Graph-based breaking news detection on Wikipedia

Extended Abstract

Ana Freire1,2, Matteo Manca3, Diego Saez-Trumper3, David Laniado3, Ilaria Bordino4, Francesco Gullo3 and Andreas Kaltenbrunner2

1Universitat Pompeu Fabra, Tanger 122-140, 08018 Barcelona, Spain
2Eurecat, Av. Diagonal 177, 8th Floor, 08018 Barcelona, Spain
3UniCredit R&D, Via Molfetta 101, 00171 Rome, Italy

Abstract

Event detection in social media usually exploits information from social-networking platforms, such as Twitter or Facebook. However, previous research has suggested that Wikipedia constitutes a valuable source of information for the task of detecting breaking news. In this work we adapt a graph-based algorithm to the Wikipedia context, and compare it to the state-of-the-art Wikipedia real-time monitoring method. The main idea behind the proposed method is to extract breaking news by looking at unusual activity in the Wikipedia edit stream. We assess the performance of the two competing algorithms by measuring the percentage of true events correctly identified. Results show that the proposed graph-based method achieves better accuracy and coverage.

Introduction

• Wikipedia as valuable source for detecting breaking news [3].

• Existing works are based on spike-detection approaches (number of page views or revisions of an article).

• Contribution: An adaptation to the Wikipedia context of a graph-based approach, traditionally used for detecting events from online user-generated content.

Algorithms

Spike-Detection algorithm VS Graph-based Detection algorithm.

Spike-Detection

• Inpired by the Wikipedia Live Monitor (WLM) [4].

• Monitor Wikipedia articles in real time to discover concurrent edit spikes.

• A Wikipedia article is identified as a potential event if and only if the following constraints are satisfied:

  – Number of concurrent edits \( ed > n_1 \)

  – Number of concurrent editors \( ed > n_2 \)

  – Time between two consecutive edits \( D \leq t \)

  – Revision length (in bytes) \( r > \text{avg}(n) > 140 \)

  – \( \text{minor} \_\text{edit} = \text{false} \): an edit is not considered if it is marked as a minor edit.

Graph-based Detection

Iterative densest-subgraph extraction approach [1]. We build an input graph where:

• Vertices correspond to Wikipedia pages.

• Two pages are connected by an edge if and only if they have been edited consecutively by the same user within a considered time slot.

• For each time slot, every edge is weighted by the number of common editors.

• Extracts the subgraph achieving maximum density and considers it as an event.

• Repeat until the desired number of events has been detected or the input graph has become empty.

Results

Graph-based algorithm achieved both higher precision (0.70 vs. 0.67) and coverage (168 vs. 73 events detected).

Table 1: Comparison between the two competing methods involved in the comparison.

<table>
<thead>
<tr>
<th>Category</th>
<th>Spike-Detection</th>
<th>Graph-based Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sport</td>
<td>49</td>
<td>117</td>
</tr>
<tr>
<td>Entertainment</td>
<td>15</td>
<td>46</td>
</tr>
<tr>
<td>Social&amp;Political</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Biography</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Tech&amp;Science</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Disasters</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Number of breaking news detected per category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Events (Spike)</th>
<th>Events (Graph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sport</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Social&amp;Political</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Biography</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tech&amp;Science</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Disasters</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

• Graph-based algorithm to identify breaking news in Wikipedia.

• This method improves both precision and the absolute number of breaking news detected with respect to the state-of-the-art Wikipedia event-detection algorithm.

• The graphs were built using the co-editions occurred during the whole day, while the Spike-detection method was performed in real-time.

Future work

• Make our method work in real-time (reducing the time granularity and working with incremental updates of the graph).

• Compare our work with other state-of-the-art approaches [2] [3].

• Include a crowdsourcing evaluation process for labeling the events.

References


Acknowledgment

This research has been co-funded by the EC SUPER (FP7-606853) project.