Understanding Social Movement by Tracking the Keyword in Social Media

Virach Sornlertlamvanich, virach@gmail.com
Eakasit Pacharawongsakda, sit.ake@gmail.com
Sirindhorn International Institute of Technology (SIIT), Thammasat University
Thatsanee Charoenporn, thatsanee@gmail.com
Burapha University

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
THAILAND

67,091,000
TOTAL POPULATION

34%
URBAN
66%
RURAL

24,000,000
INTERNET USERS

36%
INTERNET PENETRATION

16,834,140
USERS ON TOP SOCIAL NETWORK

25%
SOCIAL NETWORKING PENETRATION

78,667,910
MOBILE SUBSCRIBERS

117%
MOBILE PENETRATION

http://tulaneict4d.wordpress.com/2013/04/05/social-media-in-thailand/

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Thailand Monthly Active Users on Social Network and Messenger App in 2013

Data Preparation

• Word segmentation
• Keyword: extracted from topic related documents (training set)
• Tweeter inquiry: using the prepared topic related list of keywords
• Text similarity: using GETA algorithm
GETA: Association Search Engine

- Proposed by Akihiko Takano (Takano, 2003)
- Back-and-forth searches between words and documents spaces
  - Set of keywords → The top-n ranks of associated documents
  - Set of documents → The top-m ranks of associated keywords
- Require a **Word Article Matrix (WAM)**, and a similarity function

![Diagram of GETA model](Diagram.png)

Keywords → GETA → Top-n Ranked Associated Documents
Top-m Ranked Associated Keywords → GETA → Documents

**Word Article Matrix (WAM)**

**A similarity function**

**Word space**

**Document space**

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Word Article Matrix (WAM)

Creating WAM

Wikipedia Articles

Page Contents

Word seg. & Lemmatization
Thinking -> Think

Word Article Matrix (WAM)
A similarity function

Keywords

Top-Ranked Associated Keywords

G ETA

Top-Ranked Associated Documents

Keywords

Word space

Document space

Word list

Pages\Words | “Twitter” | “Tennis” | “Dollar” | “Google” | ...
---|---|---|---|---|---
IT | 2 | 0 | 1 | 4 | ...
Sport | 0 | 2 | 1 | 0 | ...
Economics | 0 | 0 | 2 | 0 | ...
...

Wikipedia WAM

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Similarity Functions

- **Ranking weight** in both words and documents spaces
- **Equation form**

\[
\text{SIM}(d, q) = \sum_{t \in q} \frac{w_q(t, q) \cdot w_d(t, d)}{\text{norm}(d, q)}
\]

- \(w_d(t, d)\) = Weight of term \(t\) in document \(d\), \(w_d(t, d) = 0\) if \(t \not\in d\)
- \(w_q(t, q)\) = Weight of term \(t\) in query sentence \(q\), \(w_q(t, q) = 0\) if \(t \not\in q\).
- \(\text{norm}(d, q)\) = Normalization function due to the different length of \(d\) and \(q\).
- **Smart measure** (Singhal et.al., 1996)

\[
\frac{1}{\text{avg}(f_d) + \theta(f_d - \text{avg}(f_d))} \sum_{t \in q \land t \in d} \log\left(\frac{N}{f_t}\right) \cdot \frac{1 + \log(f_{d,t})}{1 + \log(\text{avg}_{w \in d}(f_{d,w}))} \cdot \frac{1 + \log(f_{q,t})}{1 + \log(\text{avg}_{w \in q}(f_{q,w}))}
\]

- **Dot Product** (Wilkinson et.al., 1996)

\[
\sum_{t \in q \land t \in d} (w_{q,t} \cdot w_{d,t})
\]
Example GETA Calculation

SIM Function: Dot Product

$$\sum_{t \in \text{doc}} (w_{q,t} \cdot w_{d,t})$$

“Twitter has 800M dollars”

Word seg. & Lemmatization dollars -> dollar

“Twitter” “Tennis” “Dollar” “Google”

1 0 1 0

(n=2) Select Top-n Most Associated Documents Ranking

<table>
<thead>
<tr>
<th>Pages</th>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wikipedia WAM

Word Article Matrix (WAM) A similarity function

Keywords

Top-Ranked Associated Documents

Word space

Document space

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Example GETA Calculation

SIM Function: Dot Product

\[ \sum_{t \in \text{query}} (w_{q,t} \cdot w_{d,t}) \]

“Twitter has 800M dollars”

Word seg. & Lemmatization dollars -> dollar

<table>
<thead>
<tr>
<th>Pages/Words</th>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wikipedia WAM

(n=2) Select Top-n Most Associated Documents Ranking

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Example GETA Calculation (Cont.)

SIM Function: Dot Product \( \sum_{i \in q \land d} (w_{q,i} \cdot w_{d,i}) \)

Top-2 Most Associated Documents

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
</tbody>
</table>

Output of GD

| Pages\Words | “Twitter” | “Tennis” | “Dollar” | “Google” | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wikipedia WAM

(m=3) Top-m Most Associated Keywords

<table>
<thead>
<tr>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>
Example GETA Calculation (Cont.)

SIM Function: Dot Product $\sum_{t \in q \cap d} (w_{q,t} \cdot w_{d,t})$

Top-2 Most Associated Documents

<table>
<thead>
<tr>
<th>Pages</th>
<th>Dot Product Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>3</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>2</td>
</tr>
</tbody>
</table>

*Sport Pages = 0, since only Top-2 is selected from previous slide.

“Twitter has 800M dollars”

Google, Dollar, Twitter

Wikipedia WAM

(m=3) Top-m Most Associated Keywords

<table>
<thead>
<tr>
<th>“Twitter”</th>
<th>“Tennis”</th>
<th>“Dollar”</th>
<th>“Google”</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>
Modified WAM for Social Media Text Classification

Wikipedia / Online news document

Word segmentation
Stematization
TF/IDF for keyword

i-WAM

Keyword

Social media text

Manual annotation

Domain annotated text

Word segmentation
Stematization
TF/IDF for keyword

TF merging

m-WAM

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
# A Part of Modified WAM

<table>
<thead>
<tr>
<th></th>
<th>i-WAM</th>
<th>m-WAM</th>
<th>m-WAM2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ชื่อเพล 'intel'</td>
<td>ชื่อเพล 'intel'</td>
<td>ชื่อเพล 'intel'</td>
</tr>
<tr>
<td></td>
<td>ศัลยกรรม 'surgery'</td>
<td>ศัลยกรรม 'surgery'</td>
<td>ศัลยกรรม 'surgery'</td>
</tr>
<tr>
<td></td>
<td>เมตตา 'mercy'</td>
<td>เมตตา 'mercy'</td>
<td>เมตตา 'mercy'</td>
</tr>
<tr>
<td></td>
<td>คลาตด 'cloud'</td>
<td>คลาตด 'cloud'</td>
<td>คลาตด 'cloud'</td>
</tr>
<tr>
<td></td>
<td>คลิป 'clip'</td>
<td>คลิป 'clip'</td>
<td>คลิป 'clip'</td>
</tr>
<tr>
<td></td>
<td>เม็ต 'network'</td>
<td>เม็ต 'network'</td>
<td>เม็ต 'network'</td>
</tr>
<tr>
<td>Life</td>
<td>0.052</td>
<td>0.026</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>0.205</td>
<td>0.275</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>0.160</td>
<td>0.268</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td>0.168</td>
<td>0.177</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.036</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.040</td>
<td>0.034</td>
</tr>
<tr>
<td>Education</td>
<td>0</td>
<td>0.013</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>0.025</td>
<td>0.018</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.187</td>
<td>0.019</td>
</tr>
<tr>
<td>Technology</td>
<td>0.103</td>
<td>0.164</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.019</td>
</tr>
</tbody>
</table>
F-Measure

<table>
<thead>
<tr>
<th></th>
<th>Life</th>
<th>Education</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-WAM on text1</td>
<td>86.79%</td>
<td>36.09%</td>
<td>55.79%</td>
</tr>
<tr>
<td>m-WAM on text1</td>
<td>86.64%</td>
<td>34.57%</td>
<td>56.01%</td>
</tr>
<tr>
<td>m-WAM on text2</td>
<td>91.33%</td>
<td>29.91%</td>
<td>43.17%</td>
</tr>
<tr>
<td>m-WAM2 on text2</td>
<td>94.93%</td>
<td>29.47%</td>
<td>43.72%</td>
</tr>
</tbody>
</table>
Tweet query by keyword

Word Article Matrix (WAM)

Domain specific tweets

Social movement timeline viewer

Tweeter Viewer

<table>
<thead>
<tr>
<th>Pages</th>
<th>&quot;Twitter&quot;</th>
<th>&quot;Tennis&quot;</th>
<th>&quot;Dollar&quot;</th>
<th>&quot;Google&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Economics</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Coup on May 22, 2014

• military, NCPO, country, announce, peace, power, government, coup d’état, gathering, police, situation, PM, control, seize, meeting, economy, law, war, leader, minister, election, democracy, revolution, seize the power, curfew, martial law
Tweet Query

• Search Tweets by using Restful API
  – GET search/tweets. Set q = the keyword set
  – 100 tweets/search limited
  – Repeatedly fetch data until all tweets in the coup periods are discovered

• Be able to search back to 7 days

May 22, 2014 Coup-related tweet : 339,148 tweets
Coup D'etat

Timeline Word Cloud

MAPLEX2015, February 9-10, 2015, Yamagata, Japan
Conclusion

• Key word expansion is effective to understand the short message i.e. tweet
• Key word expansion can be done using the known training corpus and GETA algorithm
• Timeline word cloud shows the development of the social movement e.g. some events are predictable

MAPLEX2015, February 9-10, 2015, Yamagata, Japan