“智慧政务”中的文本挖掘应用

摘要

随着互联网的发展与应用，网络留言成为人们相应反馈意见与建议的一种重要方式，网络留言使用简单方便，给人们生活带来极大的便利，但是大量的留言非常复杂，政府部门处理起来比较困难，因此怎样对这些留言数据进行合理分析以便于相应政府部门进行处理成为现今数据库技术的研究热点。

针对问题一：对于附件 2 所给的 9210 条留言数据，利用 *jieba* 分词、去停用词后，按照 4:1 的比例来选取训练集和测试集进行训练与测试，再对每条留言信息进行文本数据的向量化表示，形成加权矩阵，从而获取训练样本和测试样本的 $TF-IDF$ 权值向量。最后采用 *MultnomialNB* 、*GaussianNB* 、*Logistic* Re *gression* 、*MLP* 、 *RandomForest* 等分类方法进行留言分类，通过多种方法的比较，发现神经网络分类效果最佳，准确率可达 0.8977。

针对问题二：首先对附件 3 的 4326 条数据进行聚类处理，本文采用了 $k-means$ 聚类、*birch* 聚类两种方法。对于 k 值的确定，本文采取经验法对 k 值进行预估，再结合手肘法、轮廓系数法最终确定 k 值为 17 时最佳，再进行算法相似度参数计算，在确定高斯核函数的参数后，使用改进后 *birch* 的聚类算法，将之前已经归一化处理的文本向量进行聚类操作，将聚集的 17 类的留言主题提取关键词画出词云图，再对地点人群、问题描述进行总结，找出了五类影响热点的指标，对最后基于留言聚类的热点问题进行排序，得出热度排名前 5 的留言。

针对问题三：本文在相关性、完整性、基础性的基础上，增加了准确性和正规性两个指标，也即构建了五个评价指标，按比例对留言进行等级划分，并且定义评价等级，最终给出每条答复意见的评价等级，最后进行评价方案验证，将留言回复所得到的分数按常规比例划分等级，再与标准等级区域进行对比，得到本套评价方案可以作为回复质量的评定。

关键词：神经网络；$k-means$ 聚类；*birch* 聚类；文本分类

Abstract

With the development and application of the Internet, network message has become an important way of people's corresponding feedback and suggestions. Network message is simple and convenient to use, which brings great convenience to people's life. However, a large number of messages are very complex, and it is difficult for government departments to deal with them. Therefore, how to reasonably analyze these message data to facilitate the corresponding government departments to deal with them It has become a research hot spot of database technology.

To solve the problem 1: for 9210 message data given in Annex 2,after using jieba participle and removing stop words,Select training set and test set according to the ratio of 4:1 for training and testing,Then, the text data of each message is vectorize to form a weighting matrix, so as to obtain the TF weight vectors of training samples and test samples. Finally,multnomialNB、GaussianNB、LogisticRegression、MLP、RandomForest and other classification methods are used to birch classify messages. Through the comparison of various methods, it is found that the neural network classification effect is the best, with an accuracy of 0.8977.

To solve the problem 2: firstly, 4326 pieces of data in Annex 3 are clustered. In this paper, k-means clustering and birch clustering are used. For the determination of K value, this paper adopts the empirical method to estimate K value, then combines the elbow method and the contour coefficient method to finally determine the best value when the value is 17, and then calculates the algorithm similarity parameters. After determining the parameters of Gaussian kernel function, the improved clustering algorithm is used to cluster the previously normalized text vector, and extract the 17 types of message topics Key words draw word cloud map, then summarize the description of the place population and

problems, find out five kinds of indicators that affect the hot spots, rank the hot issues based on message clustering finally, and get the top 5 messages in the hot degree ranking.

To solve the problem 3: Based on the relevance, integrity and foundation, this paper adds two indexes of accuracy and normality, that is to say, five evaluation indexes are constructed to grade the messages according to the proportion, define the evaluation grade, and finally give the evaluation grade of each reply. Finally, the evaluation scheme is verified, and the scores obtained by the reply to the messages are compared with the conventional ones The evaluation scheme can be used to evaluate the quality of reply.

Keywords: neural network; k-means clustering; birch clustering; text classification

目录

1 问题重述

随着互联网的发展与应用，计算机非常重要的特征是信息化、数字化和网络化，网络留言成为人们相应反馈意见与建议的一种重要方式，它可以实现网站与客户之间及不同客户之间的交流与沟通，网络留言使用简单方便，给人们生活带来极大的便利，但是大量的留言非常复杂，政府部门处理起来比较困难，因此怎样对这些留言数据进行合理分析以便于相应政府部门进行处理成为现今数据库技术的研究热点。

本次建模目标利用网络问政平台的群众留言，按照一定的划分体系对留言进行分类，以便将其分派至相关部门处理，对文本数据进行预处理，利用 *jieba* 中文分词工具和停用词表进行文本的分词和去除停用词，根据分词过后的数据绘制词云图，使用 $tf-idf$ 将文本数据转化为向量，然后文本进行分类，构建一级标签分类模型；对留言进行某一时间段内反应特定地点或人群问题进行归类，并且定义热度评价指标体系，给出热度排名前 5 的热点问题；最后从政府对居民答复的相关性、准确性、完整性、正规性、可解释性等角度对答复意见的质量给出一套评价方案。达到以下目标：

（1） 利用分词工具对留言数据进行挖掘，将留言数据进行分类；将其划分为对应得标签，以便政府部门进行处理。分析居民们反映的问题，了解当今民情现状 ，改善人们的生活。

（2）进行留言分类，构建热度评价指标，及时发现热点问题，相关职能部门进行高效率的处理，及时解决人们经常反映的热点问题。

（3）提高答复意见的质量，更高效合理的解决居民反映的问题。

2 符号说明

符号	意义
$DF(t_i)$	出现特征词的个数

w_{ij}	j 词在 i 类文本下的 $tf-idf$ 矩阵
a_{ij}	j 词在 i 类文本下的词频
$U \in R^{d \times m}$	变换矩阵
μ_j	特征均值
$x_i^{(j)}$	对第个 i 样本的第 j 个特征
μ_i	每个类 c_k 类别中心
$J(c_k)$	各点到聚类中心 μ_i 的距离平方和
$J(C)$	各类总的平方和
$a(i)$	聚类凝聚度
$b(i)$	聚类分离度
$s(i)$	样本点轮廓系数
SSE	和方差
$N(topic)$	文本总数
$T(first)$	最早发布时间
$TN(topic)$	留言持续时间
$N(new)$	最新发布文本量
$N(tmp \mid ops)$	点赞和反对总数
$score$	留言总得分

3 文本分类

3.1 数据预处理

在附件 2 中有 7 个标签，每个标签随机抽取 600 个样本，在数据预处理之后以4:1的比例来获取训练集和测试集样本。删除重复留言，重复的留言信息。然后对于留言信息进行结巴分词处理，导入中文停用词表：哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词库以及自己的停用词表。将全部数据分割为若干个词组，利用停用词表剔除文本当中的停用词等无关内容，去除留言中的空格序列，自定义函数实现机械压缩，去除留言详情中重复词。降低后续工作的难度，提高系统的工作效率。对留言内容进行中文分词，采用 *python* 中的中文分词包 *jieba* 进行分词，主要特点：支持精确模式、全模式、搜索引擎模式三种分词模式；支持繁体分词；支持自定义词典。它所涉及的算法有基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图；采用了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；对于未登陆词，采用了基于汉字成词能力的 *HMM*（马尔科夫）模型，使用了 *Viterbi* 算法，使得能更好的实现中文分词效果。

3.2 *TF – IDF* 算法

在进行留言内容分词后，需要把这些词语转换为向量，以供挖掘分析数据使用，利用词频信息对每条留言信息进行文本数据的向量化表示，形成矩阵。在上面的 *jieba* 分词中，已经去掉了那些次数出现太少的单词，因此我们用 *TF – IDF* 权重策略，从而将留言数据向量化，其主要思想：如果某个词或短语在一篇文章中出现的频率 *TF* 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力。*IDF* 就是用来衡量一个单词在所有文本中的出现频率，其计算公式为：

$$IDF = \log(\frac{W}{1 + DF(t_i)})$$

其中W表示总文档数，$DF(t_i)$表示在文本数据集中出现特征词的文本个数。

$TF-IDF$算法是经典的关键词提取方法，是目前应用最多的基于统计信息的关键词提取方法，分为三大模块：

（1）文本预处理：先输入文本数据a，首先进行了分词等预处理操作，然后把文本a的内容看成由特征词的组合，假设文本a用特征词的集合表示为$a_i=(t_1,t_2,\cdots,t_i,\cdots t_n)$其中$t_i$是特征项。

（2）权重计算：根据各个项t_i在文本a_i中的重要性给其赋予一定的权重w_i，$TF-IDF$算法通过特征词词频(TF)和反文档频率(IDF)来计算特征词t_i的权重w_i，文本a_i的特征词的权重计算公式如下：

$$TF-IDF(a_i,t_i)=TF(a_i,t_i)\times IDF(t_i)=TF(a_i,t_i)\times\lg(\frac{W}{1+DF(t_i)})$$

$TF(a_i,t_i)$表示特征词在文本中出现的次数，$DF(t_i)$表示在文本数据集中出现特征词的文本个数，W表示文本数据集总的文本数，$IDF(t_i)$表示反文档率。

（3）提取关键信息：按照权重w_i从大到小对特征词t_i进行排序，选择前m个词作为文本a最终的关键词。

注：（1）在idf的计算公式中使用了log的原因是，如果某个词在这个文本中的出现次数太少，从而得到的值就会非常大，这样就会对计算结果影响非常大，因此我们取对数来抑制idf的影响。并增加词频信息，之后进行归一化，避免句子长度不一致，从而将留言数据转换成向量矩阵。

（2）实际分析得出$TF-IDF(a_i,t_i)$值与一个词在留言数据中出现的次数成正比，也即$TF-IDF$的值越大，代表这个词在留言文本中的重要性越高，因此我们按照权重大小来进行排序，次数最多的即为要提取的留言信息的关键词。

3.3 获取$TF-IDF$权值向量

训练样本以及测试样本生成$TF-IDF$权值向量的具体步骤：

（1）使用$TF-IDF$算法，根据权重大小选取关键词；

（2）把这些关键词合并成一个集合，计算每个留言对于这个集合的词频，有些需要并增加词频信息，之后进行归一化，避免句子长度不一致，从而将留言数据转换成向量矩阵，如果没有就记为 0；

（3）生成各个留言的$TF-IDF$权值向量，计算公式：

$$TF-IDF(a_i,t_i)=TF(a_i,t_i)\times IDF(t_i)=TF(a_i,t_i)\times \lg(\frac{W}{1+DF(t_i)})$$

3.4 留言数据分类

在提取关键词之后，为了建立留言内容的一级标签模型，我们采用$MultnomialNB$、$SGDC$、$Logistic\,\mathrm{Re}\,gression$、$MLP$、$RandomForest$等分类方法进行留言分类。

3.4.1 MLP神经网络分类器

多层感知机（MLP）也叫人工神经网络，除了输入输出层，它中间可以有多个隐层，最简单的MLP只含一个隐层，即三层的结构，如下图 1：

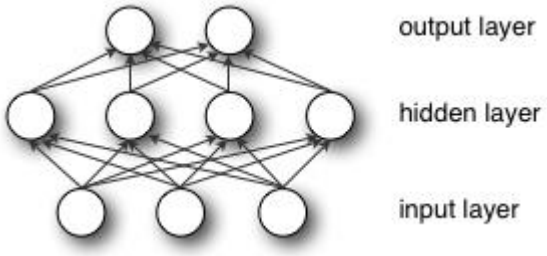


图 1 多层感知机

从上图可以看到，多层感知机层与层之间是全连接的（全连接的意思就是：上一层的任何一个神经元与下一层的所有神经元都有连接）。多层感知机最底层是输入层，中间是隐藏层，最后是输出层。

神经网络主要有三个基本要素：权重、偏置和激活函数。

权重：神经元之间的连接强度由权重表示，权重的大小表示可能性的大小。

偏置：偏置的设置是为了正确分类样本，是模型中一个重要的参数，即保证通过输入算出的输出值不能随便激活。

激活函数：起非线性映射的作用，其可将神经元的输出幅度限制在一定范围内，一般限制在$(-1,1)$或$(0,1)$之间。最常用的激活函数是*Sigomid*函数，其可将$(-\infty,+\infty)$的数映射到$(0,1)$的范围内。

神经网络的特点：

（1）高度的并行性

（2）高度的非线性全局作用

（3）良好的容错性与联想记忆功能

3.4.2 *LinearSVC* 分类器

支持向量机*LinearSVC*是*Vapnik*等人根据统计学习理论提出的一种新的机器学习方法，它是对结构风险最小归纳原则的近似，其特点是具有出色的学习性能，只需要较少的样本就可以迅速训练出具有相对较高性能指标的分类器。

支持向量机原理：设给定样本集$\{x_i,y_i\}=1$和核函数$k(x_i,y_j)$，其中$x_i\in R,y_i\in\{1,1\}$，$k(x_i,x_j)$对应某特征空间z中的内积，即$k(x_i,x_j)=\langle g(x_i),g(x_j)\rangle$，变换$g:x\to z$是将样本从输入空间映射到特征空间，支持向量机的数学模型如下：

$$\min_{w\cdot h}\frac{1}{2}\|w\|^2-c\sum_{i=1}^{m}\xi_i \qquad s.t.y_i<w,g(x_i)>+b\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\cdots,m.$$

式中：w为超平面的法向量，b超平面的偏量，ξ_i为松弛变量，c为惩罚因子。

3.4.3 高斯朴素贝叶斯

朴素贝叶斯方法是一种以贝叶斯定理为基础，以各个特征相互独立为假设的概率分类算法，朴素贝叶斯算法分类效率高，需要关注的参数少，具备良好的泛化能力，并且能够将最后的分类结果给出合理的概率解释。

对于给定的留言数据训练集中，朴素贝叶斯算法将输入特征向量定义为在输

入空间的随机变量X，输出（类标记）定义在输出空间的随机变量Y.

（1）先验概率分布： $P(Y=c_k), k=1,2,\cdots K.$

（2）条件概率分布：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k), k=1,2,\cdots,K.$$

因其假设各个特征相互独立，则条件概率可以表示为：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k).$$

（3）计算给定输入变量的后验概率分布：

$$P(Y=c_k\mid X=x)=\frac{P(X=x\mid Y=c_k)P(Y=c_k)}{\sum_k P(X=x\mid Y=c_k)P(Y=c_k)}=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)},$$

$k=1,2,\cdots K.$

（4）最后遵循期望风险最小化准则，保留后验概率最大的标记，朴素贝叶斯分类模型的输入变量既可以是离散型变量，又可以是连续型变量，其中，对于连续型变量应用最好的模型是高斯朴素贝叶斯(*Gaussian* NB)模型，其条件概率可以表示为： $P(X_j=x_j\mid Y=c_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp(-\frac{(x_j-\mu_k)^2}{2\sigma_k^2})$. 需要从训练样本集估计$\mu_k$和$\sigma_k^2$的值，$\mu_k$是在样本类别为$c_k$下，所有的$x_j$的均值，$\sigma_k^2$是在样本类别为$c_k$下所有$x_j$的方差。高斯朴素贝叶斯模型的主要参数只有一个，即先验概率$P(Y=c_k)$.通常情况下，默认为$P(Y=c_k)=m_k\mid m_m$是训练样本集总数，是输出第k类时的训练样本数。

3.5 模型评价

以*TP*表示预测为正、实际为正的样例数量*FP*表示预测为正、实际为负的样例数量, *TN*表示预测为负、实际为负的样例数量, *FN*表示预测为负、实际为正的样例数量.

(1)准确率$(TP+TN)/(TP+TN+FN+FP)$

(2) 查准率 $P = TP/(TP + FP)$

(3) 查全率 $R = TP/(TP + FN)$

(4) $F1 = 2 * P * R/(P + R)$， $F1$ 值 $(F1\ score)$，是统计学中用来衡量二分类模型精确度的一种指标，它同时兼顾了分类模型的查准率和查全率，其最大值为 1，最小值为 0，$F1$ 值可以看作是模型查准率和查全率的一种加权平均，是对模型性能进行评价的更全面的指标。

3.6 结果分析

通过对留言问题的分词，提取出训练集与测试集后向量化形成加权矩阵，采用 *MultnomialNB* 、*GaussianNB* 、*Logistic* Re *gression* 、*MLP* 、*RandomForest* 等 12 种分类后，得到相应的准确率（如下图 2 所示），最后得出人工神经网络（*MLP*）斯分类效果更好，使得训练集和测试集的匹配度更高，$f1-score$ 得分达到 0.883.

图 2 分类准确率结果图

可见 *MLP* 分类方法是效果最好的，准确率为 0.8977135980746089

通过使用对分类方法进行评价，得出的结果如下图 3

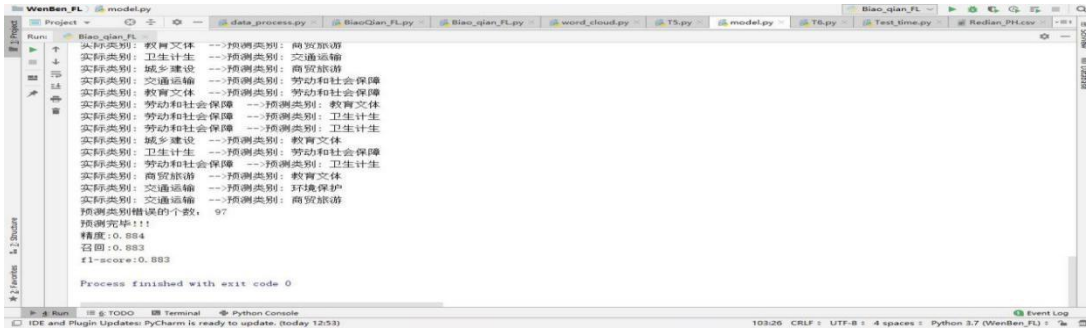


图 3 分类方法评价

4 热点聚类与排行

4.1 问题聚类

将文本处理为词频矩阵后，进行 *PCA* 降维减少噪音，再采用 $K-Means$ 聚类算法与 *Birch* 算法对文本进行聚类，通过手肘法与轮廓系数法找出聚类最佳值，对两种算法的轮廓系数值进行对比，可得出 *Birch* 算法聚类较好。

4.1.1 文本处理

首先，对留言内容进行结巴分词，去除空格、符号、停用词等。在附件 3 中，将文本中的词语转换成词频矩阵，矩阵里面的元素用 a_{ij} 表示，其中 a_{ij} 表示 j 词在 i 类文本下的词频，把每个词语的 $tf-idf$ 权值都统计出来，形成 $tf-idf$ 矩阵，里面的元素 w_{ij} 表示 j 词在 i 类文本中的 $tf-idf$ 权重，从而形成词频矩阵。

4.1.2 *PCA* 降维

由于形成的词频矩阵为 4326×7189 维矩阵，因此我们对它进行 *pca* 降维，$K-L$ 变换为 *PCA* 提供了理论基础依据。*PCA* 是一种基于线性变换进行向量降维的方法，其核心思想是通过坐标旋转（即寻找新的正交基），将数据投影到使数据方差最大化的若干个坐标轴上，得到数据在新空间的表示以消除原数据空间的多重共线性，从而达到数据降维的目的。

通常，要获得低维子空间，最简单的是对原始高维空间进行线性变换，给定 d 维空间的样本集 $X=\{x_1,x_2,\cdots,x_m\}\in R^{d\times m}$，变换之后得到 n 维空间 Z 中的样本，其中 $n<<d$ 。$Z=U^TX$ 。式中：$U\in R^{d\times m}$ 为变换矩阵（即找到新的正交基），U 是由 X 的协方差的特征值最大的前 n 项所对应的特征向量构成的正交矩阵；$Z\in R^{d\times m}$ 为样本在新空间中的投影，是高维 X 降维后的低维近似。

样本集 $X=\{x_1,x_2,\cdots,x_m\},x_i\in R^d$，低维空间维数 n，PCA 算法步骤如下：

（1）样本特征去均值化（即数据的中心化）。对样本的每个特征，使用当前特征的值减去样本集中的该特征的均值 μ_j，对第 i 个样本的第 j 个特征，

$x_i^{(j)} = x_i^{(j)} - \mu_j$，其中 $\mu_j = \frac{1}{m}\sum_{k=1}^{m} x_k^{(j)}$，全部特征均值向量 $\mu = (\mu_1, \mu_2, \cdots, \mu_d), \mu \in R^d$；

（2）计算样本协方差矩阵 $\sum XX^T$；

（3）对协方差矩阵 $\sum XX^T$ 做特征值分解，取最大的前 l 个特征值所对应的特征向量 $u_1, u_2, \cdots, u_l$ 组成转换矩阵 U；

（4）生成降维后的样本集 $Z = \{z_1, z_2, \cdots, z_m\}$，原始样本集的近似；

$\hat{D} = \{Uz_i + \mu\}, i = 1,2,\cdots,m$。

通过 *PCA* 降维，把向量维度从100% 降到99%．在降维过程中可以去除无意义或者次数只出现一次的词。

4.1.3 $K-Means$ 聚类算法

对于给定的一个包含 n 个 d 维数据点的数据集 $X = \{x_1, x_2, \cdots, x_n\}$，，其中 $x_i \in R^d$，以及要生成的数据子集的数目 K，$K-Means$ 聚类算法将数据对象组织为 K 个划分 $C = \{c_k, i = 1,2,\cdots,K\}$。每个划分代表一个类 c_k，每个类 c_k 有一个类别中心 μ_i。选取欧式距离作为相似性和距离判断准则，计算该类内各点到聚类中心 μ_i 的距离平方和

$$J(c_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \qquad (1)$$

其聚类目标就是使得各类总的距离平方和 $J(C) = \sum_{k=1}^{K} J(c_k)$ 最小。

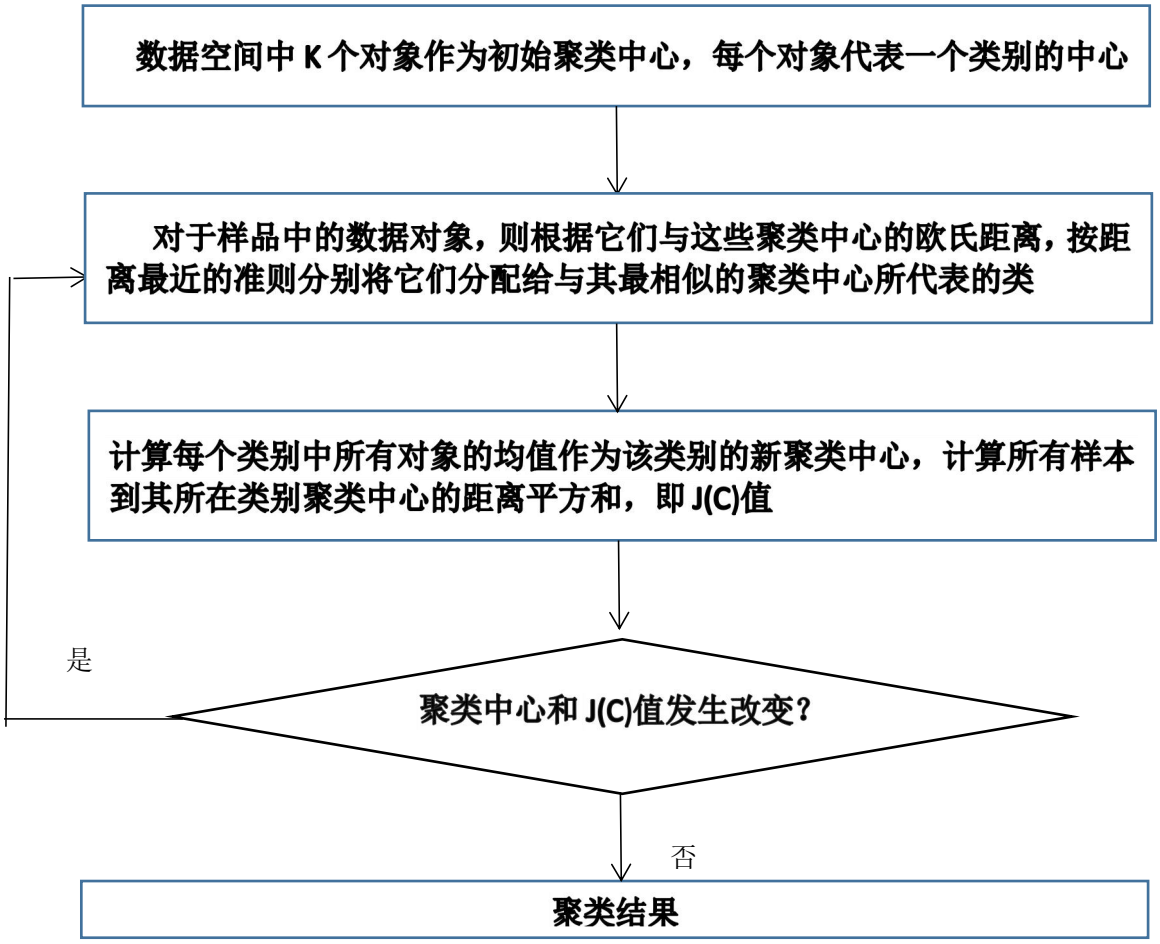
$$J(C) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|x_i - \mu_k\|^2 \qquad (2)$$

其中，$d_{ki} = \begin{cases} 1, x_i \in c_i, \\ 0, x_i \notin c_i. \end{cases}$ 显然，根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应

该取为类别 c_k 类各数据点的平均值。

$K-Means$ 聚类算法从一个初始的 K 类别划分开始，然后将各数据点指派到各个类别中，以减小总的距离平方和。因为 $K-Means$ 聚类算法中总的距离平方和随着类别个数 K 的增加而趋于减小（当 $K=n$ 时， $J(C)=0$ ）。因此，总的距离平方和能在某个确定的类别个数 K 下，取得最小值。

$K-Means$ 算法是一个反复迭代过程，目的是使聚类域中所有的样品到聚类中心距离的平方和最小，算法流程包括 4 个步骤，具体流程图如图所示



4.1.4 改进的层次聚类算法

在使用 $K-Means$ 聚类算法后，继续对分类进行改进，发现在改进之后，聚类效果更好。本文使用常见的改进的层次聚类算法，也即 *Birch* 算法， *Birch* 算

法是一种常用的层次聚类算法。是一种采用对称方法进行平衡迭代归纳的综合性层次聚类方法，该算法使用聚类特征树和聚类特征进行聚类。 *Birch* 聚类算法具有算法效率更高，聚类速度更快，并且适用于处理大规模数据集等优点，也更容易计算类簇的直径和类簇之间的距离。其中聚类特征定义为：$CF=(N,\overrightarrow{LS},SS)$,其中 N 为该类数据中的数据点数，$\overrightarrow{LS}=\sum_{i=1}^{N}\overrightarrow{X_i}$,类中数据点的线性和；$SS$ 为平方和，即 $\sum_{i=1}^{N}\overrightarrow{X_i^2}$ 。

Birch 算法工作可分为下面两个阶段：

（1） *Birch* 扫描数据库，构造一棵初始存放于内存的 *CF* 树，压缩并保留 *CF* Tree 中所有的信息；

（2） *Birch* 算法对 *CF* Tree 的叶子节点进行第二次聚类来提高精度。

注：一 棵聚类特征树 （*CF* Tree）储存了层次聚类得聚类特征 ，它是一棵高度平衡树 ，并包含两个参数：分支因子或以及一个阈值 ，每个非叶子节点最多包含 B 个项目，形式为 $[CFi,childi]$， $i=1,2,\cdots,B$， *childi* 是指向第 i 个子节点的指针 ，*CFi* 是第 i 个子节点所代表子类的 *CF*；每个叶子节点最多包含 L 个项目，每个项目是其子类的一个 *CF*，另外 ，每个叶子节点有两个指针分别指向左右两边的叶子节点 ，这些指针将 *CF* 树中所有叶子节点链接起来 ； *CF* 树中每个节点都可以看作由其个或个项目所代表的子类组成。所有叶子节点子类的半径必须小于阈值 T . *CF* 树 的大小可以通过调整 T 得到控制 ，T 值越大树越小 。

4.1.5 k 值的取定

（1）经验法：我们在进行聚类算法之前，如何确定最佳聚类数是一个关键问题，在进行程序编辑之前，我们先用人为的方法进行大致确定，也即经验法，当聚类的数量为1的时候，所有数据对象都聚成一个簇，算法没有任何意义，因此可以肯定算法中 k 值的下界为 2 。对于 k 值的上界，根据现有文献的证明和多数学者使用的经验规则，若待聚类数据总数为 n，则 k 值的范围为 $[2,\sqrt{n}]$，根据经验法，聚类的数量 k 等于对象的数量除以 2 再开平方。公式为：$k=\sqrt{n/2}$，附件 3 给出的数据为 4326，因此所求得 $k\approx 46.51$，显然，我们只能把经验法得到的数据作

为一个参考，接下来我们再结合手肘法进行确定 k 值。

（2）手肘法：手肘法是一种利用 SSE 和 k 值的关系图确认最优 k 值的方式，SSE 还可以替换为样本点到聚类中心欧式距离平均值，本文选用 SSE 。其算法思想为：数据集在 $k-means$ 算法聚类下，随着 k 的不断增大，数据被分割的更加详细，聚类中心不断增多，SSE 逐渐减小。当 k 值小于真实聚类数时，随着 k 值的增大，SSE 值的变化比较大，关系图显示两点之间的连线会比较陡峭。当 k 与真实聚类数相等时，随着 k 值的增大，SSE 值的变化较小，关系图显示两点之间的连线会变得平缓。所以 SSE 值和 k 值关系图是一个“手肘型”的折线图，“肘部”为最优的 k 值。

步骤及定义：输出：每个 k 值和对应的 SSE 值。

1. 根据 $k-means$ 进行所有 k 值的聚类。
2. 计算每个 k 值对应的误差平法和 SSE 值。
3. 利用工具绘制二维图形划出 SSE 值和 k 值对应的关系折线图。
4. 确认最优 k 值。

定义 误差平方和公式：

$$SSE=\sum_{i=1}^{k}\sum_{p\in L_i}\left\|p-q_i\right\|^2$$

公式中， p 代表第 i 个类组 L_i 中的数据对象， q_i 代表某个类组中所有数据对象的平均值， k 代表分类组个数。

由于由于 k 值范围过大，逐一增加 k 值非常浪费空间和时间，因此设定 10 为跳跃间隔，从 k 为 10 开始，依次记录 SSE 值，缩小 k 值的范围，然后在精确后的范围内逐一增加 k 值，以找到最佳聚类数。以下是 k 值逐渐增加后的手肘图。如图 4

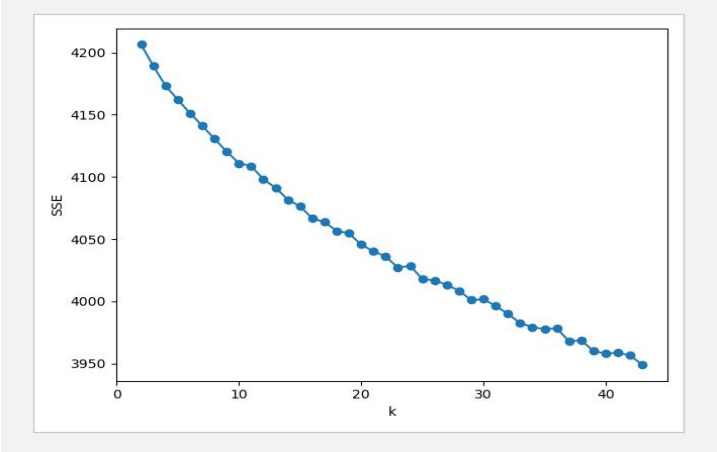


图 4

可以从图中看出k值在 0-10，10-20，20-30，30-40 的范围变化大致为 90，60，45，40 。变化不够平缓均匀，继续改进。如图 5

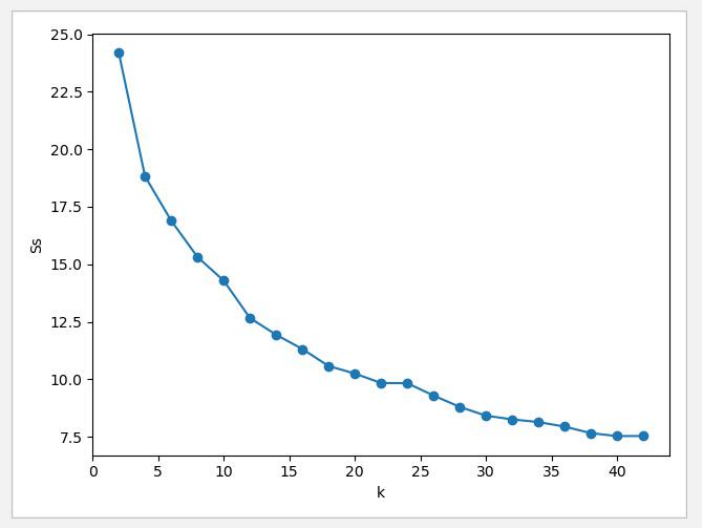


图 5

可以从图中看出k值在 0-10，10-20，20-30，30-40 的范围变化大致为 11,4,1.75,1.25，现在相对来说变化较平缓均匀，因此k值范围 10-20 之间，在 10-20 之间精确寻找k值。接下来，为了使k值更加精确，再结合轮廓系数法确定k值。

(3) 轮廓系数法:轮廓系数用于评估聚类的效果，它结合了聚类的凝聚度和分离度。轮廓系数的取值范围为$[-1,1]$，其值越接近 1，代表凝聚度和分离度越优。通过整体轮廓系数可以判断聚类效果，具体方法如下：

Step 1 计算聚类凝聚度$a(i)$，即每个样本点i与其同一簇内所有其他样本点的欧式距离的平均

Step 2 计算聚类分离度$b(i)$：选取样本点i外的一个簇b，计算i与b内所有样本点间的两两距离的平均值，遍历其他簇，K个平均值中的最小值即为聚类分离度；

Step 3 根据以上的$a(i)$和$b(i)$，确定样本点轮廓系数$s(i)$；$s(i)=\dfrac{b(i)-a(i)}{\max\{a(i),b(i)\}}$；

Step 4 计算$s(i)$的均值，即整体轮廓系数：$s=\dfrac{\sum_{i=1}^{n}s(i)}{n}$。

由整体轮廓系数量化数据聚类的紧密程度，易知$s(i)$越小，说明i与所在簇的样本点的平均距离远于最近的其他簇，即聚类效果较差；相反，若$s(i)$越大，说明聚类效果比较好。因此，可选择整体轮廓系数法最大的K值为最优的聚类个数，作为改进$k-means$聚类的最优K值选取的一种方法。

通过轮廓系数法生成的图为图 6：

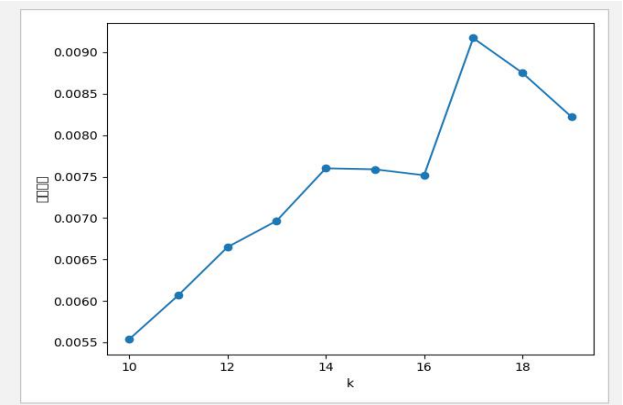


图 6

紧接着继续用手肘法对 k 值做改进，如下图 7。

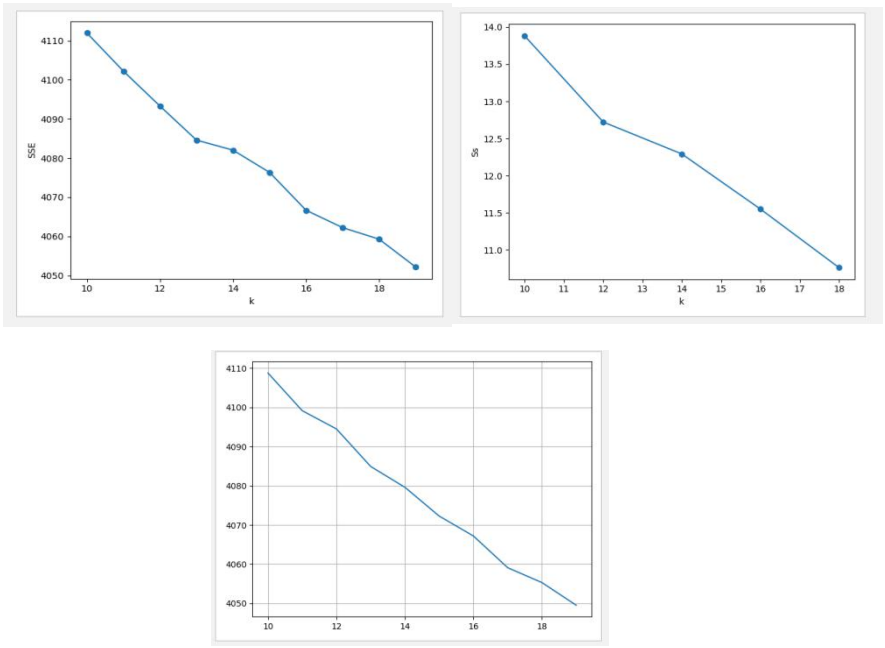


图 7

此时经过不断改进，所得的手肘图相对前几次整体均匀下降，只是在 16-18 之间，SSE 下降幅度较为明显，而在此之后又呈现平缓状态，因此我们选取 17 作为本次聚类的最佳聚类数。

4.2 热点排行

4.2.1 构建每类的词云图

对留言详情进行关键字提取，形成词云图如下图 8：

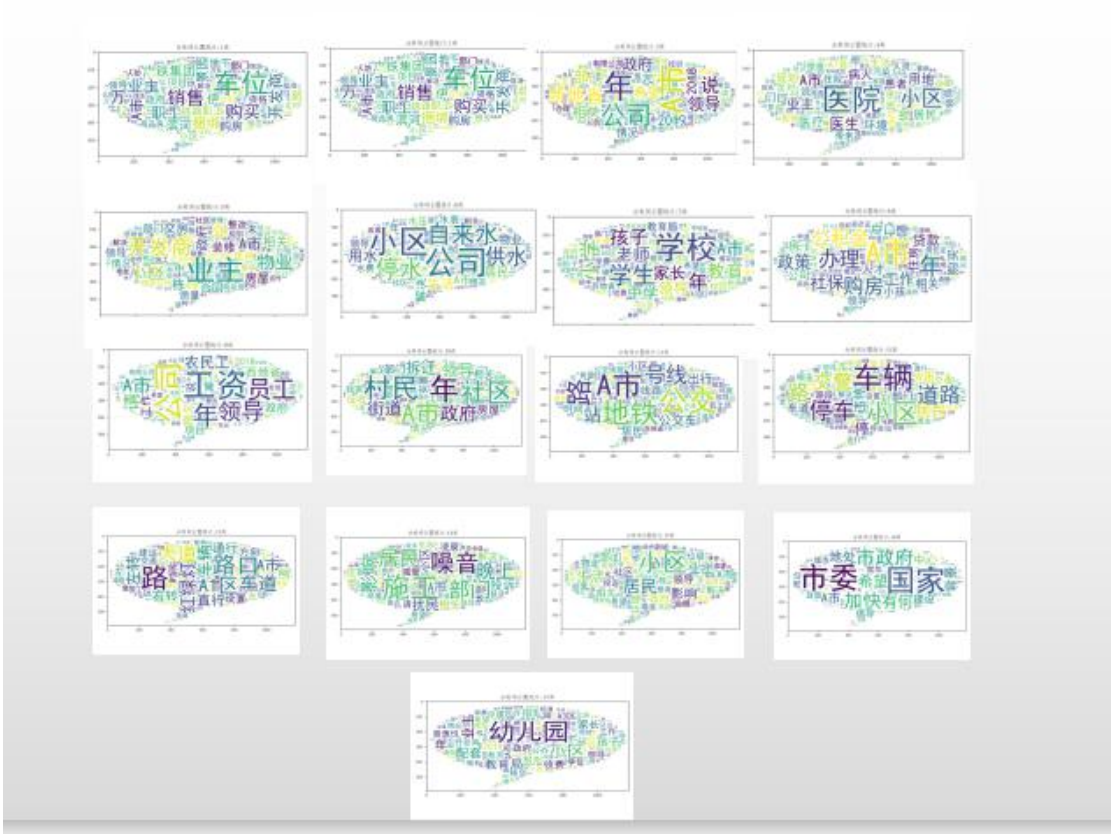


图 8　17 种词云图

4.2.2 构建评价指标

在进行热度指标排序之前，同样地对数据进行去除停用词等处理。留言文本和主题词对于热点留言发现的作用是相辅相成的，当一个热点留言出现时，与该留言密切相关的热点主题词大量涌现，与该留言密切相关的留言也大量涌现，并且这些热点主题词在留言数据中占有比较大的权重。为了能对留言热度进行比较排序，必须将影响留言热度的因素量化，使之成为能够计算的数据本文采用留言中的文本总数、最早发布时间、留言持续时间、最新发布文本量、点赞和反对总数五个因素作为热度排序依据，并将这些因素量化。

文本总数是指从首次发布时间到当前日期，有关某个留言的所有文本数量总和，数学符号记为 $N(topic)$ 。

最早发布时间是指有关某留言的文本最早发布的时间，最早发布时间越接近当前时间，表示留言越新，热度越高，日期格式可统一为“ $XXXX-XX-XX$ ”，

数学符号记为$T(first)$。

留言持续时间是指有关留言的文本中，从最初发布的日期到最后发布日期的时间长度，日期格式可统一为“$XXXX-XX-XX$”，数学符号记为$TN(topic)$。若最后发布时间为$T(last)$)，则计算公式为：

$$TN(topic)=T(last)-T(first) \qquad (4-1)$$

最新发布文本量是指当前最新一天发布的文本数量，数学符号记为$N(new)$。

点赞和反对总数是指所有有关留言的文本下的点赞数与反对数的总和。数学符号记为$N(tmp|ops)$，其中tmp表示点赞数，ops表示反对数，计算公式为：

$$N(tmp|ops)=\sum_{i=1}^{N(topic)}(tmp+ops) \qquad (4-2)$$

4.2.3 热度排序实现流程

如果仅仅只依据文本的总量或留言的点赞数和反对数，考虑涉及到的热度影响因素比较单一，这样只突出了某个因素的影响而忽略了其他因素。本文制定的热度排序策略，从多个因素对留言主题热度进行评估排序，并且能够依据自身需求设置参数，不仅全面反映留言的受关注度，还能通过调参突出某些因素的重要性。

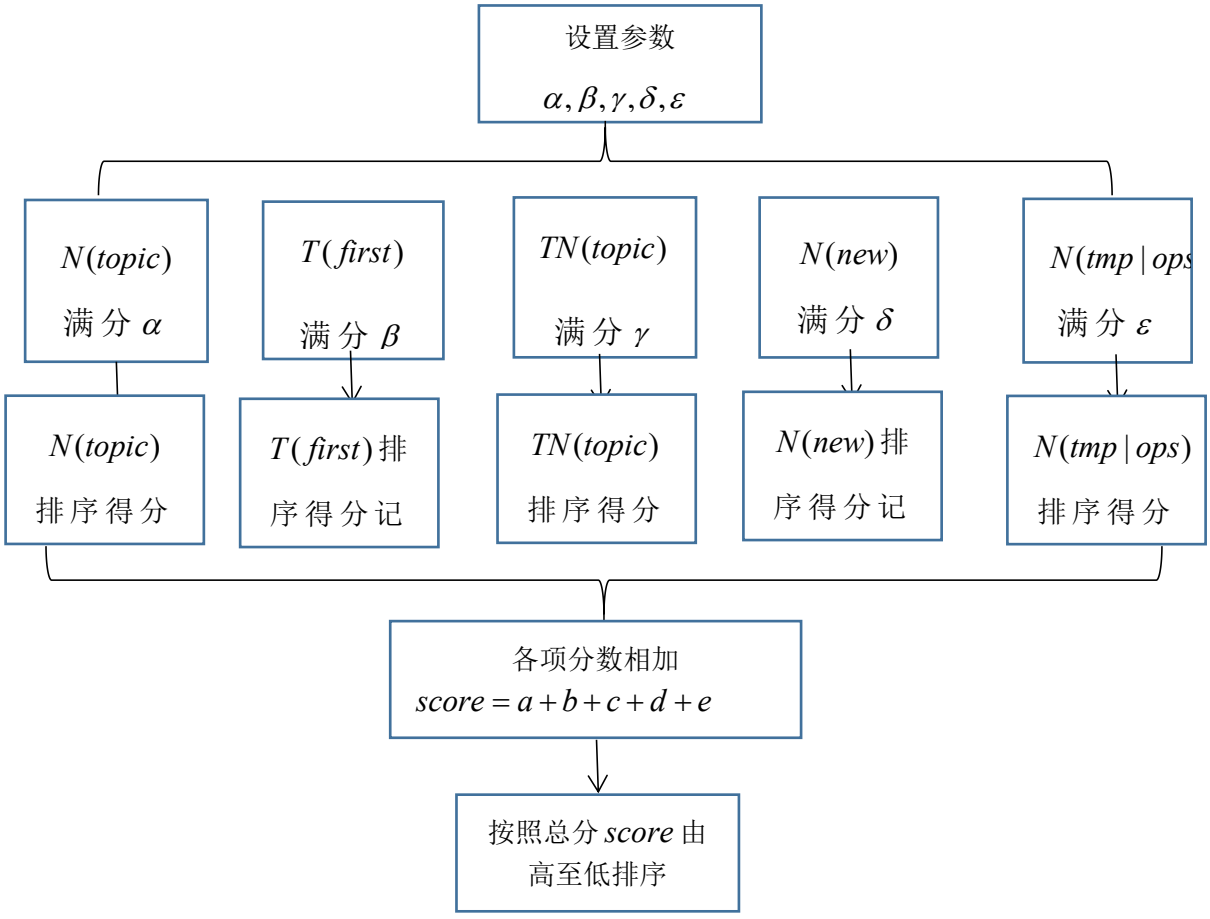
本文排序采用的策略思想是，对留言热度的总评分设置为100　分，将五个评估因素按照所设置的比例分配相应分数（每个因素20分），所有留言分别按照单个因素排序，每个因素进行归一化处理后乘上相应的分数。之后将五个因素单项分相加，总分就是留言热度排序的最终结果，并且按照最终总分进行从高到低的排序。满分分配方式如下：

$$100=\alpha+\beta+\gamma+\delta+\varepsilon \qquad (4-3)$$

总分计算方式：

$$score=a+b+c+d+e \qquad (4-4)$$

$\alpha,\beta,\gamma,\delta,\varepsilon$ 分别为五个评估因素 $N(topic)$，$T(first)$，$TN(topic)$，$N(new)$，$N(tmp|ops)$ 的单项排序满分，a,b,c,d,e 分别为一个留言在各因素的单项得分，$score$ 为留言的总得分，$N(topic)$，$TN(topic)$，$N(new)$，$N(tmp|ops)$ 按照数量由高至低自单项满分依次降低评分，$T(first)$ 按照时间由近至远自单项满分依次降低评分。流程图如下：



4.3 结果分析

$K-means$ 和改进的层次聚类算法 $birch$ 算法的比较，采取经验法对 k 值进行预估，再结合手肘法、轮廓系数法聚类最佳，再进行算法相似度参数计算，在确定高斯核函数的参数后，使用改进后 $birch$ 的聚类算法，将之前已经归一化处理的文本向量进行聚类操作通过计算他们的轮廓系数比较，发现 $birch$ 算法的结果

更大，最后采用*birch*算法聚类，最终分类结果部分如下图 9

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	225479	A0004380	A市市政建设开发有限公司违规操作	2019/7/5 1:55:26	A市市政建设开发有限公司违规操作铁广职工住宅项目，2500多名职工盼新房遥遥无期，未取得预售资	0	0
1	255507	A909195	违反自由买卖的A市伊景园滨河苑车位	2019-08-20 12:34:21	广铁集团铁路职工定向商品房伊景园滨河苑项目，广铁集团与A市政府及A市政工程有限公司要求职工限	0	0
1	258386	A909185	A市伊景园滨河苑欺压百姓	2019-08-28 00:00:00	A市伊景园·滨河苑强制要求购房者捆绑购买其根本不需要的车位，不买就取消购房资格，2千多购房者	0	0
1	199190	A0009508	关于A市武广新城违法捆绑销售车位	2019/8/1 22:32:26	武广新城为铁广集团的定向商品房，在未取得预售资格强行逼迫职工缴纳18.5万购房款且不签正式购房	0	0
1	232892	A0001533	A市万科魅力之城开发商未通知业主	2019/1/8 9:54:00	您好！我是57栋业主，2018年12月25日万科魅力之城4.1期车位开盘，开发商未通知业主就进行车位开	0	1
1	289473	A0001034	反对滨河苑房子和车位捆绑销售	2019-08-22 00:00:00	现有伊景园滨河苑在对广铁集团员工定向销售的过程中，将房子和车位打包销售，要求广铁集团职工在	0	0
1	246785	A0001592	不需要车位，不想白扔12万	2019-08-15 15:32:28	我们是还没住进武广新城片区的伊景园的业主，在未取得预售资格强行逼迫职工缴纳18.5万购房款且不签	0	0
1	205982	A909168	坚决反对伊景园滨河苑强制捆绑销售	2019-08-03 10:03:10	我坚决反对伊景园滨河苑捆绑销售车位！原本广铁集团与市政府和开发商协议，可以给铁路职工优惠定	0	2
1	285897	A909191	武广新城伊景园滨河苑违法捆绑销售	2019-08-01 20:06:52	我们是广铁集团铁路职工,武广新城片区的伊景园滨河苑楼盘是我们的定向限价商品房,在现在楼盘还未	0	0
1	218739	A909184	A市伊景园·滨河苑欺诈消费者	2019-08-24 00:00:00	A市伊景园滨河苑强行捆绑车位销售，诱骗购房者交付车位定金，还不写入协议，不给签合同。请有关	0	0
1	289588	A909183	投诉A市伊景园·滨河苑开发商	2019-08-21 21:00:21	A市伊景园·滨河苑售房时捆绑销售车位，恶意抬高房价，行为极其恶劣，严重影响广大消费者为国家贡	0	0
1	250236	A0001108	A5区保障住房小区地下停车位问题何	2019/1/16 20:02:55	尊敬的书记，您好。我是一名A5区黎锦苑小区的业主。2015年7月9日响应政策我第一时间缴纳了黎锦苑	0	0
1	218709	A0001066	A市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	伊景园滨河苑作为广铁集团定向商品房，取得预售证，不与广铁购房者签订正规购房合同，强制收取18	0	1
1	251844	A909167	投诉伊景园滨河苑项目违法捆绑车位	2019-08-20 13:34:12	投诉广铁集团强制要求职工捆绑购买12万的车位费，不购买就取消购房资格！诸大的广铁集团，竟不顾	0	1
1	233543	A0007331	A7县深业睿城小区部分业主私自霸占	2019/9/22 23:53:31	沈书记您好！我是深业睿城小区业主，最近小区因为地下停车位的事，闹得人心不安，部分业主认为地	1	2
1	254499	A0004000	购买A市龙湖楚风原著停车位一年多	2019/2/6 6:30:16	本人购买A市龙湖楚风原著（A市A6区银星路555号）停车位一年多了没有给我开发票，已催促几次没有	0	0
1	222209	A0001717	A市伊景园滨河苑定向限价商品房项目	2019-08-28 10:06:03	广铁集团与伊景园滨河苑开发商沆瀣一气，严重侵害了购房职工的合法权益，在不签购房合同的情况下	0	0
1	279941	A909177	广铁集团职工商品房真然捆绑销售	2019-08-28 09:30:20	领导好！A市广铁集团为职工提供定向商品房伊景园滨河苑的事帮帮我们吧，当初以为是集团为职工谋	0	0
1	228317	A0001113	A市高新区B4区涉外景园的人防车位	2020/1/6 20:26:46	您好，我是A市高新区B4区涉外景园C小区的居民，因为生活需求，需要停车位，但是小区的车位是人防	0	0
1	283879	A0004475	A市伊景园滨河苑项目捆绑销售车位	2019/7/18 20:27:40	关于铁广集团铁路职工定向商品房伊景园滨河苑项目是由A市政府办牵头为铁路职工谋福利好事，但现	0	0
1	246407	A0009959	举报广铁集团在伊景园滨河苑项目非	2019-09-01 14:20:22	我要举报广铁集团在伊景园滨河苑项目非法绑定车位出售牟取非法利益。伊景园滨河苑项目原本是为了	0	0
1	205277	A909234	伊景园滨河苑捆绑车位销售合法吗？	2019-08-14 09:28:31	广铁集团强制要求职工购买伊景园滨河苑楼盘时捆绑购买12万一个的车位，不买车位，不能购买房子！	0	1
1	286304	A909196	无视职工意愿、职工权益的A市伊景园	2019-08-23 10:23:23	广铁集团与A市政府及A市政工程有限公司要求职工购买房子的同时一对一购买所谓按成本价销售的12万	0	0
1	195511	A909237	车位捆绑违规销售	2019-08-16 14:20:26	对于伊景园滨河苑商品房，A市广铁集团违规捆绑车位销售至今，买房必须买车位我们反映多次一直没	0	0
1	258295	A0004893	A市富基世纪公园捆绑销售变成强买	2019/12/31 21:00:00	尊敬的领导：您好！我是富基世纪公园四期的业主，我是2018年3月份买的房子，当时说面积大的房子	0	0
1	209506	A909179	A市武广新城抗客户购房金额并且捆	2019-08-02 16:36:23	您好！由A市广铁集团发起的定向商品房伊景园滨河苑项目存在坑害客户的现象！我们作为广铁集团的	0	0
1	276016	A909181	车位属于业主所有，不应该被捆绑销	2019-08-06 00:00:00	尊敬的胡书记，您好！我叫陈玉春，身份证号**********，现实名投诉广铁集团铁路职工定向商品房	0	2
1	191001	A909171	A市伊景园滨河苑购买要求购房同时	2019-08-16 09:21:33	商品房伊景园滨河苑问题由A市政府办牵头为广铁集团铁路职工定向销售的楼盘，作为集团的一名退	1	12

Sheet1

图 9 聚类结果

热点排序结果如下表表 1

热度排名	问题id	热点指数	时间范围	地点/人群	问题描述
1	3	78.815	2018/10/2	A市A7县西地省公司	希望各相关部有效解决问题
2	15	50.7075	2019/1/2/3	A市小区	小区街道油烟垃圾影响居民
3	5	45.7392	2019/1/1/3	A市开发商	开发商交房质量不合格
4	2	34.6962	2019/1/2/3	A市城市居民	城市规划与建设
5	7	33.853	2017/6/8/3	A市A3区西地省学生家	希望教育局重视学校教育问题

表 1 热点排序结果

5 回复留言质量的综合评价方案

5.1 构建指标

先把文本特征分类分为五个特征，分别为答复意见的相关性、准确性、可解释性、完整性、正规性。根据这五个特征来来给出一套评价方案。

（1）相关性：指回复问题与提问之间的关联度，回复问题的关键词与提问关键词的重叠的数量，数学符号记为R.

（2）完整性：问题与回复长度比率，数学符号记为W.

（3）正规性：回复问题中非停用词的数量，数学符号记为N.

（4）可解释性：回复问题与提问之间的时间间隔，数学符号记为E.

（5）准确性：回复问题的长度，数学符号记为A.

提取出每条问题的回复长度，对附件 4 相关性的处理，提取出回复与提问进行分词且去除停用词后，对每个问题重叠词的数量进行统计，再对所有值进行归一化乘上相应的权重。由于有个别数据数值较大，则设置上限 1000 词，如超过 1000 词则相应权重为满分，其余数据主要分布在 0 到 1000，则每个词除以 1000 再乘上权重，以确保大部分数据得分。

5.1.1 余弦相似度

对答复意见进行相关性打分时，需要进行余弦相似度的计算，所谓余弦相似度，就是通过两个向量之间的夹角来衡量向量相似性，余弦相似度计算公式如下：

$$\cos\theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} = \frac{\sum_{i=1}^{n} A_i B_I}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

并且相似度在$(0,1)$之间，令$k = i \times 20$，从而求出各个答复意见关于这个相关性指标的分数，最后再把每条答复意见关于各个指标的分数相加，即为总共得分，从而给出评价等级。

在确定答复意见质量指标后，需要计算出每一条答复意见关于这 6 个指标的分数，先设置总分为 100 分，每个指标的总分如下定义，然后计算每条答复意见在这六个指标集中所占的分数。

注： 相关性能很好的反应答复意见是否更好的回答留言问题，因此设相关性占 30 分。通过前人的研究，准确性和完整性对于鉴别高质量答案是比较显著的，因此都设为 25 分。正规性是关于非停用词的数量，并不会很大的影响留言意见的质量，因此设为 5 分。可解释性主要是回复时间和问题时间的间隔，占有一定的重要性，因此设为 15 分。

5.2 定义评价等级

在本文中，根据生活经验及相关文献查阅，定义答复意见评价等级为优，良，一般，合格，差。见表 2

评价等级	优	良	一般	合格	差
分数	≥90	≥80	≥70	60≤	<60

表 3 评价等级表

5.3 验证评价结果的准确性

在给出每条留言意见的评价等级后，本文对其进行了验证，以便来确保评价方案的可行性，其采用验证方法的思路是将留言回复所得到的分数按常规比例划分等级，再与平常等级区域进行对比，得到本套评价方案可以作为回复质量的评定，其部分评价结果和验证结果如下图 10

	A	B	C	D	E	F	G	H	I	J
1	留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复得分	答复质量	评价
2	18413	UU00883	反对在A7	2015/12/3	经调查	网友“UU0	2016/1/7	98	优	优
3	9128	UU00842	关于A市和	2017/3/22	首先，	网友“UU0	2017/4/14	96	优	优
4	27418	UU00840	反映A5区	2018/10/1	尊敬的各	“UU00840	2018/11/2	95	优	优
5	4347	UU00819	A5区佳园	2018/10/2	尊敬的书	网友“UU0	2018/11/5	95	优	优
6	88150	UU00850	投诉J8县	2019/5/23	J市人民政	一、房屋	2019/7/3	94	优	优
7	19979	UU00814	关于水电	2014/2/21	2013	网友“UU0	2014/3/5	94	优	优
8	119181	UU00818	请求K7县	2018/11/2	幸福路连	您好！收	2018/11/3	94	优	优
9	53872	UU00817	强烈呼吁	2012/7/11	由于近	近期，	2012/7/25	94	优	优
10	92985	UU00818	反映J11市	2019/5/29	在全国和	UU00818	2019/6/6	94	优	优
11	8954	UU00834	反映A1区	2017/4/17	尊敬的	网友“UU0	2017/5/12	94	优	优
12	9744	UU00813	关于对A市	2016/10/2	小区	网友“UU0	2016/11/2	94	优	优
13	110813	UU00838	K4县金润	2017/12/2	尊敬的	2	2017/12/2	94	优	优
14	9199	UU00814	投诉A5区	2017/3/14	请求政	网友“UU0	2017/3/24	93	优	优
15	117369	UU00818	举报K市永	2019/6/25	被举报人：	“UU00818	2019/7/15	93	优	优
16	133439	UU00841	盼望L6县	2018/10/2	L6县委书	网友：您	2018/11/1	93	优	优
17	50409	A000492	关于独生	2014/5/8	赵主任	“A000492	2014/5/20	93	优	优
18	83000	UU00822	G市万达B	2018/6/13	尊敬的周	网友：您	2018/7/5	93	优	优
19	10335	UU00821	湖楚中医	2016/5/2	尊敬的	网友“UU0	2016/5/16	93	优	优
20	84572	UU00898	反映G1区	2019/12/1	G市G1区	您好！函	2020/1/2	93	优	优
21	5607	UU00877	关于打造	2018/6/21	尊敬的胡	您好！您	2018/7/6	93	优	优

Sheet1

图 10 评价结果和验证结果

6 根据结果分析，提出改进建议

随着技术的发展，文本的处理在不断的进步，本文做出的解决方案更多的是解决眼前面临的需要，随着时间推移与探索，还有很多能够提高的地方。

（1）在对文本分词时，采用的是最常规的 *jieba* 分词法，会导致一些网络用语被错误划分，因此可以在词料库添加一些网络用语，也可以采用更好的分词方法，

使分词更加准确。

（2）对于文本分类分类问题，主要采用的还是神经网络、支持向量机、高斯贝叶斯这些常规的方法进行分类，希望在之后能寻找到更好的分类方法，提高分类的准确率。

（3）对于 $K-means$ 聚类方法，对于最佳聚类的取值，还是采用的是传统的手肘法与轮廓系数法来确定，如能找到一些更好确定最佳聚类的方法，会有助于提高 $K-means$ 算法在文本聚类里面的准确度。

（4）地点人群与热点问题提取方面，本文通过关键词来确定，其中关键词在精度上还有待改进，同时希望以后可以完成关键词自动生成热点问题的工作。

7 结论

在网络高度发达的当代，信息传递的速度达到了空前的高度，正因如此，也使市民能够对自己生活中所遇见的问题进行及时反映，由于反映的问题众多，大量的文本鱼龙混杂，政府部门希望能够将问题进行分类，与此同时将热点问题进行排行，以便于及时了解市民所遇见的问题，得到良好的解决。针对以上需求，本文通过对文本的处理，基于自然语言处理中的聚类技术，提取出留言问题，并根据热度影响因素对话题排序，确保让相关部门关注民生问题。本文主要工作有以下几点：

（1）对文本分类准确率采取了多种方法进行比较。对文本进行分类后，采用了神经网络、支持向量机、高斯贝叶斯等多种算法，使用数据的百分之八十来训练模型，百分之二十进行测试，以精确率、召回率、F 值以及 CA 作为效果评价标准，根据算法作对比，验证得出神经网络的算法性能优良。

（2）改进 $K-means$ 聚类算法，即 $birch$ 聚类算法。对于 $K-means$ 聚类算法，在寻找最佳聚类的过程中，首先采用经验法缩小 k 值得范围，再结合手肘法与轮廓系数法确定 k 值，最后得出最佳聚类。改进后 $K-means$ 聚类算法，研究层次聚类算法思想并将其与 $K-means$ 算法思想相结合，自顶而下进行 $K-means$ 划分，防止聚类结果产生空簇。鉴于 $K-means$ 算法对初始质心的依赖性，摒弃随机选取初始质心的方法，而是根据数据范围规律性的确定初始质心，提高聚类效果。

使用高斯核函数作为$K-means$相似度的计算方法，将数据映射到高维空间中实现线性可分，并且降低噪声数据的干扰。

（3）根据留言问题的定义使用$TF-IDF$技术从文本中提取出关键词，确定话题并使其能够最大程度地代表文本核心内容。制定留言问题热度排序策略，从问题的文本总量、文本的最近发布时间、话题持续时间、最新的文本增长量、点赞与反对总数等影响因素进行热度评估和排序，改善了传统话题热度排序中仅仅使用话题点击量和话题更新时间作为评估因素的单一性。此外对热度影响因素赋予权重，可根据实际要求对权值参数进行调整，增加了热度排序策略的灵活性。

（4）对留言回复进行等级划分。从留言的相关性、准确性、可解释性、完整性、正规性等方面得出评定留言质量的评价方案，将留言回复所得到的分数按常规比例划分等级，再与标准等级区域进行对比，得到本套评价方案可以作为回复质量的评定。

参考文献

[1] 张德成，王杨，赵传信，等.基于贝叶斯决策的极短文本分类模型[J].重庆科技学院学报（自然科学版），2018，20（4）

[2] 牛强，王志晓，陈岱，等.基于支持向量机的Web文本分类方法[J].中国矿业大学.2006，23（9）

[3] 曲凯扬.基于支持向量机的文本分类研究[D].河南师范大学.2016，5

[4] 李兵，何华.基于$KPCA-GaussianNB$的电子商务信用风险分类[D].河北工业大学.2019

[5] 于韬，王洪岩.基于$TF-IDF$算法的文本信息提取[J].辽宁科技学院.2018

[6] 王千，王成，冯振元，等.$k-means$聚类算法研究综述[J].电子设计工程.2012，20（7）

[7] 孙平安，王备战.机器学习中的PCA降维方法研究及其应用[J].湖南工业大学学报，2019，33（1）：73-78

[8] 张国锋.在文章聚类中话题热度排序的研究与实现[D].东华大学.2019

[9] 成娅辉，张英杰.聚类算法在电信客户细分中应用效果的对比研究[J].邵阳学院学报（自然科学版），2009，6（4）

[10] 吴广建，章剑林，袁丁.基于$K-means$的手肘法自动获取K值方法研究[D].杭州师范大学，2019，40（5）:167-170

[11] 苗晴，赖承栋.基于$K-means$聚类方法对广东省各市经济发展评价的研究[J].江西电力职业技术学院学报，2018，31（11）

[12] 顾海艳，王权.基于随机森林算法的吸毒人员甄别模型研究[J].南京师大学报（自然科学版），2019，42（2）

[13] 丁一，付弦.基于核心树的增量聚类算法研究[J].湖北师范学院学报（自然科学版），2011，31（2）

[14] 湖泽.在线问诊服务回答质量评价方法研究[D].哈尔滨工业大学，2019