“智慧政务”中的文本挖掘应用

摘 要

近年来，科学技术发展日新月异，这给人们的生活、工作等各方面带来了极大的便捷。大数据、人工智能、云计算等现代技术的出现与飞速发展也给各级政府部门的相关工作带来了福音。如今的政务问题，不再像过去那样只能依靠落后的人工来解决。正是如此，“智慧政务”这一概念也应运而生。而“智慧政务”中的文本挖掘应用技术对“智慧政务”的建设和研究具有重大的意义。 本文将基于数据挖掘技术对收集自互联网公开来源的群众问政留言记录进行深入的信息挖掘，并建立模型来解决留言划分、热点整理和答复意见评价等问题。

对于问题一，本文首先将附件 2 中的留言主题进行去重去空、去 x 序列、中文分词和去停用词等数据预处理，再利用 TF-IDF 算法把留言主题转换为权重向量。最后对词向量基于支持向量机构建分类模型，通过精确率、召回率以及F_1系数对该模型进行评价并发现其准确率可达 90%，较好的实现了留言分类工作。

对于问题二，本文利用词袋模型来提取留言主题的特征向量，采用基于平均连接法的凝聚式层次聚类算法，对留言数据集进行聚类。并综合考量热点问题的影响因素，定义了依赖于用户留言数目与用户参与度的热度评价指数。随后，本文对每一个留言类通过命名实体识别方法提取地名和人群信息，查找每一类所有留言时间中的最大值和最小值进行拼接作为时间范围，取各类留言中的一条留言主题去停用词作为问题描述的结果，最后将所得信息按照热度评价指数排名顺序得到表 1-热点问题表和表 2-热点问题留言明细表。

对于问题三，本文从相关性、完整性、可解释性和及时性四个角度对答复意见的质量构建了评价指标。首先，本文将利用 HanLP 这个工具包实现 TextRank 算法，完成对留言主题和答复意见关键词的提取。然后，将这些关键词进行比对，根据比对情况，将它量化并作为相关性数值指标。其次，为构建完整性和可解释性评价指标，本文通过正则表达式检索答复文本中开头和结尾有无固定格式，以及有无出现一般格式的法律法规条文。最后参考各省规定，设置回复时间的阈值，判断留言回复是否及时并量化为及时性数值指标。答复意见综合质量的评定指标将由上述四个指标加权求和而得。

关键词：TF-IDF，支持向量机，层次聚类，命名实体识别，正则表达式

Abstract

In recent years, the rapid development of science and technology has brought great convenience to people's life and work. The emergence and rapid development of big data, artificial intelligence, cloud computing and other modern technologies have also brought good news to the relevant work of government departments at all levels. Today's government problems are no longer solved by backward manpower as they used to be. Just like this, the concept of "smart government" came into being. The application technology of text mining in "smart government" is of great significance to the construction and research of "smart government". In this paper which will be based on data mining technology, we will conduct in-depth information mining on the records of public political messages collected from the open sources of the Internet, and establish a model to solve the problems of message division, hot spot sorting and reply comments evaluation.

For the first problem , this paper first preprocesses the message subject in Annex 2, such as de duplication and de emptiness, de X sequence, Chinese word segmentation and de deactivation words, and then uses TF-IDF algorithm to convert the message subject into weight vector. Finally, the classification model of word vector is built based on the support vector mechanism. The model is evaluated by the accuracy rate, recall rate and F1 coefficient, and the accuracy rate is up to 90%, which is a good implementation of message classification.

For the second problem,this paper uses the word bag model to extract the feature vector of the message topics, and uses the clustering hierarchical clustering algorithm based on average join to cluster the message data set. Considering the influence factor s of hot issues, the heat evaluation index is defined, which depends on the number of users'comments and the degree of users'participation. Then, this paper uses the named entity recognition method to extract place name and crowd information from each me ssage class, finds the maximum and minimum of all message time in each class, and s plices them as the time range.Finally, according to the ranking order of the heat rating index, we can get the table 1-hot issues table and table 2-hot issues table.

For the third question, this paper constructs the evaluation index of the quality o f the response from four aspects: relevance, completeness, interpretability and timelin ess.First , this paper will implement TextRank algorithm by using HanLP toolkit to co mplete the extraction of the message topic and the comment keyword. Then, these key words are compared, according to the comparison, it is quantified and used as a correl ation value index. Secondly, in order to construct the index of integrality and interpret ability, this paper searches the text of the reply by regular expression to find whether t here is a fixed format at the beginning and the end, and whether there are laws and reg ulations in the general format. Finally, referring to the provincial regulations, set the t hreshold of reply time to judge whether the message reply is timely and quantized as a numerical index of timeliness. The indicators for assessing the overall quality of resp onses will be weighted by the above-mentioned four indicators.

Keywords: TF-IDF;Support Vector Machine;Hierarchical Clustering;Named Entity

Recognition;Regular Expression

目 录

1 挖掘目标

本次挖掘目标是利用互联网公开来源的群众问政留言记录，及相关部门对群众留言的答复意见，在对文本进行基础的数据预处理，如去重去空、中文分词、去停用词等操作后，通过建立多种合适的数据挖掘模型，来实现群众留言分类、热点问题挖掘、答复意见的评价。

2 总体流程图

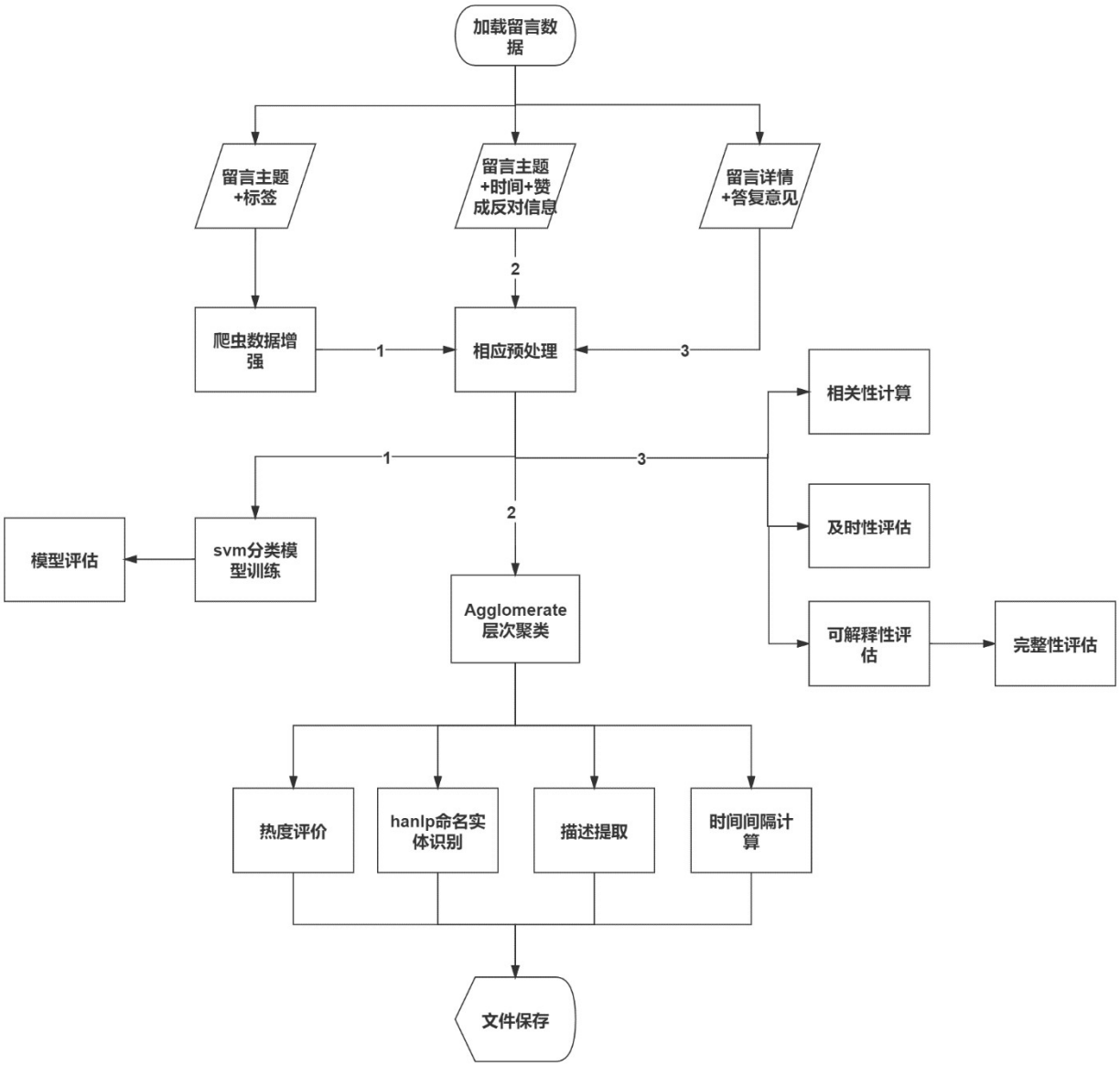


图 2-1 总体流程图

3 问题 1 留言分类

3.1 数据预处理

高质量的数据集是模型适配和优化的基石，对全部数据集进行分析处理可以促进对数据集的详细认知。对于初始数据大致要从准确性和完整性两个方面去探索。本文数据处理的过程主要分为以下几步：

3.1.1 数据增强

观察已经得到的数据集，可以发现样本数据很多。在全部数据中，七个一级标签所对应的留言数量比值为 2048:1026:653:1509:1993:1219:887。但遗憾的是，样本数目却存在着不平衡不均匀的问题。在这种情况下，能够收集到的样本数目不满足模型训练的要求，同时它导致模型对于小样本类别往往处于欠拟合的状态。为解决这一问题，本文采用回译进行文本数据增强。采用此方法的原因在于，回译不单单有近似替换同义词的能力，它还拥有在保持原意的前提下增加或移除单词并重新组织句子的能力。在这个方法中，我们通过注册百度翻译开发者账户获得翻译 api 接口，用机器翻译把小样本类别对应数据下的留言主题和留言详情翻译成其他语言，然后再翻译回中文，把获得的回译文本与原始文本相拼接，从而将每一类留言样本的数量增强到 2000 左右，获得较为平衡的训练数据。部分回译文本展示如下：

原文：A市坪塘大道施工及大车噪音严重扰民
回译结果一： 平塘大道建设及车辆噪声严重干扰a市居民生活
回译结果二： A.市平堂大道施工和车辆噪音给居民造成严重伤害。
1:1/938原文：A8县市花明楼镇郑家冲的贵庭住工沥青味太重
回译结果一： a8县华明楼镇郑家冲贵庭住宅柏油味太浓
回译结果二： A8。县华明路邑廷家忠的贵亭商住站的沥青味道太重了。
1:2/938原文：A8县市花明楼镇花明楼村贵庭住工毁我池塘，生产沥青毒气！
回译结果一： A8县华明楼镇华明楼村，贵亭小区工程毁了我的水塘，产生了沥青瓦斯！
回译结果二： A8。县城华明路镇华明路村的贵政居民破坏了我的池塘，生产反青气体！

图 3-1 部分回译文本展示图

3.1.2 数据去重、去空及去 X 序列

在题目给出的数据中，出现了很多重复的留言记录。例如：同一个用户重复地发布同样的或极其相似的留言。考虑到这种重复性的留言会干扰到数据集的有

效性。本文在去重的时将针对每一个用户取更新时间最晚的留言，去掉历史留言。考虑到 python 中的字典在保存数据时，key 不变，value 取值为最后更新的值。所以在读取数据时，本文按时间升序把留言用户当为 key，把所有留言信息作为 value 保存在 value 中。最后再将字典中的内容写入文本即可。

与此同时，群众留言数据里存在一些空缺的或不完全的记录，如果把这些留言数据也引入进行中文分词、文本聚类、文本分类等一系列操作，势必会对中间的分析过程形成消极影响，而据此得到的结果必然存在极大的问题。那么，在使用这些群众留言数据之前就一定要把大量的这些空缺的、不完全的留言记录清除。为有效解决这一问题，本文将采用直接过滤的方法对留言数据进行去空。

另外在留言中包含许多对分类无意义的包含有数字信息（银行卡号，电话，生日，时间等）和包含字母的城市信息的 x 序列，同样利用正则表达式将其去除。

3.1.3 中文分词

欲通过机器来进行数据分析等操作，直接将原始未经加工的文本投喂给机器显然是荒唐的。在人与人的交流中，相比较于英文等其他语言，中文的理解要难得多。原因主要在于，在中文当中词和词之间是没有空格的，这样的语言习惯导致了歧义等问题的产生。人尚且如此，机器就更难直接理解中文了。所以，在自然语言处理领域中，很基础但又极为重要的一项技术就是中文分词技术。正如著名专家郝玺龙所说:“没有中文分词，其他一切深入的中文信息处理都无从谈起”。该技术是将中文连续的字序列按照一定的规则重新组合成词序列。

中文分词技术的发展大致经历了三个方向，最初的方向是依赖词典，后来逐渐转为基于语法和规则的分词法的方向，近年来又转为基于统计的分词法方向。国内外的技术研究者在中文分词领域进行了全面的探索和研究，也发表了众多合理且有效的算法。如今，比较受大众认可的分词方法有以下三种：基于字符串匹配的分词方法、基于统计的分词方法、基于知识理解的分词方法。下面，本文将对这几大主流方法进行简要的介绍：

1.基于字符串匹配的分词方法

它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。按照扫描方向的不同，字符串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大匹配和最小匹配；按照是否与词性标注过程相结合，可以分为单纯分词方法和分词与词性标注相结合的一体化方法。该方法简单、容易理解、易于实现，但过于依赖词典且对歧义和未登录词处理效果不佳。

2.基于统计的分词方法

该方法是在给定大量已经分词的文本的前提下，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。基于统计的分词方法所应用的主要的统计量或统计模型有：N 元文法模型（N-gram），隐马尔可夫模型（HiddenMarkovModel, HMM），最大熵模型（ME），条件随机场模型（ConditionalRandomFields, CRF）等。这些统计模型主要是利用词与词之间的联合出现的概率作为分词判断的信息。在实际的应用中，基于统计的分词系统都需要使用分词词典来进行字符串匹配分词，同时使用统计方法识别一些新词，即将字符串频率统计和字符串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

3.基于知识理解的分词方法

该方法主要基于句法、语法分析，并结合语义分析。通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和词义信息，用来对分词歧义进行判断。这类方法试图让机器具有人类对理解能力，需要使用大量的语言知识和信息。由于汉语的复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于知识理解的分词系统还处在试验阶段。

综上所述，本文选用基于统计的分词方法，利用 python 中的 jieba 库，对留言实现分词。jieba 采用基于前缀词典实现的高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。对于未登陆词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。在经过前面所述的文本数据清洗处理后，将所得留言文本进行分词，通过针对性的加入部分留言场景相关词汇，逐步完善分词结果，最后得到高准确度的留言内容的分词序列，为留言分类模型的建立做好准备。以第一条留言主题为例，分词效果如下：

分词前：A 市经济学院体育学院变相强制实习

分词后：A 市 经济学院 体育学院 变相 强制 实习

3.1.4 去停用词

留言详情在经过去重、去空、中文分词后，并非所有的剩下的词语都可以作为特征词，里面还有一些包含的信息量很低甚至没有信息量的词语，需要将它们

过滤掉，否则会影响下文的分析正确率。为了后面模型训练的效率以及准确性，在模型构建之前，下载停用词表，同时针对留言文本添加停用词典去除停用词。停用词结果示例：

分词后：A 市 经济学院 体育学院 变相 强制 实习

停用词过滤后:经济学院 体育学院 强制 实习

在分词处理后，原始的文本会被处理成一系列由空格隔开的单个词。接下来，可以使用词频统计等方法，在构建语言模型的过程中得到词向量，从而实现对留言的进一步处理。

3.2 分类模型建立

3.2.1 构建词向量空间

文本是非结构化数据，而分类算法所能处理的是结构化的数字数据。所以把文本从非结构化转化为结构化这一过程是整个文本分类工作的基石，这一转化过程的好坏将直接影响到最终的分类结果。目前常用的文本表示方法有：布尔模型（Boolean Model）、概率模型（Probabilistic Model）、向量空间模型（Vector Space Model），其中又以向量空间模型（简称 VSM）的使用最为简单。

VSM 模型中有三个重要概念：一是特征项（Term）， 特征项是经过特征选择所得的最能表达文本内容的词语；二是特征项权重（Weight），特征项权重是使用相应的特征权重算法对特征词赋值所得的权值；三是特征向量（Feature Vector)，特征向量是用特征项和特征项权重共同表示的文本数字化向量。VSM 把每篇文档都表示为特征词-权重向量，把文本看作是一系列特征项t的集合，对每个特征项赋予对应的权值。特征项$t_1,t_2,...,t_n$可以看作是一个 n 维坐标系，而权值$w_1,w_2,...,w_n$表示其对应的坐标值，每篇文档d_i映射为该向量坐标空间中的一个特征向量$V(d_i)=(t_1,w_{i1};t_2,w_{i2};...;t_n,w_{in})$。文档集总体的 VSM 表示见图 3-2。

	t_1	t_2	t_3	$\cdots$	t_n
d_1	w_{11}	w_{12}	w_{13}	$\cdots$	w_{1n}
d_2	w_{21}	w_{22}	w_{23}	$\cdots$	w_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
d_m	w_{m1}	w_{m2}	w_{m3}	$\cdots$	w_{mn}

图 3-2 文档集总体 VSM 表示

本文使用 TF-IDF 算法把留言记录信息转换为权重向量。TF-IDF 算法的具体

原理如下：
第一步，计算词频，即 TF 权重(Term Frequency)。

$$\text{词频(TF)}=\text{某个词在文本中出现的次数} \tag{1}$$

考虑文本有长短之分，为便于不同文本的比较，进行“词频”标准化。

$$\text{词频(TF)}=\frac{\text{某个词在文本中出现的次数}}{\text{文本的总词数}} \tag{2}$$

或

$$\text{词频(TF)}=\frac{\text{某个词在文本中出现的次数}}{\text{该文本出现次数最多的词的出现次数}} \tag{3}$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率(IDF)}=\log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数+1}}\right) \tag{4}$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

$$\text{TF-IDF}=\text{词频(TF)}\times\text{逆文档频率(IDF)} \tag{5}$$

实际分析得出 TF-IDF 值与一个词在留言详情中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，TF-IDF 值最大的即为要提取的留言记录中的文本关键词。
生成 TF-IDF 向量的具体步骤如下：
（1）使用 TF-IDF 算法，找出每条留言记录的前 6 个关键词；
（2）对每条留言记录提取的 6 个关键词，合并成一个集合，计算每条留言记录描述对于这个集合中词的词频，如果没有则记为 0；
（3）生成每条留言记录的 TF-IDF 权重向量，计算公式如下：

$$\text{TF-TDF=词频(TF)}\times\text{逆文档评率(IDF)} \tag{6}$$

一个好的词向量可以实现词义相近的一组词在词向量空间中也是接近的，可以通过显示词向量空间中相近的一组词并判断它们语义是否相近来评价词向量构建的好坏。

3.2.2 支持向量分类机

支持向量机（Support Vector Machine,简称 SVM）是在统计学习理论基础

上提出的一种新型通用的机器学习方法。自 1995 年 Vapnik 提出 SVM 作为模式识别的新方法之后，SVM 一直备受关注。它建立在结构风险最小化原则基础上，具有很强的学习能力。

SVM 可看作一种广义的线性分类器，其基本思想是：通过非线性变换将输入空间变换到一个高维的特征空间，并在新空间中寻找最优的线性分界面。对于线性可分的情况，可用图 3-3 说明。图 3-3 反映的是支持向量机的两类（Two Class）问题模型，其中的“+”和“-”分别表示两类训练样本，x_1和x_2为样本的两个特征项，H 为分界面，H_1和H_2为分别过两类样本中离分界面最近的点且平行于分界面的平面。在支持向量机模型中，要确保经验风险最小，因此为取得最优分界线时不仅要求该分界线能正确的将两类数据分开，而且要使得他们之间的分类间隔（Margin,图中 M）最大。因此，图中的H′虽然也能正确分开两类数据，却得不到最大的分类间隔，因此不适合作为分界线使用。

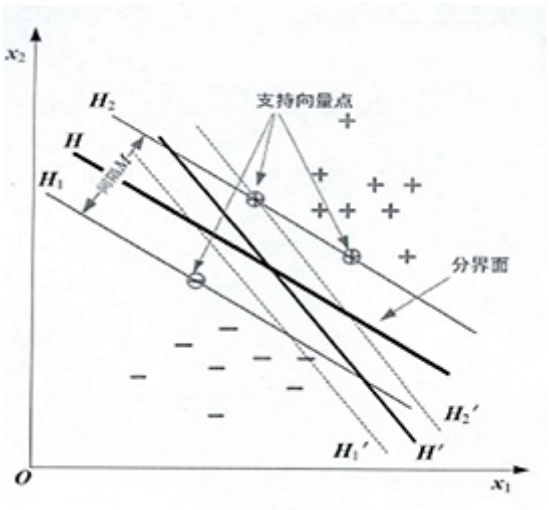


图 3-3 支持向量机分界面示意图

设两类问题的线性可分样本集为$\{(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_N, y_N)\}$,其中$\mathbf{x}_i \in R^d$（即数据为 d 维），类别标号$y_i \in \{1,-1\}, i \in [1,N]$。则该 d 维的输入空间中的线性判别函数的一般形式为:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{1}$$

其中$\mathbf{w}$为 d 维向量，b 为常量。图中的最大分类间隔为 M，所取得的分界面需要满足下面的不等式：

$$\mathbf{w}^T\mathbf{x}+\mathrm{b}\begin{cases} > \dfrac{M}{2} & \text{for } \mathrm{y}_i = 1 \\ < \dfrac{M}{2} & \text{for } \mathrm{y}_i = -1 \end{cases} \tag{2}$$

将上述不等式归一化，使所有样本都满足$|y(\mathbf{x})| \geq 1$，并且距离分界面最近的样本满足$|y(\mathbf{x})| = 1$，因此有

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1,...,N \tag{3}$$

据此得到的分类间隔 M=$\frac{2}{\|\mathbf{w}\|}$，因此，要使得分类间隔最大，就必须使得$\|\mathbf{w}\|$最小。同时，为了使目标函数成为二次规划问题，取$\|\mathbf{w}\|^2$最小，因此有

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1,...,N \end{aligned} \tag{4}$$

利用 Lagrange 方法，可以得到其对应的 Lagrange 函数如下:

$$L(\mathbf{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) \tag{5}$$

根据 Karush-Kuhn-Tucker 条件（简称 KKT 条件），有

$$\frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial \mathbf{w}} = 0 \tag{6}$$

$$\frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial b} = 0 \tag{7}$$

$$\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) = 0 \quad \text{for } i = 1,...,N \tag{8}$$

$$\alpha_i \geq 0 \quad \text{for } i = 1,...,N \tag{9}$$

在将式（6）、（7）带入式（5）后得到原目标函数的 Wolfe 的对偶问题

$$\begin{aligned} &\min_{\alpha} \quad \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\alpha_i \\ &\text{s.t} \quad \sum_{i=1}^{N} y_i\alpha_i = 0, \quad \alpha_i \geq 0 \quad \text{for } i = 1,...,N \end{aligned} \tag{10}$$

其中，α_i 是 Lagrange 乘子。求解该问题后得到最优解$\boldsymbol{\alpha}^* = [\alpha_1^*,...,\alpha_N^*]^T$，可计算出最优超平面

$$y(\mathbf{x}) = \sum_{i=1}^{N}\alpha_i y_i < \mathbf{x}\cdot\mathbf{x}_j > + b \tag{11}$$

其中，$b = y_i - \sum_{i=1}^{N} y_i\alpha_i^* < \mathbf{x}_i\cdot\mathbf{x}_j >$。这里的 y_j 对应于$\boldsymbol{\alpha}^*$ 的任意的一个正分量α_j^*。

该超平面对应的决策函数 $g(\mathbf{x})=sign(y(\mathbf{x}))$。很明显，只有$\boldsymbol{\alpha}^*$中的那些大于零的

分量对应的训练集样本才对构造超平面或决策函数有实际意义，这些样本就是支持向量（图 3-3 中用圆圈圈住的 3 个支持向量点）。

前面式（10）描述的是线性可分的情况，而实际需要处理的经常都是线性不可分的数据集，即非线性可分问题。因此，分类过程不可避免地会出现一些错分的情况发生。为此，有两种方式用于减小经验风险：一是通过非线性变换将输入空间的非线性可分问题映射到更高维的特征空间中转化为线性可分问题；二是通过引入松弛因子 ξ_i 来调整分类面允许分类过程存在一定的错分样本，并使用惩罚因子 C 控制错分的损失。对于前者，由于寻找这样的非线性映射函数 $\phi(\cdot)$ 非常复杂，不易实现。仔细观察式（10）和（11），可发现训练和决策过程都仅仅与训练样本间的内积 $<\mathbf{x}_i \cdot \mathbf{x}_j>$ 有关，因此可直接引入核函数完成从线性问题推广到非线性问题。在使用支持向量机解决问题时，核函数的选择相当关键，常常需要根据具体的问题构造相应的核函数。常见的核函数如下表 3-1 所示：

表 3-1 常见核函数

核函数	计算公式
线性核	$K(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i^T\mathbf{x}_j$
多项式核	$K(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i^T\mathbf{x}_j+b)^d$
高斯核（径向基函数核）	$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(\|\mathbf{x}_i \quad \mathbf{x}_j\|^2)$
Sigmoid 核	$K(\mathbf{x}_i,\mathbf{x}_j)=\tanh(\mathbf{x}_i^T\mathbf{x}_j+b)$

综合前述方法，可得到的非线性不可分支持向量机，也称为一次软间隔支持向量机（L1 Soft-Margin Support Vector Machine,简称 L1-SVM）的目标函数如下：

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}+\mathrm{C}\sum_{i=1}^{N}\xi_i \\ &s.t. \quad \mathrm{y}_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1-\xi_i \qquad \xi_i\geq 0 \quad \text{for } i=1,...,N \end{aligned} \tag{12}$$

其中，C 为惩罚系数（或正则化系数），起平衡模型复杂度和损失误差的作用。结合 Lagrange 方法和对偶原理，给目标函数可转化为：

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)-\sum_{i=1}^{N}\alpha_i \\ &\text{s.t.} \quad \sum_{i=1}^{N}y_i\alpha_i=0, \quad \mathrm{C}\geq\alpha_i\geq 0 \quad \text{for } i=1,...,N \end{aligned} \tag{13}$$

其对应的最优超平面 $y(\mathbf{x})=\sum_{i=1}^{N}\alpha_i y_i K(\mathbf{x},\mathbf{x}_j)+b$，决策函数 $g(\mathbf{x})=sign(y(\mathbf{x}))$。

求解该对偶问题后，得到的对应于 $\alpha_i^* \neq 0(i=1,...,N)$ 的训练点为支持向量。进一步结合 KKT 条件分析可发现，对应于 $\alpha_i^* = C$ 的样本点 $\mathbf{x}_i$ 会因为 $\xi_i \geq 0$ 而位于图 3-3 的间隔区域（H_1和H_2之间）之内。虽然因为其 Lagrange 乘子值大于零而参与了分界面的描述，但实际上却可能包含一些被错误划分的数据（当 $\xi_i \geq 1$ 时），因此被称为限定支持向量（Bounded Support Vector，简称 BSV）。另外一些对应于 $C > \alpha_i > 0$ 的训练样本则因为对应的 $\xi_i = 0$ 而刚好落于间隔区的边界上（如图 3-3 中的圆圈标识样本），因为它们代表刚好能被正确分类，因而被称为非限定性支持向量（Unbounded Support Vector）。

因为 **w** 向量中的每个元素$\mathbf{w}_k(k=1,\ldots,d)$都对应于训练集中 **x** 的一个特征维度，因此在寻找最小化‖**w**‖的同时，研究人员也希望能寻找到使得 **w** 最为稀疏的表达，以达到同时对数据进行降维（Dimension Reduction）的目的，如基于 p 模的支持向量机。

3.2.3 支持向量机处理多分类问题

SVM 本身是一个二值分类器，它最初是为二值分类问题设计的，所以当处理多类问题时，就需要构造合适的多类分类器。目前，构造 SVM 多类分类器的方法主要有两类：直接法、间接法。

（一）直接法

直接在目标函数上进行修改，将多个分类面的参数求解合并到一个最优化问题中，通过求解该最优化问题“一次性”实现多类分类。这种方法看似简单，但其计算复杂度比较高，实现起来比较困难，只适合用于小型问题中。结合本题数据情况，显然直接法对处理大量的留言数据是不可行的。

（二）间接法

主要是通过组合多个二分类器来实现多分类器的构造，常见的方法有 one-against-one 和 one-against-all 两种。

（1）一对多法（one-versus-rest，简称 OVR SVMs）

训练时依次把某个类别的样本归为一类，其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。分类时将未知样本分类为具有最大分类函数值的那类。

假如有四类要划分（也就是 4 个 Label），他们是 A、B、C、D。于是在抽取训练集的时候，分别抽取：

（i）A 所对应的向量作为正集，B，C，D 所对应的向量作为负集；

（ii）B 所对应的向量作为正集，A，C，D 所对应的向量作为负集；

（iii）C 所对应的向量作为正集，A，B，D 所对应的向量作为负集；

（iv）D 所对应的向量作为正集，A，B，C 所对应的向量作为负集。

使用这四个训练集分别进行训练，然后得到四个训练结果文件。在测试的时候，把对应的测试向量分别利用这四个训练结果文件进行测试。最后每个测试都有一个结果 f1(x),f2(x),f3(x),f4(x)。于是最终的结果便是这四个值中最大的一个作为分类结果。

基于“一对多”的方法又衍生出基于决策树的分类。这一种分类首先将所有类别分为两个类别，再将子类进一步划分为两个次级子类，如此循环下去，直到所有的节点都只包含一个单独的类别为止，此节点也是二叉树树种的叶子。该分类将原有的分类问题同样分解成了一系列的两类分类问题，其中两个子类间的分类函数采用 SVM。具体流程如图 3-4 所示。

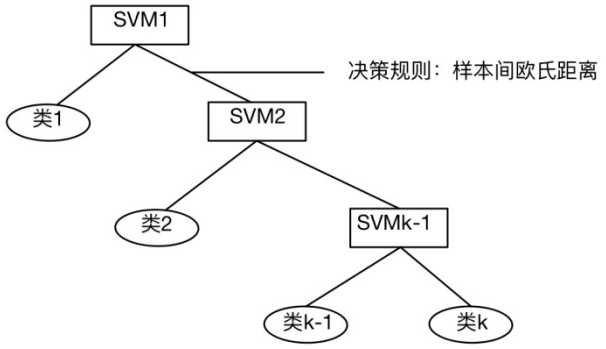


图 3-4 决策树的分类

（2）一对一法（one-versus-one,简称 OVO SVMs 或者 pairwise）

其做法是在任意两类样本之间设计一个 SVM，因此 k 个类别的样本就需要设计 k(k-1)/2 个 SVM。当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。Libsvm 中的多类分类就是根据这个方法实现的。

假设有四类 A,B,C,D 四类。在训练的时候选择 A,B;A,C;A,D;B,C;B,D;C,D 所对应的向量作为训练集，然后得到六个训练结果，在测试的时候，把对应的向量分别对六个结果进行测试，然后采取投票形式，最后得到一组结果。投票是这样的：

A=B=C=D=0;

(A,B)-classifier 如果是 A win,则 A=A+1;otherwise,B=B+1;

(A,C)-classifier 如果是 A win,则 A=A+1;otherwise,C=C+1;

...

(C,D)-classifier 如果是 A win,则 C=C+1;otherwise,D=D+1;

The decision is the Max(A,B,C,D)

基于“一对一”的方式出发，出现了有向无环图（Directed Acyclic Graph）的分类方法。具体流程如图 3-5 所示：

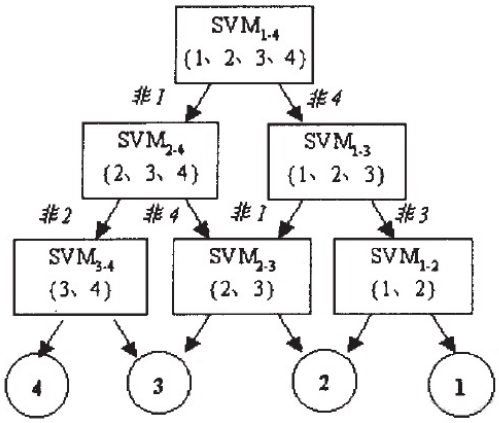


图 3-5 有向无环图的分类方法

综上所述，直接方法尽管看起来简洁，但是在最优化问题求解过程中的变量远远多于第一类方法，训练速度不及间接方法，而且在分类精度上也不占优。当训练样本数非常大时，这一问题更加突出。正因如此，间接方法更为常用，因而在本文中，对留言分类采用间接法。

3.3 分类模型评价

对于分类模型的评价，评估指标有很多，例如精确率（Precision）和召回率（Recall）等。但是对于本题，需要处理的数据量很大，这时，这两个指标往往是相互制约的。所以在本题中我们还需要综合权衡这两个指标，即采用F_1系数，它是综合考虑精确率和召回率的调和值：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}(\text{其中}P_i\text{为第}i\text{类的查准率},R_i\text{为第}i\text{类的查全率。}) \quad (1)$$

我们常常使用 Precision 和 Recall 去评价二分类的模型，但对于本题，我们需要对一个多分类模型进行评价，不妨将多分类问题转换为二分类模型来做。针对本题我们将引入如图 1 的一个 7 行 7 列的判断矩阵。在这个判断矩阵中，每一行之和表示该类别的真实样本数量，每一列之和表示被预测为该类别的样本数量。根据概念我们知道，在转换为二分类问题的过程中，我们重点在意的是怎样去做正样本，其实我们可以把每个类别单独视为“正”，所有其他 6 类型视为“负”那么对于这个判断矩阵我们就可以计算它的每一个类别的精确率P_i、召回率R_i和F_1系数。具体公式如下：

精确率（*Precision*）：$P=\frac{TP}{TP+FP}$ （2）

召回率（*Recall*）：$R=\frac{TP}{TP+FN}$ （3）

通过程序实现，本文模型在全部数据上的测试分类精确率高达 90%，各分类 F1 系数和召回率均在 0.8 以上，评价结果及其判断矩阵如图 3-6 和图 3-7 所示：

accuracy 0.900609756097561

	precision	recall	f1-score	support
城乡建设	0.86	0.80	0.83	394
环境保护	0.92	0.94	0.93	540
交通运输	0.94	0.95	0.94	485
教育文体	0.91	0.86	0.88	293
劳动和社会保障	0.88	0.86	0.87	360
商贸旅游	0.88	0.91	0.90	699
卫生计生	0.91	0.92	0.92	509
accuracy			0.90	3280
macro avg	0.90	0.89	0.90	3280
weighted avg	0.90	0.90	0.90	3280

图 3-6 评价结果

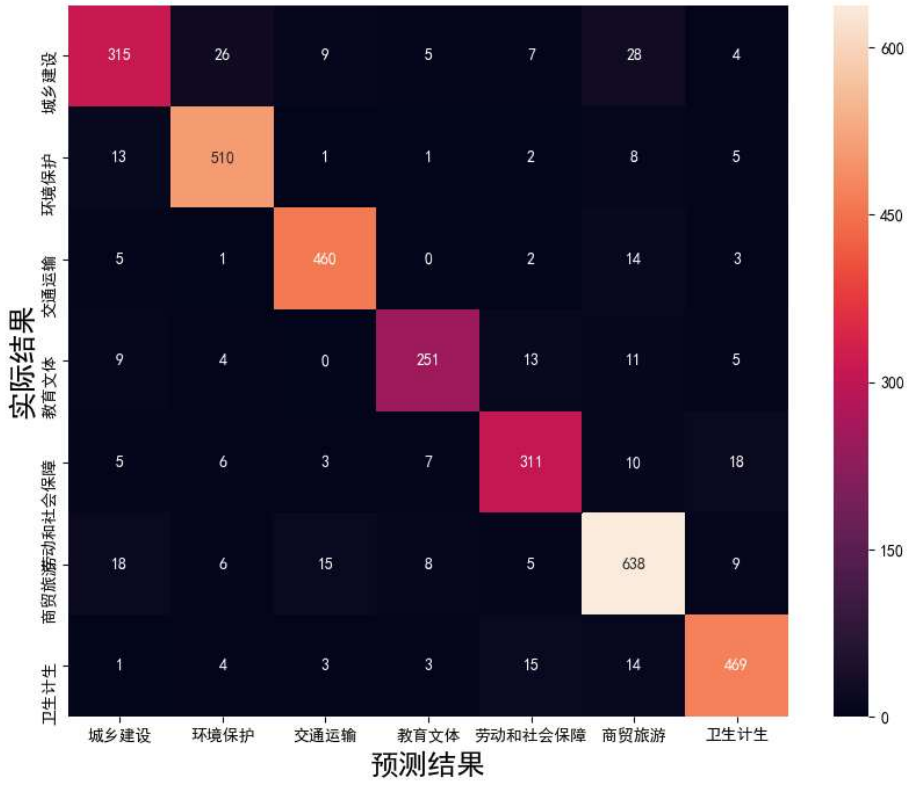


图 3-7 判断矩阵

4 问题 2 热点问题挖掘

4.1 文本聚类

4.1.1 文本聚类概述

文本聚类(text clustering,也称为文档聚类或 document clustering)指的是对文档进行的聚类分析，被广泛用于文本挖掘和信息检索领域。最初文本聚类仅用于文本归档，后来人们又探索出了它许多的新用途，例如生成同义词、文本去重等等。总之，文本聚类的用途十分广泛。

文本聚类的基本流程可以分为特征提取和向量聚类两大步。如果能将文档表示为向量，就可以对其运用聚类算法。这种表示过程称为特征提取。而一旦将文档表示为向量，剩下的算法就与文档无关了。这种抽象思维无论从哪个角度考虑都显得十分简洁有效。

基于上述分析，本文将根据附件 3 的留言信息，对留言主题进行问题一中的数据预处理，然后进行特征提取和向量聚类两大步。

4.1.2 文档的特征提取

文档是一系列单词的有序不定长列表，这些单词的种类数无穷大，且可能反复出现。单词本身已然千变万化，它们的不定组合更加无穷无尽。从细节上完全地表示一篇文档并不现实，我们必须采用一些有损的模型。有关文本表示模型在前文已有说明，这里不再赘述。针对问题 2，本文将采用词袋和层次聚类模型进行解决。

4.1.2.1 频数加权的词袋向量

词袋(bag-of-words)是信息检索与自然语言处理过程中最常用的文档表示模型，它将文档想象成一个装有词语的袋子，通过袋子中每种词语的计数等统计量将文档表示为向量。由于词袋模型不考虑词序，词袋模型的计算成本非常低。也正是这个原因,词袋模型损失了词序中蕴含的语义。比如,对于词袋模型来说，“人吃鱼”和“鱼吃人”的词袋向量是一模一样的。这听上去很荒谬，但在实际工程中，词袋模型依然是一个很难打败的基线模型。

当然，词袋模型并非只能选取词频作为统计指标，而是有很多选项。常见的

统计量还包括如下几个。

o 布尔词频：词频非零的话截取为 1，否则为 0。

o TF-IDF：参考本文问题一中的叙述。

o 词向量：如果词语本身也是某种向量的话，则可以将所有词语的词向量求和作为文档向量。得到词向量的途径有很多。

它们的效果与具体的数据集有关，需要通过实验验证。一般而言，词频向量适合主题较多的数据集；布尔词频适合长度较短的数据集；TF-IDF 适合主题较少的数据集；而词向量则适合处理 OOV 问题严重的数据集。

通过几种方式比较，本文最终以词频作为统计指标，用词袋模型来提取文档的词频特征向量，并通过对词频矩阵的每一列都乘以该列数值和的平方来扩大特征词权重。到此，特征提取介绍完毕。这样，留言主题根据词袋组成了同一维度的词向量，得到稀疏的词汇-文本词频矩阵，进一步为防止聚类效果产生偏差，对矩阵进行标准化处理后，即可用于聚类模型聚类。

4.1.3 聚类算法

从文本到向量的转换已经转换完毕。转换后我们得到了一系列文本向量，或者说是一系列数据点。接下来我们将使用聚类算法将这些数据点聚集成不同的簇。常用的聚类算法有：

1.基于划分：给定一个有 N 个元组或者纪录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，K<N。

特点：计算量大。很适合发现中小规模的数据库中小规模的数据库中的球状簇。

算法：K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法

2.基于层次：对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。

特点：较小的计算开销。然而这种技术不能更正错误的决定。

算法：BIRCH 算法、CURE 算法、CHAMELEON 算法

3.基于密度：只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。

特点：能克服基于距离的算法只能发现“类圆形”的聚类的缺点。

算法：DBSCAN 算法、OPTICS 算法、DENCLUE 算法

4.基于网格：将数据空间划分成为有限个单元（cell）的网格结构,所有的处理都是以单个的单元为对象的。

特点：处理速度很快，通常这是与目标数据库中记录的个数无关的，只与把数据空间分为多少个单元有关。

算法：STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法

针对问题 2 的留言文本聚类问题，本文尝试了基于划分的 K-MEANS 算法和层次聚类法将其划分为 k 类。

4.1.3.1 K-means 聚类

假设有一个包含 n 个 d 维数据点的数据集 $X=\{x_1,x_2,...,x_i,...,x_n\}$，其中 $x_i \in R^d$，K-means 聚类将数据集 X 组织为 K 个划分 $C=\{c_k,i=1,2,...,K\}$。每个划分代表一个类 c_k，每个类 c_k 有一个类别中心 μ_i。选取欧式距离作为相似性和距离判断准则，计算该类内各点到聚类中心 μ_i 的距离平方和

$$J_{(c_k)}=\sum_{x_i \in c_k} \| x_i-\mu_k \|^2 \quad (1)$$

聚类目标是使各类总的距离平方和 $J(C)=\sum_{k=1}^{K} J_{(c_k)}$ 最小，

$$J(C)=\sum_{k=1}^{K} J_{(c_k)}=\sum_{k=1}^{K}\sum_{x_i \in c_i} \| x_i-\mu_i \|^2=\sum_{k=1}^{K}\sum_{i=1}^{n} d_{ki} \| x_i-\mu_i \|^2 \quad (2)$$

其中，$d_{ki}=\begin{cases}1, 若 x_i \in c_i \\ 0, 若 x_i \notin c_i\end{cases}$，所以根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。

K-means 算法的具体步骤如下：

1.随机选取 K 个样本作为类中心;

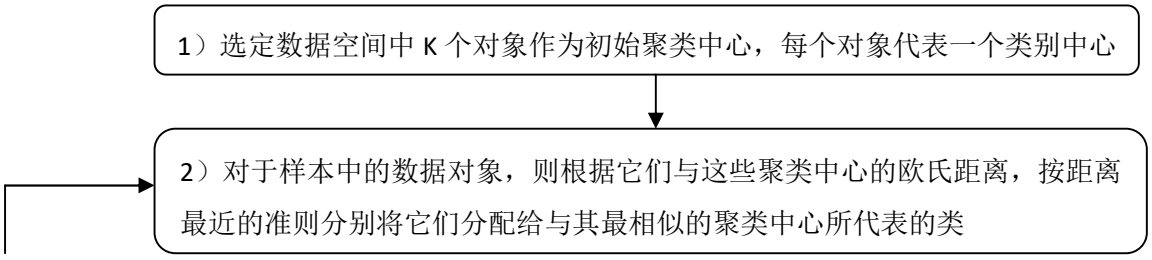
2.计算各样本与各类中心的距离;

3.将各样本归于最近的类中心;

4.求各类的样本均值，作为新的类中心;

5.判定：若类中心不再发生变动或达到迭代次数，算法结束，否则返回到第 2 步。

K-means 聚类的算法流程图如下：



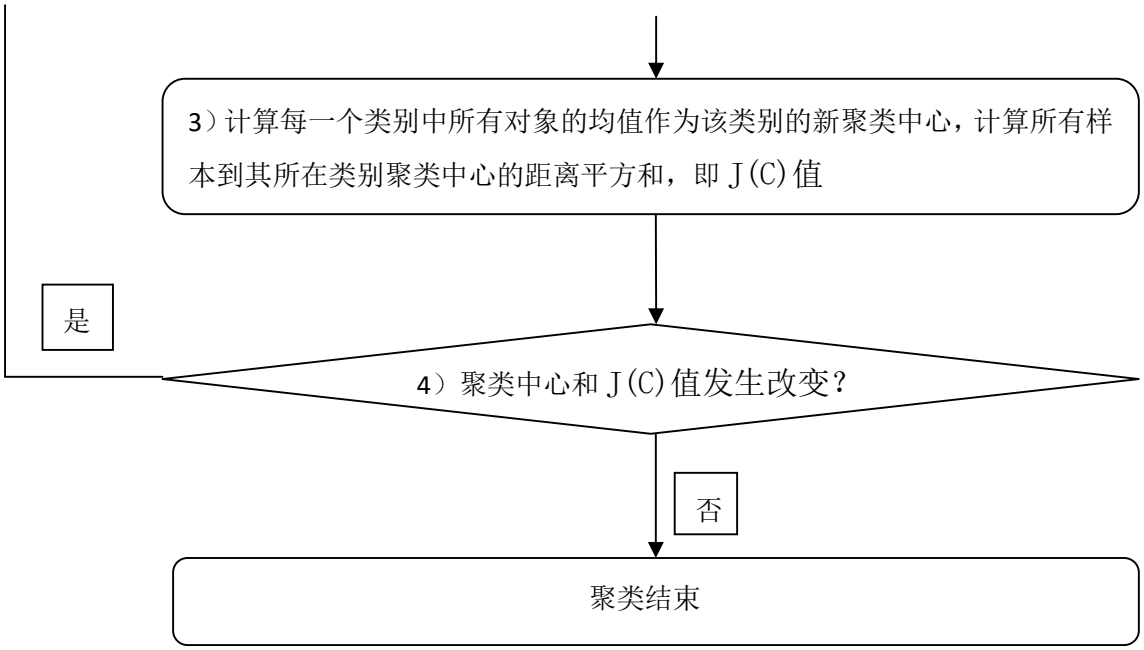


图 4-1 聚类算法流程图

K 值的确定：

在 K-means 算法中，我们常常采用探索法确定 k 值，即给定不同的 k 值，对比某些评估指标的变动情况，进而选择合理的 k 值。

4.1.3.2 层次聚类

层次聚类算法是比较常用的一种聚类方法，该算法实质上是产生一个聚类层次，也就是生成一棵层次聚类树，通过反复分裂或合并操作，得到满足一定要求的聚类结果。层次聚类方法可分为凝聚式和分裂式两种基本形式，其中凝聚式层次聚类算法是通过由下而上的策略，首先将每一个数据视为一个类别，直到所有的数据凝聚为一个类别或是满足一定条件的宗旨条件，聚类结束；分裂式层次聚类算法是通过由上而下的策略，首先将所有的数据看作一个类别，直到每个数据成为一个类别或是达到某些终止条件，聚类结束。它们的算法如图 4-2 所示：

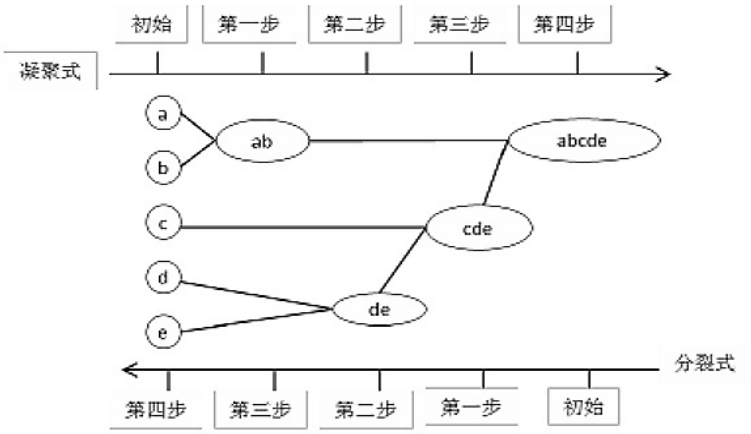


图 4-2 两种层次聚类算法

与分裂式层次聚类方法相比较，凝聚式层次聚类算法从一个层次到另一个层次所需要的计算方法比较简单快捷，且能够取得较好的聚类效果，因此本文主要研究凝聚式层次聚类算法。下面给出凝聚式聚类算法的具体步骤：

假设待聚类的样本数据大小为 N，

1.将每一个数据看作是一类，总共有 N 个类；

2.计算各类之间的距离，生成一个距离矩阵；

3.找到距离矩阵中的最小值，将对应的两个类合并为一类，这样就产生新的聚类；

4.重新计算各类之间的距离；

5.重复步骤 3 和 4，直到所有的数据都合并为一个类为止，此时这个类包含 N 个数据。

考虑到 K-means 算法对初始选取的质心点是敏感的，不同的随机种子点得到的聚类结果完全不同，对结果的影响很大，并且它对噪音和异常点比较的敏感。更值得注意的是它采用迭代方法，可能只能得到局部的最优解，而无法得到全局的最优解。相比于 K-means 算法，层次聚类的距离和规则的相似度容易定义，限制也比较少。并且层次聚类是可以通过设置不同的相关参数值，得到不同粒度上的多层次聚类结构；在聚类形状方面，层次聚类适用于任意形状的聚类，并且对样本的输入顺序是不敏感的。综合以上因素以及实际分类情况，本文最后选择了层次聚类对留言信息进行聚类操作。

4.1.3.3 距离度量

在前面我们通过构建“词袋”模型，将文本数据对象进行表示。若 $D=\{d_1,d_2,...d_n\}$ 表示文本集合，$W=\{w_1,w_2,...,w_m\}$ 表示出现在 D 中的所有特征单词的集合，则每一个文本可以用一个 m 维的向量来表示：

$$W_d = (wf(d_i, w_1), wf(d_i, w_2), ..., wf(d_i, w_m)) \quad (1)$$

其中 $wf(d,w)$ 为单词 $w \in W$ 在文本 $d \in D$ 中出现的频率。在词袋模型的基础上，对于两个文本数据对象 W_a, W_b 之间的相似性度量有很多种表达形式，本文在这里列举了几种最常见的，具体如下：

1.欧式距离（Euclidean Distance）：欧式距离是一种几何距离度量，是欧几里得空间中两点间“普通”（即直线）距离。

$$D_E(W_a, W_b) = (\sum_{i=1}^{m} | wf(a, w_i) - wf(b, w_i) |^2)^{\frac{1}{2}} \quad (2)$$

2.杰卡德系数（Jaccard index）：杰卡德系数的取值范围是[0,1],即当两个文本相同时，取值为 1；当两个文本完全不同是，取值为 0.

$$SIM_J(W_a, W_b) \frac{W_a \cdot W_b}{|W_a|^2 + |W_b|^2 - W_a \cdot W_b} \quad (3)$$

3.皮卡逊相关系数（Pearson Correlation Coefficient）：皮卡逊相关系数的取值为[-1,1]，当两个文本完全一样是取值为 1。

$$SIM_P(W_a, W_b) = \frac{m(W_a \cdot W_b) - \sum_{i=1}^{m} wf(a, w_i) \cdot \sum_{i=1}^{m} wf(b, w_i)}{\sqrt{\left(m|W_a|^2 - \left(\sum_{i=1}^{m} wf(a, w_i) \right)^2 \right) \left(m|W_b|^2 - \left(\sum_{i=1}^{m} wf(b, w_i) \right)^2 \right)}} \quad (4)$$

4.余弦相似度（Cosine Similarity）：余弦相似度的取值范围为[0,1]，它的特点是测量出的相似性度量与文本的长度无关。

$$SIM_C(W_a, W_b) = \frac{W_a \cdot W_b}{|W_a| \times |W_b|} \quad (5)$$

以上距离在机器学习中都可以用来计算相似程度。欧氏距离是最常见的距离度量，而余弦相似度则是最常见的相似度度量。很多其他的距离度量和相似度度量都是基于这两者的变形和衍生。本文采用余弦相似度计算样本之间的距离。

4.1.3.4 层次聚类连接方式选择

在进行聚类时，我们不仅要度量个体与个体之间的距离，还要度量类与类之间的距离。凝聚式层次聚类算法在类间距离度量上主要有三种方法：单连接法，完全连接法，平均连接法。下面对这三种方法做简单介绍：

1. 单连接法：用两个类中抽取的每对样本间的最小距离表示，故又称为最短距离法。对任意两个聚类c_i,c_j ，单连接法的计算公式为

$$dist(c_i,c_j)=\min\left\{dist(x_i,x_j)\middle|x_i\in c_i,x_j\in c_j\right\} \tag{1}$$

当c_i,c_j合并为一类时，则新类c_r 与其他类c_k 的距离为

$$\begin{aligned}&dist(c_r,c_k)=\\&\min\{\min\{dist(x_i,x_k)\mid x_i\in c_i,x_k\in c_k\},\min\{dist(x_j,x_k)\mid x_j\in c_j,x_k\in c_k\}\}=\\&\min\{dist(c_i,c_k),dist(c_j,c_k)\}\end{aligned} \tag{2}$$

2. 完全连接法：选取两类数据样本之间距离的最大值，也称为最大距离法。对任意两个聚类c_i,c_j，完全连接法的计算公式为

$$dist(c_i,c_j)=\max\left\{dist(x_i,x_j)\middle|x_i\in c_i,x_j\in c_j\right\} \tag{3}$$

当c_i,c_j合并为一类时，则新类c_r 与其他类c_k 的距离为

$$\begin{aligned}&dist(c_r,c_k)=\\&\max\{\max\{dist(x_i,x_k)\mid x_i\in c_i,x_k\in c_k\},\max\{dist(x_j,x_k)\mid x_j\in c_j,x_k\in c_k\}\}=\\&\max\{dist(c_i,c_k),dist(c_j,c_k)\}\end{aligned} \tag{4}$$

3. 平均连接法：两个类之间的距离选取类别间任意两个样本之间的平均距离。对任意两个聚类c_i,c_j，分别含有n_i,n_j个样本，则平均连接法的计算公式为

$$dist(c_i,c_j)=\frac{1}{n_in_j}\sum_{x_i\in c_i}\sum_{x_j\in c_j}dist(x_i,x_j) \tag{5}$$

当c_i,c_j合并为一类时，则新类c_r（含有n_r个样本数据）与其他类c_k（含有n_k个样本数据）的距离为

$$\begin{aligned}&dist(c_r,c_k)=\frac{1}{n_rn_k}\sum_{x_r\in c_r}\sum_{x_k\in c_k}dist(x_r,x_k)=\\&\frac{1}{n_rn_k}(\sum_{x_i\in c_i}\sum_{x_k\in c_k}dist(x_i,x_k)+\sum_{x_j\in c_j}\sum_{x_k\in c_k}dist(x_j,x_k))\end{aligned} \tag{6}$$

其中$n_r=n_i+n_k$。

经过简化得到

$$dist(c_r,c_k)=\frac{n_i}{n_i+n_j}dist(c_i,c_k)+\frac{n_j}{n_i+n_j}dist(c_j,c_k) \tag{7}$$

比较上述几种类间距离度量（基于完全连接法的聚类不适用于余弦距离度量），基于单连接的聚类效果通过关注局域连接，常常得到一些奇怪的类，正是由于其过于极端，所以效果也不是太好。但研究发现基于研究证明基于平均连接法的凝聚式层次聚类算法比较稳定，所以本文采用基于平均连接法的凝聚式层次聚类算法，对留言数据集进行聚类分析。

4.1.3.5 最佳聚类数分析

确定取多少类是聚类的关键，本文将留言数据分为 k 类，k 的确定将依据轮廓系数（Silhouette Coefficient）来确定。轮廓系数是聚类效果好坏的一种评价方式。最早由 PeterJ.Rousseeuw 在 1986 提出。它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。具体方法如下：

1.计算样本 i 到同簇其他样本的平均距离 a_i 。a_i 越小，说明样本 i 越应该被聚类到该簇。将 a_i 称为样本 i 的簇内不相似度。簇 C 中所有样本的 a_i 均值称为簇 C 的簇不相似度。

2.计算样本 i 到其他某簇 C_j 的所有样本的平均距离 b_{ij}，称为样本 i 与簇 C_j 的不相似度。定义为样本 i 的簇间不相似度：$b_i = \min\{b_{i1}, b_{i2}, ..., b_{ik}\}$，$b_i$ 越大，说明样本 i 越不属于其他簇。

3.根据样本 i 的簇内不相似度 a_i 和簇间不相似度 b_i，定义样本 i 的轮廓系数：

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{1}$$

4.判断：s_i 接近 1，则说明样本 i 聚类合理；s_i 接近-1，则说明样本 i 更应该分类到另外的簇；若 s_i 近似为 0，则说明样本 i 在两个簇的边界上。

所有样本的 s_i 的均值称为聚类结果的轮廓系数，是该聚类是否合理、有效的度量。本文通过绘制轮廓系数与簇的个数之间的折线图（图 4-3），确定了示例样本的最佳分类数 k 为 18。

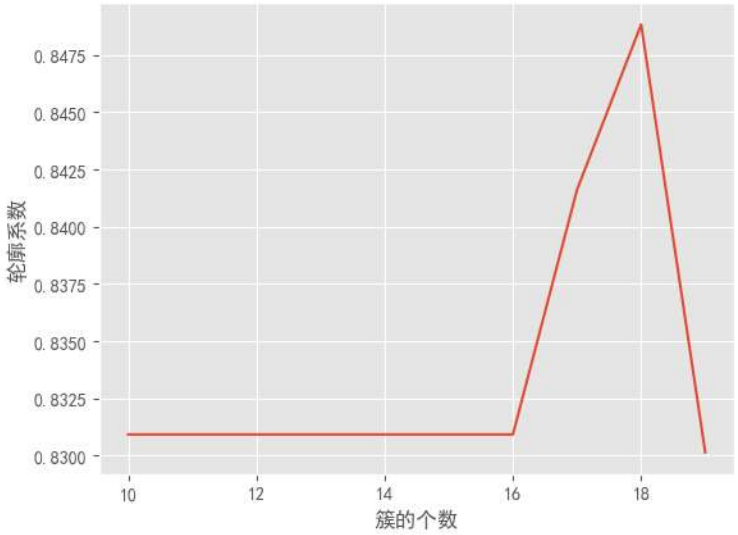


图 4-3 轮廓系数与簇的关系图

聚类完成后，将各个数据的类别标签赋值给问题 ID 列名称，并按照问题 ID 排序。

4.2 热点问题排序

4.2.1 定义热度评价指标

聚完类后，计算出 k 类留言中每一类留言问题的热度评价指数，赋值给热度指数列，并对热度指数排序，取排名前五的留言数据制成表 2-热点问题留言明细表。有关热度评价指标的定义详细情况如下。

考虑到留言问题的热度与留言问题提出的数目和用户的关注度有关，现定义依赖于留言问题提出数目与用户关注度的热度评价指标。在留言问题中，留言问题的提出数目由用户的留言数目直接决定。而每类留言的点赞数与反对数均可反映该类问题的受关注程度，但点赞数目能够更大程度上反映对该类问题表示赞同的人群，因此放大点赞数的特征，即对每一类留言问题的点赞数之和做取平方操作，综合考虑后定义第 i 类留言问题的热度指数 h_i 为：

$$h_i = N_i + \frac{(\sum_{j=1}^{N_i} l_{ij})^2}{\sum_{j=1}^{N_i} (l_{ij} + d_{ij})} \tag{1}$$

其中 N_i 为第 i 类留言问题的留言总数，l_{ij} 为第 i 类留言问题中第 j 个留言的点

赞数，d_{ij}为第 i 类留言问题中第 j 个留言的反对数。

4.2.2 制作热点问题表

对每一个相似留言类，本文通过命名实体识别方法提取地名和人群信息，计算其在该类留言中的频次，取每一类别中频数最高的词作为地点或者人群。同时查找每一类所有留言时间中的最大值和最小值进行拼接作为时间范围。最后，本文将选取每一类留言问题中的一条主题，去除停用词后作为该类留言的问题描述。于是，将所得信息以及相应的留言描述作为汇报结果，按照热度排名顺序保存为表 1-热点问题表。

4.2.2.1 命名实体识别

命名实体识别（Named Entity Recognition，简称 NER）是指识别语言中人名、地名、组织机构名等命名实体。通常包括两部分：实体的边界识别、确定实体的类型（人名、地名、机构名或其他）。中文的命名实体识别与英文的命名实体识别相比，挑战更大，目前未解决的难题更多。但中文命名实体识别作为中文切分任务的延续，是中文信息处理领域的一个基础任务，被广泛且成功地应用于信息抽取、信息检索、信息推荐和机器翻译等任务中。

命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法等。

1.基于规则和词典的方法

基于规则的方法多采用语言学专家手工构造规则模板,选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词(如尾字)、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。基于规则和词典的方法是命名实体识别中最早使用的方法，一般而言，当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。基于规则的方法的另外一个缺点是代价太大，存在系统建设周期长、移植性差而且需要建立不同领域知识库作为辅助以提高系统识别能力等问题。

2.基于统计的方法

基于统计机器学习的方法主要包括：隐马尔可夫模型(Hidden Markov Moder,

HMM）、最大熵模型(Maxmium Entropy Moder,MEM)、支持向量机(Support Vector Machine,SVM)、条件随机场(Conditional Random Field,CRF)等。

四种学习方法的优缺点分析：

（1）最大熵模型结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，有时甚至导致训练代价难以承受，而且最大熵模型需要归一化计算，导致开销比较大；

（2）条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题；

（3）一般说来，最大熵和支持向量机在正确率上要比隐马尔可夫模型高一些，但是隐马尔可夫模型在训练和识别时的速度要快一些，主要是由于在利用 Viterbi 算法求解命名实体类别序列的效率较高；

（4）隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，如短文本命名实体识别。

基于统计的方法对特征选取的要求较高，需要从文本中选择对该项任务有影响的各种特征，并将这些特征加入到特征向量中。依据特定命名实体识别所面临的主要困难和所表现出的特性，考虑选择能有效反映该类实体特性的特征集合。主要做法是通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。有关特征可以分为具体的单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征等。基于统计的方法对语料库的依赖也比较大，而可以用来构建和评估命名实体识别系统的大规模通用语料库又比较少，这是此种方法的又一大制约。

3.基于混合的方法

以上方法都是基于单一模型，这些模型往往由于自身的缺点不能达到很好的效果，因而有很多研究者提出将多个模型或多个方法进行整合，从而提升模型的性能。

本文利用 HanLP 工具来实现对留言数据中的地名及人群的识别。HanLP 是一系列模型与算法组成的 NLP 工具包，目标是普及自然语言处理在生产环境中的应用。HanLP 具有功能完善、性能高效、构架清晰、语料时新、可自定义的特点。它主要功能包括分词、词性标注、关键词提取、自动摘要、依存句法分析、命名实体识别等等。通过命名实体识别，提取出各类留言的地名与人群，计算其在该类留言中的频次，取每一类别中频数最高的词作为地点或者人群。相关词性标注如表 4-1 所示：

表 4-1 词性标注表

字母	描述
ni	机构相关（不是独立机构名）

nic	下属机构
nis	机构后缀
nit	教育相关机构
nnd	职业
nnt	职务职称
nt	机构团体名
ntc	公司名
ntcb	银行
ntcf	工厂
ntch	酒店宾馆
nth	医院
nto	政府机构
nts	中小学
ntu	大学
nx	字母专名
ns	地名
nsf	音译地名

5 问题 3 答复意见评价

“智慧政务”建设的目的，其一就是及时获取民情以助力政府精准把脉，其二是提供快捷途径帮助相关职能部门反馈信息给民众。本文将从相关性、完整性、可解释性和及时性这四个方面制定评价指标，对附件 4 中的答复意见做出评价。答复意见综合质量的评定指标将由上述四个指标加权求和得到。

5.1 答复意见相关性

显然，相关性是答复意见质量评价的根本。群众通过网络问政系统向政府反映社会现象、家庭困难等等。他们所反映的问题种类繁多，诉求不一。相关性和相似性虽有不同，但考虑到职能部门的反馈意见要尽可能的做到有针对的进行答复，不能答非所问，而且目前在文本挖掘领域中人们对相关性的研究还不够成熟。

故本文借助相似性的比较来完成相关性的判断。

5.1.1 TextRank 关键词提取

TextRank 算法可以用于提取文本关键词和生成摘要，其思想主要来源于PageRank。实际上，TextRank 就是 PageRank 在文本中的应用。

PageRank 是一种用于排序网页的随机算法，它的工作原理是将互联网看作有向图，互联网上的网页视为节点，节点V_i到节点V_j的超链接视作有向边，初始化时每个节点的权重$S(V_i)$都是 1，以迭代的方式更新每个节点的权重。每次迭代权重的更新表达式如下：

$$S(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{1}{|\mathrm{Out}(\mathrm{V}_j)|}S(V_j) \qquad (1)$$

其中 d 是一个介于（0,1）之间的常数因子，在 PageRank 中模拟用户点击链接从而跳出当前网站的概率，$In(V_i)$表示链接到V_i的节点集合，$\mathrm{Out}(\mathrm{V}_j)$表示从$V_j$出发链接到的节点集合。可见，并不是外链越多，网站的 PageRank 就越高。网站给别的网站做外链越多，每条外链的权重就越低。因为根据（1）中的分式$\frac{1}{|\mathrm{Out}(\mathrm{V}_j)|}S(V_j)$，外链权重跟外链总数成反比，与提供外链的网站权重成正比。如果一个网站的外链都是这种权重很低的外链，那么在迭代中它的 PageRank 会下降。同时，它给出去的外链权重也会降低，造成不良的连锁反应。正所谓物以类聚，与垃圾网站交换外链的往往也是垃圾网站。PageRank 公式恰好捕捉了这一点，因此能够比较公正地反映网站的排名。

将 PageRank 应用到关键词提取，就是将单词看作节点。另外，要注意的是，每个单词的外链来自自身前后固定大小的窗口内的所有单词。如图 5-1 所示例子：

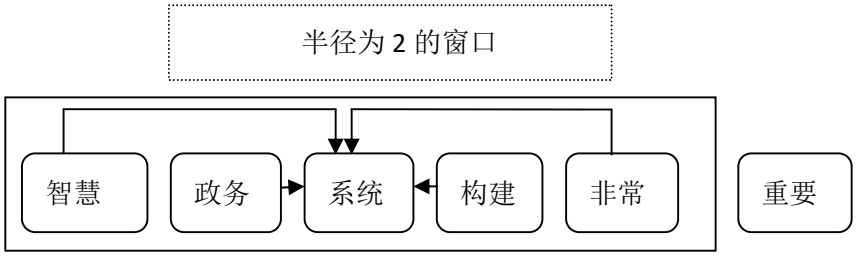


图 5-1 TextRank 中的窗口

在图 1 中，假设窗口半径为 2，对于单词“系统”而言，它的外链来自“智慧”“政务”“构建”“非常”这 4 个单词。同理，对其他每个单词都以它为中心建立窗口，让窗口内的每个单词链接到它。这样做的目的是模拟“解释说明”

这种语言现象，窗口内的词语常常用来解释中心词语，相当于为中心词语投了一票，每一票的权重等于窗口词语的权重被投出去的所有票平分。中心词这种左右搭配越多，给自己投票的词语就越多。另一方面，单词频次越高，给它投票的机会就越多，这一点与词频统计类似。然而在 TextRank 中，高频词不一定权重高，因为每一票还必须考虑投票者的权重。

同问题 2 中的命名实体识别一样，本文将利用 HanLP 工具包实现 TextRank 算法，完成对留言主题和答复意见关键词的提取。然后，将这些关键词进行比对，如果检索到一致的就计相关性为 1，否则，计相关性为 0。部分留言关键词提取情况如图 5-2 所示：

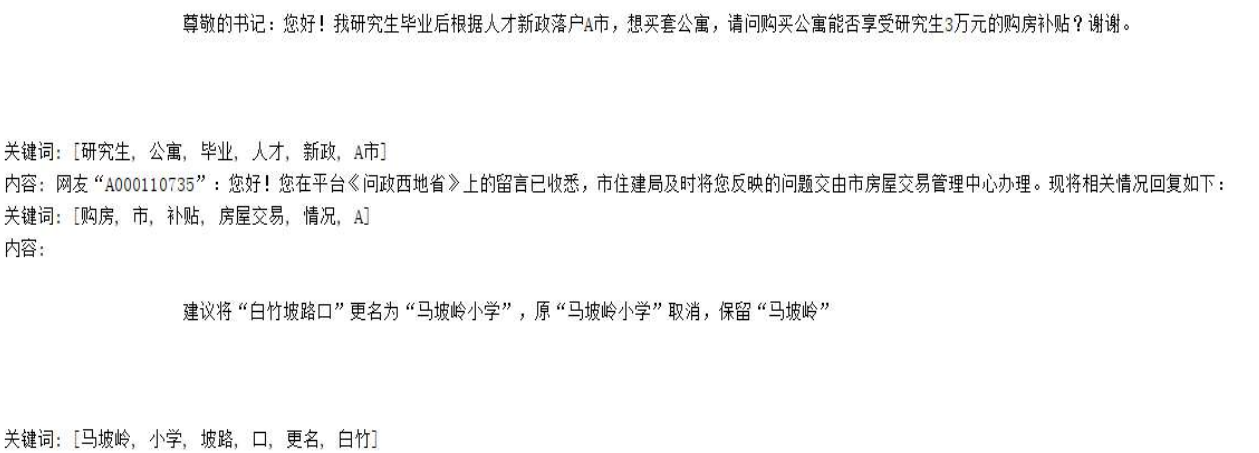


图 5-2 部分留言关键词提取情况

5.2 答复意见完整性及可解释性

答复意见的完整性和可解释性很大程度上决定了政府在民众心中是否有威望以及是否值得信任。下面，本文将分别介绍有关答复意见完整性和可解释性的评价指标构建方法。

针对完整性评价指标，我们先定义标准格式为：xx 用户：您好/你好!.....现回复如下：..... 谢谢您的关心与支持！随后，本文将基于这个标准格式对答复意见的完整性进行量化。具体步骤如下：

1.对某一个文本，定义一个数a，设置初始值为 0。这个数a代表该条答复意见的完整性度量；

2.在各个答复意见文本中，检索是否存在“：您好”或者“：你好”，如果返回值为 TRUE，则认为该条答复意见文本符合标准格式开头处的标准，并令$a=a+\frac{1}{3}$，同时将“：您/你好 “从文本中删除。若返回值为 false，答复意见文本就不做处

理。

3.对处理后的答复意见，寻找第一个以逗号或者句号开头并以冒号结尾的一句话，如果能提取出来并且其中有“复 ”字，则认为符合标准格式中“现答复如下：”的格式，并令$a=a+\frac{1}{3}$;

4.由于感谢一般都写在结尾，所以本文用正则表达式寻找包含“谢”字的句子，即满足：在“谢”字之前有0～8个中文字符，“谢”字之后0～8个中文字符的句子。如果能提取出来，则认为该答复意见符合标准格式结尾处的标准，并令$a=a+\frac{1}{3}$;

5.以最终的a值，作为该条答复意见的完整性数值；

6.单个答复意见文本的完整性数值在0～1之间。所有答复意见文本的完整性数值要对所有单个答复意见文本的完整性数值加权求和，这里权重取1：1：...:1（因为各个答复文本同等地位）。

针对可解释性评价指标，本文首先定义本例中的可解释性为答复意见中是否具有一定的理论支撑，例如针对留言用户的问题，答复内容是否能够根据具体的法律法规或者政府机关公文等等对问题的标准做出明确的界定，能否指明用户所咨询问题的法律法规出处。高的可解释性要求工作人员的答复做到专业化，而非仅仅摆出一定的事实依据或者处理措施。通过对可解释性的要求，我们希望以高度的可解释性保证用户对留言答复的确信与认同。对可解释性的度量主要依靠留言答复中是否有政府机关公文或者法律法规等来体现。同时，考虑到公共场合政府的留言答复对法规的描述虽然格式不一，但在公共场合下，政府的留言答复应具有最基本的规范，即对法规的描述至少应该具有书名号“《》”，而不是毫无格式可言或草草几个字描述。可解释性低的答复会让用户对政府工作人员的工作作风以及确信度产生怀疑。综上，本文对留言答复的可解释性度量主要依靠是否能在文中检索到以书名号包裹的法律法规条文。

通过对数据的简单筛选排查，发现留言答复中存在的符合基本格式的法律法规以及正式条文通知等可分为三大类，具体如下：

1.《某某法律法规或通知》…（…（ 年份四位数字 ）..号）

《某某法律法规或通知》…（…【年份四位数字】..号）

《某某法律法规或通知》…【年份四位数字】..号某某文件

《某某法律法规或通知》…（..年..月..日…）

对以上四种形式出现的法律法规通知条文可利用正则表达式对双字节字符的约束，先对留言答复文本做简单的处理，即将所有英文输入状态下的小括号中

括号以及大括号转换成中文输入状态的小括号中括号以及大括号，由于双字节字符中包括中文状态下的符号而不包括英文状态下的符号，因此在该操作后，可对上述四种情况进行统一正则表达式检索，即整合为第一大类情况：“《1-15 个中文汉字》0-15 个双字节字符 1-4 个数字”。

2.某某法律法规或通知《两个字母-四位数字-四位数字》

对以该种形式出现的法律法规或通知可通过正则表达式检索：“《两个英文字母-四个数字-四个数字》”。

3.根据/依据/按照/参照/由《某某法律法规或通知》

对该种形式的法律法规或通知可整合为第三大类情况，即：“据《1-15 个中文汉字》”或“照《1-15 个中文汉字》”或“由《1-15 个中文汉字》”，对该种形式的检索只需要用 Python 中的 in 即可。

具体实施步骤：

1. 定义一个数 A，设置初始值为 0。这个数 A 代表答复意见的可解释性度量;

2.采取投票机制，在所有留言答复文本中检索有无以上三大类共五种格式中的某一种，在一个留言答复文本中检索出第一个满足条件的法律法规条文，则令 A+1，同时在下一个留言答复文本中进行检索;

3.最终得到的 A 值除以总留言数即为最终的可解释性度量。

5.3 答复意见及时性

答复意见不仅要保证具有较高的相关性、完整性、可解释性，还必须要及时。民众往往希望自己的留言能够迅速得到回复，也正因为如此，政府能否及时回复留言用户在答复意见评价方案中也具有至关重要的作用。

针对答复意见的及时性，本文参考了各省规定，设置回复时间为 14 天的阈值，将它作为判断回复是否及时的标准。计算每一条答复意见对应时间与留言时间之差，若这个差大于 14，则认为及时性数值为 0，若这个差小于或等于 14，则认为及时性数值为 1。

5.4 综合评价指标

通过前文的探索，我们分别得到了相关性、完整性、可解释性、和及时性的四个评价方案。根据经验，相关性的权重应为最高，及时性其次，再次为可解释性，最后为完整性。所以，本文用对这四个评价指标归一化后按照 4：3：2：1 加权求和后除以权值总和的商作为该留言对应的答复意见的评价质量，根据质量数的范围区间，规定 0.8 及以上为‘高’，0.6 至 0.8 为‘中’，0.6 以下为‘差’，

确定综合质量指标。综上所述，我们对实例留言样本进行计算整理后，提取前六条留言及答复评价展示如图 5-3 所示：

留言时间	留言详情	答复意见	答复时间	相关性	完整性	可解释性	及时性	综合质量
2019/4/2	2019年4月	现将网友	2019/5/1(	1	1	1	0	高
2019/4/2	潇楚南路,	网友“A0(	2019/5/9	0	1	0	1	高
2019/4/2	地处省会	市民同志	2019/5/9	1	1	1	1	高
2019/4/2	尊敬的书	网友“A0(	2019/5/9	0	0.666667	1	1	高
2019/4/2	建议将“	网友“A0(	2019/5/9	1	1	0	0	高
########	欢迎领导	网友“A0(	2019/5/9	0	1	0	0	中

图 5-3 前六条答复评价情况

5.5 建议

政府优化自身服务功能的有效途径之一是推进“智慧政务”建设工作进程,这是促进当前经济社会长期、稳定、可持续发展的重要举措。目前,现代化的“智慧政务”建设工作还有一些留白,只有彻底解决这些问题,才能对政府治理能力现代化发展建设起到关键性影响。所以,在智慧政务建设过程中要注重信息技术应用、社会资源有效整合等。针对相关结果，以下是给相关政府职能部门“智慧政务”建设的一些建议：

（1）各级政府部门应积极参与“智慧政务”建设。当社会发展变快，但是相关的政府管理工作跟不上来,这会对社会发展产生恶性影响。从前,现代技术落后，政务问题大多都是通过人工来完成，不仅费力费时费事，而且也不一定能及时并准确解决政务问题。如今，现代网络技术日益提升，我们处于一个大数据时代，这给政府的管理和施政提供了很好的技术支持。作为各级政府部门应当合理运用网络技术，积极参与“智慧政务”建设，顺应创新发展新趋势。

（2）民众在网络上应合理反馈民生问题。网络反馈建议的方式给群众带来了极大的便利。但是我们不难发现有一些人正是因为有了这一方便和低成本，滥用这一途径，比如：一条建议反复刷；发的内容与主题不符；措辞低俗缺乏逻辑性等等。这些现象给政府部门的管理工作带来了极大的不便，也给绿色网络环境的创建带来了不好的影响。因而，广大群众应该认识到即使是在网络上发言也要经过自己的思考，对自己的留言建议要担负得起责任。

（3）政府职能部门要对网络上反馈的信息及时回复并且是有针对性地回复。“智慧政务”给政府部门的管理和施政带来了极大的便利和方便，大大的降低了政府部门的工作量。但是我们要强调的是应该有效运用节省的资源和时间，而不是把节省出来的时间和资源拿去消遣混日子。如何运用呢？显然，政府部门可以花更多的时间在反馈和施政上，让反馈越来越人性化，让施政越来越合理化。

6 结语

总结本次“智慧政务”中文本挖掘的相关任务，我们利用互联网公开来源的群众问政留言记录，及相关部门对群众留言的答复意见，在对文本进行基本的数据预处理后，通过建立多种数据挖掘模型，利用支持向量机模型实现了群众留言分类任务，通过层次聚类实现热点问题挖掘任务、构建四个角度的评价指标实现答复意见的评价任务。

回顾本次的文本挖掘过程，每一步都是通过建模、编程、模型评价、结果分析这几大步骤来实现。在模型的构建过程中，本文建立了多种模型进行对比分析。例如，对于问题 2 中的文本聚类任务，本文尝试了 K-means 聚类、层次聚类、LDA 主题聚类模型等等。在得到各种模型的结果后，本文综合题目要求和实际情况，筛选出最佳模型。在整个实验过程中，我们依然有很多需要改进的地方。例如，在问题 3 中对于答复意见的相关性评价指标构建，本文采取的方案是用相似度代替相关性，这种方法缺乏对语义内涵的考虑以及相关与相似之间的关系进行深入的研究。在后面的学习中，我们将针对这些还不够完善的地方继续进行探索和分析，以期得到较好的模型，完善解决文本挖掘任务。

参 考 文 献

[1]曹卫峰.中文分词关键技术研究[D].南京理工大学,2009.

[2]叶雪梅，毛雪岷，夏锦春，王波.文本分类 TF-IDF 算法的改进与研究[J].计算机工程与应用,2018.

[3]琚晓辉，徐凌.基于 SVM-Adaboost 裂缝图像分类方法研究[J].公路交通科技,2017.

[4]于瑞萍.中文文本分类相关算法的研究和实现[D].西北大学,2007.

[5]平源.基于支持向量机的聚类及文本分类研究[D].北京邮电大学,2012.

[6]陈珍,夏靖波,杨娟,韦泽鲲.基于 MapReduce 的支持向量机态势评估算法[J].计算机应用,2016.

[7]何晗.自然语言处理入门[M].人民邮电出版社,2019.

[8]徐衍鲁.基于改进的 K-means 和层次聚类方法的词袋模型研究[D].上海师范大学,2015.

[9]张跃，李葆青，胡玲，等.基于 K-means 文本聚类的研究[J].中国教育技术装备,2014.

[10]白雪.聚类分析中的相似性度量[J].北京交通大学,2012.

[11]江会星. 汉语命名实体识别研究[J].北京邮电大学,2012.
[12]孙镇，王慧临.命名实体识别研究进展综述[J].现代图书情报技术,2010.
[13]吴广财.HMM 增量学习算法在中文命名实体识别中的应用[D].华南理工大学,2011.
[14]项雪峰. 基于关键词相关度的计算机辅助定密技术研究[D].北京交通大学,2017.