Bridging the Information Gap between Urban and Rural Lifestyle
---Collect, Analyze, Visualize---

Virach Sornlertlamvanich
virach@gmail.com
Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand

Global Expert Meeting on Multilingualism in Cyberspace
Khanty-Mansiysk, Russia Federation
July 4-9, 2015
Purpose

• To deliver local news to complement the deteriorating local newspaper
• To promote data usage in rural area

To improving rural livelihood through its reputation extraction!
What kind of ‘Smart’ in NEWS publishing?

• News are collected from public sources i.e. online published news, government portal news, wikipedia, social media, classified, etc. specified by location.

• News reported for the same incident are grouped automatically by ‘keyword based text similarity’ algorithm.

• Similar news from several publishers and sources are aggregated and summarized as ‘word cloud’.

• Incidents are sorted based on reported time and can be automatically extended to a timeline publishing.

• News is automatically classified into its plausible categories i.e. location-wise and content-wise.
# 4 Sources of Data / 3 Processes

## Source of Data
- **Information web**
  - Ecommerce, Community, Agriculture, Education, etc.
- **Knowledge web**
  - Open Data=>Linked Open Data
  - Wikipedia=>DBpedia=>Knowledge web
- **Social media**
  - Twitter, FB, LINE, etc.
- **News web**
  - Online news, tabloid, magazine, etc.

## Process
- Collect
- Analyze
- Visualize
1-Collect

• Web crawling i.e. web archiving
• Wikipedia Infobox harvesting
• Topic based message extraction
• Template based crawling
Collecting information

- Information Database
- Fact Database
- Social Database
- News Database
- Web spider
- Reliable Knowledge Downloader
- Social Media Querier
- News aggregators
- Gov
- Others
- Wiki
- FB
- Twitter
- Nation
- Daily news
- Others
- News Publishers

Internet

Ugra Global Expert Meeting on Multilingualism in Cyberspace Khanty-Mansiysk, Russia Federation, July 4-9, 2015
2-Analyze

• **NEWS Classification**
  – Word segmentation
  – Named Entity Recognition
  – NEWS domain and province classification and ranking based on TF/IDF technique

• **Information Extraction**
  – Template based extraction

• **Infobox**
  – Celebrity i.e. politician, idol, actors, singers
  – Province

• **Social Media**
  – Trending topic
  – Leader and follower
  – Opinion polarity
News Classification
3-Visualize

- Classified news
- **Word cloud**
- Document similarity based ranking
- *Infobox*
### News Portals

<table>
<thead>
<tr>
<th>Source</th>
<th>Amount of news captured</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bangkokpost Manager</td>
<td>20,602</td>
<td>March 15, 2014</td>
<td>September 2, 2014</td>
</tr>
<tr>
<td><a href="http://www.bangkokpost.com">www.bangkokpost.com</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nation Manager</td>
<td>90,114</td>
<td>January 1, 2014</td>
<td>September 2, 2014</td>
</tr>
<tr>
<td><a href="http://www.manager.co.th">www.manager.co.th</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thairath</td>
<td>43,275</td>
<td>February 14, 2013</td>
<td>September 2, 2014</td>
</tr>
<tr>
<td><a href="http://www.thairath.co.th">www.thairath.co.th</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thaipost</td>
<td>40,420</td>
<td>January 30, 2014</td>
<td>September 2, 2014</td>
</tr>
<tr>
<td><a href="http://www.thaipost.net">www.thaipost.net</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dailynews</td>
<td>26,699</td>
<td>December 21, 2011</td>
<td>July 13, 2014</td>
</tr>
<tr>
<td><a href="http://www.dailynews.co.th">www.dailynews.co.th</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naewna</td>
<td>59,284</td>
<td>June 7, 2009</td>
<td>September 2, 2014</td>
</tr>
<tr>
<td><a href="http://www.naewna.com">www.naewna.com</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naewna</td>
<td>18,343</td>
<td>April 29, 2013</td>
<td>September 1, 2014</td>
</tr>
</tbody>
</table>

Ugra Global Expert Meeting on Multilingualism in Cyberspace Khanty-Mansiysk, Russia Federation, July 4-9, 2015
<table>
<thead>
<tr>
<th>Section</th>
<th>Source</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTOP</td>
<td>OTOP Product <a href="http://www.cddopc.com/otopselect2013">www.cddopc.com/otopselect2013</a></td>
<td>3,207 products</td>
</tr>
<tr>
<td>Property</td>
<td>DDProperty <a href="http://www.ddproperty.com">www.ddproperty.com</a></td>
<td>89,304 places or locations</td>
</tr>
<tr>
<td>Job</td>
<td>JobThai <a href="http://www.jobthai.com">www.jobthai.com</a></td>
<td>17,477 jobs</td>
</tr>
<tr>
<td>Procurement</td>
<td>Thai Government Procurement <a href="http://www.gprocurement.go.th">www.gprocurement.go.th</a></td>
<td>62,795 posts</td>
</tr>
<tr>
<td>Weather</td>
<td>OWM <a href="http://www.openweathermap.org">www.openweathermap.org</a></td>
<td>5,937 days</td>
</tr>
<tr>
<td>Restaurant</td>
<td>Wongnai <a href="http://www.wongnai.com">www.wongnai.com</a></td>
<td>99,568 restaurants</td>
</tr>
<tr>
<td>Movies</td>
<td>Major Cineplex <a href="http://www.majorcineplex.com">www.majorcineplex.com</a></td>
<td>1,091 Movie schedules</td>
</tr>
<tr>
<td>Exchange rate</td>
<td>Exchange rate (daily)</td>
<td>599,433 exchange rate (THB &lt;-&gt; Other currency for 10 years)</td>
</tr>
<tr>
<td>Gold price</td>
<td>Gold price (daily)</td>
<td>3,702 days</td>
</tr>
<tr>
<td>Stock market</td>
<td>SET Index (daily)</td>
<td>1,095 days</td>
</tr>
</tbody>
</table>

Ugra Global Expert Meeting on Multilingualism in Cyberspace Khanty-Mansiysk, Russia Federation, July 4-9, 2015
## Other Portals

<table>
<thead>
<tr>
<th>Section</th>
<th>Source</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel</td>
<td>Tourism Authority of Thailand (<a href="http://www.tourismthailand.org">www.tourismthailand.org</a>)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Accommodation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Attractions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Food and drink</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Shopping</td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td>• Facebook</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Instagram</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Google Plus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Twitter</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Youtube</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Wikipedia</td>
<td></td>
</tr>
</tbody>
</table>
Demo

• http://nation.iisilab.org/
Uber: The world’s largest taxi company, owns no vehicles.

Facebook: The world’s most popular media owner, creates no content.

Alibaba: The world’s largest accommodation provider, owns no real estate.

Airbnb: The most valuable retailer, has no inventory.

Something interesting is happening.

TOM GOODWIN
Summary

• Database created to access to a collection of NEWS, Knowledge, Information and Social media

• A portal of Hyper Local NEWS responding to the local daily life i.e. news, restaurants, movies, shopping, weather, celebrity, etc.

• News harvesting from the existing (growing) sources