Microsoft Research
微软亚洲研究院

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

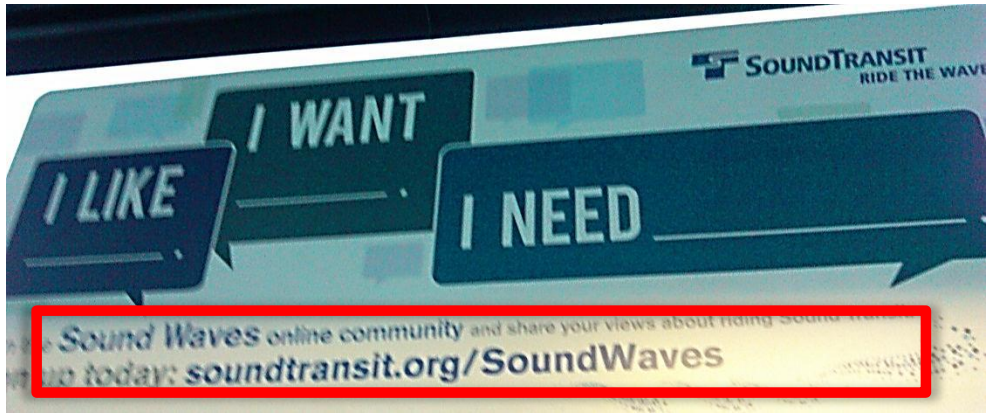
July 16, 2013

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?

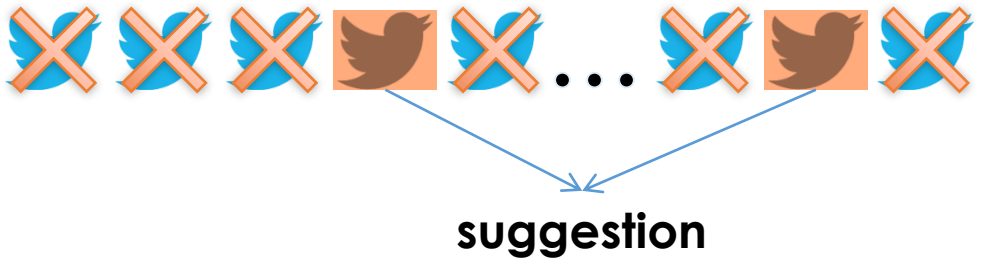
- A novel & useful task for Business Intelligence
  - Listen to your customers
  - Help on further improving the products
  - Extension for sentiment analysis

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

Weight: dot product of two k dimension vectors

$$\widehat{y}(\mathbf{x}) := w_0 + \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} \sum_{j'=j+1}^{p} \mathbf{v}_j{}^T \cdot \mathbf{v}_{j'} x_j x_{j'}$$

- Compare with polynomial kernel SVM

$$\widehat{y}(\mathbf{x}) := w_0 + \sqrt{2} \sum_{j=1}^{p} w_j x_j + \sum_{j=1}^{p} w_{j,j} x_j{}^2$$

Weight: for each intersection

$$+ \sqrt{2} \sum_{j=1}^{p} \sum_{j'=j+1}^{p} w_{j,j'} x_j x_{j'}$$

Steffen Rendle (2012): Factorization Machines with libFM, in ACM Trans. Intell. Syst. Technol., 3(3), May.

Model

- Objective function

$$OPT(S,\lambda) := \underset{\Theta}{argmin} \left(-\sum_{(\mathbf{x},y)\in S} \ln \sigma\left(\hat{y}(\mathbf{x}|\Theta)\, y\right) + \sum_{\theta\in\Theta} \lambda_\theta \theta^2 \right)$$

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

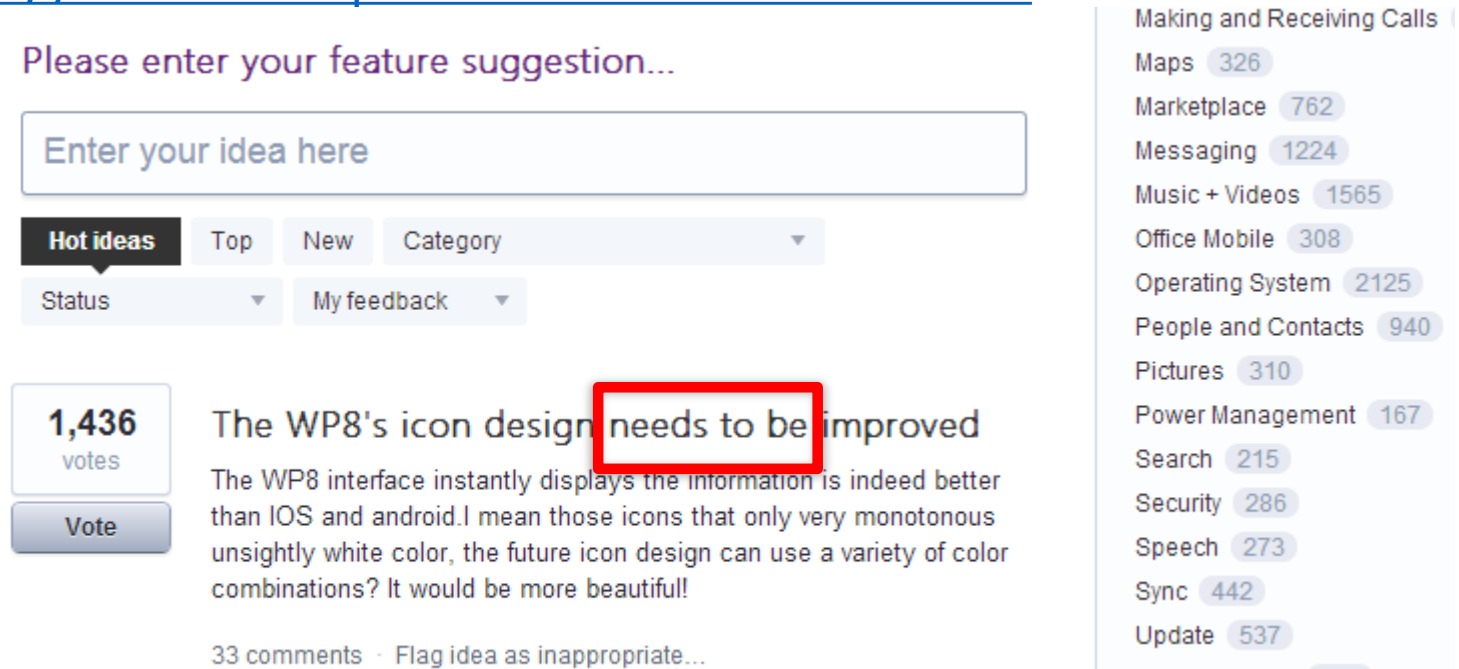
Steffen Rendle (2012): Factorization Machines with libFM, in ACM Trans. Intell. Syst. Technol., 3(3), May.

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 τ

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

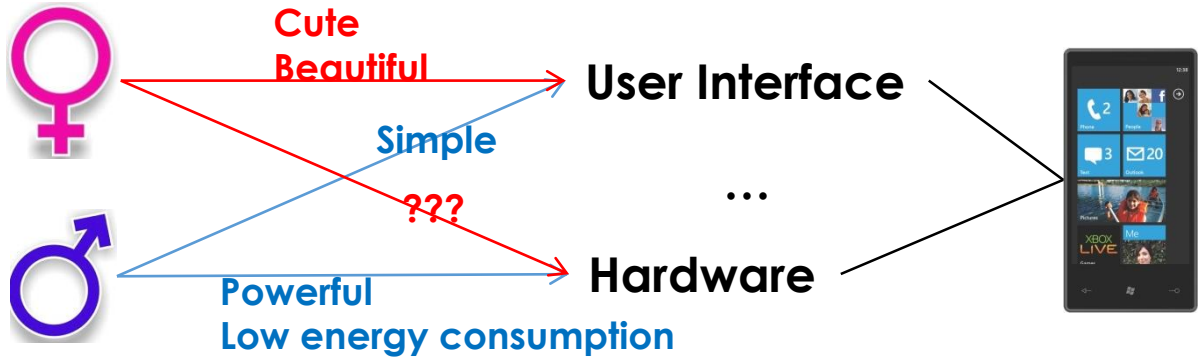
	Method	Suggestion Tweets			Non-suggestion Tweets			Acc.
		Prec.	Rec.	F-1	Prec.	Rec.	F-1	
SVM with bag-of-words	SVM1	56.96	52.18	54.47	95.82	96.53	96.17	92.33
+cost-sensitive	SVM2	56.79	57.27	57.03	96.33	96.27	96.30	93.21
+all features	SVM3	63.68	60.63	62.12	96.66	97.04	96.85	93.78
+cost-sensitive + all features	SVM4	63.76	65.35	64.55	97.02	96.85	96.93	94.49
+cost-sensitive + all features + polynomial kernel	SVM5	62.25	64.42	63.32	96.90	96.59	96.74	94.30
FM with bag-of-words	FM1	85.74	24.48	38.09	93.96	99.63	96.71	88.40
+cost-sensitive	FM2	60.89	60.79	60.84	96.54	96.57	96.55	93.68
+all features	FM3	85.37	43.00	57.19	95.37	99.34	97.31	91.18
+cost-sensitive + all features	FM4	71.06	67.86	**69.42**	97.21	97.46	97.33	**94.86**

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

Future Work

- Target/Aspect Identification
  - Target: "Microsoft" Aspect: app
    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
  - Target: #microsoft #WindowsPhone7 Aspect: multitasking
    #microsoft #WindowsPhone7 I'd like multitasking please
- Suggestion Summarization
  - Who suggest How to What, When?



THANKS!

Q&A

Any suggestions?