第八届“泰迪杯”数据挖掘挑战赛

题 目 “智慧政务”中的文本挖掘应用

摘 要:

近年来，随着网络的日益普及，“网络问政”作为一种新型的议政问政方式，已经成为政府了解民情、听取民声、体察民意、汇聚民智的一个桥梁，本文利用自然语言处理和文本挖掘方法完成群众留言分类、热点问题挖掘及答复意见的评价。

针对任务一，对文本数据进行 jieba 分词、采用哈工大及人为添加词库去停词、利用 TF-IDF 提取关键词并用 Doc2vec 将文本关键词向量化，建立了基于 XGBoost 算法的分类模型，采用网格搜索进行局部调参，得到最佳的分类模型，其F_1值为 0.87。

针对任务二，引入基于条件随机场序列标注的命名实体识别方法将地点和人群作为关键词提取出来，并将其放入关键词列表中。建立了基于 K-means++算法的热点问题确定模型，利用手肘法确定了最优聚类数目，实现相似留言聚类，并通过问题关注度、问题支持度与问题集中度三个方面对留言问题进行热度评估，解决了任务二中热点问题的排名问题。

针对任务三，本文提出了一种客观有效的答复意见质量分级方法。从答复意见的相关性、完整性、时效性三个角度选取了相关指标，对指标进行量化以及归一化处理。提出了基于层次分析法的答复意见质量综合评价模型。通过对答复意见文本质量的评估，评选了前十个高质量答复意见。

关键词： TF-IDF Doc2vec XGBoost 算法 命名实体识别 K-means++聚类分析 层次分析法

1 问题重述

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各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。同时，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系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 问题提出

1.2.1 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分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。建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

1.2.2 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。总结出排名前 5 的热点问题，并给出相应热点问题对应的留言信息。

1.2.3 答复意见的评价

针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案并实现。

2 问题假设与符号说明

2.1 问题假设

(1) 假设附件中数据真实可靠；

(2) 假设写问题留言的人不给自己点赞。

2.2 符号说明

表 2-1 符号说明

符号	说明
F_1	模型评价值
C	热度值
k	聚类个数
c_i	聚类中心的序列

3 问题分析

3.1 任务一的问题分析

建立关于留言内容的一级标签分类模型本质上是个文本分类问题，我们需要对附件二的数据进行分割。并对文本进行向量化表示和类别映射。将文本的分类问题转化为常规的机器学习分类问题。从而集合文本表示和分类建模两部分，共同组成一个整体通用模型，以满足任务一提出的现实需要。技术步骤主要包括 j 分词、构建停词表（基于哈工大停词库加人为添加词汇）、去停词、TF-IDF 提取特征、dco2vec 模型训练、文本向量的预测、XGBoost 分类算法建模、网格搜索调参等。

3.2 任务二的问题分析

考虑到热点问题往往与“地点/人群”非常相关，因此在分词和文本的关键词表示的基础上，引入基于条件随机场序列标注的 NER 方法提取地点和人群关键词，并将其添加进文本并使用上一问得到的 doc2vec 模型预测文本向量。这样训练出来的文本向量对地点和人群的特征选择有更大的权重。随后使用 K-means++ 算法对热点问题分类，同时建立类别的热度评估模型，基于此进行类别热度排序。最终可得到热点问题表和热点留言表等。

3.3 任务三的问题分析

针对答复意见质量的评定标准主观性过强、考虑指标单一的问题，本文提出一种客观有效的答复意见质量分级方法，即 AHP 层次分析法。

从三个方面衡量答复意见的质量，相关性、完整性与时效性。答复的相关性是指答复意见的内容是否与问题相关；答复的完整性是指是否满足某种规范；答复的时效性是指是否在短时间内及时进行答复。考虑质量评价方案如下：首先确定评价指标，接着对评价指标进行量化、归一化处理，采用层次分析法赋予权重，最后计算出评价结果。

4 任务一的模型建立

4.1 文本预处理

处理文本数据前需要对文本数据进行分词、去停用词、关键词提取等处理。

4.1.1 中文分词

中文文本中词与词之间没有明显的分隔符，为有效理解文本，第一步需要进行分词。中文分词算法主要有基于字符串匹配、基于理解和基于统计。

本文采用当前广泛使用的基于 Python 的中文分词工具 jieba，jieba 中文分词工具内置多个算法，支持多种模式进行分词，能有效解决未登陆词和歧义词。选择使用精确模式对附件 2 中“留言主题”与“留言详情”文本内容进行分词，精确模式将句子最精确地切开，适合文本分析。

4.1.2 去停用词

停用词的过滤可以有效降低待处理文本的维度，剔除掉无用的词汇，只保留可以表示文本的词语，从而提高词语搜索的效率[1]。停用词的处理对文本分类的准确性也有很大的帮助[2]。在一般实验应用中，存在一个存放停用词的集合，叫做停用词表，其中的停用词往往由人工根据经验知识加入，具有通用性[3]。

在分词步骤结束后，新的文本成为词的组合，但是并不是所有的词都与文本内容相关，这些词往往是语言中一些表意能力很差的辅助性词语，比如中文词组中类似“的”,“得”,“我们”,“了”等词汇。除此之外，一些标点符号，比如省略号，下划线，破折号等，也需要进行处理。这些词汇和符号往往存储在停用词表的文档中，可以参照停用词表对文本进行停用词处理。

本文建立停用词表，在哈工大停用词词库的基础上加了一些与文本相关的无用词，如“尊敬”、“领导”、“谢谢”等，采用字符匹配的方式扫描分词词典进行删除。

4.1.3 关键词提取

根据数据的标注情况，可以将关键词提取技术分为无监督学习方法和有监督学习方法。目前比较常用的关键词提取算法都是基于无监督算法。基于无监督学习的方法将关键词提取看作一个排序问题，这类方法主要利用词语的统计特征或语法规则对其进行重要性排序。经典的基于无监督的关键词提取算法包括 TF-

IDF，TextRank，YAKE 等。

TF-IDF 算法中包含 TF 和 IDF 两部分。

TF 即词频，表示词语在某一篇文档中出现的次数，通过计算词频可以有效获得全文的重要词汇。TF 的数学公式表示为：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中，$n_{i,j}$表示该词语t_i在某一文本d_i中出现的次数；$\sum_k n_{k,j}$是该文本d_j中所有出现过的词语次数k的总和。

IDF 即逆文本频率，是通过使用集合中的总文本数除以包含某一词语的文本数量并取对数获得，IDF 表示某一词的特异性。若某一词语在整个文本集合中的大部分文章中出现，那么这个词语的特异性较小，即 IDF 值较小。反之，IDF 值较大，该词语特异性较强。IDF 的数学公式表示为：

$$idf_i = log\frac{|D|}{1 + |\{j: t_i \in d_j\}|} \tag{2}$$

其中，$|D|$为语料库中文档的个数，分母表示包含词语t_i的文件个数，也就是$n_{i,j} \neq 0$的文件个数，分母加上 1 避免了词语不在语料库中从而导致分母为 0 的情况。

TF-IDF 实际上就是将 TF 的值和 IDF 的值相乘，将某一文本内的高频词且在整个数据集中出现频率较低的词语筛选出来，得到高权重的 TF-IDF 值：

$$tf - idf = tf_{ij} \times idf_i \tag{3}$$

本文采用 TF-IDF 算法，利用 jieba 分词系统中的 TF-IDF 接口抽取“留言详情”分词、去停用词后的关键词。再将“留言主题”去停用词结果与“留言详情”关键词去除重复数据后合并，得到初始的文本特征集合。

4.2 文本特征提取

4.2.1 Doc2Vec

2013 年，Google 开源了 Word2vec 算法[4]，该算法通过神经网络模型实现词语的向量表示模型。2014 年，Mikolov 等[5]又进一步提出了文本向量化的深度学习算法 Doc2Vec，将向量特征的计算从词语层面扩展到句子（段落）层面。Doc2Vec 是在 Word2Vec 的基础上提出的一种用于计算长文本向量的深度学习算法，与 Word2Vec 不同的地方是，在神经网络输入层，Doc2Vec 增加了一个句子向量，在每次训练过程中，将长文本作为一个特殊的段落 ID 引入语料中。在训练过程中，算法结合上下文、单词顺序和段落特征，训练词向量出现的概率分布。为此，算法在计算句子向量的同时也可计算词向量。

Doc2Vec 主要通过两种模型进行训练：DM 和 DBOW。两种模型均以神经网络语言模型为基础， 去掉隐含层， 利用上下文和段落特征来预测某词语出现的概率分布。段落向量与词向量是其训练过程的副产物。

4.2.2 利用 Doc2Vec 进行文本向量化

对于有标签的数据，利用 Doc2Vec 可以用监督学习的方法进行文本分类。因此，本文采用 Doc2Vec 对文本关键词训练向量化。

4.3 构建基于 XGBoost 算法的分类模型

4.3.1 XGBoost 算法

Boosting 是一种非常有效的集成学习算法，采用 Boosting 方法可以将弱分类器转化为强分类器，从而达到准确的分类效果。其步骤如下所示：

（1）将所有训练集样本赋予相同权重；

（2）进行第 m 次迭代，每次迭代采用分类算法进行分类，采用公式计算分类的错误率：

$$err_m = \frac{\sum \omega_i I(y_i \neq G_m x_i)}{\sum \omega_i} \tag{4}$$

式中ω_i代表第i个样本的权重，G_m代表第m个分类器；

（3）计算$\alpha_m = \log\left(\frac{1-err_m}{err_m}\right)$；

（4）对于第 m+1 次迭代，将第i个样本的权重ω_i重置为$\omega_i \times e^{\alpha_m \times I(y_i \neq G_m x_i)}$；

（5）完成迭代后得到全部的分类器，采用投票方式得到每个样本的分类结

果。其核心在于每次迭代后，分类错误的样本都会被赋予更高的权重，从而改善下一次分类的效果。

Gradient Boosting 是 Boosting 的一个改进版本，经证明，Boosting 的损失函数是指数形式[6]，而 Gradient Boosting 则是令算法的损失函数在迭代过程中沿其梯度方向下降，从而提升稳健性。

XGBoost[7]是一种 Gradient Boosting 算法的快速实现，它能够自动利用 CPU 的多线程进行并行，同时在算法上进行改进以提高精度。XGBoost 对损失函数做了二阶的泰勒展开，在目标函数之外加入了正则项，整体求最优解，用以权衡目标函数的下降和模型的复杂程度，避免过拟合。

4.3.2 基于 XGBoost 算法的分类模型实现

1. XGBoost 分类建模

本文采用 XGBoost 算法的 Python 语言版本进行分类建模，并对其结果进行可视化，如图 4-1 所示：

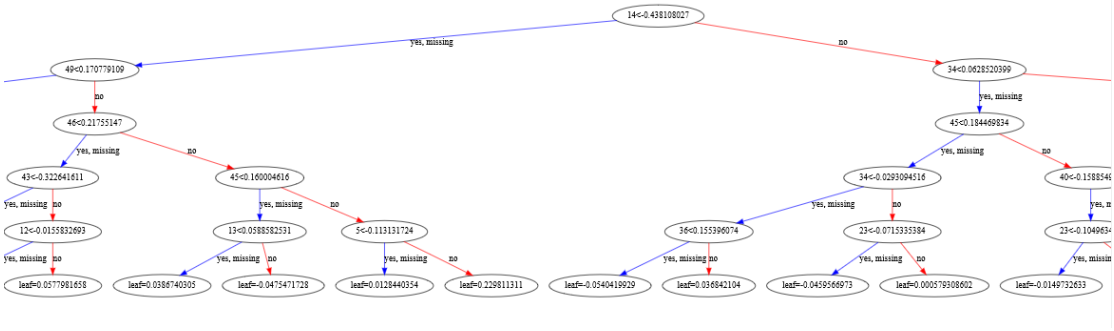

图 4-1 树的部分展示图

2. 通过网格搜索进行局部调参

先使用 xgboost.cv 来确认参数的范围，采用 np.arange 作为参数的备选值，对影响最显著的参数 n_estimators 进行局部参数调节。最终得到测试集的最佳结果F_1值为 0.87，其对应参数如表 4-1 所示：

表 4-1 最优分类模型参数详情

参数	参数说明	设定值
n_estimators	弱分类器的数量	100
colsample_bytree	特征采样的比例	0.8

learning_rate	学习率	0.1
max_depth	树的最大深度	5
min_child_weight	子节点最小样本权重和	1
subsample	从样本中进行采集的比例	0.8
Seed	随机种子	0

5 任务二的模型建立

5.1 文本预处理

5.1.1 基础处理

同 3.1 的方法， 首先采用 jieba 工具对附件 3 中的“留言主题”和“留言详情”文本数据进行分词，再以哈工大停用词表为基础建立停用词表去停用词，接着用 jieba 工具基于 TF-IDF 算法对“留言详情”去停用词结果进行关键词提取，将“留言主题”去停用词结果与“留言详情”关键词去除重复数据后合并，得到初始的文本特征集合。

5.1.2 命名实体识别

命名实体识别(Named EntitiesRecognition, NER)是自然语言处理(Natural LanguageProcessing, NLP)的一个基础任务。其目的是识别语料中人名、地名、组织机构名等命名实体。命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法等。

考虑最后结果地点和人群在热点事件分类中的重要影响，本文引入基于条件随机场序列标注的命名实体识别方法将地点和人群作为关键词提取出来，并将其放入关键词列表中。

5.2 文本特征提取

同 4.2 的方法，对预处理后的文本进行特征提取。采用 Doc2vec 工具进行文本向量化。

5.3 基于 K-means++算法的热点问题确定模型

5.3.1 基于 K-means++算法的问题聚类模型

1. K-means++算法原理

聚类是按照某个特定标准把一个数据集分割成不同的类或簇,使得在同一个簇内的数据对象的相似性尽可能的大,同时不在同一个簇中的数据对象的差异性也尽可能的大，也就是说， 聚类后同一类别的数据尽可能的聚集在一起,而不同的数据尽量分离。

K-means 算法是一种基于划分的聚类算法，其基本思想是：对随机选取的 K

个点为中心进行聚类，根据距离的原则将其进行归类。通过重复反馈，一次次重新选择各类的中心点的数值，直到满足精确度得到最终的聚类结果。但是根据不确定性抽取的点可能会使聚类的结果和数据的实际分布相差很大。

K-means++算法则是根据某种原则选择初始“种子点”的一种改进算法，该算法简述如下：

（1）从已知的数据中随机选择一个数矩作为第一个初始的聚类中心。

（2）对于给定的聚类中心，计算集合中每一个数据点 x 到这个中心点的距离 D(x)。

（3）比较所有计算得到的 D(x)的大小，选取 D(x)最大的点作为第二个聚类中心。

（4）重复（2）（3）步，以此类推，直到选出第 K 个聚类中心。

（5）针对数据集中每个样本X_i，计算它到 K 个聚类中心的距离，并将其分到距离最小的聚类中心所对应的类中。

（6）针对每个类别C_i，重新计算它的聚类中心，即属于该类所有样本的质心。

（7）重复第五步和第六步直到聚类中心的位置不再发生变化。

2. 聚类分组数目的确定

因为可能存在多个市民留言同样的问题，进行 K-means++聚类时，聚类个数难以确定，聚类数量较少造成聚类结果不精确；聚类数量过多计算所需空间与时间将很长。因此，引入手肘法[8]确定 K 值:

$$SSE = \sum_{i=1}^{k} \sum_{p \in L_i} \|p - q_i\|^2 \quad (5)$$

式中：k：聚类个数；p：第i个类组L_i中的数据对象；q_i：某个类组种所以数据对象的平均值分别进行多次不同聚类个数的分析，求解每种聚类个数时的SSE值，并绘制曲线如图 5.1 所示

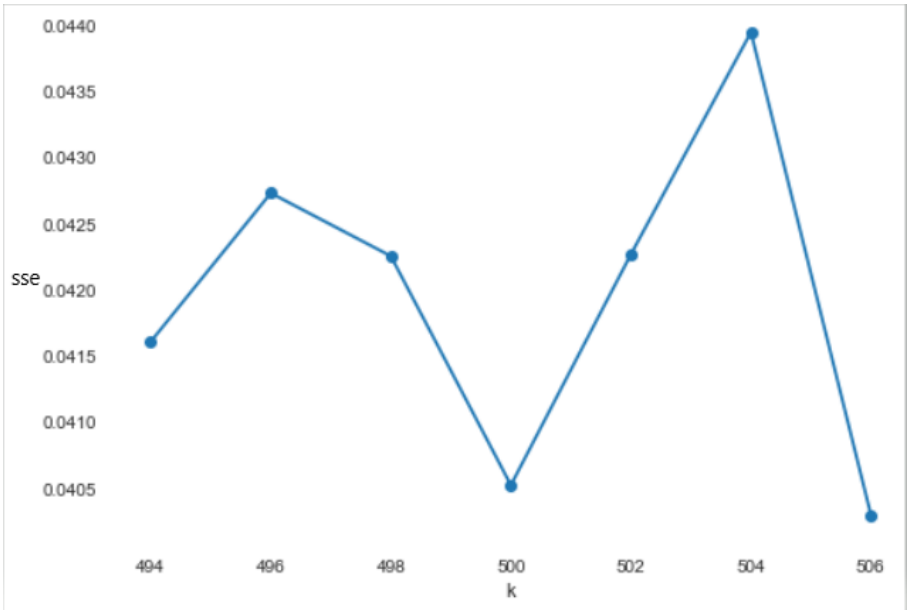

图 5-1 聚类 SSE 值随聚类个数的变化

由图可以看出在聚类个数在 500 时候，SSE值显著小，尽管聚类个数的增多，会使SSE值减小，但是造成的计算时间和空间将是个人电脑难以承受的，为了兼具准确性和便捷性，选取聚类个数为 500 进行分析。

5.4 留言问题热度评价

5.4.1 热度评价指标

本文从三个方面对留言问题进行热度评估，问题关注度、问题支持度与问题集中度。

（1） 问题关注度

问题相关留言数量体现了人群对该问题的关注度。有大量相关留言出现的问题才可以成为热点问题，所以问题相关留言的数量是热点问题的重要评价标准，一个问题在一段时间之内只有几条相关留言，那么这个问题的热度一定不会太高。问题关注度可以用问题相关留言数量来衡量。

（2） 问题支持度

留言的点赞数与反对数体现了人群对留言问题的支持度，点赞数越多，表明人群对该问题支持度越高；反之，反对数越多，人群对该问题支持度越低。问题支持度可以用留言点赞数与反对数的差来衡量。

（3） 问题集中度

考虑本文研究的留言问题对实时性要求较高，所以需要考虑一个问题是否在短时间内有大量的相关留言。如果关于一个问题的相关留言数量很多，但是分散在较长的时间段，那么可以判断其不是实时的热点问题。问题集中度可以用相关问题首次出现与最后一次出现的时间间隔来衡量。

5.4.2 热度值计算

本文按照上述的三个角度进行留言问题热度的评估，在对这三种因素综合评估之后，设计出了对于留言问题热度的计算公式，如公式所示。

$$HotValue(C) = \frac{n}{N} \times \frac{a-d}{M_a - M_d} \times \frac{n}{T_e - T_b} \tag{6}$$

其中，n表示问题相关留言数量，a表示留言点赞数，d表示留言反对数，N表示在特定期限T_b至T_e内所有留言数量，M_a、M_d分别表示在特定期限T_b至T_e内所有点赞数与反对数，T_e、T_b分别表示与问题相关的最后一条留言与首次留言的时间。

5.5 热点问题前 5 的确定

通过 Python 软件编写 K-means++聚类算法，得到 500 个聚类结果。其中类中元素最多有 38 个，最少有 1 个，通过热度评价指标得到前五名的热点问题，如下表 5-1 所示：

表 5-1 前 5 名的热度问题

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	2265	2019/1/11 至 2019/7/8	A 市 58 车贷案受害人	A 市 58 车贷案案情毫无进展
2	2	1889	2019/1/15 至 2019/11/11	A 市五矿万境 K9 县住户	A 市五矿万境 K9 县物业管理混乱，房屋质量与交通存在问题
3	3	1559	2019/1/30 至 2019/9/6	A 市绿地海外滩小区业主	A 市绿地海外滩小区距离长赣高铁太近，噪音大
4	4	1305	2019/3/26 至 2019/7/9	A6 区月亮岛路	A6 区月亮岛路沿线架设 110KV 高压电线杆
5	5	1174	2019/4/11	A 市金毛湾小区业主	A 市金毛湾配套入学问题

6 任务三的模型建立

6.1 选取评价指标

针对答复意见质量的评定标准主观性过强、考虑指标单一的问题，本文提出一种客观有效的答复意见质量分级方法。本文从三个方面，答复意见的相关性、完整性、时效性对答复意见的质量进行评价。

（1） 相关性

相关部门对留言的答复意见内容应当与留言详情内容相关，相关性可以用答复意见内容与留言详情内容之间的相似度指标衡量。

（2） 完整性

相关部门对留言的答复意见内容应当规范，开头应当为“网友：您好！您的留言已收悉。现将有关情况回复如下”，结尾应当为“感谢您对我们工作的支持、理解与监督”。统一开头与结尾使答复内容规范完整。本文用规范性指标衡量完整性。

（3） 时效性

相关部门对留言进行答复越及时，答复意见质量越好。时效性可以用答复时长指标来衡量。

6.2 评价指标的处理

由于指标的衡量标准不统一，为了能够得到理想的评价结果，需要将各个子指标进行量化和归一化处理。

（1） 文本相似度指标

计算“答复意见”与“留言详情”文本相似度之前，应当对文本进行分词、去停用词，并用 doc2vor 提取特征，使文本向量化，再采用余弦函数来计算两个文本之间的相似度。

（2） 规范性指标

对“答复意见”进行分词、去停用词，建立一个词表，词表中包含了规范用词，如“网友”、“您好”、“留言”、“收悉”、“感谢您”、“支持”、“理解”等。采用计分的方法，将“答复意见”文本与词表进行匹配，若文本中出现一个词表中的规范用词，则规范值加 1 分，扫描整个词表，计算出每条“答复意见”的规范

值，设置最大规范值为词表中的最大词数。

（3） 答复时长指标

用答复时间与留言时间的时间间隔差来衡量答复时长，单位为天数。

6.3 基于 AHP 的答复意见质量评价模型

层次分析法（Analytic Hierarchy Process）是一种定性分析与定量分析相结合的多目标评价决策方法。该方法将一个复杂的多目标决策问题作为一个系统，将目标分解为多个目标或准则，进而分解为多指标（或准则、约束）的若干层次，形成多层决策结构模型。依据该结构计算得到各个子指标对决策目标的影响权重，从而能够对各个目标进行综合评价。AHP 计算流程如图 6-1。

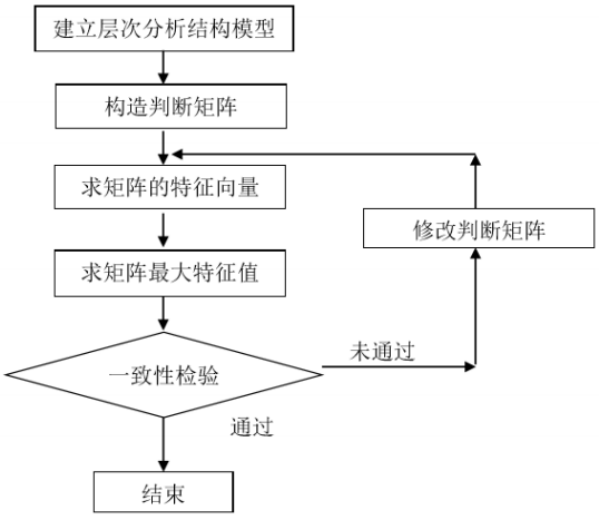

图 6-1 AHP 流程图

6.3.1 计算方法

（1）构造两两判断矩阵

采用 9 标度法（含义见表）对各指标的权重加以判别，进行两两指标间的相对比较，可得目标层与准则层、准则层与指标各层各指标间的判断矩阵 $A = (aij)m \times n$。

表 6-1 9 标度法评价规则

判断尺度	评价规则
1	两指标相比，重要性相同
3	两指标相比，一个比另一个稍微重要
5	两指标相比，一个比另一个明显重要

7	两指标相比，一个比另一个强烈重要
8	两指标相比，一个比另一个极限重要
2、4、6、8	介于上述两个相邻判断尺度中间
倒数	指标 i 与 j 比较得判断值 a，指标 i 与 j 比较得判断值 1/a

（2）用和积法计算判断矩阵

将判断矩阵的每一列向量归一化得到 B = $(bij)m\times n$。

$$b_{ij}=\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}} \qquad i,j=1,2,3\ldots\ldots n \tag{6}$$

再计算 B 的行向量元素的算数平均数$\omega_i=\frac{1}{n}\sum b_{ij}(i=1,2,3\ldots..n)$，$W=(\omega_1,\omega_2,\omega_3,\ldots\ldots\omega_n)^T$即为所求的特征向量。

（3）计算最大特征值

$$\lambda_{max}=\frac{1}{n}\sum_{i=1}^{n}\frac{(A\omega)_i}{\omega_i} \tag{7}$$

式中$(A\omega)_i$为 Aw 的第 i 个分量。

（4）一致性检验

计算一致性比率CR ，$CR=\frac{CI}{RI}$。当$CR\geq 0.1$时，认为判断矩阵 A 不具有满意一致性，需要进行修正；相反，当$CR<0.1$时，认为判断矩阵 A 有满意一致性。

其中，$CI=\frac{\lambda_{max}-n}{n-1}$为一致性指标，$n$为矩阵的阶数；$RI$为平均随机一致性指标，取 Satty 计算指标值见表 6-2。

表 6-2 RI 指数

矩阵阶数	2	3	4	5	6	7	8	9	10	11	12	13
RI	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.54	1.56

6.3.2 构造矩阵与权重确定

由于一级指标相对应的二级指标均只有一个，不需要考虑二级指标权重的问

题。确定综合评价指标对一级指标的判断矩阵，再对矩阵进行计算，得到以下结果：

表 6-3 判断矩阵计算结果

	A1	A2	A3	W	λ_max=3.0387
A1	1	5	3	0.6333	CI=0.01936
A2	1/5	1	1/3	0.1062	RI=0.5800
A3	1/3	3	1	0.2605	CR=0.0334

6.3.3 各指标对答复意见质量评价的贡献度

答复意见质量评价综合指标V

$$V = \sum_{i=1}^{n} \omega_i v_i$$

式中：v_i为第 i 个二级指标的值，ω_i为第 i 个二级指标的权重。

6.3.4 评价结果

通过本文建立的 AHP 层次分析法综合评价模型对附件 4 中的答复意见进行评价，得到评估得分区间为[0.000208,0.741896]。

列出评价质量排名前十的答复意见及其得分如表 6-4。

表 6-4 排名前十的答复意见得分

排名	留言编号	相似度	规范值	答复时长	评价得分
1	96757	1.0000	1.0000	0.0092	0.7419
2	18248	0.8660	1.0000	0.0110	0.6575
3	170734	0.8660	0.4615	0.0100	0.6001
4	97073	0.7071	1.0000	0.1369	0.5897
5	98326	0.8660	0.3077	0.0211	0.5866
6	138410	0.7071	1.0000	0.0875	0.5768
7	7457	0.7071	1.0000	0.0213	0.5596
8	90208	0.7071	0.8462	0.0052	0.5390
9	177915	0.7071	0.8462	0.0030	0.5384
10	119178	0.7071	0.7692	0.0086	0.5318

参考文献

[1]化柏林. 知识抽取中的停用词处理技术[J]. 现代图书情报技术，2007(8) : 48 —51.
[2] Silva C, Ribeiro B. The importance of stop word removal on recall values in text categorization[C]//Neural Networks, 2003. Proceedings of the International Joint Conference on. IEEE, 2003, 3: 1661-1666.
[3] Hao L, Hao L. Automatic identification of stop words in Chinese text classification[C]//Computer Science and Software Engineering, 2008 International Conference on. IEEE, 2008, 1: 718-722.
[4] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. [2018-07-24]. http://ling. .Snu.Ac.kr/class/AI_Agent/lecture//07-1-WordRepresentationinVectorSpace.pdf
[5]Le Q, Mikolov T. Distributed Representations of Sentences and Documents[C]//Proceedings of International Conference on Machine Learning. 2014.
[6] Theofanis Sapatinas. The Elements of Statistical Learning[M].Springer，2001：192-192.
[7] Chen T， He T， Benesty M. xgboost ： Extreme Gradient Boosting[J].2016，5（9）：222-208.
[8]王建仁,马鑫,段刚龙.改进的 K-means 聚类 k 值选择算法[J].计算机工程与应用,2019,55(08):27-33.