泰迪杯数据挖掘 C 题论文

摘要

社交媒体逐渐成为政府提供公共信息服务及其与公众沟通的新渠道，对社交媒体相关留言的挖掘有助于“互联网+”背景下政府精准服务的实现和网络舆情管理。本文利用自然语言处理相关技术，针对留言文本分类、热点问题挖掘、留言答复评价这三个具有很强现实意义的问题，给出了比较切实可靠的方案，方案具有较好的泛化能力、精确性、合理性、可实现性。对建立智慧政务平台具有一定的参考意义。

对于问题一，我们首先对给定的文本进行数据预处理，去除无意义的字符，利用 jieba 中文分词工具对留言详情文本进行分词；后将得到的分词结果进行 TF—IDF 向量化表示；接着对不平衡的一级标签样本进行数据不平衡处理以提升精度；最后选择以 Linear 为核函数的支持向量机算法来进行分类，将训练好的模型应用到测试集得到预测值并进行比对分析。

对于问题二，首先将留言详情文本进行向量化表示；然后向量化的文本进行聚类并贴上标签；接着我们针对问题构建热度评价模型；再然后对按热度顺序输出的聚类进行检查，剔除其中效果较差的类别并由后面的热点问题顺位递进。

对于问题三，首先我们选择从回答格式、单词数、相似度、法律条文引用以及效率五个方面评分并对各个方面进行权重分配；接着分别从每个方面进行详细的分值计算；最后对得到的分数分析找出不合理的计算方面（相似性）并对其进一步改进，最终得出符合正态分布的得分。

关键词　分词；TF-IDF；聚合聚类；降维；层次分析法

Abstract

Social media has gradually become a new channel for government to provide government to provide public information services and communicate with the public. The mining of comments on social media is conducive to the realization of accurate government services and the management of online public opinions in the context of "Internet +".In this paper, with the help of natural language processing technology, a more practical and reliable scheme is presented, which has better generalization ability, accuracy, rationality and realizability.It has certain reference significance to establish the platform of intelligent government affairs.

For question one, we first perform data preprocessing on the given text to remove meaningless characters, and use the jieba Chinese word segmentation tool to segment the message detail text; then the resulting segmentation results will be TF-IDF vectorized; then The unbalanced first-level label samples are processed with unbalanced data to improve accuracy. Finally, the support vector machine algorithm with Linear as the kernel function is selected for classification, and the trained model is applied to the test set to obtain predicted values and comparative analysis.

For question two, first vectorize the message details text; then cluster the vectorized text and label it; then we construct a heat evaluation model for the problem and check the clusters output in order to remove The less effective categories are followed by the subsequent hot issues.

For question three, first of all, we choose to score and assign weights to the five aspects of answer format, word count, similarity, legal provisions, and efficiency; then we will calculate the scores from each aspect separately; The analysis of the scores finds unreasonable computational aspects (similarity) and further improves them, and finally obtains a score that conforms to the normal distribution.

Keywords: Participles;TF - IDF;Polymer clustering;Dimension reduction;Analytic hierarchy process

目录

1 绪论

1.1 研究背景

随着新公共管理运动的推动和展开，电子政务逐渐兴起，进而对政府传统的管理模式产生了冲击，电子政务将促进社会的整体创新，不仅仅包括技术，还包括制度，管理，理念，体系创新，目的是从根本上提高政府的服务水平和实现流程再造。社交媒体平台可以集合不同地域，不同空间领域的人的意见，扩大公民参与政务讨论的范围，并引导舆论的流向。

近年来，随着微信、微博、市长信箱等平台的广泛使用，涉及公民反映的社情民意相关的文本数据量呈几何级上升，这对主要依靠人工进行划分整理留言问题的相关部门是极大的挑战。

随着大数据，云计算，人工智能等技术的发展，建立基于自然语言处理技术的智慧政务平台已经是创新社会治理的新方式。

2 问题一的求解

2.1 数据预处理

数据预处理是指在对数据进行数据挖掘主要的处理以前，先对原始数据进行必要的清洗、集成、转换、离散和规约等等一系列的处理工作，以达到挖掘算法进行知识获取研究所要求的最低规范和标准[1]。

文本预处理是文本分类的基础，是实现分类必不可少的步骤。它包括中文分词和去除停用词两个过程。

为了在程序中更好的读取数据，对给出的附件 2 中"留言详情"列的数据进行预处理，将换行符，空格符号删除。另外注意到，数据中包含的一些词语，比如：某个小区，某市长，出现的地点等等都对后面模型建立的分析没有决定性作用，遂将其删除。

中文分词，指将一句用汉语表达的句子分解成单个词（成语）或词组。目前常用的分词系统有 ICTCLAS、结巴分词、Boson NLP、百度 NLP、腾讯文智，阿里云 NLP 等[2]。这些分词系统的研究已经较为成熟完善。本次应用的是开源的 jieba 分词。

停用词，指使用十分广泛甚至是过于频繁的一些单词，比如英文的”i”、”is”、”what”,中文中的”我”、”就”之类的词几乎在每个文档上都会出现。此外还有文本中出现频率很高，但实际意义不大的词，主要包括语气助词、副词、介词、连词等。去除之后对文章语义不会产生影响。去除停用词能更好的分析文章语义，提高文本分类的正确率。

2.2 生成词云图

“词云”就是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词云图”。对所有的留言详情中的内容进行预处理之后，形成词云图，能够直观的观察出关键词的分布，为每个标签的分类做准备。对 9210 条留言进行分词后，获得的词云图如下:

图 1. 附件二“留言详情”词云图

从上面的图中可以看出领导、政府、公司、学校等词出现的评率最高，经过进一步的统计得出排名前十的高频词汇为:领导 5447 次，政府 4601 次，公司 4084 次，学校 4027 次，部门 3405 次，相关 3103 次，西地省 2906 次，国家 2797 次，情况 2714 次，学生 2641 次。从统计的结果来看，大致可以看出已有几类分类比较明确，如公司等可以分为城乡建设一类，而学校，学生等可以分到文体教育一类。

2.3 文本向量化表示

文本表示是对文本分类中对文本结构化的过程。经过预处理之后的文本，计算机并不能识别，需要将文本语言转化成自然语言，即将文本向量化表示。本次使用的是 TF-IDF 向量化表示。

TF 表示每个文档的词频，IDF 表示逆文档频率。

$$TF = \begin{cases} \text{某个词在文章中出现的次数} \\ \dfrac{\text{某个词在文章中出现的次数}}{\text{文章的总次数}} \\ \dfrac{\text{某个词在文章中出现的次数}}{\text{该文中出现次数最多的词的出现次数}} \end{cases}$$

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

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

在 python 中，附件二中的“留言详情”经过 TF-IDF 转化后的矩阵维度多达 70000 多维，显然需要一定程度的降维处理。因此我们取最大的特征维度为 10000 维，即 max_features=10000，因此规模为 9210×10000。

2.4 数据不平衡处理

通过统计发现，一级标签中各个类别所含样本的个数不同，可能会在分类时造成多个标签误分类为同一个类别，所以在分类之前预先进数据不平衡预处理。本次采用处理方法为 SMOTE 方法。在进行训练集和测试集划分之后，使用 SMOTE 方法，使得每个类别在采样时是均匀的。

在本次实验中，发现“城乡建设”类的留言有 2009 条，“环境保护”类的留言有 938 条，“交通运输”类的留言有 613 条，“教育文体”类的留言有 1589 条，“劳动和社会保障”类的留言有 1969 条，“商贸旅游”类的留言有 1215 条，“卫生计生”类的留言有 877

条。显然最多的“城乡建设”类与“交通运输”类之间存在着数据的不平衡，如果不进行处理，那么“城乡建设”与“劳动和社会保障”可能存在过拟合的现象，在验证时其他类的留言会被错误的分为这两类。

2.5 标签分类

进行一级标签的划分的最为关键和主要的部分为分类算法。对待测文本，即留言详情进行分类预测，最终实现一级标签的划分。目前常见的分类算法有朴素贝叶斯 NB，线性最小二乘拟合法，K 最近邻 KNN 方法以及神经网络方法[2]。SVM 是一种不具备先验知识的条件下，以最小化结构风险为目标，对有限样本进行学习的机器学习方法。SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而非样本空间的维数，这在某种意义上避免了“维数灾难”，同时 SVM 学习问题是一个凸优化问题，因此一定可以得到全局最优解，此外基于结构风险最小化原则可以避免过度学习问题，泛化能力很强。故本次选用的分类方法为以 Linear 为核函数的支持向量机算法。

非线性支持向量机学习算法[4]：

输入：训练数据集$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$，其中$x_i\in\chi=R^n, y_i\in\Upsilon=\{-1,+1\}$，i =1,2,…,N:

输出：分类决策函数。

（1） 选取适当的核函数$K(x,z)$（这里为线性核），和适当的参数 C，构造并求解最优化问题

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$0\le\alpha_i\le c, i=1,2,\dots,N$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\dots,\alpha_N^*)^T$。

(2) 选择α^*的一个正分量$0\le\alpha_j\le c$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^*\,y_i K(x_i,x_j)$$

(3) 构造决策函数：

$$f(x)=sign\left(\sum_{i=1}^{N}\alpha_i^*\,y_i K(x,x_i)+b^*\right)$$

2.6 结果分析

第一问的步骤流程图如下:

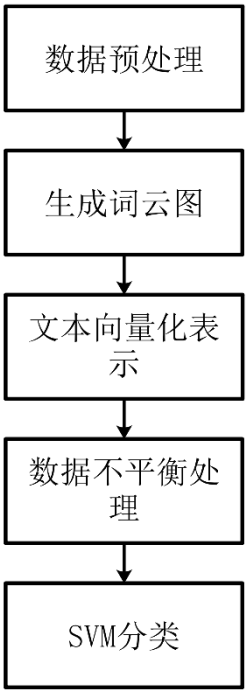


图 2 问题一求解流程图

将全部的数据 70%设为测试集，30%设为验证集，验证集数据共 2763 条，在进行数据不平衡处理后，利用支持向量机 SVM 进行分类后，得到验证集的精度为 90.5%。最终得到的 F-Score 结果为:

表 1 一级标签分类的 F-Score

一级标签	F-Score
交通运输	0.85
劳动和社会保障	0.95
卫生计生	0.92
商贸旅游	0.84
城乡建设	0.88
教育文体	0.93
环境保护	0.93
平均	0.90

对应的混淆矩阵为:

$$\begin{bmatrix} 153 & 2 & 0 & 9 & 12 & 1 & 0 \\ 1 & 566 & 7 & 3 & 9 & 10 & 0 \\ 0 & 10 & 230 & 7 & 3 & 1 & 0 \\ 6 & 2 & 11 & 298 & 20 & 10 & 1 \\ 18 & 10 & 2 & 35 & 555 & 7 & 21 \\ 4 & 11 & 1 & 5 & 8 & 431 & 1 \\ 1 & 0 & 0 & 7 & 7 & 1 & 266 \end{bmatrix}$$

从混淆矩阵中可以看出:“交通运输”和“商贸旅游”类的 F-Score 最低，其被错误的分为“城乡建设”的数目最多，主要原因可能是因为其中出现了大量的相同词汇，如下表所示：

表 2 三个“一级标签”前 10 词汇表

交通运输		城乡建设		商贸旅游	
词汇	词频	词汇	词频	词汇	词频
出租车	557	电梯	968	业主	1908
公司	341	公司	740	政府	1333
快递	308	部门	606	领导	1024
司机	272	业主	539	开发商	1013
政府	252	相关	523	部门	919
领导	219	传销	480	房屋	903
部门	198	政府	456	规划	827
的士	186	领导	439	居民	808
希望	185	旅游	418	相关	788
车辆	176	希望	383	公司	774

“交通运输”类中的“政府”“领导”“公司”与“城乡建设”中的重合，“商贸旅游”类中的“相关”“领导”“政府”“公司”“部门”与“城乡建设”中的重合，且词频出现了很大的差异，所以“交通运输”类与“商贸旅游”类的 F-Score 的值较低。

但是最终的平均 F-Score 高达 0.9，分类准确率也高达 90.5%，所以 SVM 最终的分类效果还是比较可观的。

3 问题 2 求解

附件三中的文本信息有两列：留言主题与留言详情。因为留言主题内容文本信息较少，很难用于文本聚类，所以使用留言详情解决问题 2。
解决思路如下：

（1） 将留言详情文本进行向量化表示

（2） 对于向量化后的文本，利用聚类算法进行聚类，并贴上相应的标签。

（3） 构建热度评估计算方法

（4） 将聚类后的每个类看作一个热点，对其进行热度计算，并按降序排列输出。

（5） 检查输出结果是否存在聚类效果较差的类，并判断是否要将其剔除或是需要进一步聚类。若剔除，则由后面的热点问题进行顺位递进。

（6） 重复步骤（5）至输出 5 类热点及其明细。

3.1 文本表示

3.1.1 分词

在市民留言中，往往会使用尊称或问候语，如“尊敬”、“领导”、“您好”和“谢谢”等，这类词语并不能反应文本内容的具体特征，所以将其去掉。
另外，值得注意的是，在以往的文本处理中，可能会将英文字母或是数字进行剔除，但是在本次文本中，英文字母往往有重要的地点信息，如 A 市、B 县等，这类地名可能能够反应热点问题的特定地点，有助于聚类效果的提升，所以将其保留。

3.1.2 向量化表示

常用的文本向量化表示方法有：词频矩阵、TF-IDF 矩阵、word2vec 等。词频矩阵一般用于样本数目较少的任务，而 word2vec 是将每个文本看成一系列单词的序列，然后将每个文本表示成一个矩阵。通常来说 word2vec 精度更高，但是训练所需要的开销较大。受限于硬件，我们小组采取的依然是与第一问相同文本向量化方法：TF-IDF 矩阵。

3.2 文本聚类

对于文本聚类，常用的方法有 K-means 算法和层次聚类法等。K-means 算法在聚类算法中有着较快的速度，但是理论上来讲，K-means 的最优求解问题是 NP 困难问题，现实中采用迭代方式进行求解，这就不可避免地容易陷入局部最优。同时，目前 python 中主流的机器学习模块 scikit-learn 中的 K-means 算法只能使用欧式距离，而在文本聚类中常用的是余弦距离。层次聚类的优点是操作简单，且适用于任意形式的相似度或距离表示形式，缺点是运算速度慢。这次的数据数量并不算很大，时间开销并不是特别长，所以我们小组最终选择的是层次聚类中的聚合聚类方法。

聚合聚类算法如下[4]：

输入：n 个样本组成的样本集合及样本之间的距离；

输出：对样本集合的一个层次化聚类

（1） 计算 n 个样本两两之间的距离

（2） 构造 n 个类，每个类只包含一个样本

（3） 合并类间距离最小的类，其中最小距离为类间距离，构建一个新类

（4） 计算新类与当前各类的距离。转至步骤（3），直至满足终止条件

一般常用欧式距离表示两个样本间的距离度量，用新类内样本各维度的均值作为新类的中心点。但是对于文本聚类而言，每个样本的向量是个高维向量，用夹角余弦表示两样本间的距离更合适。当采用夹角余弦作为类间距离度量时，若两个不同的类内含有多个样本，则用这两个类内样本的夹角余弦的均值作为这两个类的类间距离。

举个例子。如图 1 所示，在二维平面里有 5 个样本，其中样本 1、2、3 属于 A 类，样本 4、5 属于 B 类。为了计算 A 类与 B 类的夹角余弦，则需要计算类 A 内所有样本与类 B 内所有样本的夹角余弦，即$\{cos_{ij}\}, i \in \{1,2,3\}, j \in \{1,2\}$, cos_{ij}表示 A 类第 i 个样本与 B 类第 j 个样本的夹角余弦。那么 A 类与 B 类夹角余弦为：$cos_{A,B} = \frac{1}{3\times2}\sum_{i,j} cos_{ij}$

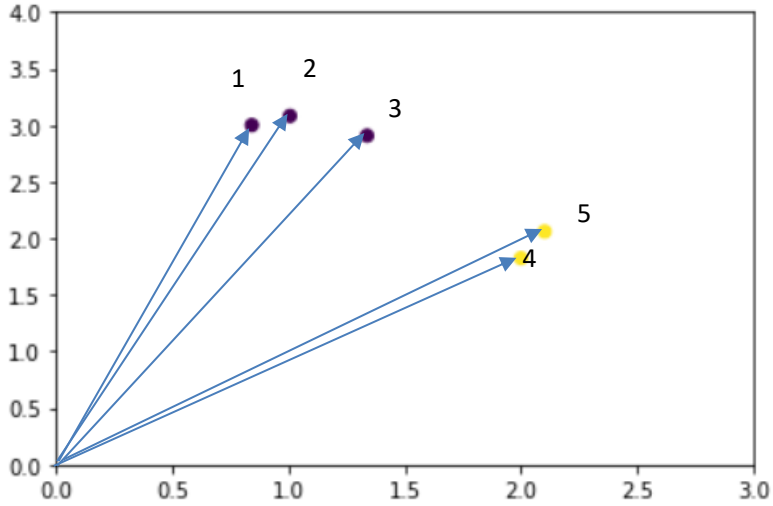


图 3

聚合聚类的最佳类数目可由实验得到。根据以往的经验，聚类数目先取数据集样本数目的一半，观察聚类结果再做进一步的调整。这里先取 2000，聚类完成后，发现聚类效果还不错，接着在 2000 左右再进行聚类。最后聚类数目确定为 2200。

3.3 热度指标

3.3.1 热度评分的相关因素

关于热度指标的建立可以从下面三个方面思考：热点问题的样本数目、热点问题的点赞反对数以及该问题讨论激烈程度。

热点问题的样本数目，即类内样本数目，显然应该与热度评分成正相关关系。当类内数目越多时，热度评分应该越高。

热点问题的点赞反对数反映了问题的市民关注程度，同样也应该与热度评分成正相关关系。这里，我们不区分市民对待问题的态度，即不区分市民是赞成还是反对，只关注热点问题的点赞与反对之和。

关于问题的讨论激烈程度的衡量，我们采用市民对待问题的态度进行衡量。当热点问题的点赞数与反对数非常接近时，则表示该热点问题的讨论激烈程度越高。

3.3.2 热度评分公式

根据上面三个方面建立以下热度评分公式：

$$score = \left(k_1 \times \frac{n}{N} + k_2 \times \frac{x+y+1}{M} \log\left(\frac{1}{|x-y|+1} + e\right)\right) \times 1000$$

其中：n 为某热点问题所包含的留言数目；N 为样本总数；x，y 分别为某热点问题所包含的所有点赞与所有反对；M 为所有的点赞与反对之和；k_1与k_2为待定系数，分别代表两项的权重，；e 为自然常数。

由上式可以看出，热度指标主要由两项组成：热点问题所包含的样本数目在所有样本中的占比以及该热点问题点赞反对在所有样本中的占比。第一项主要刻画热点问题在整体数据集中的重要程度，而第二项主要刻画市民对该热点问题的关注程度及讨论激烈程度。当一个热点问题的点赞反对总数高，且点赞与反对的人数十分接近时，就代表该热点问题讨论得越激烈。（x+y+1）表示的就是该热点问题的点赞反对总数（加上 1 的目的是防止 x+y=0 导致第二项整体为 0），|x-y|表示点赞人数与反对人数的差距。最后整体乘以 1000 是为了使得分能在较合理的区间。

3.3.3 待定系数的确定

关于权重k_1与k_2的确定，我们作了一简单假设，如下：

$$k1\frac{\text{最多的类内样本数目}}{\text{数据集样本总数目}} = k2\frac{\text{最多的类内点赞与反对之和}}{\text{数据集的点赞与反对之和}}$$

经过聚类结果分析，得到以下数据：

数据集样本总数	4327
数据集上点赞与反对之和	13022
最多的类内样本数目	47
最多的类内点赞与反对之和	2387

根据上述假设很容易得到$\frac{k_1}{k_2}$约为 16。为了防止权重分配误差过大，我们最后取$k1 = 15, k2 = 1$。

3.4 结果分析与改进

根据上面确定聚类数目及热度指标公式，输出热点问题明细表。经观察，发现有两个热点问题明细情况不是很好，分别是关于小区物业管理差和幼儿园入学难、学费高。

根据明细表显示，关于“小区物业管理差”这一主题的留言很多来自不同的小区，并不能反应出热点问题的特定地点这一特性。聚类算法之所以将这些留言聚为一类，是因为样本

中都大量出现了“小区”、“地处”等词。为了尽可能将这些样本区分开来，现将这些词在分词的时候就去掉。

同样的，关于幼儿园的留言也是来自不同县市的不同小区，且反映的问题也呈现多样化，有反映幼儿园入学难的，也有反映幼儿园学费高的，同样也有反映幼儿园非法办学的。经统计关于幼儿园的留言共 44 条，其中关于入学难的有 17 条，关于学费高的有 16 条，关于幼儿园非法办学的有 5 条，其他内容有 6 条。算法将这些留言聚为一类是因为这些留言都出现了“幼儿园”这一单词，而在构建 TF-IDF 矩阵时，在该单词所处维度赋予了较大的权重。

如果对关于幼儿园的留言再进行更细致的聚类，那么原关于幼儿园的 44 条留言将会被拆成上述几条内容，且需要对新的聚类结果重新进行热度计算，操作比较繁琐。

经过对热点问题明细表的观察，注意到关于幼儿园留言的点赞数较少，而排在幼儿园后面的热点问题的数目明显高于幼儿园留言拆分的数目，所以采用最直接的办法：将幼儿园这一聚类效果差的热点剔除，由剩下的聚类效果好的热点进位。

最后的输出结果如下:

表 3：热点问题明细表

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	239	2019-01-08 至 2019-07-08	58 车贷受害人	受害人要求发布案件进展
2	2	168.18	2019-05-22 至 2019-08-19	汇金路五矿万境	小区物业管理差
3	3	165.07	2019-07-02 至 2019-09-01	伊甸园	捆绑车位销售
4	4	142.82	2019-04-11 至 2019-09-25	梅溪湖金毛湾	金毛湾学位至今还没有着落
5	5	105.82	2019-11-13 至 2020-01-15	暮云街道丽发新城	搅拌站粉尘大、噪音大

4 问题 3 求解

4.1 权重分配

关于答复的评价，可以从相关性、完整性、可解释性以及时效性，这四个角度进行评分。

相关性指的是，答复与留言意见的相关程度；完整性指答复是否符合一定格式，整体是否完整；可解释性指答复内容是否有一定依据，如法律条文等；时效性则指隔了多久回复市民留言。

从人的主观来看，这四个角度并不是完全独立的，如相关性和可解释性就可以看作完整性的一部分，这就给这四个角度的权重划分带来了难度。为解决这个问题，我们将这四个角度再做进一步细分，共分成五个小角度：回答格式评分、单词数评分、相似度评分、法规条文评分和效率评分，并引入层次分析法，如下图所示。

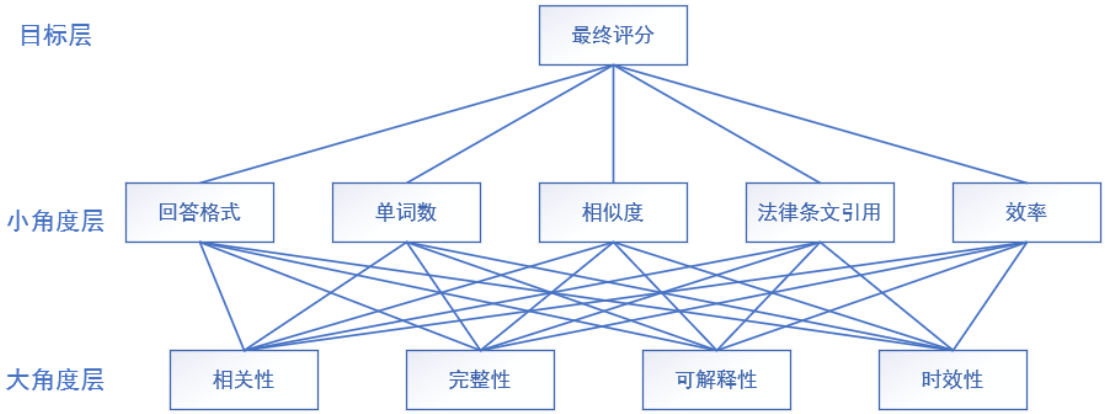


图 4：层次分析法

构建从小角度层到目标层的成对比较阵$A=[a_{ij}]_{5\times5}$，其中a_{ij}表示第 i 个小角度和第 j 个小角度对最终评分的影响之比。矩阵A是正互反阵，其内部值如下：

$$A=\begin{bmatrix}1.0000 & 0.3333 & 0.2000 & 0.5000 & 0.6667\\ 3.0000 & 1.0000 & 0.5000 & 1.2500 & 2.0000\\ 5.0000 & 2.0000 & 1.0000 & 2.5000 & 2.5000\\ 2.0000 & 0.8000 & 0.4000 & 1.0000 & 1.2500\\ 1.5000 & 0.5000 & 0.4000 & 0.8000 & 1.0000\end{bmatrix}$$

求得它的最大特征值为 5.0195，其对应的特征向量经归一化处理后为$w=\begin{bmatrix}0.08\\0.23\\0.40\\0.16\\0.13\end{bmatrix}$。检验其一致性：$\frac{\frac{5.0195-5}{5-1}}{1.12}=0.004<0.1$，通过一致性检验。

最终评分采取百分制，根据上述权重得回答格式占 8 分，单词数占 23 分，相似度占 40 分，法律条文引用占 16 分，效率占 13 分。

若想要得到各样本在四个大角度的评分，需要构建 5 个不同的大角度层到小角度层的成对比较阵，操作较繁琐，且题目要求给出评分即可，所以这一步骤省略。

4.2 固定格式评分（8 分）

答复的固定格式，即开头是否有简单的问候语（如：网友您好），文末是否有答复时间，如果满足这些要求则得 4 分；另外，也要判断答复是否意见是否给出一些建议或是否经过调查，如果满足，则得另外 4 分。这里因为牵涉到对特殊单词得查找，所以在分词前要先对停用词表进行简单得查看，剔除这些特殊单词。

4.3 字数个数评分（23 分）

先对分词后的整个数据集的各样本单词数做一下统计，表格如下：

单词个数区间	样本数	累计占比
<5	56	0.02
5-50	788	0.30
50-100	840	0.60
100-200	783	0.88
200-300	225	0.96
300 以上	124	1

简单起见，若某个样本的单词个数落入以上某个区间，则该样本的单词个数评分为其相应的累计占比与 23 分的乘积。

4.4 相似度评分（40 分）

相似度主要是用来体现答复与留言的相关性。为了计算相似度，需要将文本向量化表示。这里需要比较的是答复与留言的相似度，因为答复数据与留言数据属于不同的数据分布，所以不适合用前两问的 TF-IDF 矩阵表示方法。

我们采用类似 one-hot 编码的方式对文本进行向量表示。每个词所在维度的取值只能是 0 或者 1。如下面的例子所示：

Txt1：我家水管坏了，请报修。

Txt2：你家水管已报修。

去除停用词后的语料库：我家，水管，坏了，报修，你家。

	我家	水管	坏了	报修	你家
Txt1	1	1	1	1	0
Txt2	0	1	0	1	1

接着采用问题 2 聚类相似的方法，计算两者的余弦值。余弦值是处于[0,1]区间的，可以看作一个百分比。所以用余弦值乘以相似度权重（40 分），即得到相似度评分。

余弦值计算方法：

$$Cosine = \frac{x \cdot y}{|x| \cdot |y|}$$

其中 x,y 分别表示留言详情与答复意见的向量，|x|与|y|分别表示模长。

4.5 法律条文引用评分（16 分）

若答复意见是根据某条法律法规做出回应，则表示该答复意见的可解释性较强。同时注意，当文本引用某条法律法规时，往往会出现“根据”，“依据”，“遵守”，“依法”，“按照”等字眼，所以将这些词作为是否引用法律条纹的判断依据。每出现一次这样的词，可得 4 分。如果这类词出现 4 次及以上，则得满分。

4.6 效率评分（13 分）

通常而言，回复时间间隔越短代表解决问题的效率越高，经统计，回复时间差数据如下：

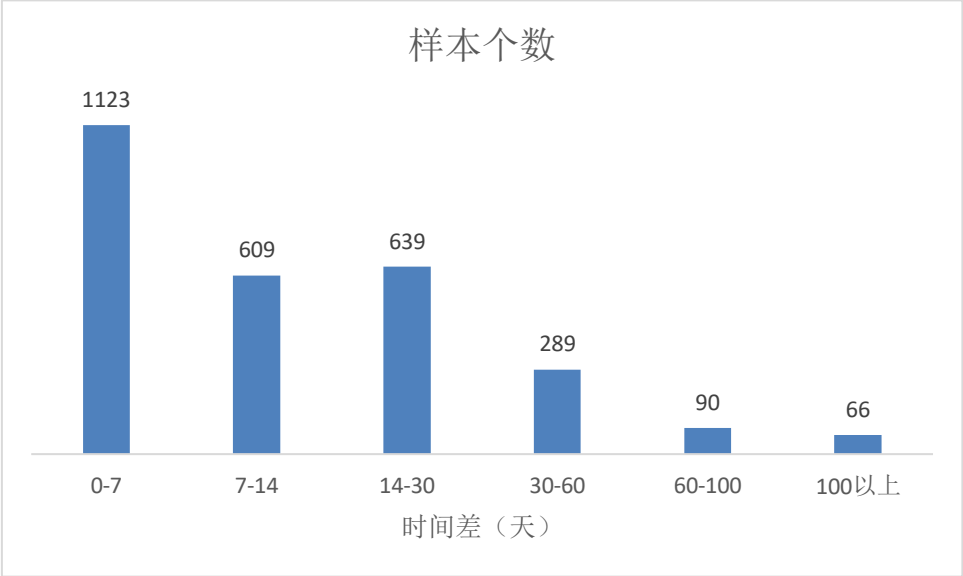


图 5：回复时间差柱状图

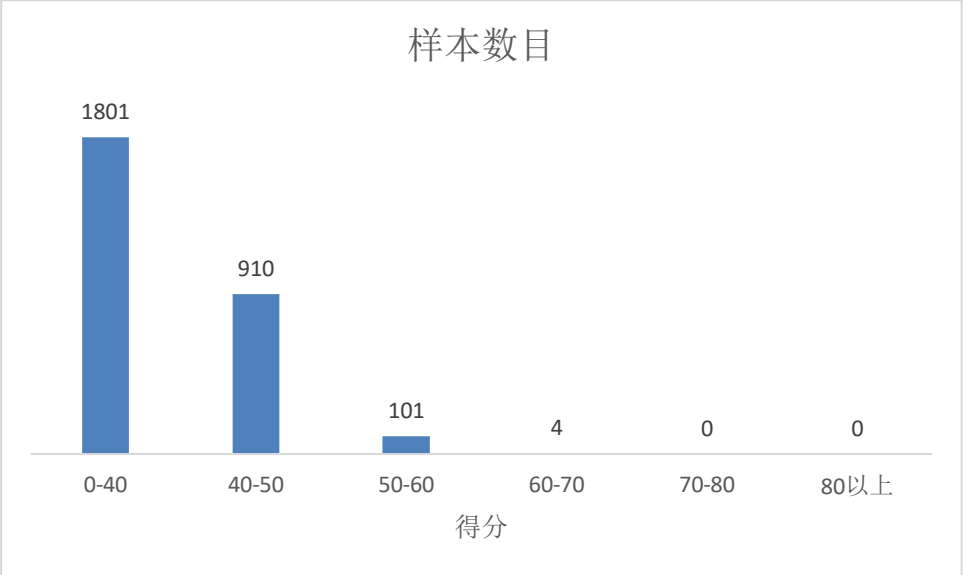
经观察，可以发现在 7 天内回复的样本数目最多，在 7-14 天内和在 14-30 天内回复的样本数量差不多。

我们给这几个时间段赋予相应的分数，如下：

时间差（单位：天）	效率得分
0-7	13
7-14	10
14-30	8
30-60	6
60-100	2
100 以上	0

4.7 结果分析与改进

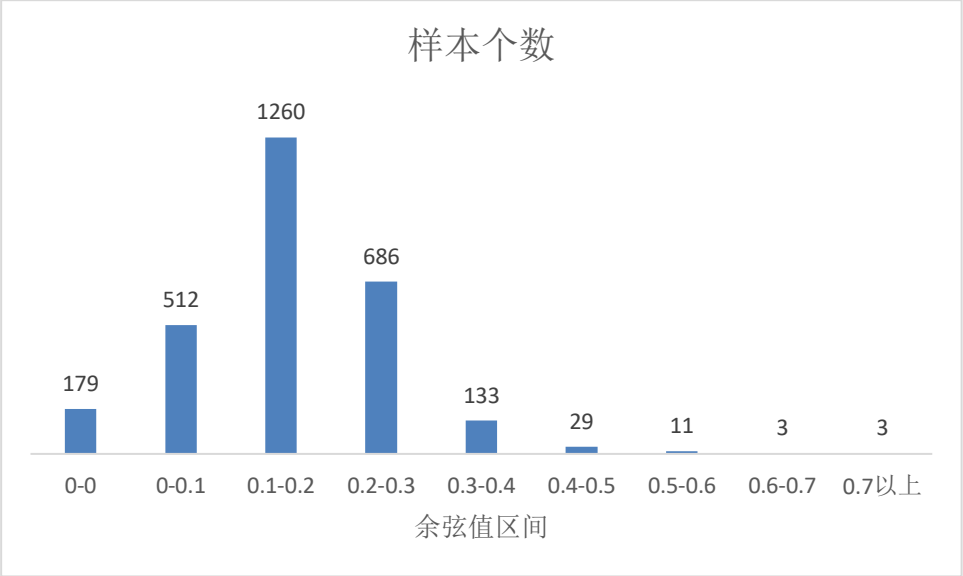
经过上述评分过程，得到以下得分分布：



图六：样本得分柱状图

很明显，样本的得分集中在 40 分以下。按照正常生活中的经验，得分应该呈一个正态分布，均值约为 70 分。而现在的结果是十分不合理的。

统计各样本的余弦值如下：

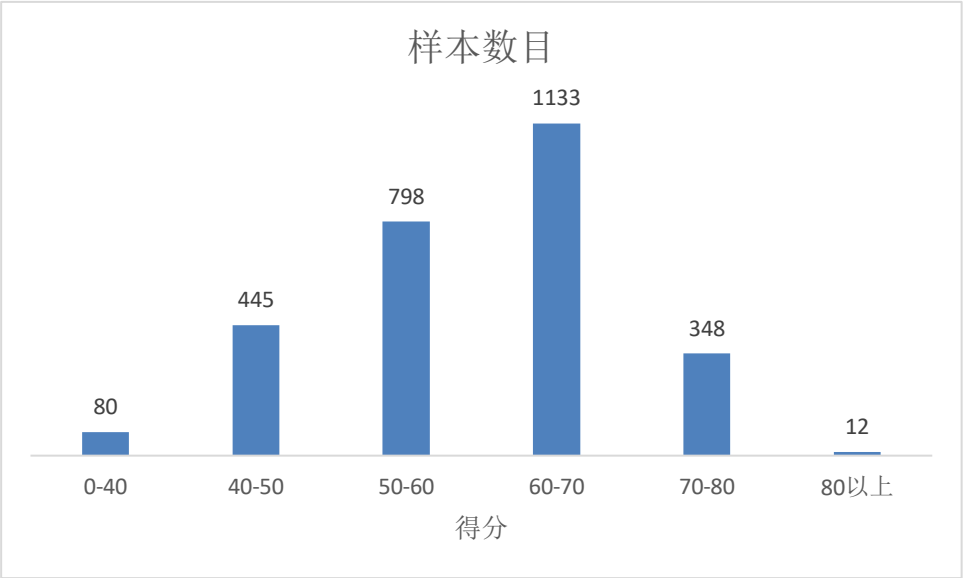


图七：各样本余弦值柱状图

观察上述结果，其实不难发现，样本余弦值普遍较小。这是因为只有当某一样本的留言与答复的单词相同的数量非常多时，才会使余弦值处于较合理的范围。但是在实际生活中，很难让答复与留言的单词相同，因为从数据分布来说，答复与留言属于不同的分布，所以它们的相似度普遍较小。同时，又由于相似度权重较大，导致得到了上述不合理的得分分布。

为了修正余弦值，我们采取了较简单做法：为原来计算得到余弦值加上一个修正值 ε，使得最后的得分分布成一个均值在 70 左右的正态分布。根据余弦值统计情况，仅有（29+11+3+3）=46 个样本的余弦值超过 0.4。我们假设这 46 个样本为相似度最好的样本，它们修正后的余弦值为 1，由此得到修正值 ε 为 0.6.修正后，会有极少部分样本的余弦值大于 1，可理解为这些样本相似度非常好从而获得了附加分。

修正后的得分分布如下：



图八：各样本修正余弦值柱状图

可以看出，修正后的得分符合实际情况。

5 参考文献

[1]方洪鹰. 数据挖掘中数据预处理的方法研究[D].西南大学,2009.
[2]吴萍萍.基于信息熵加权的 Word2vec 中文文本分类研究[J].长春师范大学学报,2020,39(02):28-33.
[3]段国仑,谢钧,郭蕾蕾,王晓莹.Web 文档分类中 TFIDF 特征选择算法的改进[J].计算机技术与发展,2019,29(05):49-53.
[4]李航 统计学习方法[M]