

“智慧政务”中的文本挖掘应用

摘要

以自然语言处理技术为基础的智慧政务系统可以大大提高政府工作效率，提高管理水平。本文采用自然语言处理和文本挖掘的方法建立模型，解决“智慧政务”中的文本分类、热度评价以及答复质量评价问题。

针对问题一，需要根据附件 1 的划分体系利用附件 2 的数据建立关于留言内容的一级标签分类模型。我们在必要的的数据预处理基础上采用三种向量化的算法进行初始特征的提取，同时比较逻辑回归、支持向量机等多个分类器分类性能。接着通过潜在语义分析（LSA）等特征选择方法确定最终的特征为 LSA-TF-IDF，利用遗传算法（GA）对支持向量机（SVM）模型主要参数进行优化改进。实验显示，基于特征选择和遗传算法的改进的 SVM 模型一定程度上提高了预测精度，模型的 F-Score 达到 0.896。

针对问题二，需要根据附件 3 将某一时间段内反映特征地点和人群的问题进行分类，定义热度评价指标提取出排名前五的热点问题。我们基于命名实体识别（NER）结合预处理后的重组的数据，采用层次聚类的方法进行求解，得到了 2317 类反映特定地点和人群的分类结果。建议“用户特征影响力”、“内容特征影响力”以及“传播特征热度”三个一级指标，分别对一级指标建立各自的二级指标。接着利用因子分析验证指标的正确性并且进行调整，结果显示提取的三个公共因子和我们调整的三个维度的指标刚好吻合，可以较好的反映热点问题热度的总体信息。

针对问题三，需要根据附件 4 建议关于答复意见质量的评价模型。我们首先对文本数据进行数据重组，接着定义包括“相关性”、“完整性”、“可解释性”、“时效性”、“探究性”在内的五个评价指标，其次建立基于指标评价值的客观赋权方法变异系数法这一评价模型，通过求解得到五个指标权重分别为 0.28，-0.083，0.411，0.198，0.188，根据权重可以判断“可解释性”指标对结果影响较大。最终，经过计算可以得到各个答复意见所对应的评价指标。

关键词：特征提取、GA、SVM、NER、层次聚类、因子分析、变异系数法。

Abstract

Intelligent government system based on natural language processing technology can greatly improve the efficiency of government work and management level. This paper uses natural language processing and text mining to build a model and solve the problems of text classification, heat evaluation and response quality evaluation in "intelligent government affairs".

For the first problem, it is necessary to establish a first-level label classification model for message content based on the classification system in appendix 1 and the data in appendix 2. On the basis of necessary data preprocessing, we used three vectorization algorithms to extract the initial features and compared the classification performance of multiple classifiers such as logistic regression and support vector machines. Then, through feature selection methods such as latent semantic analysis (LSA), the final feature is determined to be lsa-tf-idf, and the main parameters of support vector machine (SVM) model are optimized and improved by genetic algorithm (GA). The experiment shows that the improved SVM model based on feature selection and genetic algorithm improves the prediction accuracy to a certain extent, and the f-score of the model reaches 0.896.

For the second problem, it is necessary to classify the problems reflecting characteristic locations and people in a certain period of time according to annex 3, define the heat evaluation index and extract the top five hot issues. Based on the named entity recognition (NER) and the reconstructed data after preprocessing, we used hierarchical clustering method to solve the problem and obtained the classification results of 2317 categories reflecting specific places and populations. It is suggested that the three first-level indicators, namely "user characteristic influence", "content characteristic influence" and "spread characteristic heat", should establish their own second-level indicators for the first-level indicators respectively. Then, factor analysis was used to verify the correctness of the indexes and adjust them. The results showed that the three common factors extracted were just in line with the indexes of the three dimensions we adjusted, which could better reflect the overall information of the heat of hot issues.

For the third problem , an evaluation model for the quality of response comments is suggested in annex 4. We first to data reorganization, the text data and then definition includes "relevance" and "integrity", "can be interpreted", "timely", "explore", five evaluation indexes, and secondly establish weight method based on the index value of variation coefficient method, the evaluation model is obtained by solving the five index weight were 0.28, 0.083, 0.411, 0.198, 0.188, according to the weight can be judged "interpretability" data greatly influenced the results. Finally, the evaluation indexes corresponding to each response can be obtained through calculation.

Key Words: feature extraction, GA, SVM, NER, hierarchical clustering, factor analysis, Coefficient of variation method

“智慧政务”中的文本挖掘应用	1
摘要	1
Abstract	2
一 简介	5
1.1 挖掘意义	5
1.2 问题的提出	5
二 符号说明	5
三 问题分析	5
3.1 问题一的分析	5
3.2 问题二的分析	6
3.3 问题三的分析	6
四 文本特征向量化	6
4.1 向量化流程	6
4.2 数据预处理	6
4.3 词云图及标签占比展示	8
4.4 初始特征提取	9
4.4.1 TF-IDF	9
4.4.2 Hash	10
4.4.3 Doc2vec	10
4.4.4 结果分析	11
4.5 特征选择	12
4.5.1 LSA 特征降维	13
4.5.2 Poly 多项式构造	14
4.5.3 特征拼接	15
4.5.4 最终特征选择	15
五 留言内容一级标签分类模型	16
5.1 分类流程	16
5.2 分类器模型及参数选择	16
5.2.1 K 近邻模型	16
5.2.2 逻辑回归模型	17
5.2.3 支持向量机模型	18
5.3 遗传算法改进的支持向量机	18
5.4 改进模型训练与验证	20
六 热点问题挖掘模型	20
6.1 挖掘流程	20
6.2 中文命名实体识别	21
6.3 层次聚类	22
6.3.1 算法介绍	22
6.3.2 算法求解	23
6.4 热度评价指标	24
6.4.1 指标构建	24
6.4.2 指标提取	24
6.4.3 指标检验	25
6.5 因子分析热度评价模型	26

	6.5.1	KMO 测度和巴特利球体检验	26
	6.5.2	热度指标公式提取	26
	6.5.3	评价结果	28
七		相关部门答复意见质量评价模型	28
	7.1	评价总体流程	28
	7.2	答复意见质量指标	29
	7.2.1	指标构建	29
	7.2.2	指标提取	31
	7.2.3	指标检验	32
	7.3	变异系数法答复意见质量评价模型	33
	7.4	给相关部门的建议	35
八		总结与模型评价	35
	8.1	问题一	35
	8.2	问题二	35
	8.3	问题三	36
九		未来的工作	36
十		参考文献	36
十一		附录	37

一 简介

1.1 挖掘意义

政府了解民意、汇聚民智、凝聚民气的重要手段包括微信、微博、市长信箱等网络问政平台，但是依靠人工进行留言划分和热点整理的相关部门由于文本数据量的上升迎来工作上更大的挑战。

所以，在大数据、云计算等技术日新月异的当下，可以设计以自然语言处理技术为基础的智慧政务系统推动政府管理水平和施政效率的提高。

1.2 问题的提出

问题一：参考附件 1 给出的标签分类体系，根据附件 2 的数据建立关于群众留言内容的一级标签分类模型，并使用 F-Score 进行评价。

问题二：及时发现热点问题，根据附件 3 将某一时间段内反映特定地点和人群问题的留言进行分类，同时定义合理的热度评价指标对热点问题给出评价结果，进而提出前五个热度问题及其对应的留言信息。

问题三：提取出附件 4 答复意见的相关性、完整性、可解释性等内容，定义合理的评价方案，对相关部门的答复意见进行质量评价。

二 符号说明

符号	符号意义
A	文本特征向量矩阵
D_i	附件 i 数据
C	惩罚因子
γ	核参数
T	参差聚类阈值
X_i	第 i 个热度指标
F_i	第 i 个公共因子
F	热度指数
ω_i	第 i 个质量评价指标

注：部分符号在下文具体说明

三 问题分析

3.1 问题一的分析

通过对赛题的分析，本文可以判定这是一个分类问题，需要代替人工的方式按照一定的划分体系对留言进行分类。而解决这个问题，需要用到文本向量化以及分类模型的理论 and 知识。本文进行数据预处理之后采用 TF-IDF、Doc2vec、Hash 算法进行文本向量化，进而可以利用向量化后的数据采用支持向量机等分类模型进行求解。

3.2 问题二的分析

问题二是热点挖掘问题，需要将地点和人群进行归类，定义热度评价指标，给出评价结果。而解决这个问题，和问题一样需要用到文本向量的方法。除此之外，我们需要通过数据预处理后进行中文命名实体识别，再对向量化后的数据利用无监督的聚类方法进行分类。接着，可以根据建立的热度评价指标得出热点问题。

3.3 问题三的分析

问题三是答复不同角度给出答复一个评价方案，直接的想法是答是所问，即答复得关键词和问题的关键词要高度相似，并考虑答复时间等因素，提出相关性、完整性、可解释性，时效性，探究性五个指标，利用提取出的指标建立一个较为完善的评价方案。

四 文本特征向量化

4.1 向量化流程

在文本分类中，本文首先需要将文本数据进行预处理。对于中文文本，需要去除标记信息及标点符号，接着可以进行分词得到相对纯净的文本，最后经过去停用词、词长过滤后进行特征提取与选择转化为 N 维的空间向量形式。总体流程图如下所示：

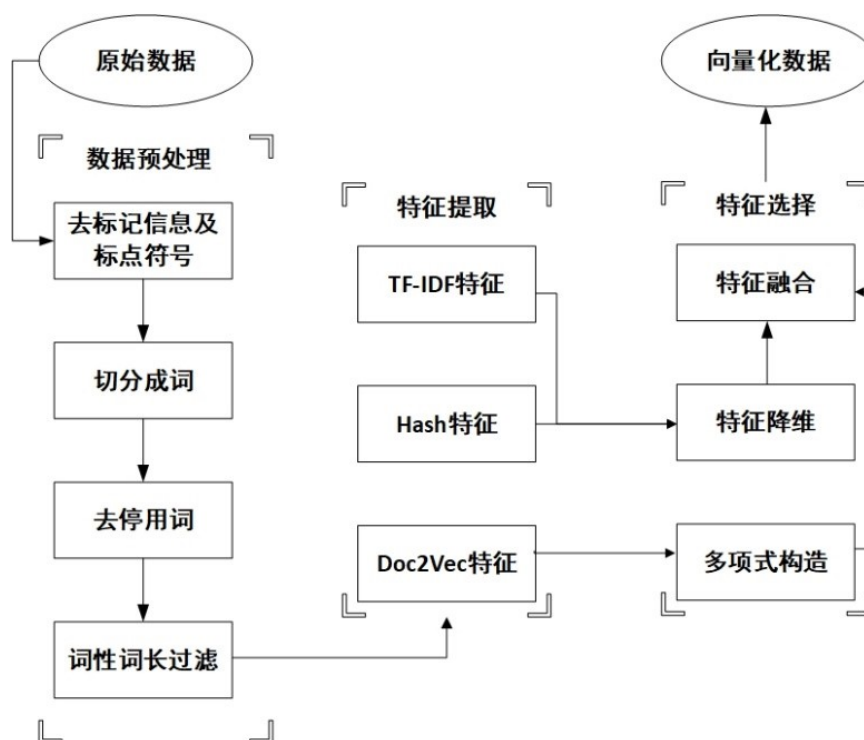


图 5.1 向量化总体流程图

4.2 数据预处理

在进行数据挖掘前，首先需要对数据进行预处理。在真实情况下，数据在格式和规范上存在较大的差异，含有大量的噪声。对这些噪声数据进行预处理，将它们变为符合规范的格式。针对本中文文本数据预处理主要包括切分成词、去停用词、向量化。

Step1: 去标记信息及标点符号

从切分结果可以看出存在大量的换行符、制表符以及标点符号等等，由于这些数据无法表示出有效信息，将大大影响分类识别效果，为了确保分类器的准确率和效率，采用正则表达式过滤这些噪声数据^[1]

Step2: 切分成词

词代表最小的语义单位，进行文本挖掘首先需要将句子划分成词。对于中文文本分词，目前比较主流的工具具有以下几种：1) jieba 分词；2) 清华中文词法分析工具包 (THULAC)；3) 雪自然语言处理库；4) 中国科学院汉语分词系统 (NLPIR) 等本文选用 Jieba 分词器。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，得到所有切分可能，根据切分为止，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，也就得到了最终的切分方式。其中，分词有三种模式，分别是精确模式、全模式、搜索引擎模式。本文采用精确模式。

Step3: 去停用词

本文首先采取去掉文本数据的常用词，比如日常使用的自然语言中存在的大量的功能词，包括连接词、语气词、副词、介词、数字符号、标签符号以及使用频率很高的单子等。这些功能词没有具体的实际意义，为了提高分类性能和处理效率，过滤这些词语得到较纯的文本。通过比较百度替换词表、哈尔滨工业大学停用词表以及四川大学机器智能实验室替换词表^[3]，选用哈尔滨工业大学停用词表作为使用的停用词表。

从分词结果可以看出由于所给数据存在分隔符、换行符、制表符等，分类结果存在“\n”，“\t”，“\u3000”等分词结果，这些可以通过 python 中的 replace 函数和 strip 函数给予删除。通过去除停用词可以得到较为科学的分词结果

Step4: 词性词长过滤

研究表明，一个文本的关键词的词长一般大于 2，所以可以将词长小于 2 的词语过滤掉。虽然关键词词长越长，包含的信息越多，但是关键词词长一般也不超过 6，因此同样将词长大于 6 的词语过滤^[4]。同时，将分好的词语标注词性，标注结果为介词、连词、助词、拟声词等的词语由于其无法表征有效信息，同时增加分类工作量，所以将这些词语过滤。^[5]

(Step5:) 数据重组

为了便于后面词向量构建和模型的训练，有选择性的将预处理后的文本数据重新组成一句完整的话。

为了直观的表现数据预处理的实际效果，截取原始数据部分文本，以“用户编号为 24 的留言为例，预处理前后的效果比对如表 5.1 所示

表 5.1 预处理前后对比

留言主题		留言详情	
处理前	处理后	处理前	处理后
A 市西湖建筑集团占道施工有安全隐患	西湖 建筑 集团 施工 安全隐患	<p>A3 区大道西行便道，未管路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。</p>	<p>大道 西行 便道 未管路口 加油站 路段 人行道 包括 路灯 西湖建筑 集团 燕子 安置项目 施工 围墙 每天 尤其 上下班 期间 路上 人流 车流 安全隐患 非常 强烈 请求 文明城市 尽快 整改 文明 路段</p>

通过对比前后结果,可以看出分词结果较令人满意,同时也可以粗略的看书留言主题内容可能包含于留言详情,这也为特征的提取做下准备。

4.3 词云图及标签占比展示

在进行分类任务之前，本文对处理后的数据按照一级分类的不同进行词云图的绘制。以城乡建设为例，可以得到如图 5.3、图 5.4 所示的词云图（附件 2 绘制的所有词云图见附录）



图 5.3 城乡建设留言主题词云图



图 5.4 城乡建设留言详情主题词云图

从绘制的词云图可以看出“留言主题”和“留言详情”存在重复的内容，同时对于城乡建设的留言内容主要针对小区、业主、规划、问题等。

对附件 2 数据一级标签进行柱状图绘制，结果如图 5.5 所示，城乡建设以及劳动和社会保障的一级标签内容内容最多，交通运输最少。

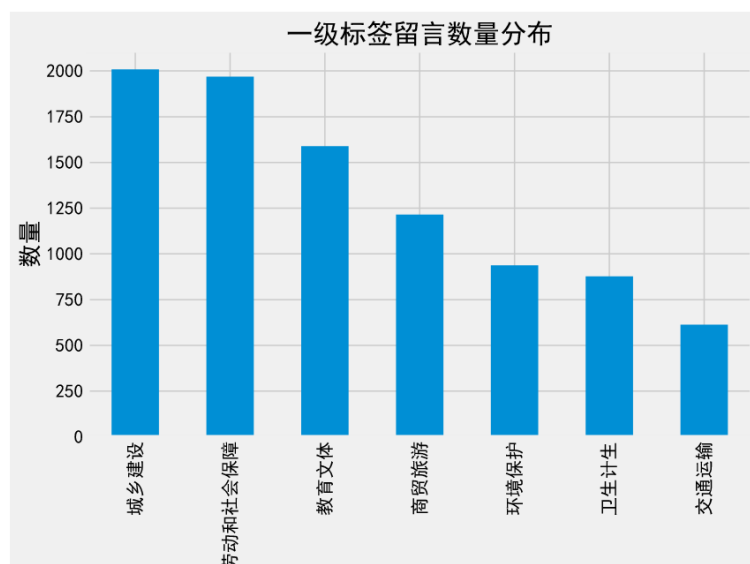


图 5.5 标签占比图

4.4 初始特征提取

4.4.1 TF-IDF

子文本中，某些单词会经常出现（例如“我”，“他”，“是”），这些词语几乎无法表示出文档内容的有效信息，直接提供给分类器将影响权重。为了将计数特征重新加权为适合分类器使用的情况，可以采用 TF-IDF 变换。

词频 TF (Term Frequency) 是词语在文档中出现的词频，由于各条文本的长度不同，因此需要进行规范化使得这些频次差距缩小，从而满足这些频词在同等环境下进行对比的条件。逆文本频率指数 IDF (Inverse Document Frequency) 表示词语的重要性，其主要思想是特征项在不同类别的文档中出现的频率有所差异，即当该特征项在某一类文档中出现的次数较多，但在其余类别文档中出现次数忽略不计的情况下，该特征项具有良好的类别区分能力，使其具有较大的权重。假设所有单词重要性相同，为了对抗出现频率高的词语，用一个系数将权重变小。TF 与 IDF 乘积即为 TF-IDF 值。

假设留言数据集共有 n 条留言， m 个特征词，从而可以构建出 $m \times n$ 的矩阵 $A = [a_{ij}]$ ，矩阵中元素 a_{ij} 表示第 j 个特征词在第 n_i 条留言中的权重。由于 TF-IDF 算法需要调用两次，首先生成词典再创建特征向量，在处理长文本时会有一定的不足之处，为了加快处理速度、使得向量维度变小，本文运用 TF-IDF 算法计算权重最终可以得到“留言主题”的特征向量，结果如图 5.5 所示。

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

图 5.6 TF-IDF 特征向量

4.4.2 Hash

TF-IDF 比较简单，但由于它在词语权重和语料库中该词出现的次数之间建立了联系，需要调用两次，先创建词典在创建特征的方式会导致占用内容过大、速度缓慢的问题。

在处理文本高维稀疏性向量矩阵时引入“哈希技巧”可以克服这些问题，提高分类过程的时间和空间效率。它是一种快速且空间利用率高的特征向量化方法，可以将词块用哈希函数确定它在特征向量中的索引位置，并且无需创建词典。哈希技巧在提高空间效率的同时，可以实现线上流式传输，并且可以进行并行处理，为大规模分词系统创造了条件。

最终构建的“留言详情”特征向量如下图所示：

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

图 5.7 Hash 特征向量

4.4.3 Doc2vec

Doc2vec 的方法是 NLP 邻域 word2vec 的一个衍生模型，word2vec 指词向量表示，即将一个文本中的词语用一个固定维度的向量表示。Doc2vec 主要采用两种模型，其中本文采用 DM 模型进行训练，示意图如下：

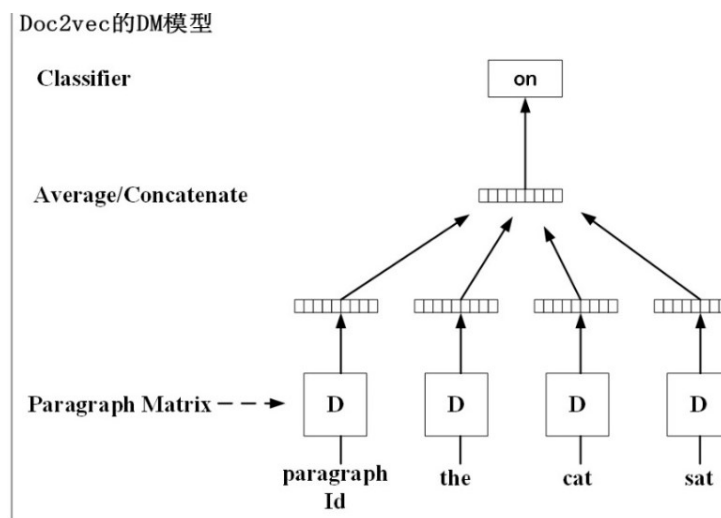


图 5.8 DM 模型示意图

该算法通过一个单层的神经网络结构建立模型，在模型的训练中得到副产物段落向量，同时增加一个向量作为段落的向量表示，与词向量通过求平均或者拼接的方式作为输入进入网络，网络通过梯度下降的方式进行优化。

不同于 TF-IDF 算法依赖于相同词语的出现的是，Doc2vec 方法的好处是可以不依赖于相同词汇的出现，同时反映出语义信息，将语义相似的两句话识别出来^[8]，这依赖于其良好的段落向量映射。

最终得到的“留言详情”特征向量如下图所示：

```
array([[ 0.02015193, -0.01692119, -0.00805672, ..., -0.00826775,
        0.00570985,  0.01442296],
       [-0.00020814, -0.03621939, -0.03923923, ..., -0.02637641,
        -0.00348128, -0.02120337],
       [ 0.04462942, -0.00859354, -0.14447823, ..., -0.04879132,
        -0.05250673,  0.06624547],
       ...,
       [ 0.01051524, -0.01797257, -0.05311875, ...,  0.0155513 ,
        0.0392946 ,  0.02598974],
       [ 0.02057061, -0.02620897, -0.04824081, ..., -0.00486129,
        0.01858059,  0.01129172],
       [ 0.02232585, -0.01363939, -0.035737 , ..., -0.00141808,
        0.03091889,  0.00719671]])
```

图 5.9 doc2vec 特征向量

4.4.4 结果分析

本文对较短字段的“留言主题”采用 TF-IDF 方式提取特征向量，对“留言详情”特征向量采用 Hash 和 doc2vec 的方式提取。虽然结合“哈希技巧”后的提取速度大大提高，但是从提取出的特征结果可以看出，与 TF-IDF 特征向量相同，Hash 特征向量仍然具有高纬度、稀疏的特点，这也将影响接下来的分类器识别准确度。而 Doc2vec 克服了这一缺陷，表现较良好。

提取特征后，为了比较特征、分析特征，最终确定用于分类的特征，本文首先对利用 TF-IDF、doc2vec、Hash 方法生成的三个特征用下文 6.1 介绍的多个模型进行训

练。由于数据量较大，同时比较不同分类器性能，本文通过 10 折交叉验证使用 F-Score 对其进行评价。

其中，k 折交叉验证如图 5.8 所示，将数据集 D 划分为 k 个大小相似的互斥子集，即 $D = D_1 \cup D_2 \cup \dots \cup D_k, D_i \cap D_j = \Phi (i \neq j)$ [9]

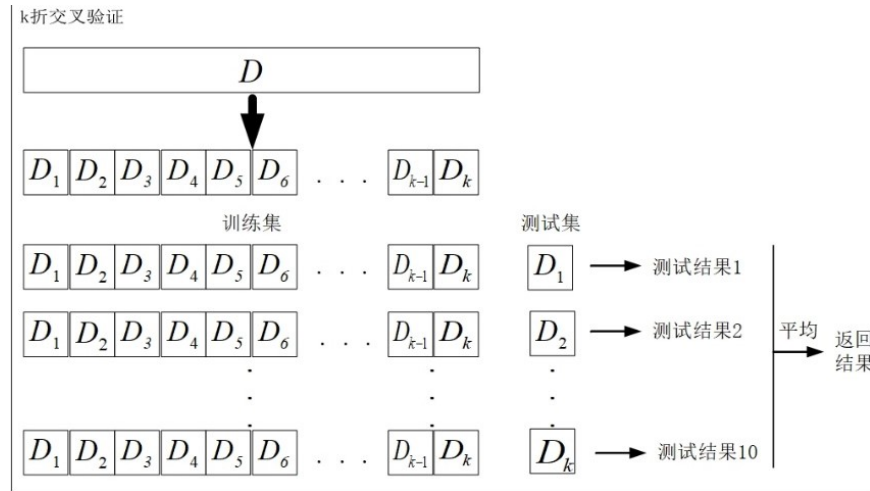


图 5.10 k 折交叉验证

F1-Score 公式如下：

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

$$P_i = \frac{(TP)_i}{(TP)_i + (FP)_i}$$

$$R_i = \frac{(TP)_i}{(TP)_i + (FN)_i}$$

其中， P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

接着，将 6.1 介绍的 k 近邻模型（KNN）、逻辑回归模型（LR）、支持向量机模型（SVM）、多项式朴素贝叶斯模型（NB）初始化基本参数进行训练，得到以下结果：

表 5.2 初始特征得分

	KNN	LR	SVM
TF-IDF	85.61(2)	58.32	86.48(1)
Doc2vec	72.53	84.14(1)	80.83(2)
Hash	75.37(2)	54.94	84.92(1)

从表中可以看出，SVM 模型表现良好，可以暂时作为特征筛选的模型，并且可以看出单纯的初始特征表现均不良好，需要进一步优化。

4.5 特征选择

特征选择能够影响分类精度，是文本分类的重要组成部分。在 5.4 提取出的特征中，主要由“相关特征”（relevant feature）和“无关特征”（irrelevant feature）两部

分组成，其中前者是指对于分类有用的特征，后者则是影响分类精度的特征。从原始特征集合中选择相关特征，组成新的特征子集的过程就是特征选择。通过特征选择可以从原始特征集中选择使得评价准则最大化的最小特征子集，剔除对数据分类影响最小的数据特征降低数据维度，从而更快地获得分类模型，提高分类性能，

本文对 5.4 提取出的 TF-IDF、Hash 特征通过选择合适的维度通过潜在语义分析（LSA）方法进行降维得到 LSA-TF-IDF 特征和 LSA-Hash 特征，接着对 Doc2vec 进行多项式构造后得到 Poly2-Doc2vec 特征；拼接若干个特征可以得到 Concatenation - Feature 特征，最后经过实验比较得到最终的向量化特征数据。

确定最终的向量化特征数据的过程中，本文采用的方法如下：1）对提取的、暂未处理的单个特征在单个模型上进行交叉验证得到单特征单模型的平均得分，得到该特征最适合的模型；2）将该特征在特定模型下选择最佳参数降维以及多项式构造；3）比较所有特征在最佳模型上的平均得分 4）最后由所有特征的平均得分确定最终的向量化特征数据。

4.5.1 LSA 特征降维

由 5.4 分析可知构建的特征向量存在高维度、稀疏性的特点，为此本文采用潜在语义分析（LSA），这也被称为潜在语义索引（Latent Semantic Index, LSI）。其主要思想是基于线性代数矩阵理论中的奇异值分解（Singular Value Decomposition, SVD）技术。该方法可以加强特征之间的关联性，削弱非相关特征之间的关联性（语义结构），从而将高维空间中的向量利用数学转换映射到低维的潜在语义空间中。

假设 A 为得到的特征向量矩阵，运用奇异值分解技术可以将 A 转化为两个正交矩阵和一个对角矩阵的乘积，如式下式所示。

$$A=U\Sigma V^T$$

U 是 $A^T A$ 的正交特征向量，也是 A 的左奇异值向量，也是 $m \times r$ 的正交矩阵 ($U^T U = I$)； V 是 $n \times r$ 的正交矩阵 ($V^T V = I$)，也是矩阵 A 的右奇异值向量以及 AA^T 的正交特征向量； Σ 为 $n \times r$ 的正交矩阵 ($V^T V = I$)，也是矩阵 A 的奇异值矩阵， Σ 中的元素按降序排列，为了进行降维，选取对角矩阵 Σ 中前 K 个最大的奇异值（设 $m > n$ ， $K < r$ 且 $K \ll \min(m, n)$ ），其余元素置 0，保留 U 和 V 的前 K 列。从而得到 A 的 K -秩近似矩阵 A_k ，如下式所示：

$$A_k=U_k \Sigma_k V_k^T$$

其中， $\Sigma_k = \text{diag}(\sigma_1, \sigma_2, \sigma_3, \dots, \sigma_k)$ 由 Σ 中的前 K 个元素构成， U_k 和 V_k 分别由 U 和 V 的前 K 列（ V^T 的前 K 行）组成。从而使得矩阵的秩由 r 降到 K ，并且在 K 维的特征空间中描述了原始矩阵潜在的语义结构。对原始特征空间 A^T 利用下面的公式进行映射。

$$A_{lsa} = A^T U_k \Sigma_K^{-1}$$

其中， A_{lsa} 为新形成的特征空间， A^T 为原始的特征空间。

这样就可以对前文生成的 TF-IDF、Hash 特征进行降维，同时通过循环确定了最佳降维维度。

对这 TF-IDF 特征、Hash 特征进行特征降维，特征的得分与降维维度的关系实验如图 5.11、5.12 所示。

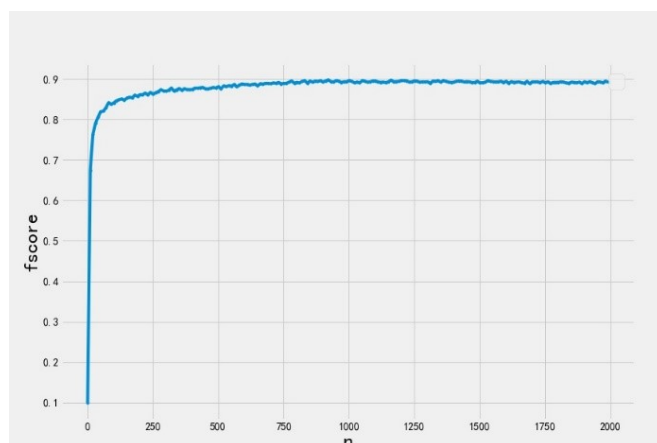


图 5.11 LSA-TF-IDF 降维得分图

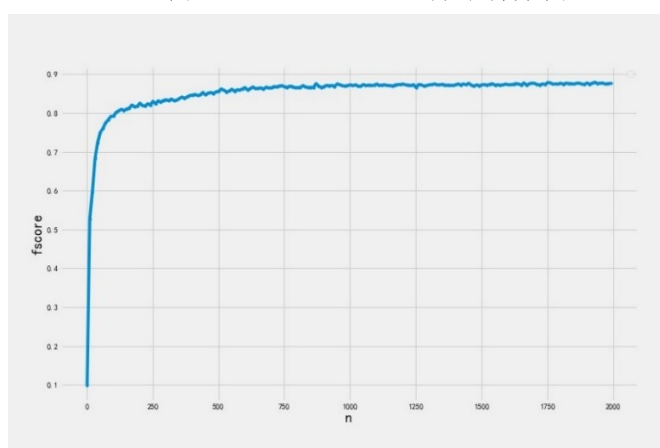


图 5.12 LSA-Hash 降维得分图

表 5.3 特征对比图

特征	TF-IDF	LSA-TF-IDF	Hash	LSA-Hash
F1 得分/%	84.48	87.23	82.92	84.93
训练时长/s	123.67	60.52	117.41	60.78

从图中可以看到，随着维度的上升特征得分得到了一定的提高。同时当 TF-IDF 维度降至 100 时得分高于 0.8，同时之后得分趋于稳定；当 Hash 特征维度降至 200 时得分高于 0.8，同时之后得分趋于稳定。并且从表 5.3 可以看出降维后的特征得分得到了一定提高同时降低了训练时长。

4.5.2 Poly 多项式构造

由 5.4.3 得到的 Doc2vec 特征可以通过使用 `sklearn.preprocessing.PolynomialFeatures` 来进行特征的构造，为了防止过拟合的发生同时防止维度过高，选用的构造阶数为 2。从而将最初的两个特征 $[a, b]$ 转化为六个特征 $[1, a, b, ab, a^2, b^2]$

将 2 阶多项式构造的特征与原始进行比较，结果如表 5.3 所示：

表 5.3 多项式构造对比图

特征	Doc2vec	Poly2-Doc2vec
F1 得分/%	72.44	72.55
训练时长/s	8.80	420.89

从实验结果可以看出虽然对 Doc2vec 构造了二阶多项式，但得分没有得到明显的提高，同时大幅度提高了训练时长。

4.5.3 特征拼接

经过降维后的 LSA-Doc2vec 特征、LSA-Hash 特征以及多项式构造的 Poly2-TF-IDF 特征可以随机选取两个或者随机选取三个拼接得到 Concatenation -Feature 特征。

将拼接后的特征进行比较如下表所示：

表 5.4 拼接特征对比图

拼接特征	LSA-TF-IDF- LSA-Hash	LSA-TF-IDF- Doc2vec	LSA-Hash- Doc2vec	LSA-TF-IDF- LSA-Hash- Doc2vec
F1 得分/%	87.83	73.53	73.43	74.15
训练时长/s	114.18	53.24	52.34	94.94

从实验结果可以看出拼接后的特征得分大部分没有得到提升，而其中一个虽然得分得到一定提高，但训练时长大大增加。

4.5.4 最终特征选择

将以上结果进行集中展示如表 5.5 所示：

特征	F1 得分/%	训练时长/s
TF-IDF	84.48	123.67
Hash	82.92	117.41
Doc2vec	72.44	8.80(1)
LSA-TF-IDF	87.23(2)	60.52
LSA-Hash	84.93	60.78
Poly2Doc2vec	72.55	420.89
LSA-TF-IDF-LSA-Hash	87.83(1)	114.18
LSA-TF-IDF-Doc2vec	73.53	53.24
LSA-Hash-Doc2vec	73.43	52.34(2)
LSA-TF-IDF-LSA-Hash-Doc2vec	74.15	94.94

从实验结果可以看出大部分拼接后的特征得分提升并不明显，降维后的 TF-IDF 特征和 Hash 特征进行拼接得到的特征，虽然在一定程度上提高了得分，但也导致训练时长大大增加。

最后经过综合考虑，本文选择 LSA-TF-IDF 做为最终的特征，此时的特征维度为 771，得分为 89.83，运行时长为 60.52s。

五 留言内容一级标签分类模型

5.1 分类流程

对向量化后的数据可以通过各个模型进行分类，我们文献^{[9][10][11]}从三个常用的分类模型下手进行分类，它们分别是 k 近邻、逻辑回归以及支持向量机模型，通过横向以及纵向比较首先确定出建模的特征，接着选取表现良好的模型利用遗传算法进行优化。

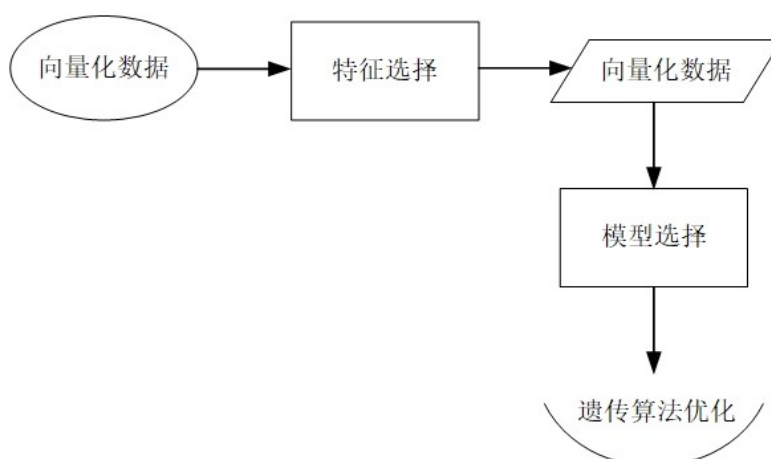


图 6.1 分类流程图

5.2 分类器模型及参数选择

5.2.1 K 近邻模型

K 近邻（K-Nearest Neighbor，简称 KNN）模型工作机制如下：给定测试样本，基于某种距离度量找出训练集中与其最靠近的 k 个训练样本，基于这 k 个样本进行预测。在分类中可以使用“投票法”，选择这 k 个样本中出现次数最多的类别标记为预测结果。

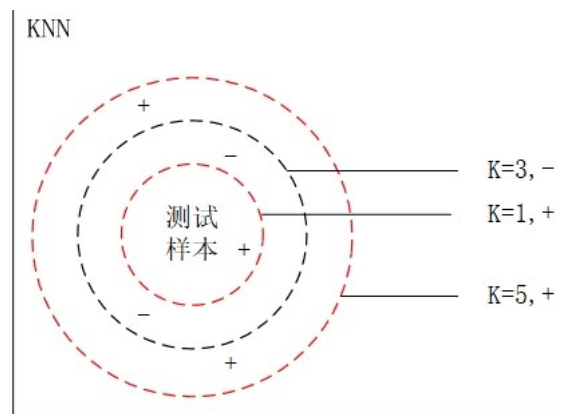


图 6.2 KNN 二分类示意图

K 近邻涉及到的超参数主要邻居数 k 、距离计算公式以及功率参数 p ，设定邻居数 k 取值范围 $[1,10]$ ，距离计算公式为“uniform”或者“distance”，功率参数 p 取值范围 $[1,5]$ 。

通过实验以及参考文献^[10]可以确定最佳参数为 $k=3$ ，距离计算公式取“distance”， $p=2$ ，此时对于最佳得分为 0.57，也可以看出融合后的特征在某些模型上的表现可能会得到降低，这是融合特征所带来的弊端，但总体上可以通过模型的处理使得弊端消除。

5.2.2 逻辑回归模型

逻辑回归是一种分类学习方法，直接对分类可能性进行建模，无需事先假设数据分布，避免假设分布不准确所带来的问题。它的求解目标是直接对任意阶可导的凸函数，有很好的数学性质，现在很多数值优化算法都可以直接用其求解最优解。

逻辑回归用于多分类时，本文主要考虑正则强度 C 对其准确率的影响，通过设定正则强度区间 $[0.01,5.0]$ 等间隔的选取 200 个点进行实验，实验结果如图 6.2 所示。

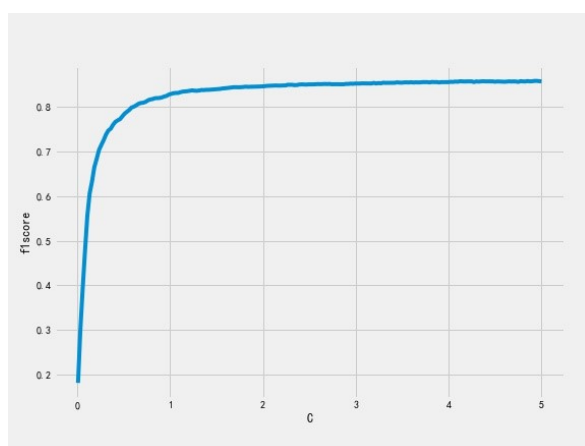


图 6.3 逻辑回归正则强度的影响曲线

从图中不同的 n 值对应的得分可以看出，在 $[0.01,5.0]$ 的区间内得分呈现剧烈浮动。当正则强度为 2.26 时，得分达到 0.839，之后的数值趋于平缓。因此，由实验结果本文最终确定逻辑回归的最佳参数 2.26。

5.2.3 支持向量机模型

支持向量机的基本思想是通过适当的核函数 K 将 n 维原始输入向量 $x \in R^n$ 映射到高维（或无穷维）特征空间 H 中，在此特征空间 H 中找到具有最大 *Margin* 的最有哦分界面将原始输入向量最大化的分开。不同的核函数将原始输入向量映射的高维特征空间不同，这样得到的学习能力也不同。

针对多分类问题，SVM 主要包括惩罚因子 C 、核参数 γ 需要设置。其中，惩罚因子 C 控制对错分样本的惩罚程度，若 C 为无穷大，则所有的约束条件都必须满足，也就导致分类面复杂，算法复杂度过高。核参数 γ 用于将原始空间映射到优化的 Hilbert 空间。^[11]

在这一部分的实验中，通过参考文献^[12]，首先确定了核参数 γ 为线性核参数，接着规定惩罚因子 C 的范围为 $[0.01, 5]$ ，最终得到实验结果如图 6.3 所示。

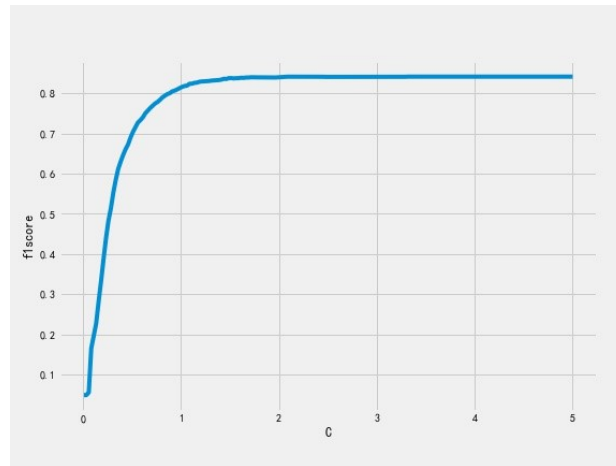


图 6.4 支持向量机惩罚因子的影响曲线

从图中可以看出，当惩罚因子取 2.19 时，得分趋于稳定，此时对应的得分最大值为 0.841。

5.3 遗传算法改进的支持向量机

从实验结果（表 5.2）以及相关文献^[14]可以看出，支持向量机在处理高维度、非线性等应用领域有较好的性能，通过 6.1.3 可以找到一定意义下 c 的最佳参数，但这种情况是基于核参数不变的条件。在解决实际问题时，支持向量机难以选取到最佳参数组成，通常采用经验值进行赋值，并且需要不断修正参数取值，影响算法运行效率。遗传算法的全局搜索性能，能够从潜在的所有核参数以及惩罚因子的组合中选出最优的参数组合^[14]

遗传算法（genetic algorithm，简称 GA）通过模仿自然界生物进化进程在问题解空间求解问题最优化解，该算法主要由三部分组成，分别是选择、交叉、变异^[13]。

本文采用单目标差分进化算法，算法流程如图 6.4 所示：

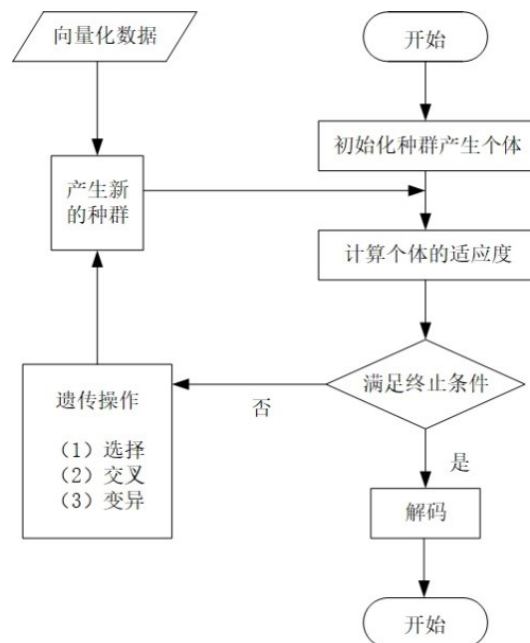


图 6.5 遗传算法流程图

Step1: 初始化参数，如种群大小、迭代次数和交叉变异概率等

Step2: 利用 SVM 模型对经 LSA 降维处理的 TF-IDF 特征进行训练，并计算个体适应度，本文采取的个体适应度为 F1 得分。

Step3: 选择差分变异的基向量，对当前种群进行差分变异，得到变异个体。

Step4: 将当前种群和个体合并，采用二项式分布交叉方法得到实验种群。

Step5: 再当前种群和实验种群之间采用一对一生存者选择方法得到新一代种群。

Step6: 判断是否满足终止条件。若满足返回最佳参数并进行分类预测，否则进行遗传操作返回 Step2 重新执行^[15]

遗传算法参数设置如下表所示：

表 6.1 重要参数设置

种群大小	编码类型	遗传代数	交叉概率	变异概率	C	γ
80	RI	30	0.55	0.01	[0.1,50]	[0.1,2]

遗传算法的寻优过程见下图

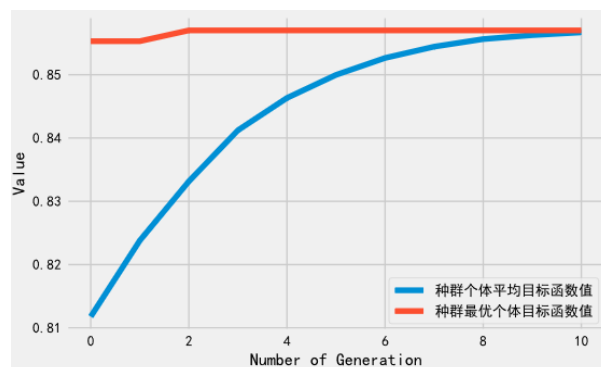


图 6.6 进化过程

经过实验可以看出遗传算法最优的一代是第三次迭代，得出支持向量机最优参数组合如下表所示：

表 6.2 优化结果

核参数类型	C	γ
RBF	3.34	2.23

5.4 改进模型训练与验证

LSA 与 GA 算法优化的 SVM 分类模型最优参数为 $C=3.34, \gamma=2.23$ ，将上文生成的最终特征 LSA-TF-IDF 作为本文的训练样本，从训练时长、F1 得分两方面对 SVM、LSA-SVM、GA-SVM、LSA-GA-SVM 四种模型进行对比，结果如表 6.3 所示：

表 6.3 仿真实验结果

算法	训练时长/s	F1 得分/%
SVM	13.6	81.5
LSA-SVM	104.8	88.7
GA-SVM	14.1	85.7
LSA-GA-SVM	107.9	89.6 (1)

从以上实验结果可以得出，LSA-GA-SVM 的 F1 得分最高，预测准确度最高；训练时长方面，与 LSA-GA-SVM 模型相比，虽然其余算法训练时长较低，但是从特征角度来看可以有效的对数据集的冗余特征进行处理。

六 热点问题挖掘模型

6.1 挖掘流程

在热点问题挖掘中，我们可以通过中文命名实体识别提取出相应的地点和人群。根据这些提取后的文本，利用第五章向量化的方法，得到向量化数据从而利用层次聚类进行归类。最后通过建立因子分析模型，构建三个一级指标以及六个二级指标进行热度评价，从而提取出热点问题。

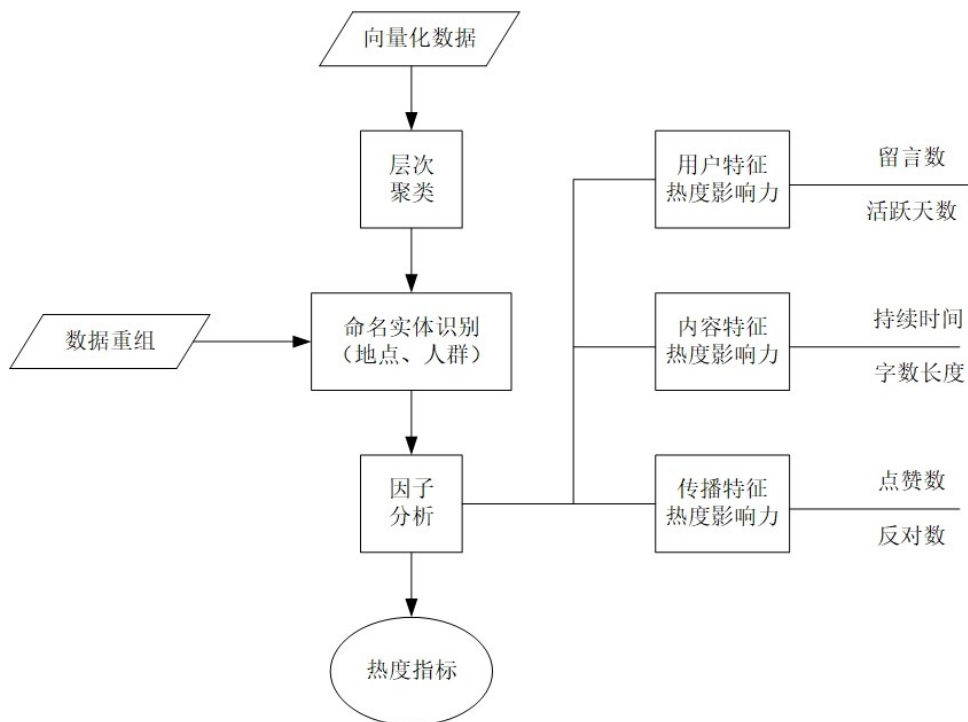


图 7.1 挖掘流程图

6.2 中文命名实体识别

主要任务是识别出文本中的人名、地名、机构名等专有名称并对其进行分类，常用的机器学习方法包括隐马尔可夫模型、最大熵、支持向量机、条件随机场等，这些均需人工提取特征，特征选择对结果有很大影响。这里选择用 HanLP 的预训练的 Bert 模型神经网络方法进行命名实体识别。

Bert 模型采用了双向的 Transformer，它的特征表示在所有层中共同依赖于左右两侧的上下文,该模型融合了其他模型的优点,并摒弃了它们的缺点,在诸多自然语言处理的后续特定任务上取得了良好的效果，模型图如下所示：^[16]

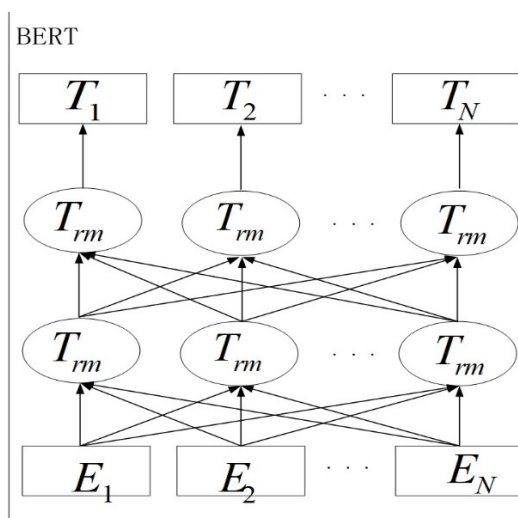


图 7.2 Bert 模型图

利用 HanLP 库可以对文本进行命名实体识别，通过编写程序可以提取出各个词语的词性，并且可以计算出各个词性的词语数量。

表 7.1 NRE 处理结果示例

预处理后的留言详情:	NRE 后的留言详情:
['您好', '这份', '非常感谢', '雅西', '地省', '小区', '传销', '东方', '航标', '栋楼', '传销', '窝点', '北京', '工作', '两天', '来市', '同学', '旅游', '娱乐', '心情', '第二天', '同学', '朋友', '朋友', '生意', '收入', '传销', '问题', '同学', '洗脑', '据说', '同学', '传销', '害人', '您们', '陷阱', '谢谢您']	[('航标 栋楼 传销 窝点 北京 工作', 'NS', 8, 14), ('同学 旅游 娱乐 心情 第二天', 'NS', 16, 21), ('传销 问题', 'NS', 26, 28)]
['师大附中', '隔音', '措施', '二层', '露天', '操场', '天天', '法律', '街道', '噪音', '标准', '上点', '小时', '隔音', '措施', '声音', '小点', '喇叭', '市民']	[('师大附中 隔音 措施 二层', 'NT', 0, 4)]

6.3 层次聚类

6.3.1 算法介绍

层次聚类在不同的层次对数据集进行划分，有“自底而上”的凝聚策略和“自顶而下”的分裂策略。这里采用“自底而上”，思路：一开始将数据集的每个样本看作一个初始聚类簇，然后找出两个聚类最近的两个簇进行合并，不断重复，直到达成某个条件或者归为一类。^[17]

基于凝聚的层次聚类算法根据类间距离计算方法的不同可以分为三类：

最小距离法（single-linkage）：类间距离等于两类对象之间的最小距离。

最大距离（complete-linkage）：类间距离等于两类对象之间的最大距离。

平均距离（average-linkage）：类间距离等于两类对象之间的平均距离。

我们采用的是最小距离法，计算两个文档之间的余弦相似度。

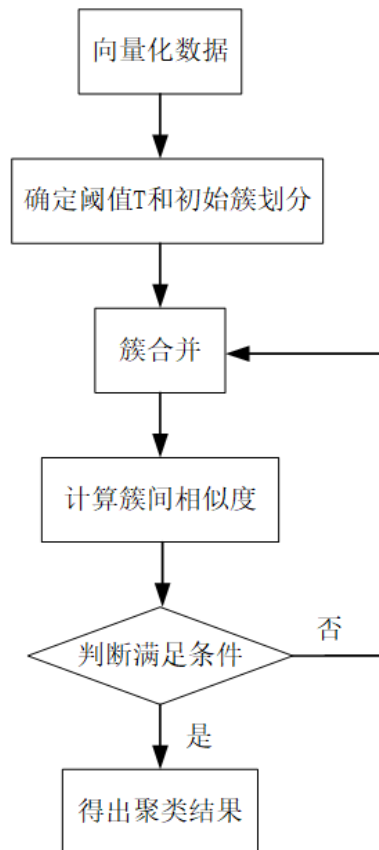


图 7.3 层次聚类流程图

具体思路是：初始聚类簇为每一个文本，确定初始阈值 T ，接着计算任意两个簇之间的相似度，相似度大于 T 的簇进行合并，同时减小阈值 T ，再继续合并相似度大于 T 的簇，不断重复此过程，直至阈值 T 或者簇划分数达到一定的条件，返回最后的簇划分集合即为层次聚类算法的聚类结果。^[18]

6.3.2 算法求解

通过认为设定阈值 $T=0.8$ ，可以将附件 3 数据进行分类为 2399 类，绘制聚类结果图如下图所示，分类后的数据详见附件。

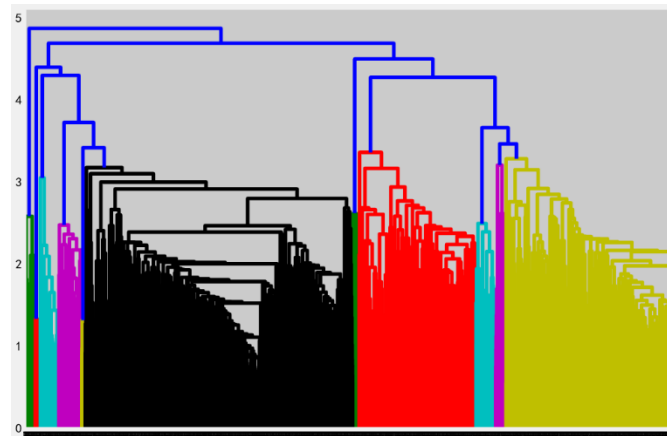


图 7.3 聚类结果图

6.4 热度评价指标

6.4.1 指标构建

根据数据组成结构，将热度评价指标体系分成三个一级指标即：用户特征指标、内容特征指标和引发关注度指标。

- （1） 用户特征影响力：留言者的信息和特征对热点问题有很大的影响，类似博客领袖影响力的研究，我们选取以下指标作为二级指标：①留言数，②活跃天数。
- （2） 内容特征影响力：留言本身就对热度有重要影响，文本的长度、包含的标签数等都可以影响热度，我们选取如下指标作为内容特征影响力的二级指标：①字数，②持续时长。
- （3） 传播特征热度：相关留言可以引起别人的关注，从而转化为热点问题。因此，我们选取的传播特征热度二级指标有：①点赞数，②反对数。



图 7.4 热度评价指标体系说明图

以下为各个指标说明：

表 7.2 热度评价指标说明表指标

一级指标	二级指标	定义
用户特征影响力	留言数	热点问题总的留言数
	活跃天数	留言用户总的活跃天数
内容特征影响力	字数	热点问题总字数
	持续时长	热点问题持续时长
传播特征热度	点赞数	热点问题留言点赞人数
	反对数	热点问题留言反对人数

6.4.2 指标提取

将计算后的评价指标进行抽取，部分数据如下表所示：

表 7.3 因子分析热度评价指标表

问题 ID	留言数	活跃天数	字数	持续时长	点赞数	反对数
1	1	1	307	1	1	3
2	1	1	318	1	8	1
3	2	2	360	215	3	3
4	3	3	3918	160	0	0
5	5	5	1495	222	18	0
6	2	2	816	63	0	0
7	3	3	510	146	1	0
8	1	1	64	1	2	0
9	2	2	150	8	0	0
10	3	3	3243	119	0	0

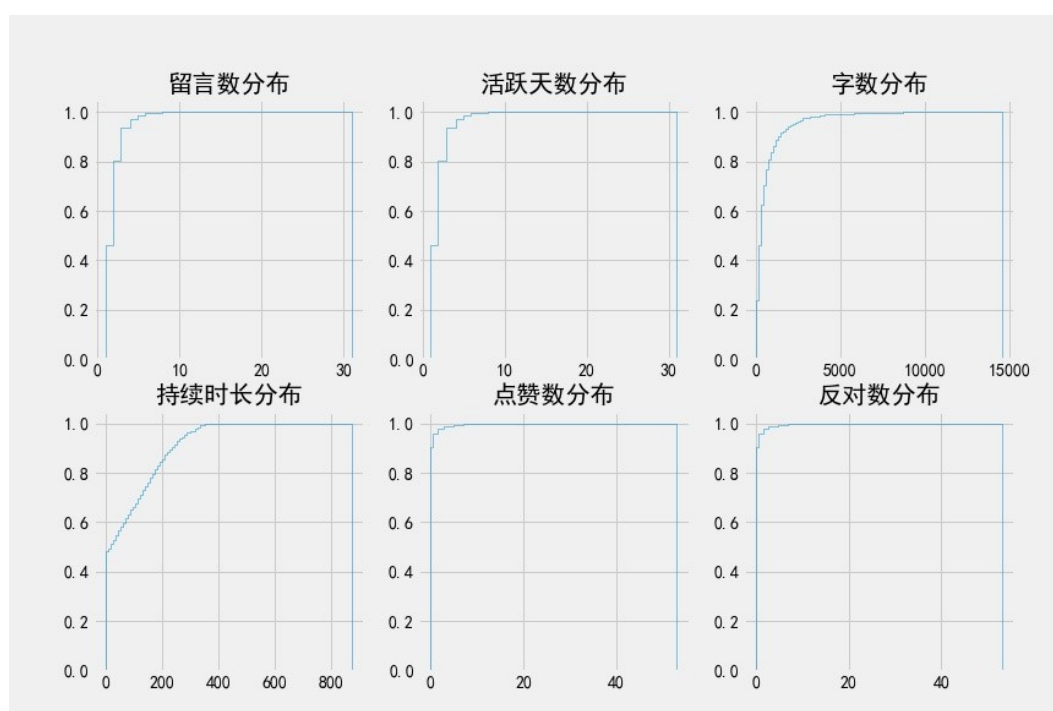


图 7.5 因子分析指标分布图

构建后的指标绘制分布图如图 7.5 所示，从图中可以看出，数据存在分布不均衡的表现，留言数指标主要集中与 0~10，活跃天数指标主要集中于 0~10，字数指标主要集中于 0~5000，持续时长指标主要集中于 0~400，点赞数指标主要集中于 0~10，反对数指标主要集中于 0~7。

6.4.3 指标检验

由于六个指标不服从正态分布，所以我们对这六个指标采用 kendall（一致性系数）进行相关分析，得到结果表 7.5。

表 7.5 指标相关性

协方差矩阵	留言数	活跃天数	字数	持续时长	点赞数	反对数
留言数	1.000000	1.000000	0.360420	0.793723	0.280052	0.136334
活跃天数	1.000000	1.000000	0.360420	0.793723	0.280052	0.136334
字数	0.360420	0.360420	1.000000	0.300214	0.114261	0.059163
持续时长	0.793723	0.793723	0.300214	1.000000	0.235114	0.108053
点赞数	0.280052	0.280052	0.114261	0.235114	1.000000	0.245408
反对数	0.136334	0.136334	0.059163	0.108053	0.245408	1.000000

通过协方差矩阵可以判断部分指标之间存在相关性，为此，删除相关性为 1 的指标“活跃天数”，同时需要进一步检验其是否可以用于因子分析。

6.5 因子分析热度评价模型

6.5.1 KMO 测度和巴特利球体检验

运用 SPSS 对上表进行 KMO 测度和巴特利球体检验，检验结果如表 7.6 所示：

表 7.6 KMO 测度和巴特利球体检验表

KMO 和巴特利特检验		.604
巴特利特球形度检验	近似卡方	1587.763
	自由度	10
	显著性	.000

从以上结果可以看出 KMO 的值为 0.604，并且巴特利球体检验的卡方统计值的显著性为 0，小于 1%，所以可以判断数据具有相关性，从而可以进行因子分析。

6.5.2 热度指标公式提取

接着对原始指标进行因子分析中的降维处理，提取三个公共因子结果如表 7.5 所示：

表 7.5 公共因子提取表

成分	初始特征值			旋转载荷平方和		
	合计%	方差%	累计%	总计	方差	累计%
1	1.928	38.559	38.559	1.883	37.654	37.654
2	1.019	20.371	58.930	1.005	20.095	57.749
3	.944	18.875	77.805	1.003	20.056	77.805
4	.745	14.905	92.710			
5	.365	7.290	100.000			

从表中可以看提取后的三个公共因子分别能够解释热度信息的 37.6%、20.1%、20.0%，累加和为 77.8%，能够较好的反映出热点问题热度信息。

接着，可以根据旋转后的成分矩阵（如表 7.6 所示）进行进一步分析。

表 7.6 各成分得分系数矩阵表

	公共因子		
	1	2	3
留言数 X_1	0.459	0.021	0.014
字数 X_2	0.362	-0.016	-0.061
持续时长 X_3	0.443	-0.085	-0.008
点赞数 X_4	-0.052	1.002	-0.026
反对数 X_5	-0.038	-0.025	1.001

由上表可知，留言数与第一个公共因子相关系数为 0.459，具有较强的相关性，字数与第一个公共因子也具有较强的相关性。同理，第一个公共因子与留言数、字数、持续时长具有较强的相关关系，第二个公共因子与点赞数具有较强的相关关系，第三个公共因子与反对数具有较强的相关关系。

从而将初始构造的指标体系修改为：内容特征影响力，点赞影响力以及反对影响力。这三个影响力分别作为公共因子 F_1 F_2 F_3 其中，内容特征影响力分为“留言数”、“字数”以及“持续时长”指标；点赞影响力为“点赞数”指标，反对影响力为“反对数”指标，指标说明图如下：

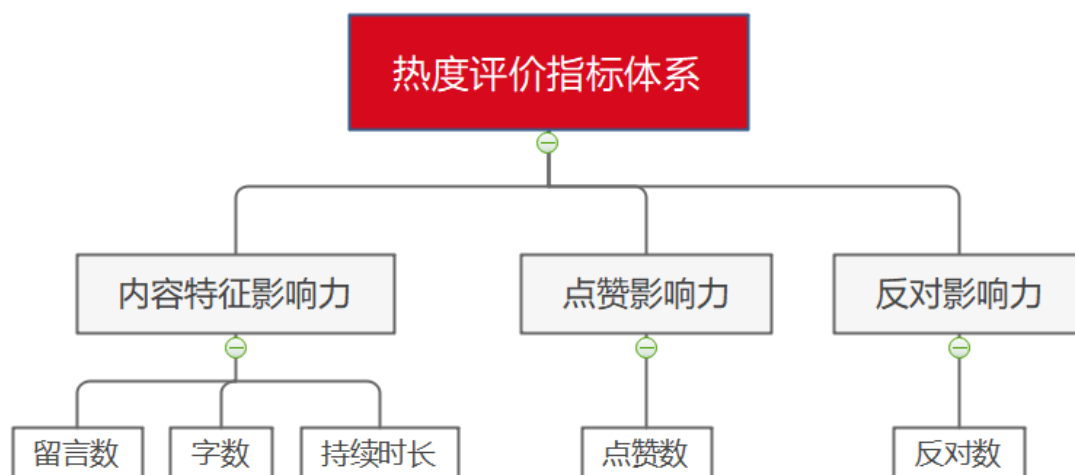


图 7.6 指标体系

最终可以得到各个公共因子表达式：

$$F_1=0.459 * X_1 + 0.362 * X_2 + 0.443 * X_3$$

$$F_2=1.002 * X_4$$

$$F_3=1.001 * X_5$$

以及根据表 7.5 得到的关于综合因子热度指数 F 公式

$$F = 0.376 * F_1 + 0.200 * F_2 + 0.200 * F_3$$

6.5.3 评价结果

利用综合因子热度指数公式计算各个热点问题的热度指数，结果及其排名如表 7.7 所示。

问题 ID	F_1 值	排名	F_2 值	排名	F_3 值	排名	F 值	排名
1	112	582	1	51	3	10	42	335
2	116	578	8	44	1	12	45	332
3	226	469	3	49	3	10	86	291
4	1490	37	0	52	0	13	560	37
5	641	183	18	34	0	13	244	145
6	324	376	0	52	0	13	121	256
7	250	446	1	51	0	13	94	283
8	24	670	2	50	0	13	9	368
9	58	636	0	52	0	13	21	356
10	1228	50	0	52	0	13	461	49

根据热度指数排名提取出排名前 5 的热点问题（地点/人群选择相同类别出现次数超过两次的 ciyu1，问题描述结合人工提取），同时提取出对应热点问题的留言信息，提取后的数据详见附件“热点问题表.xls”和“热点问题留言明细表.xls”。

通过观察表 7.7，可以看出初始的聚类后的热点问题数据进行了新的排名，观察公共因子可以发现内容特征影响力方面表现好的数据热度排名往往较高，所以我们定义的热度评价体系更有利于那些内容特征影响力大的问题，也就是问题相关的留言数越多、相关留言字数越长同时持续时间长的留言更容易成为排名靠前的热点问题。

七 相关部门答复意见质量评价模型

7.1 评价总体流程

在评价过程中，首先需要对提出“答复意见”的各个指标。通过参考前人研究成果并且结合要求，我们将评价指标概括为答复意见的完整性、相关性、可解释性、时效性以及探究性。接着，将量化后的各个指标按照一定规则进行映射，从而提取出最终的映射指标。最后，利用得到的映射指标，根据变异系数法赋权建立评价模型。

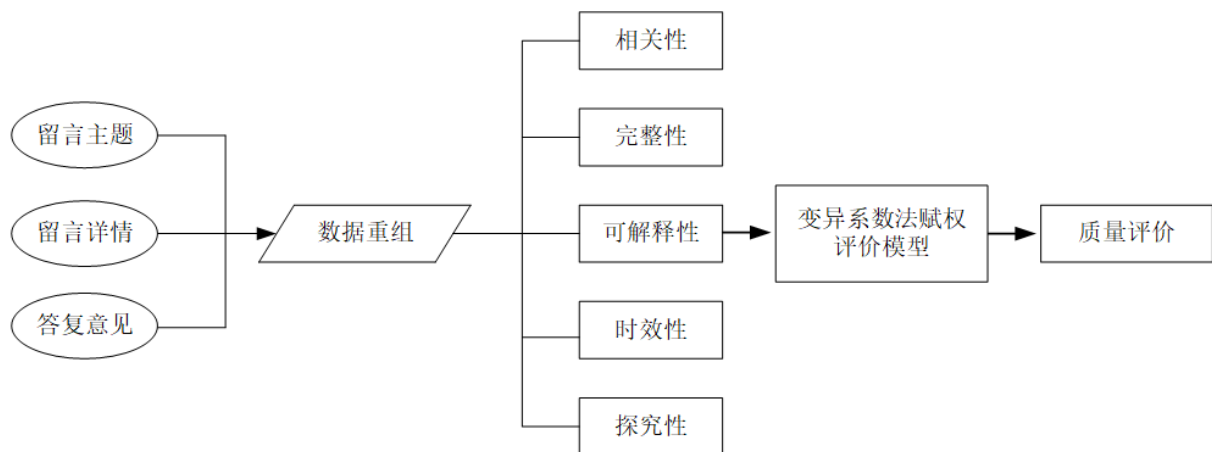


图 8.1 评价流程图

7.2 答复意见质量指标

7.2.1 指标构建

基于问题一中数据预处理的工作，将附件 4 数据进行重组数据后进行特征提取。由于“留言主题”包括于“留言详情”中，可以将通过“留言详情”与“答复意见”的关系提取指标。

根据文献^{[16][21]}的设计题目的要求和理解，假设答复质量与表 8.1 所示的五个维度的指标相关，进而可以对每个指标的具体特征项进行量化评分。

表 8.1 答复意见质量评价指标阐释

指标	定义	具体特征项
完整性	简述问题的现象、给出问题的解决方案	评论各种类型词语的数量
相关性	答复意见与留言问题的相关程度	特征词频次
可解释性	用语简洁规范	字符数长度
时效性	留言时间与答复时间的时间差	时间差
探究性	对问题分析、探究、总结的程度	具有探究性词语的词量

假设 TC (Text Content) 为答复意见的内容，TP (Text Phrase) 为答复意见内容分词后的短语对象集合，由于标点符号也可以影响质量评价，因此 TP 集合不去除标点符号， $n = |TP|$ ，CS (Concept Set) 某条留言答复意见所对应的留言详情集合，Me 为留言时间，Re 为答复时间。

以下为各个指标的计算公式：

(1) **完整性 (integrity)**：通常人们认为完整的句子应该包含主谓宾等部分，在本文中，答复意见完整性中应该包括各种类型的词语，例如词语总数 n ，名词数 nn ，形容词数 $nadj$ ，副词数 $nadv$ ，动词数 $nverb$ 。公式如下：

$$\text{inte}(\text{TP}) = \frac{nn + nadj + nadv + nverb}{n}$$

(2) **相关性 (relativity)**：相关性根据答复意见与留言问题的相关程度进行度量，相关性越高说明答复意见的内容越针对改留言问题。计算方法是答复意见分词后的短语对象集合占留言问题词典的比例，公式如下：

$$\text{real}(\text{TP}) = \frac{\sum_{tp_i \in \text{TP}} f(tp_i)}{|\text{CS}|}$$

式中， tp_i 为 TP 中的短语； $|\text{CS}|$ 为 TP 短语的数量和，函数 $f(tp_i)$ 公式如下：

$$f(tp_i) = \begin{cases} 1, & tp_i \in \text{CS} \\ 0, & tp_i \notin \text{CS} \end{cases}$$

例如，某答复意见中出现了“业主大会”一词，并且属于该答复意见对应的留言详情集合 CS，则 $f = 1$ ；否则 $f = 0$ 。遍历答复意见中的集合，取累加后的值与答复意见短语数量和的比值记为相关性指标得分

(3) **可解释性 (interpretability)**：可解释性高的答复意见通常字符数量较多，由于部分答复意见重复留言主题内容外设计内容较少，而留言主题通常不超过 20 个字符，所以定义可解释性指标公式如下所示：

$$\text{inte}(\text{TC}) = \begin{cases} 0, & \text{length}(\text{TC}) \leq l_0 \\ k, & \text{length}(\text{TC}) \in (l_{k-1}, l_k] \\ k+1, & \text{length}(\text{TC}) > l_4 \end{cases}$$

其中， $k = 0, \dots, 4$ ， l_k 可以根据实际情况进行调整，由于答复意见内容中最少长度为 3，最大长度为 7883，平均长度为 579，分布图如下图所示。获取答复意见文本长度并且绘制分布图如图 8.2 所示，所以我们将 l_k 初始设置为 20, 100, 500, 1000, 2000。

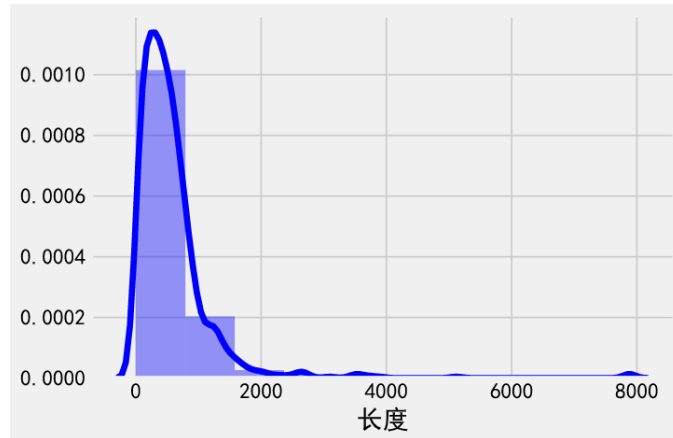


图 8.2 答复意见长度分布图

(4) **时效性 (timeliness)**：时效性高的说明相关部分回复及时，答复质量高。因此可以将留言时间和答复时间的时间差作为时效性进行量话，作为度量后的指标。相关法律指出，“有关行政机关收到度信访事项后，问能够当场答复是否受理的，应当

当场书面答复；不能当场答复的，应当自收到信访事项之日起 15 日内书面告知信访人。但是，信访人的姓名（名称）、住址不清的除外”。因此我们以 15 天为界限，公式为：

$$\text{time}(\text{TP}) = \begin{cases} 1, & \text{Re-Me} \leq 15 \\ \frac{15}{\text{Re-Me}}, & \text{Re-Me} > 15 \end{cases}$$

（5）探究性（exploration）：探究性较强的回复通常伴随着强烈的观点表达特征，这些词语包括“我”、“觉得”、“想”等词语，这类词语通常伴随着明显的观点表达特征，因此探究性根据这类词语所占答复意见的比重来衡量将这类词语记为 S（symbol），令集合 $S = \{\text{“我”}, \text{“发现”}, \text{“想”}, \text{“探究”}, \dots, \text{“总结”}\}$ ，具体的集合 S 详见附录，计算公式如下：

$$\text{expl}(\text{TP}) = \frac{\sum_{tp_i \in S} g(tp_i)}{n}$$

其中，n 为答复意见分词后的词语数量；函数 $g(tp_i)$ 定义如下：

$$g(tp_i) = \begin{cases} 1, & tp_i \in S \\ 0, & tp_i \notin S \end{cases}$$

7.2.2 指标提取

通过编写 python 程序，可以得到部分数据的特征指标，绘制累加曲线的指标特征分布图如下图所示，从图中可以看出，数据分布比较分散，所以将各指标分别根据表 8.2 映射成相应的级别。

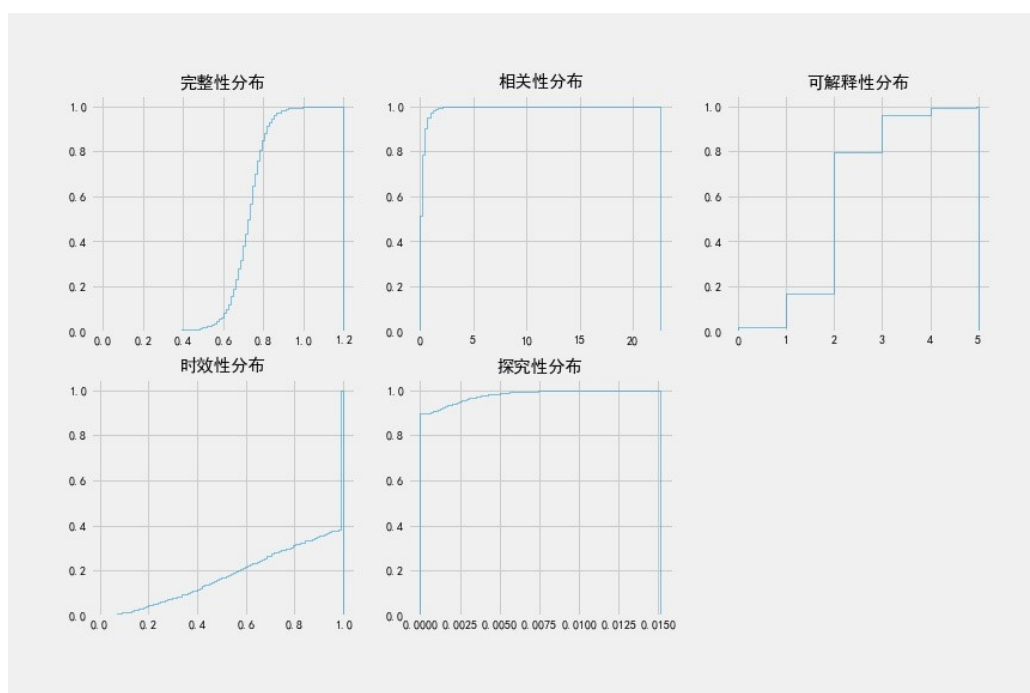


图 8.3 指标分布图

表 8.2 评价指标分值对应表

对应分值	指标				
	完整性	相关性	可解释性	时效性	探究性
0	0	0	0	>0.8	0
1	0~0.4	0~0.2	1	0.6~0.8	0~0.001
2	0.4~0.6	0.2~0.4	2	0.4~0.6	0.001~0.002
3	0.6~0.7	0.4~0.6	3	0.2~0.4	0.002~0.003
4	0.7~0.8	0.6~0.8	4	0~0.2	0.003~0.004
5	>0.8	>0.8	5	0	>0.004

通过对评价指标特征进行映射可以得到最终的五个映射指标，通过表 8.3 进行结果举例，同时绘制映射指标分布图可以看出数据分布情况。

表 8.3 映射指标举例

留言编号	完整性	相关性	可解释性	时效性	探究性
2549	4	2	2	0	0
2554	3	1	2	0	0
2555	4	3	2	0	0
2557	4	5	2	0	0
2574	3	5	2	0	0

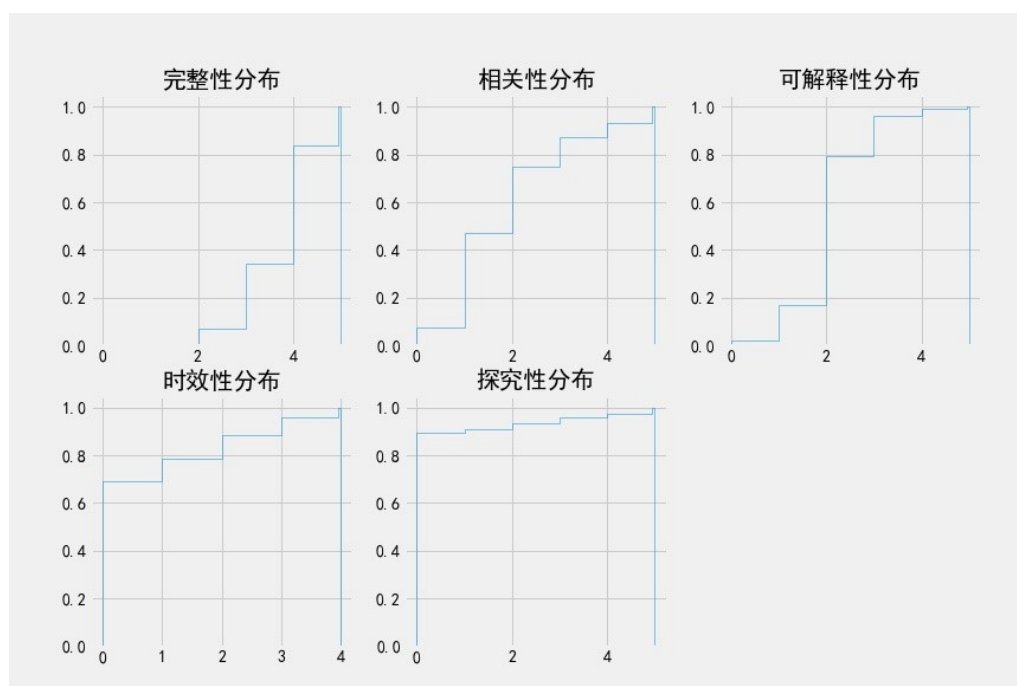


图 8.4 映射指标分布图

通过观察，可以看出政府部门答复意见总体上呈现出较好的状态，完整性、相关性、可解释性的得分都较高，但是时效性和探究性较低。

可以初步判断政府部门回答比较专业，解决问题也比较实际，但是回答不太及时，不够深入探究

7.2.3 指标检验

我们对五个指标的评分进行相关分析，得到表 8.5，绘制图 8.5。

表 8.5 指标相关性

协方差矩阵	ω_1	ω_1	ω_1	ω_1	ω_1
ω_1	1	-0.09047	0.155755	0.131133	0.218311
ω_1	-0.09047	1	-0.03233	0.342051	0.042107
ω_1	0.155755	-0.03233	1	0.107249	-0.00495
ω_1	0.131133	0.342051	0.107249	1	0.086213
ω_1	0.218311	0.042107	-0.00495	0.086213	1

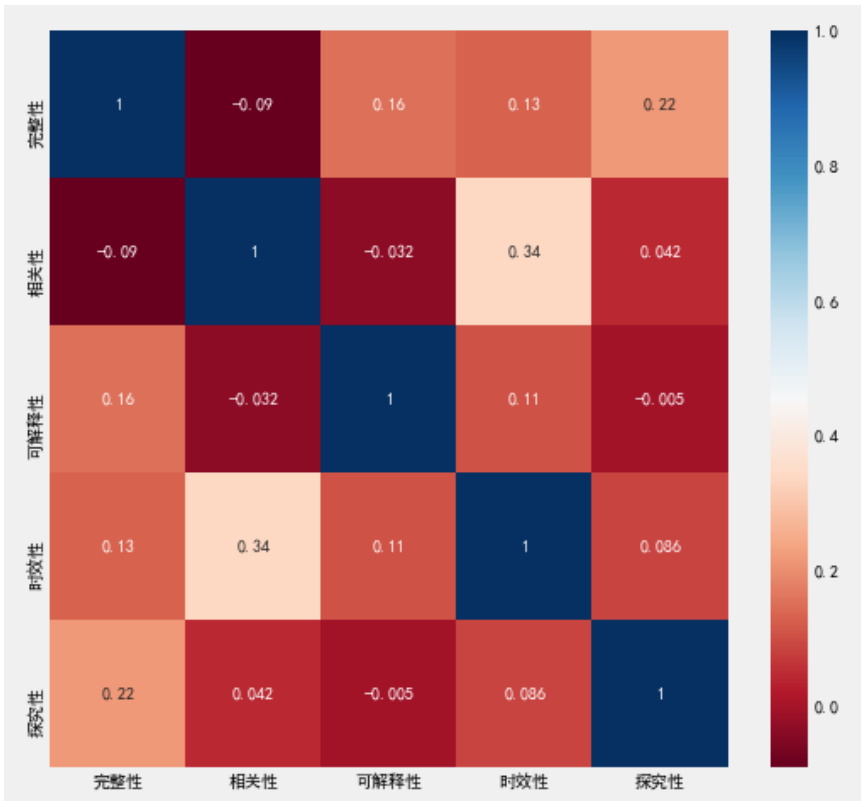


图 8.5 指标相关性图

通过协方差矩阵可以判断五个指标间无线性相关关系，可以进行变异系数法的模型建立与求解。

7.3 变异系数法答复意见质量评价模型

答复意见评价模型采用变异系数法（Coefficient of Variation Method）的方法。该方法是基于指标评价值的客观赋权方法，可以充分利用指标数据自身的信息从而得到相应的矩阵。核心思想是用指标的变异系数来评价各个指标的差异程度，从而对各个指标设立不同的权重：对评价结果影响大的赋予较大的权重，对评价结果影响小的赋予较小的权重。[23][24]

Step1: 计算各指标变异系数:

可以根据上面的五个指标，采用线性加权的方式得到总的指标，

$$y_j = \sum_{i=1}^5 \omega_i y_{ij}$$

其中， y_j 是每条答复总的得分， ω_i 是每个指标的权重， y_{ij} 是每条答复对应指标的得分。

Step2: 计算变异系数权重:

总指标要充分反映每条答复意见的质量的差别，因此对变化程度大的指标要赋予更大的权重，所以考虑采用变异系数法赋权。变异系数法赋权的主要步骤是先对各个指标的得分进行无量纲化处理，进而求出变异系数权重 ω_i ，计算公式如下：

$$\omega_i = \frac{\delta_i}{\sum_{i=1}^N \delta_i}$$

其中：

$$\delta_i = \frac{\sigma_i}{\bar{y}_i}$$

$$\bar{y}_i = \frac{\sum_{k=1}^n y_{ik}}{n}$$

$$\sigma_i = \sqrt{\frac{\sum_{k=1}^n (y_{ik} - \bar{y}_i)^2}{n}}$$

求解:

根据上述公式，运用 Matlab 计算得到权重分别为

表 8.6 指标权重表

指标	ω_1	ω_2	ω_3	ω_4	ω_4
权重	0.28	-0.083	0.411	0.198	0.188

从权重可以看出，完整性和可解释性所占的权重较大。

答复得分详见见附件 results 中的答复意见得分表。下面列出部分得分如表

表 8.7 部分答复意见得分表

留言编号	得分
2549	1.794852
2554	1.593596
2555	1.711584
2557	1.545046
2574	1.260521
2759	1.989919
2849	2.472606
3681	2.318296
3683	1.838704
3684	1.593596

根据得分的情况我们判断，得分排名前 20%，即得分在 2.36 以上可以认为是比较好的答复，排名在后 20%，即得分在 1.55 以下可以认为答复的质量较差。分析这两类不同的答复得分不同的原因，可以发现存在以下问题：1) 政府部门不够深入探究，无法做到有问必答；2) 留言答复不及时

7.4 给相关部门的建议

通过智慧政务系统的推进，将为政府部门现代化管理提供方便，可以大大提高管理水平和施政效率。从未来的发展趋势来看，原始的“以政府为中心”的管理模式将逐渐转向“以公众为中心”的管理模式^[25]。随着越来越多的民众参与到公共事务管理中，目前相关部门也存在一些改进之处。

最应该改进的是答复不及时的现象，国家政策说明有关部门应该在群众留言十五天内给出答复，但是从结果来看远超 15 天，这将降低解决问题的效率同时消磨群众的热情。所以政府部门应该提高回复效率和质量，提高行政能力，畅通服务渠道，解决群众更多的问题。

八 总结与模型评价

8.1 问题一

对于问题一，本文采用了三种方法进行构造特征，同时将特征降维和构造一阶多项式后的特征进行拼接，完成特征的融合。采用表现较好的支持向量机模型进行有监督的学习，同时为了提升分类效果，采用遗传算法对这一模型进行改进。从实验结果可以看出，基于特征降维与遗传算法改进的 SVM 算法可以有效的处理数据集的冗余特征，提高标签分类器效果。

8.2 问题二

对于问题二，我们通过命名实体识别提取出地点和人群后进行层次聚类。同时定义两层评价指标作为热度评价指标，利用因子分析作为我们的热度评价模型提出出排名前五的热点问题。该模型可以对指标体系进行合理性的检验，同时可以确定提取因子的权重得到最终的结果。

8.3 问题三

对于问题三，我们通过对“留言主题”、“留言详情”、“答复意见”进行数据重组后构建五个指标进行指标求解与检验，同时采用变异系数法的方式确定指标权重，

九 未来的工作

- 1、对于问题一，可以在哈尔滨工业大学停用词表的基础上针对特定情境制作相应的停用词表，同时在遗传算法调节上引入新的方法进行比较。
- 2、对于问题二，层次聚类可以利用收敛性进行阈值自动搜索，问题描述的提取可以采用新的模型
- 3、对于问题三，可以对质量评价指标进行优化，同时对指标权重排序灵敏度开展研究。

十 参考文献

- [1] 杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系 语料库构建 [J]. 软件学报, 2016, 27(11): 2725—2746
- [2] 武文婷. 基于微博的公安舆情监控系统研究与实现 [D] . 长 春: 吉林大学, 2014.
- [3] 官琴, 邓三鸿, 王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2017, 1(03): 72-80.
- [4] 牛永洁, 田成龙. 融合多因素的 TFIDF 关键词提取算法研究[J]. 计算机技术与发展, 2019, 29(07): 80-83.
- [5] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(S1): 167-170+180.
- [6] 叶雪梅. 文本分类 TF-IDF 算法的改进研究[D]. 合肥工业大学, 2019.
- [7] 姚芳. 基于 python 的中文文本分类研究[D]. 华中科技大学, 2016.
- [8] 贺益侗. 基于 doc2vec 和 TF-IDF 的相似文本识别[J]. 电子制作, 2018(18): 37-39.
- [9] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016
- [10] Polikar R, Zhang C, Ma Y. Ensemble machine learning: Methods and applications[M]. 2012.
- [11] 刘丽娟. 支持向量机超参数调节方法的研究及其在人脸识别中的应用[D]. 重庆大学, 2010.
- [12] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. International Conference on computer vision & Pattern Recognition (CVPR'05). IEEE Computer Society, 2005, 1: 886-893.
- [13] HOLLAND J H. Adaptation in Natural and Artificial Systems [M]. Michigan: Michigan University Press, 1975
- [14] 张小琴. 基于特征选择与遗传算法改进的支持向量机入侵检测方法[J]. 龙岩学院学报, 2020, 38(02): 18-23.
- [15] 张好勇, 张东亮, 赵雨, 刘景宇, 王正. 基于 PCA 和 GA 算法优化最小二乘支持向量机的开关柜温度预测[J]. 电气应用, 2020, 39(02): 59-63.
- [16] 祖木然提古丽·库尔班, 艾山·吾买尔. 中文命名实体识别模型对比分析[J]. 现代计算

机,2019(14):3-7.

[17] 朱翔,史晓东,陈毅东. 基于层次聚类的中文人名消歧方法研究[J]. 心智与计算,2010,4(04):236-241.


[18] 向继. 一种自动搜索阈值的中文文本层次聚类方法[A]. 信息产业部互联网应急处理协调办公室. 全国网络与信息安全技术研讨会论文集(上册)[C].信息产业部互联网应急处理协调办公室:,2007:7.

[19] 何跃,蔡博驰. 基于因子分析法的微博热度评价模型[J]. 统计与决策,2016(18):52-54.

[20] 邹沁含,庞晓阳,黄嘉靖,刘司卓. 交互文本质量评价模型的构建与实践——以 cMOOC 论坛文本为例[J]. 开放学习研究,2020,25(01):22-30.

[21] 王洪伟,孟园. 在线评论质量有用特征识别:基于 GBDT 特征贡献度方法[J]. 中文信息学报,2017,31(03):109-117.

[22] 马龙军. 基层政府网络回应机制创新研究——以 Z 县政府网站 2012-2018 年留言回复为例[J]. 未来与发展,2018,42(11):58-63.

[23] 李京峰,项华春,严雅榕,李正欣. 基于离差最大化的组合赋权评价方法及其应用[J/OL]. 火力与指挥控制:1-7. 

[24] 李浩. 基于变异系数法的智能变电站通信系统可靠性评估方法[J]. 电工电气,2018(11):22-26+39.

[25] 竺乾威. 理解公共行政的新维度:政府与社会的互动[J]. 中国行政管理,2020(03):45-51.

十一 附录





图 11.3 交通运输词云图



图 11.4 教育问题词云图



图 11.5 劳动和社会保障词云图



图 11.6 商贸旅游词云图



图 11.7 卫生计生词云图

