IDENTIFYING THE INFLUENTIAL BLOGGERS IN A COMMUNITY

Nitin Agarwal, Huan Liu, Lei Tang
Computer Science & Engineering
Arizona State University
Tempe, AZ 85287-8809

Philip S. Yu
University of Illinois at Chicago
Chicago, IL 60607
OUTLINE

- Introduction
- Importance
- A Preliminary Model
- Challenges
- Experiments and Results
- Future Work
INTRODUCTION

- Past 15 years Computers and Internet have revolutionized the communication.
- People can connect with each other beyond all geographical barriers, across different time zones.
- Humongous mesh of social interactions: Social Network.

- Web 2.0 has catalyzed this process with easy-to-use interface and desktop like experience.
### Blog Sites

- Individual blogs
- Community blogs

<table>
<thead>
<tr>
<th>Individual Blog Sites</th>
<th>Community Blog Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owned and maintained by individual users.</td>
<td>Owned and maintained by a group of like-minded users.</td>
</tr>
<tr>
<td>More like personal accounts, journals or diaries.</td>
<td>More like discussion forums and discussion boards.</td>
</tr>
<tr>
<td>No or almost negligible group interaction.</td>
<td>High degree of group discussion and collaboration.</td>
</tr>
<tr>
<td>No or almost negligible collective wisdom.</td>
<td>Enormous collective wisdom and open source intelligence.</td>
</tr>
</tbody>
</table>
Physical and Virtual World

Domain Expert

Friends

Online Community

Physical World

Virtual World
INTRODUCTION

Inspired by the analogy between real-world and blog communities, we answer:

Who are the influentials in Blogosphere?
Can we find them?

Active Bloggers = Influential Bloggers

• Active bloggers may not be influential
• Influential bloggers may not be active
Why are the influentials interesting

- Market Movers: “word-of-mouth”, trust and reputation
- Sway opinions: Government policies, campaign
- Customer Support & Troubleshooting
- Market research surveys: “use-the-views”
- Representative articles: 18.6 new blog posts per sec
- Advertising
SEARCHING THE INFLUENTIALS

- Active bloggers
  - Easy to define
  - Often listed at a blog site
  - Are they necessarily influential

- How to define an influential blogger?
  - Influential bloggers have influential posts
  - Subjective
  - Collectable statistics
  - How to use these statistics
INTUITIVE PROPERTIES

- Social Gestures (*statistics*)
  - **Recognition**: Citations (incoming links)
    - An influential blog post is recognized by many. The more influential the referring posts are, the more influential the referred post becomes.
  - **Activity Generation**: Volume of discussion (comments)
    - Amount of discussion initiated by a blog post can be measured by the comments it receives. Large number of comments indicates that the blog post affects many such that they care to write comments, hence influential.
  - **Novelty**: Referring to (outgoing links)
    - Novel ideas exert more influence. Large number of outlinks suggests that the blog post refers to several other blog posts, hence less novel.
  - **Eloquence**: “goodness” of a blog post (length)
    - An influential is often eloquent. Given the informal nature of Blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so.

- *Influence Score* = $f$(*Social Gestures*)
A PRELIMINARY MODEL

Additive models are good to determine the combined value of each alternative [Fensterer, 2007]. It also supports preferential independence of all the parameters involved in the final decision. A weighted additive function can be used to evaluate trade-offs between different objectives [Keeney and Raiffa, 1993].

\[
\text{InfluenceFlow}(p) = w_{in} \sum_{m=1}^{\lvert \Theta \rvert} I(p_m) - w_{out} \sum_{n=1}^{\lvert \Theta \rvert} I(p_n)
\]

\[
I(p) \propto w_{comm} \gamma_p + \text{InfluenceFlow}(p)
\]

\[
I(p) = w(\lambda) \times (w_{comm} \gamma_p + \text{InfluenceFlow}(p))
\]

\[
i\text{Index}(B) = \max(I(p_1))
\]
UNDERSTANDING THE INFLUENTIALS

- Are influential bloggers simply active bloggers?
- If not, in what ways are they different?
  - Can the model differentiate them?
- Are there different types of influential bloggers?
- What other parameters can we include to evolve the model?
- Are there temporal patterns of the influential bloggers?
How to Evaluate the Model

- Where to find the ground truth?
  - Lack of Training and Test data
  - Any alternative?

- About the parameters
  - How can they be determined
  - Are they all necessary?
  - Are any of these correlated?

- Data collection
  - A real-world blog site
  - “The Unofficial Apple Weblog”
Active & Influential Bloggers

- Active and Influential Bloggers
- Inactive but Influential Bloggers
- Active but Non-influential Bloggers

We don’t consider “Inactive and Non-influential Bloggers”, because they seldom submit blog posts. Moreover, they do not influence others.
Lesion Study

- To observe if any parameter is irrelevant.
OTHER PARAMETERS

- Rate of Comments

"Spiky" comments reaction

"Flat" comments reaction
TEMPORAL PATTERNS OF INFLUENTIAL BLOGGERS

- Long term Influentials
- Average term Influentials
- Transient Influentials
- Burgeoning Influentials
VERIFICATION OF THE MODEL

- Revisit the challenges
  - No training and testing data
  - Absence of ground truth
  - Subjectivity

- We use another Web 2.0 website, Digg as a reference point.

  “Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that’s important to you!”

- The higher the digg score for a blog post is, the more it is liked.

- A not-liked blog post will not be submitted thus will not appear in Digg.
Verification of the Model

- Digg records top 100 blog posts.

- Top 5 influential and top 5 active bloggers were picked to construct 4 categories

- For each of the 4 categories of bloggers, we collect top 20 blog posts from our model and compare them with Digg top 100.

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 17</td>
<td>S2: 7</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 3</td>
<td>S4: 0/1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 71</td>
<td>S2: 14</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 8</td>
<td>S4: 7</td>
</tr>
</tbody>
</table>

- Distribution of Digg top 100 and TUAW’s 535 blog posts
**Verification of the Model**

- Observe how much our model aligns with Digg.
- Compare top 20 blog posts from our model and Digg.
- Considered last six months

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All-in</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>No Inlinks</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No Comments</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>No Outlinks</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>No Blog post length</td>
<td>12</td>
<td>14</td>
<td>11</td>
<td>15</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

- Considered all configuration to study relative importance of each parameter.
- **Inlinks > Comments > Outlinks > Blog post length**
**POTENTIAL APPLICATIONS**

- Improving the preliminary model
  - Can we involve more parameters?
  - Quality vs. Quantity of comments
  - "Goodness" of blog post estimation techniques
  - Can we learn the model weights given various statistics
    - Each weight parameter likely follows its own distribution

- Community evolution
  - How does a community evolve around the influentials?
  - Do the influentials cause topic drift and how?
  - Can we experimentally study the roles and impact of the influentials?
POTENTIAL APPLICATIONS

- Trust and reputation
  - How can this work help in studying trust and reputation
    - Intuitively, an influential one is usually trustworthy
  - Trust initialization
    - Existing work focus on trust propagation
  - Is trust a serious issue on the blogosphere?
    - Splogs and collective wisdom
    - Important and sensitive in friendship networks

- Expert identification
  - Identifying the influentials on a set of blog sites of common topic theme: Experts
  - Comparing the influentials from different blog sites
    - Normalizing various collectable statistics across different blog sites
CONCLUDING REMARKS

- Ample opportunities for influential bloggers
- Influence: A subjective concept
- Challenges:
  - Model development
  - Evaluation & Verification
  - Data collection