Identifying the Large-Scale Structure of Blogosphere

Makoto UCHIDA
School of Engineering, the University of Tokyo.

In collaboration with;
N. Shibata, Y. Kajikawa, Y. Takeda, S. Shirayama,

@ Contributed Talk Session, NetSci07, New York.
Summary of this Talk

- We analyzed...
  - a “network” in the Japanese Blogosphere.
  - the large-scale structure of the networks.

- We found that...
  - the weblog networks are sparse, but highly modularized (very strong community structure).
  - each community has individual topic.
Weblogs are Getting More Social Impact...

- Number of weblogs and bloggers are increasing rapidly..
  - 8M bloggers and 35M readers (in Japan).
  - Weblogs are …
    - written by a large number of individuals.
    - reflecting people’s up-to-date interests.

⇒ Influences of the Blogosphere on our society can no longer be underestimated,
... as a datasource of web data mining, or commercial activities.
Especially in Japan..

Q4 2006 - Posts by Language

- japanese 37%
- english 36%
- chinese 8%
- italian 3%
- spanish 3%
- russian 2%
- french 2%
- portuguese 2%
- german 1%
- farsi 1%
- Other 5%

37% of weblog posts are written in JAPANESE
Features of Weblogs

Weblogs are a subset of the World Wide Web.

✓ Each post has its own permanent URL (permlink).
✓ Hyperlinks to/from other weblog posts.
✓ Trackback ping protocol.

⇒ Weblog posts form a kind of network, where individual posts are the nodes. It turned out to have...
  • small-world property.
  • disassortative mixing pattern.
  • community structure.


Networks formed by Trackbacks

• Trackbacks are hyperlinks with stronger intention.
  – Both authors of citing and being cited weblogs can recognize the relationship.

Linked by trackback.

We investigate the structure of the network of weblog posts and trackbacks.
Dataset #1: Snowball Sampling

- Snowball sampling from a randomly-selected seed.
  - The resulting network is connected.
Dataset #2: Large-Scale Collection

- Collecting all posts from ping servers.
  - Trackbacks among the posts are analyzed afterwards.
  - The resulting network may be unconnected.
Datasets

- **Dataset #1**: collected in Dec. 2004.
  - 25,668 posts, 67,828 trackbacks.
  - single connected component.
  - $L = 14.88$, $C = 0.24$ ($L_{\text{rand}} = 7.59$, $C_{\text{rand}} = 1.4 \times 10^{-4}$).

- **Dataset #2**: collected in Sep. 2006.
  - 14,733,908 posts, 163,167 trackbacks.
  - 99% of the posts are isolated (with no trackbacks).
  - the largest connected component includes 0.2% of whole entries.

- **Dataset #2’**: The Largest Connected Component of #2.
  - 29,957 posts, 66,552 trackbacks.
  - single connected component.

Highly clustered, but not so small average path length.
Community Structure.

- We apply a community extracting method based on modularity to the datasets [Newman2004].
  - Nodes are divided into communities so as to maximize the modularity $Q$ (Usually $Q_{\text{max}}$ is 0.4-0.7).

$$Q = \sum_i \left( e_{ii} - a_i^2 \right)$$

where

$$\begin{cases}
  e_{ij} : \text{fraction of edges between the community } i \text{ and } j. \\
  e_{ii} : \text{fraction of edges within the same community } i. \\
  a_i = \sum_j e_{ij} : \text{fraction of edges connected to nodes in the community } i.
\end{cases}$$
Visualization of the dataset #1

• Edges in the same community are colored by the same color.
Visualization of the dataset #2
Visualization of the dataset #2’
Is that Firework?

http://www.yakei-kabegami.com/cgi-bin/kabegami/5509.html

(c) http://Kumanoji.t.land.to/
or Galaxy!

Community Structure (result)

- Extremely large modularity.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Entries</th>
<th># Edges</th>
<th># Communities</th>
<th>$Q_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>25,668</td>
<td>67,828</td>
<td>127</td>
<td>0.94</td>
</tr>
<tr>
<td>#2</td>
<td>14,733,908</td>
<td>163,167</td>
<td>21,915</td>
<td>0.99</td>
</tr>
<tr>
<td>#2’</td>
<td>29,957</td>
<td>66,552</td>
<td>210</td>
<td>0.98</td>
</tr>
</tbody>
</table>

i.e.)

$Q_{\text{max}} = 0.713$ for a collaboration network between physicists.

$Q_{\text{max}} = 0.745$ for an Amazon.com purchasing network.

Origins of the large modularity

- Labeling the topic of communities by featured terms of contents.

[Featured terms of entries]
- Terms $i$'s of $m$ largest $w_{i,j}$ value in entry $j$ are considered to represent the topic of entry $j$ (Featured terms of entry $j$)

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

- $tf_{i,j}$: number of occurrence of term $i$ in entry $j$.
- $df_i$: number of entries which contains the term $i$.
- $N$: number of all entries.

[Featured terms of communities]
- Terms $i$'s of largest $w_{i,k}^*$ value in community $k$ are considered to represent the topic of community $k$ (Featured terms of community $k$)

$$w_{i,k}^* = tf_{i,k}^* \times \log \left( \frac{C}{df_i^*} \right)$$

- $tf_{i,k}^*$: number of occurrence of term $i$ in featured terms of entries in community $k$.
- $df_i^*$: number of communities, featured terms of entries in which contains the term $i$.
- $C$: number of all communities.
## Labeling the communities

<table>
<thead>
<tr>
<th>Event</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-Niigata earthquake</td>
<td>Chuetsu (中越), earthquake (地震), Niigata pref. (新潟県), victims (被災者)</td>
</tr>
<tr>
<td>Japan Professional Baseball strike</td>
<td>Livedoor, baseball team (球団), strike (スト), professional (プロ), baseball (野球)</td>
</tr>
</tbody>
</table>

Network and communities on Blogosphere, as of Dec. 31, 2004

Temporal analysis on communities
Conclusions

- Blogosphere is...
  - sparse, but highly clustered network.
  - extremely modularized, strong community structure.
  - like FIREWORK or GALAXY :-)

- The strong community structure is due to topics being discussed in each community.
  - Topics can be extracted by linguistic filtering technique.
  - Useful for detecting trends in the Blogosphere.