Front. Electr. Electron. Eng. China 2009, 4(1): 20–23
DOI 10.1007/s11460-009-0002-5

RESEARCH ARTICLE

Yadong ZHOU, Xiaohong GUAN, Qindong SUN, Wei LI, Jing TAO

Approach to extracting hot topics based on network traffic content

Abstract This article presents the formal definition and description of popular topics on the Internet, analyzes the relationship between popular words and topics, and finally introduces a method that uses statistics and correlation of the popular words in traffic content and network flow characteristics as input for extracting popular topics on the Internet. Based on this, this article adapts a clustering algorithm to extract popular topics and gives formalized results. The test results show that this method has an accuracy of 16.7% in extracting popular topics on the Internet. Compared with web mining and topic detection and tracking (TDT), it can provide a more suitable data source for effective recovery of Internet public opinions.

Keywords hot topic extraction, network traffic content, Internet public opinion analysis

1 Introduction

Currently, the Internet is one of the most important platforms of communication. For the purpose of public opinion analysis, how to extract hot topics that attract the public has become a quite interesting issue.

The related studies [1–6] can be divided into two parts. One is the research on topic detection and tracking [1,2], and the other is on web mining [4]. These studies select data from web sites as the data source, and the analyses present what Internet media is concerned about instead of the users. Therefore, we need to select other data sources for acquiring true Internet public opinion.

In this article, we select network traffic as the data source, which corresponds to users' network behavior, and by this means we can get more precise analysis results. We also attempt to recover and comprehend users' behavior, analyze hot topics that users are concerned about, and finally obtain actual Internet public opinion based on the formula description of hot topics on the Internet.

The article is organized as follows: Section 2 presents the formulation of hot topics on the Internet. Section 3 proposes the algorithm of word correlation. Section 4 proposes the method of building the description of hot topics and the experimental results and discussion are presented in Sect. 5. Conclusions are given in the last section.

2 Formulation of hot topics on Internet

Although some researchers have proposed the general definition of topic [6], there is no specific definition for Internet public opinion research. Thus, we propose a definition of hot topics on the Internet below.

Definition 1 Users pay attention to a hot topic on the Internet widely and continuously, which is defined as a set of certain types of information, including the semantic description of the meaning and the diffusion mode of the topic.

We propose a multidimensional vector as the formulation of hot topics on the Internet, and set hot words, key title and the web sites as the basic elements of the vector. A hot topic P can be described as

$$P=(W_1, W_2,..., W_l, T_1, T_2,..., T_m, S_1, S_2,..., S_n), \quad (1)$$

where W_i is a hot word, which is related with the hot topic and describes its meaning; T_i is a key title, which is a word or short sentence and describes the key meaning of the

Translated from *Journal of Xi'an Jiaotong University*, 2007, 41(10): 1142-1145 [译自: 西安交通大学学报]

Yadong ZHOU (✉), Wei LI, Jing TAO
MOE Key Lab for Intelligent Networks and Network Security, State Key Lab for Manufacturing Systems, Xi'an Jiaotong University, Xi'an 710049, China
E-mail: yadongzhou@gmai.com

Xiaohong GUAN
Department of Automation, Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing 100084, China

Qindong SUN
School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

topic; and S_i is the URL of a web site, which publishes pages related with the topic.

3 Algorithm of word correlation

The content of a hot topic would appear frequently in network traffic. Hence, hot words that describe the meaning of the topic would emerge with a large frequency. A hot topic could be described by several hot words, among which relationships can be found. Thus, we propose an algorithm to calculate the value of the correlation between two hot words of the topic, which is the key step of our approach.

First, we calculate the frequency of each word appearing in the traffic content, and divide the words into high frequency words, middle frequency words and low frequency words. We calculate the frequency by

$$w = (W, f), \tag{2}$$

where w is the statistical result of a word, W is a word, and f is the total frequency of the word. We set the thresholds of high frequency, middle frequency and low frequency, and get three sets of words:

$$\begin{cases} WS_{\mathrm{h}} = (w_{\mathrm{hs}1}, w_{\mathrm{hs}2},..., w_{\mathrm{hs}r}), \\ WS_{\mathrm{m}} = (w_{\mathrm{ms}1}, w_{\mathrm{ms}2},..., w_{\mathrm{ms}s}), \\ WS_{\mathrm{l}} = (w_{\mathrm{ls}1}, w_{\mathrm{ls}2},..., w_{\mathrm{ls}t}), \end{cases} \tag{3}$$

where WS_{h}, WS_{m}, WS_{l} indicate the high frequency, middle frequency and low frequency word set respectively; $w_{\mathrm{hs}i}$, $w_{\mathrm{ms}i}$, $w_{\mathrm{ls}i}$ indicate the high frequency, middle frequency and low frequency word respectively.

A hot topic could be regarded as a set of web pages on the Internet, and all their related hot words would exist in these web pages. When users read these pages via the HTTP protocol, the content of one page is the content of one TCP connection, and the hot words exist in the content of the TCP connection. Hence, the correlation of any two hot words could be calculated by the number of times the two words exist in the same connection. The more two words exist in the same connection, the larger the value of correlation would be, and the greater the probability that the two words belong to the same hot topic.

The models of network flow include a train model [7] based on the TCP connection [8], and a flow model proposed by Claffy [9]. Based on these two models, we propose a topic flow model to recover the situation of hot topic diffusion.

Definition 2 A topic flow is a set of network packets with the same four-tuple, with the time interval between them less than a toehold, and with semantic content.

The topic flow could be described as

$$\Gamma_P = (\mathrm{id}, t, \mathrm{ip}_{\mathrm{src}}, p_{\mathrm{src}}, \mathrm{ip}_{\mathrm{dest}}, p_{\mathrm{dest}}, C, T, S), \tag{4}$$

where id is the identity No. of each flow; t indicates the arriving time of the flow; $\mathrm{ip}_{\mathrm{src}}$, $\mathrm{ip}_{\mathrm{dest}}$ indicate the source and destination host IP address respectively; p_{src}, p_{dest} indicate the source and destination host port respectively; C indicates the content type of the flow; T indicates the title of the page which is transferred by the flow; and S indicates the web site that publishes the page.

Based on the above definition of topic flow, a hot word in the traffic content can be described as

$$w = (W, f, \mathrm{id}_1, f_1, \mathrm{id}_2, f_2,..., \mathrm{id}_n, f_n), \tag{5}$$

where W indicates the word, f indicates the total frequency of the word, f_n indicates the frequency existing in the flow n, and id_n indicates the id of flow n.

The correlation of two words $\rho(w, w')$ could be calculated by the number of times the two words exist in the same flow. If the two words exist in the same flow once, the value of correlation would be 1; if the two words exist in the same flow n times, the value of correlation would be n; if the two words do not exist in any flow, the value of correlation would be 0. The calculating process for word w and w' is

$$w = (W, f, \mathrm{id}_1, f_1, \mathrm{id}_2, f_2,..., \mathrm{id}_l, f_l), \tag{6}$$

$$w' = (W', f', \mathrm{id}'_1, f'_1, \mathrm{id}'_2, f'_2,..., \mathrm{id}'_m, f'_m), \tag{7}$$

$$\text{If } \mathrm{id}_{i1} = \mathrm{id}'_{j1}, \mathrm{id}_{i2} = \mathrm{id}'_{j2},..., \mathrm{id}_{in} = \mathrm{id}'_{jn}, \text{ then}$$

$$\rho(w, w') = n, \tag{8}$$

where id_{in}, id'_{in} indicate the identity of word w and w' respectively; and $\rho(w, w')$ indicates the correlation of word w and w'. The more two words exist in the same connection, the larger the value of correlation would be, and the greater the probability that the words belong to the same hot topic.

4 Method for building description of hot topics

The correlation of two hot words $\rho(w, w')$ could indicate the probability that the two words belong to a hot topic, i.e., the larger the correlation, the shorter the distance between the words. Thus, we select the density-based spatial clustering of application with noise (DBSCAN) method to group the words into some clusters [10]. The value of correlation would be large in a cluster. These clusters correspond to the hot words in Eq. (1):

$$c = (w_1, w_2,..., w_n), \tag{9}$$

where c is the cluster which includes n hot words; w_i is a hot word which is mentioned in Eq. (5).

We can get the key title and web sites based on these clusters:

$$T_{\mathrm{C}} = (s_{T_{\mathrm{C}}},\ f_{T_{\mathrm{C}}},\ \mathrm{id}_1,\ \mathrm{id}_2,\ldots,\ \mathrm{id}_m), \tag{10}$$

$$S_{\mathrm{C}} = (s_{S_{\mathrm{C}}}, f_{S_{\mathrm{C}}},\ \mathrm{id}_1,\ \mathrm{id}_2,\ldots,\ \mathrm{id}_k), \tag{11}$$

where $s_{T_{\mathrm{C}}}$ indicates the character string of the key title, $f_{T_{\mathrm{C}}}$ indicates the total frequency of the key title; $s_{S_{\mathrm{C}}}$ indicates the URL of the web site, $f_{S_{\mathrm{C}}}$ indicates the total frequency of the web site; and id_i indicates the identity of flow which corresponds to the web site.

We select some hot key titles and hot web sites by setting thresholds, combining the hot word clusters mentioned in Eq. (9), and getting the final result of hot topics mentioned in Eq. (1).

5 Experiment results

We selected the mirror traffic data as the data source, which is supported by the network center of Xi'an Jiaotong University, and processed the traffic data offline. The data process server is an Acer Altos G530, and the CPU is a P4 Xeon 3.2 with memory of ECC 4 GB and hard disk of SCSI 320 GB. We selected C ++ to program, and the OS is Windows 2003 server.

In the experiment, we set the middle frequency threshold as 3000 and high frequency threshold as 10000. We obtained 665 high frequency words, 1047 middle frequency words and 1899 low frequency words. We selected the DBSCAN method to cluster the high frequency words, and set the domain radius ε as 500 and minority density threshold min$q$ as 5. After the process of clustering, we obtained 48 clusters with 8 semantic topics and 40 other results. In the process of building the description of hot topics, we set T_{C}=500, S_{C}=500, and obtained the description of 8 hot topics.

There are 8 hot topics in the final result, including information about enrolling new students of the university, school anniversary, the history and survey of the university, and the introduction of Chung Kong professors of the university. We selected two to represent the experimental result in Table 1.

The description of topic 1, related with the information about enrolling new students and introduction of majors, includes 81 hot words, 3 key titles and 1 web site in Table 1. The main content of topic 1 is about the courses and majors of Xi'an Jiaotong University, including the names of majors and departments, and the topic is diffused by the web site of the University. The description of topic 2, related with the history and introduction of the university, includes 35 hot words, 3 key titles and 2 web sites, as shown in Table 1 (We translate the Chinese words into English).

There are two differences between topic 1 and topic 2. First, the hot words of topic 2 are 43.2% as many as topic 1, which indicates that the content of topic 2 is more centralized. Second, topic 2 is diffused through two web sites, which indicates that topic 2 would affect people more widely than topic 1.

By our approach, we get 8 hot topics in the experiment from the traffic data, which could partly describe Internet public opinion, and support useful information for network administrators. However, we could not get a clearer description of hot topics because of the performance limit of the natural language process method. We need to analyze the final results by humans to get a clearer comprehension of Internet public opinion.

6 Conclusions

This article analyzes the features of existing network information researches, proposes the definition and formalized description of hot topics on the Internet and brings forward a method to analyze the traffic content data for hot topic extracting arithmetic. This method includes network traffic flow analysis, traffic content statistical analysis, and correlation analysis of words in the traffic content. The experimental results show that the current method can effectively access network hot topics, and the result partially reflects the real status. Generally, this method is efficient for studying the spread character of a network hot topic.

In future work, we will improve the performance of our approach, and research methods to filter noise words in the traffic content.

Table 1 Description of hot topics

P		$(W_1,W_2,\ldots,W_l)$	$(T_1,T_2,\ldots,T_m)$	$(S_1,S_2,\ \ldots,S_n)$
1	$l=81$ $m=3$ $n=1$	(biology, electric, energy, science, business, commerce, accountant, automation, computer, pathology, software, algorithm,…, economy)	(enrolling new students, intro to the majors, enrollment mark of masters in 2003)	http://www.xjtu.edu.cn
2	$l=35$ $m=3$ $n=2$	(predecessor, Shanghai, Jiaotong, Xi'an, history, industry, address of university, a hundred years, tradition, western China, creation, finance, run a school, Nanyang, higher education, internal, history of university,…, sites)	(introduction to Xi'an Jiaotong University, history of Xi'an Jiaotong university, 110th school anniversary)	(http://www.xjtu.edu.cn, http://newsxq.xjtu.edu.cn)

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 60574087), the Hi-Tech Research and Development Program of China (2007AA01Z475, 2007AA01Z480, 2007A-A01Z464), and the 111 International Collaboration Program of China.

References

1. Allan J, Carbonell J, Doddington G, Yamron J, Yang Y. Topic detection and tracking pilot study: final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kaufmann Publishers, 1998, 194–218
2. Yu M, Luo W, Xu H, Bai S. Research on hierarchical topic detection in topic detection and tracking. Journal of Computer Research and Development, 2006, 43(3): 489–495 (in Chinese)
3. Kosala R, Blockeel H. Web mining research: a survey. ACM SIGKDD Explorations Newsletter, 2000, 2(1): 1–15
4. Wang Z, Jin F, Li X, Wang G. Web data mining technique and realization. Journal of Harbin Institute of Technology, 2005, 37(10): 1403–1405 (in Chinese)
5. Li B, Yu S. Research on topic detection and tracking. Computer Engineering and Applications, 2003, 39(17): 7–10 (in Chinese)
6. Topic detection and tracking (TDT) evaluation workshop. The 2002 topic detection and tracking task definition and evaluation plan. [2006-04-20]. ftp://jaguar.ncsl.nist.gov/tdt/tdt2002/
7. Jain R, Routhier S A. Packet trains–measurements and a new model for computer network traffic. IEEE Journal on Selected Areas in Communications, 1986, 4(6): 986–995
8. Mogul J C. Observing TCP dynamics in real networks. ACM SIGCOMM Computer Communication Review, 1992, 22(4): 305–317
9. Claffy K C, Braun H W, Polyzos G C. A parameterizable methodology for Internet traffic flow profiling. IEEE Journal on Selected Areas in Communications, 1995, 13(8): 1481–1494
10. Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, USA: AAAI Press, 1996, 226–231