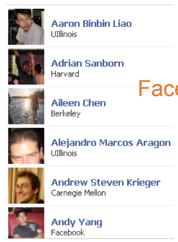
Keyword Extraction for Social Snippets

⁺ Zhenhui Jessie Li ˣ Ding Zhou ˣ Yun-Fang Juan ⁺ Jiawei Han
⁺ University of Illinois at Urbana-Champaign
ˣ Facebook Inc.

1. Motivation

Advertisers

advertisement targeting



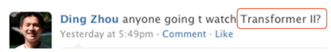
Facebook users

In general, two ways to target at users.

Structured information: *Location, age, gender, …*

Non-structured information: keywords

www.Match.com
Target at 20-30, female, single.



Pros: Well structured
Cons: Limited Info

Pros: Tons of up-to-date dynamic info
Cons: "Hidden"

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?

# Statistics	Facebook	Random web pages
# of social snippets	1,830	2,000
# of words	39,249	2,151,500
# of words / # of social snippets	21.45	1075.75
# of words in Brown corpus	33,823	1,954,383
# of words in Brown corpus / # of words	86.18%	90.84%

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

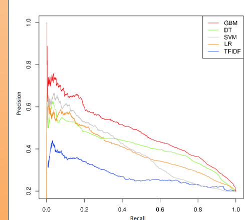
❖TFIDF
❖lin (linguistic feature)
❖pos (relative position)
❖len (length of keyword)
❖DF (document frequency)
❖capital (capitalization)

Classification Model

❖Gradient Boosting Machine
❖Decision Tree
❖Support Vector Machine
❖Linear Regression
❖TFIDF

4. Experiment

Model comparison



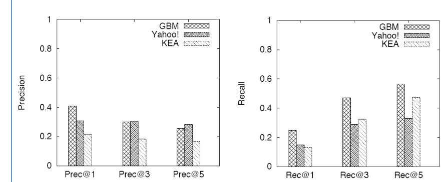
❖GBM performs the best

Feature importance

features	relative influence
TFIDF	31.28%
lin	23.91%
len	17.60%
pos	13.46%
lenText	10.32%
DF	2.61%
capital	0.82%

❖*TFIDF* does not dominate the importance
❖*lin* shows to be important

Compare with other methods



❖Yahoo! api prunes many stopwords (high precision, low recall)
❖KEA is based on Naïve 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?