A Visual Framework for Clustering Memes in Social Media

Authors: Anh Dang\textsuperscript{1}, Abidalrahman Moh’d\textsuperscript{1}, Anatoliy Gruzd\textsuperscript{2}, Evangelos Milios\textsuperscript{1}, and Rosane Minghim\textsuperscript{3}

Presenter: Anh Dang

Advanced in Social Networks Analysis and Mining 2015
Date: August 28, 2015

\textsuperscript{1}\{anh,amohd,eem\}@cs.dal.ca, Dalhousie University
\textsuperscript{2}gruzd@ryerson.ca, Ryerson University
\textsuperscript{3}rminghim@icmc.usp.br, University of Sao Paulo-USP
Overview

- Problem
- Motivation
- Related work
- Contributions
- The meme detection framework
  - Internal Centrality-Based Weighting
  - Similarity Score Reweighting with Relevance
  - User Feedback
- Results
- Conclusions
Problem

- A **meme** is a **unit of information** that can spread from person to person through online social networks (OSNs).
- OSNs can be used for spreading spam, slander, and rumors.
  - Swine flu pandemic memes in 2009.
- **How to detect memes or emerging events in OSNs effectively?**
Motivation

- Detecting memes in OSNs is a challenging task due to the limitations of textual content (e.g., Tweets 140 character limit).
- Manually labelling memes is infeasible on a large scale.
  - Clustering is a simple and efficient approach.
- Lexical content similarity does not work well for memes that are related but not using the same words.
  - Polysemy or synonymy.
Leskovec et al. (2009) proposed Meme-tracking system.
- studied the signature path and topic of each meme by grouping similar tweets in Twitter.

Cataldi et al. (2010) monitored the real-time spread of emerging memes in Twitter.
- constructed a navigable topic graph connecting related memes.

JafariAsbagh et al. (2014) introduced the problem of meme clustering.
- grouped structurally related tweets in Twitter.
Our contributions

• Automatically annotate a meme dataset about five popular topics in Reddit.

• Leverage various types of external content for computing similarity between social network text.

• Propose, implement, and evaluate two similarity combination strategies for meme clustering tasks.
  • Internal Centrality-Based Weighting (ICW).
  • Similarity Score Reweighting with Relevance User Feedback (SSR).

• Explore the use of semantic similarity score for meme clustering tasks.

1. GTM - http://ares.research.cs.dal.ca/gtm/
The meme detection framework is to:
• collect and analyze data in Reddit automatically.
• cluster similar Reddit submissions into the same meme using semantic similarity score.
Reddit submission

- A Reddit submission contains the following elements:
  
  - Title
  - Comments
  - URL or Image

![Reddit Submission Example](image-url)
External Content

URL:

IMAGE:
Similarity Measures

Google-trigram method (GTM\(^1\)) is used to:

- compute the semantic similarity between two submissions.
- produce a score from 0 to 1 between two texts.

- **Title similarity**: is the GTM score between two titles.
- **Comment similarity**: is the GTM score between their comments that are concatenated together.
- **URL similarity**: is the GTM score between their URL texts.
- **Image similarity**: is the GTM score between their image contents.

\(^1\) GTM - http://ares.research.cs.dal.ca/gtm/
Explore different similarity score combination strategies:

- **MAX**: selects the highest value among the four pairwise similarity scores.
- **AVG**: calculates the average of the four pairwise similarity scores.
- **LINEAR**: experiments a linear combination among the four pairwise similarity scores.
Combinations (cont.)

• **Internal Centrality-Based Weighting:**
  - calculates the weight factors for each element of a submission by considering its surrounding context between two submissions.

• **The algorithm:**
  - **For each weight factor and each submission:**
    - compute the semantic similarity score between the selected element with the other 3 elements.
    - **Multiple the calculated values between two submissions.**
    - Normalize the scores for all the four weight factors.
Combinations (cont.)

- Similarity Score Reweighting with Relevance User Feedback

Weights

Two selected submissions

Enter value

<table>
<thead>
<tr>
<th>TITLE</th>
<th>COMMENTS</th>
<th>IMAGE</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Big Conservative here. WITHOUT bashing Romney, (which is all I ever see on here), give me your reason to re-elect Pres. Obama.

Reddit URL: [Link]
Evaluation Metrics

- 2439 submissions, 1003 URLs, and 230 images about 5 popular topics in Reddit from Oct to Nov 2014.

- Results returned from each search were labeled with their corresponding keywords.

- Hierarchical clustering is used for the evaluation.

- Purity is used as an evaluation metric.
  - The number of correctly assigned submissions.

<table>
<thead>
<tr>
<th>No.</th>
<th>Topic</th>
<th>Submission Counts</th>
<th>Comments</th>
<th>Submissions with a URL</th>
<th>Submissions with an image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EBOLA</td>
<td>495</td>
<td>89394</td>
<td>218</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>FERGUSON</td>
<td>495</td>
<td>83912</td>
<td>203</td>
<td>87</td>
</tr>
<tr>
<td>3</td>
<td>ISIS</td>
<td>488</td>
<td>76375</td>
<td>190</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>OBAMA</td>
<td>490</td>
<td>139478</td>
<td>142</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Trayvon Martin</td>
<td>471</td>
<td>93848</td>
<td>250</td>
<td>30</td>
</tr>
</tbody>
</table>

The experiment ground-truth dataset
Semantic vs. Syntactical

- TF-IDF and Euclidean distance are used as a baseline similarity distance.

- GTM as a similarity distance outperforms Euclidean-based TF-IDF vectors.

- GTM scales better with an increasing vocabulary size.

GTM vs. Baseline
Experiment – Combination

ICW outperforms MAX and AVG.

SSR improves the clustering results for ICW, MAX and AVG.
Conclusions

• Present a framework to tackle the problem of meme detection in Reddit.

• Define and propose several pairwise similarity scores between elements of two submissions using semantic similarity.

• The proposed Internal Centrality-Based Weighting and Similarity Score Reweighting with Relevance User Feedback strategy achieves the best result.
Future Work

• Extend this proposed framework to other social network websites, such as Twitter, Facebook, and Google Plus.

• Compare the spread of rumour-driven memes between Reddit and other OSNs.
  • provide a more holistic view of rumour spread.

• Visualize how a rumour related meme is discussed and spread in Reddit.
References


• Questions