Keyword Extraction for Social Snippets

Zhenhui Jessie Li  Ding Zhou  Yun-Fang Juan  Jiawei Han

University of Illinois at Urbana-Champaign  Facebook Inc.

1. Motivation

In general, two ways to target at users.

- Advertisers
- Facebook users

Structured information:
Location, age, gender, ...

Non-structured information:
keywords

2. Social Snippets

What are social Snippets?
Text generated for social purposes (e.g., Facebook status updates or Tweeter posts):
- updating friends about one’s current status (e.g., “attending WWW conf at Raleigh”)
- initiating or engaging conversations around a topic (e.g., “anyone bought iPad?”)
- expressing the state of the mood (e.g., “is excited for the concert today”)

What are the differences between social snippets and normal documents?

<table>
<thead>
<tr>
<th># Statistics</th>
<th>Facebook</th>
<th>Random web pages</th>
</tr>
</thead>
<tbody>
<tr>
<td># of social snippets</td>
<td>1,830</td>
<td>2,000</td>
</tr>
<tr>
<td># of words</td>
<td>39,249</td>
<td>2,151,500</td>
</tr>
<tr>
<td># of unique words</td>
<td>21,45</td>
<td>1,075.75</td>
</tr>
<tr>
<td># of words in Brown corpus</td>
<td>33,823</td>
<td>1,954,383</td>
</tr>
<tr>
<td># of words in Brown corpus</td>
<td>86.18%</td>
<td>90.81%</td>
</tr>
</tbody>
</table>

Extremely short and considerably noisy

What are the major contributions of this work?
- Define social snippets, a newly emerging type social text data calls for special attention on various applications (keyword extraction, topic modeling, sentiment analysis, …)
- Experimental study of keyword extraction on social snippets (feature engineering and model selection)

3. Keyword Extraction Method

The problem is modeled as a classification problem.

Generate keyword candidates
1. Original Text: I am going to bay area this weekend.
2. Tokenize: I | am | going | to | bay | area | this | weekend
3. Remove stopwords: I | am | going | to | bay | area | this | weekend
4. Generate uni- and bi-grams: {bay, area, weekend, bay area}

Features
- TFIDF
- lin (linguistic feature)
- pos (relative position)
- len (length of keyword)
- DF (document frequency)
- capital (capitalization)

Classification Model
- Gradient Boosting Machine
- Decision Tree
- Support Vector Machine
- Linear Regression
- TFIDF

4. Experiment

Model comparison

Feature importance

Compare with other methods

- Yahoo! api prunes many stopwords (high precision, low recall)
- KEA is based on Naive Bayes model.

5. Future Work

Mining latent interest
- The status or wall posts people "liked".
- People commented are also interested in this topic.
- Extract keywords from the conversation.

Propagate keywords
- Keywords can be propagated to friends.
- How to measure the common interest between two users?
- How to deal with efficiency issue on big social network?