基于SVM的百度新闻热搜词风险分类研究①

胡　洋，唐锡晋

（中科院 数学与系统科学研究院，北京 100190）

摘　要：网络文本挖掘在地震预测、突发事件处理等领域得到了广泛应用，其中文本分类是研究十分重要的一部分。已有的网络文本分类研究主要分析新兴网络媒介如微博、消费点评网站等内容，本文选择百度热搜词进行风险分类研究，为社会风险的及时探测提供了新方法。实验结果证明，对于百度热搜词的多风险分类问题，利用SVM(Support Vector Machine，支持向量机)技术能够取得较高的准确度。

关键词：百度热词；SVM；多风险分类

SVM-Based Risk Categorization of Baidu Hot News Words

Hu Yang, Tang Xijin

(Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Web text mining is a popular issue being applied in areas such as earthquake warming and terrorism detecting, amongst which text categorization is a critical part to analyzing web text and research on social wicked problems; While previous researches focus on social media such as Weibo, consumer review site, etc, our research uses Baidu hot news words as a new method for timely detection of societal risk. Our experiment results show that the classification of multi-risk categories on Baidu hot news words based on SVM can achieve satisfying precision.

Keywords: Baidu hot news words; SVM; multi-risk category

① 基金项目：国家重点基础研究发展计划项目(2010CB731405)国家自然科学基金(71171187)

1 引言

高速发展的网络技术在加快信息传播的同时，也为研究社会问题提供了新方向。一方面，用户可以方便快捷地获得信息；另一方面，用户的网络行为如搜索、发微博等记录也成为社会问题的研究资料。

对于微博等新兴网络媒介的网络文本挖掘主要通过处理短文本，同时利用附带信息如URL、地理位置等，进行群体情绪分析、地震预测及恐怖袭击监控等研究[1-3]。对于搜索引擎(Google等)的网络文本挖掘则通过分析相关搜索记录进行流感预测等研究[4]。我们对百度热搜词的风险分类研究正是结合了两者，一方面，百度作为目前国内最大的中文搜索引擎，每天处理高达亿万次的搜索请求，因而其提供的百度热搜词能够反映整体搜索数据的即时性；另一方面，应用文本分类等社交媒体网络文本挖掘的办法，可以反映整体的社会风险水平。

2 百度热搜词新闻正文抽取

我们选取百度新闻页面(http://news.baidu.com)作为抓取对象，通过java网络爬虫每隔一小时抓取新闻热搜词，然后按照每小时新闻热搜词的排序给予0～20的热度得分，将24小时热搜词及其热度分值汇总后抓取全天热搜词对应百度搜索结果的第一页，通过JSP页面用户可以查看指定时间段的热词，搜索包含关键词的热搜词以及利用iView、CorMap技术分析话题[5]。图1是百度热搜词各个页面之间的对应关系，百度新闻热搜词每5分钟更新一次，每一条热搜词链接的新闻搜索结果的第一页提供5～20条对应新闻网页的链接。我们将详细说明对百度热搜词搜索结果第一页的新闻网页正文的抽取工作。

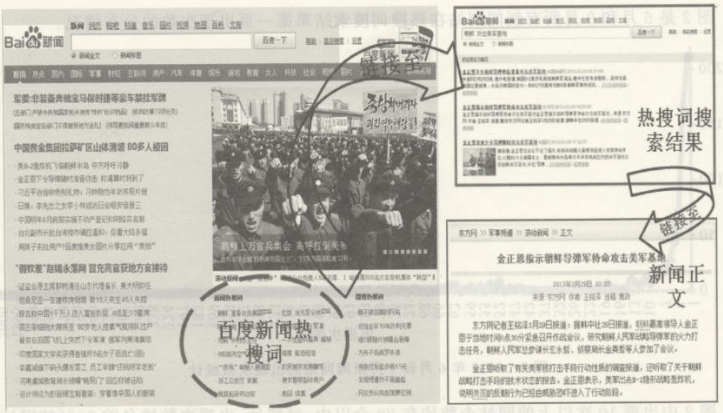


图1 百度热搜词各个页面之间的对应关系

2.1 百度热搜词新闻网站统计

由于百度热搜词的新闻来源网站众多，各大门户网站如新浪(http://www.sina.com.cn)等拥有多家子网站如新浪娱乐(http://ent.sina.com.cn/)、新浪体育(http://sports.sina.com.cn/)等，而且这些门户网站会经常改动网页设计，所以对所有网站分别进行模板式抽取无法确保程序的稳定运行，因此我们首先统计了百度热搜词搜索结果第一页上的新闻来源网站。表1是2012年6月和9月均排名前10的新闻网站及贡献的新闻条目数，可以看到在这两个月前10的网站没有变化。

表1　2012年6月和9月排名前10的网站

新闻网站	6月排名(出现次数)	9月排名(出现次数)
人民网(www.people.com.cn)	5(536)	1(1 134)
搜狐(www.sohu.com)	1(1 224)	2(1 108)
21CN(www.21cn.com)	9(355)	3(715)
凤凰网(www.ifeng.com)	2(745)	4(669)
第一视频(www.v1.cn)	6(444)	5(665)
财讯(www.caixun.com)	10(314)	6(663)
新浪(www.sina.com.cn)	4(719)	7(538)
和讯网(www.hexun.com)	3(735)	8(531)
山东新闻网(www.sdnews.com.cn)	7(440)	9(497)
网易(www.163.com)	8(375)	10(422)

图2是6月和9月所有新闻网站在热搜词搜索结果第一页出现次数的分布情况。

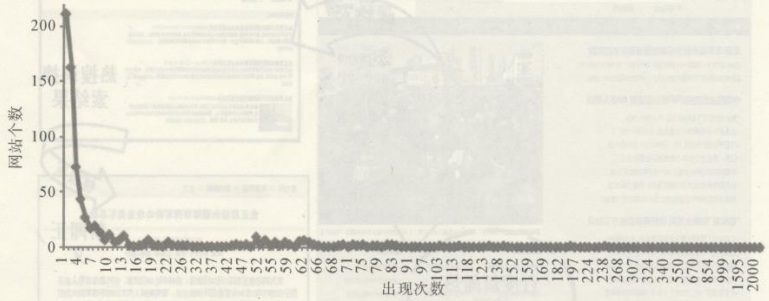


图2　2012年6月和9月新闻网站的出现次数统计

图2中出现10次以上的网站个数均在20个以内，经统计，出现次数排名前20%的网站提供了超过80%的百度热搜词搜索结果第一页的新闻，最终我们将分别在这两月排名前20%的共138个新闻网站作为抽取正文程序的处理对象。

2.2 基于通用正则方法提取新闻网页正文

为了获取新闻网页的正文，我们使用了基于通用正则方法的抽取程序，抽取过程如下：

输入：
S_i：所有百度热搜词搜索结果第一页的新闻网页集
循环：
对于 S_i 中单个 html 网页 x：
抽取 x 中的所有 div 块；
去掉含中文字符数少于阀值(默认为 300)的 div 块；
选取余下 div 块中含中文字符数比例最高的 div 块 y；
利用正则表达式抽取 y 中的正文
循环结束

我们将抽取后的正文存成 xml 格式的文件，用日期和新闻 id 命名如“YYYY-MM-DD/id.xml”。xml 中的各条新闻正文包括 6 个子项，分别是新闻标题 title、新闻网页地址 link、网站名称 site、发表日期 date、对应的百度热搜词 id 以及抽取新闻网页正文 txt。

〈item〉
〈title〉传金正恩下令导弹部队待命攻击美军基地〈/title〉
〈link〉http://news.hexun.com/2013-03-29/152626268.html? from=rss〈/link〉
〈site〉和讯〈/site〉
〈date〉2013-03-29 07:00:00〈/date〉
〈id〉13〈/id〉
〈txt〉环球外汇 3 月 29 日讯在美国连续两次向出动 B2 隐形轰炸机飞往韩国，向朝鲜当局展现军力后，据路透社报导，朝鲜领导人金正恩周五(3 月 29 日)在紧急会议中命令国内导弹部队随时待命，准备攻击韩国和太平洋(601099，股吧)的美军基地。〈/txt〉
〈/item〉

2.3 百度热搜词风险人工标注

在抓取每日百度热搜词后，我们将 2011 年 11 月至 2012 年 10 月的百度热搜词对应的社会风险进行了人工标注。已有的国内关于社会风险预警系统的研究将风险指标分为警源、警兆和警情三个层次，通过理论分析、专家调查和层次分析法的运用对指标的完备合理性进行扩充[6-8]。我们采用了中科院心理所郑蕊博士根据 2008 年奥运会前进行社会问卷调查研究而提出的社会风险分类体系，该体系将社会风险分为国家安全、金融经济、精神文明、日常生活、社会稳定、政府执政、资源环境 7 大类别，并分别细分为恐怖主义与邪教、台湾问题、政治稳定、国家安全与对外关系、重大活动、金融问题、经济问题、伦理道德、信用与诚信、社会风气与精神文明建设、医疗与疾病、教育、就业、物价上涨、交通问题、食品与药品安全、住房、假冒伪劣、重大传染性疾病、贫富与城乡差距、安全生产、治安犯罪与群体性事件、三农问题、贪污腐败、政府执政能力、社会体制与法制建设、社会福利与保障、自然灾害、人口问题和能源短缺与环境污染共 30 个风险小类[9]。标注时我们在风险大类和风险小类里分别添加了无风险这一类别。基于对百度热搜词社会风险类别的人工标注，我们对社会风险水平能够直观认知。图 3 是 2011 年 11 月至 2012 年 10 月的社会风险水平，我们将各月被标注为有社会风险类别的百度热搜词所占频率总和除以当月百度热搜词的频率总和作为该月的风险水平。

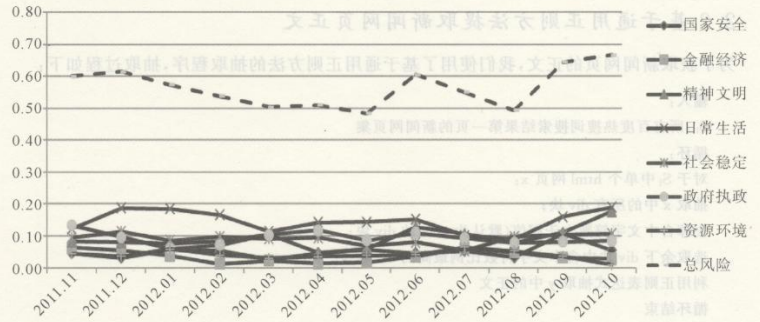


图 3　2011 年 11 月至 2012 年 10 月百度热搜词社会风险水平

图 3 中 7 大类别社会风险的风险水平均低于 0.2，其中日常生活相关的社会风险高于其他类别的社会风险，这说明日常生活相关的社会问题更多引发人们的关注。在总体风险水平上，在 2011 年 11 月至 2012 年 2 月及 2012 年 9 月至 10 月社会风险维持在 0.5 以上的较高水平，2012 年 3 月至 2012 年 8 月社会风险处于较低水平，2012 年 6 月社会风险水平有较大的提升，其中夏季地铁女性安全及着装讨论的新闻获得了更多关注，而这类新闻热搜词主要归类为精神文明相关的社会风险。在 2012 年 8 月风险水平则有较大的下降，主要原因在于人们更多地关注伦敦奥运会赛事，而与此相关的新闻热搜词普遍是无风险的。

通过对百度热搜词的实时社会风险分类，我们可以跟踪整个社会的即时舆论关注，分析搜索引擎用户所体现的社会风险水平，这为社会舆论的及时引导和社会风险的有效防控提供了重要信息。

以上百度新闻热搜词的社会风险标注均为人工处理，利用下文的 SVM 机器分类学习，我们对上述人工标注的百度新闻热搜词进行机器分类预测实验。

3　SVM 机器分类学习

SVM 和 KNN 算法、决策树、人工神经网络、朴素贝叶斯算法及 logistic 回归等在文本分类领域得到了广泛运用[10]。我们采用了 SVM 算法进行机器分类预测。图 4 是利用 SVM 算法进行文本机器分类预测的过程。

具体过程如下：

学习过程：

(1)将文本进行切词生成初始词典，这里的切词采用了 mmseg 正向最大匹配算法并过滤掉常用中文停用词，停用词表采用哈工大提供的中文停用词表(http://www.datatang.com/data/13281/)；

(2)选取初始词典中各类别卡方值前 40%的词生成最终词典；

(3)根据词典将语料文本转化为文本向量，文本向量的各个维度值采用所对应词在该文本中的 TFIDF 值；

(4)将文本向量作为训练样本进行 SVM 训练得到分类模型。

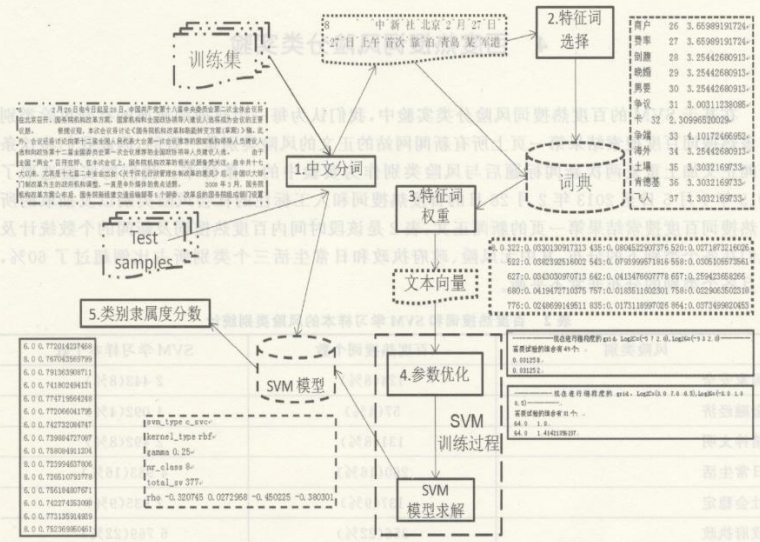


图 4　基于 SVM 的文本分类预测过程

预测过程：

(5)根据上步生成的最终词典，我们将待分类文本转化为文本向量输入；

(6)利用上步中训练得到的 SVM 模型进行分类预测。

在多类别分类问题中，样本在类别间的分布严重不平衡，以及百度热搜词中特征词语义在不同时间段存在的迁移性对于分类的准确率都有很大的干扰。以“刘翔”为例来说明特征词语义的迁移性，在 2012 年伦敦奥运会前，相关的百度热搜词主要是关于他脚伤的新闻，因此我们将这些百度热搜词的社会风险归类为日常生活大类下的医疗与疾病。而他在奥运会上跌倒后，更多对于他背后运营团队欺骗舆论的质疑成为百度热搜词，这时我们将他的社会风险类别归入精神文明下的信用与诚信。为了减小这些问题的干扰，我们使用了类别隶属度分数来确定 SVM 学习预测的结果。类别隶属度分数的公式如下：

$$score = \frac{\sum S_i}{2k} + \frac{k}{2n} \tag{1}$$

这里 S_i 和 k 是将样本分为某一类别的分类器的打分和分类器的个数，n 是全体类别的个数。由于 SVM 算法采用了“一对一”的分类策略来处理多分类问题，因此共有 C_n^2 个分类器。我们可以理解公式(1)为，将样本分为某一类别的分类器打分越高，个数越多，这个样本属于该类别的可能性就越大。最终我们选取类别隶属度分数最高的类别作为该样本的预测类别。

4 百度热搜词风险分类实验

在基于SVM的百度热搜词风险分类实验中，我们认为每一个热搜词所属的社会风险类别与该热搜词百度搜索结果第一页上所有新闻网站的正文的风险类别是一致的，因此我们将这条新闻正文加上重复两次新闻标题后与风险类别作为实验中的一条样本。我们的实验选取了2013年1月5日至2013年2月28日的百度热搜词和人工标注的社会风险类别以及抽取的所有热搜词百度搜索结果第一页的新闻正文，表2是该段时间内百度热搜词及新闻的个数统计及他们在各个类别下的分布，其中无风险、政府执政和日常生活三个类别所占比例超过了60%，而且各个类别的分布非常不平衡。

表2 百度热搜词和SVM学习样本的风险类别统计

风险类别	百度热搜词个数	SVM学习样本个数
国家安全	124(8%)	2 443(8%)
金融经济	57(4%)	1 092(4%)
精神文明	131(8%)	2 492(8%)
日常生活	260(16%)	4 983(16%)
社会稳定	137(9%)	2 635(9%)
政府执政	356(22%)	6 769(22%)
资源环境	81(5%)	1 555(5%)
无风险	459(29%)	8 863(29%)
总计	1 605	30 832

我们进行了10组实验，使用不同时间段的热搜词作为训练样本进行SVM分类模型学习，其中第一组采用前一天的热搜词分类作为学习样本，第二组则采用前两天的热搜词分类，其余依此类推。图5是2013年1月5日至2013年2月28日百度热搜词出现情况的统计。图5(a)是所有热搜词出现次数的统计，共有261个热搜词出现一次，647个热搜词出现两次，8个热搜词出现3次，5个热搜词出现4次，1个热搜词出现6次，出现3次以上的热搜词不到总数的5%。图5(b)是所有热搜词连续出现情况的统计，超过70%的热搜词连续出现2天或以上。图5(c)是每天热搜词在前N天连续出现情况的统计，每天近40%的热搜词在之前连续出现。因此我们采用前N天(N=1,2,3,…)的数据作为训练样本来对百度热搜词进行社会风险类别预测。图6是2013年1月5日至2013年2月28日的每日百度新闻热搜词搜索结果第一页上新闻正文的条目数。

我们采用文本分类中的正确率定义，公式为：

$$precision=\frac{|S_{L(SVM)=L(Manual)}|}{|S_{sample}|} \quad (2)$$

其中$S_{L(SVM)}=L(Manual)$是预测结果中SVM机器分类与人工标注结果一致的数据集，S_{sample}是实验中预测样本的数据集。我们将每一组实验中对每天学习的准确率进行了平均得到最后的平均分类准确率，为了统一比较，我们将2013年1月15日至2013年2月28日的每日样本预测正确率进行了平均。在十组实验中，采用前10天的数据进行训练得到的平均准确率最高，达到了74%。表3列出了各组的平均准确率。

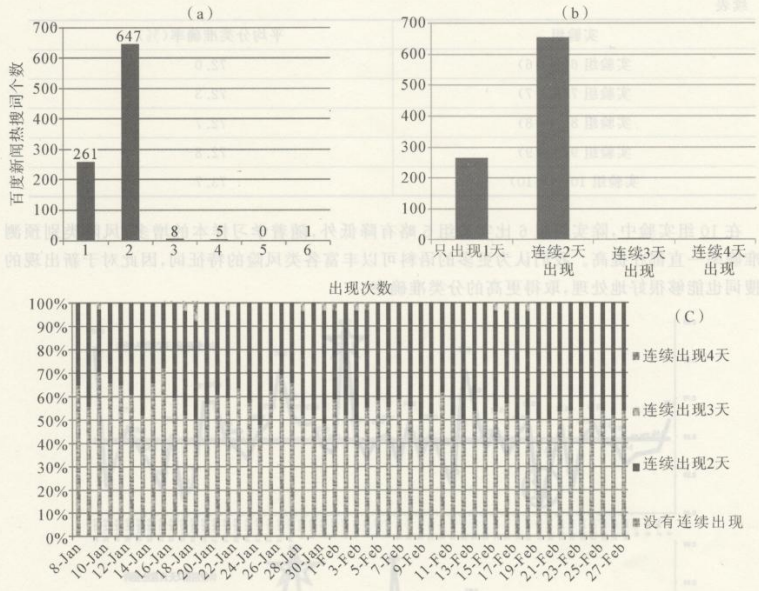


图5 百度热搜词出现情况统计(2013年1月5日至2013年2月28日)

图6 每日SVM学习样本个数(2013年1月5日至2013年2月28日)

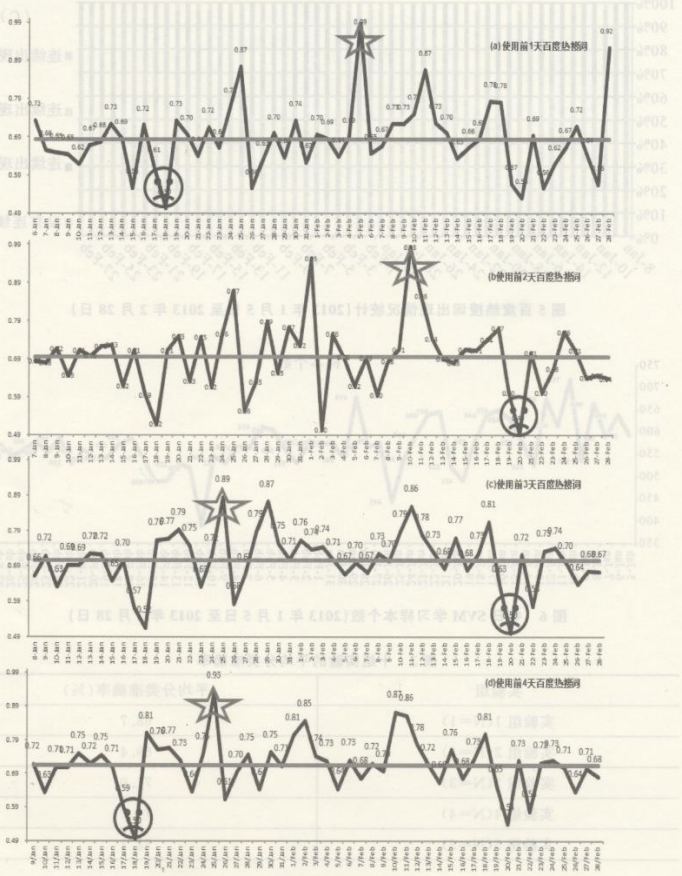
表3 十组实验的平均分类准确率

实验组	平均分类准确率(%)
实验组1(N=1)	68.7
实验组2(N=2)	69.4
实验组3(N=3)	71.0
实验组4(N=4)	71.7
实验组5(N=5)	72.1

续表

实验组	平均分类准确率(%)
实验组 6(N=6)	72.0
实验组 7(N=7)	72.3
实验组 8(N=8)	72.7
实验组 9(N=9)	72.8
实验组 10(N=10)	73.7

在 10 组实验中,除实验组 6 比实验组 5 略有降低外,随着学习样本的增多,风险类别预测的准确率一直都在提高。我们认为更多的语料可以丰富各类风险的特征词,因此对于新出现的热搜词也能够很好地处理,取得更高的分类准确率。



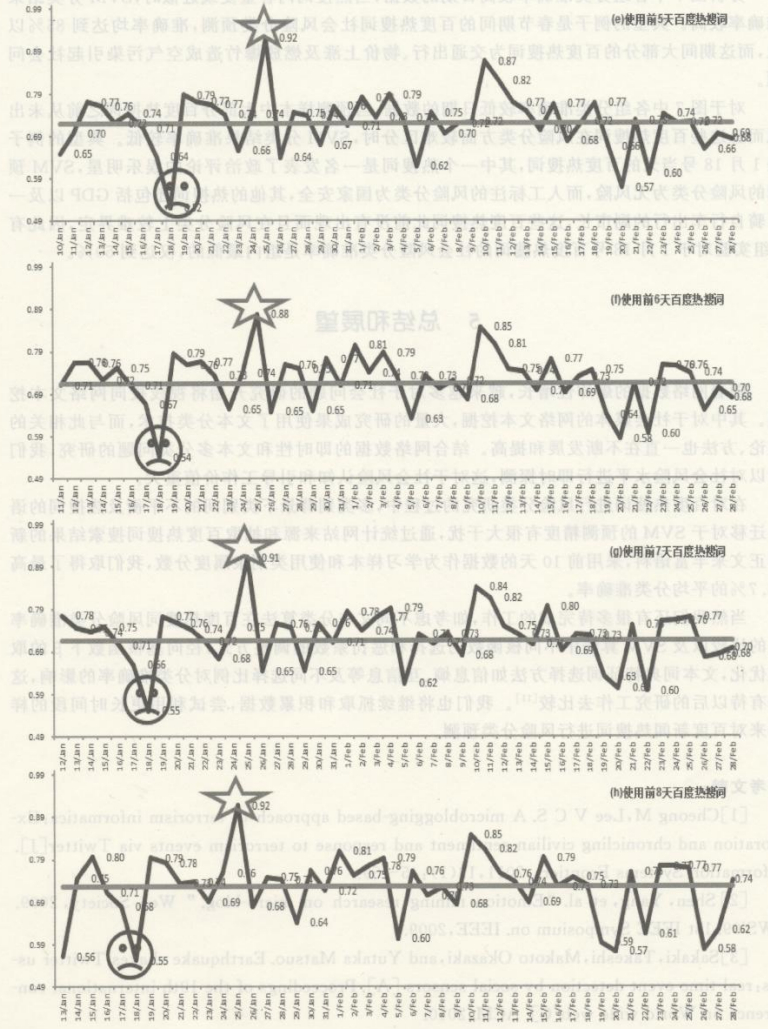


图7 10组实验中每日的分类准确率

图7是10组实验中每天的分类准确率情况。对于1月25号样本的预测，3至10组都取得了组内最高分类准确率，1组和2组分别在2月5号和2月10号是组内最高。4至10组和1组在1月18号是组内最低，2组和3组在2月20日的预测取得的分类准确率为组内最低。

分析图7中各组分类准确率较高日期的数据，当热搜词内容重复或近似时，SVM分类结果准确率较高。典型的例子是春节期间的百度热搜词社会风险分类预测，准确率均达到85%以上，而这期间大部分的百度热搜词为交通出行、物价上涨及燃放爆竹造成空气污染引起社会问题。

对于图7中各组分类准确率较低日期的数据，当预测样本中大部分百度热搜词之前从未出现而且这些百度热搜词在风险分类方面较难区分时，SVM分类结果准确率较低。典型的例子是1月18号当天的百度热搜词，其中一个热搜词是一名发表了政治评论的娱乐明星，SVM预测的风险分类为无风险，而人工标注的风险分类为国家安全，其他的热搜词还包括GDP以及一名骑自行车出行的副市长，这些百度热搜词此前没有出现而且在风险分类上较难界定，因此有8组实验对于1月18号百度热搜词的社会风险分类准确率是组内最低的，仅达到50%。

5 总结和展望

随着网络数据的爆炸性增长，越来越多对于社会问题的研究开始将视线投向网络文本挖掘。其中对于社会媒体的网络文本挖掘，大量的研究成果使用了文本分类技术，而与此相关的理论、方法也一直在不断发展和提高。结合网络数据的即时性和文本多分类问题的研究，我们可以对社会风险水平进行即时探测，这对于社会风险认知和引导工作价值重大。

在对百度热搜词进行风险分类研究的过程中，多类别背景下类别间的不平衡及热搜词的语义迁移对于SVM的预测精度有很大干扰，通过统计网站来源和抽取百度热搜词搜索结果的新闻正文来丰富语料，采用前10天的数据作为学习样本和使用类别隶属度分数，我们取得了最高73.7%的平均分类准确率。

当然我们还有很多待完成的工作，如考虑不同文本分类算法在百度热搜词风险分类准确率上的比较以及SVM算法中不同核函数的选择和惩罚系数的调整方式，径向基核函数下β的取值优化，文本词典特征词选择方法如信息熵、互信息等及不同选择比例对分类准确率的影响；这都有待以后的研究工作去比较[11]。我们也将继续抓取和积累数据，尝试利用更长时间段的样本来对百度新闻热搜词进行风险分类预测。

参考文献：

[1]Cheong M, Lee V C S. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter[J]. Information Systems Frontiers, 2011, 13(1): 45－59.

[2] Shen, Yang, et al. "Emotion mining research on micro-blog." Web Society, 2009. SWS'09. 1st IEEE Symposium on. IEEE, 2009.

[3]Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors [A]. Proceedings of the 19th international conference on World wide web[C]. ACM, 2010.

[4]Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2008, 457(7232): 1012－1014.

[5]吴丹，唐锡晋. 百度新闻热词初析[A]. 系统科学与管理科学新理论、新方法、新技术及其应用——第十一届全国青年系统科学与管理科学学术会议暨第七届物流系统工程学术研讨会论文集[C]. 武汉：武汉理工大学出版社，2011：478－483.

[6]朱恒民，苏新宁，张相斌. 互联网舆情演化的动态网络模型研究[J]. 情报理论与实践，2010，10：75－78.

[7]曾润喜. 网络舆情突发事件预警指标体系构建[J]. 情报理论与实践，2010，33(1)：77－80.

[8]宋林飞. 中国社会风险预警系统的设计与运行[J]. 东南大学学报，1999，1(1)：69－76.

[9]Zheng，R.，Shi，K.，Li，S.. The Influence Factors and Mechanism of Societal Risk Perception [A]. Proceedings of the First International Conference on Complex Sciences：Theory and Appli-cations [C]. Springer，LNICST5，2009：2266－2275.

[10]Sebastiani，F.. Machine learning in automated text categorization [J]. ACM Computing Surveys(CSUR)，2002，34，1－47.

[11]Zhang，W.，Yoshida，T.，Tang，X. J.. Multi-word extraction from Chinese text collection [A]. Proceedings of APWeb'2008 Workshops(Y. Ishikawa et al. eds.) [C]. Springer-Verlag，LNCS 4977，2008：42－53.