Advances in Applied Mathematics 应用数学进展, 2022, 11(7), 4526-4536
Published Online July 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.117479

Hans汉斯

基于向量空间和PZB服务质量模型的热点问题分析

张梓坪，林　莹，陈梦靖

闽江学院数学与数据科学学院(软件学院)，福建　福州

收稿日期：2022年6月11日；录用日期：2022年7月3日；发布日期：2022年7月14日

摘　要

本文基于K近邻与因子分析法对所收集的热点问题进行分析，并且将我国国情的特点和现代社会环境特点相结合，构建了一种新的答复质量评价指标体系。新的答复质量评价指标体系是以群众为评价主体、相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。

关键词

K邻近法，因子分析法，智慧政务

Analysis of Hot Issues Based on Vector Space and PZB Service Quality Model

Ziping Zhang, Ying Lin, Mengjing Chen

College of Mathematics and Data Science (Software College), Minjiang University, Fuzhou Fujian

Received: Jun. 11th, 2022; accepted: Jul. 3rd, 2022; published: Jul. 14th, 2022

Abstract

This paper analyzes the collected hot issues based on the K-nearest neighbor and factor analysis method, and combines the characteristics of China's national conditions with the characteristics of the modern social environment to construct a new response quality evaluation index system. The new response quality evaluation index system takes the masses as the subject of evaluation, the services provided by relevant government departments as the object of evaluation, and the masses' perception of the quality of responses provided by relevant government departments as the evaluation results.

文章引用: 张梓坪, 林莹, 陈梦靖. 基于向量空间和 PZB 服务质量模型的热点问题分析[J]. 应用数学进展, 2022, 11(7): 4526-4536. DOI: 10.12677/aam.2022.117479

Keywords

K Proximity Method, Factor Analysis Method, Smart Government

1. 引言

近年来，微信、微博、市长信箱、阳光热线等网络问卷渐渐地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。各种社会形势和舆论相关的文本数据量不断增加，相关部门的工作面临着巨大的挑战。与此同时，在新时代背景下新兴技术不断进步，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。

为了提高政府相关部门处理留言的有效性，本文基于 K 近邻与因子分析法对所收集的社区热点问题进行分析，同时将 PZB 服务质量模型(Service Quality Model)与中国国情、现代社会环境的特征相结合，进而形成评价主体是群众、评价客体是有关政府部门提供的评估服务的一项全新的反应质量评估体系。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. 研究现状

智慧政务模型

(一) DPSIR 模型

DPSIR 理论模型在资源、人口、生态可持续发展研究中具有重要作用。DPSIR 模型能够更好地反映人类活动与信息生态环境之间的关系和作用，为建立智能政府信息生态评价体系提供了基本的分析框架。该模型具有以下优点：第一，DPSIR 模型非常适合于智能政府信息生态的研究。第二，DPSIR 模型能够揭示智能政府信息生态的因果关系。第三，DPSIR 模型可以为智能政府信息生态的发展提出战略建议[1]。

(二) SERVQUAL 模型

基于“顾客服务质量感知”的研究背景，PZB 于 1985 年提出了“服务质量差距模型”。1988 年，PZB 将 10 个服务要素的数量减少到 5 个服务要素，并总结了服务质量的五要素模型，即 SERVQUAL 服务质量差距模型。1991 年，PZB 又将 SERVQUAL 测量模型重新进行了改良和修正，提高了这种模型的可靠性和效率[2]。

(三) TOE 理论框架

TOE 理论框架(Technology-Organization-Environment)，由学者 Tornatzky 和 Fleischer 于 1990 年提出，具体而言，TOE 理论框架认为应该从技术因素、组织因素和外部环境因素三个方面去探索一项组织创新技术应用的影响因素，TOE 框架为研究政务服务智慧能力关键影响因素提供了一个很好的基础框架，即从技术、组织、环境三个方面进行因素选择和探讨，然而在实际操作中需要注意，TOE 模型更多是一种因素划分或者因素归类的框架，模型本身并没有对各个维度的因素予以具体规定，即不能使用该框架作为研究模型当中因素选择的理论基础或者动机分析[3]。

本文增加新的自动答复模型，不及能够快速接收各种政务问题，而且能够及时快速回复群众，给群众以良好的使用感受。

3. 模型构建

图 1 为本文模型构建整体步骤框架，第一步使用 Rstudio 对文本数据进行分析和预处理，提取文本特征词主要使用到了文本分词方法来进行综合提取。第二步确定索引，使用信息增益算法作为所选特征词的索引，计算信息增益最大的特征词的 TF-IDF 权重，构造词权－文档矩阵。第三步使用 k-最近邻法对留言的文本内容进行综合分类。

为了对反映特定地点或特定人群在一定时间内的问题的信息进行分类，采用因子分析法确定问题热评价指标体系中各指标层次的权重，定义合理的热评价指标，并给出评价结果。

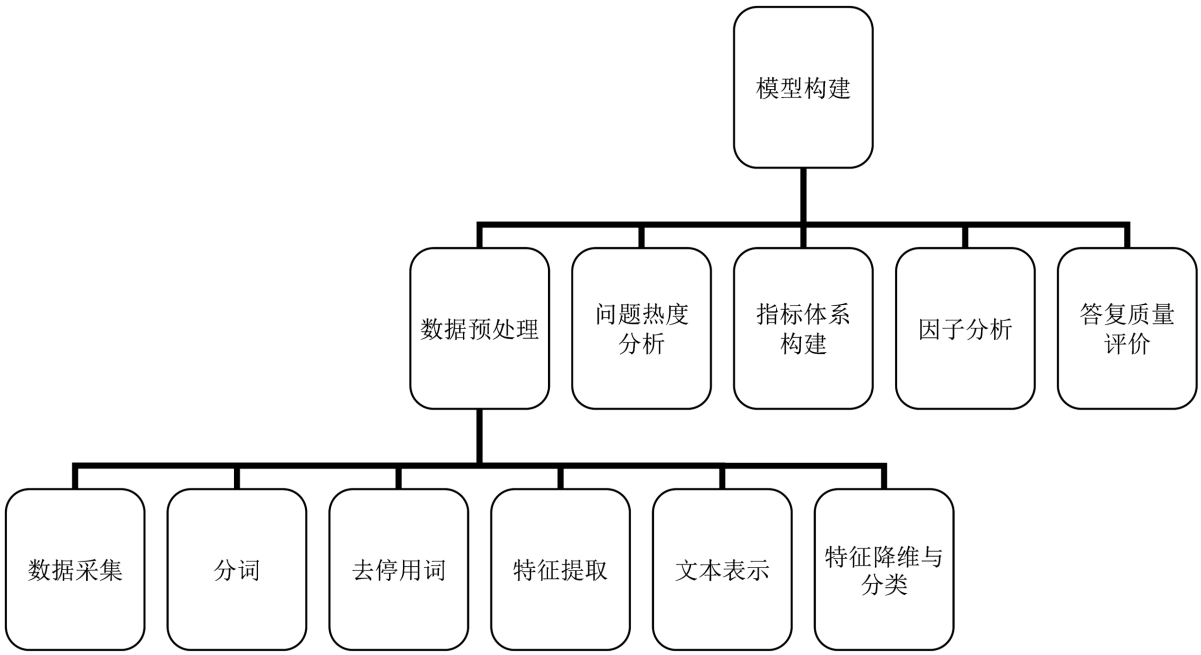

Figure 1. Model building overall framework
图 1. 模型构建整体框架

3.1. 数据预处理

(一) 数据采集

本文选择的数据集来源于中国国家统计局的社区留言总计 9210 条，采用旁置法来选择实验样本。采集并预处理后的各专题文本分布如表 1：

Table 1. Text set category distribution table
表 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

(二) 分词

由于手工编写的分词函数功能没有得到充分的优化，分词工作效率相对较低，本文采用 jieba 软件包来完成分词部分。较为常用的中文分词技术是 jieba 分词，对文本的分词准确率较高[4]。

(三) 去停用词

通过建立停用词列表来删除停用词。停用词表选取哈工大停用词表和百度停用词表组合[5]。

(四) 特征提取

采取基于特征词典的分词方法即基于特征词和字符串相互匹配的分词方法切分文本进而提取一个特征词，根据这个词典中的某些规则，找出对应的一个给定词典中单词和特征句的文本。如果当文本中的段落文字与词典中的文本对应成功，段落将被截断，并视其为一个单词，之后把这个文本中的其他组成部分重新根据此规则对其进行再次对应，直到整个文本完成。

(五) 文本表示

分词的目的是获取一组特征词。每个特征词代表向量空间中的一个维度。接下来，确定文本在各个维度上的坐标位置的数值，不妨把这个坐标位置的值叫做词权。

提取特征项并确定词权重后，文本集中的文本可以由空间中的节点表示。文本相似性通常被认为是一个文本和另外一个文本之间的关系程度的一种衡量。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。数据经过分词和去停用词后，若想要查看分词效果如何，可以绘制词云图来进行检验。画图的第一步利用 Python 先统计分词后各词语出现的频数，结果展示见表 2；接着依据频数按从大到小进行排序；最后用 Python 工具中的 wordcloud 库进行编码绘图，最终绘制出来，如表 2：

Table 2. Keyword word frequency table
表 2. 关键词词频表

热词	词频	热词	词频
请	3698	年	2213
小区	3349	村	2048
咨询	3150	相关	1863
请求	2950	公积金	1659
建议	2865	补贴	1526
公司	2654	发放	1500
镇	2568	拖欠	1425
教师	2425	噪音	1395

分析表 2 中 Top20 的词频，可以发现社区留言中有关劳动和社会保障板块的词语出现的频率较高，群众所反映的问题一般比较贴近于生活，比如“公积金补贴”、“工人工资拖欠”、“施工噪音”等热点问题，政府在解决相关问题的同时，还应该出台相应的政策，扎扎实实解决好群众最关心的利益问题和最忧虑的实际问题。

图 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，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中的特征词并没有出现在图中.除此之外，劳动和社会保障板块中一些其他重要特征词还包括“福利”、“待遇”、“保护”等。

Figure 2. Word cloud
图 2. 词云图

向量空间模型(VSM, Vector Space Model)，也称为“单词袋”方法，是由著名研究人员 Salton [6]等发布的文本表示模型。通过将文档映射为特征向量，将对象以这种方式转换为空间向量的向量运算，通过其在空间上的相似性，我们可以用一种语言来表达文本之间的相似关系，这使得我们更容易处理文本。主要广泛应用在处理文本的数据过滤、信息源的检索、搜寻和信息索引以及分析文本的信息相关性质和排名等方面[7]。向量空间模型为用户提供了将文本中的冗余数据转换为变量的技术框架。但是，在实际的数据分析、开发和文本挖掘工作过程中，对于一个文档集来说，字数是数不清的，而停用同义词表只能用来消除一小部分文本噪声，而且特征向量的文本维数非常大，一方面，很可能会消耗大量的时间和信息空间；另一方面这些冗余的文本变量也极有可能会对数据分析工作产生不利的结果影响。因此，需要在文本数据开始挖掘和进行分析之前对文本中的特征向量进行降维处理。词向量转化结果如下图 3 表示：

```
评论的词向量为：arr= [[ 0.30694512  0.14347561  0.10496541 ...  0.9198899  -2.02784281
  -0.3638035 ]
 [-0.12889441  0.15199652  0.00678514 ...  0.58964335 -1.71449155
  -0.26129071]
 [ 1.13197924 -0.01065527  0.53438331 ...  0.80065054 -2.26772316
  -0.49008371]
 ...
 [-0.01416791  0.05692092 -0.08976909 ...  0.47874524 -0.84271745
  -0.21794731]
 [-0.32877644  0.03408712 -0.03939092 ...  0.49667121 -1.32903074
  -0.199083  ]
 [-0.56847605 -0.24854683 -0.00448852 ...  0.34112264 -0.92694785
  -0.43626926]]
```

Figure 3. Convert text in message to word vector
图 3. 留言中文本转化为词向量

(六) 特征降维与分类

本文主要采用基于统计的分类算法。分类算法的基本算法步骤是：第一步准确分析和判断训练文本集中的文本。再依据分类算法的结果，建立相关关系模型，相关关系指类别与具有训练文本的特征的相关关系；第二步对测试文本集进行分类，最大限度地利用该模型，不考虑文本数据的语言结构，而是使用特定的特征词来表示文本。

特征词的降维通常采用特征选择或特征提取的方法。特征选择是指根据一定条件从一组特征集中选择一些有助于文本分析的词，这些词构成一组特征子集。它通常是指根据一定的条件从特征集中选择对文本分析有重要贡献的一部分词来构造的特征子集。此过程仅删除相对不重要的要素，不会生成新的要素项。特征选择的规则主要涉及信息的获取等。

信息增益的定义是特征文档中存在或不存在特定信息特征词后的信息熵差，它表示当文档包含特定信息特征词时，特征文档中每个类别的平均累积信息增益。公式如下：

$$IG\left(w_i,C_j\right)=\log_2\frac{p\left(w_i,C_j\right)}{p\left(w_i\right)p\left(C_j\right)}p\left(w_i,C_j\right)+\log_2\frac{p\left(\overline{w}_i,C_j\right)}{p\left(\overline{w}_i\right)p\left(C_j\right)}p\left(\overline{w}_i,C_j\right) \tag{1}$$

$$IG\left(w_i\right)=\sum_j IG\left(w_i,C_j\right) \tag{2}$$

信息增益越大，与分类相关的特征词的影响和作用越明显。因此，降低了利用信息增益进行降维的效果。

K 近邻法(KNN, K-nearest neighbor) [8] [9]由 cover 与 hart 研究人员共同发现的一种在国内已经应用广泛分类类型算法，k 近邻标签分类法主要是一种有监督的一种学习分类算法，其分类规则主要就是数据本身，并且不必再要求产生其他的分类数据来对其进行分类阐述。因此采用 k 邻近法建立一个群众留言一级标签分类算法模型，其思想简单直观。

在训练样本集中，可能有许多靠近测试样本的观察点。预测输出变量的依据应该是离测试样本最近的观测点的数量，即最近邻的数量，这是 k-最近邻法的关键。

3.2. 指标体系构建

指标体系的构成是基于不同需求和目标差异的原则。指标体系分为三个层次。指标选择层次通常有三个重要原则：客观性原则、系统性原则和敏感性原则。依据问题热度这一特殊的研究对象，在设计和构建问题热度评价指标体系的时，在上述三个原则的基础之上，还增加了两个新的原则：趋势性原则和导向型原则。

Table 3. Weibo popularity evaluation index system and the connotation of each index
表 3. 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

基于新闻学三要素及其新闻传播理论，我们将选取内容特征热影响和受众反映特征热影响两个要素，从问题传播内容和受众反映两个层面对问题热进行定量评价。

1) 内容特征热度影响力。根据流行的三大原因理论，我们可以知道留言信息内容本身的内容特点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。

2) 观众特征热影响。受众特征是指受众在收到信息后的反应和态度。指标体系如表 3 所示。

3.3. 问题热度因子分析

运用因子分析法对本研究构建的指标体系进行分析。最后，根据模型的计算结果计算热点问题的层次排序。

结果表明，使用 SPSS16.0 是必要的。它对采集的数据变量进行因子分析可行性检验，样本通过 KMO (Kaiser-Meyer-Olkin)检验统计量检验的结果为 0.876，可用于因子分析；Bartlett’s 球状检验的近似卡方值为 65.84，并且得出 P 值为 0.000，小于 0.01，表明统计数据适合进行因子分析。根据因子分析中主成分特征值提取的原理，得到两个公因子 F_1 和 F_2 (表 4)。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 0.68007 和 0.17693，所计算出来的结果与总体信息的 0.857 一样，所以可以用它来对问题热度进行评价。

Table 4. Main factor eigenvalue, cumulative variance contribution rate
表 4. 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

为了便于对各主要因子的研究，计算得到旋转因子载荷矩阵。见表 5。

Table 5. Factor loading matrix
表 5. 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，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，可以理解成为受众特征。

因为有旋转系数荷载矩阵中各指标和公因子的相关系数值可能希望列出 F_1 和 F_2 的计算公式，分别见公式(3)和公式(4)。

$$F_1 = 0.313X_1 + 0.298X_2 + 0.615X_3 \tag{3}$$

$$F_2 = 0.954X_4 + 0.977X_5 \tag{4}$$

接着表 5 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(5)。

$$F = 0.7934F_1 + 0.2065F_2 \tag{5}$$

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 Get Matching Degree (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。评分标准如表 6~10：

Table 6. Word count fulfillment score
表 6. 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

Table 7. Timeliness score
表 7. 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

Table 8. Content similarity score
表 8. 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

Table 9. Likes and dislikes score
表 9. 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 10：

Table 10. Hot question list
表 10. 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 110 kv 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

通过热力综合评分计算公式，对论证中选取的问题进行组织和排序，检验指标体系的合理性和可行性。

在现有的相关文献的基础之上，本文以新闻传播学和大众化三要素理论为基础，建立了信息传播大众化的评价体系。评价体系包括两个维度、两个要素和五个指标。这两个维度分别指接收器和通信内容。此外，通过因子分析和实例进行验证，得出指标体系是具有一定的合理性和有效性。最后，基于这个结果对不同类型问题的热源组成进行了比较分析。总结了以下结论：

1) 目前，国内外学者对微博热点和微博热点话题都给予了关注，并取得了一定的研究成果。然而，目前还没有关于问题热的研究成果，也没有关于热点问题的全面界定的文献。在评价指标的构建研究中，仅采用转发率、评论率等简单的指标，不能充分反映问题热的影响因素或形成机理。因此，有必要根据相关理论建立问题热综合评价指标体系。

2) 根据问题热度排名可以看出，评价指标体系的估算结果与问题热排序结果一致，表明所构建的评价指标体系更为合理。估计结果表明，通信内容对热度的影响最大，即在该模型中，相似度对热度的影响最大，而微博热度对接收者的影响较小。

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。

3.4. 答复质量评价体系

基于顾客感知服务质量的定义，让服务接受者、服务提供者和第三方对服务进行打分评价，并从三维评估的概念对其进行评估。其中，群众评价的主体是群众，评价对象是反应绩效；相关政府部门将成为响应提供者的评估主体，评估对象为响应过程；第三方评估的主体是第三方组织。评估对象是回复的服务能力。评估对象证明并认可回复提供者提供的回复的服务质量。三方评价完成后，制定相应的评价指标，通过对评价指标进行加权，得出评价结果。答复质量评价的整体框架如图 4 所示。

PZB 服务质量模型是由 A. Parasuraman、Valarie A. Zeithamal 和 Leonard L. Berry 等人提出，并且以此为理论依据。结合我国国情的特点与现代社会环境特点，构建了一个新的答复质量评价指标体系，新的答复质量评价指标体系是以群众为评价主体、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。具体指标如表 11 所示。

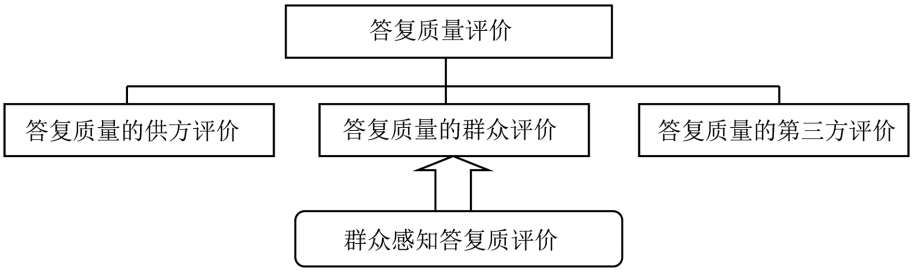

Figure 4. Overall framework for response quality assessment
图 4. 答复质量评价的整体框架

Table 11. Response quality evaluation index system
表 11. 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

指标体系中有一级指标、二级指标、三级指标。因此，在制定指标过程中，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。

指标中涉及的评估量表为 Liken7 级量表[10]，该量表有 21 个三级指标。人们会根据自己接受服务后的实际感受对服务质量进行评分。分值代表服务质量满意度，分值越高代表越满意，分值越低代表越不满意。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。

4. 结论

本文为了提高政府相关部门处理留言的有效性，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。利用互联网公开留言记录，借助 Excel、Rstudio 等软件，运用 jieba 中文分词工具对样本进行预处理，使用 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。将 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。

基金项目

2020 年福建省闽江学院校长基金(编号：103952020082)。

参考文献

[1] 张腾, 张建光, 尚进. 基于 DPSIR 模型的智慧政务信息生态评价研究[J]. 中国科技论坛, 2017(2): 186-192.

[2] Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1988) SERVQUAL: A Multiple-Item Scale for Measuring Consumer perceptions of Service Quality. *Journal of Retailing*, **64**, 12-40.

[3] 翟元甫. 基于 TOE 框架的政务服务智慧能力影响因素研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2020.

[4] 韦人予. 中文分词技术研究[J]. 信息与电脑(理论版), 2020, 32(10): 26-29.

[5] 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(3):72-80.

[6] Salton, G. (1988) Automatic Text Processing. Addison-Wesley Publishing Company, Reading.

[7] 谭静. 基于向量空间模型的文本相似度算法研究[D]: [硕士学位论文]. 成都: 西南石油大学, 2015.

[8] Soucy, P. and Mineau, G.W. (2001) A Simple KNN Algorithm for Text Categorization. *Proceedings of the* 2001 *IEEE International Conference on Data Mining*, San Jose, November 2001, 647-648. https://doi.org/10.1109/ICDM.2001.989592

[9] Chen, Y.Q., Nixon, M.S. and Damper, R.I. (1995) Implementing the k-Nearest Neighbour Rule via a Neural Network. *Proceedings of ICNN*'95—*International Conference on Neural Networks*, Perth, WA, 27 November-1 December 1995, 136-140. https://doi.org/10.1109/ICNN.1995.488081

[10] 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.