishers) to describe their campaign goals that is
  – expressive -- lets advertisers state different value for different (bundles of) contextual events
  – efficient -- is amenable to tractable allocation and pricing algorithms
    – Open and extensible – builds on OpenX

• XML syntax to express nesting contextual events and help target more specific population segments
<campaign type="DA" adv="156" id="adv156-1" budget="$1000">
  <description>sale to young adults</description>
  <creative>Buy one get one free</creative>
  <expr>
    <dimension name="age" op="=" value="[25-34]"/>
    <bids>
      <cpm val="0.4" min="0" max="500"/>
    </bids>
  </expr>
</campaign>

Other dimensions:
<dimension name="location" op="=" values="{CA, FL, WI, IL, TX}"/>
<dimension name="time" op="=" values="{am, pm}"/>
<dimension name="day" op="=" values="[Fri-Sun]"/>
<dimension name="position" op="=" value="south"/>
<creative>http://www.nike-camp.com/newSneakers.jpg</creative>
Social Search and Recommendations

• Sihem and Cong Yu + Julia Stoyanovich (Columbia), Jian Huang (Penn State)

• Integration of social sites and content sites
  – NY Times: “send to a friend”, “bookmark in delicious”,
  – OpenSocial and OpenID to enable network integration

• Dichotomy between Information Retrieval (search query) and Information Recommendation (users’ social activities)

• New information discovery paradigms:
  – Search and recommendation
    • I would like to travel to Barcelona
    • Which Web sites are hot today?
  – Leverage multiple networks
    • I ask Divesh about a dim sum recommendation, Julia about a Tacos place and Cong on recommending a football event
• Tags are not always a good model for user interest
  – (Custom) tags are sparse: long tail problem
  – Editorial tags are abundant but do not leverage tags
  – Some tags are meaningless: e.g., to-read
  – Two tags may relate to same topic: e.g., honeymoon and sunset

• Topics
  – Model users’ interests and items’ content
  – Build shared-interest networks
  – Search for content: “Romantic Beach” destination

• Co-occurrence analysis, association rules, EM
  – We use LDA (Latent Dirichlet Allocation)

• Application to Y!Travel and del.icio.us
Sample User Profiles

User: alex_mc*

Tags: architecture, art, sightseeing, weekend, romantic, road+trip, honeymoon, budget, temples, canal, ...

Tagged cities: Bangkok, Nice, Venice, Barcelona

User: mum_bhai

Tags: wine+tasting, shopping, beach, casino, adventure, outdoor+activities ...

Tagged cities: Las Vegas, Boston, Santa Monica, Laguna Beach, Los Angeles, Ojai, Napa, ...

R: Romantic, W: Seaside, A: Art, N: Nightlife
Sample Items

**Topic: Seaside/Water related**
- **Tags:** beach, scuba, summer, diving, fishing, snorkeling, island, lake ...
- **Top cities:** San Diego, Honolulu, Chicago, Miami, Seattle ...

**Topic: Romantic/Luxury**
- **Tags:** romantic, 4-star, shopping, spa, golf, luxury, honeymoon ...
- **Top cities:** Cancun, San Juan, Grand Canyon, Bangkok ...

**Topic: Nightlife, City-life**
- **Tags:** nightlife, gambling, wine, drinking, exciting, casino ...
- **Top cities:** Las Vegas, New York City, Los Angeles, San Francisco ...

**Topic: Arts/Historical/Cultural**
- **Tags:** architecture, sightseeing, art, history, culture, cathedral, castle ...
- **Top cities:** Paris, London, Boston, Istanbul, Amsterdam, Rome, Hong Kong ...

San Diego  Las Vegas  Paris  Cancun
• Incremental network computation daily

del.icio.us backup database

MySQL Extract

research9

MySQL Load

distributed analysis and index / view generation

Garcon database