“智慧政务”中文本挖掘应用

摘要：网络问政平台的群众留言的分类以往都是由人工处理的，本文针对群众留言建立了 SVM 分类和 Kmeans 聚类模型，并且构造了问题热度评价指标，构建了答复意见质量评价体系。

针对问题一，在进行数据预处理后，将类别因素引入传统计算方法 TF-IDF 构造词向量空间，再用 SVM 分类器对数据做分类训练，得出分类模型，并用 F-score 对分类方法进行评价，结果显示分类器效果良好。

针对问题二，类似问题一的处理，在构造词向量空间后，用 K-Means 聚类算法对数据进行聚类，并用热度指标提炼出目标结果。

针对问题三，利用 gensim 包分析文档相似度，从答复的相关性、完整性、可解释性、及时性等角度对构建答复意见的质量评价指标体系，使用层次分析法得出各项指标的权重，从而对答复意见质量进行评估。

关键词：SVM，K-Means，层次分析法，文本挖掘

1 绪论

1.1 研究背景、目的及意义

1.1.1 研究背景

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。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.1.2 研究目的及意义

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（参考附件 1 提供的内容分类三级标签体系）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。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。

及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

1.2 研究对象和内容

本文划分为四章，每章内容如下：

（1）第一章是绪论部分。对研究背景、目的及意义、国内外研究现状、研究对象和内容及研究方法进行系统阐述。

（2）第二章是群众留言分类部分。首先对附件 2 的数据进行清洗（去重，降维，异常值），利用机器学习中常用的评估方法（留出法、交叉验证法、自助法）划分训练集和测试集，然后通过有监督学习建立关于留言内容的一级标签分类模型，并使用 F—Score 对建立的分类模型进行评价。

（3）第三章是热点问题挖掘部分。首先根据附件 3 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，然后通过查阅相关文献定义合理的热度评价指标，并给出评价结果，最后按表 1 的格式给出排名前 5 的热点问题，按表 2 的格式给出相应热点问题对应的留言信息。

（4）第四章是答复意见的评价部分。针对附件 4 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，并尝试实现。

1.3 研究方法

（1）定性分析与定量分析相互补充.

本文首先对互联网公开来源的群众问政留言记录及相关部门对部分群众留言进行定性分析，然后对群众反映的留言进行挖掘，通过公式、模型等，建立基

于自然语言处理技术的智慧政务系统。

（2）模型与计算机软件的使用.

以数据挖掘和机器学习算法为理论基础，以 Python 和 SPSS 软件为工具，以群众留言的文本信息为样本，建立分类模型。

2 群众留言分类

2.1 文本分类的相关理论

F.Sebastiani 将文本分类表达为以下的数学形式：为得到$\Phi : D \times C \to \{T, F\}$这样的一个函数，其中$D = \{d_1, d_2, \cdots, d_{|D|}\}$表示待分类的文本信息，$C = \{c_1, c_2, \cdots, c_{|C|}\}$表示预定义的分类体系下的类别集合，$T$值表示对于$(d_j, c_i)$而言，文本$d_j$属于类$c_i$，相反，$F$值表示对于$(d_j, c_i)$而言文本$d_j$不属于类$c_i$。简单地说，文本分类的目标是需要寻找一个有价值的函数映射，它可以准确的完成从$D \times C$到T/F值的函数映射，这个映射过程本质上讲就是所谓的分类器。

文本分类的形式化定义如下：

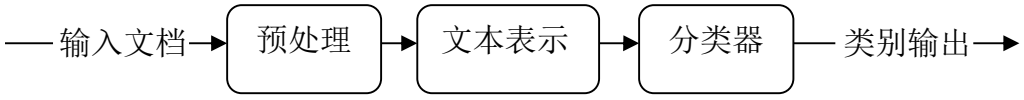
假设$i = 1, \cdots, M$为文本集合里的M条文本信息，$j = 1, \cdots, N$为预先定义的N个类别主题，可以给出这样一个分类矩阵$C = (c_{ij})$；其中矩阵中某一元素c_{ij}表示第i条文本信息与第j个类别的关系。简单地说，文本的自动分类任务可以归结为确定矩阵C的每一个元素值的过程；为此，可以使用一个布尔量 1 或 0，当c_{ij}的值为 1 时，则表示第 i 条文本属于第 j 类，否则，表示第 i 条文本不属于第 j 类，表示如下：

$$c_{ij} = \begin{cases} 1 & i \in j \\ 0 & i \notin j \end{cases} \tag{2.1}$$

我们称某条文本信息只允许被分入一个类别的任务是单类别的分类，可以增加限定条件，但是对于第j行$(j = 1, \cdots, N)$的所有元素，必须满足：

$$\sum_{i=1}^{N} c_{ij} = 1 \tag{2.2}$$

文本分类系统如下图所示：



统计机器学习(Statistical Machine Learning)中文本分类系统的组成部分包括：

1. 文本预处理模型(Text Preprocessing Model)，
2. 文本表示模型(Text Expressing Model)，
3. 特征选择模型(Feature Selection Model)，
4. 学习训练模型(Learning and Selection Model)，
5. 分类处理模型(Classification Processing Model)，
6. 性能评估模型(Performance Evaluation Model)。

对于大多数的有监督机器学习算法，学习训练模型仅需在分类预测前运行一次即可；而性能评估模型主要起到评估训练模型学习效果，衡量分类精度的作用。

文本分类的一般流程如下图所示：

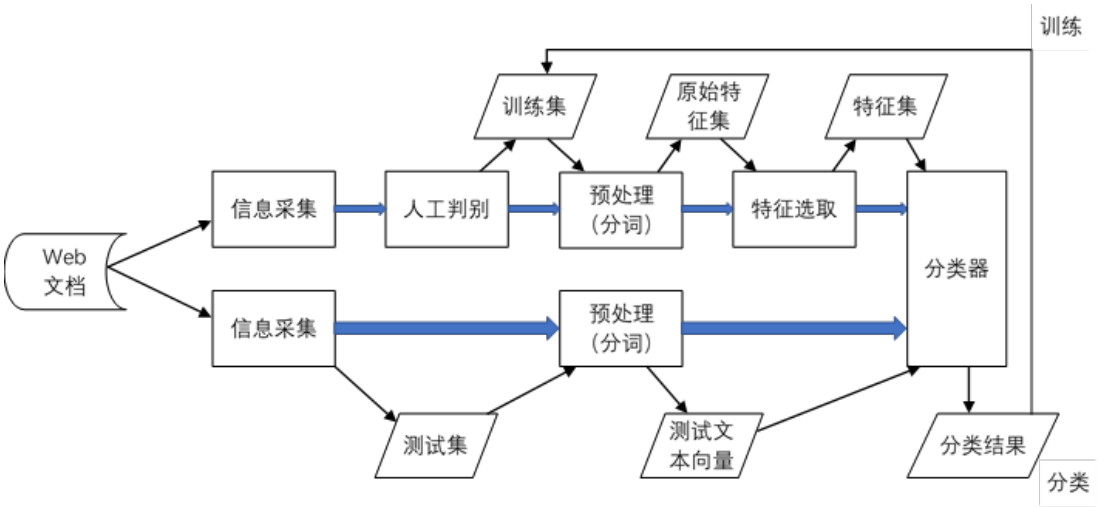


图 2.1 文本分类流程图

首先通过 Python 的 jieba 库对附近 2 的留言主题进行分词，并利用 wordcloud 库绘制词云，以下给出城乡建设和环境保护的词云图，其余词云放于附件一问题一种。

图 2.2 城乡建设词云展示

图 2.3 环境保护词云展示

过滤停用词，过滤掉已经确认对训练模型没有实际价值的分词，以提高模型的准确度和稳定性，下面仅给出城乡建设和环境保护过滤停用词后的词云展示，其余一级标签下的词云展示放于附件一问题一中。

图 2.4 城乡建设过滤停用词后的词云展示

图 2.5 环境保护过滤停用词后的词云展示

2.2 模型建立

2.2.1 SVM 模型的相关理论

支持向量机（Support Vector Machine，简称 SVM）的基本型[1]：

给定训练样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}, y_i\in\{-1,+1\}$，分类学习最基本的想法就是基于训练集 D 在样本空间中找到一个划分超平面，将不同类别的样本分开。

在样本空间中，划分超平面可通过如下线性方程来描述：

$$w^Tx+b=0 \tag{2,3}$$

其中 $w=(w_1;w_2;\cdots;w_d)$ 为法向量，决定了超平面的方向；b 为移位项，决定了超平面与原点之间的距离。划分超平面可被法向量 w 和位移 b 确定，样本空间中任意点 x 到超平面 (w,b) 的距离可写为

$$r=\frac{\left|w^Tx+b\right|}{\|w\|} \tag{2.4}$$

假设超平面 (w,b) 能够正确分类训练样本，即对于 $(x_i,y_i)\in D$，若 $y_i=+1$，则 $w^Tx_i+b>0$；若 $y_i=-1$，则 $w^Tx_i+b<0$。令

$$\begin{cases} w^Tx_i+b\geq+1, y_i=+1; \\ w^Tx_i+b\leq-1, y_i=-1. \end{cases} \tag{2.5}$$

距离超平面最近的这几个训练样本点使得等号成立，称之为“支持向量”，两个异类支持向量到超平面的距离和即“间隔”为

$$\gamma = \frac{2}{\|w\|} \tag{2.6}$$

找使得间隔最大的划分超平面，也就是要找到能满足式(4)中约束的参数 w 和 b，使得 γ 最大，即

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|} \\ &s.t.\ y_i\left(w^T x_i + b\right) \geq 1, i = 1,2,\ldots,m. \end{aligned} \tag{2.7}$$

相当于

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &s.t.\ y_i\left(w^T x_i + b\right) \geq 1, i = 1,2,\ldots,m. \end{aligned} \tag{2.8}$$

可由式(6)得到对应的“对偶问题”，即

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &s.t.\ \sum_{i=1}^{m} \alpha_i y_i = 0 \\ &\quad \alpha_i \geq 0, i = 1,2,\ldots,m. \end{aligned} \tag{2.9}$$

解出 α 后，求出 w 与 b 即可得到模型 $f\left(x\right) = w^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$

上述过程需要满足 KKT 条件，即

$$\begin{cases} \alpha_i \geq 0; \\ y_i f\left(x_i\right) - 1 \geq 0; \\ \alpha_i (y_i f\left(x_i\right) - 1) = 0. \end{cases} \tag{2.10}$$

最终模型只与支持向量有关，大部分训练样本不需要保留。通常用 SMO 算法求解对偶问题。

2.2.2 核函数

在现实任务当中，能正确划分两类样本的超平面往往不一定存在，造成了不是线性可分的情况。为了应对此类问题，往往采用将样本从原始空间映射到更高维的特征空间的方式，使得样本在此特征空间中线性可分。

令 $\phi\left(x\right)$ 表示将 x 映射后的特征向量，于是，在特征空间中划分超平面所对应的模型可表示为 $f\left(x\right) = w^T \phi\left(x\right) + b$ 其中 w 和 b 是模型参数，原问题可写成

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &s.t.\ y_i\left(w^T \phi\left(x_i\right) + b\right) \geq 1, i = 1,2,\ldots,m. \end{aligned} \tag{2.11}$$

则对偶问题为

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\phi\left(x_i\right)^T\phi\left(x_j\right)\\&s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0\\&\alpha_i\geq 0, i=1,2,\ldots,m.\end{aligned}\tag{2.12}$$

式(10)中需要计算样本映射到特征空间后的内积$\phi\left(x_i\right)^T\phi\left(x_j\right)$。又因为特征空间维数有可能比较高，甚至是无穷维，因此，直接计算$\phi\left(x_i\right)^T\phi\left(x_j\right)$往往比较困难。为避免此问题，可设想函数$k\left(x_i,y_j\right)=\left\langle\phi\left(x_i\right),\phi\left(y_j\right)\right\rangle=\phi\left(x_i\right)^T\phi\left(y_j\right)$。

x_i和y_j在特征空间的内积等于它们在原始样本空间中通过函数$k\left(\cdot,\cdot\right)$计算得出的结果，因此，不必直接计算高维特征空间中的内积，式(10)可重写为

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k\left(x_i,y_j\right)\\&s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0\\&\alpha_i\geq 0, i=1,2,\ldots,m.\end{aligned}\tag{2.13}$$

求解得

$$\begin{aligned}f\left(x\right)&=w^T\phi\left(x\right)+b\\&=\sum_{i=1}^{m}\alpha_i y_i\phi\left(x_i\right)^T\phi\left(x\right)+b\\&=\sum_{i=1}^{m}\alpha_i y_i k\left(x,x_i\right)+b\end{aligned}\tag{2.14}$$

这里的$k\left(\cdot,\cdot\right)$即为核函数。式(12)显示出模型的最优解可通过训练样本的核函数展开，这一展式即为支持向量展式。

常用的核函数有线性核，多项式核，高斯核，拉普拉斯核以及 Sigmoid 核，本文中采用的是高斯核，即$k\left(x_i,y_j\right)=\exp\left(-\frac{\left\|x_i-y_j\right\|^2}{2\sigma^2}\right)$，其中$\sigma>0$为高斯核的带宽。

本文研究的问题为多分类问题，因此本文采用多个二类支持向量机的组合来解决该问题 [3]。

首先我们对留言主题划分为训练集(70%)和验证集(15%)，然后计算其所对应的每一行，匹配总词数，计算得到每一行的每一个词的 TF-IDF，最后将提取的 TF-IDF 利用 SVM 模型中进行训练，得到以下结果：

2.3 模型评价

通常使用 F-Score 对分类方法进行评价：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}\tag{2.15}$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

表 2.1 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

查准率 P 和查全率 R 的公式定义如下：

$$P = \frac{TP}{TP + FP} \tag{2.16}$$

$$R = \frac{TP}{TP + FN} \tag{2.17}$$

其中 $TP + FP + TN + FN$ 是样本总数。

以下是利用 SVM 模型得到的分类结果：

表 2.2 类别混淆矩阵

类别	TP	FN	FP	TN	P	R
城乡建设	1887	122	478	6723	0.797886	0.939273
环境保护	834	104	37	8235	0.95752	0.889126
交通运输	374	239	9	8588	0.976501	0.610114
教育文体	1468	121	62	7559	0.959477	0.923851
劳动和社会保障	1903	66	222	7019	0.895529	0.923851
商贸旅游	1055	160	116	7879	0.900939	0.868313
劳动和社会保障	734	143	31	8302	0.959477	0.836944

计算得到 F-Score 为 F=0.88072

接下来我们赋予各个一级标签不同的权重（按照数量占比），对其提取不同个数的特征，得到改进的类别混淆矩阵，如下表所示：

表 2.3 类别混淆矩阵

类别	TP	FN	FP	TN	P	R
城乡建设	1897	112	421	6780	0.818377912	0.944250871
环境保护	857	81	38	8234	0.957541899	0.913646055
交通运输	471	142	21	8576	0.957317073	0.768352365
教育文体	1467	122	66	7555	0.956947162	0.923222152
劳动和社会保障	1854	115	129	7112	0.93494705	0.941594718
商贸旅游	1042	173	144	7851	0.878583474	0.857613169
劳动和社会保障	759	118	44	8289	0.945205479	0.865450399

同样使用 F-Score 对分类方法进行评价，得到改进后的 F-Score 为 F=0.902，与改进前相比有所提升，这也表明按权重提取特征是可行的。

3 热点问题挖掘

3.1 文本聚类的相关理论

对文本聚类主要有几个步骤：对文本分词、构建词袋模型、权值转换、计算余弦相似度、选取聚类算法、性能度量、确定聚类结束条件。如果文本所含的词汇量较多，并且已知分类的个数 k，可以选择二分 K-均值聚类算法。而层次聚类算法根据样本距离来分类，并且可以以样本距离作为分类结束的条件，比较适合 k 未知的情况。

聚类分析作为一种无监督机器学习算法，即训练样本的标记信息是未知的，其思想是尽可能的将相似的对象划分到相同的簇，将不相似的对象划分到不同的簇。使用聚类分析算法对文本进行分类，需要解决以下几个问题：

(1) 寻找合适的度量标准以衡量不同的对象是否相似；
(2) 探索合适的度量标准以衡量算法的性能；
(3) 制定分类的个数或聚类结束的条件；
(4) 选择最优的分类算法以实现理想的分类结果。

通过查阅文献，本文选择余弦相似度作为度量的标准，可以用余弦相似度即向量空间中两个向量夹角的余弦值作为衡量两个个体差异的大小。余弦值的计算公式如下：

$$\cos\theta = \frac{\sum_{i=1}^{n}\left(A_i \times B_i\right)}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

与欧氏距离相比，余弦相似度则更加注重两个向量在方向上的差异，而忽略其在距离或长度上的差异。

思路如下：首先将文本转换为权值向量，通过计算两个向量的夹角余弦值，就可以评估他们的相似度。因为余弦的取值范围在[-1,1]之间，如果余弦值越趋近于 1，则代表两个向量方向越接近；相反，越趋近于-1，代表其方向越相反。简单起见，可以将余弦值做归一化处理，将其取值范围转换到[0,1]之间，并且值越小代表距离越近。

为得到较好的聚类结果，需要满足划分到同一个簇内的样本尽可能相似，而不同簇间的样本尽可能不同，即聚类得到的结果应符合“簇内相似度”高且“簇间相似度”低的性质。

假设通过聚类得到的簇为$C=\{C_1,C_2,\cdots,C_K\}$，给出如下定义[1]：

$$avg(C)=\frac{2}{|C|(|C|-1)}\sum_{1\le i<j\le|C|} dist(x_i,x_j) \tag{3.2}$$

$$diam(C)=\max_{1\le i<j\le|C|} dist(x_i,x_j) \tag{3.3}$$

$$d_{\min}(C_i,C_j)=\max_{x_i\in C_i,x_j\in C_j} dist(x_i,x_j) \tag{3.4}$$

$$d_{cen}(C_i,C_j)=dist(\mu_i,\mu_j) \tag{3.5}$$

其中，$dist(\cdot,\cdot)$表示不同样本之间的距离，μ表示簇C的中心点，可以表示为

$\mu=\frac{1}{|C|}\sum_{1\le i\le|C|}x_i$；$avg(C)$表示簇$C$内样本的平均距离；$diam(C)$表示簇$C$内样本之间的最远距离；$d_{\min}(C_i,C_j)$表示簇$C_i$和簇$C_j$最近样本之间的距离；$d_{cen}(C_i,C_j)$表示簇$C_i$和簇$C_j$中心点间的距离。

基于以上公式可导出下面两个常用的聚类性能度量内部指标：

DB 指数（Davies-Bouldin Index，简称 DBI），公式如下：

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{cen}(\mu_i,\mu_j)}\right) \tag{3.6}$$

由公式可知 DBI 的值越小，表示簇内距离越小而簇间距离越大。

Dumn 指数（Dumn Index，简称 DI），公式如下：

$$DI=\min_{1\le i\le k}\left\{\min_{j\neq i}\left(\frac{d_{\min}(C_i,C_j)}{\max_{1\le i\le k}diam(C_i)}\right)\right\} \tag{3.7}$$

由公式可知 DI 的值越大，表示簇间距离越大而簇内距离越小。

因此，如果 DBI 的值越小，同时 DI 的值越大，可以认为聚类的效果越好。

层次聚类 [1]：

层次聚类的思想是选择不同的层次对数据集进行划分，从而形成树形的聚类结构。数据集的划分可以采用“自底向上”的聚合策略，也可以采用“自顶向下”的分拆策略。

AGNES 是一种常见的采用“自底向上”聚合策略的层次聚类算法，其思想是首先将数据集中的每个样本看作一个初始聚类簇，然后在算法运行过程中寻找距离最近的两个聚类簇进行合并，不断重复该步骤，当达到预设的聚类个数或某种条件时停止。算法的关键是计算不同聚类簇之间的距离，事实上每个簇都是一个样本集合，因此只需要计算集合的某种距离即可。下面是簇C_i和簇C_j之间距离的计算公式：

$$\text{最小距离：}\ d_{\min}(C_i,C_j)=\min_{x\in C_i,z\in C_j}dist(x,z) \tag{3.8}$$

$$\text{最大距离：}\ d_{\max}(C_i,C_j)=\max_{x\in C_i,z\in C_j}dist(x,z) \tag{3.9}$$

$$\text{平均距离：}\ d_{avg}(C_i,C_j)=\frac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{z\in C_j}dist(x,z) \tag{3.10}$$

显而易见，最小距离由两个簇的最近样本决定，最大距离由两个簇的最远样本决定，平均距离由两个簇的所有样本决定。

接下来要考虑如何确定一个合适的聚类个数或某种结束条件，具体思路是：

（1）首先选定部分样本作为测试样本，对其进行层次聚类分析。

（2）计算 DBI 和 DI 的变化趋势，与人工校验相结合，得到一个合适的聚类个数以及其对应的距离阈值。

（3）将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。

3.2 留言分类

分词：

要对中文文本做聚类分析，首先利用 python 提供的 jieba 库对文本做分词处

理，将中文长文本划分为若干个单词。

过滤停用词：

为了提高分类的准确率，还要考虑两个干扰因素：一是英文字母大小写的影响，为此我们将英文字母统一转换为大写；二是例如“影响”、“公司”、“建议”等通用的词汇，我们将这样的词汇连同“（）”、“-”、“/”、“&”等符号作为停用词，将其从分词结果中去除掉，最后得到有效的词汇组合。

统计词频并绘制词云：

在提取特征之前，可以先绘制词云，从宏观角度进一步了解文本信息。

3.3 热度评价指标体系构建

本文首先对评论中反应的所有的问题进行分类并描述。易知，热点问题与每条评论息息相关，所以，本文首先定义了每条评论的热度指数[7]，然后再根据分类后所得的问题中每条评论的热度指数以及对应问题持续时间长度构造热点问题热度指数。

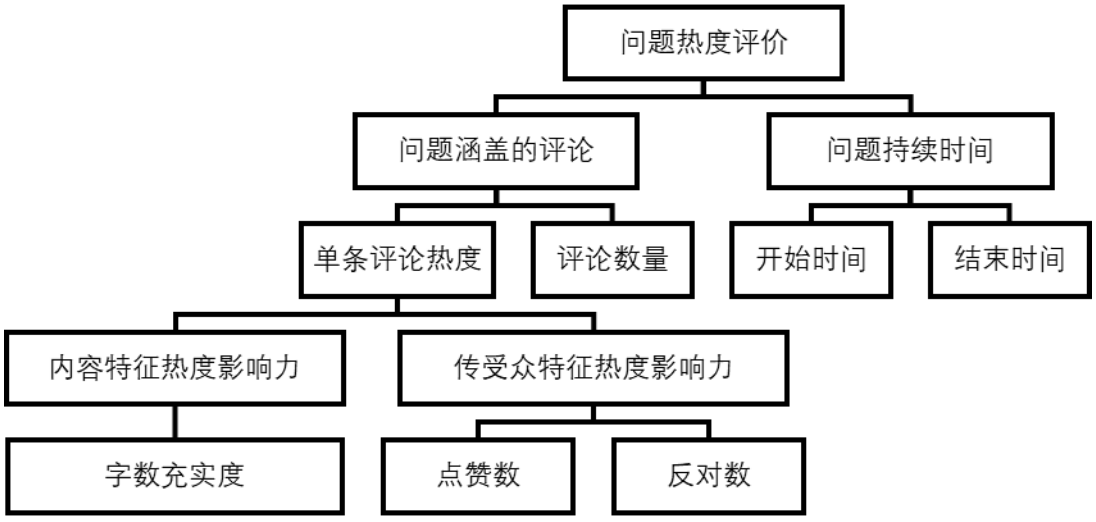


图 3.1 问题热度评价指标体系

第 i 个问题的第 j 条评论的热度用 H_{ij} 表示，则

$$H_{ij}=2^{(th_{ij}+op_{ij})}\times\log x_{ij} \tag{3.11}$$

其中，x_{ij} 表示该条评论的字数，本文用 $\log x_{ij}$ 来量化其字数充实度，th_{ij} 表示该条评论的点赞数，而 op_{ij} 表示该条评论的反对数。

第 i 个问题的热度指数用 HT_i 表示，则

$$HT_i=\left(\sum_{i=1}^{n_i}H_{ij}\right)\cdot(t_{2i}-t_{1i}) \tag{3.12}$$

其中，t_{1i} 表示第 i 个主题中评论出现的最早时间，t_{2i} 表示第 i 个主题中评论出现的最晚时间，n_i 表示第 i 个主题中评论的数量。

3.4 模型结果

表 3.1 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	5.001	2019/7/8 至 2019/8/16	A7 县诺亚山林小区门口	反对在 A7 县诺亚山林小区门口设立医院
2	2	5.00025	2019/7/18 至 2019/7/23	A7 县新国道 107	A7 县新国道 107 距离某居民家仅 3 米，政府相关部门对其拆迁申请不予同意
3	3	5	2019/8/10 至 2019/9/2	A 市	带着情怀提升 A 市规划建设水平，带动经济发展
4	4	4.00125	2019/7/21 至 2019/8/1	A5 区劳动东路魅力之城小区	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
5	5	4.001	2019/1/3 至 2019/10/24	A 市金茂府	A 市金茂府 5-4304 一房二卖

考虑到篇幅结构问题，我们将热点问题明细表放于附件一问题 2 中：

表 3.2 热点问题留言明细表

留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
210107	A00042107	反对在A7县诺亚山林小区门口设置	2019/7/8 10:38	纠纷，危害我们的安宁居住环境和可能出现	14	0
210107	A00042107	对在A7县诺亚山林小区门口设立医	2019/7/8 10:48:31	、医闹纠纷，危害我们的安宁居住环境和可	3	0
210107	A00042107	反对在A7县诺亚山林小区门口设置	2019/7/8 10:39:54	纠纷，危害我们的安宁居住环境和可能出现	1	0
226871	A000726	反对在A7县诺亚山林小区门口设置	2019/8/16 8:37:43	力低的老人、小孩的身心健康带来极大的威	1	0
226871	A000726	反对在A7县诺亚山林小区门口设置	2019/8/16 8:36:38	住环境和可能出现的人身误伤；严重影响小	0	0
273925	A00020543	07距我家仅3米，相关政府部门为何	2019/7/18 10:47:31	蒙，据不完全统计，整栋房屋开裂约有26处	3	0
273925	A00020543	07距我家仅3米，相关政府部门为何	2019/7/22 17:04:08	蒙，据不完全统计，整栋房屋开裂约有26处	2	0
273925	A00020543	107距我家仅3米，相关部门为何不	2019/7/23 11:03:10	蒙，据不完全统计，整栋房屋开裂约有26处	0	0
273925	A00020543	07距我家仅3米，相关政府部门为何	2019/7/22 17:05:04	蒙，据不完全统计，整栋房屋开裂约有26处	0	0
273925	A00020543	07距我家仅3米，相关政府部门为何	2019/7/18 10:48:28	蒙，据不完全统计，整栋房屋开裂约有26处	0	0
255614	A00094811	提升规划建设水平带动经济发展的些	2019/8/10 15:31:33	行欲望；特别是拓改和平改立等重复建设多	1	0
258379	A00010435	情怀提升A市规划建设水平，带动经	2019/8/19 22:27:22	也是挺惬意的，北方城市傲的好）；特别是	1	0
258379	A00010435	情怀提升A市规划建设水平，带动经	2019/9/2 22:04:30	也是挺惬意的，北方城市傲的好）；特别是	0	0
255614	A00094811	提升规划建设水平带动经济发展的些	2019/8/10 15:13:26	骑行欲望；特别是拓改和平改立等重复建设	0	0
258379	A00010435	着情怀提升A市规划建设水平带动经	2019/8/12 8:05:32	骑行欲望；特别是拓改和平改立等重复建设	0	0
272122	A909113	城小区一楼的夜宵摊严重污染附近	2019/08/01 16:20:02	还是觉得要维护社会和谐稳定，合法维权。	6	0
360107	A0283523	城小区一楼的夜宵摊严重污染附近	2019-08-01 16:20:02	还是觉得要维护社会和谐稳定，合法维权。	0	6
272122	A909113	魅力之城小区一楼的夜宵摊严重污	2019/07/21 10:29:36	还是觉得要维护社会和谐稳定，合法维权。	3	0
360107	A0283523	魅力之城小区一楼的夜宵摊严重污	2019-07-21 10:29:36	还是觉得要维护社会和谐稳定，合法维权。	0	3
200316	A00018085	A市金茂府5-4304一房二卖	2019/1/3 12:55:49	金茂府存在“一房二卖”违法行为。A市住	2	0
200316	A00018085	A市金茂府5-4304一房二卖	2019/1/3 12:53:28	房二卖”违法行为。A市住建委在2018年7月	1	0
200316	A00018085	再次反映A市金茂府一房二卖问题	2019/10/24 15:58:46	房二卖”违法行为。A市住建委在2018年7月	0	0
200316	A00018085	再次反映A市金茂府一房二卖	2019/3/13 21:15:16	房二卖”违法行为。A市住建委在2018年7月	0	0

4 答复意见评价

4.1 引言

根据既知信息，本文选用相关性，完整性，可解释性以及及时性等指标来综合考量答复意见质量。首先，获取指标相关信息，把评论及其对应答复按照问题一的方式操作分别得到对应的关键词以及一级分类类别；按照问题二中方式操作分别得到评论和答复的问题描述；计算应答时长。随后，根据相关信息得到评价指标。最后，用层次分析法[2]来对答复意见质量进行评价。本文所构建的答复意见质量评价指标体系如图 4.1 所示。

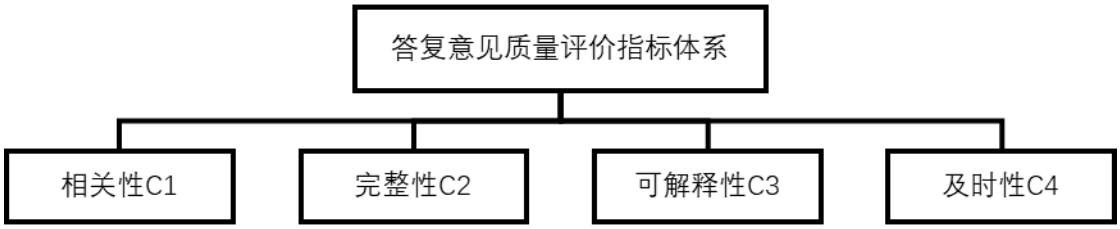


图 4.1 答复意见质量评价指标体系

4.2 层次分析步骤

1. 建立层次结构模型，其中，相关性用分类结果的一致程度表示；完整性用所属问题一致程度以及答复字数与评论字数之比描述；可解释性用评论关键词与答复意见关键词一致程度来衡量；而及时性用应答时间来衡量。
2. 构造成对比较矩阵，本文用 1-9 比较尺度（如表 4.1 所示）构造成对比较矩阵。本文构造的成对比较矩阵为

$$A=\begin{bmatrix} 1 & 5 & 7 & 3 \\ 1/5 & 1 & 2 & 1/3 \\ 1/7 & 1/2 & 1 & 1/2 \\ 1/3 & 3 & 2 & 1 \end{bmatrix}$$

3. 计算权向量并做一致性检验。成对比较矩阵最大的特征根为 $\lambda=4.1085$，对应的特征向量(归一化后)为 $w=[0.5814 \quad 0.1136 \quad 0.0810 \quad 0.2241]$，一致性指标为 $CI=\frac{\lambda-n}{n-1}=0.0362$，随机一致性指标数值 RI 如表 4.2 所示，当 $n=4$ 时 RI=0.90，一致性比率为 $CR=\frac{CI}{RI}=0.0402<0.1$，因此检验通过。所以四项指标的权重为

$$w=[0.5814 \quad 0.1136 \quad 0.0810 \quad 0.2241].$$

表 4.1 1-9 尺度 a_{ij} 的含义[2]

尺度 a_{ij}	含义
1	C_i 与 C_j 的影响相同
3	C_i 比 C_j 的影响稍强

5	C_i 比 C_j 的影响强
7	C_i 比 C_j 的影响明显的强
9	C_i 比 C_j 的影响绝对的强
2，4，6，8	C_i 与 C_j 的影响之比在上述两个相邻等级之间
1，1/2，，1/9	C_i 与 C_j 的影响之比为上面 a_{ij} 的互反数

表 4.2 随机一致性指标 RI 的数值[2]

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

5 结论与展望

针对问题一，本文通过对网络问政平台的群众留言内容清洗，特征选择及目标类别建立起关于留言内容的一级标签 SVM 分类模型。

针对问题二，通过对群众留言当中的反应的问题进行凝炼，针对留言建立 Kmeans 聚类模型，对问题进行聚类，然后定义并计算每条评论的热度指数，最后定义并计算组问题的热度指数，该指数需综合考量单条评论热度指数以及该组问题热度持续时间。

针对问题三，首先选取合适的指标并建立答复意见质量评价指标体系，对评论和答复的内容的一级标签分别进行分类，对评论以及答复所反映的问题进行凝炼并计算答复时长；其次，根据提取的信息计算各指标的值并用层次分析法计算各指标的权重；最后，根据各指标的值对答复意见进行评估。

参考文献

[1] 周志华. 《机器学习》[J]. 中国民商, 2016(3).

[2] 姜启源,谢金星,叶俊编.《数学模型（第四版）》[J]. 高等教育出版社, 2011.(1):252-256.

[3] 梁中杰. 互联网信息采集分析系统的研究及实现[D].重庆大学,2005.

[4] 王春华. 基于互联网的人力资源供求信息挖掘分析系统研究与实现[D].山东大学,2011.

[5] 伍毅敏.基于文本挖掘的两版北京城市总体规划公众意见对比分析[J].北京规划建设,2018(01):87-94.

[6] 马小龙.网络留言分类中贝叶斯复合算法的应用研究[J].佛山科学技术学院学报(自然科学版),2013,31(02):43-47+68.

[7] 李晓娴. 微博热点话题发现的研究[D]. 2014.