“智慧政务”中的文本挖掘应用

摘要

如今，在信息技术快速发展的背景下，将政府部门的运作实现信息化、数据化，可以更快捷地符合为广大民众服务。为高效处理和答复涉及社情民意的留言信息，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。那么，在处理问政平台留言的时候，首先需要将留言进行一级标签分类，所以分类模型是否准确备受关注。另外，群众在某一段时间内集中通过网络问政平台反映的问题即为热点问题，需要受到高度重视，对此，定义合理的热度评价指标挖掘热点问题，有利于高效管理从而造福百姓。最后，政府根据群众留言给出相关、完整并具有解释性的答复，是群众最为期待的结果。

对于问题 1，本文首先选择数据原本给出的一级标签数据，在随机划分测试集与训练集后，选取线性支持向量机的方法对分类模型进行评估，得到不同类别的 F1-score 分数。其中，教育文体、劳动和社会保障两类的 F1-score 分数最高，达到 0.87，商贸旅游的 F1-score 分数最低，为 0.77。

对于问题 2，本文采用层次分析法，建立一套多层次评价指标体系，最终量化分析出热度排名前五的热点问题的热度指数分别是：74.7、69.7、68.3、62.5、59.5；对应留言编号明细为：（1）191951、202575、216316、243551、263672、266931；（2）208636；（3）223297；（4）205960、209742、233743；（5）254865、262052、268250、272089。

对于问题 3，本文对数据进行分析后给出划分，计算每条留言详情和留言答复的相关性、完整性与可解释性，最后通过三维度评价法对这三个指标建立评价方案：确定某条留言数据的三个指标对应在三维坐标系的位置，若该点到原点的相对距离越远，则针对该问题给出的答复意见越好，即答复意见的相关性、完整性、可解释性越好。反之，越差。

关键词：线性支持向量机、层次分析法、三维度评价法、文本挖掘、Python

Abstract

In the context of the rapid development of information technology, if the operation of the government departments to achieve information-based, data-based management, it can serve the public more quickly. In order to efficiently process and reply to messages concerning social conditions and public opinions, the establishment of a "smart government" system based on natural language processing technology has become a new trend in the innovation and development of social governance, which has greatly promoted the government's management level and efficiency. When dealing with the message on the political platform, it is necessary to classify the message on the first level label, so the accuracy of the classification model has attracted much attention. In addition, the issues that the public have concentrated on through the network political platform within a certain period of time are hot issues and need to be paid great attention to. In this regard, a reasonable heat evaluation index to explore hot issues is conducive to efficient management, so as to benefit the people. Finally, the government can give relevant, complete and explanatory answers based on the comments of the public. This is also the result that the public are most looking forward to.

For Question 1, the first-level label data of the original data is preferred in this paper. After randomly dividing test set and training set, linear support vector machine is selected to evaluate the classification model to obtain F1-score of different categories.

For question 2, this paper uses AHP to establish a multi-level evaluation index system, and finally quantitatively analyzes the top five hot issues. The heat index is: 74.7, 69.7, 68.3, 62.5, 59.5; the corresponding message The numbering details are: (1) 191951, 202575, 216316, 243551, 263672, 266931; (2) 208636; (3) 223297; (4) 205960, 209742, 233743; (5) 254865, 262052, 268250, 272089.

For question 3, this paper classifies the data after analysis, and calculates the details of each message and the relevance, completeness and interpretability of the message reply. Finally, an evaluation scheme is established for the three indicators through a three-dimensional evaluation method: firstly, the position of the three indicators of a message data corresponding to the three-dimensional coordinate system is determined; if the relative distance from the point to the origin is further, the better the reply opinions given to the question, i.e. the better the relevance, completeness and interpretability of the reply opinions. On the contrary, the worse.

Key words: linear support vector machine, analytic hierarchy process, three dimensional evaluation method, text mining, Python

目录

1、挖掘目的

如今，通过网络问政平台对群众的留言进行处理，可以提升政府的管理水平和施政效率。然而，在处理问政平台的留言的时候，首先，需要按照一定的划分体系对留言进行分类，以便后续分配部门处理问题。因此，划分体系的一级标签的分类模型的建立则显得尤为重要，而模型是否准确也备受关注。不仅如此，当群众在某一段时间内集中通过网络问政平台反映的问题即为热点问题，需要受到高度重视，所以，定义合理的热度评价指标挖掘热点问题，有利于高效管理从而造福百姓。另外，政府根据群众留言给出相关、完整并具有解释性的答复，是群众最为期待的。

问题 1 属于模型的归类与评价问题。针对问题 1，选择数据原本给出的一级标签数据，在随机划分测试集与训练集后，选取线性支持向量机的机器学习模型对分类模型进行评估。实施计划如下：（1）对数据进行分析与预处理；（2）对预处理以后的留言主题的分词转化为 TF-IDF 向量；（3）进行卡方检验得到每一个标签的相关词语（4）最后选择线性支持向量机模型，利用混淆矩阵，预测标签和实际标签之间的差异，同时计算 F1-Score 分数，对分类模型进行评估。

问题 2 属于模型的分类与分析问题。针对问题 2，采用层次分析法，建立一套多层次评价指标体系，最终定量化分析出热点问题。实施计划如下：（1）对所有数据的留言问题和留言详情分别提取关键词；（2）根据关键词将所有数据分类；（3）对分类后的数据分别寻找相似问题，并将相似问题归类，求在关键词类下某个问题的点赞数和反对数的总数。（4）利用多层次模型，构建热度指标评价模型。（5）评估热度指标评价模型。

问题 3 属于建立评价方案问题。针对问题 3，本文将对数据进行分析后给出划分，计算每条留言详情和留言答复的相关性、完整性与可解释性，最后通过三维度评价法对这三个指标建立评价方案。实施计划如下：（1）对所有的数据进行相关性、完整性、可解释性分析；（2）建立三维评价体系，即相关性、完整性与可解释性分别代表三维直角坐标系的 X、Y、Z 轴；（3）计算每条留言的留言详情和留言答复的相关性、完整性与可解释性，将数值代入坐标系中，将坐标到原点的距离作为最终评价标准。

2、问题 1 的分析与求解

2.1 数据准备

2.1.1 数据分析

通过简单的数据处理得到数据中的一级标签包含 7 个类别，分别为城乡建设、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，每一类别数据数量分布情况如图 1 和表 1 所示。

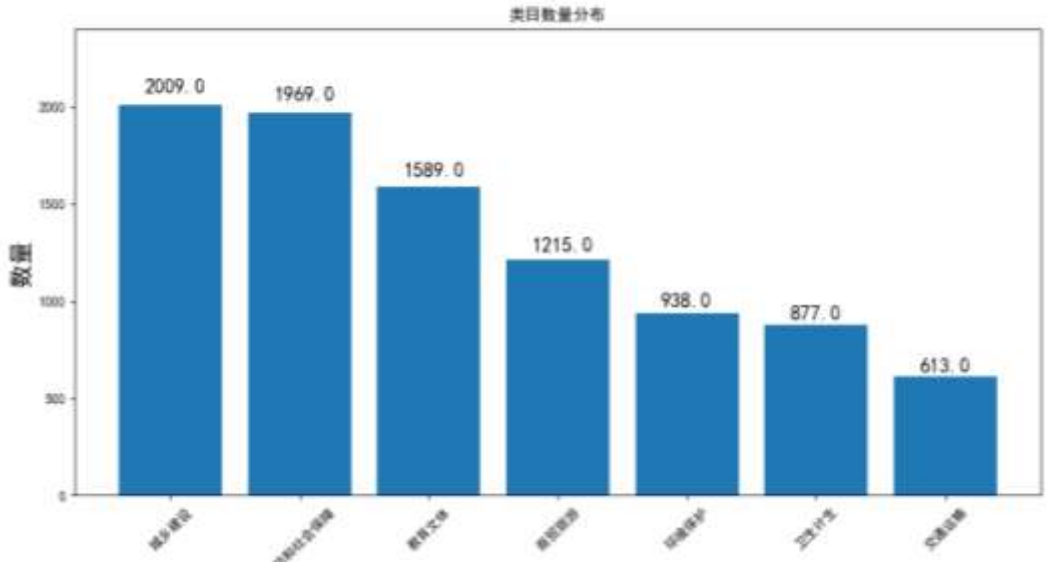


图 1 类别数量分布图

表 1 类别数量分布表

	一级标签	count
1	城乡建设	2009
2	劳动和社会保障	1969
3	教育文体	1589
4	商贸旅游	1215
5	环境保护	938
6	卫生计生	877
7	交通运输	613

2.1.2 数据预处理

（1）对数据进行清洗，发现在 9210 条数据中未出现缺失值。

（2）将一级标签转换成相应的 label_id 编码，例如“城乡建设=0”等。方便之后分类模型的训练，如表 2 所示。

表 2 类别 id 转化表

	一级标签	label_id
1	城乡建设	0
2	环境保护	1
3	交通运输	2
4	教育文体	3
5	劳动和社会保障	4
6	商贸旅游	5
7	卫生计生	6

（3）删除文本中的标点符号、中文停用词、以及一些无意义的常用词(stopword)。因为这些词和符号无法反应文本的主要内容，并且会增加计算的复杂度和增加系统开销，所以将其从文本中清除。

（4）对清洗后的数据进行分词，罗列出每个分类中前 100 个高频词，再生成这些高频词的词云图，如图 2 所示。

图 2 每类高频词 Top100 排行榜图

2.2 模型的建立

2.2.1 生成 TF-IDF 向量

计算 TF-IDF 的特征值。TF 代表词频，IDF 代表逆文本频率指数，TF-IDF 是在单词计数的基础上，降低了常用高频词的权重,增加罕见词的权重。一般生成 TF-IDF 向量是将“留言主题”转换成词频向量,然后将词频向量再转换成 TF-IDF 向量，但是本文运用了一种简化的方式是直接使用 TfidfVectorizer 来生成 TF-IDF 向量省去了转换成词频的过程。

对于某一特定文件里的词语来说，TF(词频)的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

$n_{i,j}$ 为该词在文件中出现的个数，$\sum_k n_{k,j}$ 是在文件中所有字词的出现次数之和。

IDF(逆向文件频率) 是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以 10 为底的对数得到：

$$idf_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (2)$$

然后再计算 TF 与 IDF 的乘积：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

采用 TfidfVectorizer 得到评价数据 9210 条后，全部评论中的所有词语数和词语对(相邻两个单词的组合)的总数是 56783。部分 TF-IDF 的特征值结果如图 3。

```
(9210, 56783)
--------------------------------
  (0, 34783)    0.38992145246512155
  (0, 55129)    0.38992145246512155
  (0, 29048)    0.38992145246512155
  (0, 49247)    0.38992145246512155
  (0, 23690)    0.2568946512414059
  (0, 34773)    0.2682363300472749
  (0, 55113)    0.26611652697072047
  (0, 29024)    0.26217763369454206
  (0, 49246)    0.3381483927207964
  (1, 23692)    0.3107737611568129
  (1, 22011)    0.3107737611568129
  (1, 40793)    0.3107737611568129
  (1, 9310)     0.3107737611568129
  (1, 22149)    0.3107737611568129
  (1, 20730)    0.3107737611568129
  (1, 7046)     0.14093766170266666
  (1, 22000)    0.21649752181061938
  (1, 40791)    0.26950978760106314
  (1, 9307)     0.2879426209905127
  (1, 22144)    0.24440613452037493
  (1, 20729)    0.3107737611568129
  (1, 23690)    0.20474922957592534
  (2, 52506)    0.34968334638276066
  (2, 41194)    0.3346558638327778
  (2, 15162)    0.34968334638276066
  :     :
(9207, 13486)   0.29529002879502375
(9207, 33526)   0.2245803607859515
(9207, 48905)   0.18885896370622285
(9208, 22832)   0.415022890857895
```

图 3 部分 TF-IDF 的特征值

2.2.2 卡方检验

利用卡方检验找出每个分类中关联度最大的两个词语和两个词语对，如下表 3。

表 3 每个分类中关联度最大的两个词语和两个词语对明细表

	Most correlated unigrams	Most correlated bigrams
交通运输	1、的士 2、出租车	1、出租车 管理 2、滴滴 出行
劳动和社会保	1、职工	1、社保 问题

障	2、社保	2、退休 人员
卫生计生	1、独生子女 2、医院	1、再婚 家庭 2、人民 医院
商贸旅游	1、电梯 2、传销	1、传销 组织 2、小区 电梯
城乡建设	1、公积金 2、房产证	1、拖欠 工程款 2、住房 公积金
教育文体	1、补课 2、教师	1、教师 招聘 2、培训 机构
环境保护	1、排放 2、污染	1、严重 污染 2、污染 严重

2.3 模型选择与评估

2.3.1 模型选择

本文尝试使用四种不同的机器学习模型：

随机森林(Random Forest)

线性支持向量机(Linear Support Vector Machine)

多项式朴素贝叶斯(Multinomial Naive Bayes)

逻辑回归(Logistic Regression)

对这四种方法的准确性进行了精确性评估，结果如下图 4。从图 4 的箱线图可知，线性支持向量机(Linear Support Vector Machine)模型的准确率最高。因此我们最终选取线性支持向量机模型。

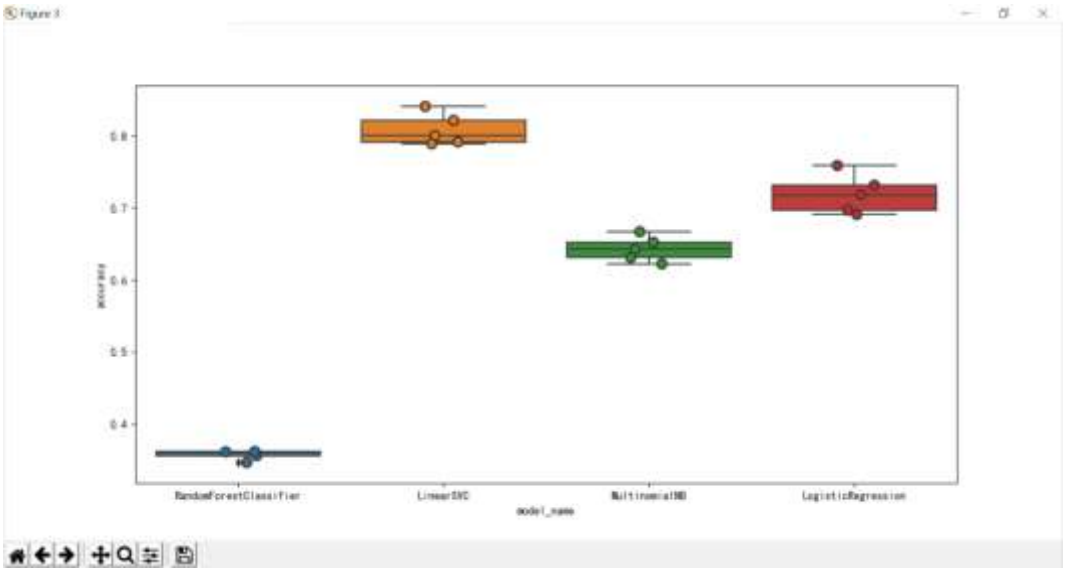


图 4 四种模型准确度对比图

2.3.2 模型评估

选取线性支持向量机(Linear SVM)的机器学习模型，支持向量机模型是根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中，以期获的最好的推广能力。

假设给定一组训练样本为

$$\left\{\left(x_i, x_j\right)\right\}_{i=1}^{l} \qquad x_i \in R^n \qquad y_i \in \{1,-1\}$$

Linear SVM 可表示为下列式子的无约束优化问题：

$$\min_{w,b,\zeta_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\zeta(w, x_i, y_i) \tag{4}$$

其中 $\zeta(w, x_i, y_i)$ 是损失函数，$C>0$ 是惩罚因子，使用的损失函数有：

$$\xi(w, x_i, y_i) = \begin{cases} \max(0, 1-y_i w x_i) & L1-SVM \\ \max(0, 1-y_i w x_i)^2 & L2-SVM \\ \log(1)\exp(-y_i w x_i)) & LR \end{cases} \tag{5}$$

通过混淆矩阵，生成显示预测标签和实际标签之间的差异的图 5，由图可见，交通运输类预测相对较为准确，城乡建设类预测的错误数量较多。

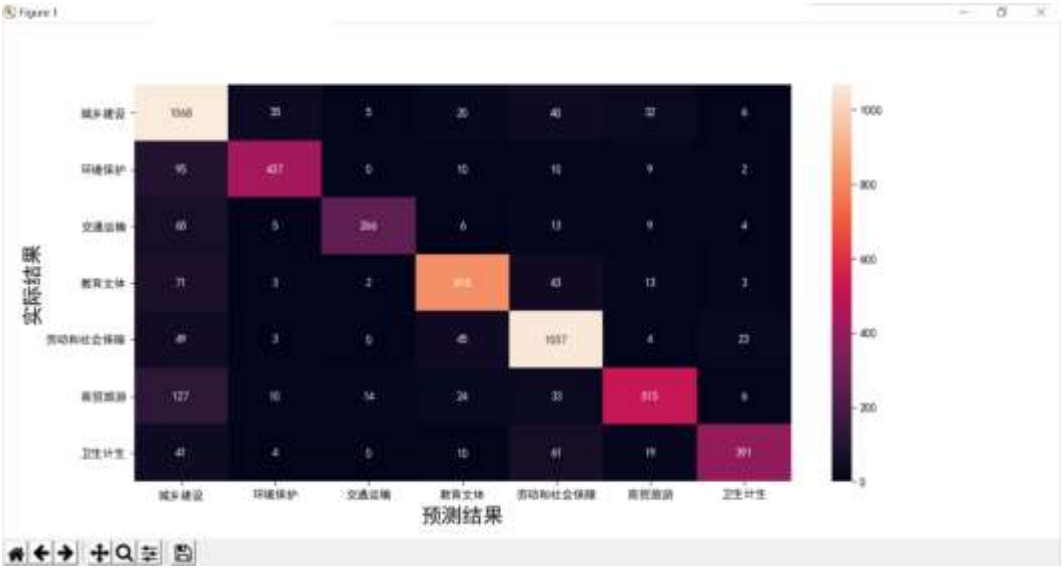


图 5 混淆矩阵图

由于多分类模型一般不使用准确率来评估模型的质量,因为准确率不能反应出每一个分类的准确性,当训练数据不平衡时，准确率不能反应出模型的实际预测精度。因此利用 F1 分数，即公式（1）评估模型，其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率，得到评估结果如表 4，根据 F1 指标，得：教育文体、劳动和社会保障的 F1 分数最大，为 0.87，即教育文体、劳动和社会保障两类预测最准确，商贸旅游 F1 分数最小，只有 0.77，即商贸旅游预测失误较多。因为这七类的数据数量分布不均衡，交通运输类数据个数只有 368，所以混淆矩阵并不准确的原因较为合理。

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{6}$$

表 4 评估结果表

	精确率 precision	召回率 recall	F_1指标 f1-score	数据个数 support
城乡建设	0.70	0.89	0.78	1206
环境保护	0.88	0.78	0.82	563
交通运输	0.93	0.72	0.81	368
教育文体	0.88	0.86	0.87	953
劳动和社会保障	0.84	0.90	0.87	1181
商贸旅游	0.86	0.71	0.77	729
卫生计生	0.90	0.74	0.81	526

3、问题 2 的分析与求解

3.1 数据准备

3.1.1 数据分析

因为留言主题是对留言内容的大致概括，所以本文利用 jieba 库，对所有数据的留言主题进行分词，提取关键词，留言主题的关键词按频次排列如表 5（表 5 只显示排名靠前的 20 个关键词），通过比较关键词出现的次数，选取了可以代表地点并且出现频次较高的前八个关键词，次数由多到少分别为：A7、A3、A2、A4、A1、A5、西地省、A6。又因为留言详情中反映了该留言人群的目的与诉求，所以本文根据留言详情再次提取关键词，留言详情的关键词按频次排列如表 6（表 6 只显示排名靠前的 20 个关键词），筛选出与人群有关且出现次数较多的关键词，次数由多到少分别为：业主、领导、开发商、政府、居民、公司、物业、学校。综上所述，最终得到 16 个关键词：A7、A3、A2、A4、A1、A5、西地省、A6、业主、领导、开发商、政府、居民、公司、物业、学校。

表 5 留言主题关键词按频次排列表（部分）

	关键词	次数		关键词	次数
1	A7	682	11	街道	199
2	小区	549	12	A5	177
3	问题	497	13	反映	176
4	A3	439	14	咨询	171
5	扰民	278	15	建议	167
6	A2	264	16	西地省	160
7	A4	236	17	噪音	147
8	严重	218	18	A6	144
9	A1	209	19	施工	140
10	投诉	202	20	社区	132

表 6 留言详情关键词按频次排列表（部分）

	关键词	次数		关键词	次数

1	业主	4096	11	情况	1279
2	小区	3871	12	解决	1229
3	领导	2126	13	影响	1222
4	部门	1812	14	已经	1201
5	开发商	1780	15	社区	1092
6	政府	1700	16	学校	1060
7	居民	1588	17	2019	1060
8	公司	1572	18	他们	1036
9	物业	1383	19	作为	1020
10	要求	1283	20	建设	1018

3.1.2 数据预处理

（1）本文利用代表地点的关键词对所有数据的留言主题进行归类，并将分类结果导成 8 个 excel 表格。通过与人群相关的关键词对所有数据的留言详情进行归类，同样生成 8 个 excel 表格。

（2）针对归类后的 excel 表格，分别对每一类的留言详情中的数据，利用 difflib 库，进行相似度比较。

（3）建立相似度索引矩阵。

若两条留言的留言详情比对其相似度后，得到相似度数值大于某一阈值，则本文认为这两条留言相似，属于一类问题，反之，不属于一类问题。其次，从每个地点或人群的 excel 表中，将代表一个问题的留言的索引存放于子矩阵中，因为一个地点或人群存在多个相似问题，所以将这些代表每个问题的子矩阵纵向合并，得到每类地点或人群中的多个问题矩阵。最终，得到 16 个相似度索引矩阵。

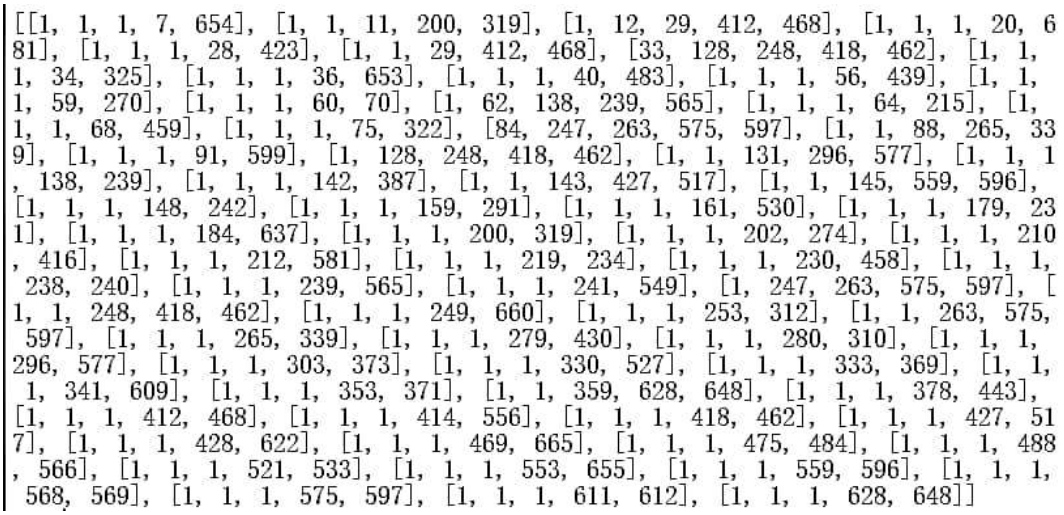
相似度索引矩阵形成实例如下：在 A2 地区 excel 表中，第 8 条留言和第 125 条留言相似，第 36 条留言和第 158 条留言相似，第 59、103 和 151 条留言相似，第 123 和 144 条留言相似，则可以得到矩阵（7）。注：数字代表对应留言的索引，第一条留言的索引为 0，所以索引=留言的序数-1；x 无实际意义。

$$\begin{bmatrix} x & 7 & 124 \\ x & 35 & 157 \\ 58 & 102 & 150 \\ x & 122 & 143 \end{bmatrix} \quad (7)$$

（4）确定相似度阈值。

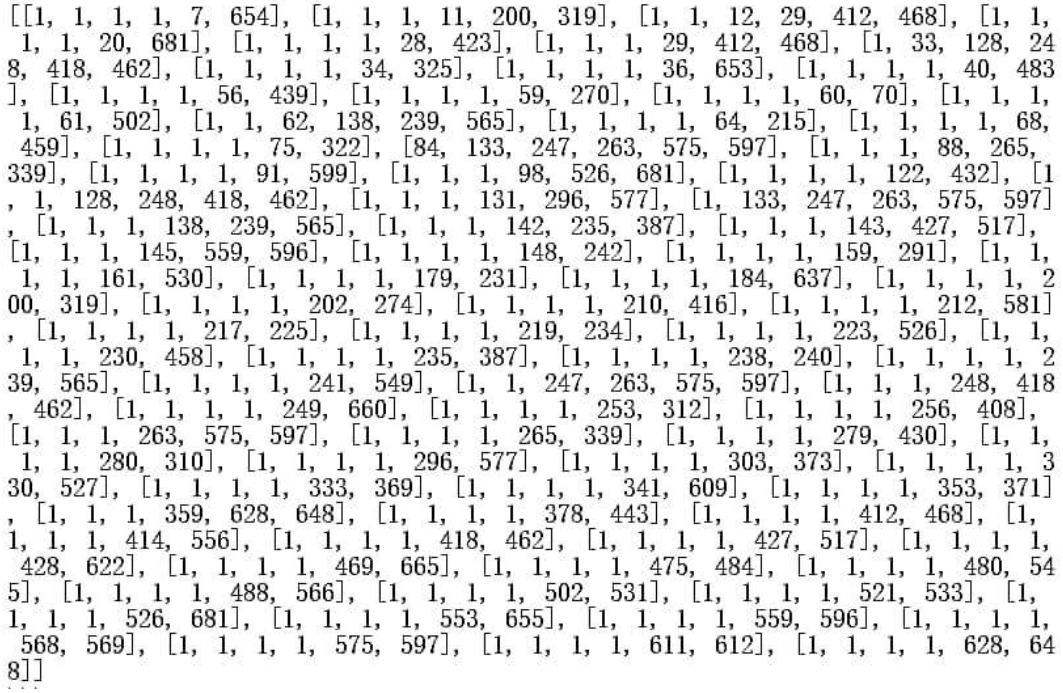
针对 A7 地区的留言，本文比较了 0.7、0.65、0.68 三个数作为相似度阈值时的情况，发现：当阈值是 0.7 时，提取出相似度索引矩阵图 6，经过验证，图 6 中的相似度索引子矩阵对应的留言都相似；当阈值是 0.65 时提取出相似度索引矩阵图 7。比较图 6 和图 7，阈值是 0.65 时相较阈值是 0.7 时多了一些相似度索引子矩阵，如：[1, 1, 1, 98, 526, 681]，结合 A7 地区留言，该子矩阵中索引 98 和 526 的留言内容不同，因此阈值为 0.65 不准确。再次尝试阈值为 0.68 时的情况，得到图 8。比较图 6 和图 8，阈值是 0.68 相较阈值是 0.7 时多了一些相似度索引子矩阵，有：[1, 1, 1, 61, 502], [1, 1,

143, 427, 517], [1, 1, 1, 217, 225], 这些子矩阵中索引对应的留言内容相同，因此阈值为 0.68 更准确。因此，确定 A7 地区留言的相似度阈值为 0.68。



```
[[1, 1, 1, 7, 654], [1, 1, 11, 200, 319], [1, 12, 29, 412, 468], [1, 1, 1, 20, 6
81], [1, 1, 1, 28, 423], [1, 1, 29, 412, 468], [33, 128, 248, 418, 462], [1, 1,
1, 34, 325], [1, 1, 1, 36, 653], [1, 1, 1, 40, 483], [1, 1, 1, 56, 439], [1, 1,
1, 59, 270], [1, 1, 1, 60, 70], [1, 62, 138, 239, 565], [1, 1, 1, 64, 215], [1,
1, 1, 68, 459], [1, 1, 1, 75, 322], [84, 247, 263, 575, 597], [1, 1, 88, 265, 33
9], [1, 1, 1, 91, 599], [1, 128, 248, 418, 462], [1, 1, 131, 296, 577], [1, 1, 1
, 138, 239], [1, 1, 1, 142, 387], [1, 1, 143, 427, 517], [1, 1, 145, 559, 596],
[1, 1, 1, 148, 242], [1, 1, 1, 159, 291], [1, 1, 1, 161, 530], [1, 1, 1, 179, 23
1], [1, 1, 1, 184, 637], [1, 1, 1, 200, 319], [1, 1, 1, 202, 274], [1, 1, 1, 210
, 416], [1, 1, 1, 212, 581], [1, 1, 1, 219, 234], [1, 1, 1, 230, 458], [1, 1, 1,
 238, 240], [1, 1, 1, 239, 565], [1, 1, 1, 241, 549], [1, 247, 263, 575, 597], [
1, 1, 248, 418, 462], [1, 1, 1, 249, 660], [1, 1, 1, 253, 312], [1, 1, 263, 575,
 597], [1, 1, 1, 265, 339], [1, 1, 1, 279, 430], [1, 1, 1, 280, 310], [1, 1, 1,
296, 577], [1, 1, 1, 303, 373], [1, 1, 1, 330, 527], [1, 1, 1, 333, 369], [1, 1,
 1, 341, 609], [1, 1, 1, 353, 371], [1, 1, 359, 628, 648], [1, 1, 1, 378, 443],
[1, 1, 1, 412, 468], [1, 1, 1, 414, 556], [1, 1, 1, 418, 462], [1, 1, 1, 427, 51
7], [1, 1, 1, 428, 622], [1, 1, 1, 469, 665], [1, 1, 1, 475, 484], [1, 1, 1, 488
, 566], [1, 1, 1, 521, 533], [1, 1, 1, 553, 655], [1, 1, 1, 559, 596], [1, 1, 1,
 568, 569], [1, 1, 1, 575, 597], [1, 1, 1, 611, 612], [1, 1, 1, 628, 648]]
```

图 6 阈值为 0.7 时 A7 地区留言相似度问题索引矩阵



```
[[1, 1, 1, 1, 7, 654], [1, 1, 1, 11, 200, 319], [1, 1, 12, 29, 412, 468], [1, 1,
 1, 1, 20, 681], [1, 1, 1, 1, 28, 423], [1, 1, 1, 29, 412, 468], [1, 33, 128, 24
8, 418, 462], [1, 1, 1, 1, 34, 325], [1, 1, 1, 1, 36, 653], [1, 1, 1, 1, 40, 483
], [1, 1, 1, 1, 56, 439], [1, 1, 1, 1, 59, 270], [1, 1, 1, 1, 60, 70], [1, 1, 1,
 1, 61, 502], [1, 1, 62, 138, 239, 565], [1, 1, 1, 1, 64, 215], [1, 1, 1, 1, 68,
 459], [1, 1, 1, 1, 75, 322], [84, 133, 247, 263, 575, 597], [1, 1, 1, 88, 265,
339], [1, 1, 1, 1, 91, 599], [1, 1, 1, 98, 526, 681], [1, 1, 1, 1, 122, 432], [1
, 1, 128, 248, 418, 462], [1, 1, 1, 131, 296, 577], [1, 133, 247, 263, 575, 597]
, [1, 1, 1, 138, 239, 565], [1, 1, 1, 142, 235, 387], [1, 1, 1, 143, 427, 517],
[1, 1, 1, 145, 559, 596], [1, 1, 1, 1, 148, 242], [1, 1, 1, 1, 159, 291], [1, 1,
 1, 1, 161, 530], [1, 1, 1, 1, 179, 231], [1, 1, 1, 1, 184, 637], [1, 1, 1, 1, 2
00, 319], [1, 1, 1, 1, 202, 274], [1, 1, 1, 1, 210, 416], [1, 1, 1, 1, 212, 581]
, [1, 1, 1, 1, 217, 225], [1, 1, 1, 1, 219, 234], [1, 1, 1, 1, 223, 526], [1, 1,
 1, 1, 230, 458], [1, 1, 1, 1, 235, 387], [1, 1, 1, 1, 238, 240], [1, 1, 1, 1, 2
39, 565], [1, 1, 1, 1, 241, 549], [1, 1, 247, 263, 575, 597], [1, 1, 1, 248, 418
, 462], [1, 1, 1, 1, 249, 660], [1, 1, 1, 1, 253, 312], [1, 1, 1, 1, 256, 408],
[1, 1, 1, 263, 575, 597], [1, 1, 1, 1, 265, 339], [1, 1, 1, 1, 279, 430], [1, 1,
 1, 1, 280, 310], [1, 1, 1, 1, 296, 577], [1, 1, 1, 1, 303, 373], [1, 1, 1, 1, 3
30, 527], [1, 1, 1, 1, 333, 369], [1, 1, 1, 1, 341, 609], [1, 1, 1, 1, 353, 371]
, [1, 1, 1, 359, 628, 648], [1, 1, 1, 1, 378, 443], [1, 1, 1, 1, 412, 468], [1,
1, 1, 1, 414, 556], [1, 1, 1, 1, 418, 462], [1, 1, 1, 1, 427, 517], [1, 1, 1, 1,
 428, 622], [1, 1, 1, 1, 469, 665], [1, 1, 1, 1, 475, 484], [1, 1, 1, 1, 480, 54
5], [1, 1, 1, 1, 488, 566], [1, 1, 1, 1, 502, 531], [1, 1, 1, 1, 521, 533], [1,
1, 1, 1, 526, 681], [1, 1, 1, 1, 553, 655], [1, 1, 1, 1, 559, 596], [1, 1, 1, 1,
 568, 569], [1, 1, 1, 1, 575, 597], [1, 1, 1, 1, 611, 612], [1, 1, 1, 1, 628, 64
8]]
```

图 7 阈值为 0.65 时 A7 地区留言相似度问题索引矩阵

[[1, 1, 1, 7, 654], [1, 1, 11, 200, 319], [1, 12, 29, 412, 468], [1, 1, 1, 20, 6
81], [1, 1, 1, 28, 423], [1, 1, 29, 412, 468], [33, 128, 248, 418, 462], [1, 1,
1, 34, 325], [1, 1, 1, 36, 653], [1, 1, 1, 40, 483], [1, 1, 1, 56, 439], [1, 1,
1, 59, 270], [1, 1, 1, 60, 70], [1, 1, 1, 61, 502], [1, 62, 138, 239, 565], [1,
1, 1, 64, 215], [1, 1, 1, 68, 459], [1, 1, 1, 75, 322], [84, 247, 263, 575, 597]
, [1, 1, 88, 265, 339], [1, 1, 1, 91, 599], [1, 128, 248, 418, 462], [1, 1, 131,
 296, 577], [1, 1, 1, 138, 239], [1, 1, 1, 142, 387], [1, 1, 143, 427, 517], [1,
 1, 145, 559, 596], [1, 1, 1, 148, 242], [1, 1, 1, 159, 291], [1, 1, 1, 161, 530
], [1, 1, 1, 179, 231], [1, 1, 1, 184, 637], [1, 1, 1, 200, 319], [1, 1, 1, 202,
 274], [1, 1, 1, 210, 416], [1, 1, 1, 212, 581], [1, 1, 1, 217, 225], [1, 1, 1,
219, 234], [1, 1, 1, 230, 458], [1, 1, 1, 238, 240], [1, 1, 1, 239, 565], [1, 1,
 1, 241, 549], [1, 247, 263, 575, 597], [1, 1, 248, 418, 462], [1, 1, 1, 249, 66
0], [1, 1, 1, 253, 312], [1, 1, 263, 575, 597], [1, 1, 1, 265, 339], [1, 1, 1, 2
79, 430], [1, 1, 1, 280, 310], [1, 1, 1, 296, 577], [1, 1, 1, 303, 373], [1, 1,
1, 330, 527], [1, 1, 1, 333, 369], [1, 1, 1, 341, 609], [1, 1, 1, 353, 371], [1,
 1, 359, 628, 648], [1, 1, 1, 378, 443], [1, 1, 1, 412, 468], [1, 1, 1, 414, 556
], [1, 1, 1, 418, 462], [1, 1, 1, 427, 517], [1, 1, 1, 428, 622], [1, 1, 1, 469,
 665], [1, 1, 1, 475, 484], [1, 1, 1, 488, 566], [1, 1, 1, 521, 533], [1, 1, 1,
553, 655], [1, 1, 1, 559, 596], [1, 1, 1, 568, 569], [1, 1, 1, 575, 597], [1, 1,
 1, 611, 612], [1, 1, 1, 628, 648]]

图 8 阈值为 0.68 时 A7 地区留言相似度问题索引矩阵

同理 A7 地区留言，针对 A3 地区留言，比较了 0.65、0.6、0.62 三个数作为相似度阈值时的情况，确定 A3 地区留言的相似度阈值为 0.62。针对 A2 地区留言，确定 A2 地区留言的相似度阈值为 0.65。针对 A4 地区留言，确定 A4 地区留言的相似度阈值为 0.6。针对 A1 地区留言，确定 A1 地区留言的相似度阈值为 0.63。针对 A5 地区留言，确定 A5 地区留言的相似度阈值为 0.65。针对西地省地区留言，确定西地省地区留言的相似度阈值为 0.6。针对 A6 地区留言，确定 A6 地区留言的相似度阈值为 0.58。针对业主留言，确定业主留言的相似度阈值为 0.6。针对领导留言，确定领导留言的相似度阈值为 0.6。

通过以上 10 类地点或人群的相似度阈值情况分析，如图 9，对其他人群的相似度阈值做出预测，去除最大阈值 0.68 和最小阈值 0.58 后，求取平均数后得出预测阈值 0.61875，即后五类人群的关键词中的相似度阈值为 0.61875。

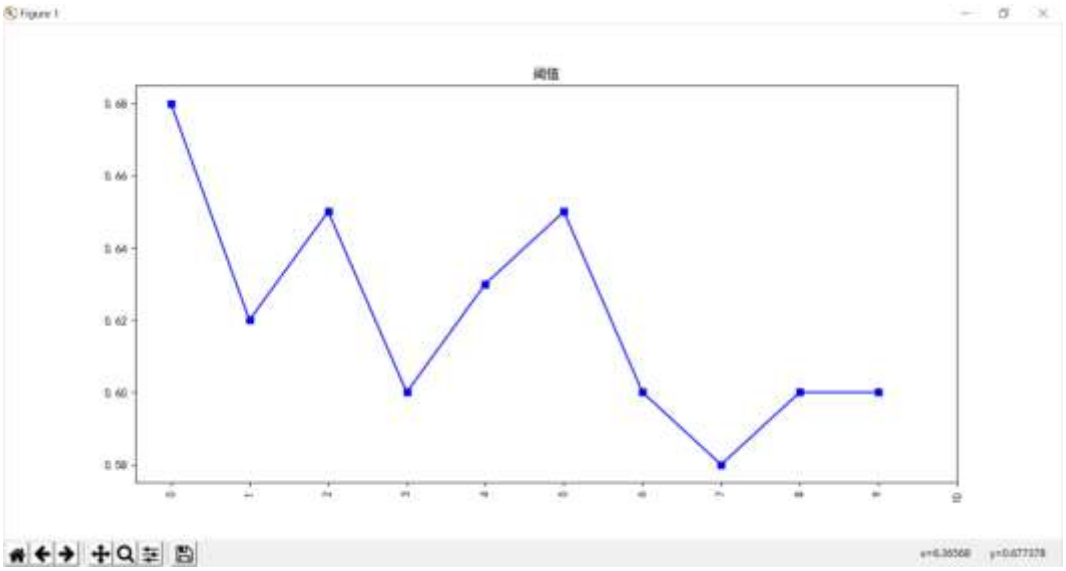


图 9 前 10 类阈值取值图

（5）求相似的留言的点赞总数和反对总数。

首先，找到点赞数和反对数均为 0 的留言的索引，使用该留言的索引填写相似度索引矩阵。其次，通过相似度索引矩阵，将留言相似的矩阵的点赞数总数和反对总数，存放到每个问题中索引最大的留言里，并将该问题的其他留言的索引的点赞数和反对数置为 0。

相似留言的点赞数和反对数求和实例如下：在 A2 地区 excel 表中，第 1 条留言的点赞数和反对数均为 0，其索引为 0，因此 x=0，从矩阵（7）形成矩阵（8）。若矩阵（8）对应的留言点赞数和反对数如表 7，则得到留言点赞矩阵（9），留言反对矩阵（11），将每个问题的点赞数和反对数求和后得到留言点赞矩阵（10）和留言反对矩阵（12），矩阵（10）和矩阵（12）求和得到矩阵（13），矩阵（13）的最后一列即为相似问题点赞数和反对数的总数。

$$\begin{bmatrix} 0 & 7 & 124 \\ 0 & 35 & 157 \\ 58 & 102 & 150 \\ 0 & 122 & 143 \end{bmatrix} \quad (8)$$

表 7 案例相似度矩阵详情表

索引	点赞数	反对数	索引	点赞数	反对数
0	0	0	122	3	1
7	8	0	124	2	0
35	2	1	143	1	0
58	0	0	150	11	0
102	1	3	157	0	2

$$\begin{bmatrix} 0 & 8 & 2 \\ 0 & 2 & 0 \\ 0 & 1 & 11 \\ 0 & 3 & 1 \end{bmatrix} \quad (9) \qquad \longrightarrow \qquad \begin{bmatrix} 0 & 0 & 10 \\ 0 & 0 & 2 \\ 0 & 0 & 12 \\ 0 & 0 & 4 \end{bmatrix} \quad (10)$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 3 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (11) \qquad \longrightarrow \qquad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 1 \end{bmatrix} \quad (12)$$

$$\begin{bmatrix} 0 & 0 & 10 \\ 0 & 0 & 2 \\ 0 & 0 & 12 \\ 0 & 0 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 10 \\ 0 & 0 & 5 \\ 0 & 0 & 15 \\ 0 & 0 & 5 \end{bmatrix} \quad (13)$$

3.2 模型的建立

3.2.1 模型表示

利用层次分析法，根据构建的多层次分析评价模型形成热度评价指标，多层次分析评价模型结构图如图 6。

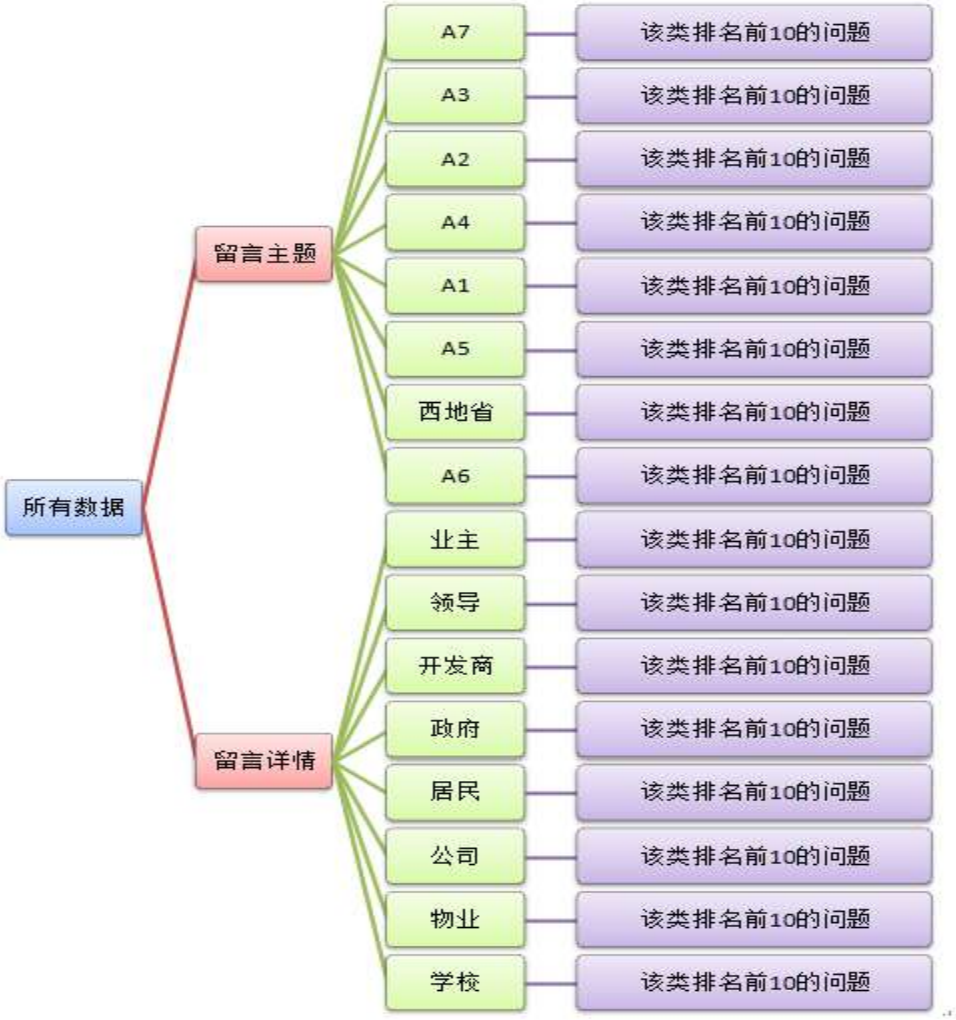


图 10 多层次分析模型结构图

3.2.2 模型中的名词解释

（1）互动数：点赞数和反对数的总和。

（2）初始分：该留言属于某关键词类的得分。

（3）附加分：该留言出现关键词的个数。

（4）互动分：互动数的得分。

（5）原始得分：其得到的排行榜用来测试有效得分排行榜。

（6）有效得分：其得到的排行榜信息最终建立热点问题表。

注：由于多个地点或人群中可能出现同一个的问题，最终取同一个问题在不同地点或人群类中不同的分数的平均分为该问题的得分。

3.2.3 热度评价指标的建立

（1）初始分的指标确立：由于本文分别根据留言主题和留言详情提取出具有先后顺序的关键词，因此本文根据关键词出现次数的多少决定初始分的高低。因此，初始分满分为 8，初始分得分指标如下表 8。

表 8 初始分得分指标表

排名	留言主题下的关键词	初始分得分	留言详情下的关键词	初始分得分
1	A7	8	业主	8
2	A3	7	领导	7
3	A2	6	开发商	6
4	A4	5	政府	5
5	A1	4	居民	4
6	A5	3	公司	3
7	西地省	2	物业	2
8	A6	1	学校	1

（2）附加分的指标确立：若某一个问题出现在多少个地区或人群中，则该问题加多少分。例如：a 问题出现 3 个关键词：A7、业主、政府，则 a 问题的附加分为 3。因此，利用 Counter 库进行了计数，结果如图 7，得到一个问题最多出现 7 个地点或人群，所以，附加分满分为 7。

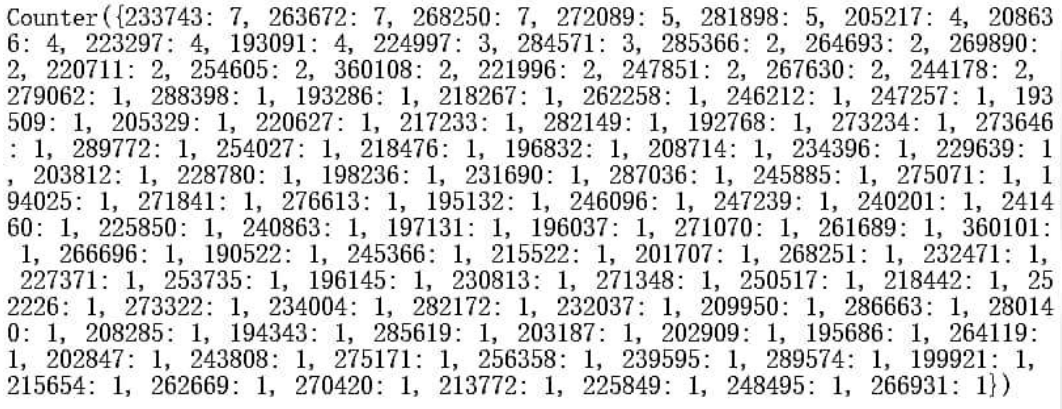


图 11 附加分得分指标结果图

（3）互动分的指标确立：根据所有关键词中的初步筛选的问题，得到互动数分布情况图，如图 8。根据图 8 可以看出，互动数分布不均匀，大部分分布在[0，500]这一区间内，所以本文对该区间的数目划分更加细致，最终建立互动分得分指标表，如表 9，因此，互动分满分为 20。

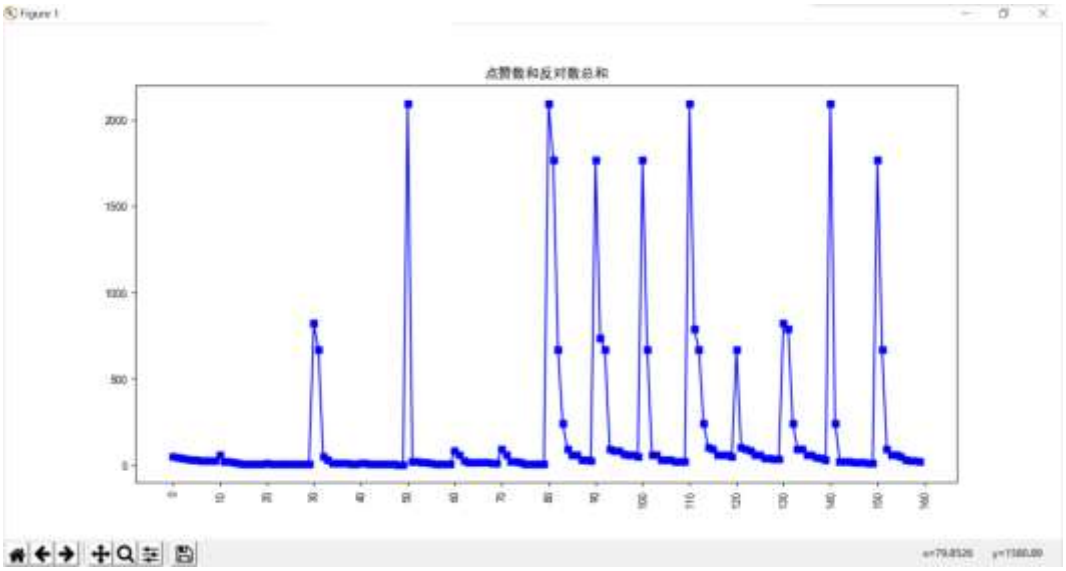


图 12 互动数分布情况图

表 9 互动分得分指标表

互动数 x 范围	互动分得分
$x \geq 2000$	20
$1000 \leq x<2000$	16
$800 \leq x<1000$	14
$600 \leq x<800$	12
$400 \leq x<600$	10
$200 \leq x<400$	8
$100 \leq x<200$	6
$80 \leq x<100$	5
$60 \leq x<80$	4
$40 \leq x<60$	3
$20 \leq x<40$	2
$0 \leq x<20$	1

（4）得分指标的确立。

考虑到影响留言热度有同一问题发布留言的条数和群众的讨论度两个因素。而初始分和附加分在原理上反映了同一问题发布留言的条数，互动分则反映了群众对问题的讨论度。

因此，本文在建立指标的时候对这两个因素分别使用不同的权重进行计算，得到原始得分与有效得分两个指标。原始得分是两个因素的权重均为 0.5 计算而来，实际上，本文对同一问题发布留言的条数和群众的讨论度两个因素权重分别赋予 0.3 和 0.7，比较而来发现最终结果雷同，因此选取了两个权重均为 0.5 这一情况作为原始得分。而有效得分即为同一问题发布留言的条数和群众的讨论度两个因素权重分别赋予 0.7 和 0.3 计算而来。

原始得分的指标确立：

$$\text{原始得分}=\left(\frac{\text{某问题的初始分}+\text{某问题的附加分}}{\text{初始分满分}+\text{附加分满分}}\right)*0.5+\left(\frac{\text{某问题的互动分}}{\text{互动分满分}}\right)*0.5$$

有效得分的指标确立：

$$\text{有效得分}=\left(\frac{\text{某问题的初始分}+\text{某问题的附加分}}{\text{初始分满分}+\text{附加分满分}}\right)*0.7+\left(\frac{\text{某问题的互动分}}{\text{互动分满分}}\right)*0.3$$

3.3 模型评估

根据热度指标分析，本文得到原始得分排行前七名的留言编号有：208636、223297、263672、220711、193091、268250、194343；有效分数排行前七名的留言编号有:263672、208636、223297、268250、233743、193091、281898（原始得分和有效得分排名根据热度从高到低排序）。

从两个排行榜中的差异性可知，该层次分析模型具有主观性，可以根据实际所需去构建需要的热点排行榜。由于本文希望在某一时间段反映特定地点或特定人群的基础上挖掘更具有代表广大居民问题的留言，所以选取有效得分排行榜前 5 的问题作为热点问题，构成热点问题表，见附件。最后，5 个热点问题对应的留言编号为：263672、208636、223297、233743、268250。

在此基础上，从所有数据的留言详情中寻找与上述 5 个热点问题对应的相似问题，经过多次调整阈值参数，找到上述热点问题相似度的阈值为 0.5，最终形成热点问题留言明细表，见附件。

4、问题 3 的分析与求解

4.1 制定评价方案的意义

目前，我国网络问政体系不断发展，网络已经成为政府搜集民意，与广大群众交流的重要平台。群众问题的回复是干部干事创业的生命线,是一个人工作能力的具体体现,更是抓落实的升级版。回复是推进工作落实的基本要求回复作为工作中的基本程序,在推进工作高效落实中起着重要作用。但是，在具体工作推进中,还存在一些“只埋头抓工作,而不及时回复”、“工作未完成,不好回复”、“工作没落实,不敢回复”的情况,对于工作高效落实产生了一定影响。而制定评价方案能够有效并直观的了解政府部门对群众回复的质量是否达到标准，同时也是监督政府在问政体系中存在的问题的方法。

4.2 模型的建立

4.2.1 评价指标体系结构

网络问政体系的构建和发展有效的提高了政府对群众的回复率，同时也更多的了解到群众们所关注的热点问题。但是政府的线上回复仍然存在一些回复不当、回复“逃避”、回复效用不大等一系列问题。因此，将根据政府对群众问题回复的数据进行等级评价，以此来更好的督促政府提高回复的质量。

本文根据留言答复数据的相关性、完整性、可解释性来进行评价。

4.2.2 建立评价模型

（1）对留言详情与答复意见分别进行关键字提取，留言详情与答复意见分别提取频次最高的 10 个关键词，利用 TF-IDF 算法将提取到的关键词向量化。

（2）对留言详情与答复意见的关键字进行余弦相似度比较，将比较得到的相似度作为数据的相关性。

（3）语义上的完整性是答复意见对留言详情所提到问题的答复是否完整，所以本文对留言详情关键字与答复意见关键字取交集，统计出同时出现在留言详情与答复意见中的关键词个数。把同时出现在留言详情与答复意见中的关键词的个数除以留言详情的关键字总数作为数据的完整性。

（4）语义上的可解释性是判断答复意见对留言用户的留言详情是否有说服力，所以我们把同时出现在留言详情与答复意见中的关键词的个数除以答复意见的关键字总数作为数据的可解释性。

（5）根据相关性、完整性、可解释性建立三维直角坐标系，其中相关性为 x 轴，完整性为 y 轴，可解释性为 z 轴，如图 13。通过对数据的获取，本文对每一条数据都可以得到相关性、完整性、可解释性所对应的数值，找到该条数据在建立好的三维坐标系的位置，通过计算该点到原点的相对距离作为评价答复意见的指标。如果该点到原点的相对距离越远，则针对这个问题给出的答复意见越好，也就是说答复意见的相关性、完整性、可解释性越好。否则，越不好。

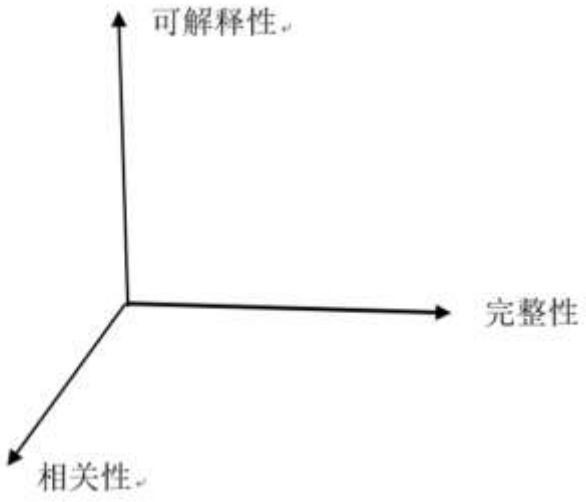


图 13 三维度评价图

4.2.3 相关性、完整性、可解释性的计算

（1）相关性： $similarity = \cos(\theta) = \frac{A * B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}$

其中，A 为留言详情关键字的向量值，B 为答复意见关键字的向量值。A_i 为 A 的分向量，B_i 为 B 的分向量。

（2）完整性： $\frac{|\text{留言详情关键字} \cap \text{答复意见关键字}|}{\text{留言详情关键字个数}}$

（3）可解释性：$\frac{|\text{留言详情关键字} \cap \text{答复意见关键字}|}{\text{答复意见关键字个数}}$

注：“｜ ｜”表示求集合中元素个数

4.2.4 评价指标计算公式

$$L=\sqrt{x^2+y^2+z^2} \quad (14)$$

其中 L 为该点到原点的距离，x、y、z 为该条数据的相关性、完整性、可解释性。

4.3 模型评估

本文从附件 4 的众多数据中，随机提取了五条数据对该模型进行评估。其中随机提取到的五条数据的留言详情与答复意见关键词如 5 条留言关键字内容图 13 所示。

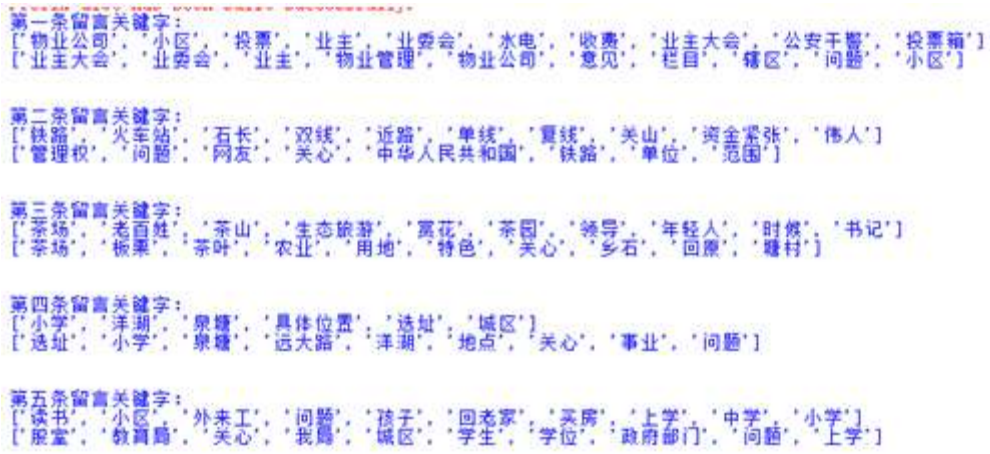


图 14 5 条留言关键字内容图

其中五条数据的相关性、完整性、可解释性如答复意见相关性、完整性、可解释性图 15 所示，其中答复意见相关性、完整性、可解释性图中的矩阵的逆对角线的值为相关性，为了计算的简便我们对数据进行四舍五入，对这五条数据的评估分别如下所示：

$L=\sqrt{0.34^2+0.5^2+0.5^2}$ =0.84 （数据一）

$L=\sqrt{0.06^2+0.1^2+0.13^2}$ =0.0869 （数据二）

$L=\sqrt{0.05^2+0.1^2+0.1^2}$ =0.07 （数据三）

$L=\sqrt{0.38^2+0.67^2+0.44^2}$ =1.0225 （数据四）

$L=\sqrt{0.11^2+0.2^2+0.2^2}$ =0.19 （数据五）

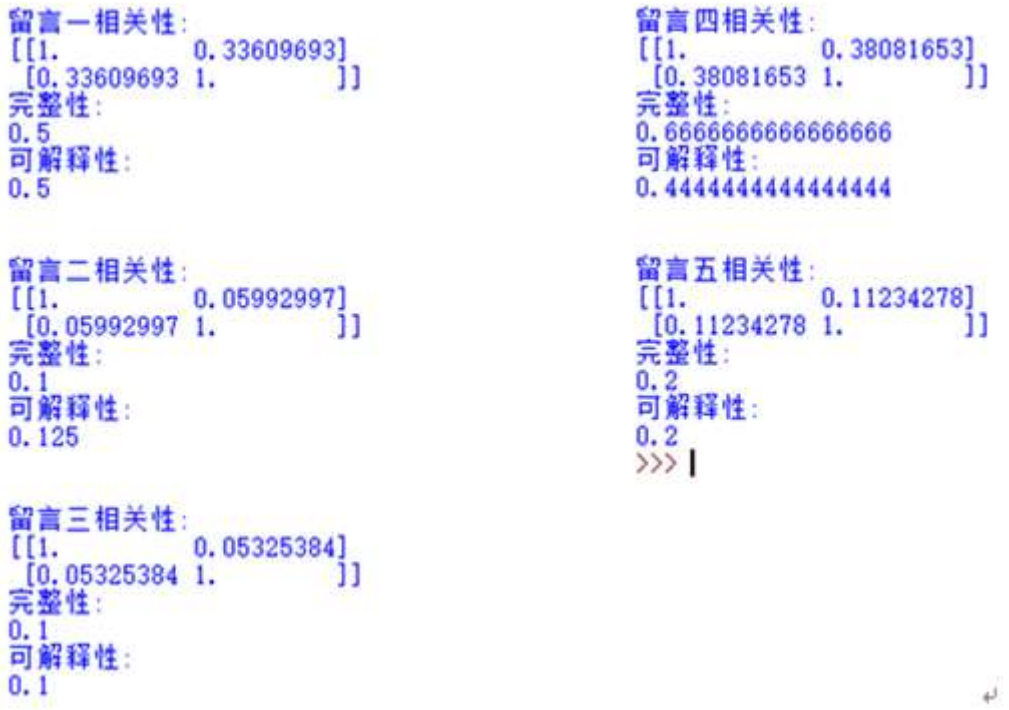


图 15 答复意见相关性、完整性、可解释性

根据所得到的相关性、完整性、可解释性，本文利用模型对这五条数据进行评估，得到图 16，其中红色代表数据一的点，深蓝代表数据二的点，紫色代表数据三的点，绿色代表数据四的点，浅蓝代表数据五的点。综上设想：在这五条留言中，针对第四条留言的留言详情所对应的答复的相关性、完整性与可解释性较好，第三条留言的留言详情所对应的答复的相关性、完整性与可解释性较差。

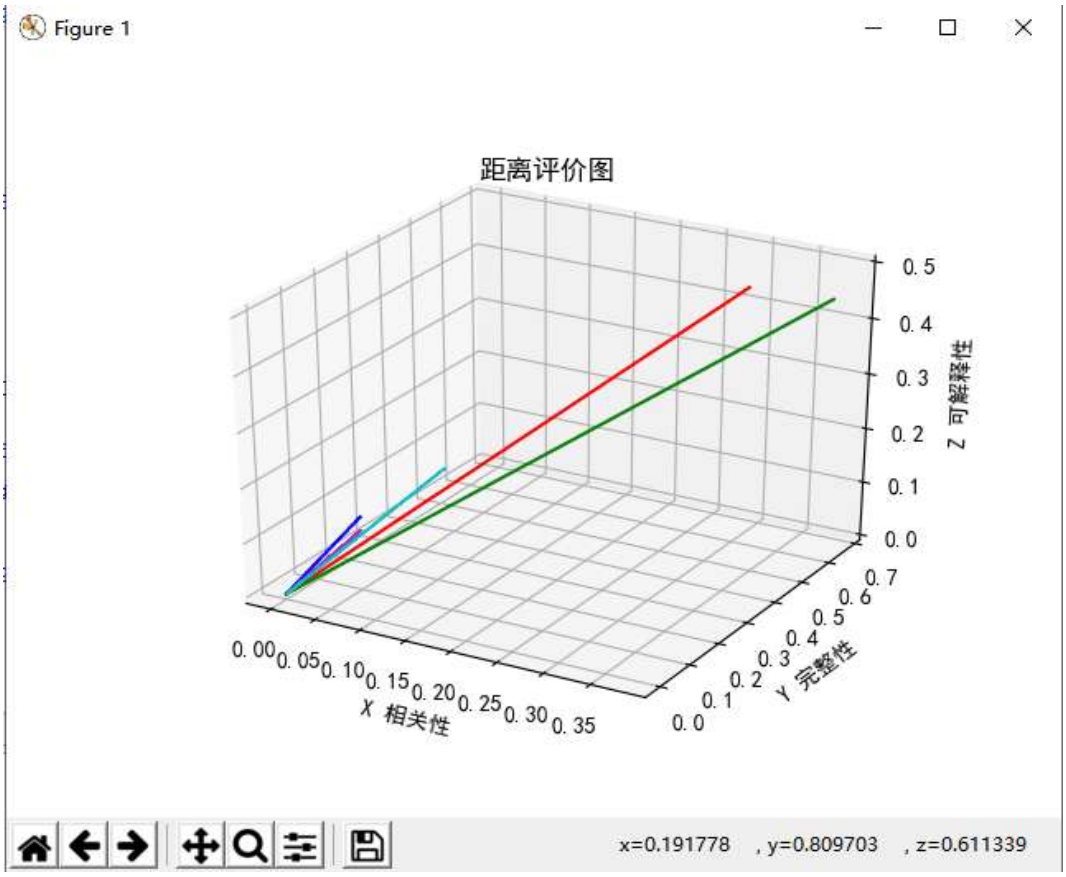


图 16 距离评价图

通过对五条数据的计算结果的分析，发现当数据的相关性、完整性、可解释性的值越大的时候，该点距离原点的距离越远，即对于该条留言详情，答复意见的相关性、完整性、可解释性越好。之后找到了这五条数据在附件 4 中的位置，通过对留言详情与答复意见之间的对比，与设想一致，认为本文给定的评价指标还不错。

5、结论

5.1 问题 1 的结论

根据线性支持向量机模型对留言内容的一级标签分类并评估如下：

一级标签	F_1指标（f1-score）
城乡建设	0.78
环境保护	0.82
交通运输	0.81
教育文体	0.87
劳动和社会保障	0.87
商贸旅游	0.77
卫生计生	0.81

5.2 问题 2 的结论

根据程序运行结果，得到热点问题表与热点问题明细表（详细请见作品附件）。热度排名前五的热点问题的留言编号如下：

Id	热度指数	留言编号
1	74.66667	191951
		202575
		216316
		243551
		263672
		266931
2	69.66667	208636
3	68.33333	223297
4	62.5	205960
		209742
		233743
5	59.5	254865
		262052
		268250
		272089

5.2 问题 3 的结论

由于每一条数据都可以得到相关性、完整性、可解释性所对应的数值，因此方便找到该条数据在三维坐标系中的位置，如图 17。计算该点到原点的相对距离作为评价答复意见的指标。若该点到原点的相对距离越远，则针对这个问题给出的答复意见越好，即答复意见的相关性、完整性、可解释性越好。反之，越差。

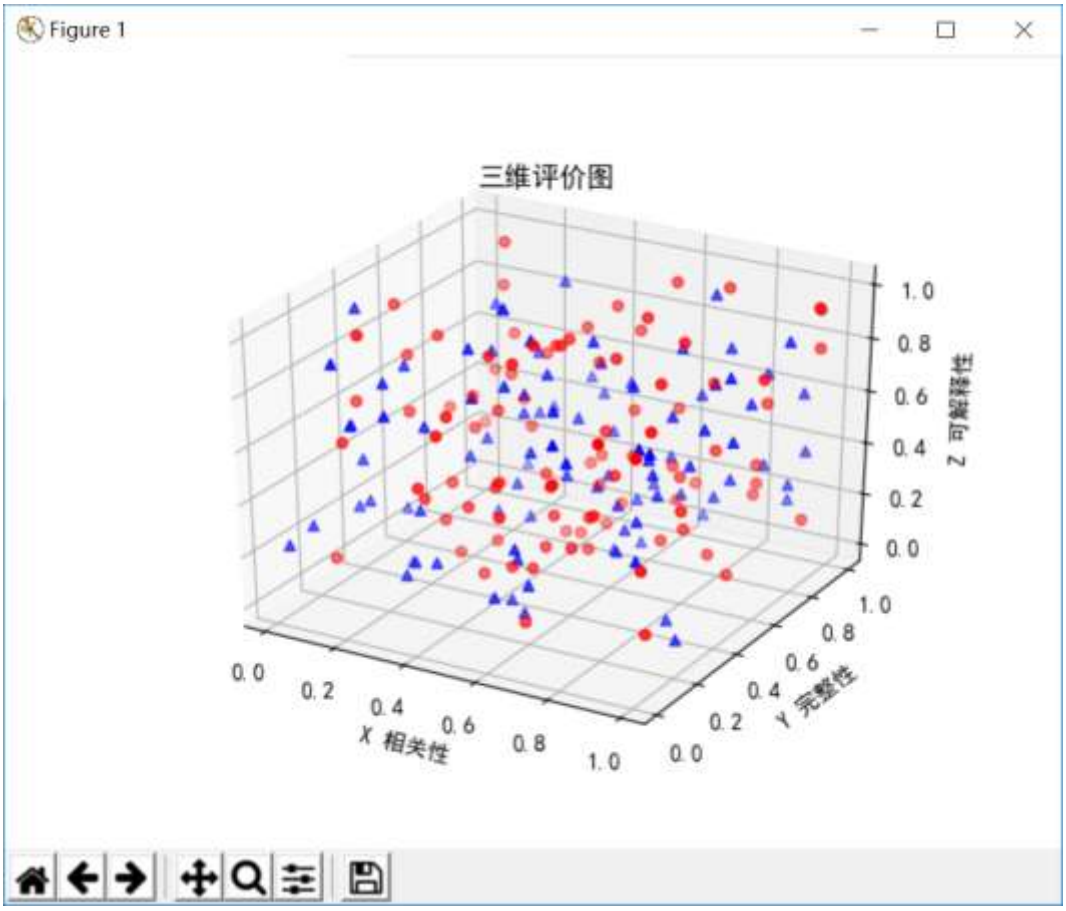


图 17 答复意见质量的三维度评价图

6 模型优缺点

6.1 问题 1 的模型优缺点

6.1.1 优点

（1）线性支持向量机方法简单，易于理解；

（2）简化了分类和回归的问题，实现了高效的从训练样本到预测样本的转导推理；

（3）利用线性支持向量机模型评估，最终结果由支持向量决定，支持向量是支持向量机的训练结果，因此在某种程度上，提高了结果的准确性。

6.1.2 缺点

（1）线性支持向量机对缺失数据敏感，所以要保证数据无缺失；

（2）在分词以及关键词提取中，存在一定的浪费和错误导致分类模型个别精确度不高；

（3）线性支持向量机(SVM)算法对大规模训练样本难以实施由于 SVM 是借助二次规划来求解支持向量，而求解二次规划将涉及 m 阶矩阵的计算（m 为样本的个数），当 m 数目很大时该矩阵的存储和计算将耗费大量的机器内存和运算时间。

6.2 问题 2 的模型优缺点

6.2.1 优点

（1）采用层次分析法，具有很强的条理性和科学性；

（2）模型处理过程中，有效地吸收了定性分析的结果，又发挥了定量分析的优势；

（3）在确定热度评价指标时可以根据决策者的侧重选择最终的结果。

6.2.2 缺点

（1）因为在确定热度评价指标时具有主观性，所以需要评价者对问题的本质等掌握的十分通透。

6.3 问题 3 的模型优缺点

6.3.1 优点

（1）对相关性、完整性、可解释性的模型求解简单易懂；

（2）本文的评估指标计算公式 L 为一个具体的数值，更直观的表现政府答复意见的合理性。

6.3.2 缺点

（1）模型中的可解释性并未运用机器学习进行运算，可能导致可解释性的值与实际值产生一定偏差。

7 模型的推广

随着网络技术的飞速发展和普及，进入了信息大爆炸的时代。在此大背景下，SVM(支持向量机）是一种很好的信息分类方法。尤其在解决小样本、非线性及高维度的模式识别问题中表现出许多优势，如文本分类、生物信息学、图像识别等领域中。另外，在其他领域，如汽轮发电机组的故障诊断，金融工程，生物医药信号处理，生物信息，自适应信号处理，手写体相似字识别，岩爆预测的支持向量机，缺陷识别等领域都有成功的应用。同样，层次分析法简洁实用，最终对目标层做出科学的评价，方便决策者进行决策。因此，被广泛应用于安全科学研究，例如油库安全性评价，城市灾害应急能力等。不仅如此，层次分析法已在大气环境研究、水环境研究等领域做出了研究。另一方面，在日常购物商品的选择，旅游目的地的安排等也可以使用层次分析法。而三维度评价法即从三个方面对某一事物进行评价，在岗位评价、效绩考核方面有突出贡献，可以应用于公司、企业等。

8、参考文献

[1]王振武，数据挖掘算法原理与实现，出版地：清华大学出版社，2015 年。
[2]CSDN，使用 python 和 sklearn 的中文文本多分类实战开发，https://blog.csdn.net/weixin_42608414/article/details/88046380，2020 年 5 月 2 日。
[3]嵩天，python 语言程序设计基础，出版地：高等教育出版社，2010 年。
[4]刘斌，黄铁军，程军等，一种新的基于统计的自动文本分类方法，中文信息学报，2002，16(6)：18-24。
[5]搜狗百科，层次分析法，https://baike.sogou.com/v354235.htm，2020 年 5 月 4 日。
[6]CSDN，Python 模块之 Difflib（文本对比，原创 source），https://blog.csdn.net/wwqqyy123456/article/details/84866470?utm_source=app，2020 年 5 月 4 日。
[7]MBA 智库 百科，三维度评价法，https://wiki.mbalib.com/wiki/%E4%B8%89%E7%BB%B4%E5%BA%A6%E8%AF%84%E4%BB%B7%E6%B3%95?from=singlemessage，2020 年 5 月 6 日。