基于文本分类的政府网站信箱自动转递方法研究*

王思迪[1,2]　胡广伟[1,2]　杨巳煜[1,2]　施　云[1]

[1](南京大学信息管理学院　南京 210023)

[2](南京大学政务数据资源研究所　南京 210023)

摘要:【**目的**】为改善政府网站领导信箱传统人工转递方式存在的人力、时间成本较高以及工作人员负担较重等问题,研究网站来信的自动转递方法。【**方法**】选择较有代表性的分类算法,包括朴素贝叶斯、决策树、随机森林以及多层神经网络,对北京、合肥和深圳的市长信箱文本数据进行对比实验,进而设计一套基于文本分类的政府网站信箱自动转递方法,并给出相应的应用建议。【**结果**】神经网络算法在市长信箱文本的分类表现最优,宏平均精确度和召回率均达0.85以上,且所有微平均指标均达0.93以上;朴素贝叶斯算法次之;随机森林算法的宏平均精确度很高,但召回率较差;决策树算法的精确度和召回率都较一般。【**局限**】未能兼顾来信数量不均衡对结果的影响,且实验时剔除了数据量过小的部门的来信数据,这在实际应用中可能会存在一定偏差。【**结论**】本文设计的政府网站信箱自动转递方法能够优化领导信箱运作机制,对提升线上政民互动效率,降低人力及行政成本具有积极意义。

关键词: 领导信箱　自动转递　文本分类　多层神经网络　流程优化

分类号: TP39　G35

DOI: 10.11925/infotech.2096-3467.2019.1182

引用本文: 王思迪, 胡广伟, 杨巳煜等. 基于文本分类的政府网站信箱自动转递方法研究[J]. 数据分析与知识发现, 2020, 4(6): 51-59.(Wang Sidi, Hu Guangwei, Yang Siyu,et al. Automatic Transferring Government Website E-Mails Based on Text Classification[J]. Data Analysis and Knowledge Discovery, 2020, 4(6): 51-59.)

1　引　言

十九大报告中提出"到2035年要基本实现国家治理体系和治理能力现代化",提升公众参与社会治理的程度以及行政回应效率是达到此目标的重要组成部分。公众线下参政议政以及提出诉求建议需要耗费大量的人力物力,而线上政民互动则提供了一条便捷高效的参与渠道。当前线上政民互动的渠道主要包括12345热线、政务微博评论版块、政府网站的领导信箱以及人民网地方领导留言板等[1-2]。领导信箱作为常设性政民互动渠道[1],每天要承接大量群众来信。

梳理各省市的领导信箱工作流程发现,目前除个别城市要求群众来信时选择相应部门(这要求公众对政府部门职能极为了解),大部分是由信访部门或政府办公室承接再根据具体诉求转递至相关责任部门,这大大增加了工作人员的劳动强度,而新兴信息技术在政务领域所形成的"大数据能力"、"智能化应用"、"机器学习方法"等没有得到充分的利用。近年,在各级政府大力推进权力清单体系建设的背景下,政府职能部门之间的权责划分更为明确[3],不同事项的来信语料特点鲜明。机器学习算法具有不断更新的学习能力,可以从大量来信文本中学习特征,进而挖掘不同政府部门之间公众来信的差异性,据此实现来信的自动分类和转递,这将大大提高政府工作效率,提升公众参与的用户体验。

鉴于此,本文致力于研究政府网站中领导信箱

通讯作者: 胡广伟, ORCID:0000-0003-1303-363X, E-mail:hugw@nju.edu.cn。

*本文系国家自然科学基金面上项目"电子政务服务价值共创机制及实现模式实证研究"(项目编号: 71573117)的研究成果之一。

万方数据

的来信自动转递方法。首先回顾特征匹配及自动转递相关应用、文本分类和线上政务信箱等方面的研究，通过信箱文本分类算法，探析利用机器学习算法构建领导信箱自动转递系统的可行性及实施效果，进而设计一套完整的来信自动转递流程，促进线上政民互动的高效化和智能化。

2 相关研究

政府网站信箱自动转递是指公众在政府网站领导信箱提交信件后，由计算机利用算法识别相关责任部门并自动进行转递的过程，其本质是对信件内容进行特征分析并自动将来信转递到最匹配相应特征的接收者。这种通过匹配特征进行自动转递的方式在某些领域已有研究与应用。王珺[4]根据电子档案的文本特征设计自动归类系统。李湘东等[5]利用文本分类方法实现了期刊常设主题栏目的自动归栏。李成铭[6]对人才简历与招聘信息分别进行文本特征提取并将两者进行自动匹配。王若佳等[7]基于在线问诊平台的实际问诊文本数据，利用多种机器学习算法实现智能分诊。

解决这类问题的关键点在于准确的文本分类，为此需选择合适的特征选择方法和分类算法。考虑到传统的文本特征选择算法大多忽略不同类别的相对文档频率，Kim 等[8]提出一种新的特征选择算法，将某个类中经常使用但很少出现在其他类中的词语赋予更高的权重。Ghareb 等[9]提出三种改进的特征选择方法，将其应用于朴素贝叶斯文本分类和关联分析中，并取得了不错的效果。这些方法对政务信箱的文本特征选择有一定的启发作用。常见的文本分类算法包括：朴素贝叶斯、最近邻算法、支持向量机和决策树等单一分类算法；AdaBoost、随机森林等集成算法。已有学者使用多种分类算法进行对比研究，如 Hartmann 等[10]将 5 种基于词汇和 5 种机器学习算法用于社交媒体文本数据分类，发现机器学习算法要优于基于词汇的方法，而且随机森林和朴素贝叶斯要优于支持向量机。此外，随着深度学习的发展，神经网络也越来越多地被用于中文文本分类。田欢等[11]通过选择合适的激活函数、设置最优的参数初始值并引入动量因子构造改进的 BP 神经网络文本分类器，较好实现了学术活动文本的分类。对于分类效果的衡量方式大致分为两种：一是算法研究类，通常选择基本算法作为对比基准，优化后的算法在准确率上要高于基本算法；二是偏重于算法的对比应用，通常选择多种算法在某一特定领域进行应用，并比较优劣。

在梳理当前根据文本特征进行自动匹配的相关研究时，笔者发现类别之间没有清晰的界定是影响分类准确率的一个关键问题，如在线问诊平台的诊室类目存在包含关系[7]、在论文自动分类中交叉学科的分类准确率较低[12]等。这说明分类算法的应用对象应具有清晰的类别，而政府部门间往往权责划分明确[3]，因此理论上可以使用分类算法实现领导信箱的自动分类及转递。

在线上政务信箱相关的研究，目前多偏重于政民互动中的回应性研究[13-14]，也有部分学者聚焦于政务信箱的来信内容研究，如孙宗锋等[1]采用描述性统计、词频分析、情感分析等多种方法对青岛市市长信箱数据进行分析。上述针对领导信箱的回应性研究和内容研究均是对信箱数据的正向利用，从数据中挖掘知识；而利用领导信箱中的历史信件构建来信的自动投递则是对数据的反向利用，将研究结果作用于运作机制本身。Ong 等[15]采用案例分析法对台北市市长信箱的运作机制进行纵向深入的研究，但这种研究仅对现有机制进行分析，未能从根本上提升政府网站信箱的运作机制。

3 研究设计

3.1 研究框架

研究政府网站领导信箱的来信如何基于文本分类进行自动转递，进而设计相应的转递方法，关键在于信箱文本自动分类的效果是否足够好。因此，本文的实验聚焦于信箱文本的部门分类，采集若干城市的市长信箱文本数据进行清洗和预处理，利用多种机器学习算法实现信箱文本的自动分类，然后对比分析不同算法的分类结果，进而找出最适用于政府网站领导信箱部门自动转递的算法，设计政府网站领导信箱的自动转递流程，并针对实验结果进行总结与讨论，提出应对实际应用中可能出现的复杂情况的建议。

万方数据

3.2　算法选择

(1) 特征选择及文本表示方法

常见的文本特征选择算法包括词频、文档频率、词频逆文档频率、信息增益、卡方统计和互信息等。几种方法各有优劣，由于卡方统计考察了词语与类别之间的相关度，且在文本特征选择中有不俗的表现[16]，因此本文使用卡方统计量进行特征选择。

最常用的文本表示方法是向量空间模型(Vector Space Model，VSM)，该模型的主要思想是：将每一个文本都映射为一组规范化正交词条矢量形成的向量空间中的一个点[17-18]。假设现在有一组领导信箱语料，对信箱文本分词并选择特征后得到 n 个特征词，那么向量空间的维度就是 n。对于其中的某一条文本 X，可被表示为一个 n 维向量：$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$。向量维度的计算目前有三种主流的方法：布尔值方法、词频统计法以及 TF-IDF 法。在本文预实验中，词频统计法与 TF-IDF 法的分类准确率表现相差无几，且词频统计的算法复杂度更低，故而选择词频统计法进行计算。

(2) 分类算法

根据文献[10]的研究结果，本文在设计领导信箱的部门自动分类识别上仅考虑机器学习算法，使用朴素贝叶斯(Naïve Bayes，NB)[19]和随机森林(Random Forest，RF)[20]算法进行实验。虽然在信箱文本分类中支持向量机的分类效果不一定逊于其他两种算法，但支持向量机属于二分类方法，在多分类问题中需要将类别二值化，且政府部门分类较多，会大大增加算法的时间成本。

朴素贝叶斯是一种单一分类算法，而随机森林是在决策树基础上发展起来的集成算法，为了更直观地比较单一算法和集成算法，在实验中也考虑使用决策树(Decision Tree，DT)算法[21]作为对比基准。此外，在实验中引入 Hinton[22]提出的基于 BP 反向传播算法的多层神经网络(Multi-Layer Perception，MLP)，探究神经网络在政府网站信箱分类中的应用效果。

4　实验及结果分析

4.1　实验数据

为验证基于文本分类设计领导信箱部门自动转递方法的可行性和有效性，需要对领导信箱的来信进行文本分类。以直辖市和省会城市为例，选择其中公开来信量较大且语料字段完整(包括来信内容、回复内容以及回复部门)的城市，最终选定北京市、合肥市以及深圳市作为分析对象。使用 Python 编写爬虫工具分别获取三个城市原始语料 26 277 条(截至回复时间 2018 年 12 月 29 日)、104 460 条(截至回复时间 2019 年 1 月 10 日)、51 540 条(截至回复时间 2018 年 10 月 31 日)，剔除无效数据后选择信件数多于 100 条的部门，数据集情况如表 1 所示。

表 1　数据集介绍

Table 1　Dataset

城市	部门数(个)	数据量(条)
北京市	16	10 703
合肥市	27	36 142
深圳市	33	37 053

4.2　实验过程

实验整体分为数据预处理、特征表示、使用训练集训练分类器以及使用测试集测试分类器效果 4 个步骤，如图 1 所示。

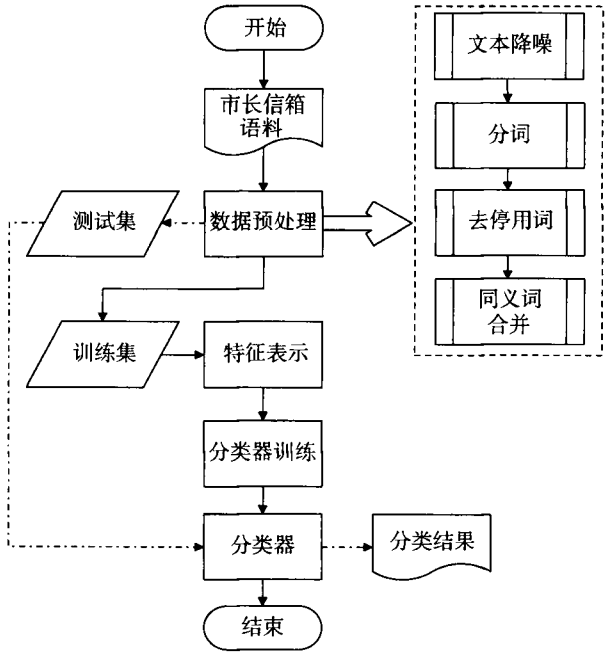

图 1　实验步骤

Fig.1　Experimental Procedure

实验预处理及分类算法均通过 Python(3.6.6 版)编程实现，使用的 IDE 为 PyCharm Community

万方数据

Edition 2019.1.3 ×64，数据预处理主要使用re和jieba开源模块，分类算法使用scikit-learn开源模块实现。

首先对所有语料进行文本预处理。预处理过程包括文本降噪、使用结巴分词进行中文分词、去停用词以及同义词合并。停用词表来源于互联网，是由多个中文停用词表整理并去重后得到，包含1 598个词语。同义词表取自哈尔滨工业大学信息检索研究室同义词词林扩展版①。文本降噪主要是除去信箱文本中特有的无意义格式段落，如“尊敬的市长，您好”等。在特征选择方面，选用的算法为卡方检验，选取卡方统计量前40%的词语作为特征词；在文本表示方面，采用向量空间模型，其中向量空间中的权重使用词频表示。最后在朴素贝叶斯、决策树、随机森林以及多层神经网络4种机器学习算法上进行分类实验和参数调优。

4.3 实验结果及分析

将三个城市每个部门的语料按照8:2的常规比例划分为训练集和测试集，分别在朴素贝叶斯(NB)、决策树(DT)、随机森林(RF)以及多层神经网络(MLP)4种机器学习分类算法上进行分类实验。

(1) 分类效果评价指标

由于三市的市长信箱中各部门数据量不均衡，分类算法会倾向于将数据量较小的类识别为数据量较大的类，因此采用准确率(Accuracy)进行分类效果的衡量可能会有失偏颇。本文采用精确度(Precision)、召回率(Recall)和调和平均值F1作为衡量分类效果的指标，计算方法如公式(1)-公式(3)所示，并分别计算各指标的宏平均值和微平均值。虽然理论上精确度与召回率都是越高越好，但两者在某些情况下存在矛盾。因此需要引入两者的调和平均值进行综合判断，只有精确度和召回率都较大才能保证F1值较高。指标的宏平均值是类的算术平均值，微平均值是数据集中实例的算数平均值。

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

其中，TP表示判断为某个类别的样本中实际也属于该类别的样本数，FP表示判断为某个类别的样本中实际不属于该类别的样本数，TN表示判断为不属于某个类别的样本中实际也不属于该类别的样本数，FN表示判断为不属于某个类别的样本中实际属于该类别的样本数。

此外，将ROC曲线下面积(Area Under Curve，AUC)也作为衡量分类效果的指标之一。ROC曲线的横轴表示负正类率(False Positive Rate，FPR)，计算如公式(4)所示。FPR越大，预测正类中实际负类越多；纵轴表示真正类率(True Positive Rate，TPR)，其计算公式同召回率，TPR越大，预测正类中实际正类越多。

$$FPR = \frac{FP}{TN + FP} \tag{4}$$

理想目标是达到TPR=1，FPR=0，即ROC曲线越靠拢图中(0，1)点，越偏离45度对角线越好。AUC可以直观评价分类器的分类效果，在0~1取值范围内，AUC值越大越好。

(2) 总体分类结果分析

实验结果如表2所示。图2是以深圳市为例4种算法的ROC曲线(北京市和合肥市4种算法的ROC曲线和深圳市类似，不再列出)。可以看出，多层神经网络算法在市长信箱文本的分类表现要优于传统机器学习算法，精确度和召回率的宏平均指标都能达到0.85以上，且所有微平均指标均达0.93以上，分类的ROC曲线也是最优的，AUC值均超过0.99；朴素贝叶斯算法在传统机器学习算法中表现最优，所有指标可达到0.80以上；随机森林算法的宏平均精确度很高，但召回率较差；决策树算法的精确度和召回率都比较一般。同时，从表2可以明显看出微平均指标普遍要高于宏平均指标，说明在各部门样本不均衡的情况下，样本量较小的部门分类效果要差于样本量较大的部门。

(3) 部门分类结果分析——以北京市为例

考虑到每个政府部门间的独立性，以总体分类指标衡量所有部门的分类效果是欠妥当的，因此需

①《同义词词林》由梅家驹等于1983年编纂而成，哈尔滨工业大学信息检索实验室基于该词林进行扩展，完成了《同义词词林(扩展版)》。

万方数据

表 2　分类效果指标数值

Table 2　Classification Performance

算法	分类效果指标	宏平均			微平均		
		北京	合肥	深圳	北京	合肥	深圳
NB	Precision	0.9085	0.8762	0.8470	0.9514	0.8985	0.9228
	Recall	0.9048	0.8368	0.8260	0.9514	0.8985	0.9228
	F1 值	0.9035	0.8527	0.8323	0.9514	0.8985	0.9228
	AUC	0.9952	0.9890	0.9852	0.9967	0.9946	0.9941
DT	Precision	0.8227	0.7222	0.7383	0.9052	0.8386	0.8697
	Recall	0.8037	0.7045	0.7017	0.9052	0.8386	0.8697
	F1 值	0.8103	0.7112	0.7163	0.9052	0.8386	0.8697
	AUC	0.8985	0.8490	0.8487	0.9494	0.9162	0.9328
RF	Precision	0.9621	0.9484	0.9204	0.9393	0.8590	0.9104
	Recall	0.7844	0.5880	0.6755	0.9393	0.8590	0.9104
	F1 值	0.8396	0.6659	0.7463	0.9393	0.8590	0.9104
	AUC	0.9975	0.9886	0.9912	0.9969	0.9918	0.9958
MLP	Precision	0.9367	0.9133	0.8828	0.9650	0.9347	0.9440
	Recall	0.9184	0.8893	0.8574	0.9650	0.9347	0.9440
	F1 值	0.9256	0.8999	0.8679	0.9650	0.9347	0.9440
	AUC	0.9990	0.9950	0.9940	0.9995	0.9970	0.9975

要对单个部门的分类效果进行评价。同时，对所有部门的分类效果进行逐一评价也有利于发现导致不同部门分类效果差异的原因，助力领导信箱自动转递系统的建设。

北京市的部门相对较少，便于比较说明，因此以北京市为例，对各部门的分类效果进行分析。北京市长信箱文本在 4 种算法上的部门分类结果如图 3 所示。

可以看出，市交通委的分类效果相对较差。进一步分析混淆矩阵发现，4 种算法都最易将市交通委的信件误分为市路政局信件，这说明交通委与路政局的职责范围有所重叠（经查，北京市路政局原属交通委，后撤销）。因此，清晰明确的职权范围是领导信箱自动分类转递效果良好的重要前提条件。

此外，分类所用数据类别不均衡，信件的数量可能对分类效果产生影响。对部门样本数与分类效果最好的多层神经网络算法分类结果指标进行相关性分析，结果如表 3 所示。

样本数与分类结果的 F1 值在 0.1 的显著性水平上有相关性，这与此前对宏平均、微平均结果差异性的分析一致，说明提高样本数量确实有利于提升分类准确度。

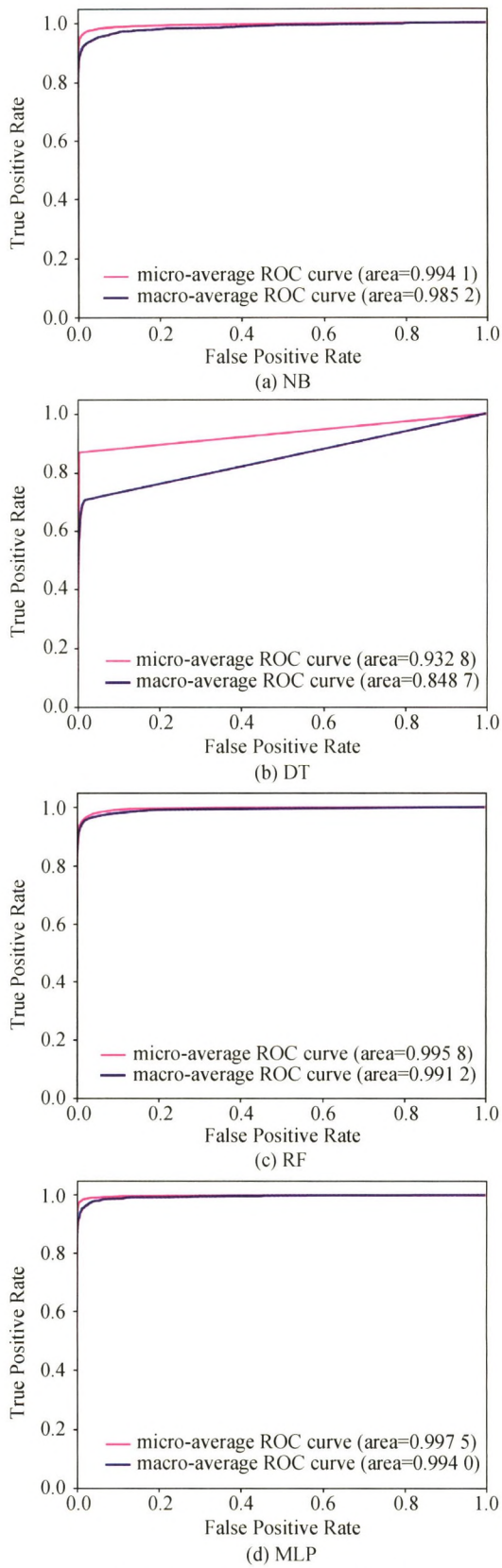

图 2　4 种算法的 ROC 曲线

Fig.2　ROC Curve of Four Algorithms

万方数据

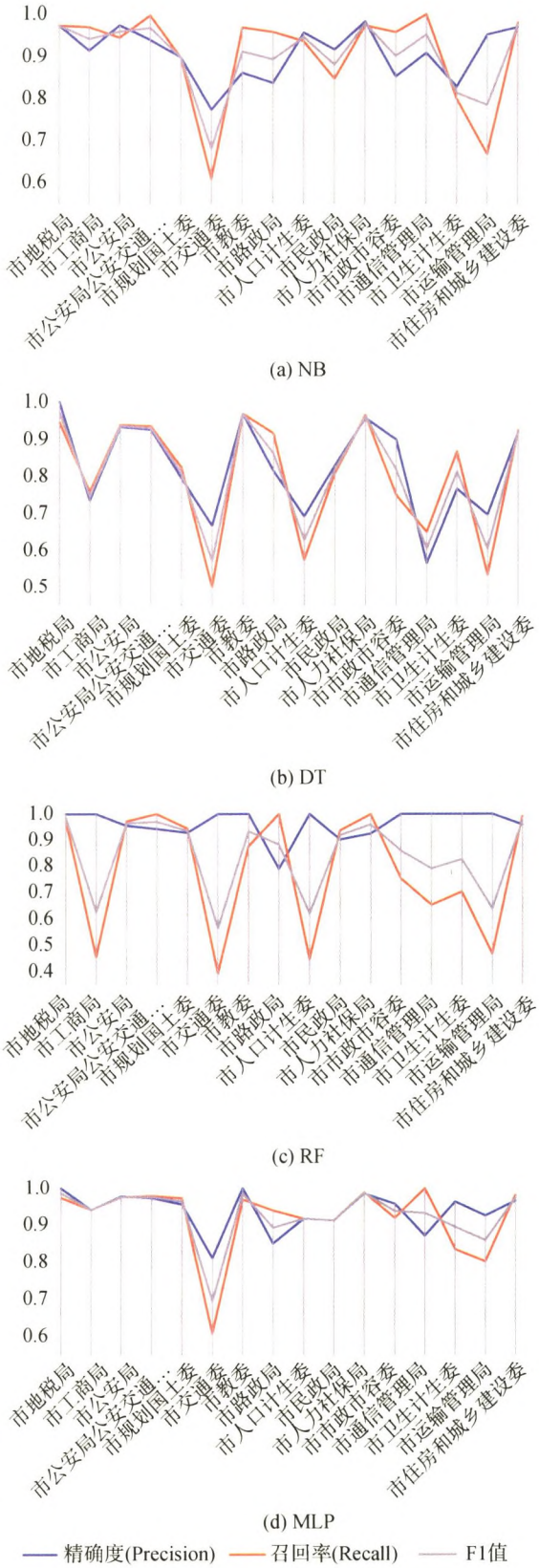

图3　4种算法的部门分类结果

Fig.3　Classification Result of Four Algorithms

表3　部门样本数与分类结果指标相关性分析

Table 3　Correlation Analysis Between the Number of Samples and Classification Result

	样本数	Precision	Recall	F1值
样本数	1.000 0	0.417 1	0.379 3	0.430 7*
Precision		1.000 0	0.601 0**	0.819 7***
Recall			1.000 0	0.950 1***
F1值				1.000 0

（注：***表示$P<0.01$（双尾）；**表示$P<0.05$（双尾）；*表示$P<0.1$（双尾）。）

5　政府网站信箱自动转递方法

传统的领导信箱来信转递需要专职工作人员，当有大量来信时工作人员负担较重，且人工转递需要一定的时间。考虑到政府各职能部门间权责划分明确，不同事项的来信语料特点鲜明，因此对各部门历史来信的语料进行学习，识别部门来信特点，进而实现自动分类转递具有理论上的可行性。对三市的市长信箱文本进行分类实验后发现，利用多层神经网络算法对政府网站领导信箱文本进行部门分类具有较高的准确率，因此对政府网站领导信箱的来信进行部门自动转递在实践层面也具有可行性。

5.1　自动转递流程设计

本文对政府网站领导信箱部门间的转递流程进行自动化设计，当有群众来信时，算法可识别出信件的责任部门并自动转递，而信箱的管理人员只需处理少量的误分信件（根据实验结果，误分率低于0.1），从而减轻工作人员的工作量，如图4所示。

首先对领导信箱中的历史数据按照部门归类，进而训练得到一个基础分类器。当有新的群众来信时，将来信文本结构化存储在数据库中并进行文本预处理，然后使用训练得到的分类器进行部门识别，根据分类结果转递至相应的职能部门A，同时在数据库中储存来信的部门标签。部门A收到来信后进行判断，如果属于本部门的职责范围则直接进行回复；如果不属于本部门的职责范畴，则选择退回信件，领导信箱的管理人员对该信件进行人工判断后转投至相应的职能部门B，同时更新该信件的部门标签并加入训练集，对分类器进行更新，并不断迭代。

来信的自动转递方法省去了人工转递的时间差且节约了人力成本。除此之外，在对领导信箱的群

万方数据

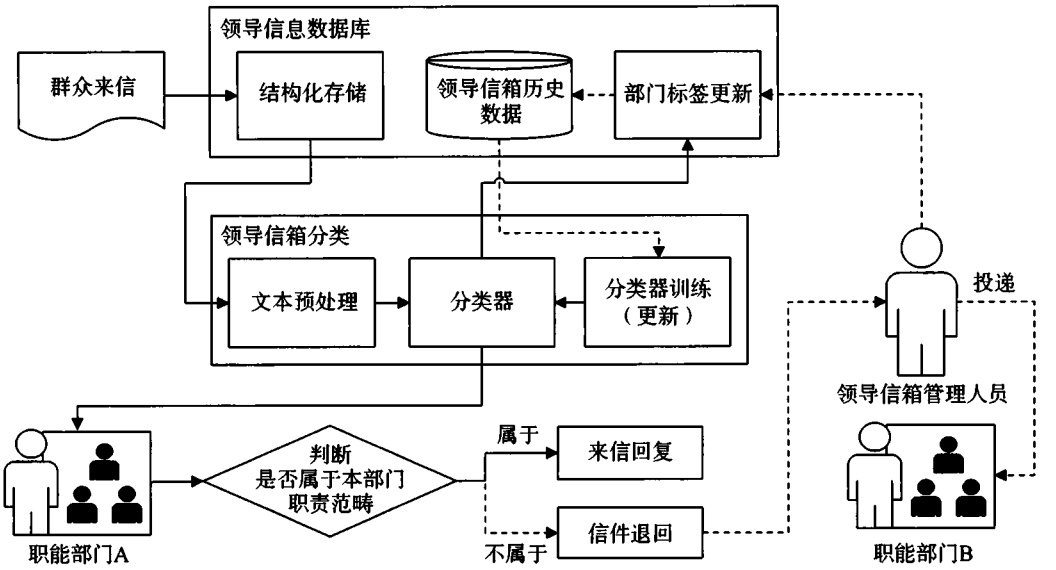

图 4　政府网站信箱部门间自动转递流程

Fig.4　Automatic Transfer Process of the Mailbox on Government Website

众来信进行分类前需要将其在数据库中进行结构化存储，这也为进一步分析群众来信的内容积累数据资源，便于从更大时空上掌握民情民意。

5.2　自动转递方法应用建议

由于政府的职能部门众多，自动转递方法在实际应用中情况较为复杂。为提升领导信箱自动分类转递的效果，结合实验分析，给出以下建议：

（1）对于历史信件数量较少的部门可结合部门职责信息构建部门分类特征词。实验发现，部门信件样本数量与分类效果间存在相关性，样本量较大的部门分类准确率也相对较高。这是由于样本量较大的部门可供算法学习的文本特征也较多。对于信件数过少的部门，可以结合部门职责信息提取特征词，丰富强化小样本特征。

（2）对于单一信件对应多部门问题可考虑设置分类概率阈值，将信件转递至多个部门。实验发现部分分类效果较差的部门是由于部门间职责存在交叉，从而增大误分率。在实际应用中还存在部分信件由多个部门进行回复的情况。因此，在应用领导信箱自动转递系统前，可对各部门职责进行梳理，厘清权责边界。此外，分类算法可输出样本属于所有类别的概率，系统可以设置阈值，当来信属于两个或多个类别的概率差值过小时，将信件转递至多个部门，从而降低误分的可能性。

6　结　语

在推进国家治理体系和治理能力现代化的进程中，公众参与治理的程度和政府回应效率是相辅相成的两个重要方面。本文从政民互动的小处着眼，考虑对线上渠道进行优化，以机器学习算法代替人工，设计一套政府网站信箱的自动转递方法。对北京、合肥和深圳三个城市的市长信箱文本数据进行分类实验，得到如下结论：

（1）神经网络算法最适合对领导信箱来信进行自动分类，分类的微平均精确度和召回率均达0.9以上。较高的分类准确率证明利用机器学习算法进行来信的部门自动识别是可行的。

（2）数据量较少的部门分类准确率要低于数据量大的部门。这说明随着样本量的增大，分类准确率会提高。

（3）个别部门间信件分类准确率较低的主要原因之一是由于部门间职责存在一定的交叉重叠。

进而，研究提炼出一套针对政府网站领导信箱的自动转递方法，并在自动转递运作过程中，对来信数据循环迭代，逐步增加学习样本，以提高分类的准确性。本文还考虑了实际应用中可能出现的复杂情况并给出应用建议。此外，自动转递过程中对群众来信进行结构化存储，数据库中的文本数据可实时

万方数据

动态更新，也有助于政府部门利用文本识别和大数据技术对群众来信进行进一步分析与处理，为政府部门在更大时空上了解民情民意提供数据支持。

本文的不足之处在于难以兼顾不同部门来信数量的不均衡现象，且在实验时剔除了数据量过小的部门来信数据，这在实际应用中可能会存在一定偏差。未来还可以在深度和广度上进一步拓展：一是拓展研究深度，如针对信箱文本数据进行分类算法的改进，针对不同部门的文本内容改进特征提取的方法，针对小样本和来信对应多部门的问题进行细化研究；二是进行研究内容的纵向延伸，如对信箱文本进行语义理解，构建领导信箱知识图谱，并尝试建立对常见问题的自动回应功能等。

参考文献：

[1] 孙宗锋，赵兴华．网络情境下地方政府政民互动研究——基于青岛市市长信箱的大数据分析[J]．电子政务，2019(5)：12-26.(Sun Zongfeng, Zhao Xinghua. A Study on the Interaction Between the Government and the People in the Internet-Based on the Big Data Analysis of the Mayor's Mailbox of Qingdao[J]. E-Government, 2019(5): 12-26.)

[2] 于君博，李慧龙，于书鳗．"网络问政"中的回应性——对K市领导信箱的一个探索性研究[J]．长白学刊，2018(2)：65-74.(Yu Junbo, Li Huilong, Yu Shuman. Responsiveness in "Governing Online"—An Exploratory Study on K City's Leader Mailbox[J]. Changbai Journal, 2018(2): 65-74.)

[3] 郑俊田，郜媛莹，顾清．地方政府权力清单制度体系建设的实践与完善[J]．中国行政管理，2016(2)：6-9.(Zheng Juntian, Gao Yuanying, Gu Qing. Practice and Perfection of Local Governmental Administrative Power List System Construction [J]. Chinese Public Administration, 2016(2): 6-9.)

[4] 王珺．基于文本特征识别的电子档案自动归类系统研究[J]．现代电子技术，2019，42(18)：45-49.(Wang Jun. Research on Electronic Archive Automatic Classification System Based on Text Feature Recognition[J]. Modern Electronics Technique, 2019, 42(18): 45-49.)

[5] 李湘东，徐朋，黄莉，等．基于KNN算法的文本自动分类方法研究——以学术期刊栏目自动归类为例[J]．图书情报知识，2010(4)：71-76.(Li Xiangdong, Xu Peng, Huang Li, et al. Research of Journals Manuscript Categorization Based on KNN Algorithm[J]. Document, Information & Knowledge, 2010(4): 71-76.)

[6] 李成铭．基于文本特征提取技术的在线人职匹配研究及应用[D]．成都：电子科技大学，2017.(Li Chengming. Research and Application of Talent Job Online Matching Based on Text Feature Extraction Technology[D]. Chengdu: University of Electronic Science and Technology of China, 2017.)

[7] 王若佳，张璐，王继民．基于机器学习的在线问诊平台智能分诊研究[J]．数据分析与知识发现，2019，3(9)：88-97.(Wang Ruojia, Zhang Lu, Wang Jimin. Automatic Triage of Online Doctor Services Based on Machine Learning [J]. Data Analysis and Knowledge Discovery, 2019, 3(9): 88-97.)

[8] Kim K, Zzang S Y. Trigonometric Comparison Measure: A Feature Selection Method for Text Categorization[J]. Data & Knowledge Engineering, DOI: 10.1016/j.datak.2018.10.003.

[9] Ghareb A S, Bakara A A, Al-Radaideh Q A, et al. Enhanced Filter Feature Selection Methods for Arabic Text Categorization[J]. International Journal of Information Retrieval Research (IJIRR), 2018, 8(2): 1-24.

[10] Hartmann J, Huppertz J, Schamp C, et al. Comparing Automated Text Classification Methods[J]. International Journal of Research in Marketing, 2019, 36(1): 20-38.

[11] 田欢，李红莲，吕学强，等．基于改进BP神经网络的学术活动文本分类[J]．北京信息科技大学学报(自然科学版)，2018，33(5)：38-44.(Tian Huan, Li Honglian, Lv Xueqiang, et al. Text Categorization of Academic Activities Based on an Improved BP Neural Network[J]. Journal of Beijing Information Science & Technology University, 2018, 33(5): 38-44.)

[12] 刘浏，王东波．基于论文自动分类的社科类学科跨学科性研究[J]．数据分析与知识发现，2018，2(3)：30-38.(Liu Liu, Wang Dongbo. Identifying Interdisciplinary Social Science Research Based on Article Classification [J]. Data Analysis and Knowledge Discovery, 2018, 2(3): 30-38.)

[13] Gauld R, Flett J, McComb S, et al. How Responsive are Government Agencies When Contacted by Email? Findings from a Longitudinal Study in Australia and New Zealand[J]. Government Information Quarterly, 2016, 33(2): 283-290.

[14] 李慧龙，于君博．数字政府治理的回应性陷阱——基于东三省"地方领导留言板"的考察[J]．电子政务，2019(3)：72-87.(Li Huilong, Yu Junbo. The Responsive Trap of Digital Government Governance-Based on the Investigation of "Message Board of Local Leaders" in Three Northeastern Provinces[J]. E-Government, 2019(3): 72-87.)

[15] Ong C S, Wang S W. Managing Citizen-Initiated Email Contacts [J]. Government Information Quarterly, 2009, 26(3): 498-504.

[16] 胡佳妮，徐蔚然，郭军，等．中文文本分类中的特征选择算法研究[J]．光通信研究，2005(3)：44-46.(Hu Jiani, Xu Weiran, Guo Jun, et al. Study on Feature Selection Methods in Chinese Text Categorization[J]. Study on Optical Communications, 2005(3): 44-46.)

[17] 张志飞，苗夺谦，高灿．基于LDA主题模型的短文本分类方法[J]．计算机应用，2013，33(6)：1587-1590.(Zhang Zhifei, Miao Duoqian, Gao Can. Short Text Classification Using Latent Dirichlet Allocation[J]. Journal of Computer Applications, 2013,

万方数据

33(6): 1587-1590.)
[18] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18 (11): 613-620.
[19] Manning C, Raghavan P, Schütze H. Introduction to Information Retrieval[M]. Cambridge University Press, 2008.
[20] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.
[21] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees[M]. CRC Press, 1984.
[22] Hinton G E. Connectionist Learning Procedures[J]. Artificial Intelligence, 1989, 40(1-3): 185-234.

作者贡献声明:

王思迪:设计研究方案,完成实验,论文撰写与修改;
胡广伟:提出研究建议,设计研究框架,提出论文修改建议,论文最终版本修订;
杨已煜:完善研究思路与方案;
施云:论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储,E-mail: w_sidi@163.com。
[1] 王思迪. 北京市长信箱数据.xlsx. 北京市长信箱原始数据.
[2] 王思迪. 合肥市长信箱数据.xlsx. 合肥市长信箱原始数据.
[3] 王思迪. 深圳市长信箱数据.xlsx. 深圳市长信箱原始数据.
[4] 王思迪. beijing.xlsx. 北京市长信箱预处理后数据.
[5] 王思迪. hefei.xlsx. 合肥市长信箱预处理后数据.
[6] 王思迪. shenzhen.xlsx. 深圳市长信箱预处理后数据.

收稿日期:2019-10-31
收修改稿日期:2020-02-11

Automatic Transferring Government Website E-Mails Based on Text Classification

Wang Sidi[1,2] Hu Guangwei[1,2] Yang Siyu[1,2] Shi Yun[1]
[1](School of Information Management, Nanjing University, Nanjing 210023, China)
[2](Government Data Resources Institution of Nanjing University, Nanjing 210023, China)

Abstract: [**Objective**] This research proposes a method to automatically transferring e-mails received by government websites, aiming to reduce labor costs of managing public email boxes. [**Methods**] First, we chose four representative classification algorithms, including Naïve Bayes, Decision Tree, Random Forest and Multi-Layer Perception, and compared their classification resutls of e-mails received by the websites of Mayor's Offices in Beijing, Hefei and Shenzhen. Then, we designed a method of automatically transferring these emails. Finally, we gave suggestions on the application of our method in the real world settings. [**Results**] Multi-Layer Perception yielded the best performance in our study, with the macro average precision and recall reaching more than 0.85, and all micro average indicators reaching more than 0.93. Naïve Bayes took the second place. Random Forest had a high macro average precision, but poor recall score. Decision Tree had an average precision and recall results. [**Limitations**] We did not examine the impacts of skewed distribution of received emails and eliminated the departments receiving few emails. [**Conclusions**] The proposed method optimizes the operation of public e-mails, which improves the efficiency of online government and reduces administrative costs.

Keywords: Leader's Mailbox Automatic Transfer Text Classification Multi-Layer Perception Process Optimization

万方数据