Event time-space Analysis Based on Multi-Factors of Network Group Information

Yang Fang[1], Lingyu Xu [1], Jie Yu[1], Lei Wang [1], Yunlan Xue [1], Yang Liu [2]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

[2]School of Computer Science and Technology, University of South, Hunan, 421001 China

saberfy@163.com

Keywords: Network group information, Time series, activity, Influence degree.

Abstract. Network group information refers to views expressed by the masses of netizens through the network carrier, for the social hot spots and important events within a certain time frame, a network domain. The number change of group information can reflect a hot issue, and from the content of group information, we can find event information. In this article, based on time series of group information activity, we find abnormal activity time through cluster algorithm. We also detect events through group information in abnormal activity time. Finally, from the perspective of time and space, we analyse influence degree of the event to the related fields.

Introduction

Network group information is actually a group behavior of netizens in the network society. Xia xueluan[1]points out that is generally around the Internet social hot issues, and event. Therefore, from the network group information, we can discover the event. Event detection often relies on the method of clustering[2].Single-pass clustering algorithm[3],k-means clustering algorithm for the discovery of the hot topics[4]. In addition, Zen yilin[5] ‘s word segmentation method. Above the hot topic algorithm based on content, it can't describe the development trend of topic, and it is difficult to meet the needs of the event analysis, so you need find the rules of the development of the event from time series. Time series is the collected data at different time points according to the time sequence. In this article, we adopt clustering through time series of post information activity, and find abnormal time point about time series, then find events with the content of posts.

Time series data clustering method in general can be divided into three kinds: cluster based on original data, based on the features[6,7], based on model[8].

Post bar[9]is the important platform of gathering network group information. In this article we choose post bar information as the network group information.

Event analysis based on multi-factors of network group information

Related Definition.

Event: There are three attributes of the event (Subject, Space, Time)

Subject: the event subject word, it is the hot topic found in the group information of network. Space: the space domain; Areas are affected by the event, including each individual; Time: the time domain; Each individual’ time from abnormal information related to the event to recovery time; An event may affect multiple individuals, so the time domain is that of each affected individual.

Activity: Activity is denoted by the law of development of post reflected by the post related parameters in a certain time interval.

Activity time series: Activity time series is an ordered set of elements made up of record values and time, the active time series is:

$$X =< x_1 = (V_1, t_1), x_2 = (V_2, t_2)..., x_i = (V_i, t_i) >$$

$V_i$is the record value of the activity parameters, t_i’s unit is day.

The choice of active parameters: Fig1 is post amounts contrast figure about three kinds of wine at the same time from two well-known site stocks during the “Jiugui Plasticizer” event. Under the background of this event, in different sites, the changes of post amounts are basically consistent with another, and in a certain period of time, post amounts have obvious changes. So We use post amounts as one of the activity parameter; At the same time post clicks also reflects the concern of post, we use clicks as another activity parameter. So Vi = (Ci, Di), Ci is amounts of related posts in ti, Di is total clicks of related post in ti.

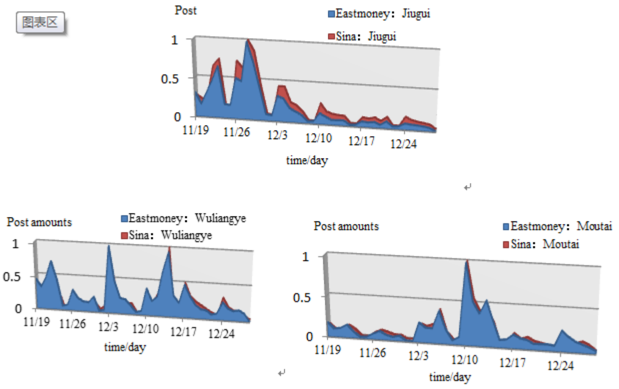


Fig 1: Jiugu, Wuliangye, Moutai post amounts in Eastmoney and Sina

Abnormal time point of activity based on clustering.

The occurrence of the event, will lead to the amount of posts increased, and we modify time sequence and take the variation of post amounts between two days:

∇X=< x_2 =(∇V_2, t_1),x_2 =(∇V_3, t_2),…， xi=（∇Vi, ti）>. ∇Vi=(∇Ci, Di),∇Ci=Ci-Ci-1;Then we take out the parameters which its variation is larger than zero, we use the them as the object of clustering. The last time sequence:

∇X={< x_2 =(∇V_2, t_1),x_2 =(∇V_3, t_2),…xi=（∇Vi, ti）>|∇vi>0};

∇X time series use hierarchical clustering method based on the center of gravity. similarity measurement use Euclidean distance.

(1)initialize the each parameter of activity time series as a cluster;(2)calculate similarity between every two clusters, choose two clusters which their similarity value is the biggest;(3)If the similarity value is greater than the given threshold, we combine these two clusters as a new cluster, and go to the second step; If the similarity value is less than the given threshold, the clustering will end.

Our clustering results will be divided into two clusters, exception cluster, or not exception cluster. we can find out those abnormal activity days, T (t1, t2..., ti);

Event detection based on the activity abnormal time point.

We mine from post contents in abnormal days, through 2.2,combining anomaly day and post contents, then we can draw event detection in abnormal day.

All the posts of every day as a text to text vectorization, each text vector Vi = (ti, x1, w (x1), x2, w (x2);...... xn, w (xn)) ; ti is time of the posts, xn as the keyword of the post , w (xn) as the weight of the keyword. The weight according to the TFIDF, TFIDF formula is as follows:

$$\mathrm{W_n} = \mathrm{TF_n}(\mathrm{x}, \mathrm{d})\log(\frac{\mathrm{N}}{\mathrm{DF(x)}} + 0.01). \qquad (1)$$

Where Wn is the weight of the nth keyword; TFn(x, d) is the word x frequency in the document d. N is documents sum; DF (x) is the document numbers containing the word x.

We select two or more than two word size nouns, and enrich the original word library to add related words, and calculate each keyword proportion to reduce the number of the keywords, and maximally retain the key information of the post.

$$\frac{\mathrm{W_i}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{W_i}} > \delta 1 \qquad (2)$$

W_i is the keyword weight, δ1is the threshold, we choose the keywords as the subject of every day, then we can build a list of subject words.

We get the activity abnormal time points in 2.1, so we can find the subject words of these time points in the list of subjects.$\mathrm{S} = (s_{t1}, s_{t2}, \dots . s_{tn})$;

Whether the subject is the event subject depending on the heat of the subject,$F_{w\mathrm{i}}$ is reported frequency of xi, or heat.

$$F_{stn} = \frac{N_{stn}}{N} > \delta 2 \tag{3}$$

$N_{s_{tn}}$is the reported number of posts of that day containing the subject word, N is number of all posts on that day. $\delta 2$ is the heat threshold.

Determination of event time domain.

We can find out the event and identify the start abnormal time about the event in 2.2. From the post, we can discover three kinds of changes of post activity: Abnormal activity on the first day the event occurs, the day the event break out, and the day the event is out of sight. So, Starting from posts' activity containing the subject word, we calculate activity relative variation of daily posts.

$$E_i = \frac{V_i}{V_1} \tag{4}$$

V_i is the ith day activity of the post containing the subject word ,$V_1$is the activity of the post containing the subject word occurring on the first day.

We cluster with their time series to divide the time series into 3 clusters. We use K means clustering method based on activity, the K is 3, and the post activity relative variation should be normalized.

$$N(E_i) = \frac{E_i - E_{\min}}{E_{\max} - E_{\min}} \tag{5}$$

Adopting the initial clustering center, $\mathrm{N}(E_1), \mathrm{N}(E_i)_{max}$, $\mathrm{N}(E_i)_{min}$, we can effectively avoid the shortcoming of K - means, make the distribution of the initial clustering center reflect the actual distribution of data as far as possible.

Event time domain judgment method: The activity relative variation will be divided into three clusters, C_s is the event happened cluster, C_b is the event outbreak cluster, C_e is the event end cluster. $C_s \cup C_b$ is the event time domain, in addition, In order to ensure continuity in time domain, the individual days can be joined the event time domain.

Determination of event space domain.

To collect group information of other post bar in the same field, using 2.2, 2.3 methods. We find out the common activity abnormal points, and analyse these posts content again, find out the same subject to find individuals affected by the event, finally we can determine event spatial domain.

Experiments

We use the web crawler to download 2012 "Jiugui Liquor" stock posts on Oriental fortune as experimental data, a total of 39718 posts. Through above algorithms, we get post topics list every day of "Jiugui Liquor" in 2012.We can conclude the year activity abnormal days on table 1 and subject list on table 2

Table 1: Activity abnormal days in 2012

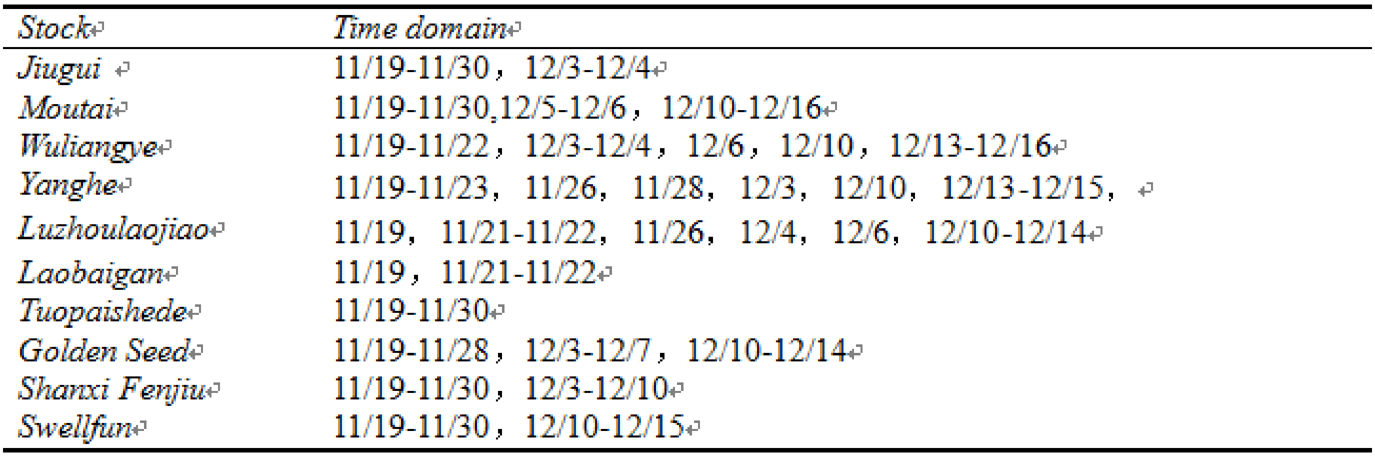
Month	*Abnormal day*	*Month*	*Abnormal day*
Jan	1/13	*Jul*	7/2
Feb	2/27	*Aug*	*8/9*
Mar	3/1	*Sep*	*9/3*
Apr	*4/5*	*Oct*	*10/6*
May	*5/21*	*Nov*	*11/19，11/26*
Jun	*6/20*	*Dec*	12/3

Table 2: Subject list

date	*Subject*	*amounts(day)*	*amounts(subject)*	*heat*
1/13	*Gold medal*	55	2	0.0364
2/27	*Yue zhibin*	73	1	0.0137
3/1	*Cost*	106	8	0.0755
4/5	*Xinxiang*	57	2	0.0351
5/21	*Sales office*	86	3	0.0349
6/20	*Prince*	143	2	0.0139
7/2	*Huafeng*	307	5	0.0163
8/9	*Sales office*	181	4	0.0221
9/3	*Race*	63	1	0.0159
11/19	*Plasticizer*	970	184	0.1897
11/26	*Plasticizer*	1674	214	0.1278
12/3	*Plasticizer*	1039	105	0.1011

Time and space domain of the event: From the keywords we can find that "plasticizer" has the most heat value, and it occurs on a number of days at most, plasticizer can be determined for the event subject, this article we are in view of the “Jiugui Liquor” plasticizing event, We select other 9 kinds of liquor stock in the same field to determine the abnormal time point, whether these stocks are affected by the influence of "plasticizer" event, and determine time-space domain Aim at the plasticizer event, from November 19 to the end of December.

Table3: time-space domain



Stock	*Time domain*
Jiugui	11/19-11/30，12/3-12/4
Moutai	11/19-11/30,12/5-12/6，12/10-12/16
Wuliangye	11/19-11/22，12/3-12/4，12/6，12/10，12/13-12/16
Yanghe	11/19-11/23，11/26，11/28，12/3，12/10，12/13-12/15，
Luzhoulaojiao	11/19，11/21-11/22，11/26，12/4，12/6，12/10-12/14
Laobaigan	11/19，11/21-11/22
Tuopaishede	11/19-11/30
Golden Seed	11/19-11/28，12/3-12/7，12/10-12/14
Shanxi Fenjiu	11/19-11/30，12/3-12/10
Swellfun	11/19-11/30，12/10-12/15

From the table 3, we can find liquor field are all affected by the plasticizer event, and in the time domain, each wine has a common time period, the starting point of time is 11/19, and the end points of time domain is different, also suggests that the effect depths of plasticizer event for each wine are different.

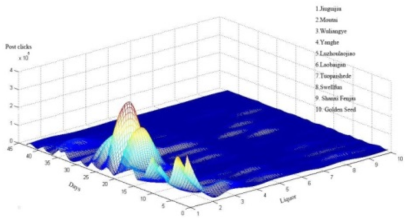


Fig2:Post amounts

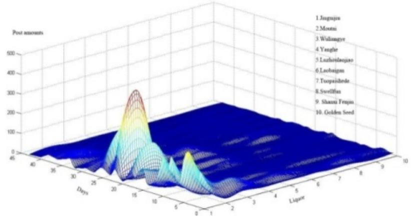


Fig3:Post clicks

Above two figures are the post amounts and clicks of 10 kinds of liquor related to plasticizer ,we can find two figure shapes are similar, and trend is consistent with the time domain. We can also conclude that the greatest influence of alcohol by the plasticizer event are “Jiugui”, “Wuliangye”, “Moutai”, although there are other liquor certainly influenced, their degree is not bigger than the 3 kinds.

Conclusion

In this article, we analyze multi-factor of network group information, we discover the impact of events on the time and space. This article mainly aims at the user's network behavior, to find the events, and events related influence degree for the related network community, which is different from the past articles adopting news reports, and emphasizing media attention to the events. So the next further work, we need combine news corpus, and the related financial data of stocks to event classification, and analyze the influence of events in the network, in reality.by setting up event model, we can estimate the event development trend ,which will have very important practical significance.

Acknowledgements

This work is supported by The Ocean Public Welfare Project of the Ministry of Science and Technology (No.201105033) and The National Natural Science Foundation of China. (No.40976108)

References

[1] Xia Xue-Luan. Constructionof Cybersociology. Journal of Peking University (Philosophy and Social Science),2004,(1)

[2] Yin Feng-Jing, Incremental algorithm for clustering texts in internet-oriented topic detection. Application Research of Computers, 2011, (1)

[3] Zhang Xiao-Yan, Research of Technologies on Topic Detection and Tracking. Journal of Frontiers of Computer Science and Technology, 2009, (4)

[4] Lei Zhen, Wu Ling-Da. Research on Event-based News Story Analysis Technology. National University of Defense Technology.

[5] Zeng Yi-ling, Xu Hong-Bo.Research on Internet hotspot information detection. Journal on Communications, 2007, (12)

[6] Zhang Hui, HoTu-Bao, Lin Mao-Song. A non-parametric wavelet feature extractor for time-series classification[C]. Berlin: Springer, 2004: 595-603.

[7] Wang X, Smith K, Hyndman R. Characteristic-Based Clustering for Time Series Data[J].Data Mining and Knowledge Discovery, 2006, 13(3):335-364.

[8] VA ITHYANATHAN S, DOM B. Model-based hierarchical clustering[C] . Stanford, California: Morgan Kaufmann, 2000: 5992608.

[9] Chang Li. Interpretation of the Communication Model of Baidu Post Bar. Press Circles, PKU CSSCI, 2007, (5).