基于文本挖掘的“智慧政务”处理方法

摘要

本文在进行了完整的数据预处理操作后，基于朴素贝叶斯算法将留言问题进行一级标签分类，并使用 F-Score 算法评价，根据热点问题聚类数量不确定的特点，选择 DBSCAN 进行文本聚类，得到热点问题排名，同时，选用了可靠的指标分析答复意见的质量，并采用组合赋权法得到指标的权重，量化了答复质量的评价。本文的工作对于“智慧政务”有一定的指导意义。

针对问题一，采用朴素贝叶斯算法、逻辑回归算法、K 近邻算法和决策树算法。首先进行文本预处理，其次采用 TF-IDF 算法计算文本特征权重，然后选取上述四种算法对模型进行评价，综合考虑 Precision,Recall,F1-Score 各项得分，选取性能最佳的 K 近邻算法为本模型的评价方法。

针对问题二，采用 DBSCAN 算法。综合考虑问题的持续时间、发生地点、相似问题的频率和点赞数各要素，建立热度评价指标。首先进行分词、词性标注和去除停用词等预处理操作来清洗数据，其次采用 TF-IDF 算法计算文本的特征权重并表示成权值向量，然后使用 LSA 模型对 TF-IDF 权值矩阵进行特征抽取，再通过 DBSCAN 算法将特征相似的问题进行聚类，最后用熵权法为各因素赋权，计算出热度指数，对热点问题进行排名。

针对问题三，采用组合赋权法。首先从答复的相关性、完整性和可解释性三个方面分析答复的质量，查阅相关文献，选择问-答关键词的余弦相似度，答复响应时间和答复详细度来定量地衡量答复的质量。然后确定各指标的权重，为使指标的赋权达到主客观相统一，使评价结果更真实可靠，采用组合赋权法，将客观的熵权法与主观的指标评分标准相结合。首先根据相关政策法规和数据的统计分析规律，制定各评价指标的评分标准，进行十分制打分，然后使用熵权法进行权重赋值，最后计算出综合评价指数。

关键词： K 近邻算法 朴素贝叶斯算法 DBSCAN 聚类 余弦相似度 熵权法

一、问题重述

1.1 问题背景

进入 21 世纪，随着互联网信息化时代的来临和智能化手机的普及，中国各地政府通过微博、微信、市长信箱、阳光热线等网络问政平台和广大群众互动已经成为社会治理的新趋势。群众通过在网络问政平台留言反映信息，相关部门对留言信息进行划分和对热点问题进行整理，同时将后续的群众留言分配至相关的职能部门处理，相关部门有针对性地进行处理并在网络平台上反馈给群众。

基于市长信箱、阳光热心、微信、微博的网络问政平台扩宽了公民参与政治的渠道，推进了社会主义民主政治的建设，有利于改进国家政府部门及相关人员的工作。然而，各类民情相关的文本数据量不断增加，给传统的依靠人工进行留言划分和热点整理的工作带来了新的挑战。因此，建立基于自然语言处理技术的智慧政务系统已成为社会治理创新发展的新趋势，对政府部门管理水平和施政效率的提高具有极大的促进作用。

1.2 要解决的问题

根据收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。利用自然语言处理和文本挖掘的方法解决下面的问题：

1.群众留言分类。建立关于留言内容的一级标签分类模型，使用 F-Score 对分类方法进行评价。

2.热点问题挖掘。将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果和排名前 5 的热点问题以及给出相应热点问题对应的留言信息。某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。

3.答复意见的评价。针对相关部门对留言的答复意见，从答复的相关性、完

整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

1.3 国内外研究现状

随着网络的普及，政府部门也越来越重视公民通过网络问政平台反映的舆情信息。国外主要基于 Twitter，Facebook 等网络平台。Budak，Alex（2008）分析了 Twitter 在奥巴马总统选举中的作用[1]。随后， Richard（2011）通过对 60 家政府机构在 Twitter 上发布的 1800 条更新内容的分析[2]。研究了政府相关人员如何利用 Twitter 与公民进行交流。国内主要基于微博、微信、市长信箱、阳光热线等网络问政平台。柳思思（2014）以“北京微博发布厅”为例[3]，对政府官方账号在微博上的交流评论等进行分析。刘星（2017）以长沙市政府电子政务平台为例[4]，针对平台数据开放等方面存在的问题，提出了相应的建议和解决对策，加快了长沙市电子政务平台建设发展。

然而，简单地依靠人工识别和分类网络问政平台的信息具有工作量繁琐、耗时长、成本高昂等缺点。因此，利用自然语言技术处理网络政务信息已成为时代发展的新趋势。其中，国外研究主要有：Herreraviedma（2006）提出了一种基于词模糊计算的网站信息质量评价方法[5]。它根据用户的感知对网站的信息质量生成语言建议。Huang（2012）采用单通道（Single-pass）聚类技术[6]，利用隐含的狄利克雷分布（LDA）模型代替传统的向量空间（VSM）模型，提取隐藏的热点话题数据。国内研究主要有：朱梦（2016）以文本分类依托[7]，对朴素贝叶斯算法进行改进，提出了结合综合特征词分布情况的 k-Bayes 算法，提升文本分类效率。同时，王明亚（2016）提出基于词向量的文本分类算法研究与改进[8]，重点研究了词向量训练工具 word2vec 的工作原理，并将其运用到传统的基于卡方检验特征选择算法的文本分类中。参考前人的研究，针对本文要解决的问题，建立文本分类、文本聚类和答复指标赋权的模型，给出了留言分类的一级标签方案，得到了热点问题的排名，并量化了答复意见的评价，对促进“智慧政务”有一定的现实指导意义。

二、问题分析

本文要求建立留言内容的一级标签分类模型，定义合理的热度评价指标，给出热点问题排名，并从相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

对问题一，充分应用文本分类有关知识，构造关于留言内容的一级标签分类模型。根据题中所给数据，群众的留言内容由留言编号、留言用户、留言主题、留言时间、留言详情、一级标签六个部分构成。由于留言内容主要由留言详情构成，且留言主题也是对留言详情的概括，因此仅考虑留言详情的分析。为构建一级标签分类模型，首先对留言详情的文本信息进行预处理，减少冗杂信息，为文本分类做准备。其次对得到的文本信息进行向量化表示和特征提取，将文本信息转化为机器能识别的语言，同时保留表征能力较强的关键词。再利用支持向量机模型对文本信息进行分类，根据训练组的文本信息将测试组的文本信息归类，从而得到分类模型，最后对分类模型进行效果评估。

对问题二，题设要求建立合理的热度评价指标并给出评价结果。本题需要综合考虑问题的持续时间、发生地点、相似问题的频率和点赞数等要素，其中的重点在于命名实体识别和问题的文本相似度判断。由于数据冗杂，能够反应热点问题的关键词多为名词和动词等，因此先进行 jieba 分词、词性标注、去重复值和去停用词等预处理操作来清洗数据。考虑到只分析词频可能会导致特征提取存在偏差，因此选择 TD-IDF 算法从词频和逆文章频率两方面来计算文本的特征权重，并将文本用向量表示。由于文本中存在的同义词和多义词现象，造成特征向量的各个分量存在一定的相关性，因此使用 LSA 模型进行特征抽取达到降维的效果。又由于问题数据量大且无法确定聚类的数量，经分析，采用 DBSCAN 算法将相似的问题进行聚类。最后综合考虑各影响因素，建立热度评价指标对热点问题进行排名，选出前 5 名的热点问题。

对问题三，考虑到需要采用具体的评价指标对答复质量进行量化分析，参考相关文献并结合具体实际，选择问-答关键词的余弦相似度，答复的响应时间和答复详细度来衡量答复的质量。然后确定各指标的权重，主、客观赋权法各具有

优缺点，为使指标的赋权达到主客观相统一，进而使评价结果更真实可靠，采用组合赋权法，将客观的熵权法与主观的指标评分标准相结合。通过分析数据，发现答复响应时间和答复长度的数据离散度很高，但它们并非可以对综合评价结果起决定作用，若直接使用熵权法，会造成很大误差，因此需要对数据进行整理。首先根据相关政策法规和数据的统计分析规律，制定各评价指标的评分标准，进行十分制打分，然后使用熵权法进行权重赋值，最后计算出综合评价指数。

三、模型的假设与约定

1、假设每条留言内容之间没有相关性，相互独立；

2、假设分类算法的计算量可以忽略，仅考虑使用 F-Score 评价分类模型效果；

3、假设当存在至少 10 个不同用户在特定时间段提出相似的问题时，就可以认为这是一个热点问题；

4、假设依据政策法规和数据统计分析规律制定指标的十分制标准是可靠的。

四、符号说明

符号	含义
TF	词频
IDF	逆文章频率
min_samples	邻域密度阈值
eps	半径参数
E_j	所有答复对指标 X_j 的贡献总量
D_j	第 j 指标下各答复贡献度的一致性程度
W_j	第 j 个指标的权重值
C_j	第 j 个指标的无量纲量化值
ESI	综合评价指数
n	评价指标个数
m	答复意见的总数

五、模型的建立与求解

5.1 群众留言分类

进入大数据时代，网络问政平台上相关社情民意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。因此我们希望建立留言分类模型，将网络群众留言自动分类并分配到具体的部门，减小相关部门工作量，提高工作效率，及时反映民情，为下一步的政府规划做准备。

为建立留言分类模型：首先将留言详情导入 Python，经过缺失值处理、重复值处理、文本内容清洗，基于理解的分词算法，利用 jieba 库对留言详情进行中文分词，得到单独的词或词语。再对分词处理后的文本去除停用词，减少冗杂信息，停用词是指在我们语句中大量出现，但却对语义分析没有帮助的词。对于这样的词汇，我们通常可以将其删除，这样的好处在于：可以降低存储空间消耗、可以减少计算时间消耗。由于机器无法理解文本信息，因此需要对去除停用词后的文本信息进行向量化及特征提取，采用文档-逆文档频率（TF-IDF）的方法，将非结构化的数据转化为结构化的数据，提取重要的关键词。再利用文本分类的方法对特征提取后的文本归类，比较不同分类方法的差异，选择最优的分类方法，最后对分类模型的效果进行评估。具体流程图如图 1 所示。

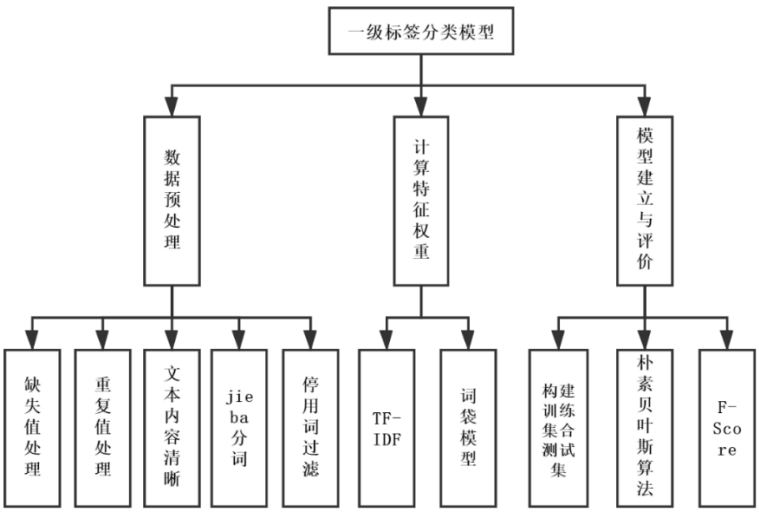

图 1 一级标签分类模型过程示意图

5.1.1 数据预处理

本文需要处理的留言详情属于非结构化数据，非结构化数据无法直接使用，需要转化为计算机能够理解的语言，因此需要对文本数据进行预处理，为进一步文本向量化和特征提取做准备。数据预处理是对群众留言文本分类的第一步，主要包括缺失值处理、重复值处理、文本内容清洗、中文分词、去除停用词。

Step 1：缺失值、重复值处理

文本数据存在空字段和多个文本数据重复的情况，对缺失值和重复值进行处理，为文本向量化和特征提取做准备，以防模型结果产生误差。

Step 2：文本内容清洗

文本数据中含有对分析作用较小的标点符号与特殊字符，对于这类词语，使用 re 库中正则匹配方法去除。

表 1 文本内容清洗前的示例数据

编号	留言详情
0	\n\t\t\t\t\t\n\t\t\t\t\t 座落在 A 市 A3 区联丰路米兰春天 G2 栋 320，...
1	\n\t\t\t\t\t\n\t\t\t\t\tA 市 A6 区道路命名规划已经初步成果公示文件，...
2	\n\t\t\t\t\t\n\t\t\t\t\t 本人系春华镇金鼎村七里组村民，不知是否有相关...
3	\n\t\t\t\t\t\n\t\t\t\t\t 靠近黄兴路步行街，城南路街道、大古道巷、一步...
…	…
4321	\n\t\t\t\t\t\t\t\t\t 关于西地省 A 市经济学院寒假过年期间组织学生去工厂工作...
4322	\n\t\t\t\t\t\t\t\t\t\t 一名中职院校的学生,学校组织我们学生在外边打工,...

表 2 文本内容清洗后的示例数据

编号	留言详情
0	座落在 A 市 A 区联丰路米兰春天 G 栋一家名叫一米阳光婚纱艺术摄影的...
1	A 市 A 区道路命名规划已经初步成果公示文件什么时候能转化成为正...
2	本人系春华镇金鼎村七里组村民不知是否有相关水泥路到户政...
3	靠近黄兴路步行街城南路街道、大古道巷、一步两搭桥小区...
…	…
4321	关于西地省 A 市经济学院寒假过年期间组织学生去工厂工作过年...
4322	一名中职院校的学生学校组织我们学生在外边打工在外省做流水线...

Step 3：jieba 分词

文本分词包括中文分词和英文分词。对于英文文本，可以以单词间空格进行分割；而对于中文文本，由于表达的复杂性、模糊性，导致中文分词不能简单的以空格作为划分。

目前，中文分词主要有三种方法：

（1）基于词典的分词。基于词典的分词方法也称为机械分词法，它将中文文本与词典中的词进行匹配，如果遇到词典中存在的词语，则匹配成功，将其标识为一个词语。常用的匹配方法包括最大正向匹配法，从文本左边向右边匹配，去掉最右边的字符；逆向匹配法，从文本右边向左边匹配，去掉最左边的字符。一般而言，逆向匹配法更符合人们阅读习惯，因为分词效率更高。

（2）基于统计的分词。基于统计的方法可以与机器学习算法一同使用，一个待切分的文本可能包含多种分词结果，相邻的字共同出现的次数越多，那么它们构成一个词的概率也就越大，将其中概率最大的那个作为字符串的分词结果。基于统计的分词优点是对歧义词和未登录词的识别都具有良好的效果；缺点是训练周期较长，计算量较大。常用的基于统计的方法主要有隐马尔可夫模型、最大熵模型、N 元统计模型等。

（3）基于理解的分词。基于理解的分词是模拟计算机学习人的思维对于句子的理解，从而剖析出句中的词汇结合方法。随着中文分词工作的推进，专家学者开发了一系列的中文分词系统和工具。1984 年，我国第一个自动中文分词系统（CDWS）被开发出来，随后各类分词系统被开发利用，例如 CASS 汉语自动分词系统、ABWS 分词系统、SEG 分词系统等。

目前较为流行的分词工具是 jieba 分词。jieba 库采用动态规划查找最大概率路径，找出基于词频的最大切分组合。同时对于未登录的词或歧义词，基于隐马尔可夫模型，使用了 Viterbi 算法，具有较好的识别效果，被认为是最好的 Python 中文分词组件。利用 jieba 库对留言详情进行分词后得到的部分结果展示如下表。

表 3 jieba 分词处理后的示例数据

编号	留言详情
0	[座落在, A, 市, A, 区联, 丰路, 米兰, 春天, G, 栋, 一家, 名叫, 一...
1	[A, 市, A, 区, 道路, 命名, 规划, 已经, 初步, 成果, 公示, 文件, 什...
2	[本人, 系, 春华, 镇金鼎村, 七里, 组, 村民, 不知, 是否, 有, 相关, 水泥...
3	[靠近, 黄兴路, 步行街, 城, 南路, 街道, 、, 大, 古道, 巷, 、, 一步, ...
…	
4321	[关于, 西地省, A, 市, 经济, 学院, 寒假, 过年, 期间, 组织, 学生, 去,...
4322	[一名, 中职, 院校, 的, 学生, 学校, 组织, 我们, 学生, 在, 外边, 打工,...

Step 4：去除停用词

在中文文本中有很多无效的词，例如“的”、“哦”、“和”等介词或连词，停用词的存在增加了文本词语的维数，容易造成维数灾难，使计算变得复杂，因此需要去除。

引入停用词库（或通过参数 stop_words 引入停用词数组），将分词后的文本数据与停用词库进行匹配，剔除匹配成功的停用词，得到去除停用词后的文本分词，称为文本特征词。得到文本特征词后，进一步绘制词云，对留言详情中出现频率较高的关键词予以视觉突出，频率越高字体越大，得到的词云图如图 2 所示。

图 2 词云图

5.1.2 文本向量表示及特征提取

留言详情经过预处理后得到文本特征词，然而计算机本质上只能识别 0 和 1，因此需要将文本特征词向量进一步地向量化表达，在语义信息尽可能完整的前提下，将其转换为一种结构化的数据格式。常见的文本向量化的方法有 One-hot 编码和词袋模型。One-hot 编码可以把文本特征词转化为向量，向量的维度即特征词的长度，向量在对应特征词的位置是 1，其余是 0，One-hot 编码省去了复杂的计算，但当特征词非常多时，One-hot 编码表示的向量将会是高维度的稀疏向量，不利于计算机的计算，且存在较大的浪费。与 One-hot 编码相比，词袋模型计算了特征词出现的次数。在词袋模型中，每条处理后的留言详情作为一个样本，每个不重复的词语作为一个特征词，通过统计特征词在词典表中出现的频率来表示每个词的向量值。

本文选择词袋模型对文本特征词向量化表示，特征词的原始维度可能很大，为了避免维数灾难，需要我们进一步特征选择，以获得文本表示能力强、更接近主题的特征。利用文档-逆文档频率（TF-IDF）方法来进行特征选择。TF-IDF 方法在词频的基础上，对每个词分配一个“重要性”的权重，最常见的词（“的”、“是”）给予最小权重，较少见的词给予较大权重。TF-IDF 的计算公式可以表示为：

$$TF-IDF = \text{词频}(TF) \times \text{逆文档频率}(IDF)$$

$$\text{词频}(TF) = \frac{\text{某个词在留言中出现次数}}{\text{留言的总词数}}$$

$$\text{逆文档频率}(IDF) = \log\left(\frac{\text{留言的总条数}}{\text{包含该词的留言条数}+1}\right) \tag{5.1.1}$$

对于留言中的高频术语而言，在包含该术语的总的留言中出现的次数越低，其对留言分类的贡献就越高。

5.1.3 模型建立与评价

对经过特征提取后的留言信息（也称词向量）进行分类：首先构造训练集和测试集，选择 80%的数据作为训练集，选择剩下的 20%的数据作为测试集。将训练集的词向量输入分类器模型进行训练，实现分类。分类器模型的作用是识别分类文本任务，可以分为基于机器的文本分类器和基于深度学习的文本分类器。本文针对问题一的群众留言分类问题，分别选择基于机器的文本分类方法的不同的方法进行比较。基于机器的文本分类方法包括朴素贝叶斯、逻辑回归、集成模型等方法。本文出于篇幅和分类效果的考虑，选择朴素贝叶斯算法、决策树算法和逻辑回归分别对留言内容进行分类，并使用 F-Score 对分类方法进行评价，选出分类效果更好的算法。

Step 1：构建训练集和测试集

经过预处理和特征提取后的数据需要按照一定的比例进行切分，将数据分为训练集和测试集。经验上，数据的切分一般采用 6：4 或 8：2 的比例，并且最好保证每个数据集所在的分布大体一致或相同。

针对问题一的留言分类问题，需要分析的文本数据共有 9104 条，随机选择将 20%的数据为测试集，其余的数据为训练集。

Step 2：朴素贝叶斯算法

朴素贝叶斯分类算法属于基于机器的分类算法，是一种基于贝叶斯定理与特征条件独立同分布假设的生成式分类器。朴素贝叶斯的前提条件是假设特征之间没有相关性，相互独立。其基本原理是：对于不能确定类别的给定样本，分别计算该样本所属类别的概率，然后选择概率最大的类别作为修改后样本数据的类别。当数据集满足假设时，利用朴素贝叶斯进行数据分类的精度很高。由于朴素贝叶斯对缺失数据不敏感，算法较为简单，因此常被用于文本分类中。

设留言问题分类中的一级标签为$\{C_1, C_2, C_3 \dots C_{15}\}$分别表示“城乡建设”类、“党务政务”类、“国土资源”类…“劳动和社会保障”类。留言集合为$D = \{d_1, d_2, d_3 \dots d_n\}$，某留言对于的文本特征向量表示为$T = \{t_1, t_2, t_3 \dots t_n\}$，留言集合的特征向量表示为$T = \{T_1, T_2, T_3 \dots T_n\}$，按照条件概率的定义，可得到公式：

$$P(c|d) = \frac{p(d|c) \times p(c)}{p(d)}$$

(5.1.2)

其中，$P(c|d)$表示留言 d 属于 c 类的概率，p(c)表示先验概率，$p(d|c)$表示在 c 类留言中，留言 d 中的每一特征向量t_i的概率。

利用朴素贝叶斯对留言信息进行分类时，它是通过计算这条留言属于每个类别的概率$p(c_i|d)$，并将最大概率值对应的类别作为该留言最终所属类别，公式表示为：

$$Z_{c_i} = \max_{c_i \epsilon c} \left\{ P(c_i) \times \prod_{j=1}^{n} P(t_i|c_i) \right\}$$

(5.1.3)

其中，$P(c_i)$代表先验概率，$P(t_i|c)$代表条件概率。

朴素贝叶斯分类算法在分类过程中时空开销小，数据处理较快，而当分类的属性相关性较强，属性之间无法独立时，分类效果较差。

Step 3：决策树算法

决策树分类算法属于基于机器的分类算法，是一种基于树形结构的分类模型，通过树结构对分类问题进行逐层的划分，再经过对一系列子决策的判断来完成分类。其基本原理是：在数据集的所有特征中选择最优特征，再选择最优候选分割值，然后根据最佳候选分割值将原始数据集分成两部分，并不断重复至达到要求。例子如图 3 所示。

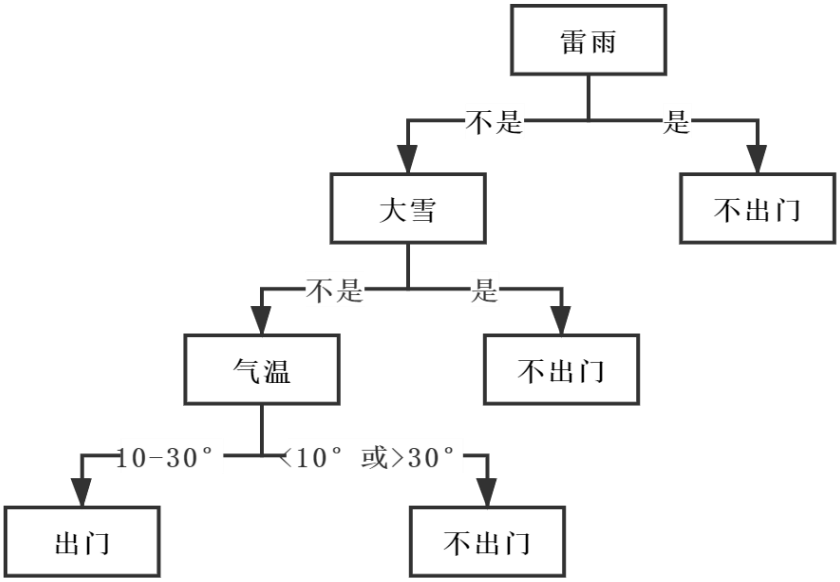

图 3 决策树算法的例子

本题选择 CART 决策树算法，采用基尼系数来划分属性，基尼系数的公式见公式

$$Gini(D) = 1 - \sum_{k=1}^{|c|} p_k^2$$

$$Gini(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

(5.1.4)

其中是D是留言集合，p_k是属于第k个类别的概率，用a来切分留言集合D，可以得到V个分支节点，得到V个离散属性a的可能取值$\{a^1, a^2, a^3 \dots a^V\}$。$D^v$表示所有离散属性$a$的值为$a^V$的留言。在分类中选择基尼系数最小的属性作为最优的划分属性。

决策树分类算法的优点是计算较为简便，可读性高；缺点是可能忽略留言文本中各个特征词之间的相关性，容易导致过度拟合。

Step 4：K 近邻算法

K 近邻算法属于基于机器的分类算法，是一种基于实例的学习算法，在许多研究领域都表现出了较好的性能。其基本原理是：假设在特征空间中，一个样本与 K 个样本最接近，则这些最接近的样本大多属于同一个类别，K 邻近算法也

将这些接近的样本判定为某一个类别。

影响 K 近邻算法分类效果的主要因素包括：K 值的选择（K 值的选择会对分类效果产生较大的影响或者造成误差。若 K 值选择较小则会出现过度拟合，若 K 值选择太大会使得计算成本增加。在实际中，一般采用交叉验证选择最合理的 K 值）、距离度量（首先将留言文本规范化处理，减小初始值或量纲对分类的影响）、分类决策规则（由与测试集 K 个最邻近的训练样本中的多数决定测试集的类别）。

使用 K 最近邻算法时，对于某一需要判别的留言样本，根据相应的计算方法将经过预处理和特征选择后的留言集合进行相似度计算，得到前 K 个最接近的留言样本，将这一需要判别的留言样本判别为这 K 个最接近的留言样本大多数所在的类别中。本题选择向量夹角余弦的方法计算留言样本之间相似性，如公式 所示。

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{c}(w_{ik} \times w_{jk})}{\sqrt{(\sum_{k=1}^{c} w_{ik}^2 \times \sum_{k=1}^{c} w_{jk}^2)}} \quad (5.1.5)$$

其中w_{ik}表示第 i 个留言样本的第 k 个属性值，$sim(d_i, d_j)$越大说明两个留言样本之间的相似度越高，越接近。

由于在分类时，K 近邻算法只需要依靠最接近的样本留言数据，不需要判别自身所属类别，使得计算容易实现，对于新的文本数据不需要重新训练。但计算时噪声数较高。

Step 5：逻辑回归算法

逻辑回归算法属于基于机器的分类算法，是一种广义的线性回归模型。其基本形式是使用逻辑函数来建立二元因变量，也称为示性变量，取值为“0”或“1”。在线性回归的基础上加一个 Sigmoid 函数就是逻辑回归。具有操作简单，计算量小等优点，但缺点是准确度不高，容易欠拟合。

Step 6：F-Score 结果比较

在完成文本分类模型的分类训练后，可以通过各种指标来评价预测效果。一般来说，常用的指标是准确率，但当分类中的数据比例很不平衡时，准确率不能

解释问题，因此可以从准确率和召回率的角度进行分析。在两个指标矛盾的情况下，F-Score 值有较好的解释，因此本文使用 F-Score 对分类方法进行评价。

F-Score 的公式可以表示为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{5.1.6}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。对比四种机器学习算法，可以总结得到表 4，如下所示：

表 4 算法 F1-Score 对比

算法类型	F1-Score
决策树	.76
朴素贝叶斯	.92
逻辑回归	.93
K 近邻	.94

比较四种方法得到的 F-Score，我们发现 K 近邻分类算法 F-Score 最高，达到的效果最好，因此我们推荐选择 K 近邻分类算法来构建一级标签分类模型。

5.2 热点问题挖掘

某一时段内群众集中反映的某一问题称为热点问题。热点问题的判断需要综合考虑问题的持续时间、发生地点、相似问题的聚类数量和点赞数等要素。首先进行分词、词性标注和去除停用词等预处理操作来清洗数据，其次采用 TD-IDF 算法计算文本的特征权重并将文本向量表示，然后用 LSA 模型进行特征抽取达到降维的效果，再通过 DBSCAN 算法将特征相似的问题进行聚类，最后采用熵权法各因素赋权，计算出热点问题的热度指数。热点问题挖掘的过程示意图如图 4 所示。

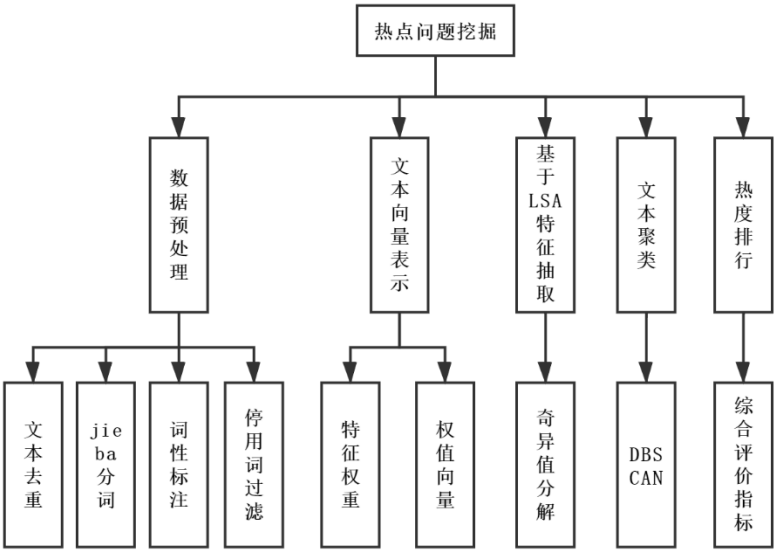

图 4 热点问题挖掘过程示意图

5.2.1 模型的建立

Step 1：数据预处理

（1）文本去重。去除重复值的目的是防止单个用户刷多条留言被误认为是热点问题。

（2）jieba 分词与词性标注。由于热点问题的发生地点或其他关键词大多数都是名词，因此采用词性标注的方法，过滤出地名、时间、其他名词和动词。

（3）停用词过滤。

Step 2：文本的向量表示

（1）基于 TF-IDF 进行特征选择。使用 TF-IDF 算法计算文本的特征权重，并在每条问题中提取出最重要的 5 个特征项。由 TF-IDF 算法所计算出的特征权重随着特征项在当前问题中出现的频率而成正比增大，而与该特征项在整个问题集合中出现的频率成反比下降。计算方法如公式 5.2.1 所示。

$$Weight = tf * idf \tag{5.2.1}$$

若某个特征项在多个问题中出现，其区分能力将变弱，重要性也随之下降，*idf* 的计算如公式 5.2.2 所示。

$$idf = \log \frac{N}{n} \tag{5.2.2}$$

其中，N 表示所有问题的总数，n 表示包含特征项 t_i 的问题数。依据 TF-IDF 提取 5 个权重最大的地名、其他名词、时间或动词，结果如图 5 所示。

```
留言编号：360113
关键词：['实习', '物流', '寝室', '暑假', '回来']
留言编号：286572
关键词：['溪湖', '游玩', '桃花', '商业空间', '拥堵']
留言编号：289408
关键词：['购房', '材料', '补贴', '提供', '业生']
留言编号：304503
关键词：['独生子', '独生', '独生子女', '生子', '子女']
留言编号：313964
关键词：['换证', '受理', '申请', '流水号', '便民服务']
留言编号：360112
关键词：['实习', '学生', '学校', '济学', '安排']
留言编号：319659
关键词：['住户', '电梯', '隐患', '开式', '推拉']
留言编号：321736
关键词：['小孩', '门诊', '报销', '医院', '应纳']
留言编号：323034
关键词：['收费', '清运', '标准', '车位', '垃圾清运']
留言编号：212323
关键词：['车位', '买不起', '购买', '感谢', '职工']
```

图 5 基于 TF-IDF 提取关键词

（2）TF-IDF 权值向量。如图 6 所示。

```
  (0, 7780)  0.21350314015206187
  (0, 6295)  0.15065749193146857
  (0, 9521)  0.08233632181708092
  (0, 11416)  0.30506815693174966
  (0, 9249)  0.14724317288080802
  (0, 8722)  0.20365358278591217
  (0, 10846)  0.22785577265169926
  (0, 29)  0.15194629740610605
  (0, 5864)  0.2632296921753869
  (0, 11108)  0.2805252995233821
  (0, 8402)  0.12507992363820092
  (0, 2600)  0.21756871179607784
  (0, 5764)  0.25231254281331716
  (0, 4319)  0.21756871179607784
  (0, 483)  0.3157882795403587
```

图 6 TF-IDF 权值向量

Step 3：基于 LSA 进行特征抽取[6]。

文本中存在的同义词和多义词现象，造成特征向量的各个分量存在一定的相关性，潜在语义索引（LSA）模型的目的是对分类任务降维，它通过挖掘文本与

特征之间潜在的高阶语义结构，将文本特征矩阵分解为一个低维的正交矩阵，实现特征空间的降维。LSA 降维主要通过奇异值分解（SVD）来实现，在 TF-IDF 矩阵建立后，就可以利用奇异值分解计算 A 的 k 维近似矩阵 A_k，结果如图 7 所示。

```
[[ 6.83978121e-01 -2.32166965e-01  4.37220725e-01 -1.09175705e-01
   9.05981046e-02 -1.51629222e-01 -1.33441654e-01  8.07983221e-02
  -2.18554053e-01  1.33858633e-01 -1.60218611e-01  1.76703681e-01
  -2.20996796e-01  2.09606084e-01 -6.56060664e-02]
 [ 6.38870593e-01 -1.03819088e-01  4.05900865e-01  2.04840486e-01
   2.20381195e-01 -1.68986536e-02  4.41184280e-01 -4.25649509e-02
   4.29763328e-02  1.53341693e-02 -1.25511543e-01 -2.03429187e-01
   2.27386662e-01 -1.28133107e-01  4.12380316e-02]
 [ 5.43431315e-01 -2.01505891e-02  2.98879722e-01 -1.52099948e-01
   2.89444651e-02 -2.40182590e-01  1.36212051e-01  2.90999789e-01
   5.27411111e-02 -2.34403082e-01 -9.91351283e-02 -9.99801619e-02
  -2.24441627e-01 -5.07813682e-01 -2.10365878e-01]
 [ 8.07140054e-01 -3.97556666e-01 -1.37694437e-01 -7.15534671e-02
   4.74794340e-02 -3.20381805e-02  1.52365238e-01  1.00989161e-01
  -4.61733855e-02  7.40149478e-03 -1.76078289e-01 -1.03185068e-01
  -1.50919344e-01 -2.18506705e-01 -1.23854517e-01]
```

图 7 降维后的 TF-IDF 权值向量矩阵

Step 4：文本聚类

聚类方法主要有划分聚类法、层次聚类法和密度聚类法、基于网格的方法和基于模型的方法等[7]。K-means[8]是划分聚类中广泛使用的基于迭代的聚类算法，适合用于已知数据可以分为几类且数据量较少的情况。基于密度的 DBSCAN 算法将簇定义为密度相连的点的最大集合，不需要预先指定聚类簇数，还可以在含有噪声数据的数据集中识别任意数量和形状的聚类，具有较强的过滤噪声的能力。

本题所需处理的数据量大并且不清楚聚类簇数，综合考虑后，选择基于密度算法的 DBSCAN 聚类算法。DBSCAN 集群模型[9]如图 8 所示。

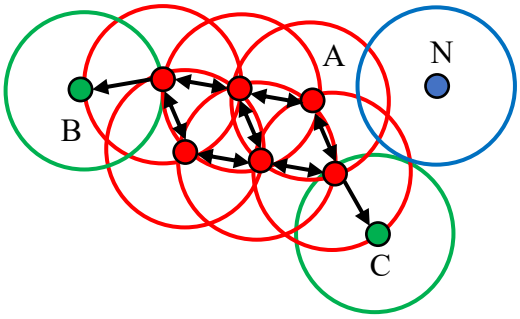

图 8 DBSCAN 集群模型图

其主要思想为通过获取密度核心，由核心向密度较高的地方延展，将相近的问题合并为同一个簇。图中的箭头表示直接密度可达性。点 B 和 C 是密度可达的，而点 N 是密度不可达的，被认为是噪声点。算法具体过程[9]如表 5 所示。

表 5 DBSCAN 聚类算法过程

```
1   标记所有对象为 undefined;
2   do
3       随机选择一个 undefined 对象 p;
4       标记 p 为 defined;
5       if p 的 ε-邻域至少有 MinPts 个对象
6           创建一个新簇 C，把 p 添加到 C;
7           令 N 为 p 的 ε-邻域的对象集合;
8           for N 中每个点 p1
9               if p1 是 undefined
10                  标记 p1 为 defined;
11                  if p1 的 ε-邻域至少有 MinPts 个点，把这些点加到 N;
12              if p1 不是任何簇的成员，把 p1 加到 C;
13          end for
14          输出 C;
15      else 标记 p 为噪声;
16  until 没有标记为 undefined 的对象;
```

聚类过程主要有两个输入参数：半径参数 *eps* 和邻域密度阈值 *min_samples*。由于当半径参数大于 0.5 时，会使属于不同类别的新闻聚在一起，而小于 0.3 时，又无法识别相同的问题，因此半径参数 *eps* 一般都设置为 0.3-0.5 之间。邻域密度阈值 *min_samples* 主要用来设定热点问题的评定门槛。本题选择 *eps* 的值为 0.3，*min_samples* 的值为 5。

Step 5：计算问题持续时间。由于所给数据中存在两种不同的时间格式，为方便进行时间处理，先将时间格式进行归一化。处理数据后，发现留言问题的持续时间是动态变化的，会在短期内到达高峰，热度期结束后，保持一个很低的频率。基于问题持续时间的动态阈值模型[10]，如下所示：

$$Threshold(T,t) = \theta + \alpha * (Time(S) - Time(T)) \tag{5.2.3}$$

其中 $Threshold(T,t)$ 表示 t 时刻问题 T 的阈值；θ 是一个常数，表示问题刚提

出时的阈值；α 是一个可调参数，表示时间信息在动态阈值中所占的比例。$Time(S)$ 和 $Time(T)$ 分别表示问题的提出次数超过 10 次[11]的时间和问题第一次提出的时间，都以天为单位。

Step 6：热度排行。综合考虑相似问题聚类大小、问题持续时间和点赞数三个因素，进行热度问题排行。

5.2.2 模型的求解

通过 Pycharm 求解 DBSCAN 模型，得到问题聚类效果散点图，如图 9 所示。问题的聚类数量为 29，噪点数为 2010，噪点在图中以黑点表示。

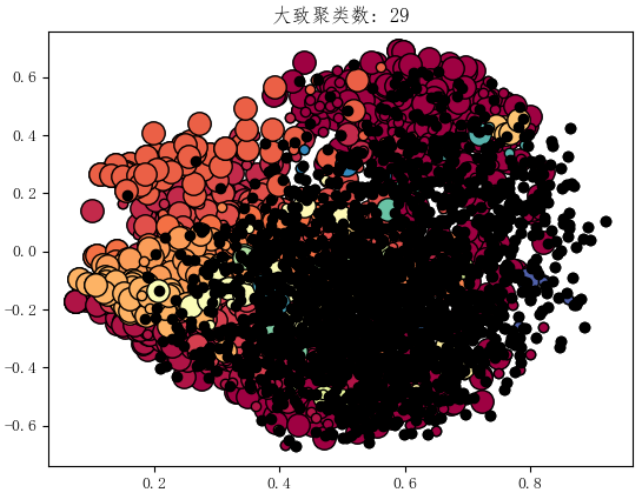

图 9 问题聚类效果散点图

问题聚类结果，如图 10 所示。“问题 ID”表示问题所属的类，一共 29 个类，其中“-1”为噪点，不属于热点问题的范畴。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
-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 11:34:04	我是A市A5区汇金路五矿万境K	2097	0
14	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖	1762	5
3	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p	821	0
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	胡市长：您好！西地省展星投	790	0
3	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您好！58车贷案发，	733	0
-1	263672	A00041448	A4区绿地海外滩小区距长赣高铁最近只有30米	2019/9/5 13:06:55	您好，近日看到了渝长厦高铁	669	0
0	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	位于A市A2区暮云街道丽发新城	242	0
-1	284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远离	2019/1/10 15:01:26	京港澳高速城区（g4）是1996	80	0

图 10 问题聚类结果

综合考虑相似问题聚类大小、点赞数和问题持续时间三个因素，依据熵权法计算各因素的权重分别为 0.41、0.57 和 0.02，计算出热点问题的热度指数。热度排行结果如图 11 所示。

热度排名	问题ID	热度指数	聚类大小	点赞数	时间范围		持续时间/天
1	3	15.31	164	1734	2019/2/2	2019/3/1	27
2	14	10.65	137	1767	2019/3/20	2019/6/20	92
3	0	9.48	736	1121	2019/1/1	2020/1/7	371
4	1	4.69	364	547	2019/3/26	2020/4/15	386
5	2	3.16	282	339	2019/8/21	2020/8/21	366
6	4	2.36	234	233	2019/1/13	2020/1/7	359
7	5	1.25	82	147	2019/1/14	2020/1/5	356
8	6	0.88	69	93	2019/1/16	2020/1/6	355

图 11 热度排行结果

前 5 名热点问题留言频率和点赞数条形图，如图 12 所示。

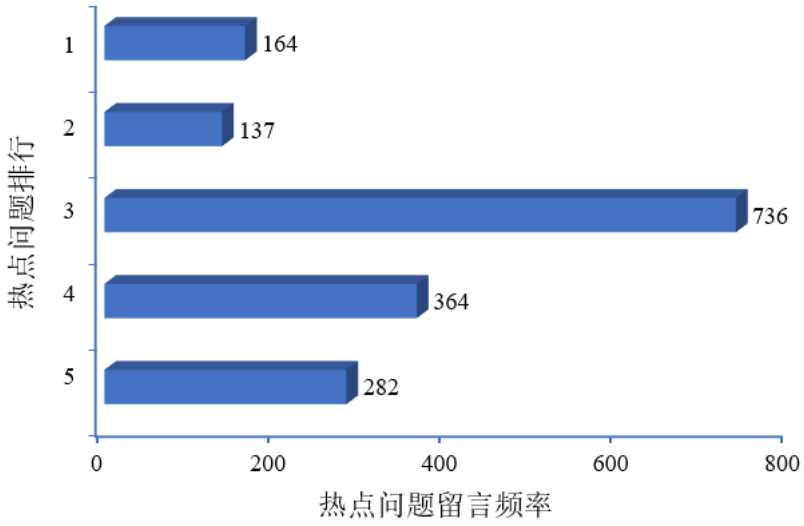

图 12(a) 前 5 名热点问题留言频率条形图

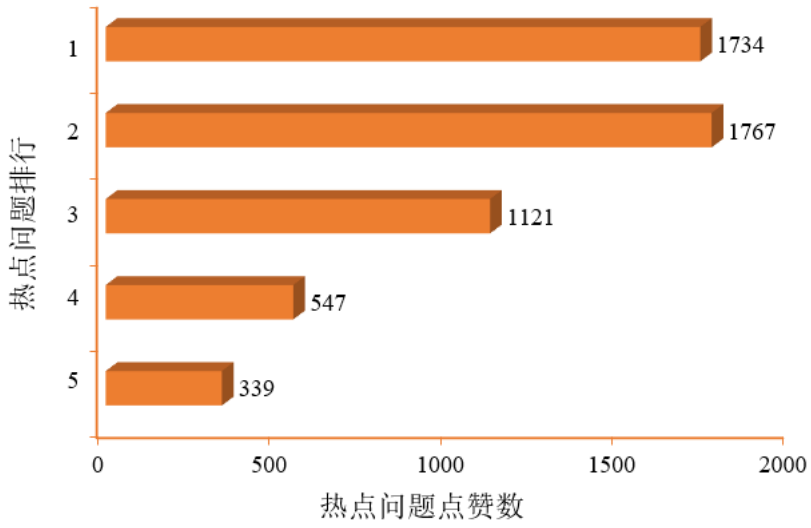

图 12(b) 前 5 名热点问题留言频率条形图

前 5 名热点问题表，如图 13 所示。

热度排名	问题ID	热度指数	时间范围		地点/人群	问题描述
1	3	15.31	2019/2/2	2019/3/1	A市A4区	请书记关注A市A4区58车贷案
2	14	10.65	2019/3/20	2019/6/20	A市梅溪湖金毛湾	反映A市金毛湾配套入学的问题
3	0	9.48	2019/1/1	2020/1/7	A市A2区暮云街道丽发新城	A市富绿物业丽发新城强行断业主家水
4	1	4.69	2019/3/26	2020/4/15	A6区月亮岛路	关于A6区月亮岛路110kv高压线的建议
5	2	3.16	2019/8/21	2020/8/21	A市	请A市依法查处“爱玩客ivankr”特大电信诈骗案

图 13 前 5 名热点问题表

5.3 答复意见的评价

根据题设要求，从以下三个角度衡量相关部门答复意见的质量。

1）相关性，即答复意见的内容是否与问题相关；

2）完整性，即是否满足某种规范；

3）可解释性，即答复意见中内容的相关解释。

为了选用具体的评价指标进行量化分析，参考了相关文献[11][12][13]，评价答复的质量可以选择问答的内容匹配度、答复的响应时间、答复详细度和答复的非重复字符数作为评价指标。由于本题中的留言均是由相关部门的工作人员回复，结合本题的具体实际，并不存在答复时重复内容凑字数的现象，因此，排除答复的非重复字符数这一指标，我们选择问-答关键词的余弦相似度，答复的响应时

间和答复详细度来衡量答复的质量。

5.3.1 模型的建立

Step 1：数据预处理。

（1）去除空格。将数据导入 SPSS 中，选择去除字符串前后的空格，避免分词时出现许多无意义的符号。

（2）文本去重。在 SPSS 中选择标志重复个案，去除同一用户的相同问题。

（5）自定义词典。为了提高关键词提取的准确性，自定义添加分词时不能自动识别的名词，如：“灵活就业”、“公积金”、“原民办教师”、“养老金”等。

（3）jieba 分词。以词为基本单元，自动对中文文本进行词语切分。

（4）停用词过滤。去掉对文本实质内容无意义的停用词。

Step 2：文本的向量表示

（1）基于 TF-IDF 提取关键词。我们尝试分别提取问题和答复的 5 个、10 个、20 个以及 50 个关键词，发现当提取 20 个关键词时，既不会因为关键词太少导致过大偏差，又可以把最重要的关键词提取出来，因此选择提取 20 个特征权重最大的关键词。

（2）OneHot 编码。将分别提取的问题和答复的 20 个关键词转化为词向量矩阵。

Step 3：计算答复详细度。根据答复意见的具体内容，发现长度较短的答复一般都是“已转交相关单位处理”这类答复，并未给出具体的解决方案，而较长的答复一般都是根据具体的法律法规做出合理的解释。答复意见的长度[12]较好地刻画了答复的详细度。利用 Python 语言自动计算答复意见的长度。

Step 4：计算余弦相似度。基于向量空间模型的方法包括匹配系数(Matching Coefficient)、余弦相似度(Cosine)、欧式距离(Euclidean Distance)等[14]。使用余弦相似度来计算不同问题之间的距离，可以理解为空间中两个向量所形成的夹角，当夹角为 0° 时，说明其方向相同、线段重合，相似度就越高，反之同理。因此，通过判断夹角的大小来判断不同向量之间的相似度。计算方法如公式 5.3.1 所示。

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{5.3.1}$$

$$= \frac{\mathrm{A} \cdot \mathrm{B}}{|A| \times |B|}$$

针对留言详情和答复意见的 OneHot 编码结果，计算其余弦相似度[13]，定量地表达留言详情-答复意见的内容相关度，如表 6 所示。

表 6 问-答关键词的余弦相似度

留言编号	答复详细度	留言 OneHot 编码	答复 OneHot 编码	问-答关键词余弦相似度
8629	217	[1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1]	[1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	80.00%
97307	310	[0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1]	[1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	75.00%
50409	654	[1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1]	75.00%
18413	1368	[1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1]	[0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0]	65.00%
48073	387	[0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1]	[1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0]	13.48%

Step 5：计算答复的响应时间。

答复的响应时间[12]也就是答复意见与留言详情之间的时间差值，由于答复时间与留言时间大部分在不同日期的不同时间，无法直接采用时间相减的办法，因此将时间差转化为小时数，精确地表达时间差。可以使用 EXCEL 中的公式快速计算。

$$(TEXT(Gi - Di, "[h]:mm") * 24) / 24 \tag{5.3.2}$$

计算结果如表 7 所示。

表 7 答复的响应时间

留言编号	留言时间	答复时间	响应时间/天
8629	2017/5/25 14:26:48	2017/6/21 18:48:42	27.2
97307	2019/11/14 12:53:19	2019/11/22 17:09:02	8.2
50409	2014/5/8 10:13:48	2014/5/20 11:16:06	12.0
9128	2017/3/22 12:55:17	2017/4/14 15:32:24	23.1
18413	2015/12/30 2:26:45	2016/1/7 10:59:47	8.4

5.3.2 模型的求解

各指标的权重对于综合评价结果有着重要意义。主观赋权法在根据指标特点确定权重方面具有优势，但客观性较差；而客观赋权法在不考虑指标实际含义的情况下，确定权重具有优势，但有时会出现确定的权重与属性的实际重要程度相悖的情况。针对主、客观赋权法的优缺点，为兼顾各指标的实际含义，同时又尽力减少赋权的主观随意性，使指标的赋权达到主观与客观相统一，进而使评价结果更真实可靠。

因此，采用组合赋权法，将客观的熵权法与主观的指标评分标准相结合。熵权法利用每个答复的各指标数据，通过熵值法得到各指标的信息熵，信息熵越小，信息的效用值越大，指标的权重越大。通过分析数据，发现答复响应时间和答复长度的数据离散度很高，但这并不意味着它们对于综合评价结果可以起到较大的决定作用，若直接使用熵权法，会造成很大的误差，因此需要对数据进行整理。首先根据相关政策法规和数据的统计分析规律，制定各评价指标的评分标准，进行十分制打分，然后使用熵权法[15]进行权重赋值，最后计算出综合评价指数。

Step 1：依据政策法规和数据统计规律，以十分制制定各指标评分标准

（1）答复响应时间的评分标准

根据《中华人民共和国政府信息公开条例》第二十四条对行政机关政府信息公开答复的期限作了明确规定，“不能当场答复的，应当自收到申请之日起 15 个工作日内予以答复，如需延长答复期限的，应当经政府信息公开工作机构负责人同意，并告知申请人，延长答复的期限最长不得超过 15 个工作日。”根据本条规

定，15 天是法律规定的答复期限，若申请延期，整体答复期限也不超过 30 天。依据此标准，并结合生活实际，制定出答复响应时间的十分制评分标准，如表 8 所示。

表 8 答复响应时间的评分标准

响应时间/天	评分	响应时间/天	评分
1-3	10	10-15	5
3-7	8	15-30	2
7-10	6	>30	1

（2）余弦相似度的评分标准

附件 4 中的问-答关键词的余弦相似度数值都在 0-80%之间，换算成十分制，每个数据都扩大 12.5 倍，作为相关性的评分。为了避免数值为 0 而导致后续步骤无法进行，当余弦相似度为 0 时，统一设定为 0.1 分。

（3）答复详细度的评分标准

根据余弦相似度的计算结果，发现当答复与问题的余弦相似度大于 50%时，答复的长度平均值约为 500 字，可以认为 500 字能够达到最优的回复效果；当余弦相似度为 0 时，答复长度绝大部分都在 100 字以下。根据统计分析规律，制定答复详细度的评分标准，如表 9 所示。

表 9 答复详细度的评分标准

答复详细度	评分	答复详细度	评分
>500	10	200-300	5
400-500	8	100-200	2
300-400	6	1-100	1

Step 2：熵权法确定指标权重

依据每个答复的各指标评分，使用熵权法确定指标权重。假设多指标决策矩阵如下：

$$M=\begin{array}{c}\\ A_1\\ A_2\\ \vdots\\ A_m\end{array}\begin{bmatrix} x_{11} & x_{11} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{5.3.3}$$

首先使用 $P_{ij}=\dfrac{x_{ij}}{\sum_{i-1}^{m}x_{ij}}$ 表示第 j 个指标下第 i 个答复 A_i 的贡献度，做出初始的 P 矩阵，如图 14 所示。

留言编号	时间差/天	答复详细度	余弦相似度	响应时间评分	详细度评分	相关性评分	P矩阵		
8629	27.2	221	80.0%	2	5	10.0	0.000137931	0.00034	0.001669
97307	8.2	310	75.0%	6	6	9.4	0.000413793	0.000407	0.001565
50409	12.0	661	75.0%	5	10	9.4	0.000344828	0.000679	0.001565
9128	23.1	895	75.0%	2	10	9.4	0.000137931	0.000679	0.001565
18413	8.4	1368	65.0%	6	10	8.1	0.000413793	0.000679	0.001356
29745	34.9	616	65.0%	1	10	8.1	6.89655E-05	0.000679	0.001356
6448	3.3	132	63.2%	8	2	7.9	0.000551724	0.000136	0.001319
17229	4.0	900	62.0%	8	10	7.8	0.000551724	0.000679	0.001294
140649	6.6	216	60.0%	8	5	7.5	0.000551724	0.00034	0.001252
119181	10.2	590	60.0%	5	10	7.5	0.000344828	0.000679	0.001252
7107	21.8	707	60.0%	2	10	7.5	0.000137931	0.000679	0.001252
8764	31.8	1972	60.0%	1	10	7.5	6.89655E-05	0.000679	0.001252
6891	33.7	821	60.0%	1	10	7.5	6.89655E-05	0.000679	0.001252
159042	54.3	24	57.0%	1	1	7.1	6.89655E-05	6.79E-05	0.001189
20893	3.9	90	56.6%	8	1	7.1	0.000551724	6.79E-05	0.00118
16781	1.2	469	55.0%	10	8	6.9	0.000689655	0.000543	0.001147
17622	1.5	159	55.0%	10	2	6.9	0.000689655	0.000136	0.001147
176604	3.5	282	55.0%	8	5	6.9	0.000551724	0.00034	0.001147

图 14 P 矩阵计算结果

再将 P 矩阵中的每个元素都变换为 $P_{ij}\ln(P_{ij})$，得到新的 Q 矩阵，如图 15 所示。

P矩阵			Q矩阵		
0.000137931	0.00034	0.001669	-0.00123	-0.00271	-0.01067
0.000413793	0.000407	0.001565	-0.00322	-0.00318	-0.01011
0.000344828	0.000679	0.001565	-0.00275	-0.00495	-0.01011
0.000137931	0.000679	0.001565	-0.00123	-0.00495	-0.01011
0.000413793	0.000679	0.001356	-0.00322	-0.00495	-0.00895
6.89655E-05	0.000679	0.001356	-0.00066	-0.00495	-0.00895
0.000551724	0.000136	0.001319	-0.00414	-0.00121	-0.00875
0.000551724	0.000679	0.001294	-0.00414	-0.00495	-0.0086
0.000551724	0.00034	0.001252	-0.00414	-0.00271	-0.00837
0.000344828	0.000679	0.001252	-0.00275	-0.00495	-0.00837
0.000137931	0.000679	0.001252	-0.00123	-0.00495	-0.00837
6.89655E-05	0.000679	0.001252	-0.00066	-0.00495	-0.00837
6.89655E-05	0.000679	0.001252	-0.00066	-0.00495	-0.00837
6.89655E-05	6.79E-05	0.001189	-0.00066	-0.00065	-0.00801
0.000551724	6.79E-05	0.00118	-0.00414	-0.00065	-0.00796
0.000689655	0.000543	0.001147	-0.00502	-0.00408	-0.00777
0.000689655	0.000136	0.001147	-0.00502	-0.00121	-0.00777
0.000551724	0.00034	0.001147	-0.00414	-0.00271	-0.00777

图 15 Q 矩阵计算结果

然后用 E_j 表示所有答复对指标 X_j 的贡献总量：

$$E_j = -k\sum_{i=1}^{m} P_{ij}\ln(P_{ij}) \tag{5.3.4}$$

其中，m 为答复的总数 2815，常数 k：

$$k = \frac{1}{\ln(m)} \tag{5.3.5}$$

计算出 k 的值为 0.126。

D_j 是第 j 指标下各答复贡献度的一致性程度，$D_j = 1 - E_j$

各指标权重：

$$W_j = \frac{D_j}{\sum_{j=1}^{n} D_j} \tag{5.3.6}$$

其中，n 为评价指标个数 3。根据公式 5.3.3，计算出答复响应时间、答复详细度和答复相关性的权重分别为 0.2905、0.2912、0.4183，计算结果如表 10 所示。

表 10 各指标的贡献总量与权重

参数	响应时间	详细度评分	相关性评分
E_j	0.975852712	0.975797851	0.965231177
D_j	0.024147288	0.024202149	0.034768823
W_j	0.290517249	0.291177285	0.418305466

Step 3：计算综合评价指数

根据参考文献[16]，综合评价指数的计算公式为：

$$ESI = \sum_{j=1}^{n} W_j \times C_j \tag{5.3.7}$$

其中，ESI 为综合评价指数，W_j 为第 j 个指标的权重值，C_j 为其无量纲量化值，n 为评价指标个数 3。

答复的综合评价结果（部分数据）如表 11 所示。

表 11 答复的综合评价

留言编号	响应时间评分	详细度评分	相关性评分	综合评价
8629	2	5	10.0	6.22
97307	6	6	9.4	7.41
50409	5	10	9.4	8.29
9128	2	10	9.4	7.41
18413	6	10	8.1	8.05
29745	1	10	8.1	6.60
6448	8	2	7.9	6.21
17229	8	10	7.8	8.48
140649	8	5	7.5	6.92
119181	5	10	7.5	7.50

六、模型的评价与推广

6.1 模型的优点

1、在构建一级标签分类模型中，分析比较了朴素贝叶斯、决策树、逻辑回归、K 近邻分类算法的优缺点，并针对 F-Score 分类效果评价方法，给出了最优的分类算法；

2、DBSCAN 算法不需要预先指定聚类簇数，还可以在含有噪声数据的数据集中识别任意数量和形状的聚类；

3、根据相关文献，答复质量的评价所选择的评价指标是可靠的，在计算指标权重时，采用了组合赋权法，结合了主客观赋权的优点。

6.2 模型的不足

1、在分类算法的选择上，由于时间、篇幅有限，只考虑了基于机器的分类算法，忽视了基于深度学习的分类算法。基于深度学习的分类算法包括循环神经网络、卷积神经网络、胶囊神经网络等；在之后的研究，考虑同时比较多个基于深度学习的分类算法的优势，并利用 F-Score 评估出效果最好的分类算法；

2、基于密度的 DBSCAN 算法对于输入参数敏感，需要对半径参数 *eps* 和邻域密度阈值 *MinPts* 联合调参，不同的参数组合对聚类效果有较大的影响；

3、在进行答复指标赋权时，各指标的评分标准是根据政策法规和数据统计规律制定的，不可避免地带来一定的误差。

6.3 模型的未来展望

1、基于 K 近邻构建的分类算法效果显著，易于理解，在非结构化数据例如留言信息、垃圾短信、网络舆情等研究分析中都可以得到应用。

2、通过 DBSCAN 算法将相似的问题进行聚类，综合考虑了各个因素对热点问题的影响，并给出了热度评价指标，为未来文本数据热度分析提供了参考；

3、对答复意见的质量采用组合赋权法评价，将客观的熵权法与主观的指标评分标准相结合，提高了评价的可靠性，为问政平台相关工作人员以后的答复提供了参考标准。

参考文献

[1] 吴广建. 面向政务微博的数据分析系统设计与实现[D]. 杭州师范大学，2020.

[2] Waters R D, Williams J M. Squawking, tweeting, cooing, and hooting: analyzing the communication patterns of government agencies on Twitter[J]. Journal of Public Affairs, 2011, 11(4): 353-363.

[3] 柳思思. 政务微博推广机制研究——以“北京微博发布厅”的推广为例[J]. 电子政务. 2015(01): 43-51.

[4] 刘星. 大数据背景下长沙市政府电子政务平台建设问题研究[D]. 广西师范大学, 2017.

[5] Herreraviedma E, Pasi G, Lopezherrera A G, et al. Evaluating the Information Quality of Web Sites: A Methodology Based on Fuzzy Computing With Words[J]. Journal of the Association for Information Science and Technology, 2006, 57(4): 538-549.

[6] Huang B, Yang Y, Mahmood A, et al. Microblog topic detection based on LDA

model and single-pass clustering[C]. Springer, Berlin, Heidelberg, 2012:166-171.

[7] 朱梦. 基于机器学习的中文文本分类算法的研究与实现[D]. 北京邮电大学，2019.

[8] 王明亚. 基于词向量的文本分类算法研究与改进[D]. 华东师范大学, 2016.

[9] 王金柱. 基于系统相似模型与持续时间的话题检测技术研究[D]. 复旦大学, 2009.

[10] 孔维泽, 刘奕群, 张敏, 马少平. 问答社区中回答质量的评价方法研究[J].中文信息学报, 2011, 25(01): 3-8.

[11] Agichtein E, Castillo C, Donato D, et al. Finding high-quality content in social media[C]. Proceedings of the 2008 international conference on web search and data mining. 2008: 183-194.

[12] Adamic L A, Zhang J, Bakshy E, et al. Knowledge sharing and yahoo answers: everyone knows something[C]. Proceedings of the 17th international conference on World Wide Web. 2008: 665-674.

[13] 许丽利. 聚类分析的算法及应用[J]. 吉林大学, 2010.

[14] 段明秀. 层次聚类算法的研究及应用[D]. 中南大学, 2009.

[15] Schubert E, Sander J, Ester M, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems (TODS), 2017, 42(3): 1-21.

[16] 王春柳, 杨永辉, 邓霏, 赖辉源. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37(03): 158-168.

[17] 贾艳红, 赵军, 南忠仁, 赵传燕, 王胜利. 基于熵权法的草原生态安全评价——以甘肃牧区为例[J]. 生态学杂志, 2006(08): 1003-1008.

[18] 胡永宏, 贺思辉. 综合评价方法[M]. 北京: 科学出版社, 2000.