doi:10.3969/j.issn.1000-436x.2013.z2.015

基于 SVM 过滤的微博新闻话题检测方法

程俊霞[1]，李芝棠[1,2]，邹明光[1]，肖津[1]
(1. 华中科技大学 计算机学院，湖北 武汉 430074；2. 下一代互联网接入系统国家工程实验室，湖北 武汉 430074)

摘 要：在基于聚类的话题检测方法上提出了一种基于 SVM 过滤的检测方法，该方法在聚类前将微博文本特征抽象成用于输入向量机的向量，对微博文本进行过滤，降低了计算量。并针对微博聚类的长尾现象提出了基于高频词排序的改进单遍聚类方法，能很好地检测孤立点的存在。实验表明，该方法在海量微博数据中能有效地检测出新闻话题。

关键词：话题检测；特征向量；SVM
中图分类号：TP311.134.3 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0074-05

Novel topic detection method for microblog based on SVM filtration

CHENG Jun-xia[1], LI Zhi-tang[1,2], ZOU Ming-guang[1], XIAO Jin[1]
(1. School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;
2. National Engineering Laboratory for Next Gerneration Internet Access System, Wuhan 430074, China)

Abstract: A detection method based on SVM filtration was proposed. The method uses text feature as imported vectors to filtrate microblog news, reducing the amount of calculation greatly. A single-pass clustering algorithm based on the improvement of high-frequency words sorting was proposed, which can detect isolated points commendably. Experimental results show that the method can detect news topics from massive microblog data efficiently.

Key words: topic detecting; characteristic vector;SVM

1 引言

微博是近年来发展非常快且影响非常大的网络全民媒体形式。自从2006年Twitter在美国上线以来，其注册用户已超过 10 亿[1]。以新浪微博、腾讯微博为代表的国内微博平台也表现出了强大的发展势头。用户可以通过网页、移动客户端、开放 API 等各种途径随时随地记录生活见闻、表达个人观点、关注好友状态，或者是了解最新时事等[2]。微博因其全民参与的草根性使得微博信息具有实时性，即能使突发新闻在微博上以很快的速度传播。同时也使微博具有数据量大、新闻信息密度高的特征。

针对微博的实时性，对微博内容进行分析和整合具有重要的实际意义，不仅可以帮助过滤无效信息，提高内容质量、改善用户体验，更能起到监测、观点挖掘、舆情控制等重要作用。另一方面，微博是一个信息流量相当大的平台，而内容格式又非常散乱、数据噪声较大，人工审视或者基本的统计方法很难有效地从海量数据中提炼出精确有用的信息，因此引入文本挖掘的方法对信息进行去重、筛选、聚类非常必要[3~5]。突发事件检测作为微博文本挖掘的一大方向，在国内外都逐渐受到关注。

但目前针对微博话题的检测研究成果还比较少，传统方法应用于微博时往往出现计算量过大，准确率较低的现象。因此需要提出一种新的高效微博新闻话题检测方法。

2 无效微博特征及过滤

据统计，仅新浪微博每天发布的微博数就达到 1.17 亿，但是大部分微博跟当天的热点话题无关，在检测微博话题之前，要先区别有效微博和无效微博。有效微博是指具有一定的新闻价值或者关注价值的微博消息，内容比较正式，形式上与传统正式

收稿日期：2013-09-08

报道的摘要内容接近，而且是有新闻价值的微博。无效微博是指个人情绪表达、推销、抽奖等消息。在话题检测之前对微博信息进行过滤，将大大减少计算数据量。

2.1　微博文本特征

经观察，那些属于个人感情宣泄或者状态更新的消息与具有新闻价值的消息相比，在URL链接数、表情符号、特殊标点符号、语句数等特定特征上有比较明显的特征。经实验统计可以得到以下结论。

1) 有效微博中，接近 50%的微博都含有链接，而且基本都只含有 1 条链接，存在一部分无效微博数据含有 2~3 个链接，多是一些推销、促销、抽奖等信息。

2) 有效微博中，只有不到 7%的微博含有表情符号，而且使用最多的是[蜡烛]、[鼓掌]，基本都只有一个表情符号，无效微博为了表达自己的喜、怒、哀、乐情感，含有的表情符号较多。从总体上看超过了 50%。

3) 用户通常使用一些特殊的标点符号来增强语气或者表达情感，其中一种就是连续性符号，无效微博中含有特殊标点符号的微博占了 15%左右；其中，最多的是连续个“>>>”，很多无效微博发URL链接之前总会加上这个符号。

4) 有效微博数据长度集中在 5~9 句，而无效微博集中在 1~4 句。实际上，表达个人信息的无效微博往往只需一两句话就说清楚，所以语句数相对较少。

2.2　无效微博过滤

本文采用 SVM 方法，根据上文的微博统计特征对采集的微博数据进行过滤。采用 SVM 分类方法有以下几个好处。

1) 有效微博与无效微博在文本盘特征上有明显的区别，将这些特征可以表示成向量空间(VSM)的模型，正好满足 SVM 的输入数据条件。

2) 过滤的目的是区分有效微博和无效微博，这是一个二分类的问题，微博的这些特征值基本都是[1,10]的整数，SVM 实现二分类问题时效果最好。

3) 由于微博数据量很大，不可能统计所有微博的特征，只能通过对少数样本的学习实现对大量数据集的分类，SVM 正是通过小样本学习实现对大数据集的检测。

4) 特征项之间并非完全独立，基于特征项独立的分类方法不可用，而 SVM 不要求特征独立。

实验中过滤需要的特征集由 10 个特征项组成，分别是链接、平台表情出现、自行输入表情、连续的特殊符号(“?”、“~”、“!”、“>”)、问号(“?”)、感叹号(“!”)、语句数。特征值则是相关特征在微博信息中出现的次数。选取不同特征集对算法正确性的影响将在第 4 节分析。

3　话题检测系统

3.1　文本表示

文本表示是指将微博文本特征即特征词提取出来形成文本特征集，每个特征有对应的特征值，最终将文本表示成空间向量模型的过程。相似度计算是根据微博文本的空间向量模型计算微博的相似度。在对文本进行分类聚类之前，需要将文本进行分词、去停用词、提取文本特征、计算特征值等步骤。微博文本特征提取流程如图 1 所示。

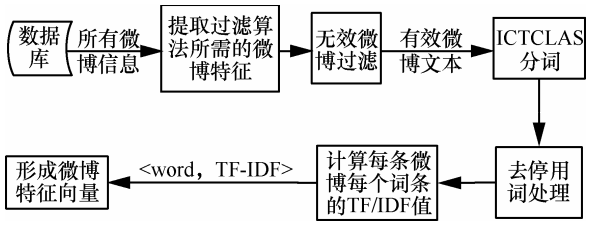

图 1　特征提取流程

将分别统计去停用词前和去停用词后不同字数词语的个数及比例，旨在分析不同字数词语对整条微博信息内容的影响。随机选取 1 000 条微博分词后的统计结果，在去除停用词的情况下，词语长度主要分布在一字词和二字词上，其中，二字词的比率超过了 50%。

通过对比分析原始语料和分词结果发现，一字词中出现较多的原因是一些词语被人为地隔开，或者为人名、地名、书名等造成的。三字词含有的信息量大，基本是类似“深圳市”、“人民币”、“CPI”、“山西省”等这样的一些关于特定地点或陈述对象的词。这些词也基本属于微博文本内容的主题词。超过 3 个字的词基本是一些与“一无所有”、“58312626”、“hobbit”等类似的成语、数字或英文单词，这些词对文本话题核心内容的表达贡献度不高。在特征值计算时，不同的词条可以给予不同的特征值贡献因子，能更好地区别不同的特征项。

本文将微博文本 m 表示为

$$m=[w_1:v_1,w_2:v_2,\cdots,w_m:v_m]$$

其中，w_i 是微博 m 中的词条，v_i 是 w_i 的权值，对于

词条 w_i 权值的计算规则如下：

1) 按照 TF-IDF 定义，计算每个词条的初始 TF-IDF 值得到 m 的初始向量为

$$\boldsymbol{m}=[w_1:v_{i_1},w_2:v_{i_2},\cdots,w_m:v_{i_m}]$$

2) 对 $\boldsymbol{m}$ 中字数不同的词条乘以不同的因子，一字词、二字词、三字词别乘以 0.5、1、1.5。

3) 对于那些出现在话题标签（“#*#”、“【*】”）内的词条，将经过 2）计算后的值乘以 2。

3.2　IPSC 话题检测方法

3.2.1　IPSC 方法描述

通过实验统计，微博的聚类结果具有长尾现象。具体表现为，在聚类结果中，只有极少类别含有的微博消息条目数较多，大部分是小类别或孤立点。对这些孤立点进行聚类本身就是无意义的，如果在聚类过程中将所有孤立点全部检测完，这将增加大量不必要的计算。从处理海量数据的效率出发，本文的话题检测方法采用单遍聚类的思想，因为单遍聚类的时间开销相对传统的聚类算法低很多，但是单遍聚类对初始输入数据很敏感。为此本文提出能检测改善初始输入和检测孤立点的改进话题检测方法，简称 ISPC 方法。

经实验发现，对输入的微博排序能够改善单遍聚类对初始数据敏感的现象。本文选择的微博排序规则是按照每条输入数据的高频词命中率从高到低排序，高频词指输入数据去停用词后的所有词语中 DF(文档频率)不小于 3 的词。因为微博文本字数少，每条微博内容的针对性较强，经常只是话题某个具体侧面的内容。因此那些文档频率越高的词越能代表话题的核心内容。如果一条微博只有低频词，那么说明与它相关的微博很少，这样的文本经过聚类将以小类别(文本数小于 3)或孤立点存在，相对于微博庞大的数据量来说，这些小类别或孤立点不属于话题检测所关心的信息，可以直接去掉。

经过排序后，孤立点和小类别微博因为包含的高频词少，基本都在排序结果的最后面。运行单遍聚类算法时，当检测微博数超过 50%左右时，几乎剩下的所有微博都以新类的形式加入到聚类结果中，即使能合并的也都是合并成了小类别。

具体的数据统计结果如表 1 所示。其中，取一天中 3 个不同时间段的微博测试，检测的停止条件是连续产生的微博类别数超过阈值或所有微博数据已经检测完。其中微博数指该时间段的微博总个数，检测数指聚类实际处理的微博数，漏检数是本身不是孤立点的微博，检测结果中被判定为孤立点的微博个数，检测率是检测数与微博数之比。分析漏检的微博，都属于那些类别大小不超过 3 的类，这些小类别的漏检对最终话题的检测几乎无影响。

表 1　　ISPC 检测率统计

时间段	微博数	检测数	孤立点		漏检数	检测率
			实际量	检测量		
10~12	2 756	1 614	1 439	1 431	8	58.57%
12~14	1 727	1 031	795	791	4	59.69%
14~16	2 414	1 401	1 253	1 245	8	56.21%

3.2.2　ISPC 话题检测总体流程

ISPC 话题检测算法的执行流程如图 2 所示。图中出现了几个重要的数据项，其含义代表如下。

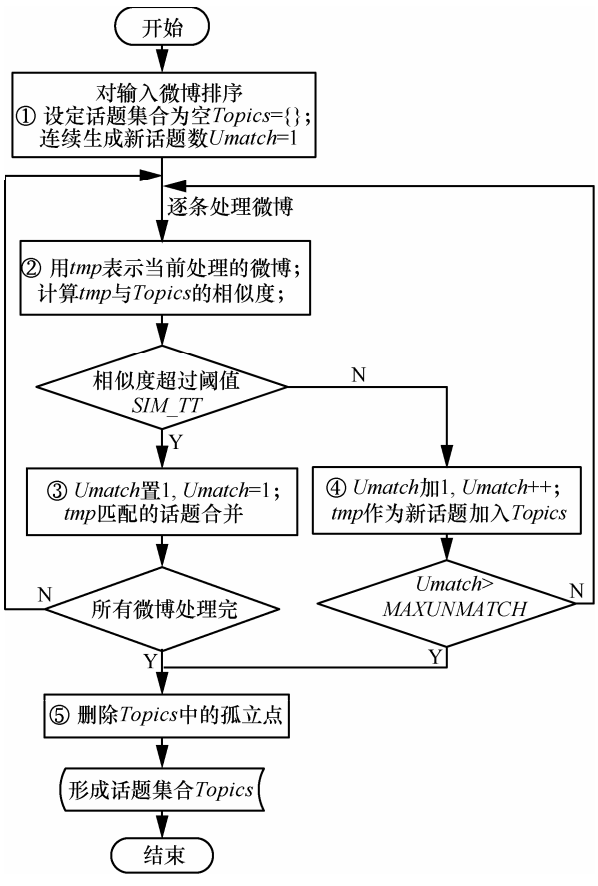

图 2　ISPC 处理流程

Topics：话题的集合，初始值为 0；

Umatch：连续生成新话题的数目，产生新话题加 1，合并则置 0；

SIM_TT：微博与当前话题相似度阈值，超过阈值则合并，实验取值为 0.6。

MAXUNMATCH：连续产生的新话题数的最大阈值，*Umatch* 超过该值时停止检测，实验取值为 100。

在图 2 的 SPTD 流程中，①~⑤是算法的 5 个核心处理过程，分别代表的含义如下：

① 利用微博排序方法排序所有输入的微博数据；

② 将当前处理的微博与 *Topics* 集合中的话题逐个计算相似度；

③ 若②中的相似度超过阈值，就将当前处理的微博与对应的话题合并；

④ 若②中相似度没有超过阈值，根据当前的微博生成一个新话题加入 *Topics* 中；

⑤ 删除聚类结果中那些孤立点类别，最终输出话题集。

4　实验与评估

4.1　SVM 微博过滤实验评估

基于 SVM 的算法评估的主要目的是，分析在选取不同的特征类别的情况下 SVM 过滤器的表现优劣，主要从准确性和效率两方面考虑。随机选择 6 000 条微博数据，其中有 1 992 个正样本。将其分成两组，3 000 条用于训练，另外 3 000 条用于测试。详细结果如表 2 所示，其中，所有特征是指包含了 URL 链接特征、表情符号特征、特殊标点符号特征、语句数特征的情况。其他情况是指去掉其中一种特征，剩下其余 3 种特征的情况。

从表 2 中可以看出，4 个特征同时选取的情况下，准确率超过 95%，有效微博的召回率达到 98.92%，如果去掉链接、表情符号、特殊标点符号 3 个特征中的任意一个，检测的准确率和召回率都有不同程度的下降。在没有语句数特征时，准确率只有 71.8%，而且有效微博的召回率只有 31.26%；可以看出语句数对 SVM 过滤算法的影响最大，是最能区分有效微博和无效微博的特征。

4.2　IPSC 算法评估

采集了 2013 年 1 月 8 日的微博数据共 34 999 条，其中，经过 SVM 过滤处理后剩余 15 527 条有效微博数据。利用 15 527 条数据进行实验，得到排名前 5 的话题如表 3 所示，其中话题内容是合并结果中的一条微博，主题词是合并话题的关键词，合并是指这个话题合并的微博数量，合并的微博数量越多，越能代表这个话题受到广泛关注。这一天的热词榜有<药品,279>、<昆明,238>、<地铁,209>、<事故,177>、<兰考,174>、<火灾,168>、<中国,159>、<热水袋,132>、<报告,117>、<列车,109>，其中的数字表示词的词频。

5　结束语

本文以文本为基础进行话题检测，目前已有不少学者分享了微博拓扑结构、用户影响力、用户拓扑等方面的研究，可以考虑将传统话题检测与微博应用的结构、用户相结合，制定出更好的微博话题检测方法。

表 2　**SVM 各个特征贡献值评估**

测试类别	特征类别				
	所有特征	无 URL 链接	无表情符号	无特殊标点符号	无语句数统计值
准确率	95.57%	90.90%	90.3%	89.26%	71.8%
有效微博召回率	98.92%	92.58%	89.49%	88.17%	31.26%

表 3　**话题检测示例**

ID	话题内容	主题词	合并
1	【昆明地铁空载试运行脱轨致 1 死 1 伤】	昆明，脱轨，地铁，运行，列车，车厢，司机，试，调查组，载，轻伤，值班，南段	90
2	【我国下月起 400 余品种药物降价，平均降幅 15%】	药品，降价，财政局，下月，降幅，品种，自筹，解热，镇痛，平均，限价，剂型	75
3	【热水袋充电变“炸弹”烫伤 5 岁男童半张脸】	热水袋，充电，烫伤，脸部，男童，炸弹，大面积，床头柜，爆炸，烧伤，脖子，孙子	57
4	【中科院报告：中国 2049 年全面超越美国实现复兴】	超越，中科院，复兴，时间表，美国，路线图，报告，实现，全面，总量，提出，预计	56
5	【兰考火灾事故 6 名相关责任人被停职检查】	责任人，民政局，兰考，停职，城关，镇长，党委，火灾，民政，科员，局长，书记	53

参考文献：

[1] LUO W H，LIU Q. Research and development of topic detection and tracking technology[A]. Research and Development of Topic Detection and Tracking Technology[C]. 2003.1437-1440.

[2] SUN W P. Research on the Detection and Tracking of Chinese Micro blog Hot Topic[D]. Beijing: Beijing Jiaotong University,2011.

[3] CRESTANI M F, RUTHVEN I. Construction of topics and clusters in topic detection and tracking tasks approach in semantic technology and information retrieval[A]. Technology and Information Retrieval International Conference[C]. Putrajaya,2011. 171-174.

[4] LI R F, GUO W B.HMM-based state prediction for Internet hot topic[A]. Computer Science and Automation Engineering International Conference[C]. Guangzhou, China,2009. 157-161.

[5] SEO Y W, SYCARA K. Text Clustering for Topic Detection Technical Report[D]. Carnegie Mellon University, 2004.

作者简介：

程俊霞（1990-），女，河南开封人，华中科技大学硕士生，主要研究方向为网络与信息安全。

李芝棠（1951-），男，湖北监利人，华中科技大学教授、博士生导师，主要研究方向为计算机系统结构、网络与信息安全、P2P 网络。

邹明光（1986-），男，江西抚州人，华中科技大学硕士生，主要研究方向为网络与信息安全。

肖津（1990-），女，湖北武汉人，华中科技大学硕士生，主要研究方向为网络与信息安全。