Microblog Bursty Feature Detection Based on Dynamics Model

Yanyan Du Wei Wu
School of Computer
Wuhan University
Wuhan, China

Yanxiang He Nan Liu
Key Lab of Information Network Security
Ministry of Public Security
Shanghai 201204, China

***Abstract*—Microblog is becoming more and more popular in our life. Due to the numerous information on this platform, it is very useful to detect bursty topic in real-time to help people get essential information quickly. As a necessary stage, detecting busty feature effectively is important for bursty topic detection. Based on dynamics model, we propose a new microblog bursty feature detection method. Firstly, we compute term weight taking account of both term frequency and tweet weight, where tweet weight factors include retweet number, comments number and time fading factor. After computing all terms' weight, a bursty feature detection method is proposed based on dynamics model. On the analogy of physical dynamics model, we compute each term's momentum by using MACD (Moving Average Convergence/Divergence) and determine whether it is a bursty feature in a given time interval. We employ our method to detect the bursty terms of Sina tweets with a series of experiments. It is demonstrated that our method is able to detect bursts for news terms accurately and efficiently.**

Keywords-microblog; bursty feature; topic detection; dynamics model

I. INTRODUCTION

Nowadays, data stream burst detection techniques are widely used in many fields, for example, in financial field we can detect high stock trading volume or significant price fluctuation in short time to explore the developing trend of price, in network management we can detect too much accessing requests timely to avoid attacks and explore truth people pay much attention to, and so on. So this is a meaningful field. As the occurrence of microblog platform, it is become more and more essential to provide hot topic for people timely and succinctly, that's because microblog has large amount of information and millions of users, so when logging on it every day, people want to focus on current affairs more conveniently and quickly.

As the basic task of topic detection, the accuracy of bursts detection will infect the accuracy of topic detection directly, so just considering bursts detection on microblog platform, we make a new computational method of term weight which consider both term frequency and tweet weight, and then propose a bursty feature detection method based on MACD. The paper is organized as follows: first present the current state of art on data stream burst detection in section 2, and introduce burst detection method based on MACD in section 3, then introduce Microblog bursty feature detection implementation in section 4, later describe related experiments and evaluation in section 5, at last give conclusion in section 6.

Support by the National Science Foundation of China (NSFC) under Award 61070083, 90820005, 61070082; Youth Science and Technology Morning program of Wuhan under Award 201150431105

II. RELATED WORK

Since TDT (Topic Detection and Tracking) task were proposed in 1946, there has been a great deal of research work, in which HMM (Hidden Markov Model)[1] Aging Theory[2], Time Series Analysis[7]and LDA Latent Dirichlet Allocation) model[4] have become the classic models in TDT.

As an important part of TDT, data stream burst detection techniques also develop rapidly. According to studying object, we can divide them into two species: one is based on time series, which mainly are applicable to such fields as finance, communication and so on. For this species, according to differences in theoretical basis, we can also divide them into four species further: the first one is based on Wavelet Analysis, [9] propose a method of detecting data stream on elastic windows and design a new data structure SWT (Shifted Wavelet Tree) to detect different lengths of aggregate burst in different time scales. Paper [3] adapts Lasting Factor and Abrupt Factor to describe two burst types, namely Lasting Burst and Abrupt Burst, and detect burst on a new data structure called Two-Layered Wavelet Tree. The second one is based on Self-similarity, [13] adapt The b-model to simulate burst time series according to similarity principle, then estimate the model by entropy, while [11] adapt the self-similarity of aggregate function values in different sliding window and detect burst in monotonic searching space. The third one is based on ratio, [12] design adaptive aggregate burst detection algorithm, which can realize burst detection in any time window size. The forth one is based on moving average, [9] adapt a method of setting threshold dynamically by comparing the ratio difference among each protocol of flows. The other one is based on content, which mainly are applicable to text stream, it can be divided into two species further: the first one is based on automaton, the most classic burst algorithm is proposed in [10], which define burst as change in states and adapt infinite state automaton to model the change, [5] generate a static topological graph based on analyzing Blog's structure, and take the number of links of blog space as input of automaton. The second one is based on burst feature, which establish a statistical model to detect distribution of each feature according to time, then define burst standard and cluster bursty features to detect the optimal bursty event.

According to analyze current burst detection techniques, most of them are applicable to such data which are mainly long, normative and uniform in distribution, however, microblog platform data has such problems as short text, sparse data and redundant information and so on, traditional techniques may are not fully applicable to microblog platform,

 978-1-4673-0199-2/12/$31.00 ©2012 IEEE

so we introduce a new burst detection method based on MACD.

III. BURST DETECTION BASED ON MACD

In our method, compared with many traditional definitions about burst, we give burst an impactful definition. Refer to physical dynamics model in [6], similar to author's topic bursts definition, we redefine burst as a dynamic phenomenon and construct burst as the rate of change of momentum, further to say, if this rate of a term is positive over a set time interval, the term may be considered as a burst. As defined in physics, momentum is the product of mass and velocity, where velocity is the rate of position change. In our method, we redefine mass as the current importance of the tweet which term belonging to and position as its frequency.

To measure burst accurately and efficiently from these known values, we adapt technical stock market trend analysis tools to analyze the change of momentum in our method, such as EMA (Exponential Moving Average) and MACD[8]. In stock market, EMA and MACD are used to spot changes in the strength, direction, momentum, and duration of a trend in a stock's price. EMA highlight recent changes in a stock's price, MACD is a computation of the difference between two (EMAs) of closing prices, by comparing EMAs of different periods, the MACD line illustrates changes in the trend of a stock, then by comparing that difference to an average, an analyst can chart subtle shifts in the stock's trend. Briefly, by using differences of smoothed values (moving averages), users can estimate derivatives (rates of change), then to analyze trends. For stocks, position can be a measure of variable like stock price and mass can be a measure of market importance. Analogy to these concepts, we can also adapt these techniques in our method and redefine basic concepts here.

EMA: for a variable $x = x(t)$ in discrete time intervals $x = \{x_t \mid t = 0,1,...,\}$ the n-day EMA can be defined as follows:

$$\begin{aligned} EMA(n)[x]_t &= \alpha * x_t + (1-\alpha) * EMA(n-1)[x]_{t-1} \\ &= \sum_{k\geq 0}^{n} \alpha * (1-\alpha)^{k} x_{t-k} \end{aligned} \quad (1)$$

Where the factor $n > 0$ and the factor α is often taken to be $\alpha_n = 2/(n+1)$.

MACD: for a variable x_t, its MACD can be defined by comparing the difference of n_1 -day and n_2 -day EMA as follows:

$$MACD(n_1, n_2) = EMA(n_1) - EMA(n_2) \quad (2)$$

This value can estimate derivative of x with respect to t, namely $\Delta x / \Delta t$, hence can estimate velocity.

There exists many definitions about burst, traditional methods mainly rest on term's arrival rate, which only consider one-dimensional intuitive feature, but in our method, considering data specific characteristics on microblog platform, we both consider term's frequency and the importance of tweet which the term belonging to, hence improve the accuracy of term weight. Moreover our burst definition is based on the change rate, by which we can filter out hot features. Since it hard to measure burst from these known values directly, we refer to popular stock market trend analysis technology, on the one hand, it can solve the problem of calculating derivatives in the discrete time values and smooth the noise influence on the values of x_t, one the other hand, since EMA and MACD are both linear functions, our method can be efficient and practical, avoiding expensive computation of many existing burst model. Thus it can improve the accuracy of burst detection.

IV. MICROBLOG BURSTY FEATURE DETECTION METHOD

Based on proposed definitions and techniques in our method, Microblog bursty feature detection implementation can be described as follows:

Step 1. Separate the timeline into equal time intervals. We define the i-th considered time interval I_i as follows:

$$I_i = (t_i, t_i + l) \quad (3)$$

Where t_i represents the starting point of the i -th considered time interval, l represents the length of time interval, when $i = 0$, t_0 represents the first time point.

Step 2. Compute term weights of each tweet in each time interval. We define term weight as follows:

$$weight_t^{tw} = (tf_t^{content} + tf_t^{title} \times boost_{title}) \times weight_{tw} \quad (4)$$

Where t represents each term of all tweets, tw represents each tweet, $tf_t^{content}$ represents term t 's frequency in the content of one tweet tw, tf_t^{title} represents term t 's frequency in the title .

$$weight_{tw} = \frac{\alpha \times retweet_{tw} + (1-\alpha) \times reply_{tw}}{\beta} \times (\gamma^{time_{tw} - \theta} + 1) \quad (5)$$

Where $retweet_{tw}$ represents the retweet number of one tweet, $reply_{tw}$ represents the comments number of one tweet received, $(\gamma^{time_{tw} - \theta} + 1)$ represents the influence degree of time factor on tweet weight, $time_{tw}$ is time fading facto which represents the difference value between the time interval which a tweet belonging to and the first time interval, the factor α, β, γ, θ are all training parameters, where α represents the influence degree of $retweet_{tw}$ on tweet weight , $(1-\alpha)$ represents the influence degree of $reply_{tw}$ on tweet weight , β represents the influence degree of both $reply_{tw}$ and $reply_{tw}$ on tweet weight, γ and θ are used to adjust the influence degree of time fading factor, they can be set by experience, and $0 \leq \alpha \leq 1$, $\beta > 0$, $0 < \gamma < 1$, $\theta > 0$.

In the above formula (4), since we believe that one term in title will be more influent on bursty feature detection than that in content, we use an enhancement factor to represent this impact factor.

In the above formula (5), through analyzing microblog's characteristics, just considering current news on this platform, we select three more effective parameters: $retweet_{tw}$,

$reply_{tw}$ and $\gamma^{time_{tw}-\theta}$ to depict tweet features. Since news is real-time, there are few users to collect them, we don't take the collection number feature into account; and since as time went by, the retweet number and comments number of a tweet may increase, but this can't demonstrate this tweet is talking something bursty directly, so to eliminate the cumulative effect of time, we introduce exponential function to depict time impact.

Step 3. Compute term weights of each time interval. For each time interval considered: $I_i=(t_i,t_i+l)\ \ i=0,1,2...$, we use the following formula to compute each term weight in the corresponding time interval:

$$weight_t^{I_i}=\sum_{tw\in I_i} weight_t^{tw}\ (i=0,1,2...) \qquad (6)$$

Where t represents each term, tw represents each tweet, $weight_t^{tw}$ represents each term weight in corresponding time interval.

Step 4. Compute each term's dynamics value series in all time intervals. Now, we have got each term's weight value series in all time intervals, it means that, for a term t, its weight value series can be described as follows:

$$t(i)=[weight_t^{I_0},weight_t^{I_1}...weight_t^{I_i}...](i=0,1,2...) \qquad (7)$$

To estimate the rate of change of momentum, namely dynamics value, for each term series value $t(i)$ as above definition (7), first, refer to formula (1) in section 3, we replace x_t with $weight_t^{tw}$ in step 2; then refer to formula (2) in section 3, we can estimate the rate of change of mass*position, namely mass*velocity; next, we should estimate the derivative of the MACD by using EMA tools again, which can be calculated as follows:

$$d(n1,n2,n3)=MACD(n_1,n_2)-EMA(n_3)[MACD(n_1,n_2)] \qquad (8)$$

Where $d(n1,n2,n3)$ represents dynamics value.

At last, we compute each term's dynamic value series by above formula (8) and describe each term t's dynamics value series as follows:

$$[dynamics_t]\ =[d_{I_0},d_{I_1}...d_{I_i}...](i=0,1,2...) \qquad (9)$$

1/1 5:35-3:35		1/1 15:35-13:35		1/2 21:35-19:35		1/3 3:35-1:35		1/3 15:35-13:35	
term	dynamics value	term	dynamics value	term	dynamics value	term	dynamics value	term	dynamics value
朝鲜 Korea	2.611587	日本 Japan	15.4895539	网络 network	11.07746	网站 website	2.101768	芦苇 reed	17.24338
遗训 teachings of the deceased	2.377244	地震 earthquake	13.8442674	复生 resurge	9.902933	铁道部 Ministry of Railways	1.799442	钢筋 rebar	17.24249
金正恩 Zhengen Jin	2.233943	发生 happen	10.9493462	原配 first wife	7.445333	购票	1.739291	江苏 Jiangsu	17.15673
金正 Jinzheng	1.927126	东京 Tokyo	10.8981796	死亡 death	7.3603	带宽 bandwidth	1.731556	车祸 traffic accident	13.50716
失去 lose	1.655207	震感 tremble	9.71591111	人大代表 deputies to the National People's Congress	6.115199	登录 log on	1.714864	豆腐渣 jerry-built	12.95468
获得 obtain	1.64742	地区 area	7.44027184	涉嫌 suspect	5.692548	增加 enhance	1.680321	工程 project	12.93928
司令官 commandant	1.643491	关东 northeast China	5.1741037	广东省 Guangdong Province	5.667793	改善 improve	1.090792	筑堤 embankment	12.85072
收藏 collection	1.157763	消息 information	3.75939075	中毒 poisoning	5.068813	旅客 passenger	0.955699	空姐 airline stewardess	8.864617
会议 conference	1.042004	深度 depth	3.61923857	干部 cadre	4.255933	巴马 bama	0.781629	接生 accouche	7.053085
广东 Guangdong	0.901436	气象 weather	3.19573333	曝光 exposal	3.360711	总统 pres.	0.725421	孕妇 gravida	7.033795
事业 enterprise	0.897384	海啸 tsunami	3.14096163	遗书 posthumous papers	2.936533	公安部 Ministry of Public Security	0.663384	飞机 airplane	5.269602
金正日 ZhengriJin	0.878811	震源 hypocenter	1.94168889	黄光 Guang Huan	2.8672	增长 increase	0.62689	应该 should	5.197304
禽流感 avian flu	0.869287	孩子 children	1.93527549	龙利源 Liyuan Long	2.750222	美国 USA	0.616132	男子 man	4.819058

Figure 1. Partial terms and dynamics values of bursty features on Microblog

Step 5. Select bursty features in each time interval. To get bursty features, set a threshold τ_i for each time interval, bursty features set of each time interval can be calculated as follows:

$$\{bursts\}_{I_i}=\{t\,|\,d_{I_i}^t>\tau_i\}(i=0,1,2...) \qquad (10)$$

Where $d_{I_i}^t$ represents a term t 's dynamics value in time interval I_i, then we can get bursty features sets of all time intervals.

V. EXPERIMENTS AND EVALUATIONS

We have constructed several experiments to analyze and verify the preciseness and efficiency of our method. The experiments' data contains about 15 representative Sina micoblog users' tweets of four days from December 31, 2011 to January 3, 2012, whose total number is about 2000. Because we focus on news bursty topic detection, we chose users who write new tweets frequently and have numerous followers. In the first phrase of experiment, we get 9756 segmented terms of each tweet in total by using ICTCLAS. Then, we compute each term's weight by the formula of section 3. According to experience, we set $boost_{title}$, α, β, γ, θ to 2, 0.35, 100, 0.5, 1, respectively. $time_{tw}$ is computed by hour. In the second phrase of experiment, we set the length of time interval l to two hour and compute each term's weight in each time interval. In the last phrase, we set n_1, n_2, n_3 to 4, 8, 5 to compute each term's dynamics value series based on MACD, and simply choose the top 13 terms to be bursty feature as our threshold standard in each time interval. The total running time is 59.8414004447s .Due to the space limitation, we only choose the top 13 terms of bursty features by dynamics value in each time interval and only five time interval's experiments' result is shown in figure 1.

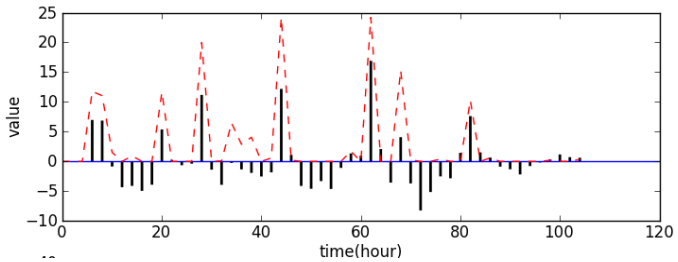

Figure 2. Comparison between frequency and dynamics value of "金正恩 (Zhengen Jin)"

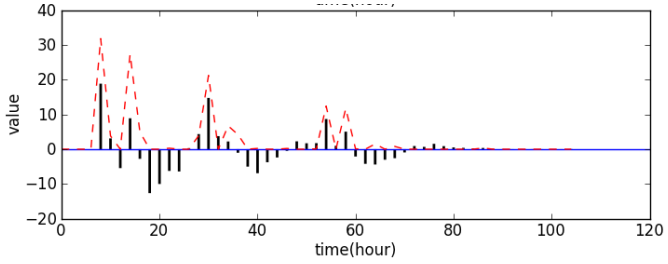

Figure 3. Comparison between frequency and dynamics value of "禽流感 (avian flu)"

From the figure, we can see that the term's dynamics value represents the bursty degree and one term may have different dynamics value in different time interval. In intuition, when a term's frequency is increasing sharply and abruptly in a time interval, it can be considered as a bursty feature and therefor it has a big dynamics value. Here we show two terms' frequency and dynamics value series in figure 2 and figure 3. From the figure, we can see that the change of the dynamics value corresponds to intuition.

To show our method's accuracy, we compare the experiment result with billboard of Baidu on January 5, 2012. In order to do this, we first simply get the tweets which contain the experiment result's bursty features. Then, we surround the news or search terms which contain the title of the selected tweets in figure 4 by red lines. Due to space limitation, we only exhibit partial result by comparing our method result and top ten recent events of billboard. We can draw that our method can detect most of the bursty features which represent the bursty topic.

Figure 4. Comparison between our method result and billboard of Baidu

TABLE I. TRANSLATION OF FIGURE 4

billbooad of baidu> events> ranking list of recent events		
ranking list of recent events		
Ranking	Key words	trend
One	First wife who was driven to death by the third wheel resurges	upward
Two	Pass off reed as rebar	upward
Three	Railway workers get a pay rise	upward
Four	First wife was driven to death by the third wheel	downward
Five	A scandal about monk	downward
Six	Present state of Korea	upward
Seven	Yan Liu's clothes are perspective	upward
Eight	Real-name system is applied in buying railway tickets	upward
Nine	Tragic events of entertainment circle in Korea	downward
Ten	Kodak is warned to be out of the market	upward

VI. CONCLUSION

We proposed an effective and accurate method to detect bursty feature in micoblog. The method considers several important factors such as retweet number, time to compute the term weight. We use physical concept momentum to detect bursty and compute each term's dynamics value based on MACD. Because the computation of MACD is linear, our method is very effective. Experiments show that our method can detect bursty features accurately. In addition, our method can filter out hot features. The bursty features detected by our

method can further be used to detect relative bursty topics. In further works, we go on to propose a bursty topic detection method based on our present work.

ACKNOWLEDGMENT

Thanks for the supports by the National Key Technology R&D Program under Award 2011BAK08B03-01, and the Opening Project of Key Lab of Information Network Security of Ministry of Public Security.

REFERENCES

[1] Chen C.C, Chen M.C and Chen M.S. LIPED: HMM-based life profiles for adaptive event detection, Knowledge discovery and Data Mining, 2005, pp. 556–561.

[2] Chen C.C, Chen Y.T and Chen M.C. An Aging Theory for Event LifeCycle Modeling, IEEE Transactions on Systems, Man and Cybernetics, 2007, 37(2), pp. 237–248.

[3] Chen T, Wang Y, Fang B, Zheng J. Detecting Lasting and Abrupt Bursts in Data Streams Using Two-Layered Wavelet Tree, Proceedings of Interuational Conference on Interuet and Web Applications and Services/Advanced International Conference on Telecommunications, 2006, pp. 46–52.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, pp. 993–1022.

[5] Gabriel Pui, Cheong F. Parameter free bursty events detection in text streams. Proceedings of the 31st International Conference on Very Large Data Bases, 2005, pp. 307–320.

[6] He Dan, Parker D. Stott. Topic Dynamics:An Alternative Model of "Burst" in Streams of Topics. Knowledge discovery and Data Mining. 2010.

[7] He T. , Qu G. , Li S. and et al. Semi-automatic Hot Event Detection, Proceeding of 2ed International Conference on Advanced Data Mining and Applications, 2006, pp. 1008–1016.

[8] John Murphy. Technical Analysis of the Financial Markets. Prentice Hall. 1999.

[9] Kim S. ,Reddy,N A. L. Real-time detection and containment of network attacks using QOS regulation. Proceeding of Internatinal Conference on Communications, 2005, pp. 56–68.

[10] Zhu Y. and Shasha D. Efficient elastic burst detection in data streams. Proceeding of Special Interest Group on Knowledge Discovery and Data Mining, 2003, pp. 276–284.

[11] Kleinberg J. Bursty and hierarchical structure in streams . 8th International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, 2002, pp. 97–110.

[12] Qin Shouke, Qian Weining. Fractal-Based Algorithms for Burst Detection over Data Streams. Journal of Software, 2006, 17(9), pp. 1969–1979.

[13] Qin S, Qian W, Zhou A. Adaptively Detecting Aggregation Bursts in Data Streams. Database Systems for Advanced Applications, 2005, pp. 90–101.

[14] Wang M, Madhyastha T. M, Chan N. H. Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic. 18th International Conference on Data Engineering, 2002, pp. 12–22.