What makes a tweet relevant for a topic?

#MSM2012 Lyon
April 16th, 2012

Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben
Web Information Systems, TU Delft
Get information from Twitter

How do people use Twitter as a source of information?

- Twitter is more like a news media.

- How do people search for information?
  - Search on Twitter

<table>
<thead>
<tr>
<th>Event</th>
<th>Web Search</th>
<th>Twitter Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese earthquake and tsunami</td>
<td>5,106</td>
<td>12.00</td>
</tr>
<tr>
<td>UK Royal Wedding</td>
<td>3,966</td>
<td>1.64</td>
</tr>
<tr>
<td>Raid on Osama bin laden</td>
<td>5,530</td>
<td>15.22%</td>
</tr>
<tr>
<td>UEFA Champions League</td>
<td>6,903</td>
<td></td>
</tr>
</tbody>
</table>

Query length (chars):
- Web Search: 18.80
- Twitter Search: 12.00

Query length (words):
- Web Search: 3.08
- Twitter Search: 1.64

Is a celebrity name:
- Web Search: 3.11%
- Twitter Search: 15.22%
Research Questions

What are the challenges we are facing?

• 1. Given a topic, can we find the relevant tweets based on the characteristics of the tweets?

• 2. Are semantics meaningful for determining the relevance of the tweets to a topic?
Search on Twitter

Traditional solution

- Twitter search interface
- Ordered by time
- Keyword-based match
Syntactical feature

Is a tweet more relevant if it contains a hashtag?

Hypothesis 1: tweets that contain hashtags are more likely to be relevant than tweets that do not contain hashtags.
Syntactical feature

Is a tweet that contains a URL more relevant?

Hypothesis 2: tweets that contain a URL are more likely to be relevant than tweets that do not contain a URL.

Annette Morris  @LaFranglaise
30 Mar
The 21st International World Wide Web Conference #www2012 will take place in Lyon, France April 16-20 2012 @www2012Lyon www2012.wwwconference.org
Syntactical feature

Is a tweet that mentions @somebody more relevant?

Hypothesis 3: tweets that are formulated as a reply to another tweet are less likely to be relevant than other tweets.
Syntactical feature

Does the length of a tweet influence its relevance for a topic?

Hypothesis 4: the longer a tweet, the more likely it is to be relevant and interesting.
## Overview of features

**Short summary**

<table>
<thead>
<tr>
<th>Topic sensitive</th>
<th>Topic insensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword-based relevance</td>
<td>Syntactical features</td>
</tr>
</tbody>
</table>

**Are there further features that allow for estimating the relevance?**
Semantic features

Find semantics in a tweet to estimate the relevance

Challenge the future
Semantics

Is a tweet with more entities more interesting?

• 5 entities extracted.

Hypothesis 5: the more entities a tweet mentions, the more likely it is to be relevant and interesting.
Semantic features

Do the types of semantics influence the relevance?

- Person : 1
- Artifact : 1
- Event : 1
- Location : 2

Hypothesis 6: different types of entities are of different importance for estimating the relevance of a tweet.
Semantics

How many types are there in the entities?

- 4 types of entities

Hypothesis 7: the greater the diversity of concepts mentioned in a tweet, the more likely it is to be interesting and relevant.
Semantics

Was the author of the tweet happy or not?

- Sentiment: Neutral

Hypothesis 8: the likelihood of a tweet’s relevance is influenced by its sentiment polarity.
### Overview of features

What do we have now?

<table>
<thead>
<tr>
<th>Topic sensitive</th>
<th>Topic insensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword-based</td>
<td>Syntactical</td>
</tr>
<tr>
<td>?</td>
<td>Semantics</td>
</tr>
</tbody>
</table>

Semantics in the given topic?
Semantic-based relevance

Expand the queries to match more tweets.

Tim Berners-Lee, who invented WWW, is going to give a keynote talk at WWW2012, Lyon, France.

Query: “TBL WWW keynote”

New Query: “TBL WWW keynote” or “Tim Berners-Lee WWW keynote” or ...

dbp:International_World_Wide_Web_Conference

dbp:Tim_Berners-Lee
Semantic-based relevance

Find the meanings in the tweets!

Query: “TBL WWW keynote”

dbp:Tim_Berners-Lee

dbp:International_World_Wide_Web_Conference

Tim Berners-Lee, who invented WWW, is going to give a keynote talk at WWW2012, Lyon, France.
# Overview of features

By now, we have 4 types of features.

<table>
<thead>
<tr>
<th>Topic sensitive</th>
<th>Topic insensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword-based</td>
<td>Syntactical</td>
</tr>
<tr>
<td>Semantic-based</td>
<td>Semantics</td>
</tr>
</tbody>
</table>

Can we utilize the contextual information of tweets?
Contextual features

Does the number of posts influence the relatedness?

Hypothesis 9: the higher the number of tweets that have been published by the creator of a tweet, the more likely it is that the tweet is relevant.
Contextual features

Is this tweet too old?

Hypothesis 10: the lower the temporal distance between the query time and the creation time of a tweet, the more likely is the tweet relevant to the topic.

Tim Berners-Lee, who invented WWW, is going to give a keynote talk at WWW2012, Lyon, France.

March 31

April 16
## Summary of Features

The features

<table>
<thead>
<tr>
<th>Topic sensitive</th>
<th>Topic insensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword-based</td>
<td>Syntactical</td>
</tr>
<tr>
<td>Semantic-based</td>
<td>Semantics</td>
</tr>
<tr>
<td></td>
<td>Context</td>
</tr>
</tbody>
</table>
Analysis

• Research Questions:
  1. Which features are more influential on predicting the relatedness of a tweet to a certain topic?
  2. Which types of features are more important? Are semantics meaningful?
  3. What’s the performance that we can achieve by utilizing these features?

• Setup
  • Consider the search problem as a classification task
  • Classification algorithm = Logistic Regression
Dataset
From TREC 2011 Microblog Track

- Twitter corpus
  - 16 million tweets (Jan. 24th, 2011 – Feb. 8th)
  - 4,766,901 tweets classified as English
  - 0.14 million distinct entities

- Relevance judgments
  - 49 topics
  - 40,855 (topic, tweet) pairs
  - 60.31 relevant tweets per topic (on average)
## Results

### Which type of features matters?

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword relevance</td>
<td>0.3040</td>
<td>0.2924</td>
<td>0.2981</td>
</tr>
<tr>
<td>semantic relevance</td>
<td>0.3053</td>
<td>0.2931</td>
<td>0.2991</td>
</tr>
<tr>
<td>topic-sensitive</td>
<td>0.3017</td>
<td>0.3419</td>
<td>0.3206</td>
</tr>
<tr>
<td>topic-insensitive</td>
<td>0.1294</td>
<td>0.0170</td>
<td>0.0300</td>
</tr>
<tr>
<td>without semantics</td>
<td>0.3363</td>
<td>0.4828</td>
<td>0.3965</td>
</tr>
<tr>
<td>all features</td>
<td>0.3674</td>
<td>0.4736</td>
<td>0.4138</td>
</tr>
</tbody>
</table>

Overall, we can achieve the precision and recall of over 35% and 45% respectively by applying all the features.
Results

Which features?

Keyword-based

Syntactical

Semantics

Contextual
Conclusions & Future work

1. We constructed 17 features, including keyword-based relevance, semantics-based relevance, syntactical features, semantic features and contextual features.

2. Semantic features and topic-sensitive features are meaningful.

3. The contextual features have little impact on the prediction.
Future work

• We plan to leverage the implementation of search engine for Twitter based on the work done in this paper.

• The progress on this work can be found at: http://wis.ewi.tudelft.nl/twinder/
QUESTIONS?

April 16th, 2012

http://slideshare.net/to-be-added

k.tao@tudelft.nl

http://ktao.nl/

THANK YOU!