Burst Events Detection in Text Streams by Using Keyphrases

Sheng-Hsiang Chen[1] Roung-Shiunn Wu[2]
Department of Information Management
National chung cheng University
168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan
[1]welkin2019@gmail.com
[2]roungwu@ccu.edu.tw

Abstract

Mining text streams for hot topics and events has attracted extensive attention in the world because of its broad applications. Since keyphrases have more expressive power than single term and keyphrases can be utilized to represent documents more semantically. In this research, we try to detecting burst events by using keyphrases.

We give a formal definition to the above problem and present the frameworks with five steps to solve the problem: (1) use KP-Miner to extract keyphrases from text streams as features set; (2) cluster keyphrases with synonymy or hypernymy into groups; (3) calculate occurrence frequencies of the groups in sliding windows; (4) evaluate burst groups; (5) burst event as burst groups. We also find the problem about loosing potential burst groups in fixed time window. In order to alleviate this problem, the original time window and the shift time window are good ways to settle the problem. We evaluate the proposed framework on real Google news stream which is suitable for our research. Experimental results show that our framework can detect more descriptive burst events than external events.

Keywords: KP-Miner, keyphrases extraction, Stream mining, Text mining, WordNet

Chapter 1 Introduction

Text streams are everywhere and often naturally formed as new information is incrementally created and accumulated. It can be news, literature, blog, query, even patient records stream. Text streams have one interesting characteristic that there is often an intensive coverage of some topic with certain period, which we refer to as a burst topic pattern or burst event. For example, when a big event happens in the world, all news articles try to have the intensive coverage of the event; as a result, there would be a coverage burst of the topic lasting for a certain period [1].

There are two common methods to find burst events in text streams. One is using clustering approach to group similar texts together, like K-Means, and identifies each cluster having burst or not. Another Method is considering burst event as burst features. For example, the event “SARS” consists of the features “Outbreak”, “Atypic”, “Respire”.

1.1 Research Motivation

There are still many interesting aspects that we can plunge into. For example, [2] tried to build an online application that presents daily bursts. The other extended research [3] to micro blogging services or tried experimenting on the bodies of the blog post. Text mining method should be able to apply to real-time text stream processing [4]. More complex association semantic between the streams need to investigated further [5]. The future research can establish semantic space model [6]. These suggestions give us a direction to study semantic relationship in text streams.

Since keyphrases have more expressive power than single term, leyphrases can be utilized to represent documents more semantically. Some research tried to combine keyphrases, which normally contains a noun as its head and which can be modified in many ways, with the use of WordNet to explore better ways of representing documents semantically. However, they did not take text streams into account when they used keyphrases to analysis.

1.2 Objectives

Even thought there are some weaknesses of using keyphrases, we employ the keyphrases to text streams for burst events detection. We use keyphrases to complement the weakness of using single term for improving the quality of detecting burst events.

Two keyphrases are synonymous if they can be used to express the same meaning. Synonymous can be identified with the help of thesaurus, such as WordNet. In the keyphrase-based method, a document is represented as a set of keyphrases. Each phrase represents a concept and consists of several word stems. We measure the similarity between two keyphrases and the similarity between concepts they represented.

Our objective is to use keyphrases for detecting burst events from text streams.

1.3 Contribution

First, we extract the keyphrases to represent documents more semantically. Second, we employ the WordNet combined with hypernymy to deal with the synonymous problem of the keyphrases semantic. Finally, we overcome the gap between semantic keyphrases on text streams and improve the description of detecting burst events.

In the following chapters, we provided related works and introduce our proposed burst event detection framework. In chapter 2, we describe some concepts of text streams and related works. In chapter 3, our proposed framework is described. In chapter 4, more detail description of the framework and implementations are discussed. Finally, we evaluate experimental results in chapter 5 and discuss the conclusions in chapter 6.

Chapter 2 Related Works

There are broad range technologies used in burst detection, including Infinite-state automaton [7] [2] [3], keyword search on databases and subscription/alert service [8] [5], general probabilistic algorithm [9] [4], LDA-Based technology [10] [11] [12], feature-pivot clustering [13], compound technologies [14] [15] [6] and go on.

2.1 Compound Technologies

Yuan et al. [6] introduced an integrated approach to solve burst events detection problem over high speed short text streams. (1) Simplify the requirement by considering burst event as a set of burst features. (2) By using the ratio of the number of documents with specific feature and total number of documents during a period of time as the measurement, their solution can be applied to any kind of data distribution. (3) The proposed burst detection algorithm used Ratio Aggregation Pyramid and Slope Pyramid data structure. They used one-year short documents forum and BBS from a web site. Experimental results showed that their approach is effective. [15] [14] [12] considered burst events as burst features and used time information for a specific goal.

2.2 Keyphrase Extraction Algorithms or Systems

After introducing concepts of phrase extraction in text mining, we review some literatures related to keyphrase extraction algorithms or systems. Keyphrases are known as a list of terms. Each term is made up of one or more words and associated with the documents. Several algorithms and systems were developed to extract keyphrases, such as Extractor [16], KEA [17], and KP-Miner {18], in which different methods were used.

KP-Miner system does not have to be trained on a particular document set for achieving the task. [18] argued that gaining an understanding of the keyphrase extraction process. A system, which do not have training samples at an accuracy comparable to that of supervised machine learning extraction, can be built. KP-Miner also has advantage of being configured

as rules, in which heuristics are associated to the general nature of documents and keyphrases. This advantage implies that the users can input documents into the system and fine-tune to their particular needs. We thought the KP-Miner is a suitable choice in developing our research work [18].Three major steps are identified when KP-Miner extracts keyphrases.

Step1: candidate keyphrase selection

Step 2: candidate keyphrase weight calculation

Step 3: final keyphrase refinement

The web-based version of KP-Miner system is presented in Figure 4. This version has Get Phrases, ClearText Area and Reset All functions and can select language and enter desired number of keyphrases to be retrieved. It is easy to copy and paste document into input text box box for extracting keyphrases [19].

Chapter 3 Methodology

In this Chapter, we describe our proposed burst events detection framework. As mentioned previously, WordNet is applied in our framework for generating synonymy or hypernymy phrases.

3.1 Burst Events Detection Framework

The proposed burst events detection framework is shown in Figure 1. Keyphrases instead of single words are used to represent text streams semantically. The functions of the framework can be divided into five major components.

Component 1: News streams are collected and the KP-Miner is used to extract keyphrases from the news streams as feature set..

Component 2: WordNet tool generates synonymy or hypernymy for the extracted keyphrases which are then clustered into groups.

Component 3: The occurrence frequencies of the groups are calculated based on sliding windows.

Component 4: Burst groups evaluation

Component 5: Burst events detection

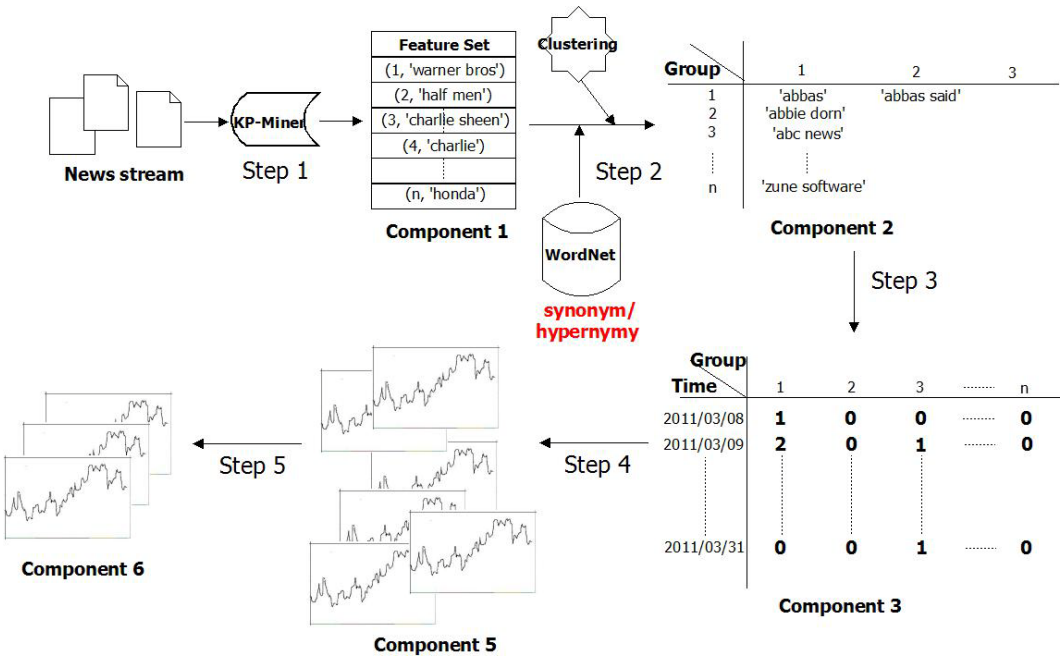

Figure 1. The proposed burst events detection framework

In step1, keyphrases are extracted from news streams. The extracted keyphrases satisfy single term or certain syntactic relations, such as verb-objective, noun-verb, or adjective-noun, from the documents. In order to extract syntactic phrases effectively, we employ the KP-Miner to extract keyphrases. These keyphrases are organized into a feature set.

In step2, we apply the WordNet tool to identify synonymous and hypernymous phrases for the extracted keyphrases.

In step 3, we calculate group occurrence counts by time windows. Each group is represented as keyphrases with similar semantic relationship. In order to use time as a dimension, the time stamps of news streams are indexed for associated keyphrases. For each group, the keyphrases occurrence frequencies from news stream are summed up as total group counts in the sliding window. Note that a sliding window may contain multiple documents. However, the number of keyphrases is always high, ranging from several hundreds to thousands. It makes sense to group keyphrases into a low keyphrase space.

In step 4, burst groups detection is based on sliding windows. The basic idea is to run computations on all of the news streams seen in a time window. Within a time window, the incoming news streams which represented by keyphrases are used for detecting burst groups. Line charts are used to identify burst group patterns inside time windows. According to literatures, the methods for burst detection are different, such as DFIDF [14], strength [9], popularity [3], probability [13] [4], and query frequency [2]. However, these methods are not suitable for our research work because of using keyphrases. Unlike keyphrase features used in previous studies, keyphrases are hot single term extracted from news streams in our study. Hence, keyphrases do not need to be categorized whether they are important.

In step 5, burst events are detected by considering these burst group patterns which have similar line charts. For example, the event “fukushima nuclear power plant” consists of the burst groups “fukushima plant”, “nuclear reactor”, “plant”, “radioactive iodine”, in which have similar burst patterns of line charts in news streams.

Chapter 4 Experimental Design and Results

In this chapter, we describe dataset which were collected for our experiments, evaluate our proposed framework to detect the burst events, and present experimental results.

4.1 Data Preparation

The collected documents must be in English for the KP-Miner to extract keyphrases. Google English news is used to evaluate our proposed framework. Google news is a popular text streams which collecting from different source of news stories such as CNN, Reuters, and other countries' newspaper [20]. We have archived three-week, from 2011/03/08 to 2011/03/31, news stories from Google English news on Internet. To collect news stories, <graph>, <related news>, <light statement>, and <contact information> fields are deleted and <title>, <content>, <source> and <publication> fields are retained. We do not conduct document pre-processing to remove punctuation, digits, and stop words. Totally, there are 287 news stories are collected for experiments. The experiments were implemented in Java and performed on Intel Pentium M notebook running Window XP with 768 MB of memory.

4.2 Experiment on the Framework Step 1

We employ the KP-Miner to extract keyphrases from each of news streams in dataset as feature set. We extract top 5, top 10, and top 15 keyphrases from each news story separately. The results show that top 10 and top 15 keyphrases are mostly single terms which are less important to represent the news stories semantically. Therefor, Top 5 keyphrases, which is the default number in the KP-Miner system, is extracted to evaluate the framework. Top 5 extracted keyphrases of 287 news stories are listed in Excel. The length of extracted keyphrases has no limit, but extracted keyphrases are rarely exceeding three terms.

4.3 Experiment on the Framework Step 2

After extracting keyphrases, we employ WordNet tool to generate synonymy or

hypernymy words for each keyphrase. We cluster keyphrases with synonymy and hypernymy semantic relationship into groups. Hypernymy is defined as the semantic relation of being superordinate or belongs to a higher rank or class [21]. Clustering keyphrases into groups can reduce keyphrases space that results in increasing the possibility to detect burst events. First, all keyphrases are collected to a feature set replace with duplicates eliminated. Second, we cluster keyphrases with synonymy and hypernymy semantic relationships into groups.

4.4 Experiment on the Framework step 3

In step 3, we calculate occurrence frequencies of the groups in different time windows. Top 5 keyphrases of each news story instead of entire content of news story are used to identify it belongs to which groups, since keyphrases are semantically used to represent the entire news story. By using keyphrases, there is no need to find whether the keyphrases are important. To achieve this task, we apply Java to implement and achieve this task.

It is vital to decide what size of time window is more suitable for detecting burst events. In the beginning, we calculate occurrence frequencies of the groups with one hour (i.e., TW=1), and enlarge the size of time window gradually. From the results, 32 hours window size is a better choice for burst period. It is clear to see the burst pattern of groups with 32 hours window size.

4.5 Experiment on the Framework step 4

The statistical methods are applied to detect burst groups. To measure the grouped data's central tendency, mean, median, and mode are usually used. Mean has arithmetic mean, weighted arithmetic mean, and geometric mean. Here, arithmetic mean is more suitable in our experiments to detect burst groups. Arithmetic mean of group's occurrence frequencies is 1.76. There are 180 groups exceeding 1.76. The highest occurrence frequency of these groups is 15.

4.6 Experiment on Framework Step 5

In this subsection, the burst events are detected with similar burst pattern of line charts as burst groups. [15] analyzed feature trajectories for event detection and applied spectral analysis to categorize features, which equal groups in this research, for different event characteristics: important and less-important, periodic and aperiodic. Although different method used to detect burst event from news stream, feature trajectories can be used to identify our groups. Figure 2 shows the four group sets for events. This four group sets include HH (aperiodic groups for important aperiodic events), HL (periodic groups for important aperiodic events), LH (aperiodic groups for less- important aperiodic events) and LL (noisy features). The burst pattern of each group set is presented aside.

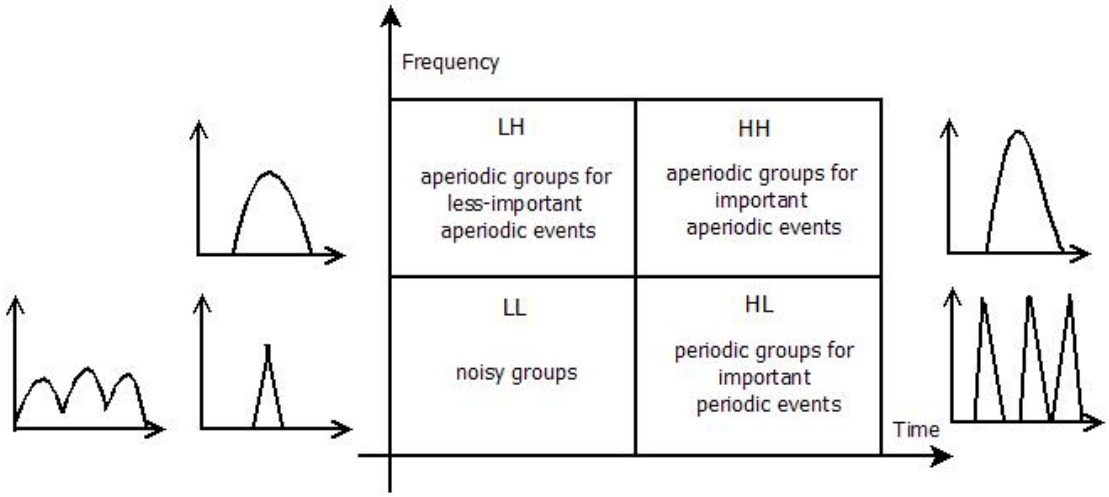

Figure 2. the four group sets for events

Since only groups from HH, HL and LH are meaningful and interesting. They could potentially be representative to events. We discard 95 burst groups classified as LL. Then, we compare and classify similar burst patterns among 85 groups classified as HH, HL, and LH. Table 1 shows the illustration of the method that we detect burst events as burst groups. We classify 2-g38, 3-g11 and 3-g55 as an event. The burst periods of these groups are from 2011/3/13 to 20011/3/17 and from 2011/3/22 to 2011/3/29 and the burst pattern of these

groups are similar.

Table 1. The method that we detect burst events as burst groups.

	1-g236	1-g249	2-g38	2-g130	3-g8	3-g10	3-g11	3-g55
2011/3/8	0	0	0	0	0	0	0	0
2011/3/10...	0	0	0	0	0	0	0	0
2011/3/13...	1	0	1	2	0	0	1	0
2011/3/14...	2	1	1	3	0	1	1	4
2011/3/16...	2	0	0	1	1	0	1	0
2011/3/17...	2	0	1	0	1	1	2	0
2011/3/18...	0	0	0	1	1	0	0	0
2011/3/21...	0	0	0	1	0	0	0	0
2011/3/22...	0	0	2	0	1	1	4	3
2011/3/24...	0	2	1	1	2	2	1	1
2011/3/25...	1	0	1	0	2	0	2	2
2011/3/26...	0	0	0	0	0	0	0	0
2011/3/29...	0	0	0	0	1	2	1	1
2011/3/30...	0	2	0	0	0	0	0	0

We present four burst events as described in the following. Four events comprise important aperiodic events and less-important aperiodic event. Important periodic events are not detected because our data is short of stream mining. Figure 3 shows a line chart of the burst event composed of group 292, group 520, group 564, and group 598. As shown in Table 2, this important aperiodic event is talked about fukushima daiichi nuclear power plant which includes keyphrases such as “fukushima plant”, “nuclear reactor”, “plant”, “radioactive iodine”. Group 292 and 520 have a peak from 2011/3/22 16:00 to 2011/3/24 00:00. Group 564 has a peak from 2011/3/14 16:00 to 2011/3/16 00:00. Group 598 has a peak from 2011/3/25 08:00 to 2011/3/26 16:00.

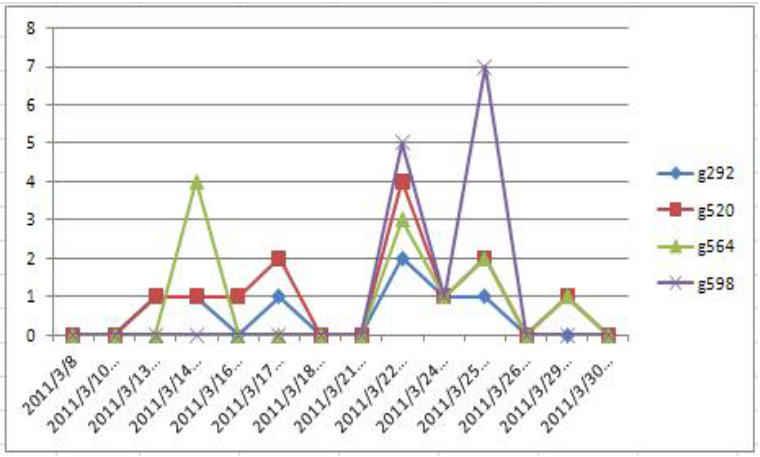

Figure 3. Line chart of the burst event composed of group 292, group 520, group 564, and group 598

Table 2. The keyphrases of group 292, group 520, group 564, and group 598

Group 292	fukushima	fukushima daiichi	fukushima plant
Group 520	reactor	nuclear reactor	nuclear reactors
Group 564	plant		
Group 598	radiation	radioactive	radioactive iodine

This important aperiodic event is talked about Japan's tsunami caused by earthquake is shown in Figure 4 which is composed of group 236, group 384, group 770, and group 750. The keyphrases are "Friday's catastrophic quake", "Japan", "Tokyo", and 'ensuing tsunami" as shown in Table 3. Group 236 remains steady from 2011/3/13 08:00 to 2011/3/18 16:00. Group 384 and 770 have a peak from 2011/3/14 16:00 to 2011/3/16 00:00. group 749 has a peak from 2011/3/24 00:00 to 2011/3/25 08:00.

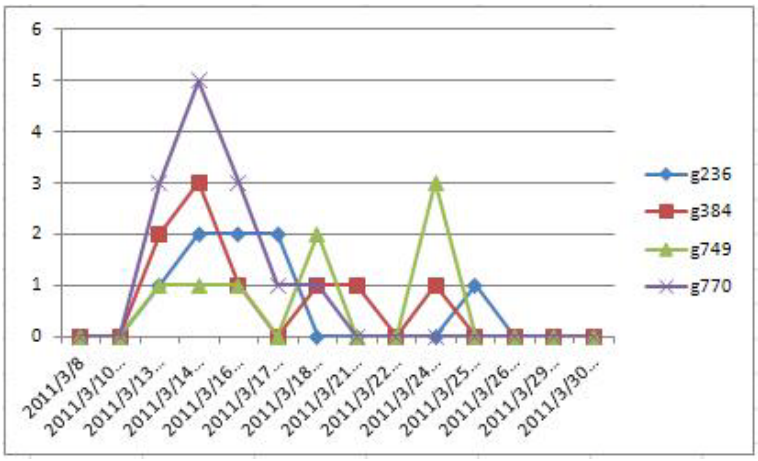

Figure 4. Line chart of the burst event composed of group 236, group 384, group 749, and group 770

Table 3. The keyphrases of group 236, group 384, group 749, and group 770

Group 236	quake	tremor	earthquake	friday's catastrophic quake
Group 384	japan	japan's	japanese	
Group 749	tokyo			
Group 770	tsunami	ensuing tsunami		

Figure 5 shows Line chart of the burst event composed of group 22, group 59, group 162, and group 604. Table 4 shows that this less-important aperiodic event is talked about airstrikes between government forces and rebel forces and having keyphrases such as “air strikes”, “government forces” and “rebel forces”. Group 22 and 604 have a peak from 2011/3/26 16:00 to 2011/3/28 00:00.

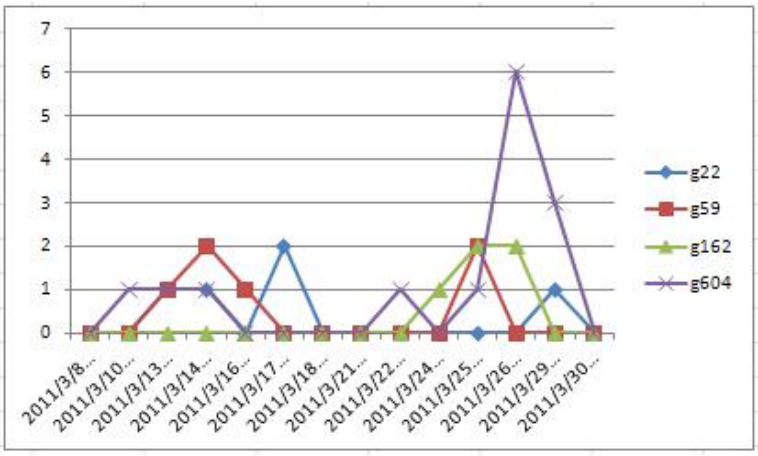

Figure 5. Line chart of the burst event composed of group 22, group 59, group 162, and group 604.

Table 4. The keyphrases of group 22, group 59, group 162, and group 604.

Group 22	air strikes	airstrikes	coalition airstrikes
Group 59	authorities	government	government forces
Group 162	col gaddafi	colonel qaddafi	
Group 604	rebel	rebel forces	

Figure 6 shows Line chart of the burst event composed of group 294, group 425, group 512, and group 781. Table 5 shows that this important aperiodic event is talked about no-fly zone between libya and united states and is having keyphrases such as “gaddafi”, “Libyan people”, “no-fly zone”, “united states”. Group 294, group 425 and group 512 have a peak from 2011/3/25 08:00 to 2011/3/26 16:00. Group 781 has a peak from 2011/3/17 08:00 to 2011/3/18 16:00.

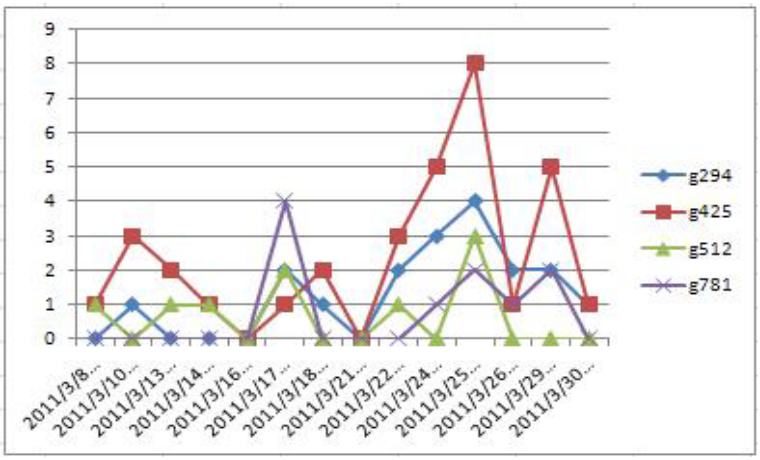

Figure 6 Line chart of the burst event composed of group 294, group 425, group 512, and group 781

Table 5 The keyphrases of Group 294, Group 425, Group 512, and Group 781

Group 294	gaddafi	gadhafi	gaddafi's	gaddafi's regime
Group 425	libya	libyan	libyan people	
Group 512	no-fly	no-fly zone		
Group 781	State	united states		

4.7 Discussion

In the experiments, we extracted four events, which have similar patterns as burst groups. The trend of four events can be seen from the line charts of burst groups. For example, the problem of fukushima daiichi nuclear power plant led to the release of radioactive iodine derived form the first event. Tokyo was damaged by Friday's catastrophic quake and ensuing tsunami derived from the second event. The rebel force extremely rose after air strike between government forces and rebel forces derived from the third event. Libya people had an action after united state set on-fly zone derived from the fourth event.

We learned that fixed time window has an inherent problem. Therefore, we provided the shift and intersections method to alleviate the problem. In order to prevent loosing potential burst groups, the original time window and the shift time window are a good way to settle the problem.

Chapter 5 Evaluation Method

We evaluate the bursts events to see what their causality and quality by manually matching external events. The external events, which are the chosen news stories, are coming from March, 2011 in Wikipedia [22]. Table 6 shows Wikipedia events in 2011 March. From the table, it can be seen that four major event bursts is composed of the occurrence date, and description for the events. The first event is talked about Japan's earthquake and tsunami. From the second to fourth events are all talked about Arab spring and Libya civil war.

Table 7 shows the detected events from our framework. Comparing these two tables, our general observation is that detected burst from our framework is more focusing on Japan's earthquake and fukushima nuclear power plant than Arab Spring and Libya civil war. Furthermore, we can see that the earthquake occurred on Friday and fukushima nuclear power plant had trouble about radioactive iodine. Another general observation is that detected burst from our framework is more focusing on Arab Spring and Libya civil war than king of Bahrain Hamad. We can see that the war and air strike are between government force and rebel force. United state created a no-fly zone on Libyan who is related to gaddafi.

In sum, the detected events from our proposed framework have more adjectives to describe the events than external events.

Table. 6 Wikipedia events in 2011 March

Date	Description of events
2011/3/11	A 9.1-magnitude earthquake and subsequent tsunami hit the east of Japan, killing over 15,000 and leaving another 8,000 missing. Tsunami warnings are issued in 50 countries and territories. Emergencies are declared at four nuclear power plants affected by the quake.
2011/3/15	Arab Spring: King of Bahrain Hamad bin Isa Al Khalifa declares a three-month state of emergency as troops from the Gulf Co-operation Council are sent to quell the civil unrest.
2011/3/17	Arab Spring and Libyan civil war: The United Nations Security Council votes 10-0 to create a no-fly zone over Libya in response to allegations of government aggression against civilians.
2011/3/19	Arab Spring and Libyan civil war: In light of continuing attacks on Libyan rebels by forces in support of leader Muammar Gaddafi, military intervention authorized under UNSCR 1973 begins as French fighter jets make reconnaissance flights over Libya.

Table 7. The detected events from our framework

Event 1	fukushima plant, nuclear reactor, plant and radioactive iodine
Event 2	Friday's catastrophic quake, japan, Tokyo and ensuing tsunami
Event 3	air strikes, government forces and rebel forces
Event 4	gaddafi, Libyan people, no-fly zone, united states

Chapter 6 Conclusion

We make conclusion about our experimental results for our research work in this chapter. The objective of this research work is to use keyphrases for detecting burst events from text streams. Thus, we developed keyphrase-based and semantic-based burst events extraction method in the proposed framework. In the experiments, we have shown that our framework improve the description of burst events. With the keyphrases-based method, the framework provides a structure to detect burst events among hot topic seekers.

However, our research still has some limitation and disadvantages. First, Experimental data is three week Google English news which consisting of different source of news stories from 2011/03/08 to 2011/03/31. The data set seems to be too short of stream mining. It is hard to analyze several months of data with our framework. Second, WordNet can find the synonyms phrase and hypernymys phrase of input word. But, input word should be single term. Although we alleviate this problem by clustering keyphrases with hypernym manually, it is still too subjective to cluster keyphrases. Event though our research improved the description of detect burst events, we still have some disadvantages to overcome. Therefore, we are concerned how to deal with the disadvantages that we have and to modify our framework into a more complete framework.

The framework, which used keyphrases to improve the quality of detecting events, has shown some good results. In the experiments, we have shown that our framework can find burst events through keyphrase-based and semantic-based burst events extraction method. Since there is an inherent problem with fixed time window, we detected events in different time windows. This problem can be alleviated by using the original time window and the shift time window. In evaluation our experimental results, we compare our detected burst events to the external events. This framework indeed enhances the descriptive of burst events by using keyphrases.

Reference

1. J. L. Solka, "Text Data Mining: Theory and Methods," *Statistics Surveys*, vol. 2, 2008, pp. 94-112.
2. N. Parikh and N. Sundaresan, "Scalable and Near Real-Time Burst Detection from eCommerce Queries," *Knowledge Discovery and Data Mining*, 2008, pp. 972-980,.
3. M. Platakis, D. Kotsakos and D. Gunopulos, "Discovering Hot Topics in the Blogosphere," *Department of Informatics and Telecommunications*, 2008.
4. X. Wang, X. Jin, K. Zhang and D. Shen, "Mining Common Topics from Multiple Asynchronous Text Streams," *Web Search and Data Mining*, 2009.
5. V. Hristidis, O. Valdivia, M. Vlachos and P. Yu, "Information discovery across multiple streams," *Information Sciences*, 2009, pp. 3268-3285.
6. Z. Yuan, Y. Jia and S. Yang, "Online Burst Detection Over High Speed Short Text Streams," *The International Conference on Computational Science*, 2007, pp. 717-725.
7. J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Data Mining and Knowledge Discovery*, vol. 7, 2003, pp. 373–397.
8. V. Hristidis, O. Valdivia, M. Vlachos and P. Yu, "Continuous Keyword Search on Multiple Text Streams," *International Conference on Information and Knowledge Management*, 2006, pp. 802-803.
9. X. Wang, C.X. Zhai, X. Hu and R. Sproat, "Mining Correlated Bursty Topic Patterns from Coordinated Text Streams," *Knowledge Discovery and Data Mining*, 2007.
10. L. AlSumait, D. Barbara' and C. Domeniconi, "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," *International Conference on Data Mining*, 2009.
11. X. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous- Time Model of Topical trends," *Knowledge Discovery and Data Mining*, 2006.
12. F. Wei, S. Liu, Y. Song, S. Pan, M. Zhou, W. Qian, L. Shi, Li. Tan and Q. Zhang, "TIARA : A Visual Exploratory Text Analytic System," *Knowledge Discovery and Data Mining*, 2010.
13. G. P. C. Fung, J. X. Yu, P.S. Yu and H. Lu, "Parameter Free Bursy Events Detection in Text Streams," *Very Large Data Base*, 2005, pp. 181-192.
14. W. Chen, C. Chen, L. J. Zhang, C. Wang and J. J. Bu, "Online detection of bursty events and their evolution in news streams," *In Computers & Electron*, 2010, pp. 340-355.
15. Q. He, K. Chang and E. P, Lim, "Analyzing Feature Trajectories for Event Detection," *Special Interest Group on Information Retrieval*, 2007, pp. 207-214.
16. P.D. Turney, "Learning Algorithms for keyphrase Extraction," *Information Retrieval*, vol.2, no. 4, 2000, pp. 303-336.
17. I.H. Written, et al., "KEA: Practical automatic keyphrase extraction," *The Fourth ACM Conference on Digital Libraries*, 1999.

18. S. R. El-Beltagy and A. Rafea, “KP-Miner: A keyphrases extraction system for English and Arabic documents,” *Information Sciences*, 2009, pp. 132-144.
19. KP-Miner system, www.claes.sci.eg/coe_wm/kpminer/
20. Google news, news.google.com.tw/news?edchanged=1&ned=us
21. H. T. Zheng, B. Y. Kang, H. G, Kim, “Exploiting noun phrases and semantic relationships for text document clustering,” *Information Sciences*, 2009, pp. 2249-2262.
22. Wikipedia events in 2011 March, en.wikipedia.org/wiki/2011

使用關鍵片語及詞組從文字資料流找出突發事件

陳聖翔[1]　　吳榮訓[2]
資訊管理學系
國立中正大學
62102 嘉義縣民雄鄉三興村 7 鄰大學路一段 168 號,台灣
[1]welkin2019@gmail.com
[2]roungwu@ccu.edu.tw

摘要

從文字資料流找出熱門話題和事件有許多的應用，且已經引起廣泛的注意。由於關鍵片語及詞組比一個單字更有表達力，更能代表整篇文章的重點。所以在這項研究中，我們試圖使用關鍵片語及詞組從文字資料流找出突發事件。

我們對前面所提到的問題，給一個正式定義，並且提出擁有五個步驟的架構來解決這個問題，(1) 使用關鍵片語及詞組挖掘系統，從文字資料流找出關鍵片語及詞組，把找出來的關鍵片語及詞組視為特徵集合;(2) 對於相同意思或屬於同類別的關鍵詞及片語，進行分群;(3) 計算每一個群體在時間區間裡，出現的頻率;(4) 檢測是否是突發群體;(5) 從突發群體找出突發事件。我們也發現在固定的時間區間裡，會有遺失潛在突發群體的可能。為了減輕這個問題，原本的時間區間加上位移過的時間區間是比較好的解決方法。我們使用谷歌新聞資料流，來去驗證我們提出的架構，實驗結果顯示，我們找出來的突發事件比起外來的突發事件，更加地具有描述性。

關鍵詞： 關鍵片語及詞組挖掘系統, 關鍵片語提取, 資料流挖掘, 文字挖掘, 語法資料庫