



Graph-Based Clustering Approach for Economic and Financial Event Detection Using News Analytics Data

Sergei P. Sidorov^(✉), Alexey R. Faizliev, Michael Levshunov,
Alfia Chekmareva, Alexander Gudkov, and Eugene Korobov

Saratov State University, Saratov, Russian Federation
sidorovsp@info.sgu.ru

Abstract. In recent years, one of the most extensive research topics in social media analysis has been event detection. Most of the approaches use fixed temporal and spatial resolutions to detect events. In this paper, we employ a procedure for the detection of economic and financial events using news analytics data. We use an algorithm to compute a data similarity graph at chosen scales and detect economic and financial events simultaneously by a single graph-based clustering process. Experimental results on real world data collected from news analytics providers demonstrate the effectiveness of the event detection procedure based on real-time news analytics data.

Keywords: Event detection · Spatiotemporal analysis
News analytics · Clustering

1 Introduction

News agencies, organizations, social networks, enterprises, etc. generate a huge amount of information and news in real time. The study of the characteristics and features of the news flow has become one of the recent topics in social informatics. The news stream is publicly available and reflects processes which are going on in the world. Therefore, the examination of structural properties of the news flow and the development of the methods and algorithms for processing such data are of great interest.

In this paper we restrict our analysis to the study of economic and financial news. News analytics providers (such as Media Sentiment, Thompson Reuters and Raven Pack) collect, aggregate and pre-process news from different sources including news agencies and social media (blogs, social networks, etc.). They also use so-called pre-news, i.e. SEC reports, court documents, reports of various government agencies, business resources, company reports, announcements,

The work was supported by RFBR (grant 18-37-00060).

industrial and macroeconomic statistics. The survey of applications for news analytics tools can be found in books [14, 15]. In recent years news analytics tools have been developed and used in social network analysis [4, 9, 12, 18].

Thus, the huge amount of news analytics data enables the study of some research problems. One of such problem is the analysis of company co-mention network which has been addressed in [20, 23]. This paper concerns with another such problem which is event detection. Note that news analytic data raise advantages for event detection. Using the real-time nature of news analytic services, both traders and governments may boost their awareness of real world events much quicker than with offline analysis based on traditional media sources.

Our research interest is close to the problem of event detection in Twitter. It should be noted that event detection in Twitter has long been a research topic [24] and the main problem has been detecting those events that Twitter users are discussing most. Most of the researches supposed that an increase in the usage of some related words would show when an event is happening. An event is therefore commonly characterized by a number of keywords displaying sharp burst in appearance count [10, 24]. Since then the problem of event detection in Twitter has drawn a serious amount of enthusiasm from the data mining researchers [1, 2, 5, 6, 17]. The common understanding is that events in social media platforms can be loosely defined as real world happenings that occur within similar time periods and geographical locations, and that have been mentioned by the online users. Recent works include [3, 7, 8, 11, 16], among many others.

In this paper, we use an approach based on the ideas of the paper [24]. We use news analytics data for detecting events that are of similar scales and localized in time and referred to a particular company. Thus, we employ the characteristic of news items which are reflect localization in time and the similarity of their attributes.

2 Events

In this paper we define events as real world happenings that are reflected by news items that are concentrated in time and connected with a particular company or a group of companies, or an economic sector. We suppose that an important offline event attracts interest of news and information agencies as well as other news providers and news exchange participants. The more important the event, the more intense will be the corresponding news flow. We assume that an important event should cause a cascade of news. We define a cascade of news as the sequence of news items published by various news providers, which are published in the same time frame and related to one company and have one topic. As a rule, such a cascade has an initial news message (trigger), which triggers this cascade. If some news turned out to be important, other news providers and news agencies reprint this news close to the original form. In addition, they can publish news items related to the development of events caused by this news. If a published news item seems to be not important, it may be reprinted by a very few news providers, or it may not be reprinted at all. The news that did not provoke interest and

did not lead to a cascade of news, in our opinion, is hardly connected with an important event. Therefore, the detection of such cascades in the flow of news items allows us to detect events in one way or another. Each such cascade is a component of connectivity in the graph of news reports. Thus, the identification of cascades and their corresponding events is reduced to the task of finding the related subgraphs (clusters) of the news graph. It can be assumed that the more news are in the cascade, the more significant the corresponding event is.

Of course, different events can exhibit different temporal scales, i.e. they can span different intervals in time.

In this paper we consider a news analytics data stream with temporal and text information. The goal of this paper is to employ an approach to identify events using the stream data of news analytics providers. We treat event detection as a graph-based clustering problem, where the vertices of the graph represent the news, and the edges reflect their similarities.

3 Event Detection via Temporal and Attributive Constraints

Economic and financial events presented in news flow may have different localization behavior in time and can have different attribution to a particular company (group of companies, sector, country). When the events are localized in the two dimensions (temporal and attributive), event detection can be adequately implemented by imposing temporal and attributive constraints on the data. In this section, we detail an approach for detecting economic and financial events that are localized both in time and attribution. It can be reduced to a clustering problem, where we assemble together the news items that describe the same real world event. We define the similarity measure between different news items n_i and n_j as follows:

$$S(n_i, n_j) = \begin{cases} 1, & \text{if } t(n_i, n_j) \leq T_t \text{ and } A(n_i) = A(n_j), \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where $t(n_i, n_j)$ is the temporal difference in minutes between n_i and n_j and $A(n)$ is the value of attribute A (company name and news topic) for news item n .

The threshold T_t enforces strict temporal constraints of the events and A implements the shared attributes. Under such constraints, two news items n_i and n_j seem to refer to the same economical and financial event.

The graph connectivity components are found by means of the depth search approach. The main drawback of this clustering algorithm on big graphs is the high demand on the amount of RAM. However, the problem is hugely reduced due to an eminent sparsity of our graph.

Different approaches to find similarity measure for twits have been used in papers [6, 13].

Let $S(n_i, n_j)$ be the pairwise similarity between news items defined in (1). Let us define matrix W as follows:

$$W(n_i, n_j) = \begin{cases} S(n_i, n_j), & \text{if } i \neq j, \\ 0, & \text{if } i = j. \end{cases} \quad (2)$$

Matrix W is a symmetric sparse matrix. We can construct an sparse undirected $G = (V, E, W)$ for which W is the adjacency matrix. The vertices V of the graph will represent news items, $E = V \times V$ is the edge set. An edge connects two news if they are about the same company and have the same topic. If $W(n_i, n_j) \neq 0$ then there is an edge between two vertices v_i and v_j and the weight $w_{ij} = 1$.

With such a graph representation of W , event detection can then be reduced to a graph partitioning task, i.e. to divide the graph G into subgraphs. Our goal is to subdivide the vertices of the graph into disjoint clusters in such a way that each cluster would contain news items that probably correspond to the same economical or finance event. The constraints in (1) imply that the events should be localized in time and be attributed to the same company (or group of companies, or to the same economical sector, country).

Algorithm 1. Event Detection via Local and Attributive Constraints

Input: N : a set of news item with temporal, attributive, and text information;
 T : temporal threshold; A : attributive information

begin

- Compute the pairwise similarities $S(n_i, n_j)$ between all news items in N using (1);
- Compute the adjacency matrix W using (2);
- Apply the graph partitioning algorithm to W , and retain the meaningful clusters $\{c_i\}_{i=1}^m$ after postprocessing steps;

end

Output : $\{c_i\}_{i=1}^m$: clusters that correspond to events that are localized in time and share the same attribute

4 Empirical Result

We now test the performance of Algorithm 1 for real world event detection problems. We first describe the data and some implementation details, and then present the event detection results.

4.1 Data Description

Providers of news analytics obtain and aggregate data from different sources (including news agencies and business reports) and social media. Our data cover the period from January 1, 2015 to January 31, 2015 (i.e. 22 trading days). We consider all the news released during this period. Initially we performed data selection and cleaning process.

Descriptive statistics of news flow can be found in Table 1. The total amount of news items issued in January was 134199. Table 1 shows that the average amount of news published per day was 4329. There were days with the very low intensity of news flow (e.g. only 30 economic and financial news items were published on January, 1 of 2015), while some days exhibit a much higher amount of published news items (e.g. 10238 news items were issued on January, 29 of 2015).

Table 1. Descriptive statistics of news flow in January, 2015

Days	31
Total amount of news items	134199
Mean per day	4329
Minimum	30
Maximum	10238
St. deviation	3122.8
Median	5360

Figure 1 shows that the average amount of holiday news items is much less than that of holidays. Therefore, we eliminate all holiday news from the study. The news flow is highly noisy and time series of news flow data exhibit self-similarity and fractality [19, 21, 22]. Examining the problem of handling the noisy information in the Twitter data, the paper [6] uses a homogeneous Poisson process as a statistical model.

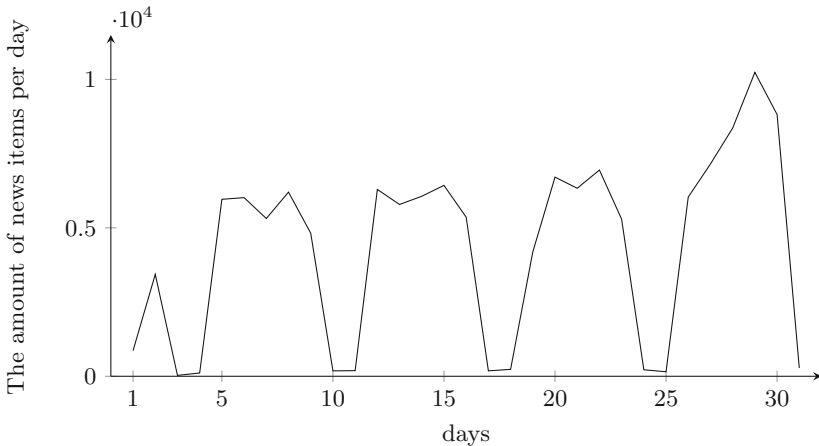


Fig. 1. Daily dynamics of the news flow in January, 2015

4.2 Event Detection Results

We would like to group similar news items into the same cluster in such a way that they could correspond to a real world event. We implement the event detection algorithm and obtain empirical results for different values of temporal threshold $\theta = 0.5, 1, 3, 6, 12$ h for the detection of event clusters. The results are presented in Tables 2 and 3. Table 2 exhibits the distribution of amount of cascades with dependence to the size of the cascade for different values of θ . Table 3 presents the sizes of connectivity components and temporal durations of corresponding news cascades for different values of θ .

As can be seen from Table 2, the distribution of the sizes of connectivity components is fairly stable regardless of the parameter θ . At the same time, Table 3 shows that most of the connectivity components correspond to small-sized cascades (ranging from 2 to 10 news). This small-sized cascades can not reflect an event. It is obvious that an increase of the parameter θ must lead to the appearance of larger-sized connectivity components, and corresponding cascades may last the entire period under consideration. In our opinion, taking the value of θ more than 12 h could lead to the unification of the whole news flow for a quite large company in one cascade. In our opinion, if an event occurred, then the discussion of it in the news flow will occur quite intensively. At the same time, the continuity of the working day of the world financial market will not allow to forget the event, and news will be published in Asian, European and American financial markets, and so on. Therefore, in our opinion, in order to detect events it is quite enough to take θ in the interval from 3 to 6 h.

Table 2. The amount of cascades N and the average duration of cascades T with a given size (the amount of news items in cascade), for different values of θ

Cascade sizes	$\theta = 0.5$		$\theta = 1$		$\theta = 3$		$\theta = 6$		$\theta = 12$	
	N	T	N	T	N	T	N	T	N	T
11–20	770	26 m	847	49 m	985	2 h 13 m	1071	3 h 59 m	1132	7 h 50 m
21–30	84	36 m	107	1 h 24 m	151	3 h 53 m	159	7 h 6 m	209	17 h 1 m
31–40	29	1 h 0 m	33	1 h 28 m	41	4 h 12 m	49	9 h 13 m	62	17 h 53 m
41–50	9	1 h 2 m	14	2 h 15 m	18	6 h 25 m	23	12 h 1 m	24	1d 3 h 23 m
51–60	8	1 h 1 m	10	1 h 56 m	13	5 h 15 m	16	11 h 31 m	24	21 h 6 m
61–70	1	2 h 43 m	1	2 h 43 m	6	11 h 40 m	7	22 h 28 m	7	1d 16 h 59 m
71–80	2	58 m	2	58 m	2	58 m	9	15 h 1 m	10	1d 6 h 56 m
81–90	0	-	0	-	0	-	0	-	1	2d 20 h 45 m
91–100	0	-	0	-	0	-	0	-	1	4d 6 h 51 m
101–110	0	-	0	-	0	-	0	-	1	1d 8 h 32 m

The analysis of the clustering results for Algorithm 1 show that the clusters detected by the algorithm correspond to meaningful economic and financial real world events of interest.

Below we describe a few cases of news cascades found by Algorithm 1.

Table 3. Characteristics of connectivity components' sizes and temporal durations of corresponding news cascades for different values of θ

	Temporal threshold θ				
	0.5 h	1 h	3 h	6 h	12 h
Total amount of connectivity components (with 2 and more items)	21038	21276	21382	21610	21657
Mean duration of connectivity components (with 2 and more items)	8 m	15 m	39 m	1 h 19 m	2 h 36 m
Mean size of connectivity components (with 2 and more items)	3.98	4.10	4.31	4.43	4.57
Median of connectivity components (with 2 and more items)	3	3	3	3	3
Total amount of cascades (with 11 and more items)	903	1014	1216	1334	1471
Mean duration of cascades (with 11 and more items)	29 m	56 m	2 h 37 m	4 h 57 m	10 h 32 m
Mean size of cascades (with 11 and more items)	16.16	16.49	17.05	17.54	18.29
Median of cascades (with 11 and more items)	13	14	14	14	15
Max cascade	74	74	74	76	101
Max duration	3 h 18 m	7 h 52 m	21 h 31 m	1d 16 h 59 m	4d 10 h 40 m

- (1) Yahoo Inc. decided to transfer its 15% of the Chinese Alibaba Group shares to a separate company. At the closure of trading on January 26th the shares' value was estimated at \$40 billion, as follows from a message on the company's website. The cascade began at 28.01.2015 02:36:07 and lasted till 28.01.2015 18:04:04. Duration 15 h 27 min 57 s. The number of news in the cascade is 22. The first news was printed by Forbes.
- (2) Facebook reports fourth quarter and full year 2014 results. Its users' community was continuously growing. Revenue for the fourth quarter of 2014 totaled \$3.85 billion, an increase of 49%, compared with \$2.59 billion in the fourth quarter of 2013. The cascade began at 21.01.2015 21:01:12 and lasted till 22.01.2015 15:17:24. Duration 18 h 16 min 12 s. The number of news in the cascade is 48. The news was printed by Facebook.
- (3) Netflix ended 2014 with 57.4 million subscribers worldwide topping expectations for overseas growth. The No. 1 subscription video-on-demand provider reported quarterly revenue of \$1.48 billion, up 26% year over year. Netflix shares were up more than 13% in after-hours trading, to \$395.15 per share, on the results. The cascade began at 20.01.2015 21:01:11 and lasted till 21.01.2015 15:21:30. Duration 18 h 20 min 29 s. The number of news in the cascade is 52. The first news was printed by Variety.

- (4) On the morning of January 28, 2015 at 05:19:31 the BBC news service informed about record Apple’s income. Apple reported about its quarterly profit of 18 billion dollars. It is a historical record for a public company. After that, the price of Apple shares increased by 5%. This statement aroused a news cascade with duration of 12 h 14 min 38 s. The cascade contained 25 news items.
- (5) Tesco PLC is a British company and one of the oldest transnational trade networks. Tesco PLC owns a large trade network in British. This company controls about 30% retail of foodstuffs. On the morning of January 8, 2015 at 7:00:28 one of the largest American publications The Wall Street Journal the news was published about lower prices and closing unprofitable stores. Tesco PLC planned to close about 43 unprofitable stores and to sell a range of assets. After that, Tesco’s shares increased at once by 14,97% and have reached 2.09 pounds. This information aroused a news cascade with duration of 21 h 31 min. The cascade contained 66 news items (Fig. 2).

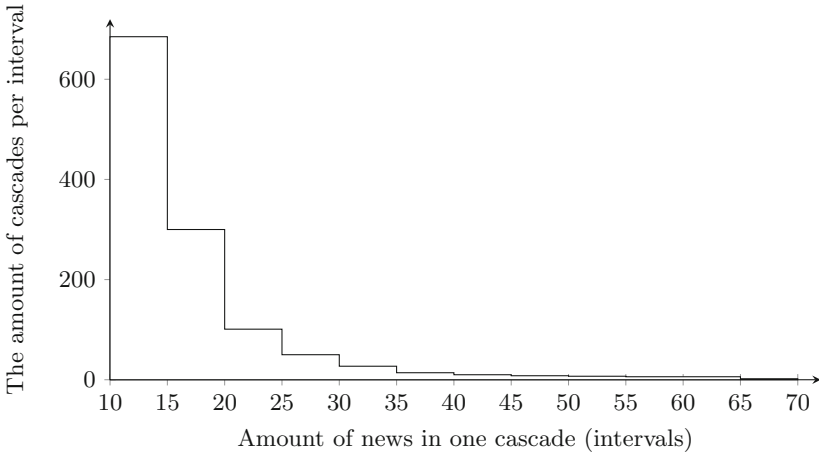


Fig. 2. The histogram for the amount of connectivity components (with 11 and more items), threshold $\theta = 3$ h

5 Conclusion

In this paper we suppose that an important economic and financial event should lead to a cascade of news. By this reason, the mining of such cascades in the news flow may be exploited to detect events. Each such cascade forms a cluster in the graph of news reports. Therefore, the mining of cascades (and associated events) can be reduced to the problem of finding the corresponding clusters in the graph. The clustering algorithm employs a similarity measure between news items that uses both the temporal and textual features of a news item. It is supposed that two news items that are generated by news agencies that describe the same event

should be closely connected with a particular company, or a group of companies, or economic sector, or country. In this paper, to measure the similarity between news we use approach based on temporal and affiliation constraints. Empirical results on real world data collected from news analytics providers exhibit the capability and the efficiency of the event detection approach.

References

1. Aggarwal, C.C.: Mining text and social streams: a review. *SIGKDD Explor. Newsl.* **15**(2), 9–19 (2014)
2. Aggarwal, C.C., Subbian, K.: Event detection in social streams. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 624–635 (2012). <https://epubs.siam.org/doi/abs/10.1137/1.9781611972825.54>
3. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. *Comput. Intell.* **31**(1), 132–164 (2015)
4. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. *AI Soc.* **30**(1), 89–116 (2015)
5. Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010*, pp. 291–300. ACM, New York (2010)
6. Dong, X., Mavroeidis, D., Calabrese, F., Frossard, P.: Multiscale event detection in social media. *Data Min. Knowl. Discov.* **29**(5), 1374–1405 (2015)
7. Hasan, M., Orgun, M.A., Schwitter, R.: Real-time event detection from the twitter data stream using the TwitterNews+ framework. *Information Processing & Management* (2018, in press)
8. Huang, Y., Li, Y., Shan, J.: Spatial-temporal event detection from geo-tagged tweets. *ISPRS Int. J. Geo-Inf.* **7**(4), 150 (2018)
9. Khan, W., Daud, A., Nasir, J.A., Amjad, T.: A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait J. Sci.* **43**(4), 95–113 (2016)
10. Kleinberg, J.: Bursty and hierarchical structure in streams. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002*, pp. 91–101. ACM, New York (2002)
11. Kolchyna, O., Souza, T.T.P., Treleaven, P.C., Aste, T.: A framework for twitter events detection, differentiation and its application for retail brands. In: *2016 Future Technologies Conference (FTC)*, pp. 323–331, December 2016
12. Manaman, H.S., Jamali, S., AleAhmad, A.: Online reputation measurement of companies based on user-generated content in online social networks. *Comput. Hum. Behav.* **54**(Suppl C), 94–100 (2016)
13. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, New York (2008)
14. Mitra, G., Mitra, L. (eds.): *The Handbook of News Analytics in Finance*. Wiley, Chichester (2011)
15. Mitra, G., Xiang, Y. (eds.): *The Handbook of Sentiment Analysis in Finance*. Albury Books (2016)
16. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.* **25**(4), 919–931 (2013)

17. Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009). AAAI (2009)
18. Schuller, B., Mousa, A.E., Vryniotis, V.: Sentiment analysis and opinion mining: on optimal parameters and performances. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **5**(5), 255–263 (2015)
19. Sidorov, S.P., Faizliev, A.R., Balash, V.A.: Fractality and multifractality analysis of news sentiments time series. *IAENG Int. J. Appl. Math.* **48**(1), 90–97 (2018)
20. Sidorov, S.P.: QAP analysis of company co-mention network. In: Bonato, A., Pralat, P., Raigorodskii, A. (eds.) WAW 2018. LNCS, vol. 10836, pp. 83–98. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92871-5_7
21. Sidorov, S.P., Faizliev, A.R., Balash, V.A., Korobov, E.A.: Long-range correlation analysis of economic news flow intensity. *Phys. A: Stat. Mech. Appl.* **444**, 205–212 (2016)
22. Sidorov, S., Faizliev, A., Balash, V.: Scale invariance of news flow intensity time series. *Nonlinear Phenom. Complex Syst.* **19**(4), 368–377 (2016)
23. Sidorov, S., Faizliev, A., Balash, V., Gudkov, A.A., Chekmareva, A.Z., Anikin, P.K.: Company co-mention network analysis. In: Kalyagin, V., Pardalos, P., Prokopyev, O., Utkina, I. (eds.) *Computational Aspects and Applications in Large Scale Networks*, vol. 247, pp. 83–98. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96247-4_26
24. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998, pp. 28–36. ACM, New York (1998)