Identifying Influential Users’ Professions via the Microblogs They Forward

Yuan Wang, Hangyu Mao, Zhen Xiao
Peking University, China
SocInf2017, 19 Aug 2017
Outline

• Background
• Data
• Our Method
• Evaluation
• Conclusion
Introduction
users are mainly organized by their professions
Message forwarding

Weibo messages

Forwarded messages 60%

Non-forwarded messages 40%

@Terminal
3月7日 14:19 来自 iPhone 6
第二个工作我们也在做。

@王威廉
今天陪了谷歌Jeff Dean一天，总结几个有趣的事情：1）谷歌邮箱自动回复功能早在2009年还是愚人节玩笑，到2015年后被实现应用了。2）谷歌正在用强化学习研究如何给GPU和CPU分别分配计算任务。3）谷歌TIP论文刚被ISCA 2017接受了，论文马上放出来。🤔
Challenge

Too many messages
Challenge

Weibo messages

Forwarded messages

Hot Forwarded messages

Non-Hot Forwarded messages

Non-forwarded messages

60%

40%
Our Framework

Fig. 1. The framework of PIFB
Our Data

41,531 manually annotated influential users

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Media</td>
<td>26.3</td>
</tr>
<tr>
<td>2</td>
<td>Entertainment</td>
<td>10.1</td>
</tr>
<tr>
<td>3</td>
<td>Estate</td>
<td>9.1</td>
</tr>
<tr>
<td>4</td>
<td>Finance</td>
<td>8.6</td>
</tr>
<tr>
<td>5</td>
<td>Government</td>
<td>8.5</td>
</tr>
<tr>
<td>6</td>
<td>IT</td>
<td>8.4</td>
</tr>
<tr>
<td>7</td>
<td>Sports</td>
<td>6.4</td>
</tr>
<tr>
<td>8</td>
<td>Fashion</td>
<td>6.2</td>
</tr>
<tr>
<td>9</td>
<td>Education</td>
<td>5.9</td>
</tr>
<tr>
<td>10</td>
<td>Literature</td>
<td>5.4</td>
</tr>
<tr>
<td>11</td>
<td>Game</td>
<td>5.1</td>
</tr>
</tbody>
</table>
Absolutely Hot Message

- Number of publisher’s followers
- Number of been forwarded times

<table>
<thead>
<tr>
<th>Message 1</th>
<th>Message 2</th>
<th>Message 3</th>
<th>Message 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot</td>
<td>Standard</td>
<td>Hot</td>
<td>Non-hot or Hot</td>
</tr>
</tbody>
</table>

Number of publisher’s followers

Non-hot or Hot

Number of been forwarded times
Relatively Hot Message

Number of publisher’s followers
Number of been forwarded times

Message 1
Message 2
Message 3
Message 4
Group Similar Hot Messages Together

- Simhash
- Paragraph2vec
- Matrix factorization
Simhash

1. feature, weight
   - \( w_1 \)
   - \( w_2 \)
   - \( w_n \)

2. hash, weight
   - 100110 \( w_1 \)
   - 110000 \( w_2 \)
   - 001001 \( w_n \)

3. \( w_1 - w_1 - w_1 w_1 w_1 - w_1 \)
   - \( w_2 w_2 - w_2 - w_2 w_2 - w_2 \)
   - \( -w_n - w_n w_n - w_n w_n - w_n \)

4. add
   - sign
   - 13, 108, -22, -5, -32, 55

5. fingerprint
   - 110001
Paragraph2Vec
User-Weibo Matrix Factorization

User \rightarrow \text{forward} \rightarrow \text{Message}

\approx \begin{pmatrix} U(i) \end{pmatrix} \times \begin{pmatrix} M(j) \end{pmatrix}

User-weibo relationship matrix M
Evaluation

Users forward how many massages in their latest 500 messages

- Number of forwards
- Professions: Me, En, Es, Fi, Go, IT, Sp, Fa, Ed, li, Ga

Users forward how many massages in their latest 500 messages

- Number of users
- Forward how many massages: 0 to 500
Evaluation

- Absolute hot
- Relative hot

Table 2. Number of absolutely hot messages

<table>
<thead>
<tr>
<th>No.</th>
<th>Threshold</th>
<th># before filter</th>
<th># after filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
<td>731,150</td>
<td>100,219</td>
</tr>
<tr>
<td>2</td>
<td>2000</td>
<td>426,019</td>
<td>82,339</td>
</tr>
<tr>
<td>3</td>
<td>10000</td>
<td>74,308</td>
<td>32,955</td>
</tr>
</tbody>
</table>
### Table 3. Number of messages under different merging strategies

<table>
<thead>
<tr>
<th>No.</th>
<th>Merging Strategy</th>
<th># before</th>
<th># after</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simhash</td>
<td>162,025</td>
<td>57,624</td>
</tr>
<tr>
<td>2</td>
<td>P2V</td>
<td>162,025</td>
<td>32,118</td>
</tr>
<tr>
<td>3</td>
<td>UWMF</td>
<td>162,025</td>
<td>27,129</td>
</tr>
<tr>
<td>4</td>
<td>Simhash+P2V+UWMF</td>
<td>162,025</td>
<td>17,196</td>
</tr>
</tbody>
</table>
Evaluation

Hot messages

17,196

162,025

tweet
tweet
tweet
tweet
retweet
retweet
retweet
retweet
merge

XGBoost

17,196

162,025
Baseline

Words in original messages
Words in forwarded messages
URLs in messages
User id
Hashtag

Discard the sparse feature

Words in original messages
Words in forwarded messages
URLs in messages
User id
Hashtag

9200 features
XGBoost

Chi-square statistic
## Evaluation

<table>
<thead>
<tr>
<th>No.</th>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Baseline</td>
<td>62.38</td>
<td>64.03</td>
<td>60.29</td>
<td>62.10</td>
</tr>
<tr>
<td>2</td>
<td>Simhash</td>
<td>69.24↑6.86%</td>
<td>70.88</td>
<td>67.61</td>
<td>69.21↑7.11%</td>
</tr>
<tr>
<td>3</td>
<td>Simhash+P2V</td>
<td>73.79↑11.41%</td>
<td>73.90</td>
<td>71.28</td>
<td>72.57↑10.47%</td>
</tr>
<tr>
<td>4</td>
<td>Simhash+P2V+UWMF</td>
<td>73.98↑11.60%</td>
<td>74.81</td>
<td>72.95</td>
<td>73.87↑11.77%</td>
</tr>
</tbody>
</table>

Table 4: Evaluation results for various features and combinations. (％)
<table>
<thead>
<tr>
<th></th>
<th>Me</th>
<th>En</th>
<th>Es</th>
<th>Fi</th>
<th>Go</th>
<th>IT</th>
<th>Sp</th>
<th>Fa</th>
<th>Ed</th>
<th>Li</th>
<th>Ga</th>
</tr>
</thead>
<tbody>
<tr>
<td>Me</td>
<td>76.7</td>
<td>5.6</td>
<td>2.7</td>
<td>3.5</td>
<td>4.1</td>
<td>3.3</td>
<td>2.2</td>
<td>0.9</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>En</td>
<td>7.2</td>
<td>74.5</td>
<td>0.2</td>
<td>3.3</td>
<td>0.7</td>
<td>1.4</td>
<td>4.4</td>
<td>5.1</td>
<td>0.2</td>
<td>1.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Es</td>
<td>7.4</td>
<td>2.0</td>
<td>72.9</td>
<td>8.5</td>
<td>5.3</td>
<td>2.2</td>
<td>0.4</td>
<td>0.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Fi</td>
<td>8.4</td>
<td>0.1</td>
<td>6.4</td>
<td>70.2</td>
<td>5.3</td>
<td>6.2</td>
<td>0.2</td>
<td>1.3</td>
<td>1.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Go</td>
<td>4.9</td>
<td>2.2</td>
<td>0.4</td>
<td>4.2</td>
<td>78.2</td>
<td>2.9</td>
<td>4.1</td>
<td>0.4</td>
<td>2.5</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>IT</td>
<td>6.1</td>
<td>0.7</td>
<td>3.9</td>
<td>4.3</td>
<td>1.3</td>
<td>76.3</td>
<td>0.2</td>
<td>0.1</td>
<td>2.6</td>
<td>0.7</td>
<td>3.8</td>
</tr>
<tr>
<td>Sp</td>
<td>5.1</td>
<td>2.9</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.0</td>
<td>86.2</td>
<td>2.2</td>
<td>0.7</td>
<td>0.0</td>
<td>1.3</td>
</tr>
<tr>
<td>Fa</td>
<td>9.7</td>
<td>14.9</td>
<td>1.0</td>
<td>6.2</td>
<td>0.2</td>
<td>0.0</td>
<td>3.3</td>
<td>61.5</td>
<td>0.9</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Ed</td>
<td>5.2</td>
<td>3.9</td>
<td>3.3</td>
<td>4.6</td>
<td>2.0</td>
<td>3.2</td>
<td>0.7</td>
<td>1.8</td>
<td>68.4</td>
<td>4.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Li</td>
<td>13.7</td>
<td>7.2</td>
<td>0.7</td>
<td>1.3</td>
<td>0.6</td>
<td>1.4</td>
<td>0.4</td>
<td>3.3</td>
<td>9.8</td>
<td>60.9</td>
<td>0.7</td>
</tr>
<tr>
<td>Ga</td>
<td>5.2</td>
<td>3.9</td>
<td>0.8</td>
<td>0.0</td>
<td>0.4</td>
<td>7.1</td>
<td>1.2</td>
<td>4.3</td>
<td>0.1</td>
<td>0.3</td>
<td>76.7</td>
</tr>
</tbody>
</table>

Table 5: Distribution of identified professions in each profession.
Conclusion

• find the hot weibo messages
• group similar hot messages together
• design a classifier to predict users’ professions
Questions