Graph-based breaking news detection on Wikipedia

Extended Abstract

Ana Freire[1,2], Matteo Manca[2], Diego Saez-Trumper[2], David Laniado[2], Ilaria Bordino[3], Francesco Gullo[3] and Andreas Kaltenbrunner[2]

[1]Universitat Pompeu Fabra, Tanger 122-140, 08018 Barcelona, Spain
[2]Eurecat, Av. Diagonal 177, 8th Floor, 08018 Barcelona, Spain
[3]UniCredit 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

- Inspired 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 $edt \geq n_1$
  - Number of concurrent editors $edr \geq n_2$
  - Time between two consecutive edits $D \leq t$
  - Revision length (in bytes) $rev\_len > 140$
  - $minor\_edit =$ 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.

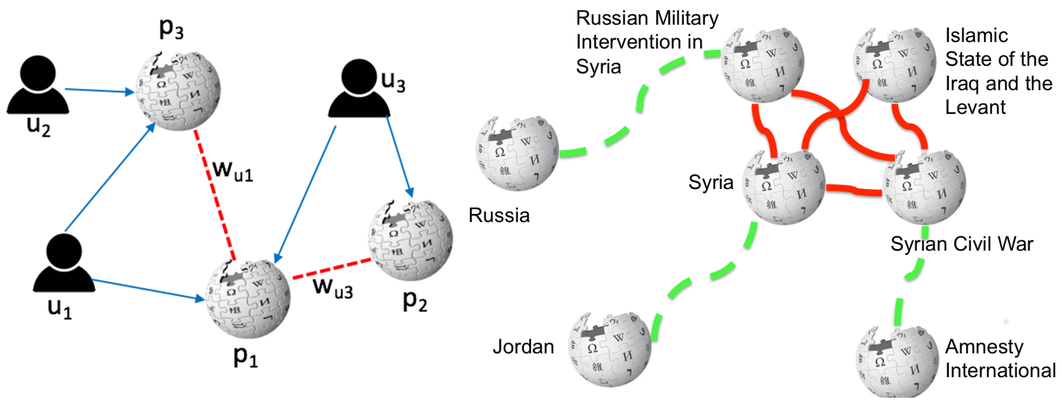

Figure 1: Left: Example of input graph: the blue lines represent the edits made by users on Wikipedia pages, while the red dotted lines represent the weighted edges between pages.
Right: Example of densest subgraph extraction.

Densest Subgraph Detection

Maximizing the average weighted degree in the snapshot under consideration.

- Remove iteratively vertex with the minimum (weighted) degree.
- Process produces a sequence of subgraphs.
- Result: subgraph with maximum (weighted) average degree.

Experiments

Dataset

A 17-days period edit stream from the English Wikipedia (2015-10-13 to 2015-10-29).

Setup

- We extracted a set of candidate events (page titles) using the two competing methods.
- Two domain engineers evaluated them manually.
- Each candidate event was labeled as:
  - True Positive (TP), if the candidate event was recognized as an actual event (i.e., if the Wikipedia pages associated to the event report an event which has happened during the time slot in which the candidate event was detected);
  - False Positive (FP), otherwise.

The Spike-Detection algorithm was tested with different parameters. The more suitable values empirically obtained are the following:

- number of concurrent editors: $edr \geq 5$ (concurrent edits parameter is implicit at this point)
- time slot: $t = 30$ min

Results

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

	Spike-Detection	Graph-based Detection
Detected Events	73	168
True Positives	49	117
False Positives	24	51
Precision	0.67	0.7

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

Sport and entertainment events represent the 74.36% of the actual detected events (true positives).

Category	#Events (Spike)	#Events (Graph)
Sport	14	41
Entertainment	15	46
Social&Political	7	13
Biography	6	1
Tech&Science	1	0
Disasters	5	16
Other	1	0

Table 2: Number of breaking news detected per category.

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

[1] Albert Angel, Nikos Sarkas, Nick Koudas, and Divesh Srivastava. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. *PVLDB*, 5(6):574–585, 2012.

[2] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. Extracting event-related information from article updates in wikipedia. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 254–266, Berlin, Heidelberg, 2013. Springer-Verlag.

[3] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR 2012 Workshop on Time-aware Information Access*, 2012.

[4] Thomas Steiner, Seth van Hooland, and Ed Summers. Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *WWW 2013*, pages 791–794, New York, NY, USA, 2013. ACM.

Acknowledgment

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

upf. Universitat Pompeu Fabra Barcelona

eurecat