第41卷 第10期 西 安 交 通 大 学 学 报 Vol.41 №10
2007年10月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Oct. 2007

流量内容词语相关度的网络热点话题提取

周亚东[1,2]，孙钦东[1,2,5]，管晓宏[1,2,3,4]，李卫[1,2]，陶敬[1,2]
(1.西安交通大学智能网络与网络安全教育部重点实验室，710049，西安；2.西安交通大学机械制造系统工程国家重点实验室，710049，西安；3.清华大学自动化系，100084，北京；4. 清华大学信息科学与技术国家实验室，100084，北京；5.西安理工大学计算机科学与工程学院，710048，西安)

摘要：针对网络舆情分析的需求，给出了网络热点话题定义及其形式化描述，分析了流量内容中热点词语与热点话题的关系，提出了流量内容中热点词语的相关度计算算法.在此基础上，采用基于高密度连接区域的密度聚类方法得到热点词语簇，结合热点词语簇相关的网页标题及网站地址信息，得出网络热点话题的属性描述.实验结果表明，该算法能够有效获取当前网络中的热点话题，话题提取有效率达到16.7%，为网络热点话题传播特性研究提供了基础.与Web挖掘、话题监测与跟踪方法相比，所提算法通过选取合适的数据源，能更大程度地还原网络用户行为，从而得到了更为准确的网络信息传播状况.

关键词：网络热点话题；流量内容；网络舆情分析
中图分类号：TP393.4 **文献标识码**：A **文章编号**：0253-987X(2007)10-1142-04

Internet Popular Topics Extraction of Traffic Content Words Correlation

Zhou Yadong[1,2], Sun Qindong[1,2,5], Guan Xiaohong[1,2,3,4], Li Wei[1,2], Tao Jing[1,2]
(1. MOE Key Lab. for Intelligent Networks and Network Security, Xi′an Jiaotong University, Xi′an 710049, China; 2. State Key Lab. for Manufacturing Systems, Xi′an Jiaotong University, Xi′an 710049, China; 3. Department of Automation, Tsinghua University, Beijing 100084, China; 4. Tsinghua National Lab. for Information Science and Technology, Tsinghua University, Beijing 100084, China; 5. School of Computer Science and Engineering, Xi′an University of Technology, Xi′an 710048, China)

Abstract: Aiming at the requirements of network public feeling analysis, the formal definition and description of the popular topic on Internet is presented, the relationship between hot words and popular topics is analyzed, and finally a hotpoint words correlation computing approach for extracting popular topics on Internet is introduced in traffic contents. Based on that, DBSCAN (Density-Based Spatical Clustering of Application with Noise) clustering algorithm is adopted to extract popular topics and formalized results are given. The test results show that this method has an availability of 16.7% in extracting Internet popular topics, which, compared to web mining and TDT (Topic Detection and Tracking), can provide a more suitable data source for effective recovery of Internet public opinions.

Keywords: popular topic on Internet; network traffic content; Internet public opinion analysis

目前，互联网已经成为人们交流信息的重要渠道，网络舆情分析研究也随之受到广泛关注.网络舆情信息具有规模巨大、凌乱无序等特点，如何从中识别、分析有价值的信息已成为研究热点.

针对这一问题，国内外均开展了相关研究[1-6]，主要包括两类，一类是话题识别与追踪研究[1-2]，另一类是Web数据挖掘研究[4]，它们都以Web站点发布的各类信息为数据源，其结果反映了网络媒体对信息的呈现状况，但却无法有效反映网络用户对信息的关注状况，因此也就很难反映网络舆情的真

收稿日期：2007-01-10. 作者简介：周亚东(1982～)，男，博士生；管晓宏(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60574087)；国家高技术研究发展计划资助项目(2003AA142060).

实情况.

本文将网络流量作为数据源，这种流量数据能直接对应于用户的网络行为，能更为准确地反映网络舆情的状况. 同时，本文着重于研究还原、理解网络用户的各种行为，从中获取用户对网络信息的关注情况，并通过网络热点话题的形式化描述得到更真实的网络舆情状况.

1　网络热点话题的形式化描述

在话题识别与追踪研究中，已对一般意义下的话题进行了定义[5]，而在网络信息分析研究中却没有明确定义，为此本文对网络热点话题作如下定义.

定义 1　网络热点话题指，以网络为传播媒介，被一定人群广泛、持续关注，并能够反映网络舆论状况的信息集合，其中包括对网络热点话题内涵的语义化描述以及话题的传播方式等.

为了突出人们关注的内容，网络热点话题可被形式化地表示为一个多维向量，并以热点词语、核心标题及信息发布网站等作为基本元素. 设当前网络的一个热点话题为

$$P=(W_1,W_2,\cdots,W_l,T_1,T_2,\cdots,T_m,S_1,S_2,\cdots,S_n) \tag{1}$$

式中：W_i 表示热点词语，即与热点话题直接相关并可用以描述话题含义的词语；T_i 表示核心标题，为可概括热点话题核心意义的词语或短句；S_i 表示信息发布网站，其中包括传播话题的网络站点源名称或地址.

2　热点词语相关度计算

网络热点话题是用户广泛关注的信息，是在网络流量中大频度出现的内容，而热点词语可以直接描述热点话题，在网络流量中其必将以较大频度出现. 一个热点话题可由多个热点词语来描述，且词语之间具有一定的相关度. 基于此，本文提出一种流量内容热点词语相关度计算方法，该方法可量化热点词语之间的相关程度，量化结果可作为网络热点话题提取的中间数据.

在处理网络流量数据时，首先计算内容中各项词语的出现频度，词语按照出现频度又分为高频词语、中频词语和低频词语. 词语的统计式为

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

式中：w 表示词语的统计值；W 表示某一词语；f 表示词语的总频度. 通过设定高、中、低频度阈值，则基于频度的词语集合

$$\left.\begin{aligned} L_{\mathrm{h}}&=(w_{\mathrm{h}1},w_{\mathrm{h}2},\cdots,w_{\mathrm{h}r})\\ L_{\mathrm{m}}&=(w_{\mathrm{m}1},w_{\mathrm{m}2},\cdots,w_{\mathrm{m}s})\\ L_{\mathrm{l}}&=(w_{\mathrm{l}1},w_{\mathrm{l}2},\cdots,w_{\mathrm{l}t}) \end{aligned}\right\} \tag{3}$$

式中：L_{h}、L_{m}、L_{l} 分别表示高、中、低频度词语集合；$w_{\mathrm{h}i}$、$w_{\mathrm{m}i}$、$w_{\mathrm{s}i}$ 分别表示高、中、低频度词语.

在网络中，热点话题可视为网络文章的集合，在集合中的所有文章都参与讨论特定话题. 描述同一个热点话题的多个热点词语，必然出现在网络文章之中，那么当用户查看这些文章时，文章的内容便是构成一次网络连接的传输内容，而热点词语也会同时出现在一次网络连接之中. 因此，流量中的任意 2 个高频度词语同时出现于网络连接之中的次数，可量化衡量词语之间的相关度，即词语同时出现的次数越多，表示它们之间的相关度越大，那么用该词语描述同一热点话题的可能性就越大.

网络流模型包括数据包列车模型[7]、基于 TCP 连接的流模型[8]，而广泛应用于 Internet 的流模型是由 Claffy 提出的[9]. 本文借鉴网络流的相关研究，定义了话题流，以重现网络连接情况.

定义 2　话题流是具有相同四元组特征、相互之间时间间隔小于一定阈值且传输内容为语义数据的数据包集合.

话题流的表达式为

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

式中：id 是流的标志号，一个标志号惟一地对应一条流；t 表示流的到达时间；$\mathrm{ip}_{\mathrm{src}}$、$\mathrm{ip}_{\mathrm{dest}}$ 分别表示话题流的源、端主机地址；p_{src}、p_{dest} 表示话题流的源、端网络端口；C 表示话题流包含的内容负载特性；T 为流内容对应文本的标题信息；S 为发布文本信息的网站.

基于话题流的定义，流量中的一个热点词语与相关属性可表示为

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

式中：W 表示词语集合；f 表示词语的总频度；f_n 表示词语在第 n 条流中出现的频度；id_i 表示包含某词语的第 i 条话题流的标志号.

词语之间的相关度 $\rho(w,w')$ 与 2 个词语流相关的程度直接关联：如果 2 个词语同时出现在一条词语流之中，称这 2 个词语与 1 条词语流相关，则这 2 个词语对象之间的相关度为 1；如果 2 个词语同时出现在 n 条词语流之中，这 2 个词语对象之间的相关度为 n；如果 2 个词语没有共同的词语流，其相关度为 0. 设 2 个词语对象 w 和 w' 的表达式为

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

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

则相关度表达式为

$$\text{If } \mathrm{id}_{i1} = \mathrm{id}'_{j1}, \mathrm{id}_{i2} = \mathrm{id}'_{j2}, \cdots, \mathrm{id}_{in} = \mathrm{id}'_{jn},$$
$$\text{then } \rho(w, w') = n \quad (8)$$

式中：id_{in}、id'_{in}分别表示词语w、w'中的流标志号；$\rho(w, w')$表示w与w'的相关度. 2个词语的$\rho(w, w')$值越大，即2个词语的相关度越大，2个词语同时出现在词语流的次数就越大，则2个词语同属于一个热点话题内容的可能性越大.

3 网络热点话题生成

任意2个热点词语的相关度$\rho(w, w')$可以描述2个词语同属于一个热点话题内容的可能性，从几何角度看，2个热点词语的相关度越大，它们的几何距离越短. 因此，本文采用DBSCAN (Density-Based Spatical Clustering of Application with Noise)[10]聚类算法将具有较大相关度(属于同一热点话题的可能性比较大)的热点词语聚合为簇，这些簇可描述各自对应热点话题(见式(1))的第1项，以簇为基础可分析相关网页标题和网站地址，从而得到如式(1)所描述的网络热点话题.

由式(1)可知，网络热点话题由3部分元素组成. 本文以聚类分析得到的热点词语簇为基础，统计每个类别中的词语流属性的核心标题T和信息发布地址S，然后用每个簇中出现次数满足一定阈值的核心标题、信息发布源及该类别的热点词语，来描述一个网络热点话题，即

$$c = (w_1, w_2, \cdots, w_n) \quad (9)$$

式中：c为聚类得到的热点词语簇，它由n个热点词语组成. 一个热点词语，其属性可由式(5)描述，其中包括词语流的标号值(每一个标号值对应一个词语流，其属性由式(4)描述).

对聚类结果c中的每一个热点词语流集合的T、S进行统计，得到核心标题及相应的信息发布网络地址，即

$$T_C = (s_{T_C}, f_{T_C}, \mathrm{id}_1, \mathrm{id}_2, \cdots, \mathrm{id}_m) \quad (10)$$
$$S_C = (s_{S_C}, f_{S_C}, \mathrm{id}_1, \mathrm{id}_2, \cdots, \mathrm{id}_k) \quad (11)$$

式中：s_{T_C}表示核心标题的字符串；f_{T_C}表示核心标题出现的总频度；s_{S_C}表示网站的字符串，该网址应在k条流中出现；f_{S_C}表示网址出现的总频度；id_i表示第i条包含网址的流的标志号.

对统计得到的核心标题及信息发布地址进行排序，选取频度可达到一定阈值的核心标题和信息发布地址，并与式(9)的聚类结果进行组合，就可以按照式(1)计算网络的热点话题.

4 实验结果分析

4.1 实验环境及数据源

将西安交通大学网络中心的多台HTTP服务器出口的镜像流量数据存储到数据分析服务器之上，然后采用离线分析的方法对90 GB流量数据进行分析、处理. 数据分析服务器为Acer Altos G530，硬件配置为P4 Xeon 3.2处理器，内存为ECC 4 GB，硬盘为SCSI 320 GB，操作系统为Windows 2003 Server，实现程序语言为C++.

4.2 实验结果及分析

在实验中，流量内容的分词处理采用了中国科学院计算所自然语言处理研究组提供的中文智能分词系统. 选取参数：中频阈值为3 000，高频阈值为10 000，相应地生成665个高频词，1 047个中频词，1 899个低频词.

利用DBSCAN聚类算法，对高频词队列进行分析，选取的领域半径$\varepsilon=500$，队列的最小密度阈值$\min q=5$，由此获得的聚类类别数为48，聚类效率为16.7%，其中含有语义信息的热点词语类别数为8，无语义信息的热点词语类别数为40. 在网络热点话题生成的过程中，$T_C=500$，$S_C=500$，由此得到8个网络热点话题的描述信息.

从聚类结果看出，有8项具有较明晰语义信息的热点话题，包括“交大招生科目信息”、“交大校庆消息”、“交大概况及校史”、“交大电气学院关于电力电工试验课程创新实践的新闻”、“交大长江学者介绍”等. 选取其中2项热点话题，通过式(1)的热点话题形式化描述对有效类别进行格式化，结果如表1所示.

在表1描述的2项网络热点话题中，第1项话题包含了81个热点词语(由于篇幅所限，不便全部列举)、3个核心标题和1个信息发布网站. 通过人工分析可知，热点话题与交大人才培养及招生录取信息有关，主要内容为交大的学科专业信息，包括一级学科、二级学科及院系名称，它们均通过交大网站向外传播. 第2项热点话题包含了35个热点词语、3个核心标题及2个信息发布网站，主要内容为交大概况及校史.

表1所示话题的区别有二：其一是第2项话题的热点词语数量只占第1项的43.2%，这表明第2项话题的内容更为集中；其二是第2项话题通过2个网站传播，这表明关注交大主页信息以及关注交

表1　网络热点话题提取结果表示

P		$(W_1,W_2,\cdots,W_l)$	$(T_1,T_2,\cdots,T_m)$	$(S_1,S_2,\cdots,S_n)$
西安交通大学人才培养及招生录取信息	$l=81$ $m=3$ $n=1$	(生物学，人文，电力，医学院，电气，热能，理学院，管理科学，贸易，会计，能源，仪器，化学系，病理学，营销，电机，软件，高压，…，经济学)	(西安交通大学招生就业，西安交通大学人才培养，2003年硕士生录取分数线)	http://www.xjtu.edu.cn
西安交通大学概况及校史	$l=35$ $m=3$ $n=2$	(前身，上海，交通，西安，交大，历史，专门，重点大学，工业，校址，历任，百年，传统，西部，创建，财经，办学，南洋，之一，高等教育，人员，国内，校史，目前，实现，利用，…，站点)	(西安交通大学概况，西安交通大学历史，西安交通大学建校110周年暨交通大学迁校50周年)	(http://www.xjtu.edu.cn，http://newsxq.xjtu.edu.cn)

大校庆新闻的很多用户均对交大概况及校史感兴趣，第2项话题主要关注用户对信息的兴趣度.

可以看出，中文词语语义的丰富性和多义性导致了中文词语聚类的有效性仅能达到16.7%，但是作为一个初步研究的成果，还是较为理想的.通过分析实验结果发现，从流量内容中提取出的8项热点话题，能够在部分程度上描述一定范围内的网络当前舆论状况，给网络管理者提供了辅助的管理信息.由于计算机的自然语言理解能力有限，暂时无法得到语义更加明确的热点话题信息，还需要由人工来解析、分析.

5　结　论

本文根据网络信息特点，定义网络热点话题并给出了其形式化描述.通过分析流量内容中的热点词语与热点话题之间的关系，提出了热点词语的相关度计算算法和网络热点话题的生成方法.该方法采用了DBSCAN聚类算法，再结合与热点词语簇相关的网页标题及网站地址信息，得出网络热点话题的属性描述.实验结果表明，本文方法能够有效地获取当前网络中的热点话题，通过选取更合适的数据源能更大程度地还原网络用户行为，获取用户对网络信息的关注情况，从而得到更为准确的网络信息传播状况.所提方法可作为研究网络热点话题传播特性的基础.

下一步的工作将研究流量内容预处理分析方法，改进网络热点话题提取算法的效率，在热点话题内容提取工作的基础上，开展热点话题动态传播规律以及相关社会网络关系的研究.

参考文献：

[1] James A, Jaime C, George D, et al. Topic detection and tracking pilot study: final report [C]// Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kaufmann Publishers, 1998:194-218.

[2] 于满泉，骆卫华，许洪波，等. 话题识别与跟踪中的层次化话题识别技术研究 [J]. 计算机研究与发展，2006，43(3):489-495.
Yu Manquan, Luo Weihua, Xu Hongbo, et al. Research on hierarchical topic detection in topic detection and tracking [J]. Journal of Computer Research and Development, 2006, 43(3):489-495.

[3] Kosala R, Blockeel H. Web mining research: a survey [J]. SIG KDD Explorations, 2000, 2(1):1-15.

[4] 王泽彬，金飞，李夏，等. Web数据挖掘技术及实现 [J]. 哈尔滨工业大学学报，2005，37(10):1403-1405.
Wang Zebin, Jin Fei, Li Xia, et al. Web data mining technique and realization [J]. Journal of Harbin Institute of Technology, 2005, 37(10):1403-1405.

[5] 李保利，俞士汶. 话题识别与跟踪研究 [J]. 计算机工程与应用，2003，39(17):7-10.
Li Baoli, Yu Shiwen. Research on topic detection and tracking [J]. Computer Engineering and Applications, 2003, 39(17):7-10.

[6] Topic Detection and Tracking (TDT) Evaluation Workshop. The 2002 topic detection and tracking task definition and evaluation plan [EB/OL]. [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 [J]. IEEE Journal on Selected Areas in Communications, 1986, 4(6):986-995.

[8] Mogul J C. Observing TCP dynamics in real networks [J]. 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 [J]. 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 [C]// Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park, USA: AAAI Press, 1996:226-231.

(编辑 苗凌)