Integrating Social with Search

Edward Chang
Director, Google Research, Beijing
Related Papers

• **AdHeat (Social Ads):**

• **UserRank:**

• **Large-scale Collaborative Filtering:**
Web 1.0
Web 2.0 --- Web with People
Outline

• Search + Social Synergy
• Search → Social
• Social → Search
• Scalability
Google Q&A (Confucius)

• Developed from 2007 till now @ Beijing
• Launched in more than 60 courtiers
  – Russia
  – HK
  – Southeast Asia
  – Arab World
  – Sub-Saharan Africa (Baraza)
Query: What are must-see attractions at Yellowstone

What are must-see attractions at Yellowstone

Three Must See Attractions at Yellowstone National Park « The View ...
Jan 15, 2008 ... Smith presents Three Must See Attractions at Yellowstone National Park posted at The View West. Interested in Yellowstone National Park? ... theviewwest.com/2008/01/15/three-must-see-attractions-at-yellowstone-national-park/ - 26k - Cached - Similar pages

Three Must See Attractions At Yellowstone National Park
Jan 15, 2008 ... Three Must See Attractions At Yellowstone National Park ezinearticles.com/?Three-Must-See-Attractions-At-Yellowstone-National-Park&id=929265 - 47k - Cached - Similar pages

Yellowstone National Park: Top Ten Attractions
YELLOWSTONE NATIONAL PARK by Yellowstone Net. Top 10 Things to See in YNP What are the "Must See" attractions to view in Yellowstone? Start here! ... www.yellowstone.net/topten.html - 16k - Cached - Similar pages

Yellowstone Must-see Attractions
Yellowstone's Must-See Attractions The locations of all sites listed below are shown on the map that you receive as you enter the park. ... www.geocities.com/dmontei/must_see.html - 8k - Cached - Similar pages

What to See in Yellowstone
Must-See Attractions -- Text Only Version · Upper Geyser Basin and Old Faithful · Grand Canyon of the Yellowstone · Fountain Paint Pots Trail · Wildlife ... www.geocities.com/dmontei/whattosee.html - 10k - Cached - Similar pages

More results from www.geocities.com »

Must See in Yellowstone National Park
Query: What are must-see attractions at Yellowstone

At first glance, Mammoth Hot Springs appear as a frozen waterfall. Large terraces abound while being connected by trickling water. The hot acidic water from the thermal aspect below ascends through ancient limestone deposits in the area. As the water dissolves the limestone, it is carried to the surface. When the suspension cools and becomes less acidic at the surface it forms the pools and the cascading features. This area is truly an amazing and dynamic work of art.

Wildlife
Query: What are must-see attractions at Yosemite
Query: What are must-see attractions at Beijing

Hotel ads

**Hotel ads**

**Beijing Travel Guide**
- Beijing Imperial Palace
- Beijing Olympic Park
- Beijing University

**Beijing Attractions**
- Beijing Forbidden City
- Beijing Temple of Heaven
- Beijing Great Wall

**Beijing Food**
- Beijing Roast Duck
- Beijing Beihai Park
- Beijing Tiananmen Square

**Beijing Transportation**
- Beijing Subway
- Beijing Bus
- Beijing Taxi

**Beijing Shopping**
- Beijing Tiananmen Market
- Beijing Wangfujing Street
- Beijing Sanlitun Village

**Beijing Nightlife**
- Beijing Houhai湖泊
- Beijing Sanlitun
- Beijing 798 Art District
Search Quality at Stake.

61 countries have Q&A or advanced forums as top 10 most clicked destination
(out of 115 countries with more than 1M session)
SNS & Mobile Also Need Q&A

• Social Networks
  – Difficult to find user intent to match ads
  – Q&A is a perfect app to learn users’ problems

• Mobile Search
  – Voice is the most convenient user interface
  – Succinct search result (or rich snippets) is desirable
Confucius: Google Q&A
Providing High-Quality Answers in a Timely Fashion

- Trigger a discussion/question session during search
- Provide labels to a post (semi-automatically)
- Given a post, find similar posts (automatically)
- Evaluate quality of a post, relevance and originality
- Evaluate user credentials in a topic sensitive way
- Route questions to experts
- Provide most relevant, high-quality content for Search to index
- Generate answers using NLP
Confucius: Google Q&A
Providing High-Quality Answers in a Timely Fashion

- Trigger a discussion/question session during search
- Provide labels to a question (semi-autonomously)
- Given a question, find similar questions (automatically)
- Evaluate quality of an answer, relevance and originality
- Evaluate user credentials in a topic sensitive way
- Route questions to experts
- Provide most relevant, high-quality content for Search to index
- Generate answers using NLP
Label suggestion using LDA algorithm.

- Real Time topic-to-topic (T2T) recommendation using LDA algorithm.
- Gives out related high quality links to previous questions before human answer appear.
Collaborative Filtering

Based on *membership* so far, and *memberships* of others

Predict further *membership*
FIM-based Recommendation

To grow the base, we need association rules

- An association rule: $a, b, c \rightarrow d$
- A Bayesian interpretation: $P(d | a, b, c) = \frac{N(a, b, c, d)}{N(a, b, c)}$
- The key is to count the occurrences (support) of itemsets $N(\ldots)$
Distributed Latent Dirichlet Allocation (LDA)

- Search
  - Construct a latent layer for better for semantic matching
- Example:
  - iPhone crack
  - Apple pie

Other Collaborative Filtering Apps
- Recommend Users → Users
- Recommend Music → Users
- Recommend Ads → Users
- Recommend Answers → Q

1 recipe pastry for a 9 inch double crust
9 apples, 2/1 cup, brown sugar

How to install apps on Apple mobile phones?

2010-12-13
Latent Dirichlet Allocation [D. Blei, M. Jordan 04]

- $\alpha$: uniform Dirichlet $\phi$ prior for per document $d$ topic distribution (corpus level parameter)
- $\beta$: uniform Dirichlet $\phi$ prior for per topic $z$ word distribution (corpus level parameter)
- $\theta_d$ is the topic distribution of document $d$ (document level)
- $z_{dj}$ the topic if the $j^{th}$ word in $d$, $w_{dj}$ the specific word (word level)
Combination of Collaborative Filtering Model (CCF)

[W.-Y. Chen, et al, KDD2008]

Communities

\[ P(c) \rightarrow z \rightarrow u \]
\[ P(z|c) \]
\[ P(u|z) \]

Communities

\[ P(c) \rightarrow z \rightarrow d \]
\[ P(z|c) \]
\[ P(d|z) \]

Communities

\[ c \rightarrow z \rightarrow P(c) \]
\[ P(z|c) \]
\[ P(d|z) \]
\[ P(u|z) \]

**Wisdom Keynote**
Confucius: Google Q&A

- Trigger a discussion/question session during search
- Provide labels to a post (semi-automatically)
- Given a post, find similar posts (automatically)
- Evaluate quality of a post, relevance and originality
- Evaluate user credentials in a topic sensitive way
- Route questions to experts
- Provide most relevant, high-quality content for Search to index
- NLQA
UserRank

- Rank users by quantity (number of links) and quality (weights on links) of contributions

Quality include:
- **Relevance.** Is an answer relevant to the Q? Measured by KL divergence between latent-topic vectors of A and Q
- **Coverage.** Compared among different answers
- **Originality.** Detect potential plagiarism and spam
- **Promptness.** Time between Q and A posting time
Outline

• Search + Social Synergy
• Social → Search
• Search → Social
• Scalability
Outline

• Search + Social Synergy
• Social → Search
• Search → Social
• Scalability
Social?

- Connecting to friends
- Knowing what friends are up to
- Connecting to strangers
  - Dating, Gaming
  - Shopping

- Making recommendations based on activities
User Latent Model

- $\alpha$: uniform Dirichlet $\phi$ prior for per user $u$ interest distribution (population level parameter)
- $\beta$: uniform Dirichlet $\phi$ prior for per interest $z$ activity distribution (population level parameter)
- $\theta_u$ is the interest distribution of user $u$ (user level)
- $z_{uj}$ the interest of the $j^{th}$ activity in $u$, $w_{uj}$ the specific activity (activity level)
Combinational Collaborative Filtering Model (CCF)
Outline

• Search + Social Synergy

• Social → Search
  – Mobilize users to improve search quality
  – Google Q&A, Facebook Like

• Search → Social
  – Use query log to help social
    • Activities → Interests → Social
    • Groupcom

• Scalability
# Prefixes

<table>
<thead>
<tr>
<th>SI prefix</th>
<th>Name</th>
<th>Power of 10 or 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>kilo</td>
<td>$10^3$ $/2^{10}$</td>
</tr>
<tr>
<td>M</td>
<td>mega</td>
<td>$10^6$ $/2^{20}$</td>
</tr>
<tr>
<td>G</td>
<td>giga</td>
<td>$10^9$ $/2^{30}$</td>
</tr>
<tr>
<td>T</td>
<td>tera</td>
<td>$10^{12}$ $/2^{40}$</td>
</tr>
<tr>
<td>P</td>
<td>peta</td>
<td>$10^{15}$ $/2^{50}$</td>
</tr>
<tr>
<td>E</td>
<td>exa</td>
<td>$10^{18}$ $/2^{60}$</td>
</tr>
<tr>
<td>Z</td>
<td>zetta</td>
<td>$10^{21}$ $/2^{70}$</td>
</tr>
<tr>
<td>Y</td>
<td>yotta</td>
<td>$10^{24}$ $/2^{80}$</td>
</tr>
</tbody>
</table>
# Prefixes

<table>
<thead>
<tr>
<th>SI prefix</th>
<th>Name</th>
<th>Power of 10 or 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>kilo</td>
<td>$10^3$ 2$^{10}$</td>
</tr>
<tr>
<td>M</td>
<td>mega</td>
<td>$10^6$ 2$^{20}$</td>
</tr>
<tr>
<td>G</td>
<td>giga</td>
<td>$10^9$ 2$^{30}$</td>
</tr>
<tr>
<td>T</td>
<td>tera</td>
<td>$10^{12}$ 2$^{40}$</td>
</tr>
<tr>
<td>P</td>
<td>peta</td>
<td>$10^{15}$ 2$^{50}$</td>
</tr>
<tr>
<td>E</td>
<td>exa</td>
<td>$10^{18}$ 2$^{60}$</td>
</tr>
<tr>
<td>Z</td>
<td>zetta</td>
<td>$10^{21}$ 2$^{70}$</td>
</tr>
<tr>
<td>Y</td>
<td>yotta</td>
<td>$10^{24}$ 2$^{80}$</td>
</tr>
</tbody>
</table>
## Prefixes

<table>
<thead>
<tr>
<th>SI prefix</th>
<th>Name</th>
<th>Power of 10 or 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>kilo</td>
<td>$10^3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{10}$</td>
</tr>
<tr>
<td>M</td>
<td>mega</td>
<td>$10^6$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{20}$</td>
</tr>
<tr>
<td>G</td>
<td>giga</td>
<td>$10^9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{30}$</td>
</tr>
<tr>
<td>T</td>
<td>tera</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{40}$</td>
</tr>
<tr>
<td>P</td>
<td>peta</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{50}$</td>
</tr>
<tr>
<td>E</td>
<td>exa</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{60}$</td>
</tr>
<tr>
<td>Z</td>
<td>zetta</td>
<td>$10^{21}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{70}$</td>
</tr>
<tr>
<td>Y</td>
<td>yotta</td>
<td>$10^{24}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$2^{80}$</td>
</tr>
</tbody>
</table>
More Data vs. Better Algorithms

Banko & Brill, 2001

Figure 2. Learning Curves for Confusable Disambiguation
More Data vs. Better Algorithms

Banko & Brill, 2001

Figure 2. Learning Curves for Confusable Disambiguation
More Data vs. Better Algorithms

Banko & Brill, 2001

Figure 2. Learning Curves for Confusable Disambiguation
More Data vs. Better Algorithms

Banko & Brill, 2001

Figure 2. Learning Curves for Confusable Disambiguation
User Latent Model

- $\alpha$: uniform Dirichlet $\phi$ prior for per user $u$ interest distribution (population level parameter)
- $\beta$: uniform Dirichlet $\phi$ prior for per interest $z$ activity distribution (population level parameter)
- $\theta_d$ is the interest distribution of user $u$ (user level)
- $z_{uj}$ the interest of the $j^{th}$ activity in $u$, $w_{uj}$ the specific activity (activity level)
LDA Gibbs Sampling: Inputs & Outputs

Inputs:
1. **training data**: users as bags of words
2. **parameter**: the number of topics

Outputs:
1. **model parameters**: a co-occurrence matrix of topics and words.
2. **by-product**: a co-occurrence matrix of topics and users.
Parallel Gibbs Sampling

**Inputs:**
1. **training data**: users as bags of words
2. **parameter**: the number of topics

**Outputs:**
1. **model parameters**: a co-occurrence matrix of topics and words.
2. **by-product**: a co-occurrence matrix of topics and users.
Observations

• Bottleneck: Communication
• Amdahl’s law caps speedup
• Words in a bag have no order
• Words on a computer node can be reordered
Example Bags / Node A

- Bag #1  w1, w2, w3, w1, w2, w3, w1, w2, w3
- Bag #2  w1, w2, w1, w2, w1, w2, w1, w2
- Bag #3  w3, w1, w3, w1, w3, w1, w3, w1

- Bundle #1  w1, w1, w1, w1, w1, w1, ...
- Bundle #2  w2, w2, w2, ...
- Bundle #3  w3, w3, w3, ...
# Two Nodes

<table>
<thead>
<tr>
<th>Node A</th>
<th>Node B</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>W2</td>
</tr>
<tr>
<td>W2</td>
<td>W3</td>
</tr>
<tr>
<td>W3</td>
<td>W1</td>
</tr>
</tbody>
</table>
Parallel Gibbs Sampling

Inputs:
1. **training data**: documents as bags of words
2. parameter: the number of topics

Outputs:
1. **model parameters**: a co-occurrence matrix of topics and words.
2. **by-product**: a co-occurrence matrix of topics and documents.
PLDA -- enhanced parallel LDA

- Take advantage of bag of words modeling: each Pw machine processes vocabulary in a word order
- Pipelining: fetching the updated topic distribution matrix while doing Gibbs sampling

![Diagram of pipeline-based Gibbs Sampling in PLDA](image)

Fig. 4: Pipeline-based Gibbs Sampling in PLDA*. (A): $t_s \geq t_f + t_u$. (B): $t_s < t_f + t_u$. 

WISE Keynote
Speedup

1,500x using 2,000 machines
Lessons Learned

• Bottleneck Matters
• Inter-iteration Matters
MapReduce
## Parallel Programming Models

<table>
<thead>
<tr>
<th></th>
<th>MapReduce</th>
<th>Project +</th>
<th>MPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFS/IO and task rescheduling overhead between iterations</td>
<td>Yes</td>
<td>No +1</td>
<td>No +1</td>
</tr>
<tr>
<td>Flexibility of computation model</td>
<td>AllReduce only +0.5</td>
<td>? +1</td>
<td>Flexible +1</td>
</tr>
<tr>
<td>Efficient AllReduce</td>
<td>Yes +1</td>
<td>Yes +1</td>
<td>Yes +1</td>
</tr>
<tr>
<td>Recover from faults between iterations</td>
<td>Yes +1</td>
<td>Yes +1</td>
<td>Apps</td>
</tr>
<tr>
<td>Recover from faults within each iteration</td>
<td>Yes +1</td>
<td>Yes +1</td>
<td>Apps</td>
</tr>
<tr>
<td>Final Score for scalable machine learning</td>
<td>3.5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
SVM Bottlenecks

Time consuming – 1M dataset, 8 days

Memory consuming – 1M dataset, 10G
Matrix Factorization Alternatives

<table>
<thead>
<tr>
<th>Factorization</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>QR</td>
<td>$O\left(\frac{4}{3}n^3\right)$</td>
</tr>
<tr>
<td>LU</td>
<td>$O\left(\frac{2}{3}n^3\right)$</td>
</tr>
<tr>
<td>Cholesky</td>
<td>$O\left(\frac{1}{3}n^3 + 2n^2\right)$</td>
</tr>
<tr>
<td>LDLT</td>
<td>$O\left(\frac{1}{3}n^3\right)$</td>
</tr>
<tr>
<td>Incomplete Cholesky</td>
<td>$O\left(p^2 n\right)$</td>
</tr>
<tr>
<td>Kronecker</td>
<td>$O\left(2n^2\right)$</td>
</tr>
</tbody>
</table>

**exact**

**approximate**
Parallelizing SVM [E. Chang, et al, NIPS 07]

Raw Data → Kernel Matrix → ICF → Matrix Multiplication

Incremental Data → Incremental Kernel Matrix → Incremental ICF → Incremental Matrix Multiplication

Matrix Summation

Newton IPM Iteration

α Tower

Cholesky Factorization
Incomplete Cholesky Factorization (ICF)

\[ \begin{array}{ccc}
  & \approx & \\
  n \times n & n \times p & p \times n \\
\end{array} \]
PSVM

Raw Data → Kernel Matrix → ICF → Matrix Multiplication

Incremental Data → Incremental Kernel Matrix → Incremental ICF → Incremental Matrix Multiplication

Matrix Summation → Incremental Linear System Solving

Cholesky Factorization → Newton IPM Iteration → α Tower
Matrix Product

\[ p \times n \times n \times p = p \times p \]
PSVM [E. Chang, et al, NIPS 07]

• Column-based ICF
  – Slower than row-based on single machine
  – Parallelizable on multiple machines

• Changing IPM computation order to achieve parallelization
Overheads
# Speedup

<table>
<thead>
<tr>
<th>Machines</th>
<th>Image (200k)</th>
<th>Speedup</th>
<th>CoverType (500k)</th>
<th>Speedup</th>
<th>RCV (800k)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1,958 (9)</td>
<td>10*</td>
<td>16,818 (442)</td>
<td>10*</td>
<td>45,135 (1373)</td>
<td>10*</td>
</tr>
<tr>
<td>30</td>
<td>572 (8)</td>
<td>34.2</td>
<td>5,591 (10)</td>
<td>30.1</td>
<td>12,289 (98)</td>
<td>36.7</td>
</tr>
<tr>
<td>50</td>
<td>473 (14)</td>
<td>41.4</td>
<td>3,598 (60)</td>
<td>46.8</td>
<td>7,695 (92)</td>
<td>58.7</td>
</tr>
<tr>
<td>100</td>
<td>330 (47)</td>
<td>59.4</td>
<td>2,082 (29)</td>
<td>80.8</td>
<td>4,992 (34)</td>
<td>90.4</td>
</tr>
<tr>
<td>150</td>
<td>274 (40)</td>
<td>71.4</td>
<td>1,865 (93)</td>
<td>90.2</td>
<td>3,313 (59)</td>
<td>136.3</td>
</tr>
<tr>
<td>200</td>
<td>294 (41)</td>
<td>66.7</td>
<td>1,416 (24)</td>
<td>118.7</td>
<td>3,163 (69)</td>
<td>142.7</td>
</tr>
<tr>
<td>250</td>
<td>397 (78)</td>
<td>49.4</td>
<td>1,405 (115)</td>
<td>119.7</td>
<td>2,719 (203)</td>
<td>166.0</td>
</tr>
<tr>
<td>500</td>
<td>814 (123)</td>
<td>24.1</td>
<td>1,655 (34)</td>
<td>101.6</td>
<td>2,671 (193)</td>
<td>169.0</td>
</tr>
<tr>
<td>LIBSVM</td>
<td>4,334 NA</td>
<td>NA</td>
<td>28,149 NA</td>
<td>NA</td>
<td>184,199 NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Scalability

• Computation
  – Parallelization
  – Approximation

• File Systems
  – Latency
  – Recovery

• Power Management
Sample Platforms

Servers
- CPUs
- DRAM
- Disks

Racks
- 40-80 servers
- Ethernet switch

Warehouse-scale Computer (WSC)
Sample Hierarchy

• Server
  – 16GB DRAM; 160MB Flash; 5 x 1TB disk

• Rack
  – 40 servers
  – 48 port Gigabit Ethernet switch

• Warehouse
  – 10,000 servers (250 racks)
  – 2K port Gigabit Ethernet switch
Storage --- One Server
Storage --- One Rack

![Graph showing performance metrics for different storage options at different levels (Server, Rack, Datacenter).]
Storage --- One Center

![Graph of Storage Analysis](image-url)
Google File System (GFS)

- Master manages metadata
- Data transfers happen directly between clients/chunkservers
- Files broken into chunks (typically 64 MB)
- Chunks triplicated across three machines for safety
- See SOSP^03 paper at http://labs.google.com/papers/gfs.html
WSC data availability: cluster file systems

- Data blocks of each stripe are placed on different fault domains
  - different disks, servers, racks
  - Data blocks are distributed across the whole WSC
    - read operations are easily load-balanced
    - recovery is highly efficient

- What affects data availability as seen by a client of a cluster file system?
Win in Scale

• Google Translate
• Voice
• Trend Prediction
  – An example benefits society
H1N1 United Nation Report

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

![Flu trends chart]

- Intense
- High
- Moderate
- Low
- Minimal

- Jul
- Aug
- Sep
- Oct
- Nov
- Dec
- Jan
- Feb
- Mar
- Apr
- May
- Jun

2009-2010
Past years ▼
Concluding Remarks

• Search + Social
• Increasing quantity and complexity of data demands scalable solutions
• Have parallelized key subroutines for mining massive data sets
  – Spectral Clustering [ECML 08]
  – Frequent Itemset Mining [ACM RS 08]
  – PLSA [KDD 08]
  – LDA [WWW 09, AAIM 09]
  – UserRank [Google TR 09]
  – Support Vector Machines [NIPS 07]
• Launched Google Q&A (Confucius) in 60+ countries
• Relevant papers
  – http://infolab.stanford.edu/~echang/
• Open Source PSVM, PLDA
  – http://code.google.com/p/psvm/
  – http://code.google.com/p/plda/
Models of Innovation

• Ivory tower
  • Only consider theory but not application

• Build it and they will come
  • Scientists drives product development

• “Research for sale”
  • Research funded by:
    – product groups or customers

• Research & development as equals
  • Research “sells” innovation;
  • Product “requests” innovation

• Google-style innovation