The Automated Acquisition of Suggestions from Tweets

Li Dong*
SKLSDE, Beihang University
donglixp@gmail.com

Furu Wei
Microsoft Research Asia
fuwei@microsoft.com

Yajuan Duan*
University of Science and Technology of China
dyj@mail.ustc.edu.cn

Xiaohua Liu
Microsoft Research Asia
xiaoliu@microsoft.com

Ming Zhou
Microsoft Research Asia
mingzhou@microsoft.com

Ke Xu
SKLSDE, Beihang University
kexu@nlsde.buaa.edu.cn

July 16, 2013
What is suggestion?

- Suggestion: The psychological process by which one person guides the thoughts, feelings, or behavior of another.
Why do suggestions matter?

- When I arrived **Seattle**, I saw this on the window of bus:
  - on the receipt of RITE AID PHARMACY:
  - Companies try to hear the voice of users.
Why do suggestions matter?

- A novel & useful task for Business Intelligence
  - Listen to your customers
  - Help on further improving the products
  - Extension for sentiment analysis
Where can we find suggestions?

- Twitter is a good data source to find suggestions.
  - User-generated content
  - Big data can lead to big intelligence

- Examples
  - I have an idea for “Microsoft”. Make an app on WP7 that can remote login into your desktop and u can do everything. Content creation I mean
  - #microsoft #WindowsPhone7 I’d like multitasking please
Task

- Task Definition
  - Input: Tweets
  - Output: Find the suggestions

- Challenges
  - Sparsity: short text
  - Imbalance: ~7.93% of tweets are suggestions (windows phone 7)
Model

- **Factorization Machines (FM)**
  - Use few parameters to model the intersection
    \[
    \hat{y}(\mathbf{x}) := w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} \mathbf{v}_j^T \cdot \mathbf{v}_{j'} x_j x_{j'}
    \]
    Weight: dot product of two k dimension vectors

- **Compare with polynomial kernel SVM**
  \[
  \hat{y}(\mathbf{x}) := w_0 + \sqrt{2} \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} w_{j,j} x_j^2
  \]
  Weight: for each intersection
  \[
  + \sqrt{2} \sum_{j=1}^{p} \sum_{j'=j+1}^{p} w_{j,j'} x_j x_{j'}
  \]
  
Model

- **Objective function**

\[
OPT(S, \lambda) := \arg\min_{\theta} \left( - \sum_{(x,y) \in S} \ln \sigma(\hat{y}(x|\Theta) y) + \sum_{\theta \in \Theta} \lambda_{\theta} \theta^2 \right)
\]

- **Optimization (off-the-shelf methods)**
  - Stochastic Gradient Descent
  - Adaptive Stochastic Gradient Descent
  - L-BFGS
  - ...

Imbalance

- Combine two meta-methods
  - Meta-method: Without modify the original model
  - Oversampling (before training)
    - Redistribute training data set
  - Thresholding (after predicting)
    - If $p > \tau$, positive; else negative;
    - Search a good $\tau$
Feature

- N-gram features
- #hashtag features
- Template features (sequential patterns)
  - Windows Phone's official web site
  - http://windowsphone.uservoice.com
Template Features

- Use PrefixSpan algorithm to mine frequent sequential patterns efficiently

Algorithm 1 Automatic Suggestion Template Extraction

Input:
- Feedback data set;
- Parameters: minimum length, maximum length, minimum support;

Output:
- Suggestion templates;

1: Tokenize the feedback data set into sentences
2: Filter domain-related words //“lumia”, “windows” as examples in our experiments
3: Remove sentences less than 4 words in length
4: `template_list=PrefixSpan(minimum length, maximum length, minimum support)`
5: `for template_it in template_list do`
6: `if (template_it are all stopwords) then`
7: `remove template_it from template_list`
8: `end if`
9: `end for`
10: `return template_list;`
Experiment

- **Data set**
  - 3,000 tweets manually
    - Keyword: windows phone 7, wp7 [September 2010 to April 2012]
  - 238 (/3,000=7.93%) of them are suggestions
    - Imbalance
## Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Suggestion Tweets</th>
<th>Non-suggestion Tweets</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F-1</td>
</tr>
<tr>
<td>SVM with bag-of-words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+cost-sensitive</td>
<td>56.79</td>
<td>57.27</td>
<td>57.03</td>
</tr>
<tr>
<td>+all features</td>
<td>63.68</td>
<td>65.35</td>
<td>64.55</td>
</tr>
<tr>
<td>+cost-sensitive + all features</td>
<td>63.76</td>
<td>65.35</td>
<td>64.55</td>
</tr>
<tr>
<td>+cost-sensitive + all features + polynomial kernel</td>
<td>62.25</td>
<td>64.42</td>
<td>63.32</td>
</tr>
<tr>
<td>FM with bag-of-words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+cost-sensitive</td>
<td>85.74</td>
<td>24.48</td>
<td>38.09</td>
</tr>
<tr>
<td>+all features</td>
<td>60.89</td>
<td>60.79</td>
<td>60.84</td>
</tr>
<tr>
<td>+cost-sensitive + all features</td>
<td>85.37</td>
<td>43.00</td>
<td>57.19</td>
</tr>
<tr>
<td>+cost-sensitive + all features</td>
<td>71.06</td>
<td>67.86</td>
<td>69.42</td>
</tr>
</tbody>
</table>

- SVM: Support Vector Machine
- FM: Factorization Machine
Summary

- Propose the task of suggestion analysis
  - Not well studied previously, but useful

- Study of suggestion classification from Tweets
  - Use to FMs to model intersection when feature space is sparse
  - Combine oversampling & thresholding to overcome imbalance

- Release the data set for research
  - [http://goo.gl/hXtRv](http://goo.gl/hXtRv)
Future Work

- **Target/Aspect Identification**
  - I have an idea for "Microsoft". Make an app on WP7 that can remote login into your desktop and you can do everything. Content creation I mean
  - #microsoft #WindowsPhone7 I’d like **multitasking** please

- **Suggestion Summarization**
  - **Who** suggest **How** to **What**, **When**?

---

**User Interface**

- Simple
- Cute
- Beautiful

**Hardware**

- Powerful
- Low energy consumption

---

**Target**

- "Microsoft"

**Aspect**

- #microsoft #WindowsPhone7
THANKS!

Q&A

Any suggestions?