doi :10.3969/j.issn.1671-1122.2014.11.006

基于随机森林算法的网络舆情文本信息分类方法研究

吴坚[1,2]，沙晶[3]

（1. 浙江大学计算机学院，浙江杭州 310058; 2. 浙江省公安厅网警总队，浙江杭州 310009;
3. 公安部第三研究所，上海 200031）

摘 要：面对海量增长的互联网舆情信息，对这些舆情文本信息进行分类成为一项非常有意义的任务。首先，文章给出了文本文档的表示模型及特征选择函数的选取。然后，分析了随机森林算法在分类学习算法中的特点，提出了通过构建一系列的文档决策树来完成文档所属类别的判定。在实验中，收集了大量的网络媒体语料，并设定了训练集和测试集，通过对比测试得到了常见算法（包括 kNN、SMO、SVM）与本算法 RF 的对比量化性能数据，证明了本文提出的算法具有较好的综合分类率和分类稳定性。

关键词：网络舆情文本；随机森林算法；文档决策树；文档分类

中图分类号：TP309 **文献标识码：**A **文章编号：**1671-1122（2014）11-0036-05

中文引用格式：吴坚，沙晶．基于随机森林算法的网络舆情文本信息分类方法研究 [J]. 信息网络安全，2014，（11）：36-40.

英文引用格式：WU J, SHA J. The Method of Classifying Network Public Opinion Text Based on Random Forest Algorithm[J]. Netinfo Security,2014,(11):36-40.

The Method of Classifying Network Public Opinion Text Based on Random Forest Algorithm

WU Jian[1,2], SHA Jing[3]

(1.*College of Computer Science and Technology, Zhejiang University, Hangzhou Zhejiang* 310058, *China;*
2. Zhejiang Province Public Security Department, Hangzhou Zhejiang 310009, *China*; 3.*The Third Research Institute of the Ministry of Public Security, Shanghai* 200031, *China*)

Abstract: Faced with massive growth of Internet public opinion information, it's very meaningful to classify these public opinion text information. First of all, this paper established the model of text document representation and selection of feature selection function. Then, it analyzed the characteristics of random forest algorithm in classification learning algorithm, and proposed to complete a series of document category by constructing decision tree. In the experiments, it collected a large number of network media corpora, and set the training and test, the common algorithm is obtained by contrast test (including the kNN, SMO, SVM) compared with the algorithm of RF quantitative performance data, this paper demonstrated that the proposed algorithm has better comprehensive classification rate and the stability of classification.

Key words: network public opinion text; random forest algorithm; document detection tree; document classification

收稿日期：2014-09-18

基金项目：国家科技支撑计划 [2012BAH95F03]

作者简介：吴坚（1980-），男，浙江，硕士研究生，主要研究方向：网络信息安全、数据挖掘；沙晶（1974-），男，上海，副研究员，硕士，主要研究方向：网络信息安全。

通讯作者：吴坚 zjga_wj@163.com

0 引言

近年来，我国网络舆情活跃，互联网应用和用户都在急速增加，根据中国互联网络信息中心的第 33 次互联网发展状况调查报告，截止 2013 年 12 月，中国网民规模达 6.18 亿、中国手机网民规模达 5 亿[1]。网络信息内容复杂多样，既有大量进步、健康、有益的信息，也有不少消极、迷信、不健康的内容。由于互联网使用方式的虚拟性、随意性、发散性，越来越多的人们都使用这种渠道来表达一些个人想法，因此，网络舆情的爆发性增长将是一个不可逆转的趋势[2]。

网络舆情出现的载体形式主要有：微博、博客、即时通讯平台、电子公告板、博客、微信等等。网络舆情通常反映的是一些人民群众生活中关心的问题，甚至是一些引起较大社会负面影响的事件，因此必须加以科学和严肃的对待。网络舆情的种类日益丰富、频次不断增加，事件关联性、次生危机衍生性特点日趋明显。当某突发事件一旦被网络媒体或网民曝光，在较短的时间周期内会立即引起社会聚焦，相应报道被反复转载、快速传播，形成网络舆情。一些重要的热点事件发生后，人们总能发现与之对应的网络舆情。在网络中出现关于事件的分析、评论及众多网络群众的诉求、抗争等舆情已是网络中的常态现象和重要景观。在这一背景之下，应对复杂多变的网络舆情已然成为检验政府执政能力水平的一个重要标尺[3]。

正因如此，互联网管理部门应该更加重视网络舆情的分析及应用，探索新形势下更好的网络舆情的引导方法[4]。文本信息是网络舆情的最主要载体形式之一，因此，对于网络舆情信息中文本信息的分类本质上还是属于一般的文本分类问题。文本分类的主要步骤有：文本模型的建立、特征的选取、分类器的选择和训练、分类结果和评价。本文以文本形式的网络舆情数据为研究对象，重点研究利用随机森林分类方法完成文本形式的网络舆情数据分类。

1 文本文档的表示模型及特征选择

为了使用机器学习方法完成网络舆情文本信息的分类，必须首先设计好文本文档的表示模型。通常情况下，会采用向量表示法来描述某个特定文本；此外，在考虑文本语义的所属层次时，仅仅考虑词汇语义信息[5]。设待处理的文本为 d_j，T 为文本特征集合，在具体实现中，第一个步骤是选定文本的特征；第二个步骤是用向量描述每一个特征的权重，可以表示为：

$$\vec{d}_j = < w_{1j}, w_{2j}, ..., w_{Tj} > \quad (1)$$

该向量中，每一个权重表示的是对应的特征对于该文本语义的贡献度[6]。在特征向量的构造过程中，有两个非常重要的因素：

1）特征的选取方法。通常，都是以词语为单位进行特征的选取，但是，已经有大量实验证明：仅仅使用词语进行特征的提取存在较多不足，因而，不少学者提出了使用短语进行特征的提取[7]。然而，仅仅使用短语完成特征的提取在实际数据的验证中并没有取得较为理想的结果。因此，有学者建议将两者结合起来，即同时考虑词语和短语[8]。

2）权重的计算方法。目前主流的计算方法有两种，分别是非二进制权重和二进制权重。前者是用 0~1 的一个小数来表达权重，而后者是用离散的数字 1 和 0 来表达权重。对于前者，学术界采用的最多的就是 Salton 和 Buckley 提出来的特征频率 - 反转文档频率函数，其表达形式如下[9]：

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#Tr(t_k)} \quad (2)$$

在公式（2）中，$\#(t_k, d_j)$ 表示的是特征项 t_k 在文档 d_j 中出现的次数，$\#T_r(t_k)$ 表示的是在所有文档中，凡是特征项 t_k 出现过的文档的数量之和。该公式所表达的含义有两点：（1）如果某个特征在某个文档中出现次数越多，则它能更好表达其内容；（2）如果某个特征在文档中出现次数越多，则它的分辨能力越差。为了使得权重值落入 [0,1] 区间，以及所有文档的权重向量具有相同长度，通常会对公式（2）中进行余弦规范化处理，得到：

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))^2}} \quad (3)$$

文本文档的特征维度通常会非常大，无法用分类器实现，因此，必须通过降维方法选择出有效的特征子集 T'。特征子集的维度远远小于原始特征集。降维方法的实施有益于改善分类器的过度拟合问题，即：提取出本质特征，抛开非本质特征。这是因为：虽然分类器对于已经训练过的数据的过度拟合可以改善分类性能，但是，对于新增的训练数据，其分类效果会很差[10]。降维过程中，可能也有一些包含了有用信息的特征被删除掉，因此，必须慎重对

待这一过程。基于信息论的特征选择方法是目前常用的方法之一，在此列举出常用的特征函数[11,12]，见表1。

表1 常用特征选择函数计算公式

特征函数	符号表示	数学表达式
信息增益	$IG(t_k, c_i)$	$\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}} P(t,c)\log\frac{P(t,c)}{P(t)P(c)}$
互信息	$MI(t_k, c_i)$	$\log\frac{P(t_k,c_i)}{P(t_k)P(c_i)}$
卡方信息	$\chi^2(t_k,c_i)$	$\frac{\lvert Tr\rvert^2[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
相关系数	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert Tr\rvert}[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]}{\sqrt{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}}$
相关分数	$RS(t_k, c_i)$	$\log\frac{P(t_k\mid c_i)+d}{P(\bar{t}_k\mid\bar{c}_i)+d}$
文档频次	DF	$N_{f,c}$

在此，对常用的特征选择函数做出简单解释：

1）信息增益 $IG(t_k,c_i)$。表示的是特征项 t_k 在训练集样本中出现的概率。IG 的值越高，对分类预测提供的信息就越多。通过设定阈值，可以将 IG 值小于阈值的特征词删除掉，从而降低特征空间的维数。

2）互信息 $MI(t_k,c_i)$。特征词条和类别的互信息有多种不同的计算方法，但总体上而言，一般都会选取互信息量最大的名词作为特征词。原因如下：该类词在某个类别中的出现概率较大，而在其他类别中出现的概率小，互信息量越大，名词和类别之间同时出现的概率也越大。

3）卡方信息。卡方信息对应的值越大，表示特征词和类别之间的相关性越强，也就是说，在训练文本集中包含特征词的文本属于某个类别的概率越大，反之亦然。

不同的特征选择函数考虑问题的角度不同，已经有大量实验表明，这些特征选择函数能够极大的降低特征维度。但是，通过有限的实验结果还是无法断定哪一个特征选择函数性能更加优异，这是因为现有的实验都是基于某种特定的分类器、语料库之上的[13-15]。因此，必须在特定的问题中做一些具体分析和改进。

2 文本文档特征提取

通过文本文档的特征选择可以完成文本文档的维度降低，但是，在文本文档分类中还存在着词语的多义性、同音多义、近义词等问题。仅仅通过筛选得到的特征函数集不一定是文档信息的最优表达，因此，有必要对原始特征函数进行进一步的综合，得到新的特征函数集，使之能够更好的表达文档内容，同时可以降低词语的多义性等问题的影响。通过原始特征综合得到的新特征的过程被称为特征提取。特征提取方法一般分为两大类：对原始特征函数进行提取的方法，又称为特征聚类；对原始特征函数进行变换的方法，又称为潜在语义索引。

1）特征聚类的主要目的是将具有较高语义关联度的词语归并到一起，用归并之后的集合代替单个特征。特征聚类与特征选择是不同的，具体体现在：前者主要用于解决特征的同义性、近义性问题，后者主要用于解决特征的冗余度问题。

2）潜在语义索引的主要目的是通过对原始特征的组合实现文档的特征向量压缩，使之成一个较低维度的空间。在具体实现中，该方法从语料库中推导出原始特征之间的依赖关系。通过对原始特征向量构成的特征矩阵进行奇异值分解，从而实现从原始特征向量到新特征向量的映射。

3 网络舆情信息的分类算法

首先，给出一般形式的文本分类算法描述方法。为了描述文本文档所属类别，预先定义一个类别状态函数 CSV，对于一个给定的文档 d_j，该函数为一个 $D\rightarrow[0,1]$ 的映射，其值描述了该文档所属类别的状态值，或者说，该函数成为了衡量 $d_j\in e_i$ 的指标。一般情况下，设类别集为 $C=\{c_1,c_2,\cdots,c|C|\}$，可通过计算单个文档的 CSV 值来判定它所属的类别。对于不同的学习方法，其 CSV 的定义不同。本节将以网络突发事件舆情信息分类为研究对象，突发事件对于社会的影响是较大的，对于网络突发事件舆情信息进行有效的分类可以为处置预案的制定及实施提供较大帮助。有学者指出：对于突发事件的正确分类对于科学地制定应急预案具有指导作用，突发事件宏观上可分为基本突发事件和非基本突发事件。其中，基本突发事件分为自然性和社会性突发事件；非基本突发事件则按照发生的空间范围、事件发生和终结的速度、诱发的原因等因素分为相应的类别。本节，将突发事件按照公共安全的四类突发事件，即：突发自然灾害、事故灾难、公共卫生事件、社会安全事件进行分类[16]。

3.1 网络舆情关键词和语料库的获取

本节首先选取了突发事件集中的高频词集，能在一定

程度上反映某个独立的网络突发事件所形成的短语。结合突发事件的分类，限于篇幅，在此简单给出部分关键词列表。

1）自然灾害类：台风、海啸、地震、山火、泥石流；

2）突发公共卫生事件：假药、过期药、医院、疫情、禽流感、手足口、医闹；

3）突发社会安全事件：上访、聚众、群体、抗议；

4）突发事故灾难：空难、脱轨、相撞、爆炸、坍塌；

之后，分别用关键词列表中的词语在搜索引擎中进行搜索，并筛选出较为有影响力的网站来源的新闻报道，共采集了2012年以来的3730篇文档，构成了突发事件的分类语料库。其中，关于自然灾害类的报道一共398篇；关于突发公共卫生事件的报道一共1092篇；关于突发社会安全事件的报道一共1121篇；关于突发事故灾难的报道一共是1119篇。语料库的各个类别所占比例参见表2。

表2 分类语料库数量列表

	自然灾害	公共卫生事件	社会安全事件	事故灾难
文档数量	398	1092	1121	1119
所占比例	10%	29%	31%	30%

3.2 基于随机森林算法的分类器实现

随机森林算法起源于20世纪90年代，以其优秀的分类效果在众多的机器算法中保持了较强的竞争力。与传统的决策树分类器相比较，随机森林有更好的分类效果和更强的泛化能力。随机森林算法的实质是一个包含多个决策树的分类器，这些决策树的形成过程是完全随机的。随机森林中的树与树之间是没有关联的。让测试数据样本进入随机森林，其本质就是让每一棵决策树进行分类，最后取所有决策树中分类结果最多的那一类作为最终的结果[17]。

作为一种分类算法，随机森林具有以下优点：对于大容量数据，具有较高的分类准确率；与目前其他分类算法相比，随机森林方法能够较好的摒弃噪声干扰的影响；利用大数定律可以得到，随机森林作为有监督的学习方法不容易过拟合；构造分类器时，可以通过单个训练集的袋外数据在内部估计模型的泛化误差；对于不平衡的分类数据集，它可以平衡误差。

结合本文所要研究的问题，首先，把3.1节中建好的语料库作为数据集。其中，随机选取3270个文档作为训练集，剩余的460个文档作为测试集。之后，为突发事件文本文档的分类设计好随机森林分类器。其中，在量化特征的选择方面，一共选用了3个特征，分别是：信息增益、卡方信息、相关系数。算法实现采用了科罗拉多大学开发的randomforest-matlab工具箱，该工具箱已经在诸多文献中得到成功应用。

算法的具体步骤如下：

1）采用Bootstrap方法重采样，随机产生N个文档训练集 $S_1, S_2, ..., S_N$；

2）依据于每个文档训练集，生成对应的文档决策树 $C_1, C_2, ..., C_N$；在每个非叶子节点（内部节点）上选择特征之前，从M个特征中随机抽取m个特征作为当前节点的分裂特征集，并以这m个特征中最好的分裂方式对该节点进行分裂；在本算法中，取 $m=\log 2M+1$；

3）每棵文档树都完整成长，不进行剪枝；

4）对于测试集中的样本Y，利用每棵决策树进行测试，得到对应的类别 $C_1(Y), C_2(Y), C_3(Y), ...C_N(Y)$;

5）采用投票的方法将T个文档决策树中输出最多的类别作为测试集样本Y所属的类别。

分类器算法的流程图如图1所示。

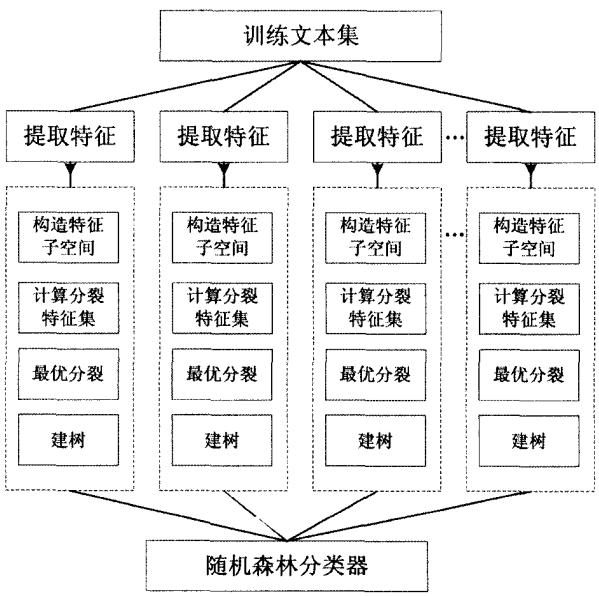

图1 随机森林分类器实现流程

4 实验及结果分析

对于文本分类算法的评价主要考察样本是否被划分到正确的类别中。对于算法性能的考察主要有两个量化指标：

精确率、召回率。精确率表达的是分类器的正确度，召回率表达的是分类器的完整度。此外，为了表达精确率和召回率的相对重要程度，通常会同等对待精确率和召回率指标，得到F1指标，该指标又被称为综合分类率。根据第3部分的数据集设定情况，对3种分类特征进行了测试，结果参见表3。

表3 三种分类特征分类结果比较

	信息增益		卡方信息		相关系数	
	精确率	召回率	精确率	召回率	精确率	召回率
自然灾害类	92.67%	89.54%	90.57%	89.75%	88.87%	87.81%
公共卫生事件	95.12%	90.13%	91.42%	84.69%	90.89%	85.32%
社会安全事件	91.60%	91.33%	90.68%	87.43%	91.23%	89.09%
事故灾难	88.91%	84.65%	86.90%	83.41%	86.56%	84.29%

从表3中可以看出，信息增益特征在分类精度上具有比较好的性能，同时，其召回率相对于其他特征也更好。

为了横向地和其他分类算法进行比较，本节选用了kNN，SMO，SVM方法[10,18,19]，在同一数据集上进行对比测试，其结果见图2。在测试中，分别计算了4个大类的F1值的情况。从图2中可以看出，各个算法不可能在所有指标上都取得最优；SMO和SVM算法的性能比较接近。在宏平均值计算结果中，SVM的性能超过了本文题出的随机森林算法，但是，随机森林算法的宏平均和微平均值都是比较接近的，也就是说，该算法比较稳定。

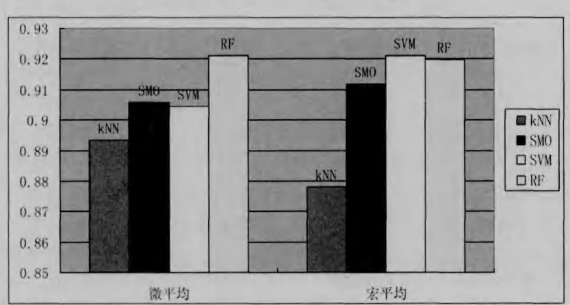

图2 四种算法的综合分类率

5 结束语

本文使用分类效果较好的随机森林算法对网络舆情文本信息的分类进行了研究。在收集大量互联网媒体的语料之后，构建了学习算法所需的训练集和测试集。通过构建文档的决策树来实现分类器。实验结果表明，随机森林算法的综合分类率性能较好，并且分类性能比其他算法稳定。（责编 吴晶）

参考文献：

[1] 中国互联网络信息中心．第33次中国互联网络发展状况统计报告[R], 2014.

[2] 许鑫，章成志，李雯静．国内网络舆情研究的回顾与展望[J]. 情报理论与实践，2009, 32(3): 115-120.

[3] 彭辉，姚颉靖．我国政府应对网络舆情的现状及对策研究——基于33件网络舆情典型案例分析[J]. 北京交通大学学报(社会科学版)，2014, 13(3): 102-109.

[4] 徐厌平，邵梦洁．公共治理视域下中国网络舆情危机及应对研究[J]. 求索，2013, (11): 250-252.

[5] 万源．基于语义统计分析的网络舆情挖掘技术研究[D]. 武汉：武汉理工大学，2012.

[6] Fabrizio Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, 2002, 34(1):1-47.

[7] Maria Fernanda Caropreso, Stan Matwin, Fabrizio Sebastiani, A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization, Text databases & document management, IGI Publishing Hershey, PA, USA, 2001, 78-102.

[8] 余一骄，刘芹．基于语义的中文网页检索[J]. 计算机科学，2012, 39(8): 79-87.

[9] Gerard Salton, Christopher Buckley. Information Processing and Management, 1988, 24(5):513—523.

[10] Busagala L.S.P., Ohyama W., Wakabayashi T., Kimura F., Multiple Feature-Classifier Combination in Automated Text Classification, 2012 10th IAPR International Workshop on Document Analysis Systems, 2012, 43-47.

[11] Norbert Fuhr, Chris Buckley, A probabilistic learning approach for document indexing, ACM Transactions on Information Systems, 1991, 9(3):223-248.

[12] Miguel E. Ruiz, Padmini Srinivasan, Hierarchical neural networks for text categorization, Proceedings of the 22nd annual international ACM SIGIR conference, California, United States, 1999, 281-282.

[13] Caropreso M F, Matwin S, Sebastiani F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. Text databases and document management: Theory and practice, 2001: 78-102.

[14] Galavotti L, Sebastiani F, Simi M. Experiments on the use of feature selection and negative evidence in automated text categorization, Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2000: 59-68.

[15] Hwee Tou Ng, Wei Boon Goh, Kok Leong Low, Feature selection, perceptron learning, and a usability case study for text categorization, Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, 1997, 31(SI): 67-73.

[16] 袁辛奋，胡子林．浅析突发事件的特征、分类及意义[J]. 科技与管理，2005，7(2)：23-25.

[17] Chen Huang, Xiaoqing Ding, Chi Fang, Head Pose Estimation Based on Random Forests for Multiclass Classification, 20th International Conference on Pattern Recognition, Istanbul, 2010, 934-937.

[18] E Wiener. A neural network approach to topic spotting, The 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas: ACM Press, 1995: 317-332.

[19] Abdul-Rahman S., Exploring Feature Selection and Support Vector Machine in Text Categorization, IEEE 16th International Conference on Computational Science and Engineering, Sydney, 2013:1101-1104.