文本分类在智慧政务系统中的应用

摘要

本文基于对文本分类的研究，利用自适应文本分类算法建立标签的分类模型，借鉴热度概念形成热度矩阵主题模型，最后运用灰色模糊评价模型对留言的答复情况做出评判。

针对问题一，提出了三种文本分类模型（fasttext 、TextCNN 和 RCNN），对附件二的一级标签提出一种基于主题相似性聚类的自适应文本分类算法，基于 CHI 与 Word Count 构建类特征词库，利用 K-means 句子相似度聚类算法和 Adaptive Strategy 算法，将所构建的类特征词库与主题相似性聚类相结合，提出自适应文本分类方法，每个类别可根据自身特征词适应于相对应的文本分类模型，从而对留言集进行分类，并且进行测试。

针对问题二，借鉴热力学中热度这一概念，将留言热度作为关键因子引入计算中，建立热度矩阵主题模型，将发现热点问题并排序的过程转变成最优化问题的求解，从而获取主题的词概率分布并且直接计算出潜在的主题热度，得到热点问题排名，进而对所有留言进行热点排序。

针对问题三，应用灰色模糊理论模型结合最大隶属度原则以及灰度原理的方法，通过讨论答复内容的解决程度、可用性以及客观原因，建立出灰色模糊综合评价体系，并通过第一问的自适应文本分类模型方法将附件 4 中的答复意见分类量化，导入体系模型中计算，得到关于答复意见质量的最终评价结果。

关键字：文本分类；自适应文本分类算法；热度矩阵主题模型；灰色模糊评价体系

目录

摘要.....1
目录.....2
一. 前言.....3
1.1 问题背景.....3
1.2 国内外研究现状.....3
二. 问题重述.....4
2.1 问题一重述.....4
2.2 问题二重述.....5
2.3 问题三重述.....5
三. 问题分析.....5
3.1 问题一的分析.....5
3.2 问题二的分析.....5
3.3 问题三的分析.....6
四、模型的假设.....6
五. 模型的建立与求解.....6
5.1 问题一的求解.....6
5.1.1 模型建立.....6
5.1.2 实验结果与分析.....10
5.2 问题二的求解.....13
5.2.1 热点问题的定义.....13
5.2.2 模型建立.....13
5.3 问题三的求解.....18
5.3.1 答复意见综合评价模型.....18
5.3.2 综合评价及数据处理.....20
参考文献.....23

一. 前言

1.1 问题背景

在日新月异的我国现代信息社会，各类行政信息的应用飞速发展，政府部门获取各种社情民意的信息方式及其来源从以前的单一方式变成了现在的更加多元化，微信，微博，市长电子信箱，阳光政务热线等各种方法层出不穷，同时随着经济发展全球化的高速推进发展，智慧公共政务平台迫切需要一个快速飞跃的发展趋势。传统智慧城市和其他政府部门是按公共业务、管理两个职责分别进行设定的，每个公共部门各司其职，这种协同处理方式仍然存在严重的公共部门管理壁垒，城市基本设施运行与大数据孤立的特点存在于不同的管理局限性，而传统智慧城市政务在基于这种共同弊端的现实情况下已经开始飞速发展。智慧公共政务主要目的是通过搭建一个政府政务信息实时公开新服务平台，供广大市民随时查询相关政府公告及其他便民服务信息，了解政府办事办理流程，追踪政府办件办理状态，随时随地便享受到智慧公共政务，智慧公共政务主要指的也就是如何可以有效实现公共政务信息服务业的高效化，数据实时精细化，响应及时化的智能政务管理工具，可以简单、高效的快速解决普通百姓的各类政务问题，数据分析统计清晰和透明了精细化。同时，智慧电子政务已经是政府面向所有我国公民和政府企业人员提供的一种可以无缝相互对接我国政府内部公共服务的智能电子政务，智慧电子政务表示政府已经需要推动政府公共服务从传统全能型服务转变成为智慧型和智能服务型[1]。

1.2 国内外研究现状

据相关研究人员发现，早在 2011 年 11 月，美国加利福尼亚州已经正式提出全国智慧城市政务体系建设总体框架，其建设目的主要是为了不断提高地方政府的公共服务管理能力[2]。在 2012 年 6 月，韩国政府也为了顺应时代需求发展相继多次提出国家智慧电子政务发展计划，而这些也使得韩国始终一直居于联合国全国电子政务质量指数行业排名系统中的绝对领先地位。而在国内，随着工业大数据，云计算，人工智能以及移动端和互联网，以及对大数据挖掘等的快速健康

发展，智慧电子政务在基于这些的技术基础上在未来不断提高我国政府服务办公室在服务上的便民已经逐渐成为一种符合时代的发展趋势，智慧电子政务已经是我们实现我国电子政务快速健康发展的重要途径，促使我国政府从一个管理型政府转变成一个智慧型，服务型。早在 2012 年，我国许多重点城市已经正式启动实施智慧新型城市政务建设总体规划，其中一些城市经济发达的重点城市北京、上海、南京、广东、浙江等，已经率先正式进入我国智慧新型城市建设实施准备阶段，而在许多智慧新型城市规划和项目建设中，智慧城市政务则一直是城市建设中的重点[3]。根据中国产业发展信息网最新发布的《2014-2018 年中国电子政务市场分析预测及发展趋势研究报告》:2013 年中国今年电子政务平台总体复合投资规模大约为 1，634.20 亿元每年人民币，2009 年以来的年均规模复合投资增长率大约为 17.35%。2014 年中国全球电子政务应用总体市场投资规模大约为 1，915 亿元每年人民币，未来五年我国全球电子政务应用市场投资规模仍将继续保持较快高速增长。到 2018 年，总体项目投资规模将首次超过 3，400 亿元。这些统计数据充分显示我国智慧电子政务在适应当今社会中越来越重要。而基于文本智慧管理政务的地方文本智慧分类系统即如何建立基于现代自然语言处理分析技术的文本智慧地方政务分类系统已经认为是我国社会经济治理科学创新技术发展的新时代趋势，这对逐步提升地方政府的政务管理水平以及施政管理效率等都具有极大的社会推动力和作用。

二．问题重述

2.1 问题一重述

随着各类网络平台不断开放，政府了解民意的方式也逐渐多元化，各类社情民意相关的文本数据量不断攀升，而这种情况给之前靠人工来划分留言和整理热点的相关部门带来了很大的挑战，并且，在云计算，人工智能，大数据等技术的飞速发展下，建立群众留言文本分类已成目前发展智慧政务的重要方法之一，基于此，本问题将会从群众留言分类着手，建立关于留言内容的一级分类模型，以便后续将群众留言分派至相应的职能部门进行处理，从而提升政府的管理水平以及施政效率，并且达到减少工作量，降低差错率的目的。

2.2 问题二重述

当群众提出问题时，政府需要敏锐的找到群众所提出的热点问题，从而高效有序的解决。热点问题是指某一时段内群众集中所反映的某一问题。相关部门及时的发现热点问题后，可以有效的，有针对性的处理，进而去提高服务效率。本问题将对附件三群众提出的问题从特定时间特定地点或者特定人群进行系统的留言分类，通过定义合理的热度评价指标，给出一套特定且合理的评价体系，从而高效率的解决群众留言中提出的相关问题，从而提高政府处理问题的能力。

2.3 问题三重述

政府在答复群众问题时，会产生以下两种情况，一是在可行性分析合理后制定方案解决了群众问题的情况，二是在法律条文规定或在现有条件限制下，不能立即解决或只能解决部分问题的情况。所以为保证政府对群众的服务质量，我们需对政府的答复意见给出一套合理的评价系统。该系统应从答复和热点问题的相关性，完整性以及对群众的可解释性等角度，进行分析，打分，加权，最后得出答复意见的质量高低。

三．问题分析

3.1 问题一的分析

针对第一问，本文提出了三种文本分类模型（fasttext 、TextCNN 和 RCNN），针对附件二的一级标签提出一种基于主题相似性聚类的自适应文本分类算法，利用 K-means 句子相似度聚类算法和 Adaptive Strategy 算法，将所构建的类特征词库与主题相似性聚类相结合，提出自适应文本分类方法，每个类别可根据自身特征词适应与相对应的文本分类模型，从而对文本进行分类。

3.2 问题二的分析

问题二中，根据已知条件我们借鉴热度的概念，将留言热度作为关键因子引

入计算，建立热度矩阵主题模型，将热点问题发现的过程转换为最优化问题求解，获取主题词概率分布并且直接计算出各潜在主题的热度，进而得到热点排名。

3.3 问题三的分析

问题三中，本节应用灰色模糊理论模型结合最大隶属度原则以及灰度原理的方法，建立灰色模糊综合评价体系，并通过第一问的自适应文本分类模型方法将附件 4 中的答复意见分类量化，导入体系模型中计算，得到关于答复意见质量的最终评价结果

四、模型的假设

⑴假设本文数据集都真实可靠，不需要进行数据集过滤

⑵忽略文本内容长度过小的留言信息。鉴于文本主题挖掘，内容较短的留言大多没有价值，并且会影响算法的效率。

五．模型的建立与求解

5.1 问题一的求解

5.1.1 模型建立

针对第一问，本文对附件二的一级标签提出一种基于主题相似性聚类的自适应文本分类算法，其模型结构如图 1 所示，主要分为以下 3 个部分：

1) 结合 x2 统计(CHI)和 Word Count 方法分别提取附件二中七个类的文本特征词，构成类特征词库。

2) 通过 K-means 算法对附件二的留言主题进行聚类，得到若干个簇，提取每个簇的特征词，构成簇特征词库。

3) 通过自适应文本分类方法 Adaptive Strategy，计算簇特征词库与类特征词库的重叠部分，然后根据重叠部分在簇特征词库中的占比，为每个簇分配一个类标签，从而选取不同的模型，得到分类结果。

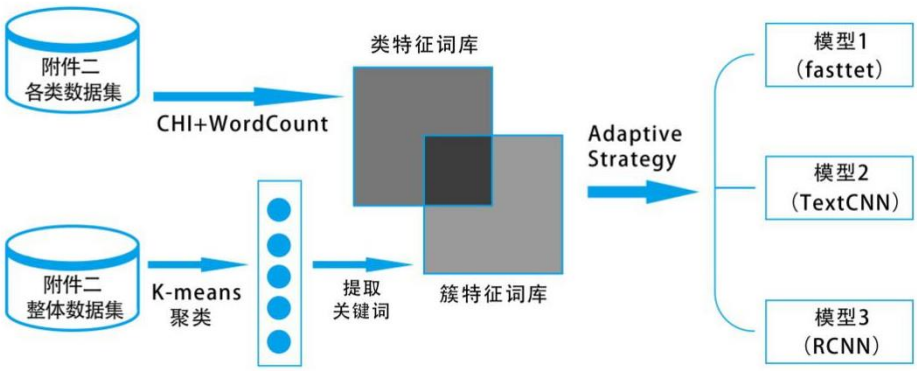

图 1 基于主题相似性聚类的自适应文本分类模型

一、基于 CHI 与 Word Count 的类特征词库构建

[4]研究表明，CHI 是较优的特征提取方法，但是其只考虑特征项出现在所有文本中的频率，而忽略了特征项在某一文本中出现的次数。增加低频词的权重，使得不同类别主题词的交叉情况变得复杂，很难分辨出该特征项在不同类别中的使用频度。因此，本文采用 CHI 方法提取类别中的文本特征，并引入词频因子构建特征词库[5]。例如，在数据集中有 2 个类别 C_i 和 C_j 同时共有特征项 t，t 在 C_i 中出现 100 次，在 C_j 中出现 1 次，则特征项 t 是类别 C_i 的特征比力是 C_j 的特征更具类别特征表示能力。 令 d_i 表示特征项 t 在类别 C_i 中出现的次数，则词频系数 β 可以用式(1)表示。

$$\beta = \frac{\sum_{i=1}^{n} d_i}{\sum_{j=1}^{m}\sum_{i=1}^{n} d_{ji}} \qquad (1)$$

其中，m 表示留言主题总数，C_i 类包含 n 个留言主题，$\sum_{i=1}^{n} d_i$ 表示特征 t 在类别 C_i 中出现的总次数，$\sum_{j=1}^{m}\sum_{i=1}^{n} d_{ji}$ 表示特征 T 在整个文本中出现的总次数因此，CHI 计算公式可以表示为如下形式：

$$x^2(w_i, C_i) = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \times \beta \qquad (2)$$

其中，a 表示包含特征词 w_i 且属于类别 C_j 的留言数，

b 表示包含特征词 w_i 且不属于类别 C_i 的留言数，

c 表示不包含特征词 w_i 且属于类别 C_i 的留言数，

d 表示不包含特征词 w_i 且不属于类别 C_i 的留言数。

词频系数越大表示特征项在类别中出现的频率越高，而在其他类别中出现的次数较少，因此可以作为本类的特征。反之，词频系数越小表示该特征项在本类别中出现的频率越低，而在其他类别中出现的次数较多，因此不适合作为本类的特征，而将类提取的特征作为类特征词库元素。对于全局特征提取，可根据式（2）计算出相关程度，再利用式（3）或式（4）计算全局 x^2 值。

$$x^2_{\max}(w_i) = \max_{1\le i\le m}\{x^2(w_i, C_i)\} \qquad (3)$$

$$x^2_{avg}(w_i) = \sum_{j=1}^{m} P(C_i)x^2(w_i, C_i) \qquad (4)$$

二、依赖于主题相似性聚类的文本标签

为了区分不同类的主题，需要将文本按照句子相似度打上聚类处理，并预先对文本进行标签，便于不同主题选取各自的模型。为了区分不同类的主题交叉情况，本文采用 K-means 算法和预训练的 word2vec 词向量计算文本相似度，以自动确定聚类中的 K 值。具体算法如算法 1 所示。

算法 1 K-means 句子相似度聚类算法：

输入包含 n 个对象的文本数据集 D，预训练好的词向量（dim=300），相似度阈值 m；

输出聚类结果（包含若干个簇）；

步骤 1 对文本数据集 D 进行预处理，使用 nltk 进行分词并去除停用词，此时文本数据集 D 表示为 $D = \{d_i \mid d_i = \{w_{i1}, w_{i2}, \cdots w_{in}\}, i = 1,2,\cdots,n\}$；

步骤 2 导入预训练的词向量，查找出文本中每个词 d_i 的词向量，对句中所有词向量求平均值，得到文本向量 $D=S_i$，其中 i=1，2，…，n；

步骤 3 随机获取句中的一个文本对象，将其向量均值作为初始的聚类中心向量 C_1，中心点为 c_1；

步骤 4 根据文本向量 S_i 与簇的中心向量 C_1 计算余弦相似度；

步骤 5 若步骤 4 中得到的余弦相似度值大于或等于阈值 m，则将文本向量 S_i 聚在一个簇中，并将簇中所有句子向量的平均值作为新的簇中心点；反之，如果

得到的余弦相似度小于阈值 m，则将这个文本向量 S_i 作为簇中心创建新的簇；

步骤 6 若簇中的文本对象不超过 10 个，则降低阈值 m，将较少的对象重新分配到已存在的簇中，并重复步骤 4 和步骤 5。

三、自适应文本分类

本文采用 fasttext 、TextCNN 和[6] RCNN （Recurrent Convolutional Neural Network）3 种分类模型。其中， fasttext 与现有的分类器不同，它是一种简单、高效且具有浅层网络的分类器，使用向量表征单词的 N-Gram 特征，并将局部词序考虑在内，以缩小线性模型和深度模型之间的差距，提高文本分类的准确率和效率[7]。TextCNN 将卷积神经网络应用在文本分类中，使用预训练的词向量完成句子级别的分类任务，并通过采用多个不同尺寸的卷积核捕捉文本中不同 尺寸卷积核的文本特征。RCNN 利用单词表示和循环结构捕捉文本上下文信息，与传统的基于窗口的神经网络相比，RCNN 减少了噪声的引入，并使用最大池化层自动判断词语在文本分类中的重要程度，以捕捉文章的关键信息[8]。

可以看出，仅使用一种网络模型进行分类，容易造成对不同类数据的敏感度不同。为了弥补这一缺陷，实现不同模型间的优势互补，本文将所构建的类特征词库与主题相似性聚类相结合，提出自适应文本分类方法 Adaptive Strategy，具体描述如算法 2 所示。

算法 2 Adaptive Strategy 算法：

输入类特征词库 $ClassFDict_{im}$（下标 im 表示第 i 类中特征项的个数为 m，i=1，2，3，4），算法 1 的聚类结果$\{c_1, c_2, \cdots, c_l\}$，l 表示簇号；

输出测试集类别；

步骤 1 使用 TF-IDF 方法提取聚类结果$\{c_1, c_2, \cdots, c_l\}$的关键词，得到簇特征词库集合$\{W_{ln}\}$，其中，下标表示第 l 个簇中特征词的个数为 n；

步骤 2 根据式（5）计算特征词库和簇特征词库的重叠部分 P_i，具体如下：

$$P_i = \{\{T_{im}\} \bigcap \{W_{ln}\}\} \qquad (5)$$

步骤 3 通过式（6）计算重叠部分 P_i 在簇特征词库 W_{ln} 中的占比，具体如下：

$$r_i = \frac{P_i}{W_{ln}} \qquad (6)$$

步骤 4 利用式（7）、式（8）计算重叠部分 P_i 在簇特征词库 W_{ln} 中的最大占

比，将该簇 l 标记为第 i 类；

$$c_{1i}=\max\{r_i\} \qquad (7)$$

$$c_{1i} \subseteq C_i \qquad (8)$$

步骤 5 根据模型数据的敏感度，选取 C_i 的模型 $Model_j$，j=1，2，3 分别对应 3 种分类模型 fasttext、TextCNN 和 RCNN；

步骤 6 利用模型 $Model_j$ 进行分类，得到最终的分类结果。

5.1.2 实验结果与分析

一、实验数据

本文实验采用留言主题分类数据集，训练集和测试集不重叠。该新闻数据集包括 7 类，分别是城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生，其中，数据集共有 9210 条数据，每类分别为 2010 条、940 条、610 条、1590 条、1970 条、1215 条、875 条。每类取 300 条数据作为测试集，共有 2100 条数据。

3.2 评测指标

文本分类常用的评测标准有查准率 P、查全率 R 和 F_1 值等。其中，查准率 P 是指文本正确分类条数 T_c 与 文本实际分类条数 T_s；的比值，其计算公式如下：

$$P = \frac{T_C}{T_S} \times 100\% \qquad (9)$$

查全率 R 是指文本正确分类条数 T_c 与原有文本信息条数 T_y 的比值，其计算公式如下：

$$R = \frac{T_C}{T_y} \times 100\% \qquad (10)$$

F_1 综合考虑查准率 P 和查全率 R 其计算公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \qquad (11)$$

二、结果分析

本文实验采用 CHI 和 WordCount 相结合的方法对每个类进行特征选择，构成类特征词库。根据式(2)计算特征项的 CHI 值，按照从大到小进行排序后选取前

30%的词作为每个类的特征词。最终在训练集中，七个一级标签，城乡建设、环境保护、交通运输、教育文本、劳动和社会保障、商贸旅游、卫生计生类分别包含 4595 个、2981 个、1812 个、3663 个、3828 个、3342 个和 2791 个特征词。本文从每类中选取 10 个特征词，见表 1。

表 1　每类提取到的特征词

类别	特征词									
城乡建设	改造	工程	建设	规划	租房	房产	建筑	公园	业主	公交
环境保护	污染	环境	排放	噪音	污水	废弃	破坏	排污	垃圾	环保
交通运输	出租	收费	快递	交通	客运	路面	司机	道路	物流	运输
教育文本	教师	小学	中学	教育	补课	学生	学校	收费	招生	培训
劳动和社会保障	职工	工资	退休	保险	员工	咨询	养老	政策	待遇	拖欠
商贸旅游	电梯	传销	收费	垄断	景区	市场	旅游	价格	故障	质量
卫生计生	医院	生育	子女	二胎	卫生	医生	计生	医疗	患者	超生

将 CHI 和 WordCount 相结合的方法与仅使用 CHI 的方法进行对比后发现，本文方法可以有效减少特征词中的低频词，降低低频词的权重，提高特征词质量。

为了研究每个模型对数据的敏感类型，本文对留言主题数据集进行预处理。在每类中随机选取 300 条数据，共 2100 条数据作为测试集，剩下的 7110 条数据作为训练集，然后使用 fasttext、TextCNN、RCNN 3 种模型对其进行训练和测试，结果如表 2一表 4 所示。

表 2　3 种模型在测试集上的查全率对比

模型	城乡建设	环境保护	交通运输	教育文体	社会保障	商贸旅游	卫生计生
fasttext	89.54	94.82	85.62	83.62	95.65	87.81	83.87
TextCNN	85.24	81.64	94.85	86.75	85.41	97.22	92.13
RCNN	87.86	93.16	86.54	90.92	91.20	83.94	91.15

表 3　3 种模型在测试集上的查准率对比

模型	城乡建设	环境保护	交通运输	教育文体	社会保障	商贸旅游	卫生计生
fasttext	90.77	94.13	85.95	88.74	95.62	85.42	85.22
TextCNN	83.99	83.44	91.65	85.91	95.85	97.75	93.45
RCNN	91.85	94.88	86.12	87.66	84.54	81.92	86.77

表 4 3 种模型在测试集上的 F1 值对比

模型	城乡建设	环境保护	交通运输	教育文体	社会保障	商贸旅游	卫生计生
fasttext	90.36	95.90	86.42	87.74	96.41	86.58	86.78
TextCNN	83.16	83.17	92.47	88.91	94.42	85.10	95.32
RCNN	89.80	95.10	95.66	89.66	83.64	92.67	92.10

在一般情况下，R 值越高，其分类模型对数据越敏感。查准率 P 值越高，分类器对数据越敏感，但在某些情况下，其与查全率结果相矛盾。例如，当 R 值为 100% 时，P 值会很低，此时可引入 F_1 值综合分析测试结果， F_1 值越高，模型对数据越敏感。因此，不同模型可根据 R、P 和 F_1 值选取不同类的数据。

根据上述理论分析以及表 2一表 4 的结果可知， fasttext 模型对于城乡建设、环境保护、劳动和社会保障类的查全率和 F_1 值均高于 RCNN 和 TextCNN 模型，因此， fasttex 模型对城乡建设、环境保护、劳动和社会保障类类比较敏感。TextCNN 对于商贸旅游、卫生计生类的查全率、查准率和 F_1 值明显高于 fasttext 和 RCNN 模型，因此，TextCNN 对商贸旅游、卫生计生类比较敏感。RCNN 模型对于交通运输和教育文体类的查全率和 F_1 值较高,但其查准率较低。从整体上考虑，本文选取 RCNN 模型对 Sci-tech 类进行分类。

表 5　模型分配情况

模型	类别
fasttext	城乡建设、环境保护、劳动和社会保障
TextCNN	商贸旅游、卫生计生
RCNN	交通运输、教育文体类

此分类模型可利用每种模型的优势，对文本进行自适应分类。将算法 1 得到的聚类结果以及CHI与WordCount相结合得到的类特征词库作为输入执行算法2，可以得到不同簇对应的不同模型，然后进行自适应分类。

因此，可利用每种模型的优势，对文本进行自适应分类。本文所选用的 3 种文本分类模型，经测试其总体准确率较为接近，如果选用更好的分类模型，则模型分类准确率差距应尽可能小，不同模型之间能够实现优势互补，达到更好的分类效果。实验结果证明，本文算法可以实现 3 种模型的优势互补，提高分类准确率。

5.2 问题二的求解

5.2.1 热点问题的定义

所谓相关热点问题，也就是具有影响力或用户关注度较高的热点问题，但对于相关留言流量数据而言，问题这个词的概念较为抽象，其中的影响力一般无法直接进行估算，通常主要是根据特定一个时间阶段范围内特定关注人群对该热门话题的某个相关热点留言关注数量与其所关注留言数量的多少而来确定。因此，热点问题的准确发现往往来说是相对滞后的，也就是说当一个相关讨论留言者的数量已经达到一定讨论规模以后，才能被准确识别并提出一个热点问题。同时，留言的热点话题众多，人们经常感兴趣的网络热点问题可能只是其中为数不多的几个，留言噪声大，如何有效应付各种大数据下的网络噪声也因此显得尤为重要。目前在一些热点问题的新发现研究领域，较为先进的数据模型和分析方法的应用效率程度受研究文档内容规模大小影响很大，并且大部分问题研究都要求是先仔细挖掘出相关数据分析中的热点话题，再对其进行一个影响力大的排名，这样不仅仅会增加了过多的冗余数据计算，还可能会因存在噪声中的数据过多而严重影响数据准确性[9]。因此，如何在没有发现相关热点的基础同时直接分析得到其中的热度评价指数与热点排名，将使其具有更大的研究价值和参考意义。

5.2.2 模型建立

根据我们查找到的相关事件在文本处理中的模型，我们不难发现以下几个主

要问题:一、文本处理模型难度高;二、现有文档模型处理方法对文档数据变量规模较敏感，当现有文档数据规模较大时，效率就会明显降低;三、很多人在文本处理中的模型大多仅仅是用到了一个词频，忽略了其他富余的属性;四、话题群和影响力群的排序通常是在一个话题群被发现之后不断进行的，会不断增加冗余数据计算和使用噪声处理数据。为有效解决上述这些问题，本章章节中我们充分借鉴了和使用传统热力学理论中的各种热度分析概念，建立了一个热度分析矩阵中的主题分析模型。该计算模型不仅可以将用户留言的优化热度作为概率计算中的因子用以作为提供留言热点问题发现的重要依据，并将留言热点问题发现的整个过程进行处理转化成一个过程解决最终要优化的热度问题，直接可以计算各板块主题的留言热度平均值，并且用户可以在其中得到各个主题概率分布，进而根据得到最终的留言热度进行排序。

具体流程为：1、将当前的留言录入文档；2、通过文本处理方法将留言逐条分析，得到词热度热度矩阵 α 、词频热度矩阵 β 以及总热度矩阵 A；3、根据热度矩阵建立最优化处理计算条件；4、得到主题热度以及主题词概率分布；5、解得热点话题排序。

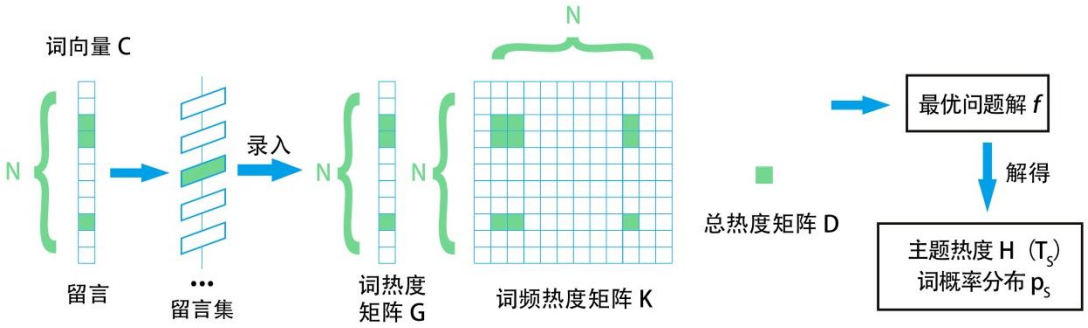

图 2　热度矩阵主题模型流程图

一、热度矩阵

定义 z 作为一个字典处理好的字典留言集，则每条字典留言通过字典处理后都应该可以被表示为形成一个属于词典的向量{$z_{(i)}$，z 规则属于定义 z}，n 为一个字典中的容量，则 1<=i<=n。词尾的向量 $z_{(i)}$可以表示英语字典中第 i 或 i 个形容词在 z 中首次出现的一个次数，|z|则表示其中留言的一个总词的次数。$h_{(z)}$

可以表示一个留言的观测热度，即你只能通过计算已知的留言观测点数值方法求出一个未知的留言目标观测值。定义总热度矩阵为M，表示Z中所有留言的总热度，它的值为Z中所有文档的热度和，即

$$M=\{\sum_{z\in Z} H(z)\} \qquad (12)$$

定义G为词热度矩阵，表示字典中每个词的热度。E中第i个词的热度计算公式为

$$G_i = \sum_{z\in Z} H(z)\cdot \frac{z_{(i)}}{|z|}, 1 \le i \le N \qquad (13)$$

定义词频热度矩阵K，N×N的矩阵，表示字典中动词的共同热度，利用动词同时出现时的热度来挖掘之间的关联性，K中元素的计算公式为：

$$K_{ij} = \begin{cases} \sum_{z\in Z} H(z)\cdot \frac{z_{(i)}^{2} - z_{(i)}}{|z|(|z|-1)}, i = j \\ \sum_{z\in Z} H(z)\cdot \frac{z_{(i)}\cdot z_{(j)}}{|m|(|m|-1)}, i \ne j, 1 \le i \le N, 1 \le j \le N \end{cases} \qquad (14)$$

以上三个热度矩阵中，总热度矩阵和词热度矩阵体现了当前留言集的总体热度和热度在词上的分布，词频热度矩阵则建立了词与词之间的相关性热度，有助于发掘文本数据中的潜在热点话题。

二、确立最优化问题

附件3中给出的留言数据，从热度的角度而言，仅有部分留言有较高的热度，因此，本节经过简单的筛选，仅保留留言数据集中热度最高的前S个主题，并定义每个潜在主题T_S表现为一个固定的词概率分布P_S，用$H(T_S)$来表示该主题的热度，热度越高，排序越靠前。由于每个潜在主题都是非齐次的泊松过程[10]，且每个主题的热度$H(T_S)$与该主题的分布期望正相关。那么由泊松过程的性质可的，文档集本身也是一个非齐次的泊松过程，因此，其期望热度可近似为所有主题的热度和。结合总热度矩阵M便可得到：

$$M=\{\sum_{s=1}^{S} H(T_s)\} \qquad (15)$$

同样地，可得到热度矩阵G和K的期望值计算公式为：

$$E(G) = \{\sum_{s=1}^{S} H(T_s \cdot P_s)\} \qquad (16)$$

$$E(K) = \{\sum_{s=1}^{S} H(T_S \cdot P_S \cdot P_S^T)\} \qquad (17)$$

综上，为了求解 $H(T_S)$ 和 P_S，只需使式（16）、（17）最小且满足式（15）。此时，由热度矩阵主题模型转换成的最优化问题为：

$$\min f = W_e \cdot e_G + W_k \cdot e_F \qquad (18)$$

$$\begin{cases} M = \{\sum_{s=1}^{S} H(T_S)\} \\ \sum_{i=1}^{n} P_{s,i} = 1, 1 \le s \le S \\ P_{s,i} \ge 0, 1 \le s \le S, 1 \le i \le N \qquad (19) \\ \text{其中，} \\ e_G = \sum_{i=1}^{N} (\sum_{s=1}^{S} H(T_S), P_{s,i} - G_i)^2 \\ e_K = \sum_{i=1}^{N} \sum_{j=1}^{N} (\sum_{s=1}^{S} H(T_S) \cdot P_{s,i} \cdot P_{s,j} - K_{i,j})^2 \end{cases}$$

f， e_G， e_K 是热度矩阵的期望和观测值的误差和。

三、热度矩阵求解

根据上述分析模型，只需预先求解一个最新的优化分析问题便能即可直接得到在大数据分析中的潜在热点话题及其关注热度。根据长尾函数效应[11]我们对其结果进行分析优化，热门搜索问题所需要涉及所用到的关键词汇中出现问题次数几乎不会很小，而一个出现问题次数远远低于某个问题阈值的关键词对于整个问题数据集的发展趋势及其影响几乎可以忽略不计，甚至可说是无任何意义的。因此，设定一个字典阈值，加入一个运算从而可以减小其在字典中的容量，降低加入运算时的成本。为方便求解，我们用一个 S×1 的矩阵 L 表示 S 个主题热度的集合矩阵:

$$L = \{H(T_S)\}_{k=1}^{s} \qquad (20)$$

然后使用梯度算法来解决该最优化问题。根据式（19）中目标函数 f 对变量的一阶导 $\frac{\partial f}{\partial L}$，$\frac{\partial f}{\partial P_s}$ 和二阶导 $\frac{\partial^2 f}{\partial L \partial L^T}$，$\frac{\partial^2 f}{\partial P_s \partial P_s^T}$，利用迭代法来更新变量。

根据附件 3，我们保留词频大于平均词频的活跃词汇，最终建立容量为 N=14754 的字典。挖掘出的热点问题及其排序（节选）如下表所示

表 6　热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	68.9	2019/4/1 至 2019/5/31	A 市部分小区/小区住户	存在一系列安全隐患问题
2	2	26.4	2019/9/1 至 2019/9/31	A3 区/住户	扰民
3	3	23.3	2019/2/20 至 2019/2/25	A 市丽发新城/小区住户	车贷、房贷、虚假业务等违法犯罪行为
4	4	17.4	2019/5/1 至 2019/5/31	A 市各个施工地区/公民	施工
5	5	13.7	2019/11/1 至 2019/11/30	A4 区绿地海外滩小区/小区住户	投诉不合理行为

表 7 热点问题留言明细表（节选）

问题 ID	留言编号	留言用户	留言主题	留言时间	反对数	点赞数
1	208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19 11:34	0	2097
1	262052	A00072424	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	2019/3/26 14:33:47	0	78
1	267630	A000100648	反映 A 市地铁 3 号线松雅湖站点附近地下通道问题	2019/5/22 23:37:38	0	42
…	…	…	…	…	…	…
2	263672	A00041448	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？	2019/9/5 13:06:55	0	669
2	284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊	2019/1/10 15:01:26	0	80
2	253369	A00074795	穿 A 市城而过的京港澳高速（长楚高速）什么时候可以外迁至远郊？	2019/11/18 15:35:11	0	29

…	…	…	…	…	…	…
3	217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	0	790
3	262052	A00072424	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	2019/3/26 14:33:47	0	78
3	202909	A00052058	A 市无良万润滨江天著牟取精装暴利	2019/1/3 19:37:36	0	28
…	…	…	…	…	…	…
4	262052	A00072424	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	2019/3/26 14:33:47	0	78
4	279062	A00027836	建议加大 A7 县东六线榔梨段拆迁力度	2019/1/17 19:25:45	1	42
4	267630	A000100648	反映 A 市地铁 3 号线松雅湖站点附近地下通道问题	2019/5/22 23:37:38	0	42
…	…	…	…	…	…	…
5	193091	A00097965	A 市富绿物业丽发新城强行断业主家水	2019/6/19 23:28:27	0	242
5	281898	A00096623	A 市长房云时代多栋房子现裂缝，质量堪忧	2019/2/25 15:17:38	5	55
5	253369	A00074795	穿 A 市城而过的京港澳高速（长楚高速）什么时候可以外迁至远郊？	2019/11/18 15:35:11	0	29
…	…	…	…	…	…	…

5.3 问题三的求解

5.3.1 答复意见综合评价模型

一、答复意见评价影响因素分析

答复意见评价影响因素分为已解决、未解决、问题不合理和客观原因导致无法解决这 4 个一级指标，以及 6 个二级指标。该体系的指标关系见下图 3。

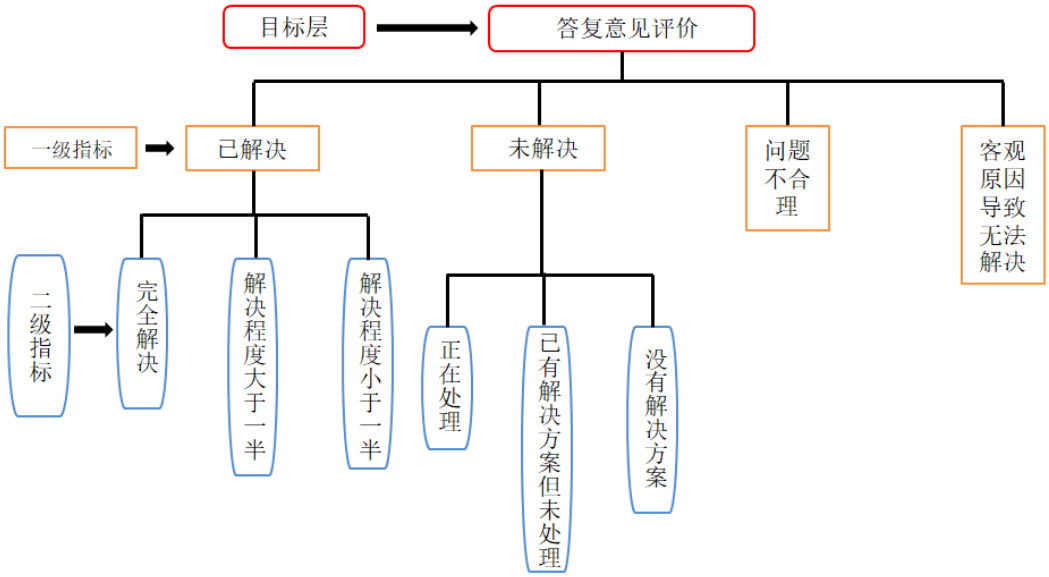

图 3 答复意见评价指标体系

假设因素集$U=\{u_1,u_2,\cdots u_n\}$;答复集$B^*=\{c_1^*,c_2^*\cdots c_n^*\}$

二、权重集与评价矩阵的建立

权重集表示了答复意见与其影响指标之间的灰色模糊关系[12]。依据上图给出的影响指标的层级关系，给出其相应的权重和，确定同层元素的相对重要性，构成如下的权重集：

$$\underset{\otimes}{\bar{A}}=[(a_1,c_1^*)(a_2,c_2^*)\cdots(a_m,c_m^*)] \qquad (21)$$

归一化处理，即

$$\sum_{i=1}^{m}a_i=1 \qquad (22)$$

单纯根据附件 4 中的答复意见我们难以准确将其量化分析，所以，依据第一问中的方法将答复意见情况大致分为 5 类：{完全解决，一般，仅有处理方案，未解决，无法解决}，对应的灰度值为:

$$\{0\sim0.2,0.2\sim0.4,0.4\sim0.6,0.6\sim0.8,0.8\sim1.0\}$$

评价的主体是答复意见的影响因素集和答复集，他们之间联系用评价矩阵(23)表示。

$$\underset{\otimes}{\bar{B}} = \begin{pmatrix} (a_{11}, c_{11}^*) & \cdots & (a_{1n}, c_{1n}^*) \\ \vdots & \ddots & \vdots \\ (a_{m1}, c_{m1}^*) & \cdots & (a_{mn}, c_{mn}^*) \end{pmatrix} \qquad (23)$$

在评价矩阵建立的过程中，既给了不同评价因素相应的隶属度，又根据数据获得的分类确定各自对应的灰度，得到评价矩阵各元素的不同分量。

5.3.2 综合评价及数据处理

一、数据处理

为保留尽可能多的信息，我们将灰色模糊评价运算分为两个部分进行。模部运算采用 $M(-,+)$ 算子，灰部运算采用 $M(/,+)$ 算子。一级综合评价得到的结果为

$$\underset{\otimes}{\bar{Q}} = \underset{\otimes}{\bar{A}} \circ \underset{\otimes}{\bar{B}} = [(q_j, c_{kj}^*)_K] = [\{(\sum_{k=1}^{m} a_k \cdot u_{kj}), \prod_{k=1}^{m} (\mid \Lambda(c_k^* + c_{kj}^*))\}] \quad (24)$$

子因素 u_i 是其上一级因素集 U 的元素，研究答复意见的影响因素集 U 与答复集 C^* 之间的灰色模糊关系为 $\underset{\otimes}{\bar{Q}} = [\underset{\otimes_1}{\bar{Q}}, \underset{\otimes_2}{\bar{Q}}, \cdots, \underset{\otimes i}{\bar{Q}}]^T$ 结合权重集 $\underset{\otimes}{\bar{A}}$ 得到答复意见想想因素评价对象 U 的综合评价向量：

$$\underset{\otimes}{\bar{Q}} = \underset{\otimes}{\bar{A}} \circ \underset{\otimes}{\bar{B}} = [(q_1, c_1^*)\ (q_2, c_2^*) \cdots (q_m, c_m^*)] \qquad (25)$$

在得到的评价结果中，研究答复意见是否解决留言是由灰度来镖师，二答复意见的可信程度则是有相应的白度表示。根据上述的研究，我们对研究结果做进一步处理。

若 $q_i \geq q_j$，则 $q_i \geq q_j$ 的可信度为 $p_{ij} = (1 - c_i^*) \cdot (1 - c_i^*)$

反之，$b_i \leq b_j$ 的可信度为

$$p_{ij} = 1 - (1 - c_i^*) \cdot (1 - c_j^*)\ (i, j = 1, 2, \cdots, m; i \neq j)$$

则在上述评价结果中，隶属度最大的可信度为：

$$p_i = \prod_{j=1}^{m} p_{ij} (i = 1, 2, \cdots, m)$$

即为最终评价结果。

二、评价结果

我们将附件 4 中的答复意见做量化处理。运用 AHP 改进，计算得到权重，做一致性检验：

$$\lambda_{\max} = \frac{1}{n}\sum_{i=1}^{n}\frac{(AW)_i}{W_i}$$

$$CI = \frac{\lambda_{\max} - n}{n-1}$$

$$RI = 0.65$$

$$CR = \frac{CI}{RI} = 0.043 < 0.1$$

由于$CR < 0.1$，即通过一致性检验。

将答复意见评价的答复集设为$C^* = \{很好, 较好, 一般, 较差, 很差\}$

表 8　评价指标、评分和评分灰度

一级指标	权重	二级指标	权重	评分					灰度
				很好	较好	一般	较差	很差	评分
已解决	0.596	完全解决	0.38			√			0.2
		解决程度大于一半	0.103				√		0.1
		解决程度小于一半	0.463		√				0.4
未解决	0.279	正在处理	0.596	√					0.5
		已有解决方案但未处理	0.463		√				0.4
		没有解决方案	0.279					√	0.1
问题不合理	0.021		0				√		0.2
无法解决	0.0046		0				√		0.2

对于一级指标评价，以已解决为例：

$$\underset{\otimes_1}{\bar{Q}} = \underset{\otimes_1}{\bar{A}} \circ \underset{\otimes_1}{\bar{B}} = [(0.38,0)(0.103,0)(0.463,0)] \circ \begin{bmatrix} (0,1) & (0,1) & (0.2,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (0,1) & (0.1,1) & (0,1) \\ (0,1) & (0,4) & (0,1) & (0,1) & (0,1) \end{bmatrix}$$

$$= [(0.26,0.3)(0.31,0.1)(0.54,0.03)(0.38,1)(0.103,0)]$$

同理得到其他一级指标的评价

$$\underset{\otimes_2}{\bar{Q}} = [(0.33,0.1)(0.24,0.2)(0.0.104,0.01)(0.1,1)(0.2,0.14)]$$

$$\underset{\otimes_3}{\bar{Q}} = [(0,1)(0,1)(0,1)(0.2,1)(0,1)]$$

$$\underset{\otimes_4}{\bar{Q}} = [(0,1)(0,1)(0,1)(0.2,1)(0,1)]$$

再由式（25）得到二级指标评价

$$\underset{\otimes}{\bar{Q}} =$$

$$[(0.596,0)(0.279,0)(0.021,0)(0.0046,0)] \cdot \begin{bmatrix} (0.26,0.3) & (0.31,0.1) & (0.54,0.03) & (0.38,1) & (0.103,0) \\ (0.33,0.1) & (0.24,0.2) & (0.104,0.01) & (0.1,1) & (0.2,0.14) \\ (0,1) & (0,1) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (0,1) & (0,1) & (0,1) \end{bmatrix}$$

$$= [(0.061,0.004)(0.215,0.0017)(0.368,0.0006)(0.214,0.0008)(0.0052,0.025)]$$

对评价结果进行分析，得到：

$$P_1 = 0.061, P_2 = 0.215, P_3 = 0.368, P_4 = 0.214, P_5 = 0.0052$$

又因为

$$g(\underset{\otimes}{\bar{Q}}) = \frac{1}{s}\sum_{i=1}^{S} g(\underset{\otimes}{\bar{Q}})$$

$$= \frac{0.004 + 0.0017 + 0.0006 + 0.0008 + 0.025}{5} = 0.0642 < 0.5$$

即灰度较小，说明我们所做的评价较充分。

根据最大隶属度原则，P_3最大，则P_3为最终评价结果，也就是说，该答复意见的质量为“仅有解决方案”。也就是说答复意见并没有非常好的解决广大居民的问题，多数是提出一个解决方案，有待进一步的实施。

参考文献

[1]卢时彻. 建设智慧政府，推动城镇信息化的战略研究[J].中国信息界 2014(6)

[2]CA Fwd. The Smart Government Framework[EB/OL]. (2011-11-03)

[3]陈丽容. 郑爱军: 智慧城市建设 首要任务是建设智慧政府[N].通信信息报。2012

[4]YANG Y， PEDERSEN J O. A comparative study on feature selection in text categorization [C]//Proceedings of International Conference on Machine Learning.New York ，USA ： ACM Press， 1997 :412-420.

[5] 周庆平，谭长庚，王宏君，等.基于聚类改进的 KNN 文本分类算法[J].计算机应用研究，2016，33(11):3374-3377.

[6] 龚千健. 基于循环神经网络模型的文本分类[D].华中科技大学，2016.

[7] 郝志峰，黄浩，蔡瑞初，温雯.基于多特征融合与双向 RNN 的细粒度意见分析[J].计算机工程，2018，44(07):199-204+211.

[8] 康雁，杨其越，李浩，梁文韬，李晋源，崔国荣，王沛尧.基于主题相似性聚类的自适应文本分类[J].计算机工程，2020，46(03):93-98.

[9]聂文汇，曾承，贾大文.基于热度矩阵的微博热点话题发现[J].计算机工程，2017，43(02):57-62.

[10] Ross S M. Stochastic Processes [M] . New York， USA :John Wiley & Sons Inc. ，1996.

[11]储姗姗. 视频网站推荐算法的研究与应用[D].北京邮电大学，2018.

[12]陈方芳.高校大学生数学素质的灰色模糊综合评价模型[J].鞍山师范学院学报，2019，21(06):6-11.