TELKOMNIKA, Vol. 11, No. 6, June 2013, pp. 3213 ~ 3219
e-ISSN: 2087-278X

Bursty Hot-Words Detection for Campus BBS

Geng Changxin*, Zhu Xiaoguang, Nie Peiyao , Lin Peiguang
School of Computer Science & Technology,Shandong University of Finance and Economics, Jinan 250014,China
*Corresponding author, e-mail: g_changxin@163.com

Abstract

In the monitoring of campus public opinions, hot words often reflect the latest burst hot topics within a certain period of time. Therefore, this paper takes in-depth research for bursty hot words detection. In the process of words weight calculation, we consider not only traditional features such as TF, IDF, but also the burstiness, part of speech, length, location in text and other factors. Consequently, the measurement formula of burstiness and the weight calculating formula based on symphysic multi-features are proposed. The weight calculating formula can identify the bursty hot-words quickly and accurately, and then discover the bursty events, finally realizing the early warning of campus public opinions effectively.

Keywords*: hot-words, bursty, weight*

1. Introduction

With the rapid development of Internet, people tend to express real thought on the Internet. The Internet is becoming the primary place for generation and dissemination of public opinion gradually, which plays an increasingly important role in the social life [1]. Today, the user groups of Internet are increasing gradually, and it exceeded 500 million at the end of December 2011, which has reached 513 million. Students are the largest groups among the Internet users, accounting for 30.2% [2]. Such a large number of university students have sensitive reaction to a lot of social phenomena, reality and issues. They like ideological communication to each other via BBS, blog, micro blog and other information platform, and disseminate public spotlight, hot issues and major issues which include international, domestic and campus by keeping abreast, posting, comments and other methods of communication.

The college students have special groupment and strong expression desire to media. Those characteristics make hot events spread forward by high speed, and university campus easily becomes the energy disperser of negative public opinions and gathering place of them. Therefore, the following research topics are very important in college management:

- Standardized management and the monitoring of network public opinion on university campus.
- Identify bursty hot words from the huge, messy and disordered college network information quickly and accurately, and then find hot topic, especially the latest bursty hot topics.
- Control the trend of hot topics, and correctly guide the college network public opinion towards a healthy direction, and thereby reduce the negative impact of the network.

Certain terms may emerge transaction over time due to appearance of sudden hot events, namely the emergence of hot words. At present, a lot of in-depth studies have processed in many aspects, such as hot words discovery, hot words analysis.

Zheng Kui et al. proposed an automatic discovery method of hot information on network public opinion of based on ICTCLAS segmentation technology. This method can read news text and process word frequency statistics after segmentation, remove stop words from word frequency table, merge multi-unit keywords to obtain keywords list of hot information of bursty events , achieve timely retrieval for network information, and then provide technical support for emergency decision of bursty event [3]. Xue Feng et al. proposed a dynamic text model-dynamic bursty vector space model, which can describe the dynamic attributes of text efficiently. Meanwhile, a method of online detect and track is proposed to combine with text clustering method [4]. A public opinion analysis system, with a kind of high efficiency improved LC frequent pattern mining algorithm data flow analysis is deigned by Chen Lizhang to analyze the

Received January 13, 2013; Revised April 5, 2013; Accepted April 20, 2013

hot spots. This system based on bypass mode of data flow distribution according to the visitors' access to forum. The content of posts on the forum is clustered by the reductive forum theme through incremental hierarchical clustering algorithm [5]. Wang Tai et al. presented a method to capture popular search words. By deploying an optical splitter on the Internet portal of the district, popular search words are extracted from the session content which assembled by data packets and filtered from the optical splitter [6]. Li Yuqin et al. processed deep research for hot-word discovering and associating technique. In the phase of word discovering, they utilize named entity recognition techniques and statistical techniques for high frequency phrase to process phrase string excavation, then take the basis of weight and weight fluctuations to compute hot-word weight. In the hot-word association period, hot words are divided based on the difference of the weight value of them, and hot-word relationship was computed from the principle of co-occurrence rate [7].

Hot word is a web vocabulary phenomenon, reflecting the widespread concerned problem within a particular coverage, such as name, place, organization and other common phrases. A lot of network hot words are new words which have not been included in the dictionary. Hot words usually have characteristics of frequent occurrence, wide distribution and sudden transaction may occur over time. In Figure 1, the word “word1” appearing with high frequency but little fluctuation, belongs to high-frequency words; the word “words2” suddenly emerges transaction growth in the Tk moment, showing a rapid growth trend, so it belongs to hot words; the word “word3” appearing with low frequency and small fluctuation changes, belongs to low-frequency words.

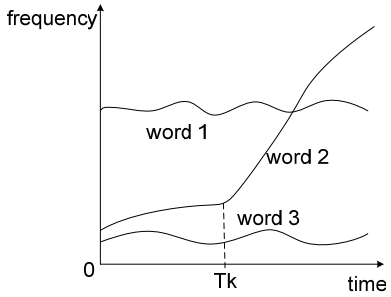

Figure 1. The Presenting Features of Hot Words

Therefore, directing at the features of hot words, the steps for their extraction are presented. Meanwhile, we develop the measurement method of burstiness and weight calculation formula which is based on the integration of multiple features; finally, a text expression method with dynamic space model is developed by use of this weight calculation formula.

2. Hot Words Analysis

The emergence of hot events will lead to sudden transaction of certain words over time. Based on this feature, a zero-copy based network packet capture platform [5, 8] combined with bypass listening way is used to capture information in colleges’ forum in real time [9, 10]. By analyzing the fluctuation exception of words contained in crawled posts, hot words can be identified, and then we can discover bursty hot topics. The main steps are as follows:

- Word segmentation and frequency statistics for captured text information.
- According to relevant rules, we achieve words filter by removing words and a number of meaningless words and symbols etc.
- Measure burstiness of words to identify hot words.
- Initialize weight for respective hot words by the use of weight calculation method based on symphysic multi-features and then sort hot words by weight. Select a certain number of hotkeys in sort list to constitute the hot keywords library.

2.1. Word Segmentation Process

A word is the smallest constituent units of a document, so that lexical analysis is the foundation and key step in information processing. In this paper, ICTCLAS developed by Institute of Computing Technology, the Chinese Academy of Sciences is used for word segmentation. This system bases on the cascading type of Hidden Markov Model, whose key functions include Chinese word segmentation, part of speech tagging, named entity recognition, the identification of new words and supporting the user dictionary. In performance, this system presents higher word segmentation accuracy and efficiency [11]. In segmentation process, location of word should be marked and single word should be removed, and at the same time, the word composed of two or more nouns should be counted. The word which exceeds a certain threshold should be added to keywords list in form of noun. e.g. “Wenchuan earthquake” are composed by two nouns, “Wenchuan” and “earthquake”. The word is added to the keywords list when the number of times exceeds the predefined threshold. After word segmentation, a list of keywords is generated with part of speech attribute. Generally, the hot words have a higher word frequency, therefore, a lower threshold needs to be set, and the word whose frequency is lower than this threshold will be removed from the list of keywords.

2.2. Words Filter

Keywords list contains a lot of words after word segmentation. Most of the words contribute little to the post, so the corresponding filter rules [12, 13, 14] need to be developed to filter keywords list.

1. Part of Speech Filter

Different parts of speech play different roles in text presentation. The semantics of a sentence primarily is expressed by nouns and verbs. Although prepositions, conjunctions and adverbs etc. have higher frequency of occurrence in the document, they have no real meanings, such as “of, the, in, thought, but” etc. For this reason, the words which have part of speech like preposition, auxiliary word etc. should be abandoned and retain only nouns and verbs.

2. Stop Words Filter

Meaningless words, punctuation, numbers and special symbols may occur in posts, such as “Ho, Aha, #, [, (” etc. These words and symbols can be added into stop words list. In this paper, the list combines and extends the stop word list of Harbin Institute of Technology.

3. Similar Words Filter

Different expression forms of words with similar or identical implications may exist in keywords list. It can be processed by statistics method to merge synonyms, such as “neglect” and “ignore” etc. By judging the similarity and word frequency, longer word in the case of same quite frequency can be retained, such as “Influenza A”, “Influenza A HINI” and “Influenza A HINI flu” etc.

4. Rule Filter

Rule filter generally is used for filtering useless string with obviously pattern, such as collection of numeral and quantifier with high frequency, common meaningless prefix and suffix.

5. Background Noise Filter

The background noise is subject unrelated and meaningless string which cannot be filtered by stop words and rule filter, such as “Beijing daily news”, “one of those”, “at last” etc. The noise has huge system and chaotic state, unable to get through manual sorting. Therefore, it needs to collect corpus as a training set to program for the extraction of background noise library.

2.3. Words Burstiness Measurement

The hot words have burstiness, like abnormal growth of word frequency within a certain period. For this reason, the word frequency distribution within a certain time period can be analyzed to determine whether the word has burstiness.

The distribution of word frequency is treated as a dimensional function f(x) in Figure 2, in which x-axis represents time, y-axis represents word frequency of a moment. Now intercept

the word frequency distribution within the Ti-Tj time period and assume that the word frequency of “word t” is Pi in Ti moment, Pj in Tj moment.

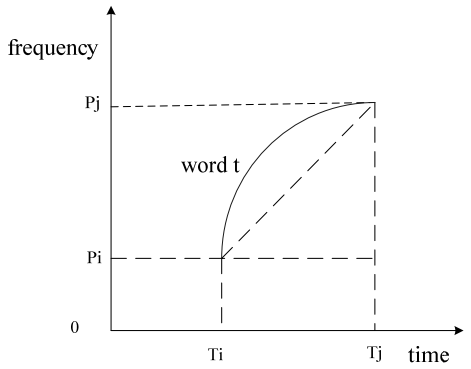

Figure 2. Word’s Frequency Distribution

Considering the distribution of word frequency may influence burstiness in a certain period time, cosine theorem can be used to measure the burstiness of word t in time Ti-Tj, as following formula:

$$B(t) = 1 - \frac{T_j - T_i}{\sqrt{(T_j - T_i)^2 + (p_b - p_a)^2}} \quad (1)$$

In formula (1), word burstiness $B(t)$ is a value between 0 to 1. The greater value it has, the greater word burstiness it will show, conversely, the smaller the burstiness is; $T_j - T_i$ is the difference between the two time periods, which can be hours, days, etc. and they can be selected according to the actual situation; $p_b - p_a$ is the difference of word frequency between time T_i and T_j.

Hot words will drift over time, i.e. a topic discuss transferred to another topic in later, i.e. the drift of topic. The current word frequency of topic which has drifted is less than the word frequency before drift. This paper focuses on the extraction of word from current bursty events, so the drifted topic is outside our consideration scope. Therefore, certain words which later frequency is less than former frequency will be removed from keywords list. So a lower threshold can be set to select words whose burstiness is higher than this lower threshold as candidate hot words list. The high-frequency words are usually more evenly distributed, so this method can remove high-frequency words within the process of bursty words selection.

2.4. Weight Calculating Method based on Symphysic Multi-features

For candidate hot words, each word has useful information, such as TF, IDF, burstiness, part of speech, location, length etc [10]. Therefore, the following factors need to be considered for hot word weight generation.

1. **TF**: known as term frequency, representing the frequency of word occurrence. The greater TF is, the higher concern degree the word has.
2. **IDF**: known as inverse document frequency, is a measure of whether the term is common or rare across all documents. Greater IDF shows greater discrimination of words and more relevant to subject.
3. **Burstiness**: hot words have the characteristics of abnormal growth in a short time, so the burstiness needs to be introduced to measure the growth of words in the period of time.
4. **POS**: known as part of speech. The named entities in post information such as names, place names and institutional names etc. contribute much more than non-named entities to the distinction of topics, so the weight growth is required for named entities, while verb takes a second place.

5. **Location**: the words in different locations make different contribution to entire post: Words in the post title make the greatest contribution; the first and last paragraph in body text also has greater contribution; the first and last sentences of each paragraph in body text also have contribution; follows are the words in reply. Therefore, the word in different locations has different weight.

6. **Length**: the longer the words are, the more information they carry.

Considering the above factors, the weight calculation formula is constructed as follows:

$$w(t,d) = a * \frac{\partial * TF(t,d) * \log_2(\frac{N}{DF(t)} + L) * weight(POS(t))}{\sqrt{\sum_{t \in D} (\partial * TF(t,d) * \log_2(\frac{N}{DF(t)} + L) * weight(POS(t)))^2}} + B(t) + b * \frac{length(t)}{avglen} * weight(position(t,d)) \quad (2)$$

In formula(2), a and b are adjustment coefficient, 0<a,b <1, a+b=1; $w(t,d)$ denotes the weight of word t in post d; D is whole posts; ∂ is adjustment coefficient of location weight; $TF(t,d)$ denotes occurrence frequency of word t in post d; $DF(t)$ denotes the number of posts which includes word t; L is experience constant; $weight(POS(t))$ is POS weight of word t, generally initializes named entities as 2, verb as 1.5; $length(t)$ denotes length of word t; $avglen$ denotes the average length of keywords; $weight(position(t,d))$ is location weight of word t in post d. $B(t)$ is the weight of burstiness factor, the calculation formula is formula (1).

After calculating the weight of above words, a certain number of words should be chosen by descending sort of weight to construct hot words feature library K_i within a certain period of time.

3. Results and Discussion

The experiment data is collected from the entrance of College network. Every day the information of college forum is captured twice through entrance, and continuous collection goes on for thirty days, and then we have thirty days historical data. For each collected documents, firstly word segmentation and word frequency statistics are processed, and carry on filter according to the rules mentioned in chapter 2.2, then we will get a candidate hot keywords list.

3.1. Results and Discussion – Burstiness Measurement

Because of too many words in keywords list, this paper only takes two words “The Olympic Games” and “Jun Zhou” as samples to measure the word burstiness within five days. Normalization is processed for convenience: word frequency is normalized to 0-100, time is normalized to 0-100 as well. Figure 3 is normalized word frequency distribution.

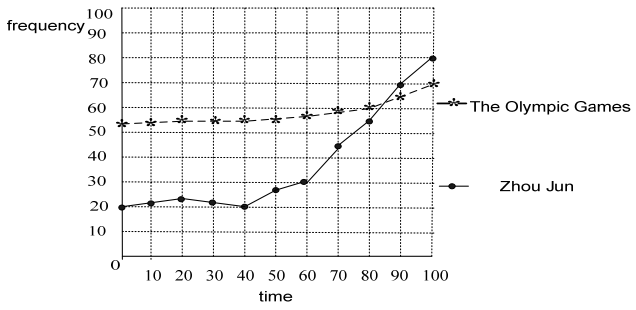

Figure 3. Word’s Frequency Distribution after Normalization

Now calculate the burstiness of these two words within the moment 40-90 by formula (1):

The burstiness of word “The Olympic Games”: $= 1 - \frac{T_j - T_i}{\sqrt{(T_j - T_i)^2 + (p_b - p_a)^2}} = 0.019419$

The burstiness of word “Jun ZHOU”: $= 1 - \frac{T_j - T_i}{\sqrt{(T_j - T_i)^2 + (p_b - p_a)^2}} = 0.292893$

In Figure 3, the burstiness of word “Zhou Jun” obviously is greater than word “The Olympic Games”, which is consistent with our experimental results. Therefore, the burstiness measurement formula proposed in this article is in line with actual requirement.

3.2. Results and Discussion - The Weight Calculating Formula based on Symphysic Multi-features

For the process of candidate hot words list with burstiness measurement, this article firstly calculates the weight of hot words by the weight calculation formula of traditional TF-IDF function and symphysic multi-features respectively, then process extraction experiments for background corpus based on the calculation results. Limited by the size of background corpus, we can’t verify the effect of keywords extraction for each document. Therefore, 500 documents are extracted randomly, and through program verification, the extraction effect of two methods for hot keywords are analyzed and compared. Compare results are shown in Figure 4.

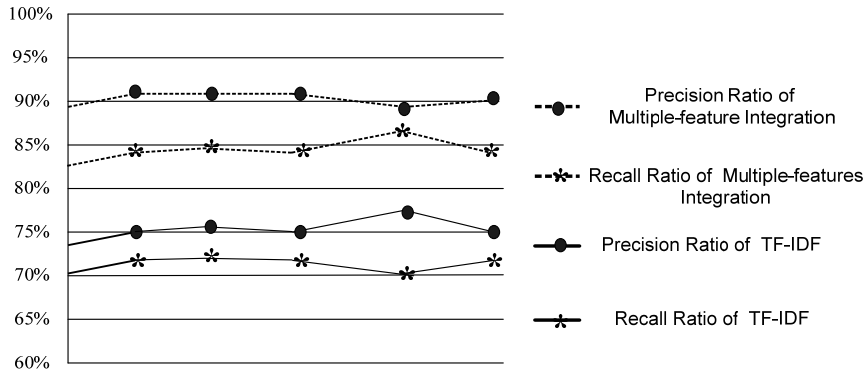

Figure 4. Effect Comparison of Two Different Extraction Methods

With the use of the traditional TF-IDF method to process extraction, average precision ratio is 76.9%, average recall ratio 72.3%; with the use of the weight calculating method based on symphysic multi-features to process extraction, average precision ratio is 90.7%, average recall ratio 85.7%. With the use of the weight calculation method based on symphysic multi-features, precision ratio increased about 13.8% above the traditional TF-IDF method, and recall ratio increased about 13.4%. From the experiment results, conclusion comes to that the weight calculation method based on symphysic multi-features is obviously better than traditional TF-IDF method.

4. Conclusion

With the rapid expansion scale of the Internet, the discovery and tracking of college network bursty hot topics has become important means for regulation of network public opinion. Identifying hot words quickly and accurately is the premise of hot events discovery. Based on the burstiness characteristic of hot words, this paper developed a method for burstiness measurement, then in the weight calculation of hot words, considering not only the higher occurrence frequency, bursty of time, but the feature information of words whose occurrence in post, such as location, POS and length etc. To research the key technologies stated above, a foundation can be established for bursty hot events, and finally achieve effective and accurate early warning for college network public opinion.

Certainly, the measurement method of burstiness proposed in this paper is not perfect. For example, in the measurement of word burstiness, only the amount of growth within a period of time is considered, without the growth rate of the words. Therefore, in practical applications, definite integration can be utilized to measure the growth rate of word frequency, which needs further study.

Acknowledgments

This work is supported by Ministry of Education, Humanities and Social Sciences Project (10YJC880076) and Shandong Province Natural Science Foundation Project (ZR2010FL008).

References

[1] CHEN Hua, LIANG Xun, RUAN Jin. Design and Implementation of correlation analysis in Cyberworld opinion. *NCIRCS'2007.* 2007; 45-49.
[2] The 29 times China Internet network development state statistic report. China Internet Network Information Center. 2012.
[3] ZHENG Kui et al. Hot Spot Information Auto-detection Method of Network Public Opinion. *Computer Engineering.* 2010; 36(3): 4-6.
[4] XUE Feng, ZHOU Yadong, GAO Feng. An Online Detection and Tracking Method for Bursty Topics. *Journal of Xi'an Jiaotong University*. 2011; 45(12): 64-69.
[5] CHEN Lizhang, LI Bin, CHEN Xiaopeng. Design and Realization of Monitoring System of Campus BBS Public Opinion*. Microprocessors*. 2012; 2(1): 40-48.
[6] WANG Tai, JIANG Guangrong, YU lixia. Capturing and analyzing popular search words in micro district. *Computer Engineering and Design*. 2012; 33(2): 556-560.
[7] LI Yuqin, SUN Lihua. Hot-Word Detection for Internet Public Sentiment. *Journal of Chinese Information Processing*. 2011; 25(1): 48-59.
[8] WANG Meng, LI Bin, SUN Chunqi. Research of Network Public Opinion Hotspots Detection Based on Frequent Items Mining. *Microcomputer information.* 2010; 26(12-3): 35-38.
[9] Nyoman Rizkha Emillia, Suyanto, Warih Maharani. Isolated Word Recognition Using Ergodic Hidden Markov Models and Genetic Algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(1): 129-136.
[10] Abeer El-Korany, Salma Mokhtar Khatab. Ontology-based Social Recommender System. *IAES International Journal of Artificial Intelligence*. 2012; 1(3): 127-138.
[11] LUO Huixia. The Network Public Opinion Monitoring System Research And Exploitation. Disertation. Taiyuan; North University of China; 2010.
[12] LI Hengxun. Key Technology Research on Web Forums Crawling and Hot Topic Detection. Disertation. Beijing: Capital Normal University; 2011.
[13] ZENG Yiling, XU Hongbo. Research on Internet hotspot information detection. *Journal on Communication*. 2007; 12(28): 141-146.
[14] LAN Kaimei. BBS Hot Topic Detection and Monitoring System. Disertation. Beijing: Beijing Jiaotong University; 2011.