Graph-based breaking news detection on Wikipedia
Extended Abstract

Ana Freire[1], **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, Room 55.228, 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 has been largely recognized as a valuable source for detecting breaking news (Osborne et al. 2012). However, to the best of our knowledge, existing works are just based on spike-detection approaches that look at the number of page views or revisions of an article. In this work we propose a novel method consisting of an adaptation to the Wikipedia context of a graph-based approach, which has been traditionally used for detecting events from online user-generated content.

Algorithms

This section describes the two algorithms we consider to detect breaking news by analyzing the Wikipedia stream: the state-of-the-art Spike-Detection algorithm and the proposed Graph-based Detection algorithm.

Spike-Detection

The Spike-Detection algorithm is the method currently used in the literature to detect events by looking at the Wikipedia stream. Inspired by the Wikipedia Live Monitor (WLM) introduced by Steiner, van Hooland, and Summers (2013), such a method consists of a module to monitor Wikipedia articles in real time in order to discover concurrent edit spikes. The algorithm is language independent, being able to perform real-time monitoring of all language versions of the Wikipedia stream.

Specifically, the algorithm analyzes the Wikipedia edit stream and identifies a Wikipedia article 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 of a given page is not considered if it is a minor edit, i.e., if it does not change the main meaning of the article (minor edits are marked as such by Wikipedia editors themselves when saving their contributions on the wiki[1]).

Each event that meet the above constraints is recognized as a candidate event.

Graph-based Detection

The method we propose here is an iterative densest-subgraph extraction approach that has been traditionally employed in the context of event detection from online user-generated content, for example in Twitter (Angel et al. 2012).

In our context we build an input graph whose vertices correspond to Wikipedia pages, and draw an edge between two pages if and only if those two pages have been edited by the same user within a considered time slot (e.g., a day). For each time slot, every edge is weighted by the number of common edits that have originated it.

We slightly adapt the traditional densest-subgraph-extraction method so as to handle graphs of this kind. Specifically, the proposed method extracts the subgraph achieving maximum density and considers it as an event. The process is repeated until the desired number of events has been detected or the input graph has become empty.

Experiments

Setup

We evaluated the aforementioned competing algorithms on a dataset downloaded from Wikimedia Labs (See `https://www.mediawiki.org/wiki/Wikimedia\_Labs`),

[1]`https://en.wikipedia.org/wiki/Help:Minor_edit`

which includes the Wikipedia stream for a 17-day period (2015/10/13 to 2015/10/29). At this preliminary stage, we only focused on the English version of Wikipedia. We plan to extend our analysis to more languages in the future.

We extracted a set of candidate events (pages' titles) from this dataset by using the two competing methods, and got them manually evaluated by two domain engineers. 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 some significant event happened during the time slot in which the candidate event was detected);
- False Positive (FP), otherwise.

Moreover, the Spike-Detection algorithm was tested with different parameter values in order to infer those that led to the best results in terms of precision. The more suitable values empirically obtained are the following:

- number of concurrent edits: $edt \geq 5$
- number of concurrent editors: $edr \geq 5$
- time slot: $t = 30$ min

Larger values of concurrent edits and concurrent editors led to a smaller number of detected events, while assigning smaller values to those parameters increased significantly the number of false positives, with a consequent loss in precision. Note that the constraint on the number of concurrent editors ≥ 5 implies that the number of concurrent edits is larger than 5 too, hence our algorithm can be simplified by discarding such a parameter.

Moreover, we defined a set of categories and manually assigned a category to each TP event in order to see the distribution of the breaking news.

Results

Table 1 summarizes the results obtained. The proposed Graph-based Detection algorithm achieved larger precision (0.70 vs. 0.67). It also evidently outperformed Spike-Detection in terms of coverage: 117 actual events detected vs. 49. On the other hand, an advantage of the Spike-Detection method is that its results have been obtained in real-time, while the Graph-Detection method has been built using the co-editions occurred during the whole day.

	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.

As shown in Table 2, sport and entertainment events represent the 74.36% of the actual detected events (true positives).

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

Table 2: Number of breaking news detected per category.

Conclusions

We proposed a graph-based algorithm to identify breaking news in Wikipedia. We showed that 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. In the future we plan to make our method work in real-time, by reducing the time granularity and devising a version of the algorithm that may exploit incremental updates of the graph, instead of periodically rebuilding it from scratch. We also envisage to compare our work with other state-of-the-art approaches (Osborne et al. 2012) (Georgescu et al. 2013), and to include a crowdsourcing evaluation process for labeling the events.

Acknowledgment

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

References

Angel, A.; Sarkas, N.; Koudas, N.; and Srivastava, D. 2012. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. *PVLDB* 5(6):574–585.

Georgescu, M.; Kanhabua, N.; Krause, D.; Nejdl, W.; and Siersdorfer, S. 2013. Extracting event-related information from article updates in wikipedia. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, 254–266. Berlin, Heidelberg: Springer-Verlag.

Osborne, M.; Petrovic, S.; McCreadie, R.; Macdonald, C.; and Ounis, I. 2012. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR 2012 Workshop on Time-aware Information Access*.

Steiner, T.; van Hooland, S.; and Summers, E. 2013. Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *WWW 2013*, 791–794. New York, NY, USA: ACM.