Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow

Dana Movshovitz-Attias
Yair Movshovitz-Attias
Peter Steenkiste
Christos Faloutsos

August 27, 2013 - ASONAM
Motivation

• Q&A networks are gaining popularity
• Most information is created by a small set of expert users.
• How to find and motivate expert users?
• Case study: StackOverflow
Strange javascript object init?

I recently came across this variable initialization in a WebGL tutorial:

```javascript
var mouse = { x: 0, y: 0 }, INTERSECTED;
```

I've never seen this format. I understand it's creating an object with an x and y property, but how is INTERSECTED related to the variable/object?

Thanks!

---

User Reputation

User

javascript variables object initialization
I may get it wrong, but I think it is just like

```javascript
var A=3,B;
```

 Defines A with value 3 and B unitialized. INTERSECTED is just another var.

The line is simply declaring two variables (mouse and INTERSECTED), and initializing mouse to `{ x: 0, y: 0}`.

INTERSECTED is not necessarily related to mouse, though clear code should only declare multiple variables together if they are highly related (even then, many prefer to declare every variable on a separate line).
StackOverflow Analysis

• In this work:
  ✓ Analysis of the SO reputation system (expert users)
  ✓ Participation patterns of expert and non-expert users
  ✓ SVD and PageRank analysis of the SO interaction graph
  ✓ Prediction of influential users using first months of activity
StackOverflow Dataset

- All actions performed in years: 2008-2012
- 3.5 M questions, 6.9 M answers, 1.3 M users
- 2.1 M accepted answers (62% of Q)
- Total votes:
  - 5.5 M for Q
  - 13 M for A
SO Reputation

- Users gain *reputation* by participating in site activities
- 2012 reputation range: 1 - 465K

<table>
<thead>
<tr>
<th>Action</th>
<th>Reputation change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer is voted up</td>
<td>+10</td>
</tr>
<tr>
<td>Question is voted up</td>
<td>+5</td>
</tr>
<tr>
<td>Answer is accepted</td>
<td>+15 (+2 to acceptor)</td>
</tr>
<tr>
<td>Question is voted down</td>
<td>-2</td>
</tr>
<tr>
<td>Answer is voted down</td>
<td>-2 (-1 to voter)</td>
</tr>
<tr>
<td>Experienced Stack Exchange user</td>
<td>onetime +100</td>
</tr>
<tr>
<td>Accepted answer to bounty</td>
<td>+bounty</td>
</tr>
<tr>
<td>Offer bounty on question</td>
<td>-bounty</td>
</tr>
</tbody>
</table>

http://stackoverflow.com/faq#reputation
SO Reputation

• Assumption: reputation indicates expertise

• Expert SO users:
  • top 1% (13087 users)
  • reputation $\geq 2400$
• '09-'10: change in reputation scheme
  • Rewarding users who provide good A rather than Q
  • Q upvote: +10 → +5
• Log-logistic pattern with some deviations:
  1. Lower-end is discretized (mixture of log-logistic functions)
2. User sharing among Stack Exchange websites

- 100 rep bonus for users with rep > 200
- New SO account: 101 rep
- Old SO account: +100 rep
SO Interaction Graph

- Nodes = Users
- Edges define interactions:
  1. \( \langle \text{User asked Q} \rangle \rightarrow \langle \text{User answered} \rangle \)
  2. \( \langle \text{User asked Q} \rangle \rightarrow \langle \text{User answered } \text{accepted A} \rangle \)
  3. \( \langle \text{User asked Q} \rangle \rightarrow \langle \text{User answered } \text{upvoted A} \rangle \)

The latter two graphs represent a more meaningful interaction, since the answerer is acknowledged of providing useful information.
PageRank: Not Correlated with Reputation

- PR is based on graph connectivity
- PR is better correlated with degree than reputation
- PR distribution is similar over all three interaction graphs
Explaining Anomalous Users with High PageRank

- **Highlighted**: 5 users with high PR and rep=1
- These users had their accounts temporarily suspended for *problematic behavior* (e.g. serial up- or down-voting)
- 4/5 have high rep online and in old SO snapshot (3K-47K)
- 1/5 still suspended
Singular Value Decomposition (SVD)

- The SVD of an adjacency matrix, $A$, is
  \[ A = U \times \Sigma \times V^T \]
- Columns of $U$: left-singular vectors
  - Eigen-vectors of $AA^T$
- Columns of $V$: right-singular vectors
  - Eigen-vectors of $A^TA$
Singular Value Decomposition (SVD)

- Using $A = U \times \Sigma \times V^T$

- Identify anomalous questioners using first columns of $U$ ($U_1, U_2, ...$)

- Identify anomalous answerers using first columns of $V$ ($V_1, V_2, ...$)
Anomalous Questioners

- Have high reputation: 1K - 3K
- Mainly earned by asking questions

Answer-to-Question Z-score

$$\frac{a - q}{\sqrt{a + q}}$$

<table>
<thead>
<tr>
<th></th>
<th>All Users</th>
<th>Anomalous Questioners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-9.84</td>
</tr>
</tbody>
</table>
Anomalous Answerers

- Among highest reputation of SO users: 194K - 465K
- Mainly earned for helpful (accepted) answers

Answer-to-Question Z-score

\[
\frac{a - q}{\sqrt{a + q}}
\]

All Users | -0.04
---|---
Anomalous Answerers | 108.63
User Contributions Over Time

Cumulative mean answers per month

Cumulative mean questions per month

Cumulative mean upvoted answers per month

Cumulative mean accepted answers per month

- Rep ≥ 2400
- 1 < Rep < 400
• Follows log-linear growth for most of the users’ activity time on site → predictable pattern of site usage

• Expert users answer/ask more Q a month
Identifying Expert Users

- The analysis shows that expert users contribute more to SO throughout their time on the site.
- This indicates that one can predict which users will become experts based on their early interaction patterns.
Identifying Expert Users

Problem Statement:
Given information of a user’s activity on SO in the first $N$ months, we classify this user into one of two classes expert, or non-expert

<table>
<thead>
<tr>
<th>Label</th>
<th>Reputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>&gt; 2400</td>
</tr>
<tr>
<td>Non Expert</td>
<td>&lt; 2400</td>
</tr>
</tbody>
</table>
Experimental Setup

- Filter out users that are not active on SO for at least a year.
- Ground truth labels are based on the current reputation.
- Train/test sets are split such that the reputation $r$ of users is

<table>
<thead>
<tr>
<th>% users</th>
<th>Min rep</th>
<th>Max rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/3</td>
<td>1</td>
<td>400</td>
</tr>
<tr>
<td>1/3</td>
<td>400</td>
<td>2400</td>
</tr>
<tr>
<td>1/3</td>
<td>2400</td>
<td></td>
</tr>
</tbody>
</table>
User Activity Model

- Answers
- Questions
- Accepted
- Upvoted
- Upvotes
- Comments
- QA Ratio
- AA Ratio
- UA Ratio
Summary

• We analyzed the SO reputation scheme:
  • PageRank is not well correlated with user expertise but is effective in detecting anomalous users
  • Both experts and non-experts exhibit log-linear growth in their engagement on the site
  • Expert users contribute drastically more as soon as they join the site
  • They can be identified reliably within a month of use