Social Computing and Its Application in Query Suggestion

Irwin King

king@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~king

Department of Computer Science & Engineering
The Chinese University of Hong Kong
Billionaires’ Shuffle

Facebook in 2004.02

2008

at 23 and $1.5 billion later...

Social Computing and Its Application in Query Suggestion, Irwin King, HUT, Finland, November 13, 2008
<table>
<thead>
<tr>
<th>Alexa as of Nov. 2008</th>
<th>USA</th>
<th>CHINA</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Google</td>
<td>Baidu</td>
<td>Yahoo</td>
</tr>
<tr>
<td>2</td>
<td>Yahoo</td>
<td>QQ</td>
<td>Google</td>
</tr>
<tr>
<td>3</td>
<td>Myspace</td>
<td>Sina</td>
<td>YouTube</td>
</tr>
<tr>
<td>4</td>
<td>YouTube</td>
<td>Google.cn</td>
<td>Windows Live</td>
</tr>
<tr>
<td>5</td>
<td>Facebook</td>
<td>Taobao</td>
<td>Facebook</td>
</tr>
<tr>
<td>6</td>
<td>Windows Live</td>
<td>163</td>
<td>MSN</td>
</tr>
<tr>
<td>7</td>
<td>MSN</td>
<td>Yahoo</td>
<td>Myspace</td>
</tr>
<tr>
<td>8</td>
<td>Wikipedia</td>
<td>Google</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>9</td>
<td>EBay</td>
<td>Sohu</td>
<td>Blogger</td>
</tr>
<tr>
<td>10</td>
<td>AOL</td>
<td>Youku</td>
<td>Yahoo.jp</td>
</tr>
</tbody>
</table>
What’s On the Menu?

• Web 2.0 and Social X
• Social Computing
• Some Interesting Problems
  • Collaborative Filtering
  • Query Suggestion
What’s On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
  - Collaborative Filtering
  - Query Suggestion
Web 2.0

- Web as a medium vs. **Web as a platform**
- Read-Only Web vs. **Read-and-Write Web**
- Static vs. **Dynamic**
- Restrictive vs. **Freedom & Empowerment**
- Technology-centric vs. **User-centric**
- Limited vs. **Rich User Experience**
- Individualistic vs. **Group/Collective Behavior**
- Consumer vs. **Producer**
- Transactional vs. **Relational**
- Top-down vs. **Bottom-up**
- People-to-Machine vs. **People-to-People**
- Search & browse vs. **Publish & Subscribe**
- Closed application vs. **Service-oriented Services**
- Functionality vs. **Utility**
- Data vs. **Value**
Social Networking
Social Search

- Social Search Engine
- Leveraging your social networks for searching
Social Bookmarking
Social Media

- YouTube
- Flickr
- Second Life

Social Computing and Its Application in Query Suggestion, Irwin King, HUT, Finland, November 13, 2008
Social News/Mash Up
Social Knowledge Sharing

Social Computing and Its Application in Query Suggestion, Irwin King, HUT, Finland, November 13, 2008
Social Marketing

- Viral marketing
- Who are the brokers?
- Who can exert the **most influence** on buying/selling?
- How **much** should one advertise?
Social/Human Computation
Human Computation
Face-off Game

- Utility Function
- Verification
- Collective Intelligence
  - Relevance Feedback
  - Pair-wise Similarity Function
Web 2.0 Revolution

The Three C’s

Connectivity

Collaboration

Communities
What’s On the Menu?

• Web 2.0 and Social X
• Social Computing
• Some Interesting Problems
  • Collaborative Filtering
  • Query Suggestion
Social Computing (SC)

• Social computing is a general term for an area of computer science that is concerned with the intersection of social behavior and computational systems. Wikipediat

• A social structure in which technology puts power in communities, not institutions. Forrester

• Forms of web services where the value is created by the collective contributions of a user population.
Issues

- **Theory** and models
- **Mining** of existing information, e.g., spatial (relations) and temporal (time) domains
- Dealing with *partial* and *incomplete* information, e.g., collaborative filtering, ranking, tagging, etc.
- **Scalability** and algorithmic issues
- **Security** and *privacy* issues
- **Monetization** of social interactions
Machine Learning in SC

- Classification, clustering, regression, etc.
- New insights on the data
  - Social relations are often hidden (latent)
  - Change data from \((x, y)\) to \((x, c_1(x), c_2(x), \ldots, y)\)
- \(c(x) = \text{context in tags, relations, ratings, etc.}\)
- Data type = binary, integer, real, cardinal, etc.
Organizational Chart
Social Network Chart

Authority vs. Importance

Social Computing and Its Application in Query Suggestion, Irwin King, HUT, Finland, November 13, 2008
What’s On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
  - Collaborative Filtering
  - Query Suggestion
What’s On the Menu?

• Web 2.0 and Social X
• Social Computing
• Some Interesting Problems
  • Collaborative Filtering
  • Query Suggestion
A Better Mousetrap?
Challenges

• Queries contain ambiguous and new terms

• apple: “apple computer” or “apple pie”?

• NDCG:?

• Users tend to submit short queries consisting of only one or two words

• almost 20% one-word queries

• almost 30% two-word queries

• Users may have little or even no knowledge about the topic they are searching for!
What is Clickthrough Data

- Query logs recorded by search engines

\( \langle u, q, l, r, t \rangle \)

<table>
<thead>
<tr>
<th>ID</th>
<th>Query</th>
<th>URL</th>
<th>Rank</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>358</td>
<td>facebook</td>
<td><a href="http://www.facebook.com">http://www.facebook.com</a></td>
<td>1</td>
<td>2008-01-01 07:17:12</td>
</tr>
</tbody>
</table>

- Users’ relevance feedback to indicate desired/preferred/target results
Joint Bipartite Graph

\[ B_{uq} = (V_{uq}, E_{uq}) \]
\[ V_{uq} = U \cup Q \]
\[ U = \{u_1, u_2, \ldots, u_m\} \]
\[ Q = \{q_1, q_2, \ldots, q_n\} \]
\[ E_{uq} = \{(u_i, q_j)| \text{ there is an edge from } u_i \text{ to } q_j\} \]

The edge \((u_i, q_j)\) exists in this bipartite graph if and only if a user \(u_i\) issued a query \(q_j\).

\[ B_{ql} = (V_{ql}, E_{ql}) \]
\[ V_{ql} = Q \cup L \]
\[ Q = \{q_1, q_2, \ldots, q_n\} \]
\[ L = \{l_1, l_2, \ldots, l_p\} \]
\[ E_{ql} = \{(q_i, l_j)| \text{ there is an edge from } q_i \text{ to } l_j\} \]

The edge \((q_j, l_k)\) exists if and only if a user \(u_i\) clicked a URL \(l_k\) after issuing a query \(q_j\).
Key Points

- Two-level latent semantic analysis

Level 1

- Consider the use of a joint user-query and query-URL bipartite graphs for query suggestion
- Use matrix factorization for learning query features in constructing the Query Similarity Graph

Level 2

- Use heat diffusion for similarity propagation for query suggestions
• Queries are issued by the users, and which URLs to click are also decided by the users

• Two distinct users are similar if they issued similar queries

• Two queries are similar if they are issued by similar users
Normalized weight, how many times $u_i$ issued $q_j$

Normalized weight, how many times $q_j$ is linked to $l_k$

$L$-dimensional vector of user $u_i$

$L$-dimensional vector of query $q_j$

$L$-dimensional vector of URL $l_k$

\[
\mathcal{H}(R, U, Q) = \min_{U, Q} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij}^{R} (r_{ij}^* - g(U_i^T Q_j))^2 \\
+ \frac{\alpha_u}{2} \|U\|_F^2 + \frac{\alpha_q}{2} \|Q\|_F^2
\]

\[
\mathcal{H}(S, Q, L) = \min_{Q, L} \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{p} I_{jk}^{S} (s_{jk}^* - g(Q_j^T L_k))^2 \\
+ \frac{\alpha_q}{2} \|Q\|_F^2 + \frac{\alpha_l}{2} \|L\|_F^2
\]
A local minimum can be found by performing gradient descent in $U_i$, $Q_j$ and $L_k$

Social Computing and Its Application in Query Suggestion, Irwin King, HUT, Finland, November 13, 2008
Gradient Descent Equations

\[
\frac{\partial H}{\partial U_i} = \alpha_r \sum_{j=1}^{n} I_{ij}^R g'(U_i^T Q_j)(g(U_i^T Q_j) - r_{ij}^*)Q_j + \alpha_u U_i, \\
\frac{\partial H}{\partial Q_j} = \sum_{k=1}^{p} I_{jk}^S g'(Q_j^T L_k)(g(Q_j^T L_k) - s_{jk}^*)L_k \\
+ \alpha_r \sum_{i=1}^{m} I_{ij}^R g'(U_i^T Q_j)(g(U_i^T Q_j) - r_{ij}^*)U_i + \alpha_u Q_j, \\
\frac{\partial H}{\partial L_k} = \sum_{j=1}^{n} I_{jk}^S g'(Q_j^T L_k)(g(Q_j^T L_k) - s_{jk}^*)Q_j + \alpha_l L_k,
\]

Only the Q matrix, the queries’ latent features, is being used to generate the query similarity graph!
Query Similarity Graph

- Similarities are calculated using queries’ latent features
- Only the top-$k$ similar neighbors (terms) are kept
Similarity Propagation

• Based on the **Heat Diffusion Model**

• In the query graph, given the **heat sources** and the **initial heat values**, start the heat diffusion process and perform **$P$ steps**

• Return the **Top-$N$** queries in terms of highest heat values for query suggestions
Heat Diffusion Model

- Heat diffusion is a physical phenomena.
- Heat flows from high temperature to low temperature in a medium.
- Heat kernel is used to describe the amount of heat that one point receives from another point.
- The way that heat diffuse varies when the underlying geometry.

\[ \rho C_P \frac{\partial T}{\partial t} = Q + \nabla \cdot (k \nabla T) \]

- \( \rho \) Density
- \( C_P \) Heat capacity and constant pressure
- \( \frac{\partial T}{\partial t} \) Change in temperature over time
- \( Q \) Heat added
- \( k \) Thermal conductivity
- \( \nabla T \) Temperature gradient
- \( \nabla \cdot v \) Divergence
Heat Diffusion Process
Similarity Propagation Model

\[
\frac{f_i(t + \Delta t) - f_i(t)}{\Delta t} = \alpha \left( -\frac{\tau_i}{d_i} f_i(t) \sum_{k:(q_i,q_k) \in E} w_{ik} + \sum_{j:(q_j,q_i) \in E} \frac{w_{ji}}{d_j} f_j(t) \right) (1)
\]

\[
f(1) = e^{\alpha \mathbf{H}} f(0) \quad (2)
\]

\[
H_{ij} = \begin{cases} 
\frac{w_{ji}}{d_j}, & (q_j, q_i) \in E, i = j, \\
-(\tau_i/d_i) \sum_{k:(i,k) \in E} w_{ik}, & \text{otherwise.} 
\end{cases} \quad (3)
\]

\[
f(1) = e^{\alpha \mathbf{R}} f(0), \quad \mathbf{R} = \gamma \mathbf{H} + (1 - \gamma) \mathbf{g} \mathbf{1}^T \quad (4)
\]

- \(\alpha\): Thermal conductivity
- \(d_i\): Heat value of node \(i\) at time \(t\)
- \(f_i(t)\): Heat value of node \(i\) at time \(t\)
- \(w_{ik}\): Weight between node \(i\) and node \(k\)
- \(f(0)\): Vector of the initial heat distribution
- \(f(1)\): Vector of the heat distribution at time 1
- \(\tau_i\): Equal to 1 if node \(i\) has outlinks, else equal to 0
- \(\gamma\): Random jump parameter, and set to 0.85
- \(\mathbf{g}\): Uniform stochastic distribution vector
Discrete Approximation

- Compute $e^{\alpha \mathbf{R}}$ is time consuming
- We use the **discrete approximation** to substitute
  \[
  f(1) = \left( \mathbf{I} + \frac{\alpha}{P} \mathbf{R} \right)^P f(0)
  \]
- For every heat source, only diffuse heat to its neighbors within $P$ steps
- In our experiments, $P = 3$ already generates fairly good results
Query Suggestion Procedure

• For a given query \( q \)
  1. Select a set of \( n \) queries, each of which contains at least one word in common with \( q \), as heat sources
  2. Calculate the initial heat values by
  
  \[
  f_{\hat{q}_i}(0) = \frac{|\mathcal{W}(q) \cap \mathcal{W}(\hat{q}_i)|}{|\mathcal{W}(q) \cup \mathcal{W}(\hat{q}_i)|}
  \]
  3. Use \( f(1) = e^{\alpha R} f(0) \) to diffuse the heat in graph
  4. Obtain the Top-\( N \) queries from \( f(1) \)

Example:

\( q = \) “Sony”

“Sony” = 1

“Sony Electronics” = 1/2

“Sony Vaio Laptop” = 1/3
Physical Meaning of $\alpha$

- If set $\alpha$ to a large value
  - The results depend more on the query graph, and more semantically related to original queries, e.g., \texttt{travel} $\Rightarrow$ lowest air fare

- If set $\alpha$ to a small value
  - The results depend more on the initial heat distributions, and more literally similar to original queries, e.g., \texttt{travel} $\Rightarrow$ travel insurance
# Experimental Dataset

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Clickthrough data from AOL search</th>
<th>After Pre-Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Period</td>
<td>March 2006 to May 2006 (3 months)</td>
<td></td>
</tr>
<tr>
<td>Lines of Logs</td>
<td>19,442,629</td>
<td></td>
</tr>
<tr>
<td>Unique user IDS</td>
<td>657,426</td>
<td>192,371</td>
</tr>
<tr>
<td>Unique queries</td>
<td>4,802,520</td>
<td>224,165</td>
</tr>
<tr>
<td>Unique URLs</td>
<td>1,606,326</td>
<td>343,302</td>
</tr>
<tr>
<td>Unique words</td>
<td></td>
<td>69,937</td>
</tr>
</tbody>
</table>
# Query Suggestions

Table 2: Examples of LSQS Query Suggestion Results ($k = 50$)

<table>
<thead>
<tr>
<th>Testing Queries</th>
<th>Suggests $\alpha = 10$</th>
<th>Suggests $\alpha = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 2</td>
</tr>
<tr>
<td>michael jordan</td>
<td>michael jordan shoes</td>
<td>michael jordan bio</td>
</tr>
<tr>
<td>travel</td>
<td>travel insurance</td>
<td>abc travel</td>
</tr>
<tr>
<td>java</td>
<td>sun java</td>
<td>java script</td>
</tr>
<tr>
<td>global services</td>
<td>ibm global services</td>
<td>global technical services</td>
</tr>
<tr>
<td>walt disney land</td>
<td>world of disney</td>
<td>disney world orlando</td>
</tr>
<tr>
<td>intel</td>
<td>intel vs amd</td>
<td>amd vs intel</td>
</tr>
<tr>
<td>job hunt</td>
<td>jobs in maryland</td>
<td>monster job</td>
</tr>
<tr>
<td>photography</td>
<td>photography classes</td>
<td>portrait photography</td>
</tr>
<tr>
<td>internet explorer</td>
<td>ms internet explorer</td>
<td>internet explorer repair</td>
</tr>
<tr>
<td>fitness</td>
<td>fitness magazine</td>
<td>lifestyles family fitness</td>
</tr>
<tr>
<td>m schumacher</td>
<td>schumacher</td>
<td>red bull racing</td>
</tr>
<tr>
<td>solar system</td>
<td>solar system project</td>
<td>solar system facts</td>
</tr>
<tr>
<td>sunglasses</td>
<td>replica sunglasses</td>
<td>cheap sunglasses</td>
</tr>
<tr>
<td>search engine</td>
<td>audio search engine</td>
<td>best search engine</td>
</tr>
<tr>
<td>disease</td>
<td>grovers disease</td>
<td>liver disease</td>
</tr>
<tr>
<td>pizzahut</td>
<td>pizza hut menu</td>
<td>pizza coupons</td>
</tr>
<tr>
<td>health care</td>
<td>health care proxy</td>
<td>universal health care</td>
</tr>
<tr>
<td>flower delivery</td>
<td>global flower delivery</td>
<td>online florist</td>
</tr>
<tr>
<td>wedding</td>
<td>wedding guide</td>
<td>wedding reception ideas</td>
</tr>
<tr>
<td>astronomy</td>
<td>astronomy magazine</td>
<td>astronomy pic of the day</td>
</tr>
</tbody>
</table>
Comparisons

<table>
<thead>
<tr>
<th></th>
<th>Top 1</th>
<th>Top 2</th>
<th>Top 3</th>
<th>Top 4</th>
<th>Top 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LSQS</strong></td>
<td>jaguar cat</td>
<td>jaguar commercial</td>
<td>jaguar parts</td>
<td>jaguarundi</td>
<td>leopard</td>
</tr>
<tr>
<td><strong>SimRank</strong></td>
<td>american black bear</td>
<td>bottlenose dolphin</td>
<td>leopard</td>
<td>margay</td>
<td>jaguarundi</td>
</tr>
<tr>
<td><strong>apple</strong></td>
<td>apple computers</td>
<td>apple ipod</td>
<td>apple diet</td>
<td>apple vacations</td>
<td>apple bottom</td>
</tr>
<tr>
<td><strong>LSQS</strong></td>
<td>ipod troubleshooting</td>
<td>apple quicktime</td>
<td>apple ipods</td>
<td>apple computers</td>
<td>apple software</td>
</tr>
<tr>
<td><strong>SimRank</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Comparisons between LSQS and SimRank

<table>
<thead>
<tr>
<th></th>
<th>Accuracy By Experts</th>
<th>Accuracy By ODP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LSQS</strong></td>
<td>0.8413</td>
<td>0.6823</td>
</tr>
<tr>
<td><strong>SimRank</strong></td>
<td>0.7101</td>
<td>0.5789</td>
</tr>
</tbody>
</table>

ODP, Open Directory Project, see http://dmoz.org
Impact of Parameter $k$

To test the extend of similarity needed

(a) Evaluation by Experts

(b) Evaluation by ODP Database

Figure 2: Impact of Parameter $k$ ($P = 3$)
Impact of Parameter $P$

To test the propagation influence

(a) Evaluation by Experts  
(b) Evaluation by ODP Database

Figure 3: Impact of Parameter $P$ ($k = 50$)
Efficiency Analysis

Figure 4: Efficiency Analysis
Summary

• Propose an offline novel joint matrix factorization method using user-query and query-URL bipartite graphs for learning query features

• Propose an online diffusion-based similarity propagation and ranking method for query suggestion
Conclusion

- Social Computing is a **paradigm shift**!
- Novel views on the **spatial** and **temporal** relationship among **social entities**!
- Great **opportunities** in a new research direction!
On-Going Research

- **Machine Learning**
  - Direct Zero-norm Optimization for Feature Selection (ICDM'08)
  - Semi-supervised Learning from General Unlabeled Data (ICDM'08)
  - Learning with Consistency between Inductive Functions and Kernels (NIPS'08)
  - An Extended Level Method for Efficient Multiple Kernel Learning (NIPS'08)
  - Semi-supervised Text Categorization by Active Search (CIKM'08)
  - Transductive Support Vector Machine (NIPS'07)
  - Global and local learning (ICML'04, JMLR’04)

- **Web Intelligence**
  - Effective Latent Space Graph-based Re-ranking Model with Global Consistency (WSDM'09)
  - Formal Models for Expert Finding on DBLP Bibliography Data (ICDM’08)

- **Collaborative Filtering**
  - Recommender system: accurate recommendation based on sparse matrix (SIGIR’07)
  - SoRec: Social Recommendation Using Probabilistic Matrix Factorization (CIKM’08)

- **Human Computation**
  - An Analytical Study of Puzzle Selection Strategies for the ESP Game (WI’08)
  - An Analytical Approach to Optimizing The Utility of ESP Games (WI’08)
Acknowledgments

• Prof. Michael R. Lyu
• Prof. Jimmy Lee
• Dr. Kaizhu Huang
• Dr. Haixuan Yang
• Thomas Chan (M.Phil)

• Hongbo Deng (Ph.D.)
• Zhenjiang Lin (Ph.D.)
• Hao Ma (Ph.D.)
• Haiqin Yang (Ph.D.)
• Xin Xin (Ph.D.)
• Zenglin Xu (Ph.D.)
• Chao Zhou (Ph.D.)
Q & A

http://www.cse.cuhk.edu.hk/~king