智慧政务中的文本挖掘分析

摘要：近年来，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，产生了大量社情民意相关的文本数据，这给主要依靠人工进行留言划分和热点整理的工作带来了极大挑战。在工业 4.0 时代，随着物联网、大数据、云计算、人工智能的发展，基于自然语言处理技术建立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积极的推动作用。本文采用文本挖掘、自然语言处理等方法对群众问政留言记录进行分类、热点挖掘，并对相关部门的回复意见进行评价。

针对群众留言分类的问题，本文首先对原始数据的留言主题和详情进行数据清洗，主要包括分词和去停用词处理，删除无实际意义的词、标点符号和数字。其次提取内容和标题，并对提取出的关键词加强权重，使用词频-逆文档频率（TF-IDF）进行特征提取，采用朴素贝叶斯将留言数据进行文本分类，建立关于留言内容的分类模型。

问题二要求对群众留言按照一定的评价指标挖掘出热点留言问题并给出前五名的排名。为了提高热度排名准确度,本文首先去除重复数据和同一用户的相似留言内容并进行分词处理。其次利用 TF-IDF 进行特征提取并同时去除数据中的离群点。最后基于相似度、留言时间和地点的综合约束条件，采用凝聚型层次聚类算法对问政留言数据进行热点挖掘，将众多留言中的相似留言归为一类，获得该类的留言明细，然后根据数据统计获取留言数量最多的前五类作为热点问题。

问题三要求对留言的答复意见从相关性、完整性、可解释性等角度给出一套评价方案对答复意见的质量进行评估。针对此问题，本文综合考虑了各种评价指标，首先对对留言回复进行文本特征提取，对预处理后的文本数据进行关联度计算，然后构建了一个衡量留言质量的分类器，利用该分类器对留言的质量进行评估。

关键词：网络问政；留言分类；热点挖掘；TF-IDF 算法；朴素贝叶斯分类；聚类算法

Text Mining Analysis in Smart Government Affairs

Abstract: In recent years, network platforms such as WeChat, Weibo, mayor mailbox, and hotline have gradually become the important channels that enable government to understand public opinions. A large amount of text data thereby generated, which brings great challenges of message classification and hot issues processing. In the Industry 4.0 era, with the development of the Internet of Things (IoT), big data, cloud computing, and AI, building an intelligent system of government affairs based on the natural language processing technology has become a new trend. This will improve the management abilities and effectivenesses of the government. This paper classifies the messages and mining their hot issues through text mining and natural language processing. Additionally, we evaluate the replies of the relevant departments.

For the problem of the message classification, we firstly apply data cleaning for the topics and details of the original data. The main techniques including sentence segmentation, stop words removal and deleting meaningless characters of words, punctuations, and numbers. Secondly, extracting contents, titles and then weighted to their keywords. Additionally, we exploit the Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction and the naive Bayesian method for text classification.

The requirements of second problem are mining hot issues based on a specific evaluation index and providing the ranking of top 5 hot issues. To improve the ranking accuracy, we firstly remove duplicate data and delete similar message contents of the same user. After the processing of word segmentation, TF-IDF is used to extract features and remove outliers simultaneously. Finally, based on the constraint of similarity, time and location, the agglomerated hierarchical clustering algorithm is used to mining hot issues to cluster similar messages as one category. According to details of messages and data statistics to obtain the most messages of the top 5 categories as hot issues.

Question 3 asks for an evaluation metric to assess the quality of the reply comments from the viewpoints of relevance, completeness, and interpretability. Firstly, we preprocess the reply text data by feature extraction and then calculate its correlation. Secondly, we build a classifier to measure the message quality.

Key words: network political, message classification, hotspot mining, TF-IDF algorithm, naive Bayes classification, clustering algorithm

目 录

1、挖掘目标

随着信息技术的不断发展，群众交流表达意见的方式从传统的纸质信件和电话来电方式发展为大数据时代下的电子信息表达。随着网络的发展，网民成为一个庞大的群体，网络上也开始兴起各种各样的意见表达方式。近年来，随着物联网、云计算、大数据等信息技术的兴起与发展，政府的运作方式发生了一些转变：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。各种社情民意的文本数据不端攀升，给相关部门的信息分类与留言处理工作带来了巨大的挑战。利用大数据、人工智能等方式建立相关的智慧政务系统，有助于提升政府的管理水平和工作效率。

本次数据建模主要针对来自互联网公开的群众留言记录以及相关部门对部分群众留言的答复意见所产生的文本数据，采用 jieba 中文分词工具、TF-IDF 特征提取算法、朴素贝叶斯分类以及凝聚型层次聚类算法达到以下目标：

（1）根据附件二的数据建立关于留言内容的分类模型，得出 F-Score 方法对分类模型进行评价的结果。

（2）对附件三的数据进行挖掘分析，对留言按照相似性大小归类，给出排名前五的热点问题，得到文件“热点问题表.xls”，同时按照题目要求的格式得到相应热点问题对应的留言信息，得到“热点问题留言明细表.xls”。

（3）对附件四中答复意见进行分析，从答复的相关性，完整性和可解释性构建一个衡量留言质量的分类器。

2、分析方法与过程

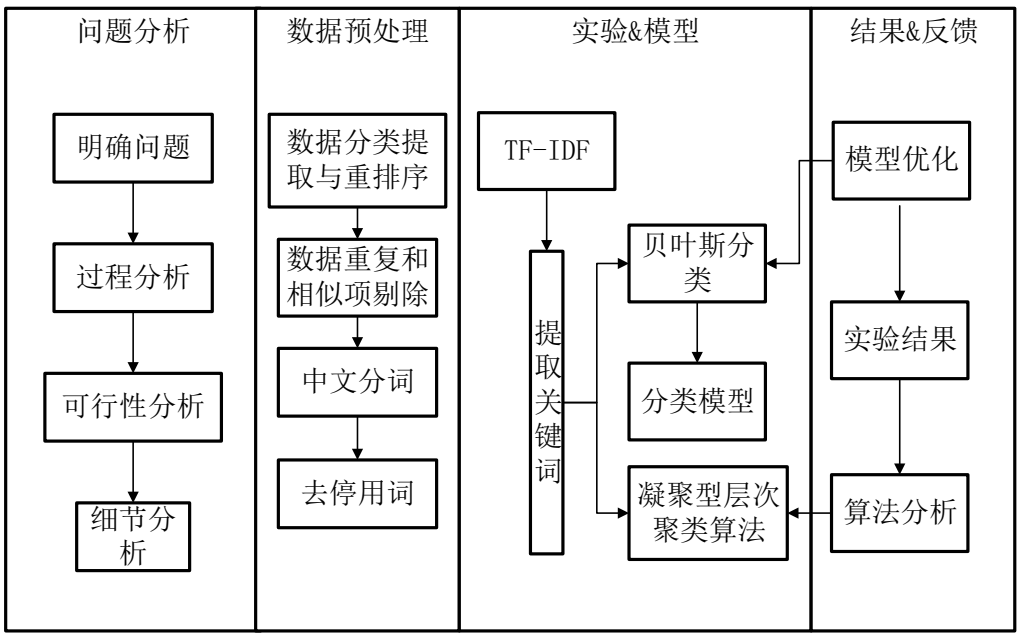

图 2-1 整体流程图

要完成所设定的挖掘目标，整体思路和步骤如图 2-1 所示：

步骤一：问题分析旨在全面地理解问题，通过分析问题题干，明确最终要解决的问题，分析解决问题的步骤和所要用到的方法，了解方法的原理并分析其可行性，在此过程中还需要明

确问题细节和方法细节。

步骤二：数据预处理：面对大量的文本数据，良好的预处理技术是完成自然语言处理任务的重要步骤。针对种类繁多、包含内容分类标签和时间的问政留言数据，将每一类对应的留言内容和标题提取出来，或按时间顺序将数据重新排列会方便程序处理。因此本文首先删除重新排序后数据中的重复词语，即去重、去空，然后对数据进行中文分词同时去掉缺乏实际意义的词语。

步骤三：进行实验，建立相关模型，解决问题。在分词之后，需要将词语转换为向量以供挖掘分析。这里采用 TF-IDF 算法提取留言内容中的关键词，将关键词信息转换为权重向量。采用贝叶斯分类算法对留言内容进行分类，得到分类模型；使用凝聚型层次聚类算法，将相似留言归为一类，按照各类留言数量，结合时间地点限制条件得到热点留言。

步骤四：根据试验结果分析问题解决的程度和质量，对整个解决问题的过程和算法进行优化，提高试验结果的质量。

2.1 问题一的分析方法与过程

2.1.1 问题一流程图

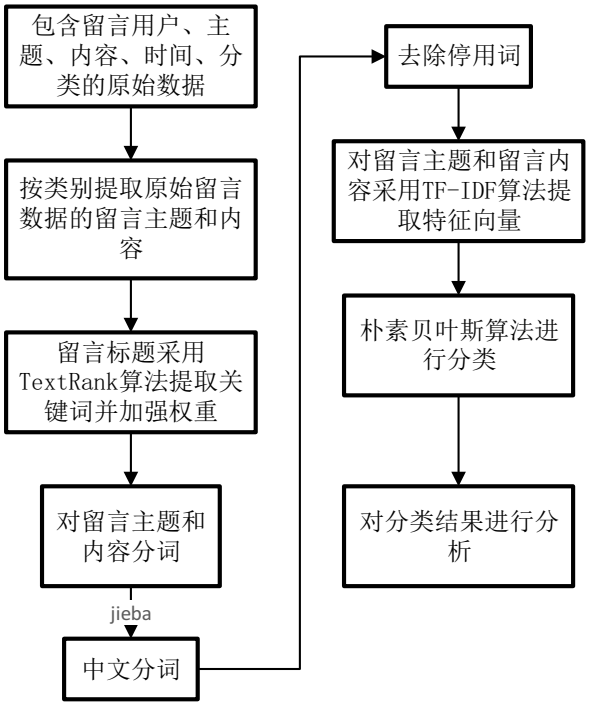

图 2-2 问题 1 流程图

2.1.2 数据预处理

题目给出的留言数据包含用户、主题、时间、内容以及每条留言对应的分类标签，对留言数据进行分类、建立分类模型时，只需要获取留言主题和留言内容及其对应的分类标签进行分类模型的训练。因此本文首先按类别提取原始留言数据并划分训练集与测试集。为了后续操作方便，先将每一条留言数据按照其对应的标签分类别提取留言主题和留言内容，重新保存为.txt 文件并放在标签文件夹下。在提取留言数据的同时，将五分之四原始数据划分

为训练集，五分之一作为测试集，相关代码为 data_process.py。

除了训练集测试集的划分，预处理还包括数据的去停用词和中文分词。在题目给出的留言数据中，有许多虚词在句中仅起到结构作用，不表示实际意义，还有一些词在语料中出现频率较高，但对实际分类的作用不大，比如‘的’、‘是’、‘在’等词语，为了提高分类的准确率，将这类词语和一些中英文标点符号和数字去掉。此外，使用计算机对文本进行自动分类，要先把文本转换成计算机能够处理的数据形式，为了便于转换，先对留言数据进行中文分词。本文采用 python 的中文分词组件 jieba 完成此工作。jieba 分词主要是基于统计词典，构造一个前缀词典，然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置构造一个有向无环图，再根据选择的模式不同，根据词典寻找最短路径后对句子进行截取或直接对句子进行截取，对于不在词典中的词使用 HMM 进行新词发现实现较好的中文分词效果。

此外，在实验过程中发现只对留言内容做训练和测试的结果与对留言内容和主题同时做测试的 F-Score 结果有所不同，而且对标题进行提取关键词并加强权重后结果会更好一些，相关数据如表 2.1 所示：

表 2.1 不同数据实验的 F-Score 结果

相关数据的实验操作	F-Score 结果
只对内容做训练和测试	0.765
对内容和标题做训练和测试	0.788
采用 TextRank 算法提取标题关键词并权重加强	0.810

2.1.3 TF-IDF 特征提取

在对留言数据进行分词后，需要将这些词语转换成向量，以供后续挖掘分析使用。未完成此任务，本文采用 TF-IDF 算法将文本数据，即留言主题和留言内容，转换为权重向量。实现 TF-IDF 算法的基本步骤如下：

第一步：计算词频，即 TF(Term Frequency)权重

词频（TF）表示词条（关键字）在文本中出现的频率，这个数字通常会被归一化(一般是词频除以文章总词数)，以防止它偏向长的文件[1]。对于在某一特定文件里的词语 t_i来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}} \quad (1)$$

$n_{i,j}$表示词语t_i在文件d_j中的出现次数，$\Sigma_k n_{i,j}$则是文件d_j中所有字词出现的次数之和。

词频的计算方法为如公式（2）或（3）所示：

$$\text{词频（TF）} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}} \quad (2)$$

$$\text{词频（TF）} = \frac{\text{某个词在文章中的出现次数}}{\text{该文本出现次数最多的词的出现次数}} \quad (3)$$

第二步：计算逆向文件频率(IDF)权重，IDF 是一个词语普遍重要性的度量，如果包含词条 t 的文档越少，IDF 越大，则说明词条具有很好的类别区分能力[2]。某一特定词语的 IDF 可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到，即：

$$\text{逆文档频率（IDF）} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right) \quad (4)$$

第三步：计算 TF-IDF（Term Frequency Document Frequency）权值

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语，其公式为：

$$\text{TF-IDF} = \text{词频（TF）} * \text{逆文档频率（IDF）} \quad (5)$$

实际分析得出 TF-IDF 值与一个词在留言信息表中出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。因此计算文本中每个词的 TF-IDF 值并排序，将出现次数最多的词语描述为留言信息类别的关键词。

生成 TF-IDF 向量的具体步骤如下：

（1） 使用 TF-IDF 算法找出描述每个类别的关键词。

（2） 将描述每个类别的关键词合并成一个集合，计算每个关键词在文本中的词频，如果没有则记为 0，有则记为 1。

（3） 生成描述各个类别的 TF-IDF 权重向量。

2.1.4 朴素贝叶斯分类算法

生成留言主题和留言内容的 TF-IDF 权重向量后，根据每个关键词的 TF-IDF 权重向量对留言主题和内容进行分类，本文采用朴素贝叶斯分类算法将留言数据分为 7 类.

朴素贝叶斯通过先验概率和类别的条件概率来估计文档d对于类别c_j的后验概率，以此实现对文档d的类别归属判断[3]。针对问题一，本文应用朴素贝叶斯方法，通过获取文本的先验概率和文本特征的条件概率，计算文本属于各类别的后验概率，并将后验概率最大的类别作为分类结果[4]。算法步骤如下：

(1)确定已知分类的待分类项集合，即训练样本集。在训练样本集和类别集合上计算每个类别的先验概率$P(c_j)$，

$$P(c_j) = \frac{\text{c}_\text{j}\text{的文档总数}}{\text{总文档个数}} \quad (6)$$

(2)统计得到在各类别下各个特征属性t_i的条件概率估计$P(t_i|c_j)$，

$$P(t_i|c_j) = \frac{t_i\text{在 c}_\text{j}\text{的文档中出现的次数}}{\text{c}_\text{j}\text{的文档中出现所有词的次数}} \quad (7)$$

(3)计算待测文本d属于每个类别c_j的后验概率[5]，取后验概率最大的类别作为文本的类别。算法流程如图 2-3 所示。

留言信息表一给出了 7 个类别共 9211 条包含用户、主题、时间、内容等特征的数据。将 7 个类别提取留言主题和内容单独分类出来，选取每个类别五分之四的数据作为训练集，剩余五分之一数据作为测试集,处理程序详见 Naive_Bayes.py。数据集经过去停用词、分词、求 TF-IDF 向量，使用朴素贝叶斯分数算法等操作得到关于留言数据 7 个分类的分类模型。

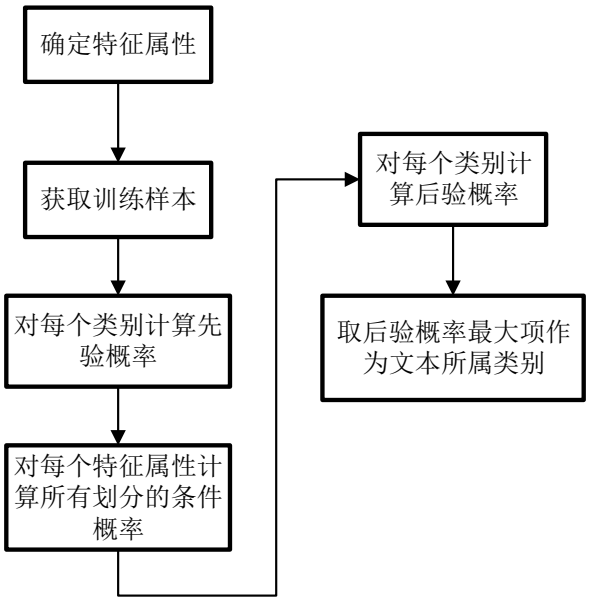

图 2-3 贝叶斯分类算法流程图

2.2 问题二的分析方法与过程

2.2.1 流程图

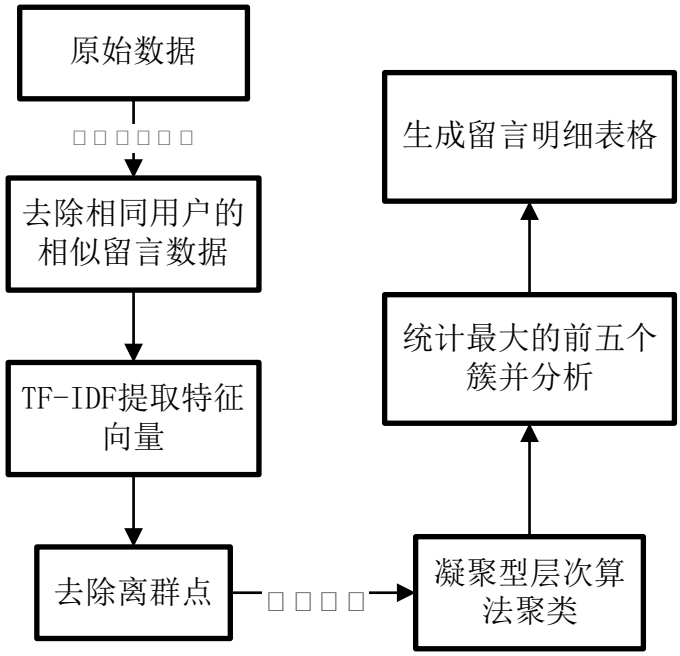

图 2-4 问题 2 流程图

2.2.2 数据预处理

原始数据的去重与重排序：本问题一共给出了关于留言数据的 4327 条数据，去除所有数据中的 118 条重复数据，对剩余的 4208 条数据，根据题目只筛选排名前五的热点问题的要求，经过大量的试验发现以下两点特征：

（1）在整个文本中，很多留言的内容出现过的次数很少，而且不是题目所要求的留言

内容且对聚类算法的性能基本没有影响，为了提升算法的性能减少运行时间，将这种出现次数极少的留言内容删除。

（2）有一部分用户会在一个时间段内重复留言相同内容，希望政府及时处理好用户所关心的事情。但是，用户重复留言的内容会影响热度排行的准确性。因此，在数据预处理过程中，此类数据被删除。

此外，对剩余的数据按照留言时间的先后顺序进行排序，这样使得在后续使用聚类算法进行相似度检测时能够减少运行时间，一定程度上提升了性能并且方便添加时间约束条件，有利于程序处理。

在进行热度排行聚类的过程中，除了相同留言的问题，发现某些用户会不停地留言内容相近、含义相同的内容以期望问题得到关注和解决。但是热度排行指标是根据不同用户相似内容的个数确定的，因此此类数据会对排行的准确度产生影响。为解决这个问题，将 excel 文件按照用户名进行排序，并依次去除相同用户的类似留言另存为 excel 表格，同一用户的相似留言一共去除 571 条数据，程序详见 User_repeat_message.py。不同用户留言相似度的评判依据是根据 TF-IDF 进行特征提取并两两计算相似度，若相似度小于阈值则保留其中一条数据，之后生成一个相似度留言列表，按照列表索引去除数据项，最后生成 excel 表格，后续数据处理操作均基于此表格。

原始数据经过去重与重排序处理之后，去掉对后续挖掘分析没有太大意义的词语，通过 jieba 分词工具进行中文分词并计算每个词语的 TF-IDF 值，将其转换为权值向量。这样的预处理操作更有利于后续判断文本的相似性。

2.2.3 聚类分析

聚类分析是一种无监督机器学习算法，它的目标是将相似的对象归到同一个簇中，将不相似的对象归到不同的簇中。如果要使用聚类分析算法对一堆文本分类，要先了解关于相似度计算的理论知识，明白什么样的聚类结果是比较好的，从而选择合适的聚类算法。

余弦相似度使用向量空间中两个向量夹角的余弦值作为衡量两个个体差异的大小。在分析解决问题的同时，发现相比欧氏距离度量，余弦相似度更注重两个向量在方向上的差异，所以更适合计算文本相似度。其公式如下：

$$\cos\theta = \frac{\Sigma_1^n (A_i \times B_i)}{\sqrt{\Sigma_1^n A_i^2} \times \sqrt{\Sigma_1^n B_i^2}} \tag{8}$$

使用余弦相似度算法时，首先将文本转换为权值向量，通过计算两个向量的夹角余弦值评估其相似度。余弦值的范围在[-1,1]之间，越趋近于 1，代表两个向量方向越接近；越趋近于-1，代表他们的方向越相反。为了方便聚类分析，将余弦值做归一化处理，将其转换到[0,1]之间，并且值越小距离越近。

在聚类过程中，发现大部分用户的留言信息内容相关性并不强，基本上不会影响算法结果。因此本文采用 Local Outlier Factor (LOF)算法去除相关性不强的数据。

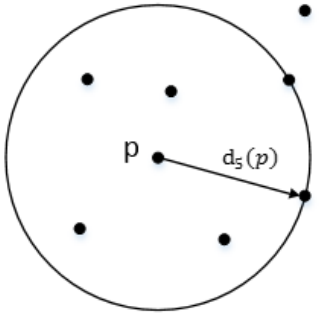

图 2-5 LOF 算法示意图

如图 2-5 所示，LOF 通过比较每个点 p 与其邻域点的密度来判断该点是否为异常点，如果点 p 的密度越低，越可能被认定是异常点。密度是通过点与点之间的距离来计算的，点之间的距离越远，密度越低；反之距离越近，密度越高。另外，为避免因为数据密度分散情况不同而错误的将正常点判定为异常点，LOF 在点的第 k 邻域而非全局计算其密度，因此得名为“局部”异常因子。LOF 需要选取合适的半径和密度以确定离群点，具体思路是：

（1）选定示例数据和一部分测试数据，对其进行统计离群点并观察是否有离群点在较大的簇里面。

（2）结合人工校验，确定一个合适的半径。

（3）将此半径作为离群计算的阈值，基于剩下的 3637 条数据，对所有样本做离群分析共剔除 487 个离群点。

凝聚型层次聚类算法的目的是在不同的层次对数据集进行划分，可以采用“自底向上”的聚类策略，也可以采用“自顶向下”的分拆策略。本文采用“自底向上”的策略，思路是先将数据集中的每个点看作一个初始聚类簇，然后找出两个距离最近的两个簇进行合并，不断重复该步骤，直到达到预设的聚类个数或某种条件[5]。聚类算法总体流程给如图 2-6：

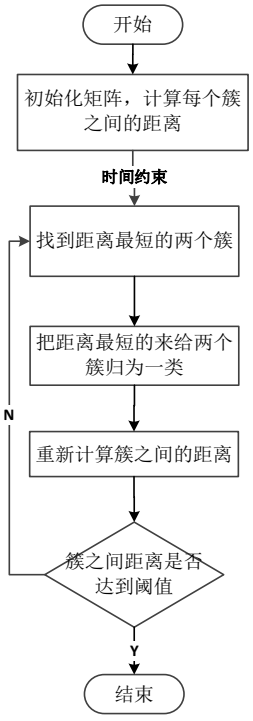

图 2-6 聚类算法流程图

与其他聚类算法类似，凝聚型层次聚类也需要确定合适的聚类个数或某种结束条件，本文采用的具体思路是：

（1）选定示例数据和一部分测试样本，对其进行层次聚类分析。

（2）结合人工校验，得到一个合适的聚类个数和对应的距离阈值。

（3）将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。此时无需再计算 DBI 和 DI 值，计算效率可以大幅提升[7]。

根据问题二的要求，希望将相似的样本尽可能划分到同一类中，同时可以接受少部分不同的样本划分到同一类，然后基于最大距离、最小距离、平均距离进行分析，计算每个簇里有多少数据并按照簇中数据量进行降序排列，排序越靠前说明该类问题的关注度越高，可将其归为热点问题，聚类程序详见 hcluster.py。

2.3 问题三的分析方法与过程

2.3.1 特征分析

本题以相关部门对留言的答复意见作为研究对象，分析了留言回复中的文本特征、统计信息、留言者和回复者之间的关系以及留言和回复的关联度。从相关部门对留言答复意见的相关性、完整性、可解释性等角度进行分析并基于此建立一个面向留言的回复质量分类器[8]。

为解决此问题，本文将从留言回复的以下 3 个方面特征进行评价:

（1）留言回复的文本特征；

（2）留言回复的统计信息；

（3）留言和回复信息的关联度。

通过在文本特征提取中对这三个特征进行评价,建立对留言回复的初级筛选，为最后的分类器能够在留言问答系统中自动选取高质量的留言回复提供基础保障。

关于文本特征，主要基于以下三个假设：

（1）句子的长度。主要包括用户反应的问题以及问题的长度，一般高质量留言的长度较长。

（2）标点符号。用户留言中肯定会存在有标点符号，但是标点符号也不会太多[9]。

（3）疑问词。对于用户留言来说，用户肯定会用疑问词。

关于统计特征分析，主要有以下几个方面的考虑：

（1）留言回复的状态.针对社会留言回复的状态有 5 种,分别是:待解决、待评价、已关闭、已解决（有令用户满意的）和已解决（没有令用户满意的）。

（2）回复的个数，包括用户满意的和不满意的。

（3）是否是用户满意的回复。

（4）最佳回复的好评率。用户满意回复的好评率高低表明留言和回复的关联度的高低,同时回复质量越高，表明反应相关问题的用户越多，也反映了留言回复的质量。

2.3.2 关联度计算

用户留言和质量较好的回复之间有较强的关联度，这两者之间一定存在关键词。假设留言的关键词集合为 M(x1,x2,…,xn),留言回复的关键词集合为 N(x1,x2,…,xm),根据集合 M

和集合 N 的关键词匹配数来计算留言和回复之间的关联度，其公式如下：

$$\mathrm{sim(Q,A)} = \frac{C(Q \cap A)}{C(Q) U C(A)} \tag{12}$$

其中 C(M)表示集合 M 中关键词个数,C(N)表示集合 N 中的关键词个数,C(M∩N)表示集合 M 和集合 N 匹配的关键词个数,sim(M,N)即留言和回复之间的关联度[10]。

2.3.3 分类器设计

在开始对留言回复的质量进行评价时，首先对留言回复信息进行文本特征提取，根据句子长度，是否含有标点符号，以及是否含有疑问词这个三个文本特征提取标准，将留言分别按 45%、15%、40%来确定其权重,初步对留言回复的质量有一个初级划分；然后进行统计信息分析,对于用户留言给予最官方，最权威的处理办法，落实解决留言反应问题的回复,本文将其认为是质量最好的,对于模糊回应的留言回复,则需要分析其它的统计信息[11]。

对于用户反应相关的问题,本文附件 4 给出的相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度分析用户的权威度并假设权威度高的用户反应的问题，或者回复留言质量相对比较高。最后通过留言与回复中的关键词来计算它们的关联度[12]。该分类器算法的流程图如图 2-7 所示。

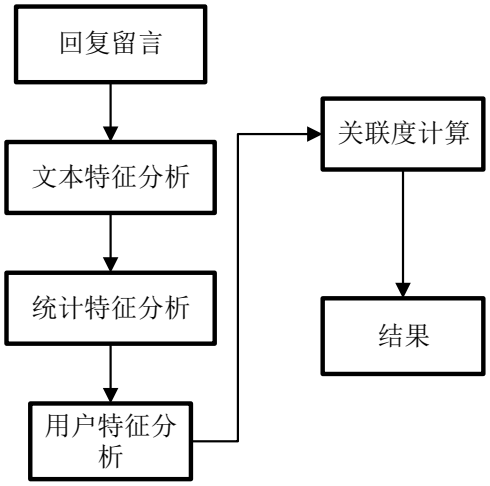

图 2-7 分类器算法流程

3、结果分析

3.1 问题一的结果分析

3.1.1 数据预处理结果分析

本问题使用的留言数据，有留言用户、主题、时间、内容以及每一条留言对应的类别标签。在分类问题中，留言主题和留言内容以及它们对应的类别标签是重要的关注点，因此本文按照类别提取留言主题和留言内容，对其进行去停用词、分词和特征提取等操作。

在只对留言内容进行分词和特征提取之后的分类效果并不是很理想，所以又考虑在数据预处理阶段对留言主题也进行了特征提取。采用 TextRank 算法对留言主题进行分词和词性

标注处理，并过滤掉停用词，只保留指定词性的单词，比如名词；然后通过词之间的相邻关系构建候选关键词图 G=（V,E），迭代计算每个节点的 rank 值，排序 rank 值即可得到关键词。

3.1.2 分类结果分析

在对留言主题和留言内容进行去停用词、分词和特征提取之后，使用贝叶斯分类器对留言内容进行分类，得出每一个分类的准确率，即查准率和查全率，然后使用 F-Score 对分类方法进行评价。这里对数据集采用五分之四作为训练，五分之一做测试进行试验，数据处理程序见 data_proceess.py，贝叶斯分类程序见 Naive_Bayes.py。查准率和查全率的结果如表 3-1 所示，通过计算的到 F-Score 的值为 0.808。

表 3-1 各类别查准率和查全率

类别	查准率	查全率
城乡建设	0.739	0.945
环境保护	0.963	0.840
交通运输	0.961	0.402
教育文体	0.892	0.906
劳动和社会保障	0.786	0.960
商贸旅游	0.916	0.720
卫生计生	0.983	0.663

根据表 3-1 的实验结果可知，除了城乡建设、劳动和社会保障两类的查准率较低以外，其他类别整体较高，平均查准率达到 94.3%。经过分析发现，城乡建设、劳动和社会保障两类查准率较低的原因是文本语义带来词语交叉使得该分类与其它分类出现了交叉识别，比如城乡建设类别中的“绿化进展”极大可能被分成是环境保护类。经过对比交通运输的分类错误数据，发现绝大多数交通运输类别被错误分类到了城乡建设与劳动和社会保障的类别，由此推测交通运输中的很多特征在城乡建设与劳动和社会保障类别中，而城乡建设与劳动和社会保障中的一些特征在交通运输类别里面并没有。

基于上面的推测，提出分别剔除交通运输和城乡建设在城乡建设中的子集以及交通运输与劳动和社会保障类别在劳动和社会保障中特征的交集，这样可以提升查全率。

3.2 问题二的结果分析

3.2.1 聚类结果分析

针对问题二经过去重后得到 4327 条数据，之后再去除统一用户得相似留言得到 3637 条数据，然后对文本进行分词、去停用词和特征提取等预处理操作后，进行离群点得去除，最后得到 3150 条数据，图 3-1 为离群点去除示意图：

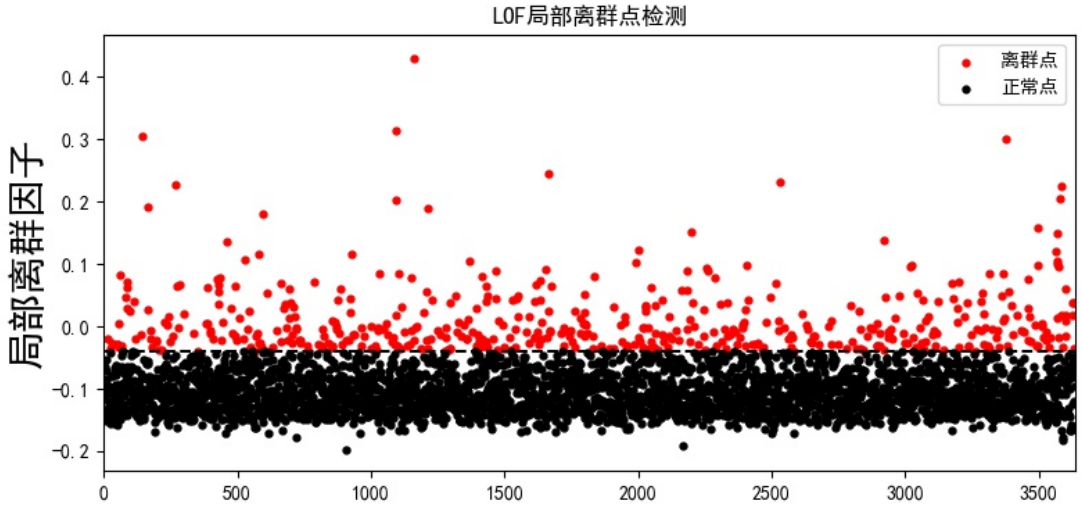

图 3-1 离群点去除示意图

由层次聚类算法得到聚类个数，其大致步骤是：

（1）利用示例数据及部分测试样本，对其进行层次聚类分析。

（2）结合人工校验，得到一个合适的聚类个数和对应的距离阈值。

（3）将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。

结合测试样本，得出将近 1000 个聚类个数，按照每个聚类簇中的数据数量从大到小排序，获取排名前五的留言明细表，其部分结果如表 3-2 所示。

表 3-2 部分热点问题留言明细表

热度排行	留言编号	留言用户	留言主题	留言时间	留言详情	赞成数	反对数
1	277363	A0001108(	A2区万芙路清政园小区地下车位外租导致业主没地方停车	2019/11/8 10:47:12	A2区万芙路清政园小	0	0
1	231153	A0003033(	A3区佳兆业云顶梅溪湖二期地下车位使用严重不符合当时购买信息	2019/6/26 11:54:19	我是A3区佳兆业云顶	0	0
1	195528	A0006953	A3区车塘河路公园尚小区物业强制买车位	2019/4/17 10:47:26	A市A3区车塘河路公	0	0
1	225859	A0008412(	A7县家和院1288个地下车位转让给非本小区业主是否合法	2019/9/19 0:40:59	尊敬的各级各部门领	0	0
1	244917	A0007479(	A7县深业睿城小区f区地下车位是否计入公摊面积?	2019/6/4 10:29:50	深业睿城小区f区地下	0	1
1	233543	A0007331(	A7县深业睿城小区部分业主私自霸占地位	2019/9/22 23:53:31	沈书记您好！我是深	1	2
1	215123	A0007479(	A7县深业睿城开发商违法出售人防车位，请政府查处	2019/8/22 17:55:04	深业睿城f区地下车位	0	8
1	245956	A0004390(	A7县金科天悦小区存在两个问题盼协调解决	2019/10/22 15:32:14	本人于2019年9月1日	0	1
1	232892	A0001533(	A市万科魅力之城开发商未通知业主就进行车位开盘销售活动	2019/1/8 9:54	您好！我是57栋业主	0	1
1	218739	A909184	A市伊景园·滨河苑欺诈消费者	2019/8/27 14:55	A市伊景园滨河苑强行	0	0
1	268626	A909186	A市伊景园滨河苑坑害购房者	2019/8/22 13:20	A市伊景园·滨河苑违	0	0
1	276460	A909170	A市伊景园滨河苑捆绑销售车位是否合理?	2019/4/22 17:45	尊敬的领导，您好，	0	0
1	258386	A909185	A市伊景园滨河苑欺压百姓	2019/8/27 14:12	A市伊景园·滨河苑强	0	0
1	251601	A909187	A市伊景园滨河苑诈骗钱财	2019/6/21 18:25	伊景园滨河苑收取购	0	0
1	258295	A0004893(	A市富基世纪公园捆绑销售变成强买强卖	2019/12/31 21:00	尊敬的领导：您好！	0	0
1	209506	A909179	A市武广新城坑客户购房金额并且捆绑销售车位	2019/8/14 21:55	您好！由A市广铁集团	0	0
1	228317	A0001113(	A市高新区B4区涉外景园的人防车位可以购买吗?	2020/1/6 15:46	您好，我是A市高新区	0	0
1	246785	A0001592(	不需要车位,不想白扔12万	2020/2/6 16:06	我们是还没住进武广	0	0
1	246407	A0009959(	举报广铁集团在伊景园滨河苑项目非法绑定车位出售	2019/1/8 18:22	我要举报广铁集团在	0	0
1	244528	A909235	伊景园滨河苑开发商强买强卖!	2019/1/6 20:26	A市广铁集团伊景园滨	0	2
1	205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！	2020/1/6 20:26:46	广铁集团强制要求职	0	1
1	209571	A909200	伊景园滨河苑项目绑定车位出售是否合法合规	2020/1/8 20:26	广铁集团铁路职工定	0	0
1	190337	A0009051(	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/1/6 20:26	投诉伊景园·滨河苑开	0	0
1	195995	A909199	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题	2020/1/6 20:26:46	尊敬的市政府领导，	0	0
1	242474	A0005143(	再次投诉A市北大资源理想家园违规租售车位	2019/10/24 10:08:32	业主联名严正诉求：	0	1
1	289473	A0001034(	反对滨河苑房子和车位捆绑销售	2020/1/6 20:26:46	现有伊景园滨河苑在	0	0

3.2.2 热点留言分析

通过对将近 1000 个聚数目进行计数排序，选取排名前 5 的留言数据进行分析，并给出热点问题表 3-3。其中，热度指数是指该问题下去除同一用户相似留言后的留言条数。

根据表 3-3，可以发现热点问题都集中在 A 市，相关问题涵盖小区物业管理、房产开发涉及的违规问题、幼儿园收费惠普性问题、小区内物业管理问题。可见 A 市涉及到许多民生问题亟待解决。

表 3-3 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	48	2019/01/08 至 2020/02/06	A市伊景园	伊景园车位捆绑销售强买强卖
2	2	34	2019/01/01 至 2019/12/26	A市房产开发	A市多地区房产开发涉及违规问题
3	3	33	2019/01/07 至 2019/11/22	A市幼儿园	幼儿园惠普性问题
4	4	30	2019/01/07 至 2019/11/08	A市物业公司	物业公司各种不合理收费
5	5	28	2019/01/06 至 2019/12/26	A市小区	小区街道管理脏乱差

3.3 问题三的结果分析

根据留言回复的文本特征对相关部门的留言进行特征提取，具体提取标准主要包括留言回复的文本特征、留言回复的统计信息以及反应内容和留言回复的关联度。基于文本特征的三个条件筛选后，根据质量高的回复和留言之间有较强关联度的特点,它们之间会有共同关键词，再对处理过的文本留言进行关联度计算。完成了留言回复信息的初步筛选，构建一个衡量留言质量的分类器，然后利用分类器，对留言的质量进行评估，通过 P(准确率)、R(召回率)来评价分类器的性能，以此来获得质量最好的留言答复。

分别计算 T,T+S ,T+S+R 情况下的 P(准确率)、R(召回率)得出实验结果。其中，T 是用户留言的文本特征，S 表示加入了统计信息的特征，R 表示加入了留言和留言回复的关联度特征。根据结果分析得出，随着加入的特征越多，实验在准确率方面有了很大的提高。但是在加入了统计信息特征后,召回率开始逐渐下降,但下降的趋势不大,经过小组成员分析得出，可能出现这种情况的原因是测试集里面的关联度计算准确度不够。

4. 结论

本文对来自互联网公开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进行分析研究。通过分词、特征提取、构建分类器等技术得到关于留言内容的分类模型，使用聚类算法对相似留言问题进行聚类得到关于热点留言问题及其明细结果。根据 F-Score 对分类方法进行评价，整体来说，各类别的查准率和查全率效果较好，根据聚类算法得到的热点留言问题明细在去除重复数据之后,结果也令人满意。除此之外,对留言回复进行特征提取、关联度比较、分类器构造等步骤对留言质量进行评估，结果表明分类器的算法性能较好。综合来说，通过对问题的分析、数据的处理与挖掘、项目试验的操作和模型的建立、对结果的分析和实验过程的改进，基本上实现了本赛题的几个挖掘目标：（1）构建关于留言内容的分类模型，（2）热点留言挖掘和热点留言明细呈现，（3）给出关于留言答复的评价方案。

由问题的结果分析可以看出，大部分群众反映的是关于城乡建设、交通运输业等方面的问题，这说明基础建设与民生息息相关。许多用户在不同时间段对相似的问题进行了多次留言，说明该问题在民众生活中比较重要，同时也说明该问题没有得到相关部门的及时回复。

5. 参考文献

[1] 隗中杰. 文本分类中 TF-IDF 权重计算方法改进[J]. 软件刊, 2018, 17(12): 39-42.
[2] 石凤贵. 基于 TF-IDF 中文文本分类实现[J]. 现代计算机, 2020, (06): 51-54+75.
[3] 郑丽香, 凌亚东, 陈泫文, 李颖, 刘馨阳. 基于改进朴素贝叶斯方法的元器件分类技术[J]. 电子产品可靠性与环境试验, 2020, 38(01): 49-53.
[4] 卢玲, 王越, 杨武. 一种基于朴素贝叶斯的中文评论情感分类方法研究[J]. 山东大学学报:工学版, 2013(06): 11-15.
[5] 何伟. 基于朴素贝叶斯的文本分类算法研究[D]. 南京邮电大学, 2018.
[6] 刘凤芹. 聚类算法研究[J]. 计算机光盘软件与应用, 2012, 000(021):60-61.
[7] 马晓艳, 唐雁. 层次聚类算法研究[C]// 2008 年计算机应用技术交流会论文集, 2008.
[8] 刘高军,马砚忠,段建勇. 社区问答系统中“问答对”的质量评价[J]. 北方工业大学学报, 2012, 24(03): 31-36.
[9] 姜雯,许鑫. 在线问答社区信息质量评价研究综述[J]. 现代图书情报技术, 2014(06): 41-50.
[10] 许世华. 社区问答系统中问题理解的关键技术研究[D]. 苏州大学, 2017.
[11] Shah C., Pomerantz J. Evaluating and Predicting Answer Quality in Community QA[C]// Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010. ACM, 2010.
[12] Hengyi F., Shuheng W., Sanghee O. Evaluating answer quality across knowledge domains: Using textual and non- textual features in social Q& A[J]. John Wiley & Sons, Ltd, 2015, 52(1).