Analyzing Patterns of User Content Generation in Online Social Networks (KDD 09)

Lei Guo, Yahoo!
Enhua Tan, Ohio State University
Songqing Chen, George Mason University
Xiaodong Zhang, Ohio State University
Yihong (Eric) Zhao, Yahoo!

Hsun-Ping, Hsieh Chien-Tung, Ho
SNA Paper Present 2009/10/13
Agenda

• Introduction
• User posting over time
• Distribution of user contributions
• Implications of user contribution distributions
• Conclusion
Agenda

- Introduction
- User posting over time
- Distribution of user contributions
- Implications of user contribution distributions
- Conclusion
Online Social Networks: Social connection

• Networking oriented OSNs (online social network)

Knowledge-sharing mainly among friends
Online Social Networks: content sharing

• Knowledge-sharing oriented OSNs

Content sharing is among all users: common interest topic
UGC in online social networks

• User generated content (UGC)
  – Users are basic elements of OSNs
  – OSNs are driven by user contributions

  User create new contents
  Contents attract new users

• Understanding UGC patterns is important
  – Business success: attract new users and clients
  – Identify and distinguish active users from spamming users
  – Predict hot spots and the trends of topics in user communities
  – Perform efficient resource management in the underlying supporting system

Advertisement

Contents attract new users

User create new contents

User generate content (UGC)
Existing studies about user contributions in online social networks

- **Wikipedia**
  - Power law: core users contribute most articles (ISSI’05)
    - Number of articles a user edited
    - Number of co-authors of a Wiki article
    - Heavy tailed, scale free: highly skewed towards top users
  - User contribution shifts from “elite” users to common users (CHI’07)
    - Log analysis from 2001 to 2006
  Power law or not: no conclusion

- **Delicious social bookmark (CHI’07)**
  - Similar shifts for user contribution as in Wikipedia
  - Power law or not: no conclusion

\[ y_i \propto i^{-\alpha} \]

\( y_i \): contribution of the user
\( i \): contribution rank of a user
Findings & contributions in this study

• User posting behavior of original content in these OSNs shows strong daily and weekly patterns

• Observe two groups of users with distinct posting behaviors
  – Steadily posting in the network
  – Inactively posting

• The analysis across three workloads from both short terms (in weeks) and long terms (in years) consistently shows that it follows the stretched exponential distribution
The stretched exponential distribution of user contributions in OSNs roughly follows the “80-20” rule.

- However, the cumulative contribution ratios of a small number of top-k users in the OSNs are much smaller in OSNs than those in standard power law distributions.
- This implies that contents in an OSN are not mainly contributed by a small number of core users.

The contributions of different UGC objects can be characterized by the stretched exponential model with different parameters.

- For example, the distribution of user contributions on high quality content tends to have a small stretch factor.
Agenda

• Introduction
• User posting over time
• Distribution of user contributions
• Implications of user contribution distributions
• Conclusion
Similar

Peak Time

Different cultures
-> different time

Smaller

Peak Time
Dynamics of user joining and posting in OSNs

The Number of new users per day
- increases with time
- **bursty** in large time scales

**User increase rate**
- decrease with time
- **bursty** in large time scales

**Post increase rate**
- decrease with time
- less bursty than user increase rate

**Implications**
- total user population and content do not increase exponentially
- User join bursts: post inc rate < user inc rate

Bursts and dynamics need to be considered for data analysis
User activity over time

**Author’s OSN age of posts**
- User’s posting frequency over time
  - The age of the user in OSN when an UGC object is posted
  - Bookmark: almost uniform distribution
  - Blog: a little skewed towards small ages
  - Answer: more skewed towards small ages

**Author’s OSN lifetime**
- User’s lifetime (active duration) in OSNs
  - The duration from the user joining time to the last user posting time
  - For user posting behavior
    - Long lifetime users
    - Short lifetime users
    - Other users: a wide range of lifetime
Agenda

- Introduction
- User posting over time
- Distribution of user contributions
- Implications of user contribution distributions
- Conclusion
**Original and non-original UGC content**

- Three kinds of UGC objects
  - Original UGC objects
  - Cut-and-paste objects
  - Spam and advertisement
- Spam: filtered out with ML model
- Cut-and-paste objects in Blog
  - Posted by a small number of users
  - No clear posting peak time
  - Focused on **recreation and social event** categories
- Spam users and cut-and-paste users are removed in our analysis
Stretched exponential distribution

- User contribution in a social network follows the stretched exponential distribution

- Rank order distribution:
  - Fat head and thin tail in log-log scale
  - Straight line in $\log x - y^c$ scale (SE scale)
Stretched exponential distribution

- Corresponding CCDF function (Weibull function)

\[ P(X \geq x) = e^{-\left(\frac{x}{x_0}\right)^c} \quad c : \text{constants} \]
\[ X_0^c : \text{constants} \]

- Let \( a = x_0^c \), \( b = y_1^c \)

\[ y_i^c = -a \log i + b \quad (1 \leq i \leq N) \]

\[ b = 1 + a \log N \quad \text{(assuming} \quad y_N = 1) \]

\( i : \text{rank of users} \quad (N \text{ users}) \)

\( y : \text{number of objects created by the user} \)
Maximal likelihood estimation error to get parameter

- The Maximal likelihood estimation by making product of probability density functions may result in non-trivial errors in the stretched exponential plot
  - While the random variables in a Weibull distribution are real numbers
  - The data to be fit are positive integer
Fit the data

- Use the coefficient of determination of the data fit

\[
SSE = \sum_{i=1}^{n} w(i)(y_i - \left(-a \log i + b\right)^{1/c})^2
\]

Smaller is better

\[
SST = \sum_{i=1}^{n} w(i)(y_i - \bar{y})^2
\]

Larger is better

\[
R^2 = 1 - \frac{SSE}{SST}
\]

Fit the data \(\Rightarrow\) close to 1
UGC creation patterns of Blog

\[ y \text{: number of original posts by the user} \]

\[ x \text{: contribution rank of user} \]

Parameters: maximum likelihood method

\[ R^2 \text{: coefficient of determination (1 means a perfect fit)} \]

\[ c = 0.418, a = 2.091, b = 27.517 \]

\[ R^2 = 0.997361 \]

\[ c = 0.32, a = 1.730, b = 22.696 \]

\[ R^2 = 0.999005 \]
UGC creation patterns of Bookmark

**x**: contribution rank of user  **y**: number of bookmark posts by the user

- **Bookmark imports**: bookmarks imported from user’s Web browser when joining the system
- **Bookmark posts**: bookmarks posted to the system by the bookmark plug-in of web browser
UGC creation patterns of Answer

\( \chi \): contribution rank of user  \( y \): number of answer posts by the user

- Best answer: the asker can select a best answer from all received answers. Best answers are high quality UGC posts since they are judged by the askers themselves.
Model validation

- **Chi-square test**
  
  \[ \chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]

  \( k \): number of bins, \( O_i \): total observed posts, \( E_i \): expected number of posts

  \( \chi^2 > \chi^2_{(\alpha,k-c)} \) rejected by the test

- **Validation on users joined the system simultaneously**
  - Users join rate increases with time
  - Some users may become inactive

- **Validation on different parts of workloads**
  - follow SE distribution with the same \( c \)
  - parameter \( c \) is the shape factor, not change for different parts of a workload

### Chi-square test results (\( \alpha = 0.05 \))

<table>
<thead>
<tr>
<th>Data set</th>
<th>( k )</th>
<th>( \chi^2 )</th>
<th>( \chi^2_{(\alpha,k-c)} )</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blog article</td>
<td>11</td>
<td>11.403</td>
<td>14.067</td>
<td>pass</td>
</tr>
<tr>
<td>Blog photo</td>
<td>12</td>
<td>14.072</td>
<td>15.507</td>
<td>pass</td>
</tr>
<tr>
<td>Bookmark (all posts)</td>
<td>10</td>
<td>11.486</td>
<td>12.592</td>
<td>pass</td>
</tr>
<tr>
<td>Bookmark (imports)</td>
<td>11</td>
<td>9.367</td>
<td>14.067</td>
<td>pass</td>
</tr>
<tr>
<td>Answer (all posts)</td>
<td>11</td>
<td>13.340</td>
<td>14.067</td>
<td>pass</td>
</tr>
<tr>
<td>Answer (best ans)</td>
<td>10</td>
<td>7.001</td>
<td>12.592</td>
<td>pass</td>
</tr>
</tbody>
</table>

\( R^2 = 0.98498 \)

\( c = 0.42, a = 2.375, b = 20.172 \)
Agenda

- Introduction
- User posting over time
- Distribution of user contributions
- Implications of user contribution distributions
- Conclusion
The “80-20” rule

• 80-20 rule of power law distributions
  – Pareto principle: 20% people own 80% social wealth
  – Internet systems: 20% web pages account for 80% requests
  – …

• In social networks
  – Blog: 20% users for 80% posts
  – Bookmark: 17% users for 83% posts
  – Answer: 13% users for 87% posts

Roughly follows the 80-20 rule

User contribution is stretched exponential

What is the difference between user contribution distribution in online social networks and user income distribution in a real society?
Asymptotical properties of top users

Highly skewed towards top users

Power law

Less skewed towards top users

Stretched exponential

The cumulative contribution ratio of top-k users among all n users in an OSN

\[
\frac{k}{n} \to 0, \quad \frac{T_{se}}{T_{pow}} \to 0
\]

A small number of top users cannot dominate the content in an OSN
The “core” users in social networks

- Looking for a threshold to identify most important users
- Stretched exponential distribution: general threshold for all systems

\[-\frac{dy}{y} : \text{decrease rate of user contribution along rank}\]
\[\frac{di}{i} : \text{increase rate of user contribution rank}\]

\[X_0 = \log k, \quad Y_0 = y_k\]

\[\Rightarrow k = \exp\left(\frac{1}{a} - \frac{1}{c}\right), \quad y_k = \left(\frac{a}{c}\right)^\frac{1}{c}\]
Creation patterns of different types of UGC

<table>
<thead>
<tr>
<th>Type</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>all posts</td>
<td>0.42</td>
</tr>
<tr>
<td>&gt; 1 KB</td>
<td>0.39</td>
</tr>
<tr>
<td>&gt; 2 KB</td>
<td>0.31</td>
</tr>
<tr>
<td>with tags</td>
<td>0.30</td>
</tr>
</tbody>
</table>

- more effort, smaller c
- longer articles need more effort to compose, adding tags needs extra effort

Blog article

<table>
<thead>
<tr>
<th>Type</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>imports</td>
<td>0.33</td>
</tr>
<tr>
<td>all posts</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Blog photo

<table>
<thead>
<tr>
<th>Type</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>all posts</td>
<td>0.25</td>
</tr>
<tr>
<td>best ans</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Bookmark

<table>
<thead>
<tr>
<th>Type</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>imports</td>
<td>0.33</td>
</tr>
<tr>
<td>all posts</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Answer

<table>
<thead>
<tr>
<th>Type</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>all posts</td>
<td>0.25</td>
</tr>
</tbody>
</table>

higher quality, smaller c
(more effort to compose)

Our conjecture: larger c, flatter user contribution distribution

Twitter

user participating effort is even smaller

WikiPedia

higher quality and more effort than best answer would have much smaller c

\[ y^c = -a \log i + b \]

small c: \[ y^c \sim \log y \]

Power law!
Discussion: UGC production vs. UGC consumption

- Internet media access patterns (PODC’08)
  - Number of requests to an media object is stretched exponential for different kinds of media systems

- Media request is content consumption (Youtube)
  - Stretch factor increases with file length (duration a user views)

- UGC creation is content production
  - Stretch factor decreases with the effort to create a UGC object
Agenda

- Introduction
- User posting over time
- Distribution of user contributions
- Implications of user contribution distributions
- Conclusion
Conclusion

• **Found in OSNs**
  – User lifetime in OSNs do not follow exponential distributions
  – User contribution distribution is stretched exponential
  – Different types of UGC content have different characteristics under the stretched exponential model

• **User contribution model: distribution of individual user behaviors**
  – Building block to understand more complex social network phenomena
  – Foundation to guide design, modeling and simulation of OSNs
Thanks for your attention
Question

• What the difference between the OSN model and Power Law model?