“智慧政务”中的文本挖掘应用

摘要

随着智慧政务系统的普及，越来越多关乎社情民意的文本数据需要被分析，自然语言处理技术的应用有效提升政府的管理水平和推动施政效率。

对于问题一，我们首先对数据分布进行观察并进行文本预处理，包括去除文本非中文部分、分词、去除停用词和文本向量化等处理。其次，根据词频矩阵，我们选取常用的朴素贝叶斯模型、knn 模型和支持向量机模型分别对附件 2 中所有留言详情进行分类训练，得到三种模型对应的 F_1 值分别为 0.94、0.74 和 0.99，运行时间分别为 0.04s、4.58s 和 0.04s，考虑到分类器的过拟合和运行速率的问题，我们认为朴素贝叶斯模型更为稳健和高效。

对于问题二，同样我们首先对留言主题和留言详情两列进行合并，再进行预处理。但考虑到位置性词语，在此我们不去除字母和数字，仅去除符号和停用词。其次，我们将对文本进行层次聚类分析，为使得聚类结果更精准，需先对留言进行区域性划分。我们选用 TF-IDF 作为文本的量化数据，在每个区域下进行层次聚类分析，并根据碎石图和轮廓系数选择最佳簇数。考虑到附件 3 所包含的数据，我们将某类问题的热度评价指标定义为该类问题留言数、反对数和点赞数之和。对热度指数进行降序排序后，我们先对每一区域排名前五的问题进行人工校对，剔除无关项并加入相关项，再对调整后的所有数据进行降序排序，对排名前二十的问题进行校对。考虑到热点问题的时间跨度，我们排除仅含有一条留言的问题，最终得出排名前五的热点问题，分别是 58 车贷案进展、五矿万境 K9 县房屋隐患、A4 区绿地海外滩小区因地铁造成的噪音困扰、A7 县松雅西地省站地下通道建设问题和广铁集团职工被捆绑销售车位，对此我们整理出热点问题表和热点问题留言明细表。

对于问题三，我们将采用基于熵权优化的主成分分析的综合模型对答复意见的质量进行评价。首先我们根据词频和时间等计算完整性、有效性和及时性指标并通过主成分分析进行降维处理，再以熵权法得到的权重综合三个主成分，以主此为得分来评价答复意见。最终求出的综合评分中大部分处于中等水平，少部分回复质量较高。

关键词：文本分类；层次聚类；热点问题；熵权法；评价模型

ABSTRACT

With the popularization of intelligent government system, more and more text data related to social conditions and public opinions need to be analyzed. The application of natural language processing technology can effectively improve the management level of the government and promote the efficiency of governance.

For problem 1, we first observed the data distribution and preprocessed the text, including removing non-chinese parts of the text, word segmentation, stop words and text vectorization. Secondly, according to the word frequency matrix, we choose the commonly used simple bayesian model, KNN model and support vector machine model separately to classify all messages in the attachment 2 for details training, corresponding to three kinds of models were F1 value was 0.94, 0.74 and 0.99, respectively, operation time were 0.04 s, 4.58 s and 0.04 s, considering the classifier of the fitting and running rate of the problem, we think that the simple bayesian model is more robust and efficient.

For question 2, we also combined the two columns of message subject and message details first, and then preprocessed them. However, in consideration of positional words, we do not remove letters and Numbers here, only symbols and stop words. Secondly, we will carry out hierarchical clustering analysis on the text. In order to make the clustering result more accurate, we need to divide the message regionally first. Tf-idf was selected as the quantitative data of the text, hierarchical clustering analysis was conducted under each region, and the optimal cluster number was selected according to the gravel diagram and contour coefficient. Considering the data contained in annex 3, we define the heat evaluation index of a certain type of problem as the sum of the number of comments, the number of objections and the number of thumb up. After ranking the heat index in descending order, we first manually proofread the top five questions in each region, removed irrelevant items and added related items, then sorted all the adjusted data in descending order, and proofread the top 20 questions. Considering the time span of the hot issues, we eliminate the problems of contains only a message in the end it is concluded that the top five hot spots, respectively is 58 car loan case progress, minmetals K9 county building hidden trouble, A4 area green space overseas beach area caused by the subway by the problems of noise, the A7 county pine iasi province station underground tunnel construction problems and guangzhou railway group worker by bundling parking, we sort out hot topic message list of tables and hot issues.

For question 3, we will evaluate the quality of the replies by using the comprehensive model of principal component analysis based on entropy weight optimization. First, we calculate the indicators of completeness, effectiveness and timeliness based on word frequency and time, and conduct dimensionality reduction through principal component analysis. Then, the weight obtained by entropy weight method is used to synthesize the three principal components, and the principal component is used as the score to evaluate the feedback. Most of the final obtained comprehensive scores are at medium level, and a few of the responses are of high quality.

Key words: Text Classification; Hierarchical Clustering; Hot Issues; Entropy Method; Evaluation Model

目　　录

一、引言.....1
1.1 挖掘背景.....1
1.2 挖掘目标.....1
二、问题分析.....2
三、问题一方案介绍.....3
3.1 文本预处理.....3
3.2 分类模型.....4
3.2.1 朴素贝叶斯模型.....4
3.2.2 knn 模型.....5
3.2.3 支持向量机（SVM）模型.....5
3.3 评价指标.....6
3.4 结果分析.....7
四、问题二方案介绍.....7
4.1 层次聚类.....7
4.1.1 聚类原理.....8
4.1.2 层次聚类流程.....8
4.1.3 轮廓系数.....9
4.2 特征选择.....9
4.2.1 TF-IDF.....9
4.3 结果分析.....10
五、问题三方案介绍.....10
5.1 主成分分析.....10
5.2 基于主成分分析和熵权法的综合评判.....11
5.3 处理方案及结果分析.....13
参考文献.....15

一、 引言

1.1 挖掘背景

近年来，随着互联网应用和数据科学技术的飞速发展，文本数据库资源得以迅速增长，带给我们具体有意义的知识和帮助却相对较少。

构建智能文本挖掘模型，学界对于机器阅读理解的研究从未止步。机器阅读理解作为目前热门的自然语言处理任务，目标是使机器在能够理解原文的基础上，正确回答与原文相关的问题。提高机器对语言的理解能力。机器阅读理解技术的发展对信息检索、问答系统、机器翻译等自然语言处理研究任务有积极作用，同时也能够直接改善搜索引擎、智能助手等产品的用户体验，因此，以阅读理解、文本挖掘为契机研究机器理解语言的技术，具有重要的研究与应用[1]。

进入 21 世纪以来，随着微信、微博、市长信箱、阳光热线的发展壮大，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，各类社情民意相关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。

因此电子政务要逐渐从原来的数据管理到目前的知识的管理和升级具有很大的阻碍，如何高校地利用和管理数量庞大而又复杂的电子政务信息资源已逐渐成为国内信息领域地重要研究课题之一。人们迫切需要有效的数据挖掘工具从海量文本数据中提取有价值的知识。

1.2 挖掘目标

目前，很多政务文本仍需人工来进行处理和分类。所以我们要构建一个智能的文本挖掘的划分体系，根据附件 1 中的内容三级标签体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理，从而能够大大提升工作效率，降低错误率，提升政府在群众心中的满意度。我们使用 F-Score 方法对文本分类和评价。

对于热点问题，政府需要及时关注和发现并进行针对性处理，我们根据附件 3 中特定人群和地点的留言进行聚类，并定义相应合理的评价指标并给出相应的评价结果。

然后我们对答复意见进行评价，从完整性、相关性、有效性和可解释性等来分析，通过分析答复时间与留言时间的时间差、有效词汇等分词数据，对提升政府的管理水平和行政能力有着很大意义。

二、问题分析

文本挖掘相比于一般的数据挖掘，其难点在于如何通过自然语言处理技术，合理地将文本量化为具体数据，从而应用于常见的数学模型并进行统计分析。

在所有分析步骤之前，我们需要先对文本进行预处理，提取有效信息。对于本文的留言数据问题，预处理的主要流程包括去除文本无效信息（符号、常见停用词等）、jieba 分词以及文本向量化。

对于问题一，我们通过文本预处理得到一个词频矩阵，它能够表示留言之间的相关信息。我们将该矩阵作为输入，训练常用的分类模型，例如朴素贝叶斯模型、knn 模型和支持向量机模型，计算三个模型的 F_1 值和运行时间，从过拟合和效率的角度进行对比。

对于问题二，我们可以通过层次聚类分析将留言进行聚类，以相同类别下问题留言数、反对数和点赞数之和作为热度评价指标。考虑到样本量和特征量维度较大，若直接进行聚类得到的结果较差。通过观察附件 3 数据，我们可以看出留言信息中会包含一些表示区域的词语，例如 A1 区、A2 区、A 市等，我们以此作为划分依据，对所有留言进行区域划分。以 TF-IDF 作为输入，对每个区域的留言进行层次聚类，并通过碎石图和轮廓系数选取最佳簇数。对于每一区域热度指标排名前五的问题进行人工校对和调整后，再对所有整合数据的热度指标排名前二十的问题进行调整。考虑到热点问题的时间跨度，我们排除仅含有一条留言的问题，从而得到最终的热度排名前五的热点问题表和热点问题留言明细表。

对于问题三，我们需要针对附件 4 中的答复意见进行评价。我们主要通过综合运用赋权法进行分析，结合主成分分析和熵权法，从而最终求得各评价的评分。首先，对于完整性、有效性和及时性指标我们需要进行量化以及无量纲处理。对于完整性，我们考虑一段回复中有效词汇的频数，一个有效词汇出现频率越多可能表示对这个的回答越完整。对于有效性，我们考虑了一段回复中相同词出现的频数（找出留言评价中出现次数多于两次的词汇个数），重复的词越少说明这段话可能越有效。而对于及时性，我们考虑回复时间与留言时间的距离大小（以总间隔天数/7 计算），考虑到间隔天数比较长，因此以星期为单位。最后，我们通过熵权法确定权重结合三个主成分，以此建立回复质量评价模型，计算每条回复的评价得分，并对所有得分分布进行分析。

三、问题一方案介绍

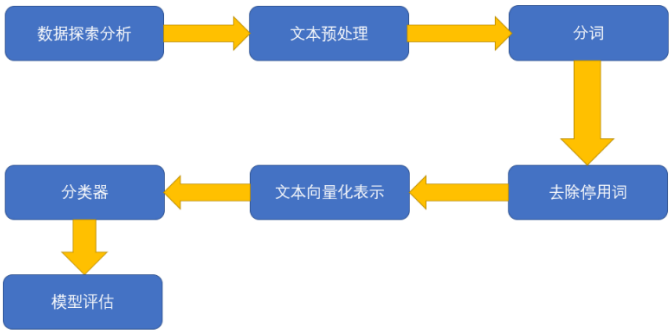

图 3-1 问题一流程图

3.1 文本预处理

在建立分类模型之前，首先我们要对附件 2 的留言数据进行预处理。根据附件 2 可知，留言内容的一级标签包括城乡建设、劳动和社会保障、教育文体、商贸旅游、环境保护、卫生计生和交通运输七个方面，我们可以通过绘制饼图直观感受数据的分布情况。由图 3-2 可知，八个标签下的数据大小较为均匀，不存在某一类占极大部分或极小部分的情况。

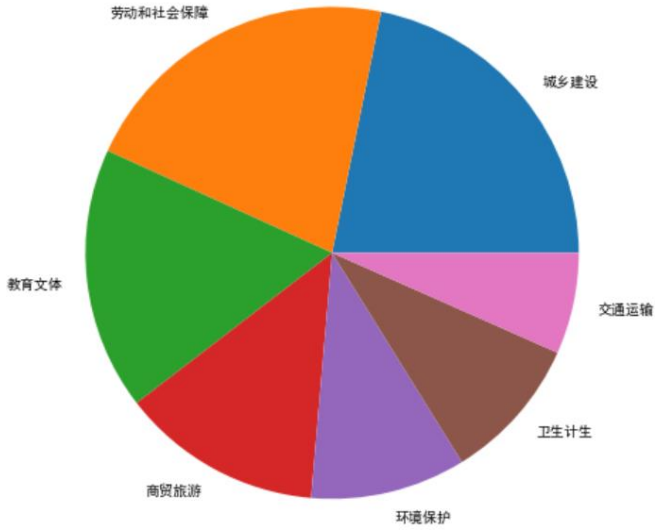

图 3-2 附件 2 中八个标签的数据分布情况

观察数据分布之后，我们对文本进行预处理的步骤包含以下四点：

1. 去除数据中非中文部分：考虑到运行效率和有效信息，我们将除去文本中的非中文内容，根据中文提供的信息进行后续分析；
2. 分词：基于 python 中的 jieba 分词包，我们将对每一条留言详情进行分词处理，即删去例如标点、英文、数字等非中文内容；

3. 去除停用词：即使删去了文本的非中文部分，留言仍含有许多无效词，例如我、你、这些、如果等，这些词对于分类并不能给予任何信息，相反，它还可能影响分类结果。因此，我们选取常用的 1208 个中文停用词，并在文本中删去；
4. 文本向量化：由于在分类模型中，我们要提供的是一个结构化的数量型数据，因此，我们需要对文本进行向量化处理，并提取特征，得到一个词频矩阵。

在进行文本预处理后，我们将处理后得到的词频矩阵进行分类模型训练。

3.2 分类模型

3.2.1 朴素贝叶斯模型

贝叶斯方法是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类。贝叶斯分类算法在数据集较大的情况下表现出较高的准确率，同时算法本身也比较简单[2]。假设特征条件之间相互独立的情况下，朴素贝叶斯分类先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。

设有样本数据集 D=$\{d_1，d_2，\cdots，d_n\}$,对应样本数据的特征属性集为$X=\{x_1，x_2，\cdots，x_n\}$, 类变量为$Y=\{y_1，y_2，\cdots，y_n\}$,D 可以分为y_m类别，不妨设 Y 的先验概率$P_{prior}=P(Y)$,后验概率$P_{post}=P(Y|X)$，由朴素贝叶斯算法可得：

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \tag{3-1}$$

朴素贝叶斯基于各特征之间相互独立，在给定类别为 Y 的情况下，上式可以进一步表示为下式：

$$P_{post}=P(Y|X)=\frac{P(Y)\pi_{i=1}^{d}P(X_i|Y)}{P(X)} \tag{3-2}$$

由于 P(X)的大小是固定不变的，因此在比较后验概率时只比较上式的分子部分即可。因此可以得到一个样本数据属于类别y_i的朴素贝叶斯计算如下式：

$$P(Y_i|X_1，X_{2，}\ldots X_d)=\frac{P(y_i)\pi_{j=1}^{d}P(X_j|y_i)}{\pi_{j=1}^{d}P(X_j)} \tag{3-3}$$

朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异。换句话说就是朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

但是，朴素贝叶斯算法也有很多缺点，比如属性独立性的条件就是朴素贝叶斯分类器的不足之处。数据集属性的独立性在很多情况下是很难满足的，因为数据集的属性之间往往都存在着相互关联，如果在分类过程中出现这种问题，会导致分类的效果大大降低。

3.2.2 knn 模型

knn 算法全称是 K 近邻算法，是通过测量不同特征值之间的距离进行分类。它的主要思想是如果一个样本在特征空间中的 k 个最相似，也即为特征空间中最邻近的样本中的大多数属于某一个类别，则该样本也属于这个类别。其中 K 通常是不大于 20 的整数，在这套算法里，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别[3]。

knn 分类算法的分类预测过程是：对于一个需要预测的输入向量 x，我们只需要在训练数据集中寻找 k 个与向量 x 最近的向量的集合，然后把 x 的类别预测为这 k 个样本中类别数最多的那一类。首先，我们根据给定的距离量度方法（一般使用欧式距离）在训练集 T 中找出与 x 最相近的 k 个样本点，并将这 k 个样本点所表示的集合记为N_k(x)。其次，根据如下所示的多数投票的原则确定实例 x 所属类别 y：

$$Y=\text{argmax}\sum_{x_i\in N_{k(X)}} I(y_i, c_j) \tag{3-4}$$

上式中 I 为指示函数：

$$I(y_i, c_j)=\begin{cases}0 & x\neq y\\ 1 & x=y\end{cases} \tag{3-5}$$

knn 算法可以处理分类问题，同时天然可以处理多分类问题和回归预测等多种问题。同时其简单，易懂，同时也很强大，对于手写数字的识别等问题来说，准确率很高。

不过 knn 算法也有其不可避免的问题，比如其效率很低，因为每一次分类或者回归，都要把训练数据和测试数据都算一遍。knn 算法对训练数据依赖度特别大，因为如果我们的训练数据集中，有一两个数据是错误的，刚刚好又在我们需要分类的数值的旁边，这样就会直接导致预测的数据的不准确，对训练数据的容错性太差。同时，knn 算法对于多维度的数据处理也不是很好。

3.2.3 支持向量机（SVM）模型

支持向量机（Support Vector Machines，SVM）是统计学习的代表方法。通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性

分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

设线性可分训练集为：T=$\{（x_1，y_1），（x_2，y_2），\cdots，（x_n，y_n）\}$，学习得到的超平面：

$$W^{*T}X + b^* = 0 \tag{3-6}$$

相应的分类决策函数：

$$f(x) = sign(W^{*T}X + b^*) \tag{3-7}$$

从而使得间隔最大化，不仅要讲正负类样本分开，而且对最难分的点（离超平面最近的点）也要有足够大的确信度将他们分开。

SVM 具有稳健性与稀疏性，是一种有坚实理论基础的新颖的适用小样本学习方法，简化了通常的分类和回归等问题，其同时考虑了经验风险和结构风险最小化，因此具有稳定性。SVM 的稳定性体现在其构建超平面决策边界时要求边距最大，因此间隔边界之间有充裕的空间包容测试样本。SVM 可以更好的处理多维问题。少数支持向量决定了最终结果，对异常值不敏感, 这不但可以帮助我们抓住关键样本、剔除了很多的多余样本,而且该方法不但算法简单，而且具有较好的鲁棒性。SVM 也可以利用已知的有效算法发现目标函数的全局最小值,有优秀的泛化能力[4]。

但是其也有缺点比如对大规模训练样本难以实施、解决多分类问题困难、对参数和核函数选择敏感。如果数据量很大，SVM 的训练时间就会比较长，如垃圾邮件的分类检测，没有使用 SVM 分类器，而是使用简单的朴素贝叶斯分类器，或者是使用逻辑回归模型分类。

值得注意的是，支持向量机性能的优劣主要取决于核函数的选取，所以对于一个实际问题而言，如何根据实际的数据模型选择合适的核函数从而构造 SVM 算法。目前比较成熟的核函数及其参数的选择都是人为的，根据经验来选取的，带有一定的随意性。

3.3 评价指标

我们在处理网络问政平台的群众留言时，首先对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。我们根据附件中给出的数据，建立了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：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{3-8}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

3.4 结果分析

我们分别采用朴素贝叶斯模型、knn 模型和支持向量机模型对留言进行分类训练，所得结果如表 3-1 所示。

表 3-1 三种模型分类结果

模型	朴素贝叶斯模型	knn 模型	支持向量机模型
F_1值	0.94	0.74	0.99
运行时间(s)	0.04	4.58	0.04

根据结果我们可以看出，从 F_1 值的角度来看，knn 模型分类结果较差，朴素贝叶斯模型和支持向量机模型结果较优，但支持向量机模型 F1 值接近于 1，可能存在过拟合的现象，因此若需要进行无标签分类预测，选用朴素贝叶斯模型可能更为稳健。从运行时间的角度来看，knn 模型的运行时间近似于其它两个模型的 100 倍，结合 F_1 值分析结果可以发现，朴素贝叶斯模型不仅分类结果更为准确与可靠，并且运行速度也较快，是一个较优的留言分类模型。

四、问题二方案介绍

4.1 层次聚类

对于无标签的留言数据，我们可以通过聚类算法将内容相似的留言分到同一类，但这样直接处理会存在一些问题，例如簇数的选择，并且所得结果较差。考虑到热点问题表现为某段时间某个地点某类人群所反映的问题，以及留言数据中大多包含 A1 区、A2 区和 A 市等这种表示城市区域的词语，我们考虑首先对留言进行地域上的划分，即仅包含 A1 区、A2 区、…、A9 区的留言作为对应区域的反映问题，其它留言作为 A 市整个范围内反映的问题。在进行区域划分后，我们通过层次聚类法对每一区域下的留言进行聚类，所得结果相比于直接对全部数据进行聚类要准确很多。

在聚类之前，我们先将文本进行预处理，处理步骤类似于问题一，但此时我们需要将留言主题和留言详情的两列内容进行合并，防止遗漏重要的地点和人群信息，并且此时不再除去所有非中文文本，防止遗漏区域信息。对于每一区域的

聚类，我们选用 TF-IDF 作为量化信息，将其作为输入进行层次聚类，并根据最大轮廓系数作为选择簇数的依据。在小样本中，我们可以通过遍历所有簇数的取值可能，选择最大轮廓系数对应的数值作为最终的结果；在大样本中，由于样本量和特征量都较大，每一次运行所需要的时间都较长，因此我们考虑通过碎石图和样本量的二分之一作为临界点，簇数每增加 m 个进行一次聚类（m 取值与样本量有关），最终选择最大轮廓系数对应的数值作为输入簇数。

4.1.1 聚类原理

聚类，就是将相似的事物聚集在一起，而将不相似的事物划分到不同的类别的过程，是数据分析之中十分重要的一种手段。层次聚类法要先计算样本之间的距离，每次将距离最近的点合并到同一个类。然后，再计算类与类之间的距离，将距离最近的类合并为一个大类。不停的合并，直到合成了一个类。其中类与类的距离的计算方法有：最短距离法，最长距离法，中间距离法，类平均法等[5]。

通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法，即凝聚的层次聚类算法和分裂的层次聚类算法，也可以理解为自下而上法和自上而下法。自下而上法就是一开始每个个体都是一个类，然后寻找同类，最后形成一个“类”。自上而下法就是反过来，一开始所有个体都属于一个“类”，然后根据 linkage 排除异己，最后每个个体都成为一个“类”。这两种方法没有孰优孰劣之分，只是在实际应用的时候要根据数据特点以及你想要的“类”的个数，来考虑是自上而下更快还是自下而上更快。为弥补分解与合并的不足，层次聚类经常要与其它聚类方法相结合，如循环定位等。

4.1.2 层次聚类流程

层次聚类的策略是先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有对象都在一个簇中，或者某个终结条件被满足。绝大多数层次聚类属于凝聚型层次聚类，它们只是在簇间相似度的定义上有所不同[6]。

我们采用欧式最小距离的层次聚类算法流程：

(1) 将每个对象看作一类，计算两两之间的最小距离；

(2) 将距离最小的两个类合并成一个新类；

(3) 重新计算新类与所有类之间的距离；

(4) 重复(2)、(3)，直到所有类最后合并成一类。

层次聚类的距离和规则的相似度容易定义，限制少，也不需要预先制定聚类

数。同时可以发现类的层次关系，也能可以聚类成其它形状。但是其计算复杂度太高，奇异值对其会产生较为严重的影响，算法很可能聚类成链状。

4.1.3 轮廓系数

轮廓系数可以用来评价聚类效果的好坏，它结合了内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

要计算轮廓系数，我们首先要计算样本 i 到同簇其他样本的平均距离a_i。a_i越小，说明样本 i 越应该被聚类到该簇[7]。将a_i称为样本 i 的簇内不相似度。簇 C 中所有样本的a_i的均值称为簇 C 的簇不相似度。第二，平均距离b_{ij}是样本 i 到其他某簇C_j 的所有样本的平均距离b_{ij}，称为样本 i 与簇C_j的不相似度。b_i定义为样本 i 的簇间不相似度：

$$b_i=\min\{b_{i1}，b_{i2}，\cdots，b_{in}\} \tag{4-1}$$

显然，b_i越大，说明样本 i 越不属于其他簇。

然后我们再通过样本 i 的簇内不相似度a_i和簇间不相似度b_i，定义样本 i 的轮廓系数：

$$S(i)=\frac{b(i)-a(i)}{max\{a(i),b(i)\}} \tag{4-2}$$

若$S(i)$接近 1，则说明样本 i 聚类合理；$S(i)$的值接近-1 的话，则表示样本 i 更应该分类到另外的簇；若$S(i)$近似为 0，则说明样本 i 在两个簇的边界上。

4. 2 特征选择

4.2.1 TF-IDF

TF-IDF 算法是基于统计的方法来衡量词或短语在文本信息中的关键性。它的主要原理:一个词在目标文本使用的次数较多，却在语料库中使用的次数较少，那么它就能够具备良好的文本区分能力。

某个词在目标文本中的 TF 值指的是该词在这个文本中出现的频率。在计算这个频率时，还需对它归一化，防止它偏向字数较多的长文本[8]。词i在文本j中的的词频 TF 值的公式如下所示：

$$TF_{i,j}=\frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}} \tag{4-3}$$

其中分子内容为词i在文中出现的次数，分母内容为文本j中所有词的总数。通过这种词频的计算方法，有效的防止了词频$TF_{i,j}$对较长文本的偏向性。

一个词的关键性的衡量可以用IDF值表示。假定一个文本语料库集合中包含词i的文本量越少，则该词的IDF_i值应该越大，这样词i的区分能力就越好，那么就越有可能作为关键词。具体公式如下：

$$IDF_i = log\frac{N}{n_i+1} \tag{4-4}$$

上式中 N 是文本集中的文本总数，n_i包含词i的文本数量。分母中的“+1”是为了处理公式中分母为 0 的情况。

经过上述方法的计算可以总结成下式。假设一个词i在某一指定的文本j中出现的频率较高而在整个文本集合中出现包含该词的文本较少，则它的 TF-IDF 值就较高，即词i较容易区分文档j，可作为关键词。

$$\text{TF-IDF}_i = TF_{i,j} \times IDF_i \tag{4-5}$$

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[9]。

4.3 结果分析

考虑到附件 3 所给数据，我们定义某类留言的热度指标为该类下的留言总数与反对数和支持数之和，根据所得结果，我们对每一区域排名前五的类别进行人工检查，即调整错误分类，经检查所有分类无误后选出最终在前五并且该类问题所包含的留言数大于 1 的热点问题，并对应导出热点问题表和热点问题留言明细表如附件所示。

排名前五的热点问题分别是 58 车贷案进展、五矿万境 K9 县房屋隐患、A4 区绿地海外滩小区因地铁造成的噪音困扰、A7 县松雅西地省站地下通道建设问题和广铁集团职工被捆绑销售车位。他们的热度指标分别为 2401、2118、698、103 和 99。

五、问题三方案介绍

5.1 主成分分析

设$U = \{x_1，x_2，\cdots，x_n\}$为研究对象的 n 因素，称之为因素集(或指标集)。而$V = \{c_1，c_2，\cdots，c_n\}$为诸因素所构成的权重评判集，那么加权之和就是

$$S = x_1c_1+x_2c_2 +\ldots+x_nc_n \tag{5-1}$$

我们希望能够更好的区分各指标间的差异，使得每个地区所对应的综合成绩s_1、s_2、s_3区分的很好，n 为学生人数。如果这些值很分散，表面区分的很好，转化到统计学中即为 Var($x_1c_1+x_2c_2 +\ldots+x_nc_n$)的值达到最大。由于方差反映了数据差异的程度，因此也就表明我们抓住了这 n 个变量的最大变异。

对于方差必须要加上某种限制，否则权值可选择无穷大而没有意义，通常规定

$$c_1{}^2+c_2{}^2+\ldots c_{\mathrm{n}}{}^2=1 \tag{5-2}$$

在此约束下，求上述方差式的最优解，这个解即代表主成分方向。一个主成分不足以代表原来的 n 个变量，因此需要寻找第二个乃至更多的主成分，要使后一个主成分不再包含第一个主成分的信息，统计上的描述就是让这两个主成分的协方差为零，让这两个主成分的方向正交[10]。

设Z_{i}表示第 i 个主成分，i=1,2,…n，可设

$$\begin{cases} Z_1 = x_1c_{11} + x_2c_{12} + \cdots +x_nc_{1n} \\ Z_2 = x_1c_{21} + x_2c_{22} + \cdots +x_nc_{2n} \\ \cdots\cdots \\ Z_{\mathrm{n}} = x_1c_{\mathrm{n}1} + x_2c_{\mathrm{n}2} + \cdots +x_nc_{nn} \end{cases} \tag{5-3}$$

其中对每一个 i，均有$c_{\mathrm{i}1}{}^2+c_{\mathrm{i}2}{}^2+\ldots c_{\mathrm{i}n}{}^2=1$，且（$c_{\mathrm{i}1}$，$c_{\mathrm{i}2}$，…，$c_{in}$）使得 Var（$Z_1$）的值达到最大；（$c_{21}$，$c_{22}$，…，$c_{2n}$）不仅垂直于（$c_{11}$，$c_{12}$，…，$c_{1n}$），而且使 Var（$Z_2$）的值达到最大，以此类推可得全部的主成分。

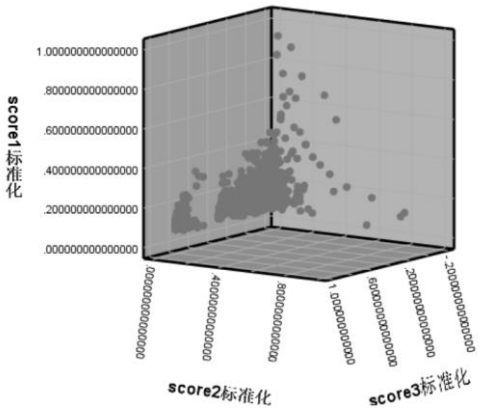

图 5-1 主成分因子示意图

5.2 基于主成分分析和熵权法的综合评判

对于留言回复的质量，完整性可以考虑一段回复中有效词汇的频数，一个有效词汇出现频率越多可能表示对这个的回答越完整；可解释性可以考虑一段回复

中因果关系和逻辑连接词出现的频数，这些词越多说明这段话可能越合理；及时性可以考虑以星期为单位的间隔时间。我们将对每条留言及对应回复计算两条内容同样的词出现的个数 x，找出留言评价中出现次数多于两次的词汇个数 y 以及对留言时间和留言回复的时间差（以总间隔天数/7 计算）。

我们主要通过综合运用主观赋权法以及客观赋权法来进行分析，将主成分分析和熵权法结合起来利用，从而最终求得各评价的评分。设$U=\{u_1，u_2，\cdots，u_n\}$为研究对象的 n 因素(或指标),称之为因素集(或指标集)。$V=\{v_1，v_2，\cdots，v_n\}$为诸因素(或指标)的种评判所构成的评判集,它们的元素个数和名称均可根据实际问题的需要和决策人主观确定.实际中,很多问题的因素评判集都是模糊的,因此,综合评判应该是 V 上的一个模糊子集 B=$(b_1,b_2,\ldots,b_k)\in$F(V)，其中b_k为评判v_k对模糊子集 B 的隶属度[11]：

$$\mu_B(v_k)=b_k(k=1,2,\ldots,m) \tag{5-4}$$

即反映了第 k 种评判v_k在综合评价中所起的作用.综合评判 B 依赖于各因素的权重,即它应该是 U 上的模糊子集 A=$(a_1,a_2,\ldots,a_i)\in$F(U)，且$\sum_{i=1}^{n}a_i=1$,其中a_i表示第 i 种因素的权重。于是,当权重 A 给定以后,则相应地就可以给定一个综合评判 B[12]。

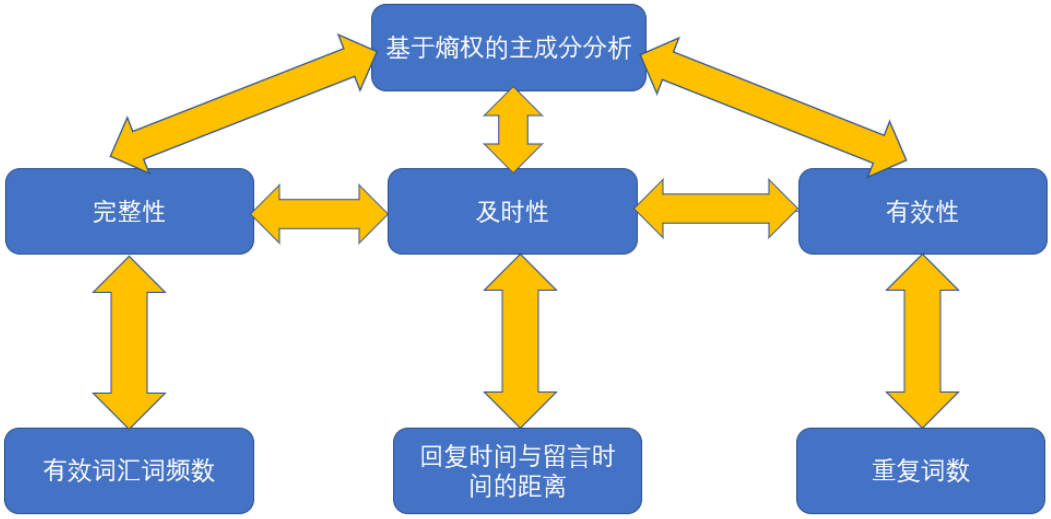

图 5-2 主成分分析和熵权法示意图

熵权法是通过计算熵值来确定各个指标的离散程度。熵值越大，说明该指标的离散程度越小，而对综合分析的作用越小，则熵值越小。我们通过计算评价指标的信息熵来获取结果：

$$e_j=-k\sum_{1}^{m}y_{ij}\ln y_{ij} \tag{5-5}$$

其中e_j为第 j 项指标的信息熵，k=1/lnm，$y_{ij}=x'_{ij}/\sum_{1}^{m}x'_{ij}$为第 i 年第 j 项指标值的比重。由此我们可以计算评价指标的熵权[13]

$$w_j=（1-e_j）/（n-\sum_1^n e_j）\tag{5-6}$$

式中，w_j为熵值法确定的第 j 项指标的权重系数，$0\leq w_j\leq 1$，且$\sum_1^n w_j=1$。

5.3 处理方案及结果分析

由题意我们需要综合较多因素，由于影响因素太多，所以我们首先选择主成分分析来将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量，从而达到降维的目的。数据处理步骤如下：

1、对指标进行无量纲处理

对所选指标作归一化处理，本文选取极差标准化法：

$$x'_{ij}=0.9*\frac{x_{\mathrm{ij}}-\min\{x_{\mathrm{ij}}\}}{\max\{x_{\mathrm{ij}}\}-\min\{x_{\mathrm{ij}}\}}+0.1\tag{5-7}$$

其中x'_{ij}表示附件 4 中第 i 行的第 j 项指标的标准化值，x_{ij}为第 i 行的第 j 项指标的原始值，其中 i=1，2,…为第 i 条回复意见；j=1，2，3 分别完整性、有效性和及时性指标。

2、计算相关系数矩阵 R

相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对主成分分析是非常重要的[14]。

$$r_{pq}=\frac{\sum_{k=1}^n \widetilde{a_{kp}}\cdot\widetilde{a_{kq}}}{n-1}\tag{5-8}$$

其中r_{pq}是第 p 个指标与第 q 个指标的相关系数。三个主成分的相关矩阵如下：

$$R=\begin{bmatrix}1.0000 & 0.4912 & 0.0281\\ 0.4912 & 1.0000 & 0.0343\\ 0.0281 & 0.0343 & 1.0000\end{bmatrix}\tag{5-9}$$

3、计算相关系数矩阵 R 的特征值和相应的特征向量

由下式计算判断矩阵的特征根

$$\lambda=\frac{1}{n}\sum_{i=1}^n\frac{(AW)_i}{w_i}\tag{5-10}$$

再结合之前熵权法的计算结果，我们得到下表。

表 5-1 特征根及贡献率

特征值	贡献率	累计贡献率
3.56	0.27	0.27
1.90	0.36	0.63
1.06	0.37	1.00

我们设留言与答复设时间差倒数为 X，留言详情和答复意见同样的词出现的个数为 Y，答复意见出现次数大于 2 的个数为 Z，3 个指标权重为 P1，P2，P3，则打分模型为 P1*X+P2*Y+P3*Z。以表 5-1 中三个主成分的贡献率为权重，构建主成分得分评价模型

$$S = 0.27p_1 + 0.36p_2 + 0.37p_3 \tag{5-11}$$

所有评价得分的分布图如图 5-3 所示，由图可知，得分在区间[0,0.5]、[0.5,0.8]和[0.8,1]的比例分别为 40.2%、48.9%和 10.9%，即大多数回复质量处于中等水平，少部分评价较低，较少部分评价优良。

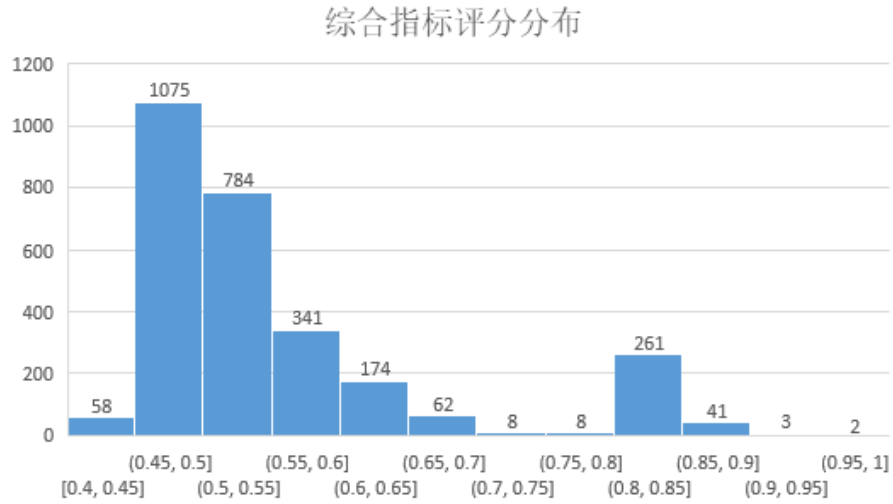

图 5-3 综合评分分布图

我们可以用熵值法确定熵值与权重后得到各个样本的综合值，这种做法比仅仅使用主成分分析法更客观。但相关性检验中，我们发现上述所处理的数据中变量之间相关度不高。由于主成分分析存在只能处理线性问题的不足，在研究实际问题时，主成分与指标之间经常存在非线性的关系，变量之间相关度不高，降维效果不明显。因此，可以对主成分分析法进行改进。在这种情况下，我们可以先绘制原始数据的散点图，根据散点图的特征，可对原始数据作对数变换，对数中心化变换或是平方根变换，来更好地达到降维的效果[15]。

参考文献

[1] 李牧南,王雯殊.基于文本挖掘的人工智能科学主题演进研究[J/OL].情报杂志:1-7

[2] 朱军,胡文波.贝叶斯机器学习前沿进展综述[J].计算机研究与发展,2015,52(01):16-26.

[3] 叶丹. KNN 文本分类及特征加权算法研究[D].湖南大学,2014.

[4] 李建民, 张钹, 林福宗. 支持向量机的训练算法[J]. 清华大学学报(自然科学版), 2003, 043(001):120-124.

[5] 李娜娜, 施龙青, 李忠建,等. 基于系统聚类法的底板突水危险性评价[J]. 矿业安全与环保, 2011, 038(002):28-30.

[6] 陈伟清,陆恩旋,曾弋戈,秦云江.基于灰色关联理论和系统聚类分析的智慧城市政府数据开放水平评价研究[J].数学的实践与认识,2020,50(06):43-52.

[7] 孙石磊,王超,赵元棣.基于轮廓系数的参数无关空中交通轨迹聚类方法[J].计算机应用,2019,39(11):3293-3297.

[8] 张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[9] 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.

[10] 曲双红,李华,李刚.基于主成分分析的几种常用改进方法[J].统计与决策,第 329 期:155-156,2011.

[11] 宋品芳,李孜军,李蓉蓉,赵淑琪,徐宇.基于熵权模糊法的高海拔矿井风机性能影响因素分析[J/OL].黄金科学技术,2020(03):1-12

[12] 殷涛,谢雄刚.矿井内因火灾的安全评价研究——基于熵值法和突变理论[J/OL].矿业安全与环保:1-5.

[13] 王珂,郭晓曦,李梅香.长三角大湾区城市群生态文明绩效评价——基于因子分析与熵值法的结合分析[J].生态经济,2020,36(04):213-218.

[14] 杨尚祯,朱小梅.湖北省农村居民家庭金融福利的测算——基于主成分分析法[J].现代商贸工业,2020,41(13):24-27.

[15] 蒋晨琛,霍宏涛,冯琦.一种基于 PCA 的面向对象多尺度分割优化算法[J/OL].北京航空航天大学学报:1-17.