BetterLife 2.0: Large-scale Social Intelligence Reasoning on Cloud

Dexter H. Hu
Yinfeng Wang
Cho-Li Wang
{hyhu,yfwang,clwang}@cs.hku.hk
Outline

- Introduction
- BetterLife 2.0 Overview
- Performance Evaluation & Analysis
- Related Work
- Conclusion & Future work
Introduction

• Many social networking websites with mobile access and recommendation service
Context-aware Service on Cloud

- Context-aware mobile applications in pervasive computing
  - Record contexts, social & environmental interactions: GPS, RFID tags, Google Calendar

- Information Surge
  - Growing individual and group behaviors in the real world
  - Difficult to find if certain information is useful
Case Based Reasoning for Intelligence

- Solve new problems by finding previous similar experiences
- K-NN Algorithm

\[ \text{Similarity}(\vec{N}, \vec{P}) = \frac{\sum_{i=1}^{n} \text{Sim}(N_i, P_j) \times W_i}{\sum_{i=1}^{n} W_i} \]

- Adopt past case solution

BetterLife 2.0: Large-scale Social Intelligence Reasoning on Cloud
CBR 4R Cycle

• **Retrieve**: Given a target problem, retrieve the most relevant or similar cases from memory to solve it.

• **Reuse**: Map the solution from the prior case to the target problem. This may involve adapting the solution as needed to fit the new situation.

• **Revise**: Having mapped the previous solution to the target situation, test the new solution in the real world (or a simulation) and, if necessary, revise.

• **Retain**: After the solution has been successfully adapted to the target problem, store the resulting experience as a new case in memory.
Social Closeness

• Diversity of Same Interest
  • In social network, users show interest by...
Common Activities

- Follow each others
- Join the same group
- Write some blogs
- Comments on each other's blog
- Comments on the same blog (not written by any of them)
BetterLife 2.0 Goal

• To provide an extensible framework to implement proactive personalized recommendation service for users in daily life by using Case-based Reasoning and social network information to analyze large amount of data on Cloud
Outline

• Introduction
• BetterLife 2.0 Overview
• Performance Evaluation & Analysis
• Related Work
• Conclusion & future work
BetterLife 2.0: Large-scale Social Intelligence Reasoning on Cloud

BetterLife 2.0 Architecture
BetterLife 2.0 Components

• Cloud Layer
  • Hadoop Distributed File System (HDFS) clusters
  • Collectively store application data represented by cases and social network information, which include relationship topology, and pairwise social closeness information

• Case-based Reasoning Engine
  • Extended from jCOLIBRi2
  • Has a data connector to Cloud Layer,
  • Calculate similarity measurement between cases to retrieve the most similar ones.

• Application Interface:
  • a master node which is responsible for handling the request query from user
  • Mobile clinet and social networking web client
MapReduce Workflow in BetterLife 2.0

- \((UserID, Timestamp, Longitude, Latitude, ShopID, ProductID, Price)\)
MapReduce Workflow in BetterLife 2.0

- Filter by ProductID,
- Calculate the global similarity except social closeness

\[
\text{Similarity}(N, P) = \sum_{i=1}^{n} \text{Sim}(N_i, P_i) * W_i \\
\text{Distance}(N, P) = 1 - \text{Similarity}(N, P)
\]

\[
\sum_{i=1}^{n} W_i = 1
\]
CBR Local Similarity Functions

- **Location Similarity**
  \[
  \text{Sim}(N_{gps}, P_{gps}) = 1 - \frac{\text{Distance}(N_{gps}, P_{gps})}{\text{MaxDistance}}
  \]

- **Timestamp Similarity**
  \[
  \text{Sim}(N_t, P_t) = 1 - \frac{\text{Diff}(N_t, P_t)}{24 \times 60}
  \]

- **Price Similarity**
  \[
  \text{Sim}(P_{price}, N_{price}) = \frac{\text{Max}_{price} - N_{price}}{\text{Max}_{price} - \text{Min}_{price}}
  \]

- **Social Closeness Similarity**
  \[
  \max_{p(i,j) \in P(i,j)} \prod_{e(u,v) \in p(i,j)} w(u, v)
  \]
Social Similarity Functions

\[
Distance(N, P) = 1 - W_{gps} \times Sim(N_{gps}, P_{gps}) \\
- W_t \times Sim(N_t, P_t) \\
- W_{price} \times Sim(N_{price}, P_{price}) \\
W_{social} \times \max_{p(i,j) \in P(i,j)} \prod_{e(u,v) \in p(i,j)} \ w(u,v)
\]

\[
Distance(C, Q)_i = c - W_{social} \times w_1 \times w_2 \times \ldots \times w_i
\]

\[
c = 1 - W_{gps} \times Sim(N_{gps}, P_{gps}) \\
- W_t \times Sim(N_t, P_t) \\
- W_{price} \times Sim(N_{price}, P_{price})
\]

At iteration \(i + 1\):

\[
Distance(C, Q)_{i+1} = c - W_{social} \times w_1 \times w_2 \times \ldots \times w_i \times w_{i+1}
\]

\[
Distance(C, Q)_{i+1} = c - (c - Distance(C, Q)_i) \times w_{i+1}
\]
Mapper Algorithm

Algorithm 1 SocialNetworkMapper (Key k, Node n)
1: if \( n.\text{Color} == \text{GRAY} \) then
2:   for all edge \( e \) of \( n \) do
3:     Node \( v\text{node} \) $\leftarrow$ new Node \((e.\text{ToID})\)
4:     \( v\text{node}.\text{Distance} \leftarrow c - (c - n.\text{Distance}) \times e.\text{Weight} \)
5:     \( v\text{node}.\text{Color} \leftarrow \text{GRAY} \)
6:     \( \text{word} \leftarrow v\text{node}.\text{Id} \)
7:     Emit $\langle \text{word}, v\text{node} \rangle$
8:     \( n.\text{Color} \leftarrow \text{BLACK} \)
9:   end for
10: end if
11: \( \text{word} \leftarrow n.\text{Id} \)
12: Emit $\langle \text{word}, n \rangle$
Reducer Algorithm

Algorithm 1 SocialNetworkReducer (Key k, Iterator V)

1: \( \text{distance} \leftarrow \text{MAX} \)
2: \( \text{color} \leftarrow \text{WHITE} \)
3: \( \text{edges} \leftarrow \text{NULL} \)
4: \textbf{for all} Node \( u \in V \) \textbf{do}
5: \hspace{1em} \textbf{if} \ u.\text{Edges}.\text{size} > 0 \textbf{then}
6: \hspace{2em} \text{edges} \leftarrow u.\text{Edges}
7: \hspace{1em} \textbf{end if}
8: \hspace{1em} \textbf{if} \ u.\text{Distance} < \text{distance} \textbf{then}
9: \hspace{2em} \text{distance} \leftarrow u.\text{Distance}
10: \hspace{2em} \{ /* \text{Save the minimum distance} */ \}
11: \hspace{1em} \textbf{end if}
12: \hspace{1em} \textbf{if} \ u.\text{Color}.\text{Ordinal} > \text{color}.\text{Ordinal} \textbf{then}
13: \hspace{2em} \text{color} \leftarrow u.\text{Color}
14: \hspace{2em} \{ /* \text{Save the darkest color} */ \}
15: \hspace{1em} \textbf{end if}
16: \textbf{end for}
17: \text{Node} \ n \leftarrow \text{new Node} \ (k)
18: \ n.\text{Distance} \leftarrow \text{distance}
19: \ n.\text{Edges} \leftarrow \text{edges}
20: \ n.\text{Color} \leftarrow \text{color}
21: \text{Emit} \ < k, n >
22: \textbf{if} \ \text{color} == \text{GRAY} \textbf{then}
23: \hspace{1em} \text{reporter.incrCounter(Counters.MOREGRAY, 1)}
24: \textbf{end if}
Outline

• Introduction
• BetterLife 2.0 Overview
• Performance Evaluation & Analysis
• Related Work
• Conclusion & future work
Experiment Setting

HKU Gideon-II Cluster with Hadoop 0.20.2

<table>
<thead>
<tr>
<th>Table 1: Node Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
</tr>
<tr>
<td>2 x Intel Quad-Core E5540 Xeon CPU, 2.53GHz, 8MB cache</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
</tr>
<tr>
<td>16GB DDR3 memory, 1066MHz, dual ranked UDIMMs</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
</tr>
<tr>
<td>2 x 250GB 7.2K RPM SATA hard disks, running in RAID-1</td>
</tr>
<tr>
<td><strong>Network Interface</strong></td>
</tr>
<tr>
<td>Broadcom 5709 dual-port</td>
</tr>
<tr>
<td><strong>OS</strong></td>
</tr>
<tr>
<td>Fedora 11</td>
</tr>
</tbody>
</table>
Data Set

• 103 user accounts in our product rating social networking website from Elgg

• Recorded activities like commenting on product, joining groups and following friends, to demonstrate a community and form historical cases.

• Locations of 7-Eleven convenient stores in Hong Kong with the social network topology of these 103 users.

• To obtain enough cases under different contexts, users’ behaviors were simulated by a set of pre-defined rules (location clusters, product type clusters, time cluster)

• Generate spam users with products of lower prices.
Application: Shopping Recommender

Send user ID, barcode, and GPS location, timestamp
Response Time

In Hadoop, CBR can run even the case base size is 25000K # of cases, while the response time only scales almost linearly (to 50s).
Effect of Social Information

When $k = 3$, accuracy in all cases is at least 70%.
For both $k = 1$ and $k = 3$, the result accuracy is improved more than 10% with social relationships.
Related Work

• Introduction
• BetterLife 2.0 Overview
• Performance Evaluation & Analysis
• Related Work
• Conclusion & future work
Conclusion

• BetterLife 2.0 is based on the
  • Case-Based Reasoning for its additive knowledge space growing, easy problem modeling
  • MapReduce framework for its large scale processing capability on cloud
  • Social network information for more relevant and trust worthy recommendation
Related Work

• Large-scale Recommender system
  • Item recommendation by collaborative filtering
  • User-centered collaborative location and activity filtering algorithm to make mobile recommendations through mining knowledge from GPS trajectory.
• Rule-based Reasoning vs Case-base Reasoning
• Social Network Analysis
  • Leskovec et al. discussed the phenomenon of information cascade
  • Relationship Closeness Inventory (RCI)
Acknowledgement

• Supported in part by Hong Kong UGC Special Equipment Grant (SEG HKU09) and China 863 grant 2006AA01A111 (CNGrid).

• Prototype by FYP students: Lo Fung, Kong Kwai Yee, Wong Kwok Kit