一种突发性热点话题在线发现与跟踪方法

薛峰，周亚东，高峰，刘霁，赵俊舟，党琪
（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）

摘要：针对在线发现与跟踪动态突发性文本流中的热点话题问题，在突发性热点词发现与度量方法的基础上提出了一种动态文本模型——动态突发性向量空间模型，用于有效描述文本的动态属性，并且结合文本聚类方法，提出了突发性热点话题的在线发现与跟踪方法. 该方法可有效解决传统的基于静态向量空间模型的热点话题发现与跟踪方法仅可分析静态文本的缺陷，并具有以下特点：在特征选择阶段动态地生成热点词特征库，利用模型统一文本和话题的表示，在文本表示时给予突发性热点词更大的权重. 基于实际网络文本流数据的实验表明，该方法对突发性热点话题发现的精确率与召回率分别达到92.75%和80.34%，显著优于传统的基于静态向量空间模型方法的实验结果，并可有效跟踪突发性热点话题，弥补了传统静态方法不能有效跟踪热点话题的不足.

关键词：突发性热点话题；话题发现与跟踪；向量空间模型
中图分类号：TP393.4 **文献标志码**：A **文章编号**：0253-987X(2011)12-0064-06

An Online Detection and Tracking Method for Bursty Topics

XUE Feng, ZHOU Yadong, GAO Feng, LIU Ji, ZHAO Junzhou, DANG Qi
(Key Laboratory of Intellectual Network and Network Security of the Ministry of Education,
Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Text representation in text mining plays an important role, but the traditional vector space model based on TF-IDF is a static statistical model and is not flexible for bursty topic detection and tracking since it could not model the bursty dynamic text flow (such as news text flow, blog text flow, etc.) effectively. A new model called dynamic bursty vector space model is proposed to model text flow, and to detect and track bursty topics based on bursty feature detection. The proposed dynamic model has several characteristics in contrast to the traditional static model: 1) The model generates features dynamically in feature selection process; 2) A unified representation of the text and topics is given; 3) The model gives more weights to temporal bursty features. The experiments of bursty topic detection and tracking demonstrate that the dynamic bursty vector space model could be able to get higher precision and recall.

Keywords: bursty topic; topic detection and tracking; vector space model

随着互联网的普及以及这一信息发布平台的逐渐完善，各类网络站点已经成为人们获取信息的主要来源之一，因此，如何能够快速准确地从规模巨大、凌乱无序的互联网信息中获取所需要的热点话题信息，特别是获取最新的突发性热点话题信息，已经成为国内外研究者关注的一个热点问题.

网络突发性热点话题是指网络中由突发性公共事件引起的、在出现之前难以预知并且传播周期较

收稿日期：2011-04-25. 作者简介：薛峰(1986—)，男，硕士；周亚东(通信作者)，男，讲师. 基金项目：国家自然科学基金资助项目(60921003，60802056，60905018)；国家“863计划”资助项目(2007AA01Z480)；国家科技支撑计划资助项目(2011BAK08B02).
网络出版时间：2011-10-08 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20111008.0833.003.html

短、对社会舆论影响较大的话题，如重大自然灾害、安全事故、社会公共事件等. 与普通热点话题相比，突发性热点话题具有因其突发性导致的先验知识缺乏以及与网络文本的时间属性有显著关联等特点. 普通热点话题由于事先已知，如“房产住房”、“国家领导人”等，因而可以通过信息检索(Information Retrieval，IR)[1]方法从网络文本中提取，而突发性热点话题不仅具有突发性，而且具有将来发生、当前未知、存在时间相对较短等特点，如“360 与 QQ 大战”、“日本核危机”等，在话题的相关事件发生之前无法预知该话题的相关信息.

突发性热点话题发现与跟踪(topic detection and tracking，TDT)[2-4]就是从网络文本集中识别出突发性热点话题，并跟踪话题的演变过程. 目前，TDT 的研究主要是在文献[5-6]的基础上进行，涉及 2 个方面：一是文本表示方式的改进，二是充分利用文本语料的时间属性. 例如，Fung 等[7-8]通过分析词语的时序分布来发现突发性特征，然后将突发性特征聚集起来发现突发性热点话题. Wang 等[9]和 Suba 等[10]尝试从数据集中挖掘突发性模式，从而构建突发性热点话题. 然而，现有的利用统计模型进行话题发现与跟踪的方法大多属于半在线的方式，利用训练集得到静态特征库，在静态特征库的基础上采用向量空间模型(vector space model，VSM)[11]对文本建模，进而进行话题发现与跟踪. 由于特征库的静态性，导致这种方法容易忽略掉网络中产生的突发性词语，如“躲猫猫”、“瘦肉精”等，使得话题发现与跟踪的精确率在很大程度上受到训练集规模、质量的影响；同时，向量空间模型没有利用文本的时间属性，无法有效地对突发性文本流进行建模.

本文针对网络突发性热点话题的特点，以网络上的海量文本流为研究对象，提出一种动态突发性向量空间模型来表示网络文本和突发性热点话题，在此模型的基础上进行话题发现与跟踪. 实验结果表明，这种方法能够在真实的网络环境中准确、快速、稳定地发现并跟踪突发性热点话题.

1　突发性热点话题的特征选择

突发性热点话题与突发性热点词之间存在着紧密的关系[12]：突发性热点话题的产生往往伴随着突发性词语的出现，具有类似突发性词语的网络文本很可能与同一个话题相关. 所以，突发性热点词可以看作是突发性热点话题发现与跟踪的启发性知识.

本文提出一种结合无监督特征选择[13]和突发性热点词语发现的双层特征选择方法，利用此方法从文本流中提取出动态特征库. 双层特征选择的过程如图 1 所示.

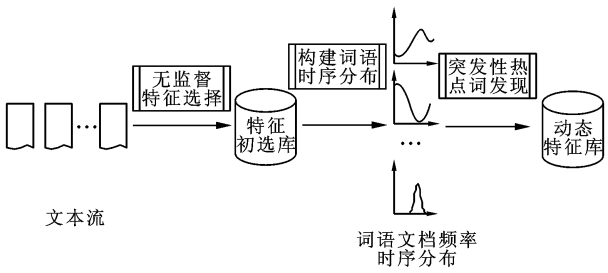

图 1　双层特征选择的过程

词语的突发性度量参考文献[14]中提出的相关方法，基于二项分布的离散事件模型，分析动态网络文本流中词语 w 在时间段 i 内的分布情况. 对于每个词语 w，可以得到表 1 所示的 2×2 相依表.

表 1　词语 w 与时间段 i 的相依表

	i	$\overline{i}$
w	A	B
$\overline{w}$	C	D

A：时间段 i 内包含词语 w 的网络文本数；B：时间段 i 之前包含词语 w 的网络文本数；C：时间段 i 内不包含词语 w 的网络文本数；D：时间段 i 之前不包含词语 w 的网络文本数.

词语 w 在时间段 i 内的突发值由下式计算

$$b_i(w)=\chi^2_{w,i}=\frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \tag{1}$$

选取突发值大于阈值 θ_b 的词语集作为时间段 i 内的动态特征库，从而完成特征的双层选择.

2　动态突发性向量空间模型

传统的文本表示方法中，采用向量空间模型将文本 $\boldsymbol{d}$ 表示成向量形式

$$\boldsymbol{d}=[v_1,\ v_2,\ \cdots,\ v_m] \tag{2}$$

其中词语 w_k 对应的权重 v_k 通过 TF-IDF[4]计算

$$v_k=f_{\mathrm{t}}(w_k,\boldsymbol{d})\log\left[\frac{(N+1)}{f_{\mathrm{d}}(w_k)}+0.5\right]\Big/\left\{\sum_{w'\in\boldsymbol{d}}\left[f_{\mathrm{t}}(w',\boldsymbol{d})\log\left[\frac{(N+1)}{f_{\mathrm{d}}(w')}+0.5\right]\right]^2\right\}^{1/2} \tag{3}$$

式中：$f_{\mathrm{t}}(w_k,\boldsymbol{d})$ 表示词语 w_k 在文档 $\boldsymbol{d}$ 中的词频；$f_{\mathrm{d}}(w_k)$ 表示文档集中包含词语 w_k 的文档频率；N 表示文档集中的文档总数.

向量空间模型中需要一个静态特征库来联系特征词与文本向量中的权重. 然而，在突发性话题的跟

踪过程中，每一次新话题集合的特征空间不一定与原话题集合的特征空间一致；另外，TF-IDF 这种权重度量方式没有体现出词语的突发性，不能有效地支撑突发性热点话题的发现. 因此，本文提出一种动态突发性向量空间模型(dynamic bursty vector space model, B-VSM)，用来表示网络文本和突发性热点话题.

2.1 网络文本的表示

通过分析突发性热点话题的特点，将网络文本 $\boldsymbol{d}$ 表示为

$$\boldsymbol{d}=[w_1:v_1,\ w_2:v_2,\ \cdots,\ w_m:v_m] \tag{4}$$

式中：$w_k\in\boldsymbol{d}$ 并且 $w_k\in F_i(1\leqslant k\leqslant m)$，$F_i$ 是时间段 i 内的特征库；v_k 是特征 w_k 的权重，结合词语 TF-IDF 与突发值进行计算，并作归一化处理，即

$$v_k=\left\{f_{\mathrm{t}}(w_k,\boldsymbol{d})\log\left[\frac{(N_i+1)}{f_{\mathrm{d},i}(w_k)}+0.5\right]+\beta b_i(w_k)\right\}\Bigg/\left\{\sum_{w'\in\boldsymbol{d}}\left\{f_{\mathrm{t}}(w',\boldsymbol{d})\log\left[\frac{(N_i+1)}{f_{\mathrm{d},i}(w')}+0.5\right]+\beta b_i(w')\right\}^2\right\}^{1/2} \tag{5}$$

其中 N_i 表示时间段 i 及其之前出现的文档总数，β 表示赋予词语 w_k 突发值 $b_i(w_k)$ 的加权系数.

式(5)在传统 TF-IDF 计算公式的基础上，通过加入突发性因子 $\beta b_i(w_k)$，能够提高突发性热点词在文本表示中的权重，使得文本聚类算法能够更准确地发现突发性热点话题.

2.2 话题的表示

话题 $\boldsymbol{T}$ 由与其相关的网络文本组成，其形式化描述如下

$$\boldsymbol{T}=\{d_1,\ d_2,\ \cdots,\ d_n\} \tag{6}$$

话题 $\boldsymbol{T}$ 的核心向量 $\boldsymbol{T}'$ 可由 DB-VSM 表示为

$$\boldsymbol{T}'=[w_1:v_1,\ w_2:v_2,\ \cdots,\ w_n:v_n] \tag{7}$$

式中：$w_l(1\leqslant l\leqslant n)$ 表示话题 $\boldsymbol{T}$ 的核心关键词；v_l 表示 w_l 在 $\boldsymbol{T}$ 中的权重，计算公式为

$$v_l=\frac{\sum_{d_k\in\boldsymbol{T}}f_{\mathrm{d}}(w_l,d_k)}{\left(\sum_{w\in\boldsymbol{T}'}\left(\sum_{d_k\in\boldsymbol{T}}f_{\mathrm{d}}(w,d_k)\right)^2\right)^{1/2}} \tag{8}$$

其中 $f_{\mathrm{d}}(w_l,d_k)$ 表示特征词 w_l 在文本 $d_k(d_k\in\boldsymbol{T})$ 中的词频.

式(4)～式(9)所描述的 DB-VSM 模型利用双层特征选择方法得到的动态特征库，通过引入突发性因子来增加突发性词语在文本向量中的权重，可以较好地体现网络文本的突发性，在此基础上进行突发性热点话题发现与跟踪具有更好的精确率与召回率(后面将进行实验验证).

3 突发性热点话题发现与跟踪方法

本节将利用 DB-VSM 来表示文本和话题，并使用 FDBSCAN[15] 文本聚类算法进行突发性热点话题的发现；在发现候选新话题后，根据话题之间的相似性进行话题合并与更新.

设网络文本 $\boldsymbol{d}=[w_i:v_i,\ \cdots,\ w_j:v_j]$，$\boldsymbol{d}'=[w_m:v_m,\ \cdots,\ w_n:v_n]$，则两者之间的相似度 ζ_d 为

$$\zeta_d(\boldsymbol{d},\boldsymbol{d}')=\sum_{w\in\boldsymbol{d}\cap\boldsymbol{d}'}v\,v' \tag{9}$$

式中：$\boldsymbol{v}$ 和 $\boldsymbol{v}'$ 分别是词语 w 在文本 $\boldsymbol{d}$ 与 $\boldsymbol{d}'$ 中的权重.

设原有话题 $\boldsymbol{T}_i$ 的 DB-VSM 为 $\boldsymbol{T}_i'=[w_{ik}:v_{ik},\ \cdots,\ w_{il}:v_{il}]$，对于新出现的文本集使用文本聚类进行话题发现得到的候选话题 $\boldsymbol{T}_j$ 的 DB-VSM 为 $\boldsymbol{T}_j'=[w_{jx}:v_{jx},\ \cdots,\ w_{jz}:v_{jz}]$，定义话题 $\boldsymbol{T}_i$ 与 $\boldsymbol{T}_j$ 的相似度 ζ_T 为

$$\zeta_T(\boldsymbol{T}_i,\boldsymbol{T}_j)=\sum_{w\in\boldsymbol{T}_i'\cap\boldsymbol{T}_j'}v_i\,v_j \tag{10}$$

式中：$\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 是词语 w 在话题 $\boldsymbol{T}_i$ 与 $\boldsymbol{T}_j$ 中的权重.

若话题 $\boldsymbol{T}_i$ 与 $\boldsymbol{T}_j$ 的相似度大于阈值 θ_T，则需要将两者合并. 设合并后的话题为 $\boldsymbol{T}$，它的 DB-VSM 为 $\boldsymbol{T}'=[w_m:v_m,\ \cdots,\ w_n:v_n]$，其中

$$\{w_m,\ \cdots,\ w_n\}=\{w_{ik},\ \cdots,\ w_{il}\}\cup\{w_{jx},\cdots,w_{jz}\} \tag{11}$$

话题 $\boldsymbol{T}$ 中关键词 w 的权重 v 通过加权平均数计算

$$v=\frac{f_{\mathrm{d}}(w,\boldsymbol{T}_i)f_T(w,\boldsymbol{T}_i)+f_{\mathrm{d}}(w,\boldsymbol{T}_j)f_T(w,\boldsymbol{T}_j)}{f_{\mathrm{d}}(w,\boldsymbol{T}_i)+f_{\mathrm{d}}(w,\boldsymbol{T}_j)} \tag{12}$$

式中：$f_{\mathrm{d}}(w,\boldsymbol{T}_i)$ 表示话题 $\boldsymbol{T}_i$ 中包含词语 w 的文档频率；$f_T(w,\boldsymbol{T}_i)$ 表示词语 w 在话题 $\boldsymbol{T}_i$ 中的词频.

最后，选取权重最大的 N 个词作为合并后话题的关键词集合，并对其对应的 DB-VSM 进行归一化处理.

突发性热点话题发现与跟踪方法的流程如图 2 所示.

4 实验与分析

为了分析 DB-VSM 文本表示方法在话题发现与跟踪上的效果，本文使用多线程的网络爬虫从网络中获取了与“360 与 QQ 大战”话题相关的 760 篇相关文本作为实验数据，评测指标采用聚类中广泛使用的精确率、召回率和 F 度量[16].

假设数据集中有 k 个话题 $T_1,T_2,\ \cdots,\ T_k$，通过话题发现方法得出 m 个类簇 $C_1,C_2,\ \cdots,\ C_m$，属于话题 T_i 并被聚类到类簇 C_j 的文本数为 n_{ij}(如表

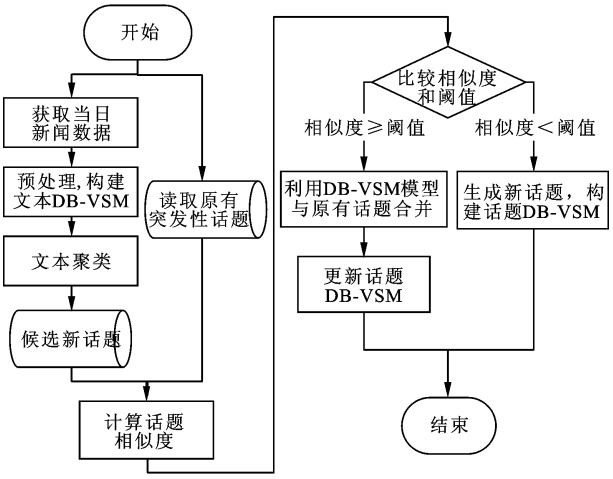

图 2　突发性热点话题发现与跟踪流程

2 所示). 关于话题 i 和类簇 j 的精确率和召回率分别定义为

$$P(i,j)=\frac{n_{ij}}{n_j} \tag{13}$$

$$R(i,j)=\frac{n_{ij}}{n_i} \tag{14}$$

式中:$n_i=\sum_{j=1}^{m}n_{ij}$;$n_j=\sum_{i=1}^{k}n_{ij}$.

表 2　聚类结果邻接表

	C_1	C_2	…	C_j	…	C_m
T_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}
T_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}
⋮	⋮	⋮		⋮		⋮
T_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}
⋮	⋮	⋮		⋮		⋮
T_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}

话题 i 和类簇 j 的 F 度量由相应的精确率和召回率定义

$$F(i,j)=\frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{15}$$

话题 i 的 F 度量定义为 $F(i,j)$的最大值,其中 $1\leqslant j\leqslant m$. 与其对应的 j 表示话题 i 通过聚类所归属的类簇标号,相应的 $P(i,j)$和 $R(i,j)$定义为话题 i 的精确率和召回率. 从而,话题 i 的 3 种评价标准分别表示为 P_i、R_i 和 F_i.

算法应用于全数据集的 3 种评价指标通过对每个话题的评价指标进行加权平均计算得出

$$P=\sum_i\frac{n_i}{n}P_i \tag{16}$$

$$R=\sum_i\frac{n_i}{n}R_i \tag{17}$$

$$F=\sum_i\frac{n_i}{n}F_i \tag{18}$$

图 3 为此话题的相关网络文本数量在时间上的分布情况,以及各主要时间段话题的讨论内容.

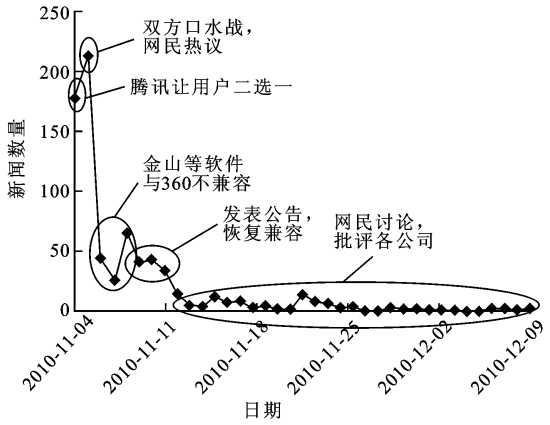

图 3　360 与 QQ 大战的网络文本数量分布图

从图 3 中可以看出,关于此话题的网络文本主要分为 5 个阶段,各阶段的时间与主要内容见表 3.

表 3　360 与 QQ 大战各阶段的时间与主要内容

起始日期	终止日期	网络文本主要内容
2010-11-04	2010-11-04	腾讯发布公告,让用户在 360 与 QQ 之间二选一.
2010-11-05	2010-11-05	双方发生口水战,网民激烈讨论.
2010-11-06	2010-11-08	金山、遨游等公司宣布与 360 公司的软件不兼容.
2010-11-09	2010-11-11	腾讯与 360 公司发布公告,双方软件恢复兼容.
2010-11-12	2010-12-10	网民对此次互联网大战积极讨论,批评双方公司.

4.1　突发性因子中词语权重的实验分析

为了对 DB-VSM 中突发性因子的参数 β 进行讨论,本文通过改变词语权重参数 β,对此数据集进行话题发现,得到的效果如图 4 所示.

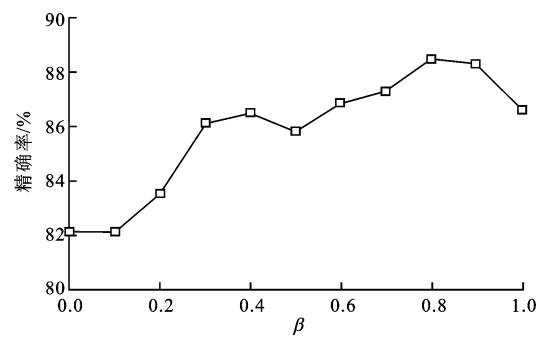

图 4　词语突发性权重对话题发现效果的影响

从图 4 中可以看出，在进行突发性热点话题发现时，应当适当地增大突发性因子中的参数 β.

4.2 DB-VSM 建立过程的实验分析

使用文本聚类进行突发性热点话题发现，保留 10 个权重最大的词语作为话题的关键词，得到 5 个阶段的话题 DB-VSM，见表 4.

表 4 话题每个阶段的 DB-VSM

日期	文本数量	话题 DB-VSM 及词语文档频率
2010-11-04	138	[360:0.568 6，QQ:0.521 2，腾讯:0.389 3，用户:0.324 5，软件:0.240 0，公司:0.130 0，一个:0.104 3，网友:0.095 7，网民:0.095 3，电脑:0.094 2] [136，138，122，122，106，81，70，57，66，89]
2010-11-05	213	[360:0.598 3，QQ:0.508 0，腾讯:0.415 5，用户:0.305 5，软件:0.183 9，公司:0.134 2，互联网:0.098 9，一个:0.093 7，网民:0.081 0，大战:0.079 7] [209，211，176，155，141，131，114，115，101，81]
2010-11-06	44	[360:0.639 4，QQ:0.487 9，腾讯:0.350 5，用户:0.314 7，软件:0.205 5，公司:0.144 9，兼容:0.102 1，网民:0.100 0，电脑:0.085 2，选择:0.083 6] [43，44，36，31，30，25，12，22，22，25]
2010-11-10	31	[360:0.647 0，腾讯:0.505 6，QQ:0.439 9，用户:0.178 0，大战:0.169 1，兼容:0.135 0，软件:0.098 0，公司:0.083 9，补偿:0.075 9，游戏:0.069 7] [31，31，31，27，25，23，24，23，21，22]
2010-11-12 后	79	[360:0.572 8，QQ:0.488 9，腾讯:0.379 3，用户:0.336 7，软件:0.229 9，公司:0.170 9，一个:0.111 3，互联网:0.103 0，电脑:0.095 1，网民:0.094 3] [79，79，68，62，55，42，50，37，38，35]

设话题相似度阈值 $\theta_T=0.8$，计算出 11 月 4 日与 5 日的话题相似度为 0.982 907，大于 θ_T，因此合并两话题，如表 5 所示.

表 5 4 日与 5 日话题合并后的话题 DB-VSM

日期	文本数量	话题 DB-VSM 及词语文档频率
2010-11-05	351	[360:0.585 4，QQ:0.512 2，腾讯：0.404 0，用户:0.313 3，软件:0.207 5，公司:0.132 3，一个:0.097 5，互联网:0.087 9，网民:0.086 5，电脑:0.082 8] [345，349，298，277，247，212，185，165，167，207]

按照这样的方法对每个阶段的话题依次合并，得到“360 与 QQ 大战”话题的最终 DB-VSM，如表 6 所示.

表 6 “360 与 QQ 大战”的最终话题 DB-VSM

时间	文本数量	话题 DB-VSM 及词语文档频率
2010-12-10 后	505	[360:0.587 7，QQ:0.498 4，腾讯:0.400 1，用户:0.305 7，软件:0.202 0，公司:0.134 1，大战:0.099 9，一个:0.098 1，竞争:0.090 8，网民:0.088 4] [498，503，433，397，356，302，106，260，35，224]

4.3 基于 DB-VSM 的突发性热点话题发现实验分析

分别利用 VSM 和 DB-VSM 表示文本，然后进行话题发现，得到的效果如图 5 所示. 从图 5 可以看出，相比于 VSM 这种传统的文本表示方法，在 DB-VSM 的基础上进行突发性话题发现能够取得更好的效果，平均精确率达到 90%以上.

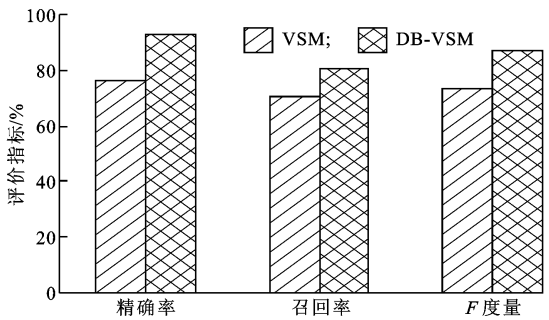

图 5 使用 VSM 和 DB-VSM 进行突发性话题发现的效果

4.4 基于 DB-VSM 的突发性话题发现与跟踪实验分析

为了研究 DB-VSM 在话题发现与跟踪中的性

能和效果,本文获取了 2010 年 11 月网络中讨论的 88 个突发性热点话题、9 170 条网络文本作为突发性热点话题跟踪的实验数据集. 基于 DB-VSM 表示方法的突发性热点话题发现与跟踪性能及效果如图 6、图 7 所示.

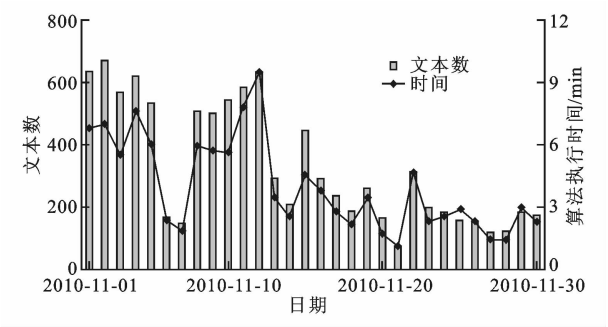

图 6　基于 DB-VSM 的话题发现与跟踪的性能

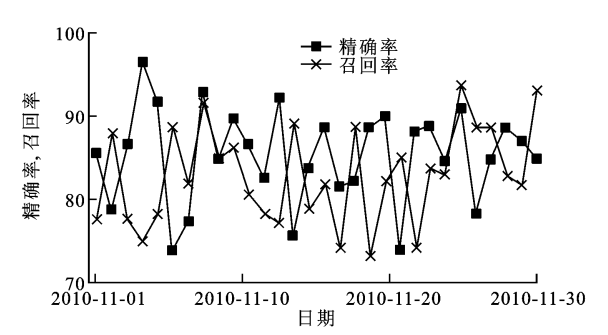

图 7　基于 DB-VSM 的话题发现与跟踪的效果

从图 6、图 7 可以看出,对于真实的网络文本,本文提出的方法能够在短时间内完成突发性热点话题的发现与跟踪,精确率为 85%左右,召回率为 80%左右,能够满足实际网络环境中的应用要求.

5　结　语

随着互联网规模的迅速扩大和网络文本的海量涌现,网络突发性热点话题的发现与跟踪已经成为网络舆情监管的一个重要手段,然而向量空间模型既无法有效地对文本流进行建模,也体现不出特征的突发性. 本文基于向量空间模型与突发性热点词发现,提出了一种动态突发性向量空间模型,用于对网络上持续发布的文本和不断动态演化的突发性话题进行建模,指导文本聚类方法发现并跟踪网络突发性热点话题. 实验结果表明,本文方法能够有效地应用于实际互联网环境.

参考文献:

[1] CROFT B, METZLER D, STROHMAN T. Search engines: information retrieval in practice [M]. Reading, MA, USA: Addison-Wesley Publishing Company, 2009: 552.

[2] LI Hong, WEI Jinfeng. Netnews bursty hot topic detection based on bursty features [C]// Proceedings of International Conference on E-Business and E-Government. Washington DC, USA: IEEE, 2010: 1437-1440.

[3] HOLZ F, TERESNIAK S. Towards automatic detection and tracking of topic change [M]// GELBUKH A. Computational Linguistics and Intelligent Text Processing. Berlin, Germany: Springer-Verlag, 2010: 327-339.

[4] JING Qiu, LIAO Lejian, DONG Xiujie. Topic detection and tracking for Chinese news web pages [C]// Proceedings of Seventh International Conference on Advanced Language Processing and Web Information Technology. Washington DC, USA: IEEE Computer Society, 2008: 114-120.

[5] ALLAN J, PAPKA R, LAVRENKO V. On-line new event detection and tracking [C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1998: 37-45.

[6] YANG Yiming, PIERCE T, CARBONELL J. A study of retrospective and on-line event detection [C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information. New York, USA: ACM, 1998: 28-36.

[7] FUNG G P C, YU J X, LIU H, et al. Time-dependent event hierarchy construction [C]// BERKHIN P, et al. Proceedings of the Thirteenth ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2007: 300-309.

[8] FUNG G P C, YU J X, YU P S, et al. Parameter free bursty events detection in text streams [C]// Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway: VLDB Endowment, 2005: 181-192.

[9] WANG Xuanhui, ZHAI Chengxiang, HU Xiao, et al. Mining correlated bursty topic patterns from coordinated text streams [C]// BERKHIN P, et al. Proceedings of the Thirteenth ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2007: 784-793.

[10] SUBA I, BERENDT B. From bursty patterns to

bursty facts: the effectiveness of temporal text mining for news [C]// Proceedings of 19th European Conference on Artificial Intelligence. Fairfax, VA, USA: IOS Press, 2010: 517-522.

[11] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. Inf Process Manage, 1988, 24(5): 513-523.

[12] HE Q, CHANG K Y, LIM E P. Using burstiness to improve clustering of topics in news streams [C]// RAMAKRISHNAN N, et al. Proceedings of the Seventh IEEE International Conference on Data Mining. Washington DC, USA: IEEE Computer Society, 2007: 493-498.

[13] RIBEIRO M N, NETO M J R, PRUDENCIO R B C. Local feature selection in text clustering [M]// Advances in Neuro-Information Processing. Heidelberg, Germany: Springer-Verlag, 2009: 45-52.

[14] SWAN R, ALLAN J. Automatic generation of overview timelines [C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2000: 49-56.

[15] LIU Bing. A fast density-based clustering algorithm for large databases [C]// Proceedings of 2006 International Conference on Machine Learning and Cybernetics. Washington DC, USA: IEEE, 2006: 996-1000.

[16] LUO Congnan, LI Yanjun, Chung S M. Text document clustering based on neighbors [J]. Data and Knowledge Engineering, 2009, 68(11): 1271-1288.

[本刊相关文献链接]

海量信息异常检测问题的异常概率排序算法. 西安交通大学学报,2011,45(4): 36-40.

应用粒子群优化-条件随机域的文本生物实体识别. 西安交通大学学报,2010,44(12):38-42.

高效的用户访问预测新算法. 西安交通大学学报,2010,44(4):28-33.

具有孤立项过滤的信息检索查询词的分析方法. 西安交通大学学报,2009,43(8):6-10.

采用并行遗传算法的文本分割研究. 西安交通大学学报,2009,43(12):40-44.

面向入侵检测系统的模式匹配算法研究. 西安交通大学学报, 2009,43(2):58-62.

结合受控词汇表的生物基因本体标注与分类. 西安交通大学学报,2008,42(2):171-174.

流量内容词语相关度的网络热点话题提取. 西安交通大学学报, 2007,41(10):1142-1145.

基于免疫记忆克隆的特征选择. 西安交通大学学报,2008,42(6):679-682.

蚁群-遗传融合的文本聚类算法. 西安交通大学学报,2007,41(10):1146-1150.

一种增量式文本软聚类算法. 西安交通大学学报, 2007,41(4):398-401.

一种新颖的基于量化概念格的属性归纳算法. 西安交通大学学报, 2007,41(2):176-179.

(编辑 葛赵青 武红江)