一种基于向量空间模型和 PZB 服务质量模型的智慧政务系统

摘要

近年来，微信、微博、市长信箱、阳光热线等网络问卷渐渐的变成政府了解人民意愿、聚集人民智慧、凝聚民心的重要途径。各类社情民意相关的文本数据量不断攀升，有关部门的工作面临重大挑战。同时，大数据、云计算、人工智能等新时代技术的不断发展，构建基于自然工程技术的智能管理系统（NLP）已成为创造和发展高动力社会治理的新政策。因此，为了提高政府相关部门处理留言的有效性，本文构建基于自然语言处理技术的智慧政务模型来解决此类问题。

对于问题 1，首先利用 Rstudio 软件对赛题提供的附件 2 文本数据进行必要的预处理之后，采用分词和不分词两种方法提取特征词，采用互信息作为筛选特征词的索引，计算具有最大互信息的特征词的 tf-idf 权重，并构建词权-文档矩阵。最后，我们使用 K 邻近法对测试留言文本进行分类。从预测结果来看，分词及不分词两种特征词提取法下的 K 近邻法均能达到比较理想的分类效果，其中分词的查准率为 78.3243%，F1-score 为 0.6463；不分词的查准率为 73.2680%，F1-score 为 0.5703。

对于问题 2，采用因子分析法对运用实例对根据附件 3 文本数据设计的热点问题评价指标体系进行分析，核心思想就是运用因子分析提取公因子，公因子尽可能包含所有指标，字数充实度（x_1）、出现及时性（x_2）、内容相似度（x_3）、点赞率（x_4）、反对数（x_5），进而以公因子作为自变量对因变量 Y 进行解释，并最终根据模型 $F=0.7934F_1+0.2065F_2$，其中 $F_1=0.313x_1+0.298x_2+0.615x_3$；$F_2=0.954x_4+0.977x_5$，计算出排名前五名的热点问题，其中在模型 F 中计算得到排名第一的热点问题为 A 市 58 车贷特大集资诈骗案，该热度指数为 64.16224，其余详细见“热点问题表.xls”及“热点问题留言明细表.xls”。

对于问题 3，我们以帕拉休拉曼(A.Parasuraman)，赞瑟姆(ValarieAZeithamal)和贝利(LeonardL.Berry) 等众多美国营销学家，他们发现 PZB 服务质量模型(Service Quality Model)与中国国情与现代社会环境的特征

相结合进行的一项新的反应质量评估, 构建了评价主体是群众，评价客体则是有关政府部门提供的评估服务。评价结果的新的答复质量评价指标体系将会变为大众对评估目标和相关政府部门的回应质量。

关键词：tf-idf；K 邻近法；因子分析法；PZB 服务质量模型

A Smart Government Affairs System Based on Vector Space Model and PZB Service Quality Model

Abstract

In recent years, online questionnaires such as WeChat, Weibo, mayor's mailbox, and sunshine hotline have gradually become an important way for the government to understand people's wishes, gather people's wisdom, and gather people's hearts. The volume of text data related to various social conditions and public opinion is constantly rising, and the work of relevant departments is facing major challenges. At the same time, with the continuous development of new era technologies such as big data, cloud computing, and artificial intelligence, building an intelligent management system (NLP) based on natural engineering technology has become a new policy for creating and developing highly dynamic social governance. Therefore, in order to improve the effectiveness of relevant government departments in processing messages, this paper builds a smart government model based on natural language processing technology to solve such problems.

For question 1, first use Rstudio software to perform the necessary pre-processing on the text data of the attachment 2 provided by the competition question, and use two methods: word segmentation and non-word segmentation to extract feature words, and use mutual information as an index to filter feature words. The tf-idf weight of the characteristic words of the information, and construct the word weight-document matrix. Finally, we use the K proximity method to classify the test message text. From the prediction results, the K nearest neighbor method under both word segmentation and non-word segmentation feature extraction methods can achieve an ideal classification effect, in which the accuracy of word segmentation is 78.3243%, and the F1-score is 0.6463; The accuracy rate is 73.2680%, and the F1-score is 0.5703.

For problem 2, the factor analysis method is used to analyze the evaluation index system of hot issues designed according to the text data in Annex 3 by using examples. The core idea is to use factor analysis to extract common factors. The common factors include all indicators as much as possible, and the number of words is full (x_1), Timeliness (x_2), content similarity (x_3), like rate (x_4), antilog (x_5), and then use the common factor as the independent variable to explain the dependent variable Y, and finally according to the model $F=0.7934F_1+0.2065F_2$, where $F_1=0.313x_1+0.298x_2+0.615x_3$; $F_2=0.954x_4+0.977x_5$, the top five hot issues are calculated, and the ranking is calculated in model F The first hot issue is the 58 car loan extra large fundraising scam in City A. The popularity index is 64.16224. For the rest, please refer to "Hot Issue Table.xls" and "Hot Issue Message List.xls".

For question 3, we use A. Parasuraman, Valarie AZeithamal, Leonard L. Berry and many other American marketing experts. They found that PZB Service Quality Model and A new response quality assessment based on the combination of China's national conditions and the characteristics of the modern social environment has established that the evaluation subject is the masses, and the evaluation object is the evaluation service provided by the relevant government department. The new response quality evaluation index system of the evaluation results will become the quality of the public's response to the evaluation objectives and relevant government departments.

Key words: tf-idf; K proximity method; factor analysis method; PZB service quality model

目 录

1．挖掘目标

本次数据挖掘的目标是利用互联网公开来源的群众问政留言记录，借助Excel、Rstudio 等软件利用 jieba 中文分词工具对留言内容进行分词和去停用词、K-means 聚类的方法及 KNN 算法 、SPSS16．0 对收集的数据变量进行因子分析等，达到以下三个目标。

1）利用 Rstudio 软件对文本数据进行必要的预处理之后，采用分词和不分词两种方法提取特征词，采用互信息作为筛选特征词的索引，计算具有最大互信息的特征词的 tf-idf 权重，并构建词权-文档矩阵。最后，我们使用 K 邻近法对测试留言文本进行分类。

2）为了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，应用因子分析法确定问题热度评价指标体系系统中每个指标水平的权重，并定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题和按表 2 的格式给出相应热点问题对应的留言信息。

3）以 PZB 服务质量模型(Service Quality Model)与中国国情与现代社会环境的特征相结合进行的一项新的反应质量评估,构建了评价主体是群众，评价客体则是有关政府部门提供的评估服务。评价结果的新的答复质量评价指标体系将会变为大众对评估目标和相关政府部门的回应质量。

2．问题 1 分析方法与过程

2.1 问题 1 分析过程

为了使智慧政务系统能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。我们提出了一种基于向量空间模型的智慧政务模型，该模型主要包括五个部分：文本预处理、特征提取、构建词权-文档矩阵以及 K 近邻法。该模型的框架图如图 1 所示。

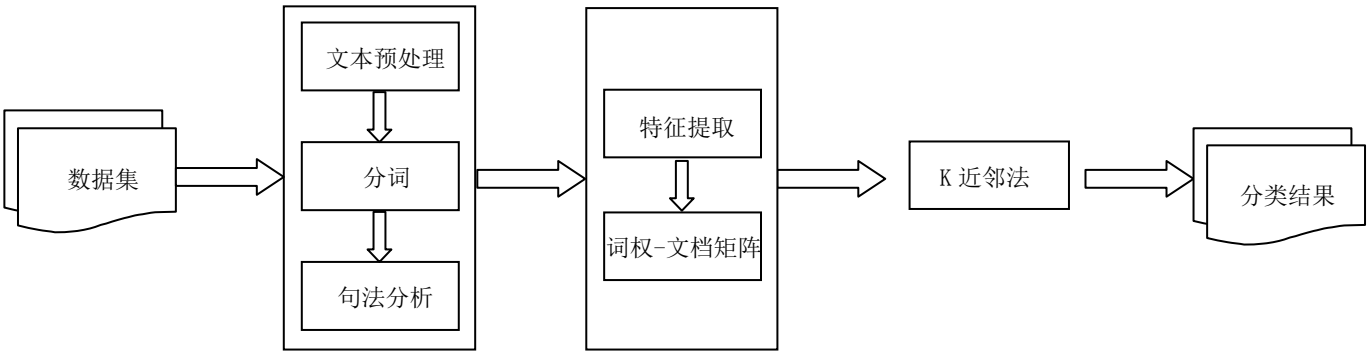

图 2-1 问题 1 流程图

本文的模型包含以下 4 个部分:

(1)文本预处理：分为分词、去停用词、文本的向量化三个流程步骤。

(2)特征提取：采用有词典及无词典情况下的提取方法提取特征词并进一步确定词的权重，将文本集转化为矩阵形式。

(3)构建词权-文档矩阵：词权即为词的权重，可以理解为特征词对某篇文本的重要程度,通过为不同的特征词分配不同的权重,可以用向量将文本表示出来。

(4) K 近邻法：是一种有监督研究的分类算法，它的规则即是数据，不需要生成其他数据来进行阐述。为了找到与每个测试样本相似度最接近的 K 个文本，我们通过分别算出其与训练样本集中每个留言文本的相似度,然后才根据加权距离和对测试留言文本所属的类别加以判断判断。

2.2 数据预处理

2.2.1 数据采集

本文选择的数据集来源于附件 2 中城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生的留言总计 9210 条。利用 Rstudio 软件读取数据,查看数据中是否出现留言为空的记录导致干扰问题的分析,若有,采取直接滤过方法去空。

文本数据以列表形式存储，列表的每一维对应一个专题，是一个由该专题下的留言文本组成的向量。利用旁置法来选取样本，其中训练样本集占总文本的 70%，测试样本集占总文本的 30%。

采集并预处理后的各专题文本分布如表 2-1：

表 2-1 文本集类别分布表

一级指标 / 样本集	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
原始文本集	2009	937	612	1588	1968	1214	876
训练样本集	1406	656	428	1112	1378	850	613
测试样本集	603	281	184	476	590	364	262

为了计算需要，与原始样本集相比，训练样本集与测试样本集不再以列表形式存储，而是转化为向量形式。

2.2.2 分词

我们可以看到，附件中留言主题和留言详情中是大量的中文文本，其特点是词词之间没有明显的分隔，因此，从文本提取词语时，需要进行词词拆分。

基于统计的分词也是中文分词技术的研究热点。其基本思想为：词语是字的固定组合，如果某两个或多个字经常在一起出现，那么这个组合就很有可能是一个词,例如：长株潭城、西湖建筑集团、蔡锷南路等固定词组。当找出所有可能的词语后，可运用决策算法和统计语言模型找到最佳的细分结果。该类方法通常和词典结合使用，并且需要大量语料来完成注释，同时还将会随着搜索空间的增加，词语间的拆分速度也会降低。不过与机械分词法相比，基于统计的分词方法也适用于无词典的情况。

由于手动编写的分词函数未经优化，分词效率较低，本文分词部分利用 jiebaR 包完成。jiebaR 是 R 语言中的软件包，可实现词语间的拆分，集成了混合模型，最大概率法，隐式马 尔科夫模型，索引模 型等分词方法，内置了分词词典。该程序包利用 Rcpp 进行开发，具有很高的分词效率。

2.2.3 去停用词

在文本处理中，功能词是人类语言组成的一部分。相对于其他一些词，功能词并没有它真正的含义。限定词是我们使用最多的功能词（″我″、″怎么″、″它″、

″好的″、和″她″），限定词有助于解释名词并在文本中表达概念。因为限定词的两个功能，所以在搜索引擎的文本处理过程中需要对它进行特殊处理。首先，这些功能词很常见。在每个文档中记录这些功能词需要占用很大的内存空间。其次，由于它们的普遍性和多功能性，这些词很少代表有关文档相关性的信息。一般情况下在预处理阶段就会删除停用词，以避免在提取文本信息时进行更多工作。本文所有的停用词，取自四川大学机器智能实验室 停用词表、哈工大 停用词表和百度停 用词表的集合。

2.3 向量空间模型

由于文本数据一般只具有有限结构甚至没有结构，计算机很难理解与处理，因此，文本数据必须在数据挖掘与分析之前，进行数据预处理。根据词法分析系统 ICTCLAS 介绍，词是可以独立活动的语言成分[1]，同时将大量词语组合起来会成文本，因此对文本的分析可以转化为对组成文本的词语的分析。

向量空间模型（VSM，VectorSpaceModel），又称“词袋”法，是由 Salton 等人研究发表的一种文本表示模型[2]，原理是将非结构化文本数据转变为结构化文本数据。目前，它是标准的文本处理模式。在向量空间模型中，文本集中的每条消息都与向量一一对应，所有可以表示文本特征的词语构成一个空间，每一个词对应空间的一个维度，每一个维度的坐标为对应词语的权重。于是对文本的向量化处理归结为确定特征词集与确定特征词权重。

中英文文本提取特征词的流程略有不同，对于英文数据来说，词是写作的最小单位，空格将单词与单词间隔开来，但是相同的单词通常采用不同的格式，因此需抽取词干，再经过去除停用词即可将一篇文档转化为多个单词的排列。而对于中文语料，书写的基本单位为字，连在一起的字或单字组成词，词与词之间在形式上并没有分隔符，在处理文本时，首先进行分词处理，其次再去除停用词。本文采用有词典及无词典情况下的特征词的提取方法提取特征词并进一步确定词权重，将文本集转化为矩阵形式。

2.4 特征提取

2.4.1 有词典特征词提取

随着互联网的普及，需要处理的文本急剧增加，中文分词技术也日益受到关注。近三十年来，分词理论方面已取得了一定的成果，主要方法可以总结为：基于词典的分词方法、基于理解的分词方法以及基于统计的分词方法[3][4]。

基于词典的分词方法即基于字符串匹配的分词方法，也称为机械分词法。要使用此方法，我们需要根据词典中的某些规则来匹配给定词典中的单词和文本。如果当文本中的一段文字与词典中匹配成功，则截取该段落并视其为一个词，之后将文本的其他部分重新按规则进行匹配，直到全部完成。机械分词法按照扫描方向不同可分为正向匹配法、逆向匹配法和双侧匹配法；根据匹配字符串长度优先程度不同可以分为最大匹配法和最小 匹配法。

一般常用的正向最大匹配法即从句子的第一个字开始。在文本中，把将要进行词语拆分的几个连续字符和词表进行一一配对，一般情况下字符串长度可取为字典中词条长度的最大值，若该字符串匹配成功，则截取出该字符串作为组成句子的第一个词，再将剩余部分作为一个完整的句子进行扫描，若字典中没有这一词条，则将字符串长度减一再进行匹配，直至单字成词，重复以上过程，直至剩余字符串长度为零。反向最大匹配法与之类似，只是扫描方向起始于句子的最后一个字。由于扫描方向不同，两种方法的分词结果也可能不同，分别按照两种分词方法分词，再根据一定规则选择分词结果的方法即双侧最大匹配法。

机械分词法经过多年发展，算法已比较成熟且易于理解与实现，不过该方法仍然存在一定的局限性。首先，分词需要事先准备一张足够全面的字典，否则无法正确切分文本进而提取特征词，即使得到一张完备的词典，对于新词也束手无策，并且还有一个更大的难题是对文本中的未登录词处理。还有一张足够大的词典在提高切分准确率的同时占据了大量的存储空间，也降低了分词的速度。虽然如此，机械分词法仍然是一种应用广泛的分词方法。

基于统计的分词也是中文分词技术的研究热点。其基本思想为：词语是字的固定组合，如果某两个或多个字经常在一起出现，那么这个组合一般情况下是一个词。当找出所有可能的词语后，可运用决策算法和统计语言模型找到最佳的细

分结果。该类方法通常和词典结合使用，并且需要大量语料来完成注释，同时还将会随着搜索空间的增加，词语间的拆分速度也会降低，不过与机械分词法相比，基于统计的分词方法也适用于无词典的情况[5]，分词、去停用词后部分实例如下图 2-2 所示。

Name	Type	Value
segword2	list [494]	List of length 494
[[1]]	character [9]	'市' '西湖' '建筑' '集团' '占' '道施' ...
[[2]]	character [11]	'市' '在' '水一' '方大' '厦人' '为烂' ...
[[3]]	character [13]	'区' '杜鹃' '文苑' '小区' '外' '的' ...
[[4]]	character [12]	'民工' '在区' '明发' '国际' '工地' '受伤' ...
[[5]]	character [7]	'县' '丁' '字街' '的' '商户' '乱' ...
[[6]]	character [11]	'县' '南门' '街' '干净' '整洁' '了' ...
[[7]]	character [11]	'县' '冷江' '东路' '蓝' '波旺' '酒店' ...
[[8]]	character [10]	'县' '九' '亿' '广场' '的' '公厕' ...
[[9]]	character [14]	'县' '石期' '市镇' '老' '农贸' '市场' ...
[[10]]	character [5]	'市域' '轨道' '交通' '规划' '建议'
[[11]]	character [7]	'关于' '市域' '轨道' '交通' '规划' '的' ...
[[12]]	character [10]	'请' '问市' '乘坐' '地跌' '是' '否可' ...
[[13]]	character [15]	'地' '铁号' '线施' '工导' '致市' '锦楚' ...
[[14]]	character [10]	'区润' '和' '紫郡' '用' '电' '的' ...
[[15]]	character [13]	'市' '锦楚' '国际' '新城' '从' '月' ...
[[16]]	character [6]	'咨询' '市楼' '盘' '集中' '供暖' '一事'
[[17]]	character [13]	'市能' '不能' '像' '北方' '一样' '给' ...
[[18]]	character [6]	'市' '可以' '实现' '集中' '供暖' '吗'
[[19]]	character [5]	'县' '坐' '公交' '车要' '元'
[[20]]	character [9]	'希望' '县' '的' '路路' '公交' '车' ...
[[21]]	character [7]	'反映' '县公' '交车' '监控' '的' '有关' ...
segword2[[3]]		

图 2-2 部分实例数据

2.4.2 无词典特征词提取

对于无词典的情况，由于要在没有词典的情况下直接在文本中提取片段，有必要明确什么样的文字片段可能构成词[6][7][8]。以两字词语为例，首先可对文本段落逐字进行拆分，如“中文分词”可拆分为“中文”、“文分”和“分词”，之后计算每个词的词频并与初始默认的阈值进行比较，删除掉词频较低的字符串，剩下的则可以认为很有可能是一个词。分别按照每两字、每三字直至给定词语的字数上限，这样文档就被拆分为若干文字片段的组合。之后计算各可能构成词的

片段的相关指标并与给定阈值比较，本文利用片段出现频数、点互信息与左右邻字熵作为文字片段的依据[9]。

片段出现频数，顾名思义，指某一片段在文本集中出现的次数，出现次数越多，该片段越有可能成为一个词，如果这一片段确实是一个词，则片段出现频率即为这个词的词频。利用词频进行筛选可以有效地去除偶然连在一起的文字片段，大大减少候选词数量。

统计词频并按词频对候选词进行筛选的 R 语言函数如下：

```
cpsx＝function(txt,n,cp.min=2)#按字数分词及按词频筛选
{
wd=mapply(function(x)substring(x,1:(nchar(x)-(n-1)),n:nchar(x)),txt)
wd=sapply(txt,function(x)substring(x,1:(nchar(x)-(n-1)),n:nchar(x))
wd=unlist(wd)[nchar(unlist(wd))==n]
tw=sort(table(wd),decreas ing=TRUE)
tw[tw>=cp.min]
}
```

其中参数 n 为汉字串长度，txt 为将训练文本集按照标点拆分后形成的文本向量，cp.min 为片段出现的最少次数，默认为 2，小于该次数的汉字片段被认为不是一个词，并对大于等于该阈值的汉字串进行词频统计。调用该函数可计算单个候选词的出现次数，并据此进行初步筛选。

不过经常连在一起的文字不一定构成词，如“的人”可能在文档集中大量出现，但这显然不是一个词，而是由于“的”与“人”均为常用字，两个字连在一起的可能性也相对较大。因此，仅仅利用词频进行判断无法取得理想的结果。

机器学习中，有关点互信息 PMI 机器学习的资料表明，常利用点互信息（PMI，PointwiseMutualInformation）作为度量两事务间相关程度的指标，公式如下：

$$\text{PMI（A,B）}=\log_2\frac{p(A,B)}{p(A,B)p(A,B)}=\log_2\frac{p(A|B)}{p(A)}=\log_2\frac{p(B|A)}{p(B)} \qquad (2\text{-}1)$$

将点互信息应用到特征词提取中，可以衡量汉字串之间结合的紧密程度，也称凝固度。记 p(A,B)为汉字串中包含“AB”两个的概率，p(A)为汉字串中包含“A”是的概率，p(B)为汉字串中包含“B”出现的概率，汉字串中包含“AB”、“A”与“B”的个数分别为 n(AB)、n(A)、n(B)，假设词频总数为 n，则 p(A,B)= $\frac{n(A,B)}{n}$ ， p(A)= $\frac{n(A)}{n}$ ， p(B)= $\frac{n(B)}{n}$ ， PMI （ A,B ）

$=\log_2 \frac{p(A,B)}{p(A,B)p(A,B)}=\log_2 \frac{n(A,B)n}{n(A,B)n(A,B)}$。

若 PMI(A,B)>0，则“A”与“B”确实具有正相关关系，结合较为紧密，当 PMI(A,B)大于事先给定的阈值时，则可以认为“AB”可能是一个词。若 PMI(A,B)≤0，则可以认为“A”与“B”不相关或负相关，这样“AB”一般不会构成词[10]。计算点互信息（词语凝固度）的 R 语言函数如下：

```
Ngd＝function（x,txt）#凝固值（点互信息）
{
k=nchar(x);n＝nchar(txt)
ztw=substring(x,c(rep(1,(k-1))),c(1:(k-1)))
cztw＝mapply(gsub,ztw””,x)
ptwl=mapply(function(x)str_count(txt,x)/n,ztw)
ptw2=mapply(function(x)str_count(txt,x)/n,cztw)
ptw=str_count(txt,x)/n
pptw＝min(ptw(ptw1*ptw2))
1og(pptw,2)
}
```

其中参数 x 为候选词，txt 为将训练文本集去除标点后拼接在一起的文本。调用该函数可计算单个候选词的凝固度，可通过循环语句实现所有候选词凝固度的计算。

利用点互信息可以判断两不相关字串多次连在一起的情况，不过即使两字串结合紧密仍然可能不构成词。如汉字串“夜宵摊”，“夜宵”与“宵摊”结合均较为紧密，有较大的点互信息，但一般认为“夜宵”是一个词，而“宵摊”不是一个词，这是由于“夜宵”的左右邻接字都相对丰富，而“宵摊”的左邻接字则很贫乏，基本只有“夜”，因此可以利用邻字熵作为字串是否可独立成词的依据。

邻字熵[11]分为左邻字熵与右邻字熵，为汉字串左右两侧所有邻接字的信息熵，可用于判断汉字串能否独立成词，也可理解为自由度，邻字熵、左邻字熵与右邻字熵分别记为 H、HL、HR，公式如下：

$$HL(w)= -\sum_{iEleft} p(i)\log_2 p(i) \quad (2\text{-}2)$$

$$HL(w)= -\sum_{iEright} p(i)\log_2 p(i) \quad (2\text{-}3)$$

$$H(w)= \min(HL(w)，HR（w）) \quad (2\text{-}4)$$

其中 w 为候选汉字串，left、right 分别为 w 的左右邻字集–，n(i)为字 i 出现

的次数，n(w)为汉字串 w 出现的次数，p(i)=$\frac{n(i)}{n(x)}$。

邻字熵可取为左右邻字熵中较小的一个，邻字熵越大，则字串左右邻字越丰富，字串越有可能独立成词。不过公式中对数函数以 2 为底，这样随着词频增大，邻字熵也可能相应增大，因此本文对邻字熵计算公式进行调整，对数函数底数取为该汉字串出现次数，对结果进行归一化以消除词频的影响，这样便可对不同词频设置统一的阈值。若汉字串的邻字熵大于给定阈值，则认为该字串可以独立成词。

计算邻字熵的 R 语言函数如下：

```
ljs=function(x,txt) #邻字熵
{
cdt=str_detect(txt,x)
cdtlc=str_locate(txt[cdt],x)
cq=substring(txt[edt],cdtlc[,1]-1,cdtlc[,1]-1)
ch=substring(txt[cdt],cdtlc[,2]+1,cdtlc[,2]+1)
cq[which(cq=＝"")]=1:length(which(cq＝＝""))
ch[which(ch=＝"")]=l:length(which(ch＝=")）
tcq=table(cq);tch=table(ch)
xxsq=-sum(tcq/sum(tcq)*log(tcq/sum(tcq),sum(tcq)))
xxsh=-sum(tch/sum(tch)*log(tch/sum(tch),sum(tch)))
cxxs=min(xxsq,xxsh)
cxxs
}
```

其中参数 x 为候选词，txt 为将训练文本集去除标点后拼接在一起的文本。调用该函数可计算单个候选词的左右邻字熵，并取其最小值作为词语的邻字熵，可通过循环语句实现所有候选词邻字熵的计算。

经过词频、点互信息与邻字熵的筛选便得到了文本集中的部分词语，不过一般不直接利用这些词语作为词典进行分词，首先，特征词的提取未考虑单字成词的情况，字数超过一定字符数的词也未考虑，另外词频筛选过程会过滤掉低频词，导致词典不完备，分词的结果恐怕也不会很理想。

无词典提取的词语可用于两个方面：直接利用这些词语作为特征词或将词语加入已有词典进行分词，之后重新提取特征词。

词频筛选过程虽然会过滤掉低频词，然而这些词本来就是特征选择中需要删除的，而字符数较多的词语也不太可能在多篇文档中大量出现，也会在特征选择

中被过滤。可以看出，这种方法对特征词提取的不利影响并没有对分词严重，文本分类中分词的目的也是提取特征词，因此对得到的特征词进行进一步降维进而利用也是一个可行的思路。不利之处在于部分词语的词频可能会重复计入，由于该方法只统计出现频率而完全不考虑词义，导致完全无法应对歧义，“依景园”会被作为“依景园”、“依景”及“景园”三个词分别计数，导致文档集无法正确表示。

未登录词是基于字典的分词方法的一大难点，虽然一些分词模型可以自动识别部分未登录词，但效果往往差强人意，因此，在条件允许的情况下，可以先利用文本集找出所有可能的候选词语，再将其加入词典以提高分词准确性。将词频方法提取出的词语加入原始词典，用新词典分词，之后词频统计进而对特征词进行降维并表示文本。这种方法可以有效识别单字词，每个汉字只会被一个词包括，词频统计结果更加可靠，不过新词发现作为特征词提取的预备工作显得用时过长，分词得到的特征词集维度也会比不分词的方法高，降维过程也会消耗更多的时间。

2.5 文本表示

分词的工作是为了得到特征词集，每个特征词都代表向量空间当中的一个维度，接下来需要确定文档在每一个维度的取值，不妨将坐标值称作词权。词权可以理解为特征词对某篇文档的重要程度，为不同特征词赋予不同的权重便可以将文档利用向量表示出来。词权的形式有很多种，对应文本的多种表示方法，假设文本集中有 n 篇文档，m 个特征词，然后我们可用一个 m 维的向量来表示每一个文档，将 n 个向量拼起来便得到一个 n×m 维的矩阵，这个矩阵便可表示整个文本集。将文本数据结构化后，便可进一步利用统计模型及机器学习算法进行数据挖掘与分析工作[12]。

词权有多种表示方法[13]。首先，一种最基本的表示方法，对于特征词 w，若在文档 D 中存在 w，则文档 D 的特征词 w 的维度坐标值就应该为 1，否则该维度坐标值为 0，对所有特征词做出一次判断，可得到一个向量，这个向量即可代表本篇文档。扫描所有文档即可得到一个矩阵，该矩阵代表整个文本集，元素值非 0 即 1。这一表示方法形式简单清楚，易于表达结构化信息，但只记录词语

出现与否，会损失一些关于词语重要程度的信息。

为了充分利用文本提供的信息，很自然地，当某个词在一篇文档中多次出现时，它可能对表达文本意义更加重要，因此可以文档中各特征词词出现的次数，即词频作为对应的坐标值。用词频代表特征权重的方法能够表示词语在文档中的的重要程度，但有些常用词在很多文档中大量出现，另一些只在某一篇文章中集中出现，在进一步的聚类和分类中，显然后者更具有区分度，但单纯以词频作为词语权重则使得两个词语在该篇文档中同等重要。

为了对特征词分布的集中与分散加以区分，引入文档频率：文档集中包含特征词的文档篇数，记为 DF；结合词频：特征词在某一篇文章中使用的频率，记为 TF；那么能代表一篇文档特征的更可能是那些在该篇文章中使用较为频繁而在其他文档中使用不是很频繁的词语，即词频较高而文档频率较低。

于是 1998 年，Salton 将词频与文档频结合起来，系统地阐释了词频-逆文档频（tf–idf）权重，它的计算方式不一，本文选取计算方法公式如下：

$$TFIDF(w)=TF(w)IDE(w)= TF(w)\log_2(\frac{|D|)}{DF(w)+\beta}) \quad (2\text{-}5)$$

其中|D|为文档总数，β为一个较小的数，为了保证分母不为零，一般取为 0. 01。作为信息获取研究中经常被使用的一种词频加权方法，tf–idf 在文本处理领域有着十分广泛的应用，本文选取 tf–idf 作为特征词权重构建文本矩阵，为了方便之后的工作，分别编写分词与未分词提取的特征词的文本表示函数，之后分别带入训练文本集与测试文本集便可得到文本集的矩阵表示。

分词的方法下文档集转化特征-文档矩阵函数如下所示：

```
trans.Matrix1=function(fenci,tzc)
{
matrix.tf=lapply(fenci,function(x)sapply(tzc,function(y)sun(x==y)))
matrix.dfx=sapply(tzc,function(x)sum(sapply(fenci.function(y)x%in%y))
)
matrix.idfx=log(n/matrix.dfx+0.01)
matrix.tfidf=t(sapply(matrix.tf,function(x)x*matrix.idfx))
matrix.tfidf
}
```

其中，参数 fenci 为存放中文分词结果的列表，tzc 为特征选择后保留的特征词。

不分词的方法下文档集转化特征-文档矩阵函数如下所示：

```
trans.matrix2=function(txt,tzc)
{
matrix.tf=lapply(txt,function(x)sapply(tzc,function(y)str_count(x,y))
)
matrix.dfx=sapply(tzc,function(x)sum(str_count(txt,x)>0))
matrix.idfx=log(n/matrix.dfx+0.01)
matrix.tfidf=t(sapply(matrix.tf,function(x)x*matrix.idfx))
matrix.tfidf
}
```

其中，参数 txt 为将训练文本集去除标点后拼接在一起的文本，tzc 为特征选择后保留的特征词。以前 20 条留言的前 10 个特征词为例。

提取出特征项及确定词权后，便可以用空间中的点表示文本集中的文本，下面规定空间的度量。文本相似度指文本集中两个文本的相关程度，通常被用来衡量文本与文本两者间的关系程度。

向量空间模型提供了将文本数据向量化的框架，但在实际的数据挖掘过程中，对于一个文档集，词语数量成千上万，停用词表只能去除很小一部分噪声，特征向量维数极大，一方面会造成时间与空间的巨大花费，另一方面冗余的变量也会对进一步的分析造成不利影响。因此，需在数据挖掘与分析前对特征进行降维处理。

在进行分类之前，可以利用词语与词权重绘制词云。文本可视化是通过分析文本内容，发现文本的关键信息，并将其以图形形式呈现出来的方法，能够直观地了解到关键的信息点。词云常用来对文本的关键字进行可视化，一般以字体的大小区分词语的重要程度，字体越大表示词语在文本中越重要。本文利用 R 语言的 wordcloud2 函数绘制词云，由于两种特征词提取方法绘制的词云非常相似，本文只给出分词方法绘制的词云，如图 2-3：

图 2-3 词云图

图 2-3 反映了全部类别的文本集不同特征词的重要程度，通过观察词云可以发现，劳动和社会保障板块词语分布较为集中，重要的词有“补贴”“公积金”“工作”“领导”等，该板块留言一般面向劳动群众，与之相比，城乡建设板块“公积金”等词的权重仍然不低但重要程度大幅下降，而环境保护及文体教育板块已见不到劳动和社会保障板块重要词语出现。除此之外， 劳动和社会保障板块其他重要特征词还包括“福利”、“待遇”、“保护”等。

2.6 特征降维与分类

文本分类是文本挖掘应用最广泛的领域之一[14]。通常意义上的分类指有监督的分类，即训练样本集各样本带有明确的标签，测试样本集被分到每一类下，分类的过程可以看做从文档集到类别集合的映射。无监督分类即聚类，样本集不带有具体类别，聚类完成后，样本集依相似性形成一定数量的类别。对于分类算法，目前国内外学者已做了非常深入的研究，方法多种多样。分类算法大致可分为基于统计的方法、基于连接的方法及基于规则的方法[15]。本文采用基于统计的分类方法，其基本思想为首先分析训练文本集中的文本，并以此构造文本特征及类别间关系的模型，之后利用该模型对测试集进行分类。这类方法不考虑文本的语言学结构，用特征词表示文本，训练分类模型一般都是采用有监督的机器学

习，得以实现将文本进行分类。常用的分类方法包括 KNN、朴素 贝叶斯、支持向量机、神经 网络、决策树及随机森林等。

将文本数据结构化后，常用的聚类和分类方法也可应用于文本挖掘，不过未经过降维的文本数据具有极高的维度而且比较稀疏，变量间又往往具有相关性，直接代入已有模型结果可能并不理想。

R 语言的朴素贝叶斯分类方法可利用 class 包的 knn 函数实现，调用前需将数据转化为函数可以识别的类型，即词权-文档矩阵。由于每个变量代表一个特征词，不同变量间并不存在量纲或数量级差异，因此可直接将词权-文档矩阵代入函数。KNN 方法中 k 值的选择会对分类结果构成较大影响，因此可以先对不同的k值进行分类并计算准确率，选取准确率最高的k值作为近邻个数代入模型。

2.6.1 特征降维

特征词的降维一般采用特征选择或特征抽取。

特征选择是指根据某些条件从特征集合中选取一部分对分析文本有贡献的词语组成特征子集。该过程只是删去了相对不重要的特征，不会产生新的特征项。选取的规则主要包括信息增益、互信息、卡方统计、文本证据权重等。

特征抽取是指将原特征空间进行组合变换后生成一个特征之间更加独立的具有更少维度的新的特征空间[14]。大多数人运用特征抽取方法包含主成分分析、非负矩阵分解和潜在语义索引等流程。

文档频率（DF）的概念在前文已做介绍，指训练样本集中，某词语出现在文档中的文档个数总和。当文档频率很低时，词语只在少数文档中出现，不具有一般性，对分类的贡献不大，应予以去除。文档频率操作简便，也可以起到降维的效果，在文本集数量巨大时具有较高的效率。

互信息（MI）描述的是变量与变量间的相关程度，经常在统计语言模型中计算两个对象的相关程度。分析词语w_i与类别C_j，它们两个的点互信息定义为：

$$\mathrm{MI}(w_i,C_j)=\log_2\frac{p(w_i,C_j)}{p(w_i)p(C_j)} \tag{2-6}$$

其中，$p(w_i,C_j)$、$p(w_i)$、$p(C_j)$可以利用词频或文档频等方法计算，本文采用文档频来计算概率，记 $\mathrm{DF}(w_i,C_j)$为类别C_j中出现词语w_i的文档篇数，$\mathrm{DF}(w_i,\mathrm{D})$

为训练样本集中出现词语w_i的文档篇数，|c|表示类别总数，|D|表示文档总数，则MI（w_i,C_j）可化为：

$$\mathrm{MI}(\mathrm{w_i},\mathrm{C_j})=\log_2\frac{\mathrm{p}(\mathrm{w_i},\mathrm{C_j})}{\mathrm{p}(\mathrm{w_i})\mathrm{p}(\mathrm{C_j})}=\log_2\frac{\frac{\mathrm{DF}(\mathrm{w_i},\mathrm{C_j})}{|\mathrm{D}|}}{\frac{\mathrm{DF}(\mathrm{w_i},\mathrm{D})}{|\mathrm{D}|}\frac{|\mathrm{c}|}{|\mathrm{D}|}}=\log_2\frac{\mathrm{DF}(\mathrm{w_i},\mathrm{C_j})}{\mathrm{DF}(\mathrm{w_i},\mathrm{D})}+\log_2\frac{|\mathrm{c}|}{|\mathrm{D}|} \quad (2\text{-}7)$$

则根据定义，词语w_i的互信息为：

$$\mathrm{MI}(w_i)=\sum_j p(C_j)\ \mathrm{MI}(w_i,C_j) \quad (2\text{-}8)$$

在实际操作中，词语 wi 的互信息常用 MI（w_i,C_j）的最大值表示，即：

$$\mathrm{MI}(w_i)=\max_j(\mathrm{MI}(w_i,C_j)) \quad (2\text{-}9)$$

若特征词与分类无关，则互信息为零，互信息越大，特征词对分类的贡献就越大。因为互信息方法没有涉及词频这一层次，这就导致互信息方法更多的会去默认的选择低频词。

信息增益定义为在文档中特定特征词出现之前和之后的信息熵的差异，表示文档中包含特定特征词时文档类别中的平均信息量。公式如下：

$$\mathrm{IG}(w_i,C_j)=\log_2\frac{p(w_i,C_j)}{p(w_i)p(C_j)}\mathrm{p}(w_i,C_j)+\log_2\frac{p(\overline{w_i},C_j)}{p(\overline{w_i})p(C_j)}\mathrm{p}(\overline{w_i},C_j) \quad (2\text{-}10)$$

$$\mathrm{IG}(w_i)=\sum_j IG(w_i,C_j) \quad (2\text{-}11)$$

信息增益越大，特征词对分类的作用就越明显，因此选择信息增益大于预设阈值的特征词作为特征子集。不过信息增益同时计算了特征词出现与不出现对分类的影响，虽然特征词不出现也会对分类有所贡献，然而实验表明，这些贡献很多情况下远小于将其考虑进来所造成的干扰[16]，这也降低了使用信息增益进行降维的效果。

衡量特征词与文档类别间的相关性还可以用卡方统计的方法，记为 CHI，公式如下[17]：

$$\mathrm{CHI}(w_i,C_j)=\frac{n[p(w_i,C_j)p(\overline{w_i},\overline{C_j})-p(w_i,\overline{C_j})p(\overline{w_i},C_j)]}{p(w_i)\mathrm{p}(\overline{C_j})p(\overline{w_i})\mathrm{p}(C_j)} \quad (2\text{-}12)$$

$$\mathrm{CHI}(w_i)=\sum_j p(C_j)\ CHI(w_i,C_j) \quad (2\text{-}13)$$

卡方统计量值越高，特征词与类别相关性越高，该词条对分类的贡献就越大，因此过滤掉卡方统计量较低的词条，保留卡方统计量大于给定阈值的词条作为特征词[14]。

2.6.2 K 近邻法

K 近邻法（KNN，K–nearestneighbor）[18][19]由 Cover 与 Hart 研究发现的一种应用广泛的分类方法。K 近邻法是一种有监督研究的分类算法，其规则就是数据，并不需要产生其他数据来阐述。我们采用 K 近邻法建立群众留言一级标签分类模型，思路简单直观，其基本思想是：为了找到与每个测试样本相似度最接近的 K 个文本，我们通过分别算出其与训练样本集中每个留言文本的相似度，然后才根据加权距离和对测试留言文本所属的类别加以判断。可见 K 近邻法并不需要指定输出变量与输入变量的具体函数形式，只需要假设输出变量的预测值是训练样本集输出变量的一个函数即可[20]。

典型的 K 近邻法将 n 个样本观测数据视为 p 维特征空间中的点（p 为输入变量个数），并根据新的观测X_0的 k 个近邻的$(y_1, y_2, \dots, y_k)$，按照函数$(y_1, y_2, \dots, y_k)$计算X_0的输出变量预测值$\widehat{y_0}$。一般函数$(y_1, y_2, \dots, y_k)$的定义为：$\widehat{y_0} = \frac{1}{k}\sum_{X_i \in N_k(X_0)} y_i$其中，$N_k(X_0)$是由$X_0$的 k 个近邻组成的集合。就回归预测问题而言，$\widehat{y_0}$为近邻输出变量的平均值，而对于分类预测问题$\widehat{y_0} = P(y_0 = m|X)$，即类别为 m 的概率，最大概率值所对应的 m 作为预测值，即取众数[20]。

训练样本集中可能有许多与测试样本临近的观测点，应选择与测试样本距离最近的多少观测作为预测输出变量的依据，即近邻个数应取为多少，是 K 近邻法的关键。

K 近邻法是一种非参数分类技术，不假设数据服从一定分布，可以承受一定噪声，对于未知及非正态数据的分类的准确率较高，概念清晰且易于实现。不过 K 近邻法是懒惰算法，其计算时空开销较大，另外也存在分类过程中未考虑特征词间的关联关系且对样本库过于依赖等问题。

2.7 实验结果

2.7.1 实验环境

在我们的模型验证过程中，我们主要基 Windows 10 的操作系统，实验环境为 256G 的内存容量，8T 的固态硬盘容量，GTX 1080Ti * 4 的 GPU，主要以

Rstudio 为开发语言，以 SPSS16.0 为数据分析软件完成模型的构建。

2.7.2 评价指标

文本分类的评价可从速度和准确性两方面入手，速度可记录分类过程的时间，而评价精度可将分类结果与实际结果进行比对，尚未分类的文本则与人工分类结果进行比较。分类结果与实际结果越接近，分类准确性越高，一般用查准率与查全率作为评价分类结果与实际结果接近程度的指标[21]。

实验评价指标：

通常使用 F-Score 对分类方法进行评价：

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

其中P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

对于本文中的模型，我们采用的评价指标主要是 F1-Score。F1-Score 指标的详细定义如下：为了方便后面符号的说明定义一个混淆矩阵，如表 2-2、表 2-3 所示：

表 2-2 指标定义

真实值 \ 预测值	Positive	Negative
Positive	True Positive(TP)	False Negative (FN)
Negative	False Positive(FP)	True Negative (TN)

表 2-3 指标说明

符号	说明	预测正确与否
TP	实际为正样例,预测为正样例的个数	对(真正类)
TN	实际为负样例,预测为负样例的个数	对(真负类)
FN	实际为正样例,预测为负样例的个数	错(假负类)
FP	实际为负样例,预测为正样例的个数	错(假正类)
TP+FP	预测为正样例的个数	
FN+TN	预测为负样例的个数	
TP+FN	实际的正样例个数	
FP+TN	实际的负样例的个数	

（1） 精准率(Precision)

精准率(Precision)是反映检索系统信号噪声比的数据，通俗的理解为检出的相关类别和检出的全部类别的百分比，衡量的是检索系统的查准率。

$$P = \frac{TP}{TP + FP}$$

（2） 召回率（Recall）

召回率(Recall)是相关类别数与标签中所有全部类别数的百分比，衡量的是检索系统的查全率。

$$R = \frac{TP}{TP + FN}$$

（3） F1-Score

F1-Score 是精确率（Precision）和召回率（Recall）的调和函数，具体计算公式为：

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

其中P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

计算混淆矩阵的 R 语言函数如下：

```
matrix.hx=function(res,ref)
{
x=matrix(0,length(unique(ref)),length(unique(ref)))
for(i in 1:length(unique(ref)))
{
for(j in 1:length(unique(ref)))
{
x[i,j]=sum(ref==unique(ref)[il&res==unique(ref)[j])
}
}
X
}
```

其中，res 为代入测试样本集计算得到的预测分类结果，ref 为各训练样本实际所属类别。

模型分类的准确性可以用已有样本集的分类准确率进行估计，不过由于很多统计模型及机器学习算法难以避免的过拟合现象，对于样本集之外的样本，分类的准确率可能会降低，导致样本集分类的准确率无法完全反映分类的结果。因此，一般将样本集分为训练样本集与测试样本集，划分可采用旁置法、留一法或交叉验证。相比之下，旁置法操作最为简便，不过训练样本集一旦选定便不会改变，分类结果受数据集划分影响较大，交叉验证法保证每一个样本均可作为训练样本与测试样本，可以随机因素造成的影响，一般常用十折交叉验证，不过当数据集很大时，运行速度会较慢，留一法实现更加复杂可以看作交叉验证法的一个特例。

从预测角度考察的降维，除了从变量自身角度（如词频、文档频等）及输入变量与输出变量相关度角度（互信息、信息增益等）之外，还可以从预测角度考察变量的降维问题。以 K 近邻法为例，可以分别取互信息最大的 100、200、300.....个特征词作为输入变量，分别计算分类的准确率、召回率、F1 值等指标，并选择使得分类效果最好的特征词作为输入变量。其中计算结果：分词的查准率为 78.3243%，F1-score 为 0.6463；不分词的查准率为 73.2680%，F1-score 为 0.5703。

3. 问题 2 分析方法与过程

3.1 问题 2 分析过程

对于热点问题的研究，在发生重大事件下，舆情的监控和预防等方面都起了很大的作用。大众评论积累的庞大信息和话题，一方面可以了解群众对时事的关注度和观念意思，另一方面专家学者可以通过这一个通道来获取挖掘所需的信息和数据。因此，学术界对于热点问题的研究也越来越广泛。

查阅了今年国内外相关资料文献，可将研究成果归纳为热门微博的挖掘和热门微博的传播。ChristianWarten 是国外最早开始对热点微博进行研究的人，主张把某一个关键特征词进行聚类分析来检测微博热度[22]。同样地，国内也有诸多学者运用多种方法对热点问题进行发现，其中有代表性的是 CURE 算法[23]箱线图识别法[24]、LDA 模型[25]。目前学术界关于热点问题进行了许多研究，可是大部分的研究都是定性的，还未发现有关于问题热度的直接研究成果，也没有文献

从正面对问题的热度作出完整定义，通过什么模型或评价指标体系可以确定在特定时间段内公众反映的问题是否是热门问题的模型或评级系统依旧是个空洞。在某些评价指标构建类资料文献中，也只是采用了简单的点赞率、反对率等指标，并无法完全反映影响热点问题的因素或形成机制。在此基础上，通过阅读大量与之有关文献，分析形成热点问题的原理，建立了热点问题综合评价指标体系，还采用了因子分析方法对指标体系进行检验，旨在能够为热点问题的界定和评价提供些许理论基础。

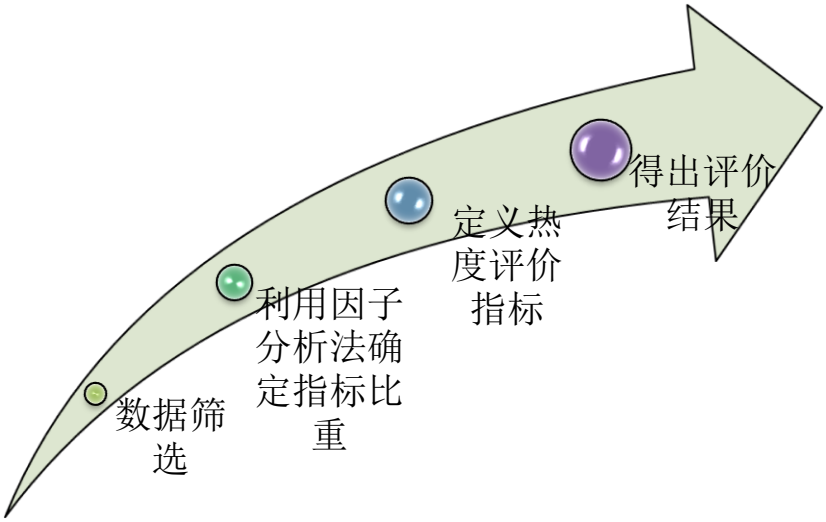

图 3-1 问题 2 流程图

问题 2 流程如图 3-1 所示，主要包括以下步骤：

步骤一：数据筛选，根据留言问题的不同种类，筛选出相对应所需的数据，并将文本数据进行整理、计算、分析。

步骤二：用 SPSS16．0 对收集的数据变量进行因子分析可行性检验，采用因子分析法，确定各层次指标的比重。

步骤三：定义热度评价指标。

步骤四：根据热度评价指标结合模型 F 得出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题 对应的留言信息，并保存为“热点问题留言明细表.xls”。

3.2 指标体系构建

3.2.1 构建原则

指标体系的组成原则一般由于不同的需求和目标的不同分为三个层面：指标选取层面通常具有三个原则:客观性原则、系统性原则和敏感性原则。客观性原则要求对象必须从客观实际出发，充分准确的反映问题的热度，并克服每个人的主观因素的影响[26]。系统性原则意味着指标体系的设计必须在整个系统中开始，并且可以包括形成热点问题的各种因素。这些指标是独立且相互关联的，形成一个有机的整体[27]。计算与操作层面一般采用数据的可用性和可操作性的原则。换句话说，在设计指标体系时，更少的指标反映了更多的实质内容，并且更易于收集和量化[28]。

针对问题热度这一特殊研究对象，在构建问题热度评价指标体系的过程中，除遵循上述三个原则外，还补充两个新的原则：趋势性原则和导向型原则。问题的热度随时都在变化，趋势性原则就是反映热点问题的趋势走向；导向型原则是指构建指标体系，不但要检测问题的热度，还要为判断热点问题提供方向指导。

3.2.2 构建理论依据

问题热度是指某一留言用户利用网络问政平台发布留言，并引起群众对该信息的广泛关注和讨论的热烈程度，它的本质是一种信息在人与人之间传递的活动。根据新闻传播学理论，信息传播活动包括 4 个因素，即信源、信宿、信道、信息。影响信息传播效果的原因有 4 个方面，即传播者、受传者、传播渠道和传播内容[29]。这类似于马尔科姆·格拉德威尔提出的流行三要素理论，他认为物体想要流行必须具备流行的基本要素，即关键人物法则、环境威力法则和内容附着力法则[30]。其实问题热度与事物流行殊途同归，关键人物法则具体到问题热度中是指留言中谁扮演着关键角色，即问题发布者、问题传播者、问题评论者。环境威力法则是指在不同环境、不同时间内发布不同的内容所引起的热效应是不同的，并且问题热度会受环境所影响。内容附着力法则取决于问题的热度是否简洁，是否可以预测预测，是否具体，是否可靠，是否具有情感和叙述性，主题是否表达清晰，长度是否合适以及是否可以引起人们注意等。

根据流行三要素和新闻传播学理论，我们将从问题传播内容和受众反映两个层面选取内容特征热度影响力、受众反映特征热度影响力两个要素对问题热度进行定量评价。

(1)内容特征热度影响力。根据流行的三要素理论，可以知道留言本身的内容特征对热度的受欢迎程度有很大的影响。 本文围绕两个方面：评论内容和内容表示方式，选取二级指标来进行反馈，它们分别为：留言信息充实度、留言出现及时率、留言内容相似度。留言信息充实度用留言详细信息的长度反映出来，内容字数越多信息量就越丰富，也越容易引发讨论。问题出现及时性主要反映问题出现的新鲜程度，大众更乐于接受关注最新的消息。留言内容相似度越高很大程度上反映这个问题越容易引起群众的关注。

(2)受众特征热度影响力。受众特征是指受众在接受到留言内容后对此所产生的反映及态度，受众的活跃度也会对问题热度产生较大影响。群众的特征可以通过受众态度的点赞率和反对数来体现。因此我们可以构建出的微博热度评价指标体系及各指标内涵如表 3-1 所示。

表 3-1 微博热度评价指标体系及各指标内涵

	一级指标	二级指标	指标内涵
热点问题热度评价指标	留言内容特征热度影响力	字数充实度	留言详情和最多字数的比率
		出现及时性	留言时间与一手时间发布时间差
		内容相似度	相似留言内容在所有留言的占比
	传受众特征热度影响力	点赞率	点赞数与所有点赞数的比率
		反对率	反对数与所有反对数的比率

3.3 问题热度因子分析

在分析问题热度因子之前，我们在进行同样可以利用词语与词权重绘制词云。通过分析文本资源，发现特定的信息，并将其以图形形式呈现出来的方法，更直观地了解到关键的信息点。词云常用来对文本的关键字进行可视化，一般以字体的大小区分词语的重要程度，字体越大表示词语在文本中越重要。如图 3-2：

图 3-2 词云图

图 3-2 反映了全部类别的文本集不同特征词的重要程度，通过观察词云可以发现，城乡建设板块词语分布较为集中，根据数据显示：“小区”词频为 511，“扰民”词频为 265，“街道”词频为 196，“投诉”词频为 191，该板块留言一般面向城乡居民群众，与之相比，劳动和社会保障板块与城乡建设板块雷同的词的权重仍然不低但重要程度大幅下降，而环境保护及文体教育板块已见不到城乡建设板块重要词语出现。除此之外，城乡建设板块其他重要特征词还包括“施工”、“噪音”、“社区”等。

多元线性回归模型只能够简单地解释因变量与自变量的关系，但是模型中只能涵盖它的部分自变量，忽略了其他变量所提供的信息，不仅造成了信息损失和资源的浪费，而且还可能导致解释结果不准确。为了解决这一问题，我们采用因子分析法对运用实例对以上设计的评价指标体系进行分析，核心思想就是运用因子分析提取公因子，公因子尽可能包含所有指标，进而以公因子作为自变量对因变量 Y 进行解释，并最终根据模型结果计算出热点问题排名。具体分析过程如下：

因子分析法提取公因子，通过因子分析的原理可得，我们需要用 SPSS16．0 对收集的数据变量进行因子分析可行性检验，样本 KMO 检验值为 0．876，可做因子分析；Bartlett 球形检验近似卡方值为 65．84，显著性概率为 0．000，小于 1%，表明统计数据适合做因子分析。按照因子分析中主成分特征值提取原则，获取 2 个公因子 F_1、F_2 (表 3-2)。可以进一步发现二个公共因子能够分别解释热

度相关信息的 68．007％和 17．693％，最后和总体信息的 85．7％相一致，由于 2 个主因子保留大部分原始数据，所以我们可以用它来对问题热度进行评价。

表 3-2 主因子特征值、累计方差贡献率

主因子	特征值	方差贡献率%	累积方差贡献率%
F_1	6.243	68.007	68.007
F_2	1.462	17.693	85.7

为准确解释命名各主因子，我们可以用最大方差法对因子载荷矩阵进行正交旋转，那么就可以得到旋转因子载荷矩阵，见表 3-3。

表 3-3 因子载荷矩阵

变量	因子载荷矩阵		旋转因子载荷矩阵	
	F_1	F_2	F_1	F_2
X_1	0.702	0.175	0.313	0.125
X_2	0.895	0.231	0.298	0.792-
X_3	0.529	0.365	0.615	0.288
X_4	0.954	0.031	0.054	0.954
X_5	0.977	-0.074	0.124	0.977

根据旋转因子载荷矩阵可知， X_1 字数充实度、X_2 出现及时性、X_3 内容相似度的载荷在第一个主因子 F_1 所占比例都很高，它表示留言的内容信息具体详情，以及表现方法和时间特点，我们可以理解为内容特征因子。X_4 点赞数、X_5 反对数载荷在第二个主因子 F_2 所占比例都很高，它表示受众对留言内容的反应情况，我们可以理解成为受众特征。

综合因子计算方程

因为有旋转因子载荷矩阵中各指标与公共因子的相关系数值，不妨列出的 F_1、F_2 的计算方程，分别见公式(3-1)和公式(3-2)。

$$F_1=0．313X_1+0．298X_2+0．615X_3 \quad (3-1)$$

$$F_2=0．954X_4+0．977X_5 \quad (3-2)$$

接着表 3 中三个公共因子的方差贡献率对留言信息的热度影响力采用累加，从而列出最终综合因子的表达式，见公式(3-3)。

$$F=0．7934F_1+0．2065F_2 \quad (3-3)$$

得到公式后，制定评分标准，并利用 Excel 软件对数据进行处理。对于 X_1 字数充实度，利用 LEN 函数进行计算；对于 X_2 出现及时性与 X_3 内容相似度，借

助函数 GetMatchingDegree(Text_a, Text_b)比较两两字符串的相似度，查找相似内容后进行时间跨度计算并评分，对于 X_4 点赞率与 X_5 反对数，将其排序后进行评分。评分标准如下表;

表 3-4 字数充实度评分

范围（字数）	0-99	100-999	1000 以上
评分（分）	取前一位为 x_1 得分	取前两位为 x_1 得分	100

表 3-5 出现及时性评分

范围（时间跨度/相隔月数）	0-1	1-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	≧18
评分（分）	100	90	80	70	60	50	40	30	20	10	0

表 3-6 内容相似度评分

范围（相似留言数/条）	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	≧90
评分	10	20	30	40	50	60	70	80	90	100

表 3-7 点赞数、反对数评分

范围（数量）	0	1-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	≧100
评分	0	10	20	30	40	50	60	70	80	90	100

通过上面求得的热度综合得分计算公式，将实证选取的问题进行排序，用来检验指标体系的合理性和实用性，结果见表 3-8。

表 3-8 热点问题表

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	64.16224	2019/1/11-2019/5/28	A 市市民	A 市 58 车贷特大集资诈骗案
2	2	63.49658	2019/3/26-2019/4/12	A6 区月亮岛路的居民	A6 区月亮岛路沿线架设 110kv 高压线杆
3	3	63.47742	2019/11/2-2020/1/26	A 市暮云街道丽发新城社区居民	A 市暮云街道丽发新城社区搅拌站灰尘，噪音污染严重
4	4	62.94992	2019/8/23-2019/9/6	A 市伊景园滨河苑居民	A 市伊景园滨河苑捆绑销售车位
5	5	58.72824	2019/7/7-2019/9/1	A5 区魅力之城小区居民	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气

3.4 不同类型热度因子构成

为进一步验证留言内容特征的热度影响力，文章选取不同类型的留言分别进行验证。我们将附件 3 中城乡建设类、教育文体类、交通运输类和民政类等 14 种类型的问题，运用热点问题热度评价指标体系和因子分析方法对影响这 14 类问题热度的两个维度进行比较分析，比较结果见图 3-3。

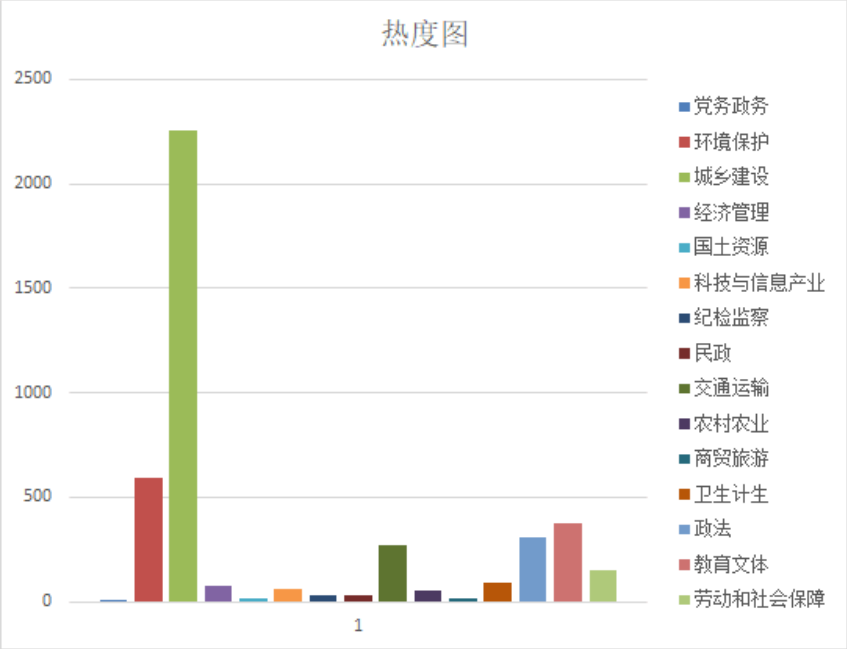

图 3-3 热度图

由图3—3可知，民众反映的问题中，城乡建设类问题占到了总问题的52.06%，其次是环境保护类占到了13.62%，教育文体类占到了7.12%，可由此增加城乡建设类方面有关的工作人员，以便有效地解决民众的问题。此结果符合表3-8 热点问题表中排名前五热点问题的类型，验证了该综合因子模型的可行性。

3.5 实验结果

文章在借鉴前人研究的基础上，依据新闻传播学和流行三要素理论从受传者和传播内容两个维度，构建了包2个要素、5个指标的评价体系，我们使用因子分析方法和实例，从实证的角度检验指标体系的合理性和效率。最后，在这种方法的基础上，对不同类型问题热度来源成分进行了对比和解析。总结以下几点结论：

(1)目前学术界关于热门微博和微博热点话题的研究成果较多，但并没有直接关于问题热度的研究成果，也没有文献可以完整定义问题的热度。在有关的评价指标构建类资料文献中，也只是采用了简单的转发率、评论率等指标，并不能综合性地反映问题热度的影响因子或形成机理，因此依据相关理论构建一套完整的评价问题热度的指标体系很有必要。

(2)通过对问题热度测度排名的发现，评价指标体系估计的问题热度排名符合，说明了指标体系的合理性。估计结果表明，传播内容对热度的影响力最大，即在本文的模型中表现为相似度对热度的影响力最大，而接收者对微博热度的影响作用较小。

(3)对城乡建设类、教育文体类、交通运输类和民政类等14种类型的留言内容进行一一检验，结果表明它们影响热度的维度都不大相同，其中城乡建设类、环境保护类、教育文体类的问题居多。

4. 问题3分析过程与结果

4.1 问题3分析过程

当今社会经济的日益发展，现代社会渐渐的进入以服务为导向的新领域，服

务这一行业在众多的行业中脱颖而出，成为新时代的一个霸主。不仅如此，越来越多的人也慢慢的开始关注服务质量问题，同时服务质量问题还是学术界研究的热点之一。本题提到的相关部门对留言的答复意见的质量可视为服务质量问题之一。早期之前，国外的学者已经在服务质量问题的领域上做了一些杰出的贡献。但是，我国服务质量问题的研究才刚刚开始，主要把重点放在了宏观层次的产业结构分析上，对综合服务质量评价体系和标准体系建设的研究还不够，具体实践指导不足。构建科学合理的服务质量体系及标准体系，不仅仅让服务业服务质量评价标准一致，而且我国服务业服务质量水平和发展潜力也随之体现出来。在它的前提之前，我们就建立起了三维质量评价体系，还对质量标准体系结构进行探究。

4.2 基于三方评价的答复质量评价体系研究

4.2.1 答复质量评价体系框架

1984 年，格罗鲁斯教授第一次提出了顾客感知服务质量的概念[31]，对比了客户意识的服务质量定义为客户对服务的期望与感知的服务性能之间的差异。从格罗鲁斯对服务质量的定义来看，服务质量包括技术质量(即服务结果)和功能质量(即服务过程)。基于此，从三维评价的概念进行评估，即从服务的接受方、服务提供方以及第三方的角度同时进行评价。其中，群众评价的主体是群众，评价对象为答复绩效；答复提供方评价的主体是政府相关部门，评价对象为答复过程；第三方评价的主体是第三方机构，评价对象为答复的服务能力，为对答复提供方所提供的答复质量进行认证和认可。完成三方评价之后，将创建相应的评价指数，并对评价指数进行加权作为评价的结果。答复质量评价的整体框架如图 4-1 所示：

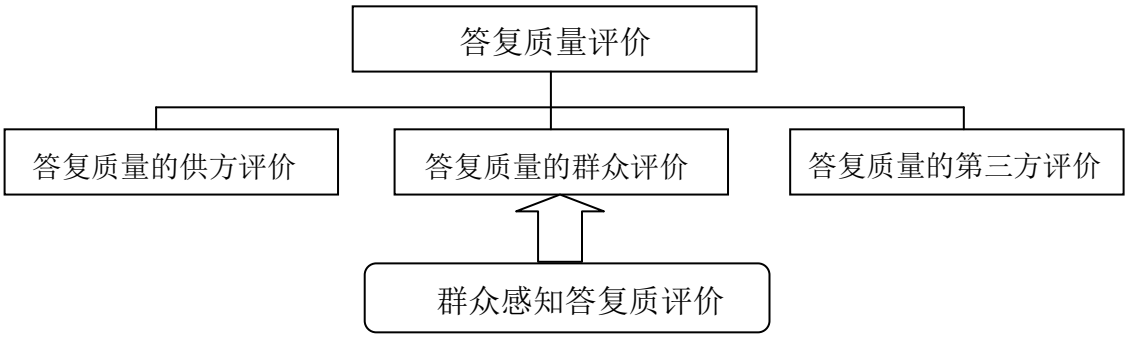

图 4-1 答复质量评价的整体框架

4.2.2 答复质量的群众评价

答复质量的结果是否满足群众的需求.需要由群众进行评价。答复质量的顾客评价即是从社会角度,从群众认知方面来进行的评价,是对答复质量的基本评价和基本测量[32]。

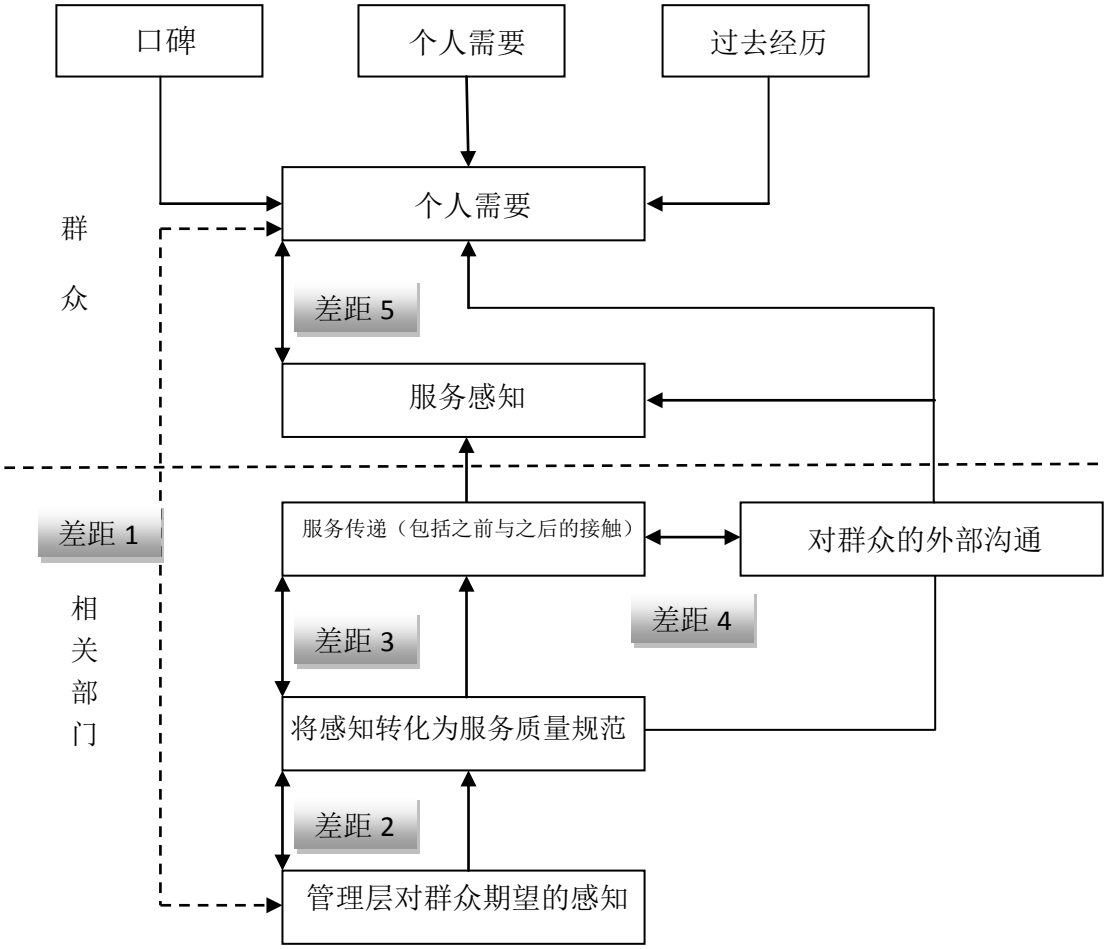

图 4-2 PZB 服务质量模型

查资料得知，以美国营销学家帕拉休拉曼(A.Parasuraman)，赞瑟姆(ValarieAZeithamal)和贝利(LeonardL.Berry)等人提出的 PZB 服务质量模型（图 4-2）为理论依据，服务质量差距模型(Service Quality Model)，也称 5GAP 模型，结合我国国情的特点与现代社会环境特点,我们形成了以群众为评价主体、有关政府部门提供的服务为评价客体、群众对政府相关部门提供答复质量的感知作为评价结果的新的答复质量评价指标体系。具体指标如表 4-1 所示:

表 4-1 答复质量评价指标体系

一级指标	二级指标	三级指标
答复质量指标体系	可靠性	1.对群众承诺的履行情况
		2.群众遇到困难时,政府职能部门给予的关注程度可靠性
		3.政府职能部门的可靠性
		4.提供所承诺的服务的相关性
		5.答复信息的可解释性
		6.相关服务资料记录与保存的完整性
	响应性	7.让群众清楚地了解提供服务的准则
		8.沟通渠道的便利性
		9.群众得到所需服务的迅速性
		10.部门服务人员帮助群众的回复态度
		11.部门服务人员提供服务的及时性
	保证性	12.部门服务人员值得信赖的程度
		13.服务过程中群众的放心程度
		14.部门服务人员的礼貌程度
		15.政府对部门服务人员提供服务的支持程度
		16.信息沟通渠道的畅通程度
	移情性	17.提供服务的个性化程度
		18.部门服务人员给予群众个别关怀的程度
		19.部门服务人员了解群众需求的程度
		20.优先考虑群众利益的程度
		21.提供的服务时间符合所有群众需求程度

指标体系中的一级指标即答复质量;二级指标；根据 PZB 服务质量 5 个维度进行修改设立,即可靠性、响应性、保证性和移情性;三级指标在 PZB 的 22 项指标基础上做出了一些改正。由于信息技术的不断发展，信息通信已成为影响服务质量的另一个重要元素，因此，在设置指标的过程中，与信息有关的指标被添加到三个维度中，它们是：可靠性，响应性和保证性。例如，可靠性维度增加了响应信息的可解释性，保证性维度中添加了信息渠道的畅通程度等。

指标的评价尺度采用 Liken7 级量表[34], 将 21 个三级指标转换为受访者易于理解与判断的语言，由群众接受服务后对答复服务质量的实际感受情况进行打分, 7 分表示非常好, 1 分表示非常不好, 中间值代表对服务质量水平不同程度的判断。

4.2.3 答复质量的供方评价

作为答复提供方，政府相关部门要对答复提供过程进行检验。答复流程中的每一道工序都为其上一道工序提供基础保障, 因此我们将评估纳入组织内部, 用来确定服务是否满足群众的需求以及它是否满足答复规范的要求，以发现答复流程中出现的差异并作出响应, 给改善答复质量指明方向。

从 PZB 的服务质量缺口理论可知, 由于企业内部存在的“五个缺口”(GAP)导致了服务质量出现了问题, 依 “五个缺口”建立评价指标体系可得, 如表 4-2 所示:

表 4-2 答复质量供方评价指标体系

指标	定义
管理者认知	评价群众期望与群众感知间的差异
管理者准则	评价群众期望的感知管理与答复质量规格间的差异
答复传递	评价答复质量规格与实际答复传递间的差异
内部沟通	评价实际答复传递与外部群众沟通间的差异
答复质量	评价群众期望的答复质量与感知的答复质量间的差异

相关政府部门的内部评估可以通过专家评分法进行，其中由主要政府的主管和主要环节的核心人物组成专家组来进行评分。

4.2.4 答复质量的第三方评价

第三方对答复质量的评价是规范的要求独立机构需要客观地进行评估，在答复提供过程到答复结果的过程对答复质量是否符合规定要求的评价。答复服务过程与答复服务结果共同组成评价指标, 不同的指标体系，我们就应该选择不同的评价方法, 可以用结构方程模型或者多元统计分析及 AHP 方法[33]等。

4.2.4.1 层次分析法的特点

20 世纪 70 年代，美国著名运筹学家、匹兹堡大学教授 T.L.Saaty 提出了一种以定性与定量相结合，系统化、层次化分析问题的方法，称为层次分析法（analytichierarchyprocess，AHP），它的主要特点是定性与定量分析相结合，以定量的形式表达人类主观判断并进行科学处理。

4.2.4.2 确定指标权重的基本步骤

4.2.4.2.1 建立层次结构模型

层次分析法中的一项重要任务就是将问题分为多个组成部分或元素，并根据不同的特征将这些元素分为不同的组，每个组构成一个层，并且这些层彼此不相交。使用同一级别的元素作为标准，并支配较低级别的全部或部分元素。当然它们还受上一层元素的约束，因此它们形成了从上到下的分层治理结构。在设计指标时，确定并建立总体系统评估指标的层次模型，从整个系统开始，对影响总体性能的相关因素进行层次分析，然后下层指标需要与上层指标相一致的原则，进行全面优化，来确定并建立总体系统评价指标递阶层次结构模型。

4.2.4.2.2 构造成对比较判断矩阵

在建立了指标的递阶层次结构后，需要比较各层次若干个元素对上一层次元素的影响，从而确定它们在目标中所占的比重。

假定上一层次的元素C_k作为准则，下一层次的元素 D_1，D_2，…，D_n 有上一层的支配关系针对C_k两个元素 D_i 和 D_j 重要程度，用数字呈现出来。由这些数值作为矩阵中的元素，而构成两两元素比较差别矩阵（见图 4-3）。其中a_{ij}表示对C_k而言，D_j 比 D_i 相对重要性的程度数值（即：$a_{ij}=D_j/D_i$）。由此可知：$a_{ii}=1$，$a_{ij}>0$，$a_{ij}=1/a_{ji}$。

$$
\begin{array}{cccccc}
C_i & D_1 & D_2 \cdots & D_3 & \cdots & D_n \\
D_1 & a_{11} & a_{12} \cdots & a_{1j} & \cdots & a_{11} \\
D_2 & a_{21} & a_{22} \cdots & a_{2j} & \cdots & a_{2n} \\
\cdots & \cdots & \cdots \cdots & \cdots & \cdots & \cdots \\
D_i & a_{i1} & a_{i2} \cdots & a_{ij} & \cdots & a_{in} \\
\cdots & \cdots & \cdots \cdots & \cdots & \cdots & \cdots \\
D_n & a_{n1} & a_{n2} \cdots & a_{nj} & \cdots & a_{11}
\end{array}
$$

图 4-3 判断矩阵的一般形式

采用 1－9 之间整数及其倒数比例标度法，见表 4-3 对元素的重要性进行判断。

表 4-3 标度的含义

标度	含义
1	表示两元素相比，具有相同重要性
3	表示两元素相比，前者比后者稍微重要
5	表示两元素相比，前者比后者明显重要
7	表示两元素相比，前者比后者强烈重要
9	表示两元素相比，前者比后者极端重要
2、4、6、8	表示上述相邻判断的中间值
上述数值的倒数	若元素 D_i 与元素 D_j 的重要性之比为 a_{ij}，则元素 D_j 与元素 Di 重要性之比为 $a_{ji}=1/a_i$

4.2.4.2.3 层次单排序及一致性检验

判断 n 阶矩阵，单根是它的最大特征根，且 $\lambda_{max}\geq n$。当 $\lambda_{max}=n$，并且其余特征根均为 0 时，可以判断该矩阵具有完全一致性。

1）计算一致性指标 CI：

$$CI=\frac{\lambda max-n}{n-1} \qquad (4-1)$$

2）查找对应的评价随机一致性指标 RI。n 的值为 1~9，RI 的取值见表 4-4。

表 4-4 阶矩阵 RI 的取值

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

要求 RI 的值，我们采用随机方法构造 600 个样本矩阵，随机从 1－9 及其倒数中抽取数字构造正互反矩阵，求得最大特征根的平均值 λ_{max}，并定义：

$$CI=\frac{\lambda max-n}{n-1} \qquad (4-2)$$

3）计算一致性比例 CR：

$$CR=\frac{CI}{RI} \qquad (4-3)$$

当 CR＜0.1 时，则判断矩阵的一致性是在合理的范围之内，相反判断矩阵应该加以修改。

4.2.4.2.4 层次总排序及一致性检验

层次单排序是一组元素对其上一层中某元素的权重向量。最后，需要获取整个目标的每个元素的权重，尤其是最底层的元素权重。总排序权重要自上而下的将单准则下的权重进行合成。

设上一层次(A 层)包含 A_1，A_2，…，A_m 共 m 个因素，它们的层次总排序权重分别为 a_1，a_2，…，a_m。又设其后的下一层次(B 层)包含 n 个因素 B_1，B_2，…，B_n，它们关于 A_j 的层次单排序权重分别为 b_{1j}，b_{2j}，…b_{nj}(当 B_i 与 A_j 无关联时，b_{ij}＝0)，则 B 层中各元素关于总目标的权重，即 B 层元素的层次总排序权重 b_1，b_2，…，b_n 按表 4-5 所示方式进行，即：

$$bi=\sum_{j=1}^{\infty} bij \cdot aj，\ i=1，\ …,n \qquad (4-4)$$

表 4-5 3B 层总排序权值

A 层 / B 层	A_1	A_2	…	A_n	B层总排序权值
	A_1	A_2	…	a_n	
B_1	b_{11}	b_{12}	…	b_{1n}	$b=\sum_{j=1}^{\infty} b1j \cdot aj$
B_2	b_{21}	b_{22}	…	b_{2n}	$b_i=\sum_{j=1}^{\infty} b2j \cdot aj$
…	…	…	…	…	…
B_n	b_{n1}	b_{n2}	…	b_{nn}	$b_i=\sum_{j=1}^{\infty} bij \cdot aj$

采用一致性检验来对层次总排序进行检验，还是遵循由上至下一层一层检验的原则。

设 B 层中与 Aj 相关的因素的成对比较判断矩阵在采用一致性检验单排序的过程中，解出单排序一致性指标为 CI(j)(j＝1，3，…，m)，与之对应的平均随机一致性指标为 RI(j)、CI(j)、RI(j)已在层次单排序时求得)，则 B 层总排序随机一致性比例为：

$$CR=\frac{\sum_{j=1}^{\infty} CI(j)aj}{\sum_{j=1}^{\infty} RI(j)aj} \qquad (4-5)$$

当 CR＜0.1 时，认为层次总排序结果具有较满意的一致性并接受该分析结果。层次分析法能统一处理规划中的定性定量因素，具有系统性、适用性、实

用性和简洁性。

4.3 答复质量标准体系研究

在我国服务质量标准化的政策和标准前提下，遵循适用性，科学性，系统性，先进性，兼容性和开放性（也称为未来或可扩展性）的原则，建立响应质量标准体系的总体结构及层次系统。如图 4-4 所示：

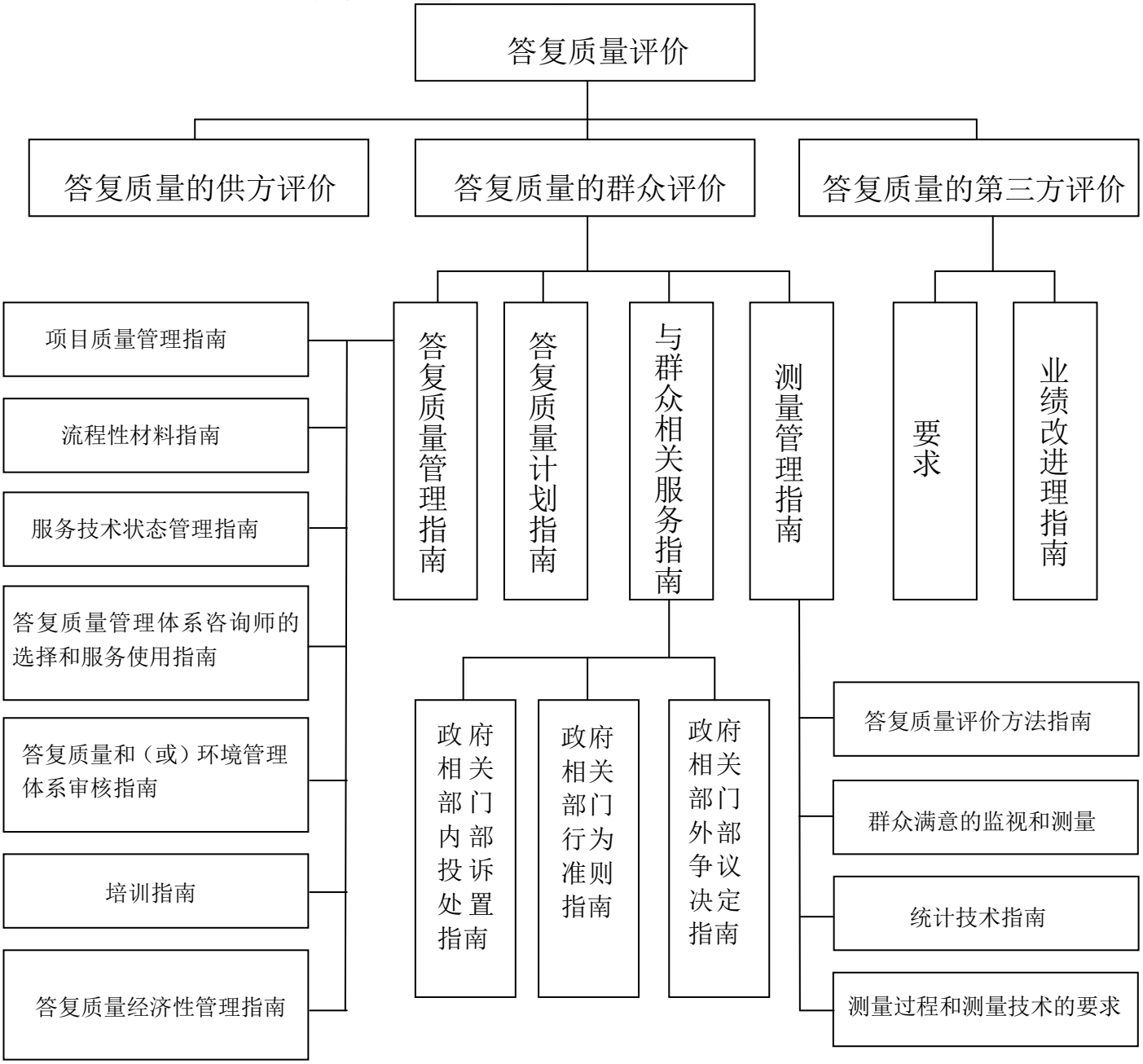

图 4-4 答复质量标准体系的总体架构及层次结构

服务质量标准体系包括三个层次：基础和术语，质量管理体系和技术支持。基本和术语级别不算在内，其他级别也分别受约束、交互，相互依存和相互补充的几个子系统。

4.3.1 基础和术语

基本知识和术语主要包括与答复质量相关的基本术语。 为了将术语，术语和技术词汇集成到答复质量管理中，响应服务行业可以分为特殊术语和基本术语。

4.3.2 质量管理体系

质量管理体系[35]包括要求与业绩改进指南两个分体系。这些要求包括答复服务质量管理的集成整体要求。 绩效改进指南提供了有关答复质量管理系统的有效性和有效性的指南，以改善组织的绩效并提高公众和其他相关方的满意度。

4.3.3 技术支持体系

技术支持体系主要为建立和实施答复质量体系提供指导与技术支持,分为 4 个类别共 14 个分体系:

(1)答复质量管理指南，包括项目质量管理指南、流程性材料指南、服务技术状态管理指南、答复质量管理体系咨询师的选择和服务使用指南、答复质量和(或)环境管理体系审核指南、培训指南、答复质量经济性管理指南；

(2)答复质量计划指南；

(3)测量管理指南，包括答复服务质量评价方法指南。群众满意的监视和测量、统计技术指南测量过程和测量技术的要求；

(4)与群众相关的服务指南，包括政府相关部门内部投诉处置指南、政府相关部门行为准则指南、政府相关部门外部争议决定指南。

5．结论

本文是团队三人两个月以来对于群众留言的文本数据挖掘分析，从认知到应用方法解决全程的总结，本次数据挖掘的目标是利用互联网公开来源的群众问政留言记录，借助 Excel、Rstudio 等软件利用 jieba 中文分词工具对留言内容进行

分词和去停用词、K-means 聚类的方法及 KNN 算法 、SPSS16．0 对收集的数据变量进行因子分析等。

文章可分为三个部分：

1）利用 Rstudio 软件对文本数据进行必要的预处理之后，采用分词和不分词两种方法提取特征词，采用互信息作为筛选特征词的索引,计算具有最大互信息的特征词的 tf-idf 权重,并构建词权-文档矩阵。最后,我们使用 K 邻近法对测试留言文本进行分类。其中分词的查准率为 78.3243%，F1-score 为 0.6463；不分词的查准率为 73.2680%，F1-score 为 0.5703。

2）为了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，通过因子分析法来确定问题热度评价指标体系中各层次指标的权重，定义合理的热度评价指标，并给出评价结果。排名在前五的热点问题主要为城乡建设类与环境保护类。

3）以 PZB 服务质量模型与中国国情与现代社会环境的特征相结合进行的一项新的反应质量评估,构建了评价主体是群众，评价客体则是有关政府部门提供的评估服务、群众对政府相关部门提供答复质量的感知作为评价结果的新的答复质量评价指标体系。

总的来说，各类社情民意相关的文本数据量不断攀升，大数据、云计算、人工智能等新时代技术的不断发展，构建基于自然工程技术的智能管理系统(NLP)已成为创造和发展高动力社会治理的新政策。因此，为了提高政府相关部门处理留言的有效性，本文构建基于自然语言处理技术的智慧政务模型来解决此类问题。

6. 参考文献

[1]朱德熙. 语法讲义. 北京:商务印书馆, 1982

[2]G.Salton.Automatic Text Processing.Wesley Publishing Company, 1988

[3]曹卫峰. 中文分词关键技术研究[D]. 南京理工大学, 2009.

[4]黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3):8-19.

[5]代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1):26-32

[6]黄萱菁, 吴立德, 王文欣, 等. 基于机器学习的无需人工编制词典的切词系 统[J]. 模式识别与人工智能, 1996, 9(4):297-102.

[7]周水庚, 关佶红, 胡运发, 等. 一个无需词典支持和切词处理的中文文档分 类系统[J]. 计算机研究与发展, 2001, 38(7):839-844.

[8]周水庚, 关佶红, 胡运发. 无需词典支持和切词处理的中文文档分类[J]. 高技术通讯, 2001, 11(3):31-35

[9]http://www.matrix67.com/blog/archives/5044.

[10]费洪晓, 康松林, 朱小娟, 等. 基于词频统计的中文分词的研究[J]. 计算机工 程与应用, 2005, 41(7):67-68.

[11]卞晓增. 微博新词发现与新词情感倾向性研究[D]. 云南大学, 2016. .

[12]杨杰明. 文本分类中文本表示模型和特征选择算法研究[D]. 吉林大学, 2013.

[13]宋惟然. 中文文本分类中的特征选择和权重计算方法研究[D]. 北京工业大 学, 2013.

[14]廖一星. 文本分类及其特征降维研究[D]. 浙江大学, 2012.

[15]卜凡军. KNN 算法的改进及其在文本分类中的应用[D]. 江南大学, 2009.

[16]Yang Y,Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]//Fourteenth International Conference on Machine Learning.Morgan Kaufmann Publishers Inc. 1997:412-420.

[17]王煜. 基于决策树和 K 最近邻算法的文本分类研究[D]. 天津大学, 2006.

[18]Soucy P,Mineau G W. A simple KNN algorithm for text categorization[M].2001.

[19]Chen Y Q,Nixon M S,Damper R I. Implementing the k-nearest neighbour rule via a neural network[C]//IEEE International Conference on Neural Networks,1995.Proceedings.IEEE, 1995:136-140 vol. 1.

[20]薛薇. R 语言数据挖掘[M]. 中国人民大学出版社, 2016.

[21]陈晓云. 文本挖掘若干关键技术研究[D]. 复旦大学, 2005.

[22]ChristianWa~ena,RogierBrussee.Topicdetectionbyclusteringkeywords[C]//WashingtonDC.USA:ProceedingsofthelthInternationalConferenceonDatabaseandExpertSystemsApplication，2008：54—58.

[23]杨长春，周猛，叶施仁，等. 基于改进 CURE 算法的微博热点话题发现[J]. 计算机仿真，2013，30(11)：383—387.

[24]郭德清，廖祥文. 基于箱线图的微博客热点话题发现[J]. 山西大学学报，2014，37(1)：19. 25.

[25]唐晓波，向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作，2014，58(5)：58—63.

[26]李文杰，化存才. 网络舆情信息的灰色预测及案件分析[J]. 情报科学，2013，31(12)：51-56.

[27]庞彦军，刘开第，张博文. 综合评价系统客观性指标权重的确定方[J]. 系统工程理论

与实践，2001(8)：3742.
[28]雯静,许鑫,陈正权.网络舆情指标体系设计与分析[J].情报科学,2009,27(7):986—991.
[29]邢雨晴，刘红翠，周瑞．微博信息传播模式及其应用的实证研究的文献综述[J]．中国外资，2012(18)：275．277.
[30]马尔科姆·格拉德威尔．流行三要素[M]．北京：中信出版社，2009.
[31].A Parsuramen,VA Zeithaml,LL Bery. A conceptualmodel of service quality and its implications for futurererh[J]. Journal of Markeing, 198.149>41-50 .
[32]高充彦， 贾建民.顾客满意度不确定性对服务质量评价的影响[J].管理科学学报2007,(10):39-47
[33]张平，汤良.基于 AHP 的服务质量评价指标方法[J].技术经济与管理研究,2006,(4):39- 40
[34]徐明， 于君英． SERVQUAL 标尺测量服务质量的应用研究[J].工业工程与管理,2001,(6):6-9
[35]任冠华，陈淑仪,魏宏,等烟草行业信息化标准体系研究[J].世界标准化与质量管理,2007,(3);17-21