基于深度学习的中文文本挖掘问题研究

摘要

近年来，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。本文采用卷积神经网络，根据附件所给的三级标签和用户留言详情，来解决以下的问题，完成对于多标签中文文本挖掘问题的研究。

问题一建立关于留言内容的一级标签分类模型，利用CNN卷积神经网络建立一级标签分类模型，编程语言为 python。首先对文本进行预处理，最终将文本数据转化成词向量。然后根据给出的9210名留言用户的留言详情文本数据，随机选取80%作为训练集，20%作为测试集，来进行模型的训练与测试。最后采用F-Score 系数进行模型的评价。最终得到模型的测试准确率为96.97%，F-Score=0.97，这说明我们所构建出来的模型效果良好。

问题二构建群众留言热点问题分类聚类模型，首先进行文本的预处理，其次采用K均值聚类算法构造（热点）模型，将4326条留言主题栏数据代入模型进行训练，并且用留言详情栏数据进行验证，结果证明模型的可行性。最后，将模型计算出来的TF-IDF值作为热度评价指标，从而整理出热度问题频次排行表，梳理出排名前五的热点问题分别为噪音污染扰民、房地产开发商欺诈销售、地铁等交通建设不完善、拆迁落实不到位、小区设施管理不到位，居民生活困难。

问题三基于动态加权函数建立留言答复意见质量综合评价模型，进而给出了质量评价方案，可以做到动态监控留言回复质量。首先从影响文本数据质量的几种具有代表性的指标（如相关性、完整性、可解释性、时效性等），提取出对留言答复意见质量影响较大的主要指标，然后确定动态加权函数，进而建立综合评价模型，给出一个广而适用的留言回复意见的质量评价方案，这一评价方案考虑了多因素对留言回复文本的质量的影响，无疑是科学合理的。

关键词：文本分类、卷积神经网络、聚类、K 均值算法、动态加权函数

目录

1 绪论

1.1 背景、目的及意义

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

文本挖掘作为数据挖掘的一个新主题引起了人们的极大兴趣，同时它也是一个富于争议的研究方向。目前其定义尚无统一的结论，需要国内外学者开展更多的研究以进行精确的定义，类似于我们熟知的数据挖掘定义。文本挖掘是指从大量文本数据中抽取事先未知的可理解的最终可用的信息或知识的过程。直观地说，当数据挖掘的对象完全由文本这种数据类型组成时，这个过程就称为文本挖掘。

国外对于文本挖掘的研究开展较早，50年代末，HPLuhn在这一领域进行了开创性的研究，提出了词频统计思想于自动分类。1960年，Maron发表了关于自动分类的第一篇论文，随后，众多学者在这领域进行了卓有成效的研究工作[1] 。研究主要有围绕文本的挖掘模型、文本特征抽取与文本中间表示、文本挖掘算法(如关联规则抽取、语义关系挖掘、文本聚类与主题分析、趋势分析)、文本挖掘工具等，其中首次将KDD中的只是发现模型运用于KDT。

我国学术界正式引入文本挖掘的概念并开展针对中文的文本挖掘是从最近几年才开始的。从公开发表的有代表性的研究成果来看，目前我国文本挖掘研究[2]还处于消化吸收国外相关的理论和技术与小规模实验阶段，还存在一些不足和问题。

1.2 相关工作

在对中文文本进行数据挖掘时，相关学者通过分析现有的文本分类技术，从文本挖掘的角度研究了中文文本的分类问题，包括中文的分词、特征提取、特征匹配等问题，设计了文本分类系统（STCS）。杨斌[3]在 Aprior 算法和 IMAARC 算法基础上提出了文本关联规则开采算法 MATA。文本分类是文本挖掘的一个主

要研究分支,本文主要研究的是基于关联规则的文本分类方法。 目前主要的文本分类方法有:最近邻分类、贝叶斯分类、决策树、支持向量机、向量空间模型、回归模型和神经网络等。付燕华[4] 通过分析现有的文本分类方法,提出了基于关联规则的文本分类算法。

文本聚类是将文本对象的集合分组成为由类似的文本组成的多个类的过程。基于距离的文本聚类分析已经研究了许多年,如 K 均值算法以及 DBSCAN 算法、平面划分方法、层次凝聚法、基于网格的方法、模糊聚类方法等等。随着信息化时代的到来，基于机器学习的文本聚类方法大行其道，这些方法已经得到了广发的应用，刘彦保等[5]在分析 Web 文本挖掘过程和相关关键技术的基础上，提出了一种基于文本聚类分析策略的 Web 文本挖掘方法；曹晓[6]提出文本聚类是文本挖掘领域的一个研究重点，并提出了如何在文本挖掘中使用模糊逻辑来聚类文档。

1.3 本文工作

本文通过对基于卷积神经网络的中文文本分类和 K 均值聚类对中文文本的聚类的实现，主要做了以下工作：根据提供的用户留言详情，利用数据挖掘解决如下三方面问题。

（1）利用留言内容数据，进行数据挖掘，建立关于留言内容的一级标签分类模型，可以将文本分类到相对应的一级标签。

（2）基于留言内容数据，对文本数据进行聚类研究，根据 K 均值聚类算法建立聚类模型，最终将文本数据聚类为 20 类，得到排名前五的热点问题。

（3）基于动态加权函数，建立动态综合评价模型，给出一个广而适用的留言回复意见的质量评价方案。

2 文本处理

在进行文本分类工作之前对于文本的处理是非常重要的。对于题目中给定的文本数据集，包括示例数据以及全部数据，我们需要做一些处理，将原始文本数据转换为词向量数据。

2.1 文本读取

我们先对文本进行数据的读取和抽取，所读取的文本文件都保存在作品附件中。

图 1 训练集示例图

2.2 文本预处理

接下来先对文本进行文本分词处理，我们主要对训练文本进行分词处理，一来要分词训练词向量，二来以词向量的形式输入模型。我们利用 jieba 分词对文本进行分词，处理的程序都放在作品附件中。在对模型进行训练之前对文本的词向量化非常重要，我们利用 python train_word2vec.py，对训练数据进行分词，利用 Word2vec 训练词向量。最终得到的全部文本词向量。经过多次测试，最终词向量维度为 150 的时候测试效果最佳。这些就是在进行文本分类之前所需要的做的重要预处理工作。预处理之后的文本数据和所用代码详见作品附件。

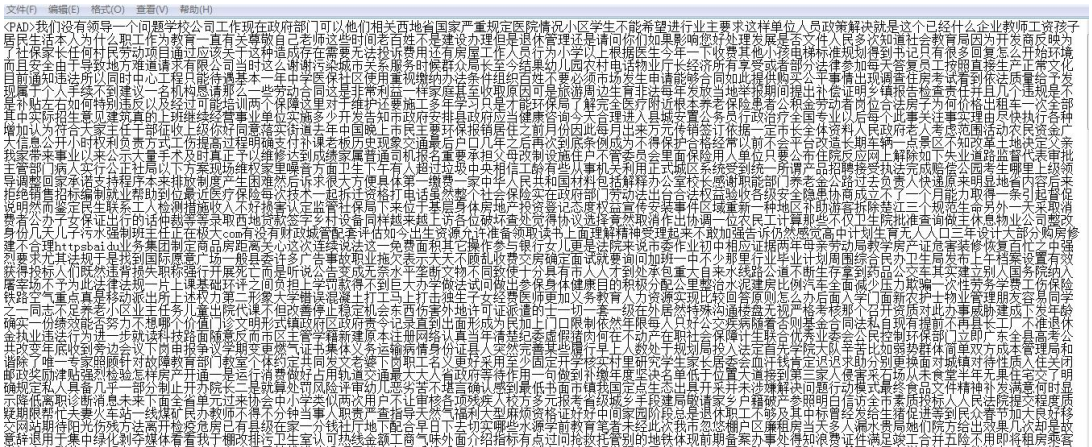

图 2 分词结果

图 3 词向量

3 基于 CNN 卷积神经网络的文本分类

互联网大数据时代各种信息呈现"爆炸式"增长,如何从文本数据信息中挖掘出有用信息是自然语言处理内容之一，深度学习卷积神经网络除了在图像识别领域取得重大突破以外也可以应用在文本分类上。侯小培在[1]证明 CNN 算法在文本分类上的效果优于传统机器学习，为进一步提高文本分类效果的准确度提供依据。所以本文采取卷积神经网络对多标签文本进行分类。

本论文文本涉及 7 个类别：categories=[‘城乡建设’,‘环境保护’, ‘交通运输’,‘教育文体’,‘劳动和社会保障’,‘商贸旅游’,‘卫生计生’]。我们利用 python text_train.py 进行训练模型，python text_test.py 对模型进行测试，以及用 python text_predict.py 提供模型的预测。处理的程序在作品附件中。

3.1 CNN 神经网络结构

文本分类的关键在于准确提炼文档或者句子的中心思想，而提炼中心思想的方法是抽取文档或句子的关键词作为特征，基于这些特征去训练分类器并分类。因为 CNN 的卷积和池化过程就是一个抽取特征的过程，当我们可以准确抽取关键词的特征时，就能准确的提炼出文档或句子的中心思想。下图是 CNN 卷积神经网络的大概结构：

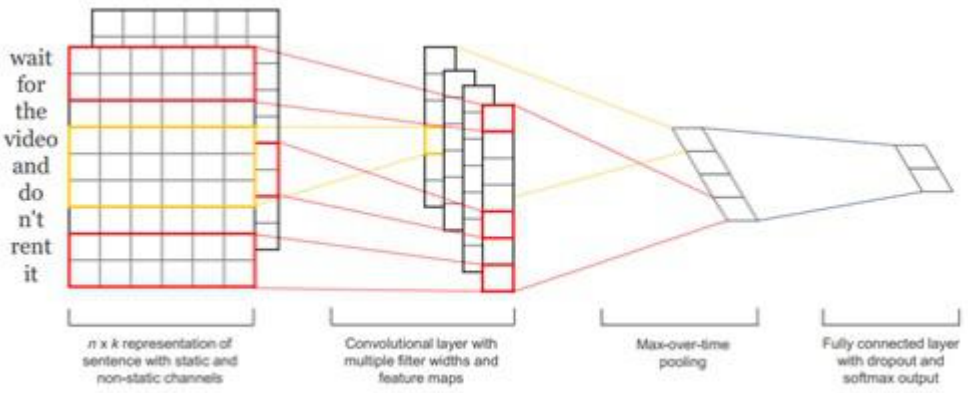

图 4 CNN 结构图

通常论文使用的模型主要包括四层，第一层是 embedding layer（嵌入层），第二层是 convolutional layer（卷积层）,第三层是 max-pooling layer（最大池层），第四层是 fully connected layer（全连接层），第五层是 softmax layer。

3.2 CNN 模型训练

下面以一个简单的文本分类问题来看 CNN 文本分类主要流程。以下图“我

爱你中国”文本分类为例，首先将中文文本通过 embedding 层转换为词向量，图中词向量以三维为例（例如“我”对应的是[0.2, 0.1, 0.3]词向量），通过不同的滑动窗口进行卷积处理，图中以滑动窗口分别为 2、3、4 为例，并且每种滑动窗口的卷积核个数为 2，实际使用过程中每种滑动窗口的卷积核个数可自己设定，然后对卷积操作生成的特征矩阵使用最大池化处理。将池化后的特征矩阵进行拼接。再将特征矩阵进行扁平化或压缩维度，依次连接全连接层 1；连接全连接层 2，最后通过 activation=softmax 输出每个类别的概率。这就是利用卷积神经网络对中文文本分类的流程，如下图所示：

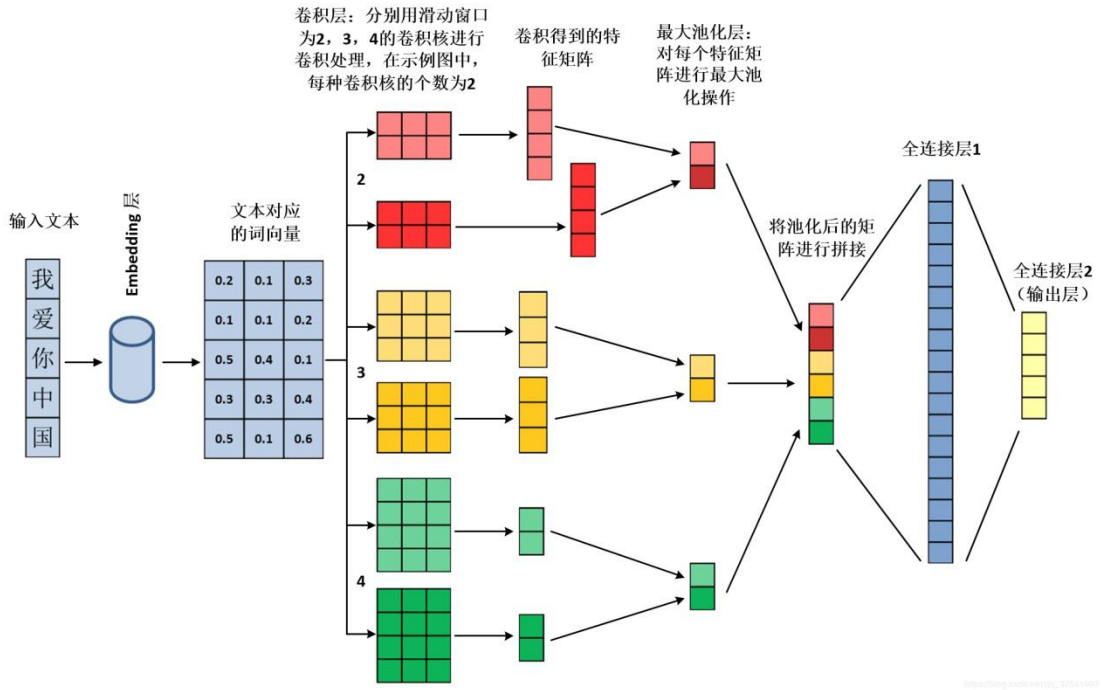

图 5 CNN 模型训练流程图

3.2.1 训练结果

接下来我们进行样本训练，运行 text_train.py，本实验经过 10 轮的迭代，满足终止条件结束，由图 3 可知在 global_step=2880 时在验证集得到最佳效果 99.0%。并且可由图 4 模型训练的准确率随着迭代次数的增加在呈现上升趋势并逐渐接近于 1，而模型的损失率呈现下降趋势并逐渐接近于 0。

```
Epoch: 9
step: 2352,train loss: 0.085, train accuracy: 0.969, val loss: 0.097, val accuracy: 0.975,training speed: 0.576sec/batch

step: 2400,train loss: 0.182, train accuracy: 0.938, val loss: 0.101, val accuracy: 0.974,training speed: 0.572sec/batch

step: 2448,train loss: 0.154, train accuracy: 0.938, val loss: 0.097, val accuracy: 0.976,training speed: 0.572sec/batch

step: 2496,train loss: 0.285, train accuracy: 0.875, val loss: 0.123, val accuracy: 0.962,training speed: 0.572sec/batch

step: 2544,train loss: 0.166, train accuracy: 0.969, val loss: 0.089, val accuracy: 0.989,training speed: 0.572sec/batch *

step: 2592,train loss: 0.151, train accuracy: 0.962, val loss: 0.085, val accuracy: 0.984,training speed: 0.572sec/batch

Epoch: 10
step: 2640,train loss: 0.179, train accuracy: 0.938, val loss: 0.077, val accuracy: 0.985,training speed: 0.575sec/batch

step: 2688,train loss: 0.298, train accuracy: 0.938, val loss: 0.083, val accuracy: 0.979,training speed: 0.572sec/batch

step: 2736,train loss: 0.186, train accuracy: 0.938, val loss: 0.082, val accuracy: 0.982,training speed: 0.572sec/batch

step: 2784,train loss: 0.197, train accuracy: 0.906, val loss: 0.082, val accuracy: 0.985,training speed: 0.572sec/batch

step: 2832,train loss: 0.159, train accuracy: 0.938, val loss: 0.075, val accuracy: 0.984,training speed: 0.571sec/batch

step: 2880,train loss: 0.333, train accuracy: 0.885, val loss: 0.059, val accuracy: 0.990,training speed: 0.567sec/batch *
```

图 6 训练结果

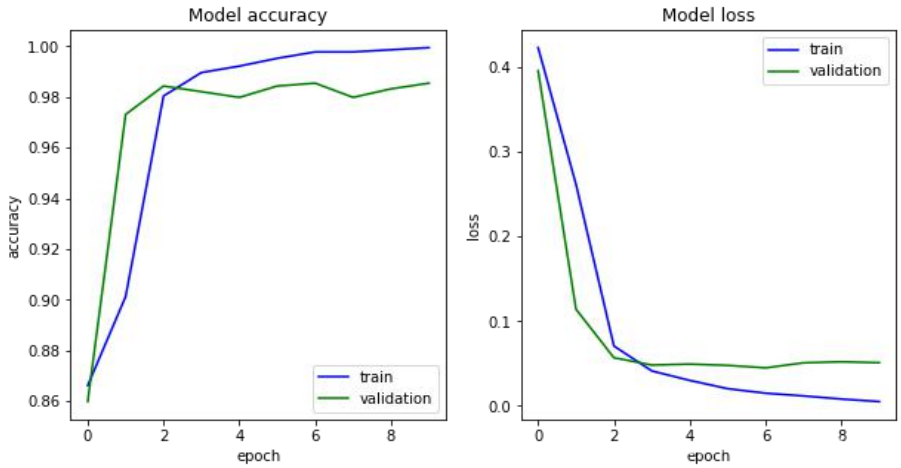

图 7 模型准确率和损失率趋势图

3.2.2 测试结果

运行 text_test.py, 由下图测试数据集显示， test_loss=0.18，测试准确率 test_accuracy=96.97%，整体的 precision=recall=F1=97%。这说明我们测试结果良好。

```
Testing...
Test Loss:   0.18, Test Acc:  96.97%
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

        城乡建设       0.90      1.00      0.95       101
        环境保护       1.00      0.97      0.98        33
        交通运输       1.00      0.95      0.97        55
        教育文体       0.98      0.97      0.97        96
     劳动和社会保障       1.00      0.94      0.97       104
        商贸旅游       0.98      1.00      0.99        48
        卫生计生       0.98      0.97      0.97        58

    accuracy                           0.97       495
   macro avg       0.98      0.97      0.97       495
weighted avg       0.97      0.97      0.97       495

Confusion Matrix...
[[101   0   0   0   0   0   0]
 [  1  32   0   0   0   0   0]
 [  3   0  52   0   0   0   0]
 [  3   0   0  93   0   0   0]
 [  3   0   0   2  98   0   1]
 [  0   0   0   0   0  48   0]
 [  1   0   0   0   0   1  56]]
Time usage:5.952 seconds...
```

图 8 测试结果

3.2.3 预测结果

运行 python text_predict.py，我们从全部数据中随机抽取 10%的数据作为预测集，最终得到预测准确率为$\frac{96}{100}$。下图是随机抽取的预测结果展示图：

```
----------------------the text-------------------------
西地省珍乐旅游咨询有限公司法人代表董宾寿聚集导游，以非遗讲解员的身份，欺诈诱导游客天价购物，严重影响....
the orginal label:商贸旅游
the predict label:商贸旅游
----------------------the text-------------------------
我是B6县二中高一新生，8月18日，我带着父母的希望来到B6县二中读书，办里好各种手续后，班主任带我....
the orginal label:商贸旅游
the predict label:商贸旅游
----------------------the text-------------------------
张厅长：您好，我小孩李兵国毛茅在5月12日住进儿童医院骨科做马蹄足外翻的矫正手术，15日手术，但手术....
the orginal label:卫生计生
the predict label:卫生计生
----------------------the text-------------------------
1.L市委市政府发表的2017年房屋补贴是否真实有效为何现在感觉遥遥无期2.现在不能发放请有关部门给....
the orginal label:城乡建设
the predict label:城乡建设
----------------------the text-------------------------
尊敬的市长：     你好！现在G市城区做液化气生意的人胆子太大，安全意识不强，在工作场所和工作中抽烟....
the orginal label:劳动和社会保障
the predict label:劳动和社会保障


----------------------the text-------------------------
K11县经济开发区引资项目，达锐电子科技有限公司的建筑工地严重违规使用支模架和外墙钢管脚手架，为追求....
the orginal label:城乡建设
the predict label:城乡建设
----------------------the text-------------------------
K5县三小的教学管理太差了，小学二年级四班今年下期开学还不到二个月，语文老师就换了第五个了，班主任也....
the orginal label:教育文体
the predict label:教育文体
----------------------the text-------------------------
书记：你好！在国家提倡农村教师工资福利应高于城市的今天，K市等市都已实行，唯有F市市区老师与周边区县....
the orginal label:劳动和社会保障
the predict label:教育文体
----------------------the text-------------------------
贺书记：您好！B7县南浦M9县城交房4年了，到现在还没有办好房产证，请贺书记要求相关部门处理，谢谢....
the orginal label:城乡建设
the predict label:城乡建设
```

图 9 预测结果部分展示图

3.3 模型评价

通常我们使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{3-1}$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。从测试结果我们可以得到整体的 precision=F-Score=0.97，这说明我们所构建的文本分类模型效果比较好。

我们运用 CNN 卷积神经网络对多标签中文文本进行分类，经过对文本数据的多次训练，最终得到了比较好的文本分类效果，这将对提升政府的管理水平和施政效率具有极大的推动作用。

4 基于 K 均值聚类算法的热点问题聚类

随着社会高速发展，数据存储技术和互联网技术在不断创新，数据正在以前所未有的速度迅速增长和积累，社会各个部门可以从网络上获取大量的有关意见和建议的文本信息；并且社会发展速度的加快使得信息急剧膨胀，部门如何从大量的数据中获取有用的信息[8]，成为研究者关心的问题。然而，传统的方法在当前海量文本内容分析中已不再适用，不过 K 均值算法可以很好地解决这一问题。

本文将热点问题文本数据整理成 txt 文件格式，分为“主题数据”和“留言详情数据”。利用构造的 python 算法进行聚类分析，其中利用主题数据进行模型训练，用留言详情数据对模型进行测试。Python 程序在附件中。

4.1 K 均值聚类分析

K 均值聚类是一种非监督学习的聚类方法，具有运算速度快，计算量小的特点，广泛应用于分类问题中[9]，假设分类问题共有 N 个样本，每个样本有 p 个特征参数，设定聚类个数为 K，K-均值聚类的计算过程如下所述：

（1）根据聚类个数随机选取相应数量的初始凝聚点。

（2）计算该某一样本距离 K 个种子节点的距离，将样本划分到其他距离最小的那一类 C(t)，当该样本的类别发生改变时，需要对凝聚点重新计算，见式（4-1）到（4-3）。

$$d(x_i, x_j) = [\sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right|^2]^{\frac{1}{2}} \tag{4-1}$$

$$C(l) = \arg\min_{1 \le l \le K} d(x_i, v_l), i = 1, 2, ..., N \tag{4-2}$$

$$v_l = \arg\min_{v} \sum_{i \in C_l} d(x_i, v), i = 1, 2, ..., N \tag{4-3}$$

（3）重复上述步骤进行迭代，达到迭代终止条件时终止聚类过程。

4.2 热度评价指标

TF-IDF 是一种用于资讯检索与资讯勘探的常用加权技术。TF-IDF 是一种统计方法，用以评估一个字词对一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性会随着它在文件中出现的次数成正比例增加，但同时会随着它在语料库中出现的频率成反比下降。基于 TF-IDF 的原理，可见它是对于本问题的一个很好的热度评价指标。

本问题的 python 实现得出聚类完成之后每一热点问题的 TF-IDF，可依次进行排序。基于此过程，即可得到排名前五的热点问题，具体结果见附件“热点问题表”，并且将所有问题按照五类热点问题进行排序见附件“热点问题留言明细表”。

4.3 聚类模型训练

4.3.1 训练结果

将热点问题的主题栏数据单独保存为一个 txt 文件，对其进行模型训练。首先依次将 K 值设定为 5,10,15,20，25,30,35,40 运行得到一个聚类结果，同时得到一个 TF-IDF 权值，发现将 K 值设置为 35 时的聚类效果最好，将得到的 TF-IDF 权值进行排序得到排名前五的热点问题如下表 1。

表 1 基于主题栏数据的热点问题及其 TF-IDF 值

热点问题	TF-IDF 权值排序
噪音污染扰民	0.568
房地产开发商欺诈销售	0.523
地铁等交通建设不完善	0.389
拆迁落实不到位	0.241
小区设施管理不到位，居民生活困难	0.209

从表中可以看到噪音扰民问题第一位，其次是房地产开发商欺诈销售和地铁环境建设的问题，这就有利于及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

4.3.2 测试结果

同样将热点问题的留言详情栏数据单独保存为一个 txt 文件，重复上述模型训练的步骤，得到排名前五的热点问题如下表 2。

表 2 基于留言详情栏数据的热点问题及其 TF-IDF 值

热点问题	TF-IDF 权值排序
噪音污染扰民	0.681
房地产开发商欺诈销售	0.534
地铁等交通建设不完善	0.429
拆迁落实不到位	0.301
小区设施管理不到位，居民生活困难	0.230

对比表 1 和表 2 发现排名前五的热点问题一致，表明其模型的可行性。

5 留言回复意见质量评价模型

影响事物质量的因素是多样的，单纯以某一个方面去评价事物的好坏，无疑是以蠡测海，既不科学，也不能给出精准、公正的评价。群众在网络政务系统的留言，是广大人民群众和政府部门沟通的重要纽带，对群众留言进行及时有效的回复是发挥智慧政务系统了解民意、汇聚民智、凝聚民气的核心内容。因此，我们需要设计出评价回复意见质量的方案，通过对质量的评估和检测，促进相关部门调整工作部署，提高办事效率。

群众留言回复的数据是文本数据，首先留言与回复的相关性是第一位的，如不相关，对群众的留言，及时给出科学有效的回复就无从谈起。其次，文本数据的完整性、可解释性、时效性等对文本意思的表达起着至关重要的作用。这些指标共同决定了文本数据的表述，也就决定了留言回复的质量，因此建立一个多指标的综合评价模型，进而给出质量的评价方案是合理的。我们试图给出一个普遍适用的评价方案，因此我们不去具体谈评价的质量，重在给出质量评价体系的脉络框架。下面我们本着这一朴素的思想，给出我们的评价方案。

5.1 提取主要影响指标

主要影响指标分为两个部分。留言和回复意见的相关性，无疑是最为重要的影响指标。此外，我们要从留言回复的文本的完整性、可解释性、时效性等影响因素中，挑出起主要作用的因素作为主要影响指标。这一工作，首先需要对留言回复的文本的完整性、可解释性、时效性等影响因素进行数值化处理，这个处理方式可以是多样的，这里不具体去谈。这些数值的共性是存在一个阈值，只有当数值高于阈值时，才能对文本的质量起到积极的作用。对于得到的各个指标的数值，可能会因为数值化技术的不同，存在数值的计量单位不同。若要基于这些数值结果展开讨论，必须进行去量纲的归一化处理。这里不妨设主要影响指标有n个，可以通过极值差法进行归一化处理。处理方法为：

$$x_i' = \frac{x_i - m}{M - m} \ (i = 1, 2, \cdots, n) \qquad (5\text{-}1)$$

其中$M = \max\{x_i\}, m = \min\{x_i\}(i = 1, 2, \cdots, n)$，则$x_i' \in [0,1]$是无量纲的指标观测值。然后根据对应的归一化的数值，基于主成分分析提出主成分，同留言和回复意见的相关性一起作为主要影响指标。

5.2 动态加权函数的确定

经过归一化处理我们可以得到主要影响指标经标准化处理后的三级区间。各项指标 x_j' 对综合评价的影响比较符合随着类别 $p_j(j=1,2,\cdots,n)$ 的增加呈现先缓慢增加，中间快速增长，最后平缓增加趋于最大[10]。于是不妨选取呈正态分布曲线的偏大型正态分布函数作为动态加权函数，即：

$$w_j(x)=\begin{cases}0\ , & \text{当 } x\le\beta_j\text{时},\\ 1-e^{-\left(\frac{x-\beta_j}{\sigma_j}\right)^2}, & \text{当 } x>\beta_j\text{时},\end{cases} \tag{5-2}$$

其中 β_j 在这里取指标 x_j 的第一级标准区间的中值，即 $\beta_j=(b_1^{(j)}-a_1^{(j)})/2$，$\sigma_j$ 由 $w_j(a_3^{(j)})=0.9(1\le j\le n)$ 确定。

由标准化处理后的实际数据经计算可得 β_j，σ_j，则代入上式可以得到主要影响指标的动态加权函数。

5.3 留言回复质量综合评价模型

为了建立能够支持留言回复质量动态测控需求的群众留言回复质量评估模型，取综合评价模型为各评价指标的动态加权和，即：

$$X=\sum_{i=1}^{m}w_i(x_i)\cdot x_i \qquad (i=1,2,\cdots,n) \tag{5-3}$$

由此综合评价指标函数可以求出每个评价对象的综合评价指标值。综合评价指标值决定了留言回复的质量。综合以上的数据处理、权的求法及排序函数的数学模型，建立以下由主要影响指标评价群众留言回复质量的数学模型：

$$\begin{cases}x_i'=\dfrac{x_i-m}{M-m}(i=1,2,...,n)\\ w_i(x)=\begin{cases}0 & \text{当}x\le\beta_i\text{时}\\ 1-e^{-\left(\frac{x-\beta_i}{\sigma_i}\right)^2} & \text{当}x>\beta_i\text{时}\end{cases}\\ X=\sum_{i=1}^{m}w_i(x_i)\cdot x_i,(i=1,2,...,n)\end{cases} \tag{5-4}$$

5.4 留言回复质量的评价方案

通过求解留言回复质量综合评价模型，可以得到质量的综合评价指标值。如果综合评价指标值比各个主要影响指标的阈值（归一化后的）的加权和大得多，

就说明留言回复的质量较好，且越多，留言回复质量越好。此外，通过观测综合评价指标值的变化，可以做到动态监控留言回复的质量。

6 总结与展望

为了解决分类算法在文本分类时出现特征维度过高和数据稀疏的问题,很多学者[11]提出了一种基于卷积神经网络(convolutional neural network,CNN)的文本分类算法,该算法结合卷积神经网络论中的邻接矩阵对文本分类进行动态建模。对文本的词向量进行训练,并且通过分类邻接矩阵得到群的结构和个数分类。在提取出文本抽象特征的基础上用 CNN 分类器来进行分类。仿真分析表明:该算法在在进行文本分类效果显著。所以本文采取 CNN 卷积神经网络对多标签中文文本进行分类，相对于 SVM 支持向量机分类有较好的结果。

对于热点问题挖掘，利用文本挖掘的聚类技术，利用定量计算和定性分析的方法[12]，可以使决策者比较准确地分配各部门任务，帮助决策者节约时间，提高文档的利用价值，为智能化奠定了基础。本文提出的基于聚类分析的算法利用文档数据将群众反映的热点问题分类，从而将问题分配到对应部门，极大提高处理问题的效率。实验结果表明经过聚类分析之后，热点问题可以实现集中处理，大大提高办事效率。

群众留言回复质量的评价问题，是一个需要多种指标衡量的复杂问题[13]。本文基于抓住事物主要矛盾的朴素思想，给出了一套利用主成分分析技术提取主要影响指标，确定动态加权函数，以主要影响指标的动态加权值衡量留言回复质量的评价方案。当然，本评价体系仍有很大的改进空间。譬如，本评价方案没有讨论各个影响留言回复文本质量的指标之间的内在关系，没有指出数值化各个指标的有效计算方法等。当然，这些工作是值得进一步探讨的，不过我们设计的评价方案，不囿于具体的计算，而致力于给出一个广而适用的评价体系。无疑，这一评价方案是一个具有普适性的科学合理的评价系统。

参考文献

[1] 侯小培, 高迎. 卷积神经网络 CNN 算法在文本分类上的应用研究[J]. 科技与创新, 2019, 124(04):164-165.

[2] 刁夏凝. 基于卷积神经网络的文本分类[D].

[3] 杨斌. 中文文本数据挖掘研究[D]. 湘潭大学.

[4] 符燕华. Web 文本数据挖掘研究[D]. 同济大学, 2006.

[5] 刘彦保,王文发,王文东.基于聚类分析策略的 Web 文本挖掘方法[J].延安大学学报(自然科学版),2007(04):22-25+29.

[6] 曹晓.文本聚类研究综述[J].情报探索，2016(01):131-134.

[7] https://github.com/cjymz886/text-cnn

[8] 王美荣. 基于卷积神经网络的文本分类算法[J]. 佳木斯大学学报(自然科学版), 2018, v.36；No.154(03):26-29.

[9] 李廷辰, 杨艳. 基于分词聚类技术的微博热点问题挖掘[J]. 教学与科技, 2013(1):8-13.

[10] 刘永芳,金菊良,魏一鸣. 城市区域综合环境质量动态评价的投影寻踪聚类模型[J]. 中国管理科学(z1):443-446.

[11] 杨东, 王移芝. 基于 Attention-based C-GRU 神经网络的文本分类[J]. 计算机与现代化, 2018.

[12] 刘旭. 博客热点话题挖掘方法[D]. 哈尔滨工业大学.

[13] ZHAO Hongfeng, LUO Lei, HOU Yubao. Cresited Ibis Habitat Assessment Using Analytical Hierarchy Processes 基于层次分析法的朱鹮栖息地质量综合评价[J]. 资源科学, 2013, 35(1):50-58.