PHCM: A Particle Horizontal Cast Movement Based Model for Bursty Events Detection of Chinese Microblog

Le Zhang[1]*, Xueqiang Lv[1], Leihan Zhang[2]

[1] Beijing Information Science and Technology University, China.
[2] Beihang University, China.

* Corresponding author. Tel.: +86-010-82316742; email: yongbuyanqi138@gmail.com

Manuscript submitted September 24, 2014; accepted March 18, 2015.
doi: 10.17706/ijcce.2015.4.6.426-438

Abstract: Microblog has witnessed an explosive development in the past few years. The research on how to detect bursty events timely and efficiently from large amounts of microblog posts has attracted much attention. The key to bursty events detection is the extraction of bursty features. We reconsider the problem from the standpoint of Kinematics theory and look on bursty events detection as a Kinematics phenomenon instead of considering bursty features separately. A novel model named PHCM (Particle Horizontal Cast Movement) is proposed in this paper. Bursty words describing the particular event tend to appear simultaneously in microblog streams. So we shuffle the bursty words appearance sequence in time scale and sort them in ascending order by frequency. Each bursty word is regarded as a moving particle over the time dimension and the movement can be decomposed into a series of consecutive flat parabolic motions with specific time interval. Then the particle's vertical velocity and horizontal velocity in each interval are taken to construct the feature vector of its trajectory. Finally, through computing similarity between different trajectory vectors, we can get bursty words clusters with similar trajectories. By retrieving microblog posts containing bursty words in the same cluster, bursty events can be obtained. Experiments show that our model can detect bursty events accurately as well as efficiently.

Key words: Microblog, bursty events, Kinematics theory, moving particle, trajectories.

1. Introduction

With the development of information technology, social media shows a general development towards diversification. As a tool of cross-platform interactive communication, microblog springs up in recent years. Users can publish short instant posts to share with the public and retweet the information posted by others freely. As an open and transparent social media platform, microblog forms the characteristics of diversity, variability and timeliness. However, it's difficult for users to insight bursty events or continuously keep track of one particular bursty event in large amounts of microblog posts. In addition, bursty events detection is also conductive to public opinion monitoring and social security. So it is meaningful to detect bursty events timely and efficiently.

Recent methods of bursty events detection can be mainly divided into two kinds [1]: methods based on text clustering and burst feature-based approaches.

Methods based on text clustering mainly refer to clustering texts by similarities. The LDA (Latent Dirichlet Allocation) topic model [2] is introduced into bursty events detection in twitter [3]. The LDA model enables bursty event detection by setting the number of hidden topics. While, if the number of implicit theme is set unreasonably, the bursty events detection accuracy will fall down. From the perspective of relevant text streams, Wang *et al*. [4] proposes a scalable probabilistic algorithm, which can find related bursty modes and bursty periods efficiently. Wang *et al*. [5] extracts bursty word set by word frequency statistics, growth rate and TF-PDF to construct feature vector to represent microblog texts and absolute clustering algorithm is applied to get event clusters.

Burst feature-based approaches [6]-[10] mainly refer to clustering bursty words according to bursty features. Sun [7] proposes a method based on SP & HA clustering for micro-blog hot events detection. The single-pass clustering algorithm and coalescing hierarchical clustering algorithm are combined to improve the accuracy of clustering. Du *et al.* [9] proposes a new microblog bursty feature detection method. They computes term weight by taking account of both term frequency and tweet weight. The tweet weight includes retweet number, comment number and time fading factor. What's more, a bursty feature detection method is proposed. The model computes each term's momentum by using MACD (Moving Average Convergence Divergence) to determine whether it is a bursty feature in a given time interval. Guo *et al*. [11] extract bursty words from texts of important users, which may omit much important information of ordinary users.

Currently, methods based on text clustering and burst feature-based approaches have the following deficiencies. Firstly, owing to the limited length of microblog text, the former will meet with the problem of data sparseness when building the text feature vectors. In addition, the text similarity calculation needs much time and space. Secondly, microblog texts contain various kinds of information, which compounds the difficulty of bursty events detection. So there exist two problems to be solved. One is how to filter noise data in microblog text stream, and the other is finding efficient bursty features.

Facing up with the problem of noise data, we build noisy vocabulary to filter useless data automatically. To extract bursty words with efficient bursty features and get the relatively high accuracy of clustering, we propose a novel model called PHCM. From the view of kinematic point, the model regards bursty words in text streams as particles. After the appearance time sequence of bursty words is disrupted and sorted in ascending order by word frequency, the transformation of bursty words in given time window is similar to particle horizontal cast movement. Bursty words with similar trajectories are more likely to describe the same bursty event. The acceleration vector and the angle between velocity vectors are proposed to quantify the similarity of particles' trajectories. Moreover, the concept of correlation rate is proposed, which considers the relationship between bursty words and the distance as they appear in the same text. We combine the cosine similarity of vectors with correlation rate to calculate the similarity. Finally, PHCM clusters bursty words according to similarity of their trajectories.

The remainder of this paper is organized as follows. Section 2 gives a detailed introduction of PHCM. Section 3 discusses the experimental result of our approach. Section 4 concludes the present paper briefly, and points out several future research directions.

2. Particle Horizontal Cast Movement Model

The frequency of bursty words hidden in text streams is time-varying. bursty words describing the same bursty event are always appearing simultaneously in text. Therefore, bursty words representing the same bursty event have a similar variation of frequency in the detection period. Based on the law, we introduce the Kinematic theory into bursty events detection and construct the PHCM model. The theoretical basis of the model is the Horizontal Cast Movement theory. First of all, we build the noisy lexicon containing advertising verbals and daily expressions to filter useless words and get candidate bursty words. From the standpoint of burstiness, the comparison lexicon is built to detect bursty words with high burstiness. In this way, we can get bursty words which make great contribution to bursty event detection. In the next step, the appearance time sequence of bursty words is disrupted and sorted in ascending order by word frequency in each time window. We look on bursty words with varying frequency in different time window as a moving particle. Consequently, trajectories of bursty words in a time window can be viewed as horizontal projectile motions. So we put forward the acceleration vector and the angle between velocity vectors to quantify the similarity of particles' trajectories. Then, the concept of correlation rate is proposed to measure the relationship between bursty words and the distance as they appear in the same text. We combine the cosine similarity of vectors with correlation rate to calculate the similarity. Finally, bursty words are clustered according to similarity of their trajectories.

2.1. Bursty Words Extraction

In the process of bursty events detection, bursty words are usually regarded as bursty features. The accuracy of bursty words extracted from text streams directly influences the detection results of bursty events. The effectiveness and burstiness are considered when bursty words are extracted. The effectiveness refers to the correlation between the candidate bursty words with the bursty event. The microblogging space is filled with amounts of noise information, such as advertising information and daily expressions, which are of no use to detect bursty events, and may further disturb the bursty words extraction. The noisy lexicon is constructed automatically to filter useless information. The burstiness is the specific attribute of bursty words. The higher the burstiness, the more likely to describe bursty events. By calculating the bursty rate, we can get bursty words having relatively high burstiness.

- **Definition 1: Bursty Word**

The word appears in large numbers suddenly along with one particular bursty event in very short time.

- **Definition 2: Bursty Rate**

Bursty rate refers to the probability of the word being a bursty word.

- **Definition 3: Noisy Word**

The noisy word refers to the word without much effective information for bursty event detection.

- **Definition 4: Time Window**

The time window refers to an independent time period.

The detection period of bursty events can be represented as T and the length of the time window is set as len . Therefore, the number of time windows is T/len .

2.1.1. Effectiveness extraction

Microblog text is filled with much jumbled information. To extract more effective bursty words, we construct automatically the noisy lexicon. With a stable and high frequency, the word is screened out as noisy words from the microblog texts of the previous year relative to the detection period to construct noisy lexicon. The formulas of noisy words' weight are as follows:

$$W(w_i) = \frac{f(w_i)}{n \times (\delta(w_i) + 1)} \tag{1}$$

$$\delta(w_i) = \frac{\sum_{i=1}^{n} [f(w_i) - \sum_{j=1}^{n} f(w_j)/n]^2}{n} \tag{2}$$

$f(w_i)$ denotes the frequency of word w_i in microblog text on a daily basis. n denotes the number of days in the detection period. $\delta(w_i)$ denotes the variance of the word w_i in the daily frequency sequence.

$W(w_i)$ is the effectiveness weight of w_i.

A few of noisy words are listed in Table 1:

Table 1. Noisy Lexicon

Advertising Verbals	Daily Expressions
分享(share),	早安(good morning), 眼花缭乱(dazzled)
时尚(fashion),	开心(happy), 容量(volume),
返利(rebate),	俯视(overlook),
正品(quality goods),	大声(loud), 好事(good deed),
肌肤(skin)	麻木(numb)
产品(product),	虚无(nihility), 食用油(cooking oil),
身材(figure)	黄昏(dusk)

2.1.2. Burstiness extraction

Both of the comparative data and the detection data are filtered by noisy lexicon. The comparative data refers to the data from the previous week before the detection date, words with high frequency in comparative data and their responding frequency are added into the comparative map $cmap$. Similarly, words with high frequency and their frequency in the detection data are added into the candidate map $dmap$.

The rules of burstiness extraction are as follows:

Rule 1: To each candidate word w_i, if its frequency is greater than the threshold λ and it is not in the comparative map, the word w_i will be added into the set S_1.

Rule 2: If both the candidate map and the comparative map contain the word w_i and its growth rate of frequency $GR(w_i)$ is greater than the threshold γ. Then the word w_i will be added into the set S_2.

The growth rate of frequency $GR(w_i)$ is calculated as follow:

$$GR(w_i) = \frac{fc(w_i) - fd(w_i)}{fd(w_i)} \tag{3}$$

$fc(w_i)$ represents the frequency of word w_i in $dmap$ and $fd(w_i)$ represents the frequency of word w_i in $cmap$.

Based on the above rules, S_1 and S_2 are constructed. Then, they are merged into the bursty set S.

2.2. Bursty Word Pairs Extraction

- **Definition 5: Co-occurrence Distance**

Co-occurrence distance refers to the distance between two words when they appear simultaneously in the same text.

Bursty words describing the same bursty event tend to appear together in the same text. The smaller the co-occurrence distance, the greater probability of the words describing one same event. For instance, the content of a text is "爸爸去哪儿，今日在湖南卫视首次播出 (The entertainment, Dad Where Are We Going, will be broadcast firstly on Hunan TV)", and the other is "爸爸一大早就去上班了，我自己也不知道去哪玩 (my dad went to work in the early morning, and I didn't know where to go.)". Both of the text contain the word "爸爸(dad)" and "去哪(where)". When only considering the co-occurrence frequency, the two sentences cannot be distinguished. If we take the co-occurrence distance into account, the two sentences can be differentiated easily.

The formula of co-occurrence distance is as follow:

$$d(w_i, w_j, t_m) = | idx(w_i, t_m) - idx(w_j, t_m) | \tag{4}$$

$idx(w_i, t_m)$ represents the position of the bursty word w_i in microblog text t_m. $d(w_i, w_j, t_m)$ represents the co-occurrence distance between w_i and w_j when they appear in t_m.

- **Definition 6: Correlation Rate**

The correlation rate refers to the degree of association between words.

The Activation Force model [12] is a statistic model. The text is converted to the sequence of words. In the sequence, the word appearing firstly has a trigger affection towards the following words. Based on the word frequency and the co-occurrence information, we build excitation intensity between words. The excitation intensity captures important information on the word network. The method can properly reflect the syntactic and semantic information.

Based on the activation force and co-occurrence, we consider the co-occurrence distance and propose the concept of correlation rate to quantify the association degree between words.

The correlation rate $R(w_i, w_j)$ between the bursty word w_i and w_j is calculated as follow:

$$R(w_i, w_j) = \frac{f(w_i, w_j)^2 / [f(w_i) \times f(w_j)]}{[\sum_{m=1}^{count} d(w_i, w_j, t_m) / count]^2} \tag{5}$$

$f(w_i, w_j)$ represents the co-occurrence frequency of w_i and w_j. $f(w_i)$ and $f(w_j)$ represent the frequency of w_i and w_j when they do not appear simultaneously. $count$ is the total number of the bursty word w_i and w_j appearing times in one same microblog text. t_m represents the *m*th microblog text t_m containing both w_i and w_j. $d(w_i, w_j, t_m)$ is the co-occurrence distance of the bursty word w_i and w_j in t_m.

For w_i and w_j in bursty set S, we compute their correlation rate $R_d(w_i, w_j)$ in detection data and $R_c(w_i, w_j)$ in comparative data. We select the bursty words pair according to their correlation rate difference in detection data and comparative data. When the condition of formula (6) is satisfied, w_i and w_j will be taken as a pair of bursty words and added into the bursty words pair set $BWPS$.

$$R_d(w_i, w_j) - R_c(w_i, w_j) > \delta \qquad \text{where } \delta > 0 \tag{6}$$

2.3. Model Quantification

The Horizontal Cast Movement in Physics is defined as: objects with a certain initial velocity is thrown in the horizontal direction and the object is only influenced by gravity, such a motion is called as Horizontal Cast Movement. The trajectory of Horizontal Cast Movement is a parabolic.

Similarly, when the bursty word particle is thrown in the horizontal direction with a certain initial velocity, it is affected by the social attention in microblog stream. And the influence can be seen as changeless in a small time interval. So the trajectory of a bursty word particle can approximately be taken as a parabolic. So we construct the acceleration vector and the angle between velocity vectors to quantify particles' trajectories.

2.3.1. Feature vectors

In different time intervals, bursty word particles do variable accelerated motion in the vertical direction. The acceleration of bursty word particle in the time interval *n* is set as a_n. The formula of a_n is as follow:

$$a_n = \frac{\sqrt{v_n^2 - v_0^2} - \sum_{i=1}^{n-1}(a_i \times t_i)}{t_n} \tag{7}$$

Here is the derivative process of a_n: v_i is used to describe the velocity of a bursty word particle in the *i*th time interval. Similarly, vy_i and vx_i describe the vertical velocity and the horizontal velocity in the *i*th time interval. And the initial vy_0 is assigned with 0.

$$vy_1 = vy_0 + a_1 \times t_1 \Rightarrow a_1 = \frac{vy_1}{t_1} = \frac{\sqrt{v_1^2 - v_0^2}}{t_1}$$

When *n* =2,

$$vy_2 = vy_1 + a_2 \times t_2 \Rightarrow a_2 = \frac{\sqrt{v_2^2 - v_0^2} - a_1 \times t_1}{t_2}$$

We can assume that when *n* =*k*-1,

$$a_{k-1} = \frac{\sqrt{v_{k-1}^2 - v_0^2} - \sum_{i=1}^{k-2}(a_i \times t_i)}{t_{k-1}}$$

Then,

$$vy_k = vy_{k-1} + a_{k-1} \times t_{k-1} \Rightarrow a_k = \frac{\sqrt{v_k^2 - v_0^2} - \sum_{i=1}^{k-1}(a_i \times t_i)}{t_k}$$

So,

$$a_n = \frac{\sqrt{v_n^2 - v_0^2} - \sum_{i=1}^{n-1}(a_i \times t_i)}{t_n}$$

Bursty word particles do variable accelerated motion in the vertical direction and the direction of velocity is changeable. The angle between velocities in the time window n is expressed with A_n. And we can get A_n as follow:

$$A_n = \arctan \frac{\sqrt{v_n^2 - v_0^2}}{v_0} \tag{8}$$

The derivative process is similar with acceleration vector.

When n =1,

$$A_1 = \arctan\frac{vy_1}{v_0} = \arctan\frac{\sqrt{v_1^2 - v_0^2}}{v_0}$$

When n =2,

$$A_2 = \arctan\frac{vy_2}{v_0} = \arctan\frac{\sqrt{v_2^2 - v_0^2}}{v_0}.$$

Finally,

$$A_n = \arctan\frac{vy_n}{v_0} = \arctan\frac{\sqrt{v_n^2 - v_0^2}}{v_0}$$

The acceleration vector of bursty word particle w_i in detection period T can be represented as $\boldsymbol{a} = (a_1, a_2, \cdots, a_k, \cdots, a_{len})$. The angle between velocity vectors of the bursty word particle w_i can be represented as $\boldsymbol{A} = (A_1, A_2, \cdots, A_k, \cdots, A_{len})$. The feature vector of the bursty word particle w_i can be represented as:

$$\boldsymbol{vec}(\boldsymbol{w}_i) = wa + (1-w)A \tag{9}$$

where, $1 \le k \le len$, and len represents the number of time windows. w is coefficient threshold value of feature vectors.

2.3.2. Similarity computation

The trajectory of the bursty word particle is quantified as a len dimensional vector $\boldsymbol{vec}$. So the similarity of trajectories can be calculated by calculating the similarity of vectors. The cosine similarity and correlation rate are combined to calculate the similarity of vectors. The similarity of trajectory is calculated as follow:

$$\begin{gathered} sim(w_i, w_j) = \frac{\boldsymbol{vec}(\boldsymbol{w}_i) \times \boldsymbol{vec}(\boldsymbol{w}_j)}{|\boldsymbol{vec}(\boldsymbol{w}_i)||\boldsymbol{vec}(\boldsymbol{w}_j)|} \times R(w_i, w_j) \\ where\ (w_i, w_j \in S) \end{gathered} \tag{10}$$

$\boldsymbol{vec}(\boldsymbol{w}_i)$ and $\boldsymbol{vec}(\boldsymbol{w}_j)$ separately represents the feature vector of bursty word w_i and w_j. $R(w_i, w_j)$ is the correlation rate between w_i and w_j. S is the bursty word set. $sim(w_i, w_j)$ is the trajectory similarity of w_i and w_j.

2.4. Bursty Event Detection

In this paper, we combine the cosine similarity with correlation rate to calculate the similarity of particles' trajectories. The agglomerative hierarchical clustering algorithm is used to cluster the similar trajectories of bursty words into different clusters. The representative bursty words in the cluster are used to represent the bursty event.

The agglomerative hierarchical clustering algorithm [13], [14] is commonly used in the field of data mining. The algorithm initializes each element which will be clustered as a cluster. Then, the clusters with the shortest distance will be merged from the bottom up. The process will stop when it meets the set condition. The detailed clustering process is shown as follows:

Input: *Bursty words*

Output: *Bursty clusters*

Step 1: The feature vector of every bursty word particle is taken as a cluster.

Step 2: Calculate the distance between any two clusters contained in the bursty words pair set $BWPS$. Then, merge the two clusters with the shortest distance.

Step 3: Recalculate the distance of every cluster and get the shortest distance between clusters.

Step 4: If the shortest distance is smaller than the threshold, the Step 2 will be executed. If not, the clustering will be stopped and output the results.

The cluster, which will be clustered, consists of the feature vector of the bursty word particle. The distance between clusters is influenced by the distance between feature vectors. So we define the distance between clusters as the inverse of the similarity of every two elements in different clusters.

The distance between clusters is calculated as follow:

$$d(c_a,c_b)=\begin{cases}\dfrac{1}{\sum\limits_{\substack{w_i\in c_a\\ w_j\in c_b}} sim(w_i,w_j)} & \sum\limits_{\substack{w_i\in c_a\\ w_j\in c_b}} sim(w_i,w_j)>0\\ \infty & else\end{cases} \tag{11}$$

$sim(w_i,w_j)$ is the similarity of bursty word particle w_i and w_j . The $d(c_a,c_b)$ represents the distance between the cluster c_a and c_b .

After the above process of clustering, bursty words are classified as different clusters. As an event should include three elements (when, where, what is going on) at least, we retain only the cluster containing more than three bursty words. For instance, the cluster {马航(MAS),飞机(flight),失联(missing)} refers to the bursty event "2014年3月8日凌晨2点40分，马来西亚航空公司称与一架载有239人的波音777-200飞机与管制中心失去联系的，该飞机航班号为MH370，原定由吉隆坡飞往北京"(At 2:40 on March 8, 2014, Malaysia Airlines said that the MAS flight MH370 carrying 239 Boeing 777-200 aircraft was missing, which was originally from Kuala Lumpur to Beijing).

3. Experiment and Analysis

3.1. Dataset and Evaluation

We get about 10.80 million weibos from March 1 to March 18 in 2014 through the open API of Sina Weibo. And the data is divided into two periods: March 1 to March 4 and March 8 to March 11. The dataset is available at http://www.datatang.com/data/46566.

In the event detection, the omission rate $Rmiss$ and the false detecting rate $Rfalse$ are generally used to evaluate the efficiency of events detection. They are put forward by the U.S. National Institute of Standards and Technology (NIST).

$$Rmiss=\frac{c}{a+c} \tag{12}$$

$$Rfalse=\frac{b}{b+d} \tag{13}$$

a represents the number of related texts in detection result. c represents the number of related texts misjudged by the model. b represents the number of uncorrelated texts in detection result. d represents the number of uncorrelated texts misjudged by the model. We define another evaluation metric P by combining $Rmiss$ and $Rfalse$ [15].

$$P = 1 - 0.5 \times Rmiss - 0.5 \times Rfalse \tag{14}$$

The threshold parameters are shown in Table 2.

Table 2. Threshold Parameters List

Symbol	Value	Instruction
λ	10	Threshold of frequency
γ	10	Growth rate of frequency
w	0.5	Eigenvector coefficient threshold
T	24	Detection period
len	3	Length of the time window
δ	10	Threshold of correlation rate

Table 3. Experimental Results

μ	$Rmiss$	$Rfalse$	P
0.1	39.54	0.0	80.23
0.3	19.11	2.0	89.43
0.5	12.28	2.0	92.85
0.7	13.81	10.62	87.78
0.9	13.31	10.16	88.26

3.2. Results and Analysis

In the agglomerative hierarchical clustering algorithm, the distance μ between clusters is the condition to decide whether two bursty words will be clustered into one topic. Considering the importance of the threshold μ, we conduct five experiments by setting different μ values with 0.1, 0.3, 0.5, 0.7, 0.9, and find the optimum value of μ is 0.5. The results are shown in Table 3.

As the Table 3 shows, the correct rate is up to the optimal value 92.85% with the threshold μ set as 0.5. The correct rate will descend when the threshold μ become smaller or bigger than 0.5. The reason is as follows: if the threshold μ is too small, it will shorten the clustering process. Therefore, bursty words describing the same bursty event cannot be clustered together, so the omission rate rose and the correct rate decreased. Inversely, if the threshold μ is too big, it will prolong the clustering process. As a consequence, irrelevant bursty words will be clustered together, which results in the improvement of false detecting rate and the decrease in correct rate.

With the threshold μ being set as the optimal value 0.5, we detect bursty events in two detection periods using the PHCM model. Fourteen clusters, which can represent correctly bursty events, are detected totally. The results are shown in Table 4.

3.3. Comparison with Other Methods

In this paper, the PHCM model is used to detect bursty events on a daily basis, which is similar with the topic modeling. The goal of the topic modeling is to discover "topics" hidden in the text streams. Therefore, we apply the widely used statistical topic model Latent Dirichlet Allocation (LDA) on the dataset. Then, the result generated from LDA is compared with our results.

Table 4. Detection Results in the First Period (2014.03.01-2014.03.04&2014.03.08-2014.03.11)

Date	Clusters	Events

2014/3/1 -2014/3/2	{暴力(voilence)，恐怖(terror)，袭击 (assault)，伤人 (hurt)，受伤(injured)，死亡(die)}
	{昆明(Kunming)，火车站(train station)，遇难(murdered)，受害者(casualty)，默哀 (silence)，祈福(bless)}
		1.云南昆明火车站暴力恐怖袭击事件。(The Kunming railway station occurred the violent terrorist attacks event.)
		2.市民在昆明火车站为遇难受害者默哀祈福。(People in Kunming Railway Station stood in silent to pray for the victims and victims.)

2014/3/3	{秩序(order)，谣言(rumour)，恐慌(panic)}
	{严惩(chastise)，可恶(hateful)，暴徒(ruffian)}
	{恐惧(fear)，愤怒(anger)，歧视(discriminate)，扭曲 (angulation)，谴责(condemn)，仇恨(hatred)}
		3.昆明事件网络谣言疯传，给民众造成了恐慌。(Network rumors concerning Kunming event pass quickly, which caused public panic)
		4.民众对昆明事件十分气愤，呼吁严惩暴徒份子。(People was very angry for Kunming event and called for punishing the terrorist severely.)
		5.部分民众将对暴徒份子的愤怒转移到对一个民族的敌意。(Some people transferred the angry for terrorist to a hostile for the nation)

2014/3/4	{海岸(seaboard)，火山(volcano)，岛屿(islands)，乐园(paradise)}
	{铭记(engrave)，仇恨(hatred)，淡忘(chillax)}
		6.绿岛成为东部的海上乐园。(The Green Island became a sea paradise of the eastern)
		7.在"将对暴徒份子的愤怒转移到对一个民族的敌意"这件事情上出现了反对的呼声。(Many opposition voices were occurred for the third event detected in 2014.3.3)

2014/3/8 -2014/3/9	{马航(MAS)，飞机(flight)，失联(missing)}
	{中国银行(Bank of China)，机组(aircrew)，家属(family members)，遇难(murdered)}
	{平安(safe)，阿弥陀佛(Amitabha)，保佑(bless)}
	{委员(committee member)，人大代表(NPC member)，政协(cppcc)，农业(agriculture)，两会(The Two Sessions)，转基因(transgenosis)}
	{子女(children)，名义(name)，买家(buyer)，好美(beautiful)}
		8.马航飞机失联。(The MAS flight was missing)
		9.中国银行为遇难家属及机组人员提供签证加急服务。(Bank of China provided expedited service of visa for the victims and families.)
		10.祈福马航平安。(blessing for MAS.)
		11.人大代表政协委员崔永元两会谈转基因被封杀。(Yongyuan Cui talked about the genetically modified in the Two Sessions.)
		12.出现以子女名义买房的潮流(The trend that parents buy a house in the name of their children is current.)

2014/3/10	{希望(hope)，雷达(radar)，感人(touching)，呼叫(call)}
	{爆炸(explode)，线索(clue)，残骸(remains)，保佑(bless)}
		13.祈福马航平安(MH370 管制雷达希望看见你---马航感人塔台呼叫)。(MH370 Control Radar hope to see you --- Malaysia Airlines touching the tower call.)

2014/3/11	{雷达(radar)，航空公司(airways)，架飞机(flight)，失事(crash)，揪心(worried)}	14.马来西亚航空公司发布新闻发布会，表明客机失事可能性大。(Malaysia Airlines issued a news conference indicating a large possibility of airliner crash.)

Table 5. Topics Detected by lda

Date	Probability	Top Words
2014/3/2	0.051	昆明(Kunming)，祈福(bless)，照片(photo)，容易(easy)，辽宁(Liaoning)，觉得（think）
2014/3/9	0.051	飞机(flight)，马来西亚(Malaysia)，得到(got)，联系(contact)，消防局(fire department)
2014/3/3	0.020	恐怖分子(terrorist)，知道(know)，清明(tomb-sweeping)，想要(want)
2014/3/10	0.020	马航(MAS)，飞机(flight)，乘客(passenger)，悲伤(sadness)，火车站(train station)
2014/3/4	0.013	牵挂(worry)，昆明(Kunming)，难受(unhappy)，现在(now)，早期 (early phase)

We regard the dataset of one day as a document, so there are eight documents in our experiment. In LDA, each document may contain various topics and each topic contains various words. The LDA assumes that the document-topic distribution has a Dirichlet prior (with hype-parameter α') and the topic-word distribution has a Dirichlet prior (with hype-parameter β'). In our experiment, the topic number T' is set as 50. The hype-parameter α', β' are separately set as 50 and 0.1. The identified top-5 topics are listed in Table 5. The "probability" in this table refers to the probability that the corresponding topic appears in a particular day.

As shown as the Table 5, we can see that topic words identified by LDA are too ambiguous to present the bursty event. For instance, the word “昆明(Kunming)” and the word “祈福(bless)” are identified as the top words for the most related topic on March 2, 2014. However, the word “照片(photo)” and the word “辽宁(Liaoning)” are mixed with them as well, which increases the difficulty of identifying bursty events. In addition, if the number of topics $T^{'}$ is reset, the LDA will return a new distribution over $T^{'}$ topics for each document even if the document hasn’t discussed about any real-life event. Therefore, to improve the results generated by LDA for event detection, further processing should be taken into account. While, according to the effectiveness and burstiness, the PHCM model can filter trivial words useless to events detection. More importantly, the number of bursty events does not need to be preseted and are generated automatically in the processing of clustering.

The method of bursty events detection approach based on burst words clustering [11] is regarded as the second comparative experiment. The weight threshold of bursty words is set as 20. The incremental clustering threshold D is set separately as 300,400,500,600,700,800. The correct rate is up to the optimal value 70.69% when the value of D is 600. At the same time, the omission rate is 48.82% and the false detecting rate is 9.79%.

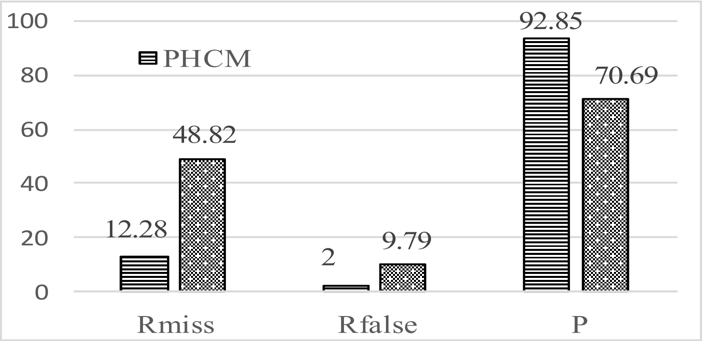

Fig. 1. Comparative results.

Table 6. Topics Detection by Baseline

Date	Clusters
2014/3/1	{庭(court)，谊(Yi)，谅(Liang)，址(Zhi)} ，{钝(Dun)，优柔寡断(indecision)，匐(Pu)，挑剔(picky)}
2014/3/2	{屠杀(massacre)，畜生(beast)，抓获(arrest)，击毙(killed)，庭(court)，暴行严惩(punish severely)，暴徒(terrorist)，无辜(innocent)，砍(cut)，昆明(Kunming)，火车站(railway station)}
2014/3/3	{默哀(silence)， 暴徒(terrorist)，惩(punish)，伤者(injured)}
2014/3/4	{举止(manners)，惧(fear)，烹(Peng)，牲(Sheng)}
2014/3/8	{降落(landing)，载有(loading)，吉隆坡(Kuala Lumpur)，马航(MAS)，航班(flight)，马来西亚(Malaysia)}
2014/3/9	{虚惊(false alarm)，机组(aircrew)，降落(landing)，吉隆坡(Kuala Lumpur)，失事(crash)，护照(passport)，马航(MAS)，航班(flights)}
2014/3/10	{ 马航(MAS),，护照(passport)，降落(landing)}
2014/3/11	无

Results of comparative experiment are shown in the Fig. 1, and the detection results (clusters) of baseline are shown in Table 6.

As shown in experimental results, the PHCM model proposed in this paper is superior to the baseline. In addition, clusters in the baseline contain much noisy information and cannot be easily interpreted into bursty events. By analyzing the experiment, the main reason can be concluded into the following two points:

Firstly, in the microblogging space, ordinary users account for a large proportion and make some contribution to the bursty event detection. While [11] only extracts the bursty word from texts of important users selected by the influence, which will omit much important information of ordinary user. In this paper, we extract bursty words according to the effectiveness and burstiness, which avoids omitting the user information.

Secondly, in this paper, we propose the PHCM model based on the Horizontal Cast Movement theory. The bursty word is taken as a particle. Bursty words are clustered according to the similarity of particle's trajectories. In addition, the relationship between bursty words and the co-occurrence distance are taken into account, which can improve the identification accuracy of similar trajectories. Moreover, bursty word pairs are extracted, which can shorten the time of clustering.

4. Conclusion

We regard the bursty detection as a Kinematics phenomenon and propose the PHCM model based on the Kinematics theory. In addition, in terms of the effectiveness and burstiness, the noisy lexicon and the comparative map are constructed automatically to get bursty words. Finally, the concept of correlation rate is presented to improve the identification accuracy of the similar trajectories. Experimental results show that the PHCM is effective on bursty events detection.

In the future work, we will research on the iterative algorithm to get the optimal value of the time window.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant (No. 61271304), Key project of Beijing City Board of education technology development and the B class key project of Beijing Natural Science Foundation (No.KZ201311232037) and the Beijing University Innovation Team Building and Teacher Occupation Development Project (IDHT20130519).

References

[1] Zhang, L. M., Jia, Y., & Zhou, B. (2012). Research microblogging incident detection method based on affective computing. *Information Network Security*, *8*, 143-145.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

[3] Diao, Q., Jiang, J., & Zhu, F. (2012). Finding bursty topics from microblogs. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*: *Vol. 1. Association for Computational Linguistics* (pp. 536-544).

[4] Wang, X., Zhai, C. X., Hu, X., *et al.* (2007). Mining correlated bursty topic patterns from coordinated text streams. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Jose.

[5] Wang, Y., Xiao, S. B., Guo, Y. X., *et al.* (2013). Detection of emergencies Chinese microblogging. *Library and Information Technology*, *29*, 57-62.

[6] Liu, F. (2011). Based microblogging outbreak detection and dissemination of information modeling. Harbin Institute of Technology.

[7] Sun, S. P. (2011). Chinese micro-blog hot topic detection and tracking technology research. Beijing Jiao tong University.

[8] Yang, G. C. (2011). Micro-blog hot topic discovery strategy research. Zhejiang University, Hangzhou, China.

[9] Du, Y., Wu, W., *et al.* (2012). Microblog bursty feature detection based on dynamics model. *Proceedings of 2012 International Conference on Systems and Informatics* (pp. 2304-2308). Yantai, China.

[10] Fung, G. P., Yu, J. X., Yu, P. S., *et al.* (2005). Parameter free bursty events detection in text streams. *Proceedings of the 31st International Conference on Very Large Data Bases*.

[11] Guo, Y. X., Lv, X. Q., & Li, Z. (2014). Unexpected word clustering based microblogging outbreak detection method. *Computer Applications*, 34, 486-490.

[12] Guo, J., Guo, H., & Wang, Z. (2011). An activation force-based affinity measure for analyzing complex networks. *Scientific Reports*.

[13] Chuang, S. L., & Chien, L. F. (2002). Towards automatic generation of query taxonomy: A hierarchical query clustering approach. *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 75-82). Maebashi City, Japan: IEEE Computer Society Press.

[14] Brandes, U., & Wagner, G. (2003). Experiments on graph clustering. *Proceedings of the 11th Annual European Symposium on Algorithms (ESA.03)*: *Vol. 2832*. *Lecture Notes in Computer Science* (pp. 568-579).

[15] Xue, S. Z., Lu, R., & Ren, Y. Y. (2013). Based on the rate of growth of microblogging hot topic. *Application Research of Computers*, *30*, 2598-2601.

Le Zhang is currently a bachelor candidate at Beijing Information Science and Technology University, Beijing, China. Her research interests include natural language processing and data mining.

Xueqiang Lv received the Ph.D. degree in Northeastern University, Shenyang, China. He is currently a professor in the Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University. His research interests include multimedia processing and natural language processing.

Leihan Zhang received his master's degree in computer applications from Beijing Information Science and Technology University, Beijing, China in 2014. He is now a Ph.D. candidate in Beihang University. His research interests focus on complex network, data mining on social network.