基于智慧政务的自然语言处理和文本挖掘的应用

摘要：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。本文将利用自然语言和文本挖掘解决以下三个问题：第一：利用改进的增强型最大匹配分词法（IMM 法）对群众留言做分词、去停用词、降维等预处理工作，在消除歧义和保证系统的效率下能够比较好的实现留言数据分析功能；由于不同的特征项对应的文本的重要程度和区分度不同，我们利用改进的 KL-Divergence 特征选取算法和特征权重的计算 TF-IDF 法对特征项赋权值，解决模型的权重问题；接着与基于 KL 的特征选取算法结合在一起实现了群众留言的自动分类模型的构造。第二：采用 jieba 库和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；使用 TF-IDF 计算权重，按权值排列顺序抽取特征词，建立 VSM 向量空间模型；再使用关联规则实现特征词的挖掘，实现了群众留言热点问题的分析与挖掘。第三：利用模糊综合评价方法和 BP 神经网络结合起来实现对答复质量评估的模糊神经网络模型构建；在 BP 神经网络确定训练目标函数值上，采用模糊综合评价方法将数据进行模糊处理，然后结合 BP 神经网络及相应算法得出该模糊神经网络与 BP 神经网络各层的对应关系为:模糊化层与 BP 神经网络的输入层相对应,模糊推理层、隐含层对应 BP 神经网络，去模糊层对应 BP 神经网络的输出层；最终实现了相关部门对留言的答复意见评价方案。

关键词：大数据、自然语言处理、文本挖掘、IMM 法、改进的 KL-Divergence 特征选取算法、特征权重计算 TF-IDF 法、BP 神经网络

Application of natural language processing and text mining based on smart government

Abstract: in recent years, with wechat, Weibo, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government. This paper will use natural language and text mining to solve the following three problems: first, use the improved enhanced maximum matching segmentation (IMM) to do word segmentation, to stop using words, dimension reduction and other preprocessing work for the masses' message, which can better achieve the message data analysis function under the condition of eliminating ambiguity and ensuring the efficiency of the system; because of the important process of text corresponding to different feature items In order to solve the weight problem of the model, we use the improved KL diversity feature selection algorithm and TF-IDF method to calculate the weight value of the feature items, and then combine with KL based feature selection algorithm to realize the construction of the automatic classification model of the crowd message. Second, we use the Jieba library and dynamic programming to find the maximum probability path and find the maximum segmentation combination based on word frequency. We use TF-IDF to calculate the weight, extract the feature words according to the weight order, and establish the VSM vector space model. Then we use association rules to realize the mining of feature words, and realize the analysis and mining of the popular message hot issues. Third, the fuzzy neural network model of reply quality evaluation is constructed by combining the fuzzy comprehensive evaluation method and BP neural network. On the basis of BP neural network to determine the training objective function value, the fuzzy comprehensive evaluation method is used to fuzzy process the data, and then the corresponding relationship between the fuzzy neural network and each layer of BP neural network is obtained by combining the BP neural

network and the corresponding algorithm For: the fuzzy layer corresponds to the input layer of BP neural network, the fuzzy reasoning layer and the hidden layer correspond to BP neural network, and the de fuzzy layer corresponds to the output layer of BP neural network; finally, the evaluation scheme of reply opinions of relevant departments to the message is realized.

Key words: big data, natural language processing, text mining, IMM method, improved KL diversity feature selection algorithm, feature weight calculation TF-IDF method, BP neural network

目录

一. 挖掘目标

本次建模目标是利用互联网公开来源的群众问政留言记录，采用 jieba 中文分词工具、IMM 分词法对留言进行分词，以及基于 KL-Divergence 特征选取算法、BP 神经网络建模算法等算法进行文本挖掘分析，主要达到以下三个目标：

（1）利用 IMM 分词等方法对群众留言进行分类，解决人工工作效率低、工作量大、差错量高等问题。

（2）对群众反映强烈的热点问题进行文本挖掘，并给出热点问题表和热点问题留言明细表，有助于相关部门进行针对性处理，提升服务效率。

（3）针对相关部门对部分群众留言的答复意见的完整性、相关性、可解释性等进行评价方案的策划并实现[1]。

二. 分析方法与过程

2.1 问题 1 分析方法与过程

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，所以建立基于自然语言处理技术的群众留言多级标签分类模型至关重要。为了实现基于自然语言处理的群众留言的多级标签分类模型，我们需要在中文文本的预处理方面加紧改进，因为要涉及到预处理，就要在其中的最大匹配分词算法中做文章，传统的最大分词算法存在较多的问题，因此诞生了改进后的 IMM 法；除此之外，为了有效地实现文本的自动分类识别，我们就要对原始数据进行变换，从而获得更准确的特征，进行特征选择和提取，从而也要在分类的过程中，应合适的选择特征提取算法，必须对将要使用的算法的复杂度和特征提取效果进行综合考虑。最终通过对贝叶斯算法、简单向量距离分类法和 KNN（K 最近邻居）算法模型进行对比，找出哪种算法模型的分类效果是最佳的，并在实际耗时上综合考虑，选出一种实用性比较好的算法模型作为群众留言的分类模型。

2.1.1 群众留言自动分类的预处理

在群众留言的三级标签分类体系中，如果要进行训练分类，那么首先要进行以下工作：

(1)对每一条留言进行分词处理，并对数字（包括中文数字和阿拉伯数字）及未登录词进行识别。例如“A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。”这是一个连续的中文字符串，可用空格或其它符号（如“/”）分割成“A3 区/大道/西行/便道，未管所/路口/至/加油站/路段，人行道/包括/路灯杆，被/圈/西湖建筑/集团/燕子山/安置房/项目/施工/围墙内。每天/尤其/上下班/期间/这条/路上/人流/车流/极多，安全/隐患/非常/大。强烈/请求/文明/城市/A 市，尽快/整改/这个/极不/文明/的/路段。”类似这样的离散的中文词的形式；

(2)为了提高准确性，降低特征维数，从而就能进行更为有效的分类判断，但是这些操作之前必须将一些文档中的所有标点符号和一些没有实际意思的词，进行过滤处理，还有一些辅助作用的词，都可以过滤掉，例如“的，地，得，啊，么，呢，……”等。

基于自然语言处理的群众留言自动分类系统的预处理就是这个过程，大概就是做上面所述的工作。

2.1.1.1 基于最大匹配分词算法的群众留言分词处理

在群众留言的自动分类系统中，大部分都是选用一个字作为特征或者是一个词语作为特征，并且用它们来描述文本的信息，但是在汉语中，就需要把其分割成独立的单词流形式，这就是所谓的文本分词处理，所以分词的速度和分词的算法上准确率的要求就至关重要了。分词的方法还是有许多的，为了最大的体现其效率，不能采用过多的分词方法，也不能使用一些过于复杂的分词方法。在多种的分类算法中，由于一般普通的机械算法并不能直接地对语法语义进行正常的分析，所以只能依靠分词词典来帮助机械算法进行分析。这样一来就可以统计信息，进行分词，而且在速度上也提升了很多，并广泛应用于实际当中。所以本系统在分词模块主要是基于词典的机械匹配法来进行实现。目前为了适应科技的进步，为了取得最好的效果，最大匹配法是比较实用的，而且也是基于机械匹配的汉语

词语切分方法。它的主要基本原理就是首先要确定符号串，它是从文本信息中选取的词典，但是不能超过词典的最大长度，接着将符号串和词典中的单词进行比较，一旦匹配不成功那就要去掉一个汉字，然后再匹配，就这样循环操作，直到找到一个相应单词为止[1]。

首先假设最开始的最大匹配长度为 $MaxLen$，指向文本的指针为 fp，取出字符串的长度为 Len。因为中文的汉字在计算机里是占有两个字节，如果在匹配的过程中没有找到相应的词，这就应该减去一个汉字，也就是减两个字节，所以取出字符串的长度要减 2，表示为 $Len = Len - 2$。算法的主要流程如图 2-1 所示：

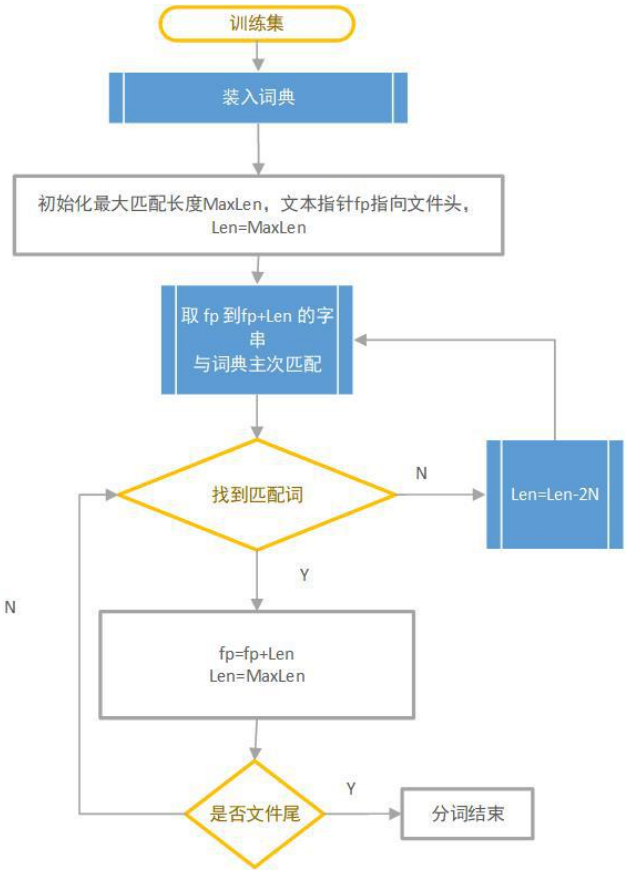


图 2-1 最大匹配分词算法流程图

最大匹配分词法整体思路还是比较清晰的，而且易于实现计算机对其进行预计分类，但也存在着一些不足，具有相当的局限性和主观性，主要表现在它试图采用一些稳定的词表来代替灵活多变的词汇，并把词表作为对判词的唯一标准。除此之外，它实际上也否定了组合递归性，所以在分词时多有歧义容易出错。

2.1.1.2 改进的增强型最大匹配分词法（IMM 法）

在实践中最大匹配法也存在较多的缺点，尤其在消除歧义这方面，在保证系统的效率下能够比较好的实现分析功能，所以还要增强分词功能，所以需要对最大匹配分词法进行改进使其成为一个增强型最大匹配分词方法（Improved Matching method），简称为 IMM 法，其实就是重点针对其特点三字长交集型歧义字段结合进行消除歧义处理及停用词处理。整个系统的分词的主要步骤总结为：对文本进行分词之前，首先对分词进行预处理，再通过使用一些书面符号，例如标点符号等对其处理的语料进行简单的一些处理，从而获得一系列的短语或者是单句，同时需要删掉那些标点符号，且识别阿拉伯数字和英文单词，其中需要把英文大写字母全部转化成对应的小写。最后结合预处理的方法以及通过中文词典和停用词表实现改进的最大匹配分词方法 IMM 法。主要步骤如图 2-2：

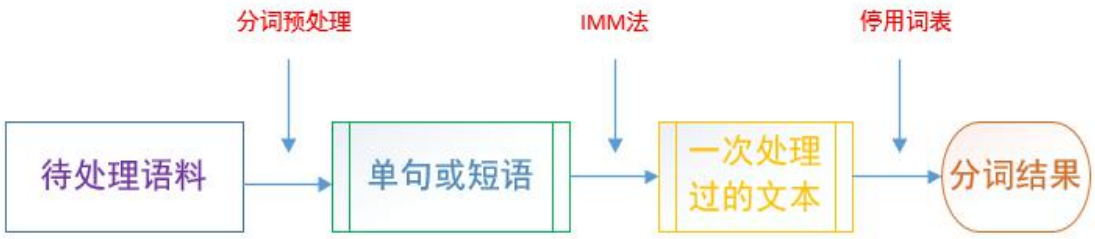


图 2-2 IMM 法分词主要步骤

在实现群众留言文本分词系统的方法中，可主要概括为两个部分，一个即预处理方法另一个就是增强型最大匹配分词 IMM 方法。

(1)分词预处理在处理中文信息的过程中，对于中文不同于英语那样，英语有着较为明显的分隔标志，而一篇普通的中文文本可以看成由连续的汉字字串组成，由此一来，计算机就难以识别出这些中文文本，所以首先需要对这个长的汉字字串进行适当的简化，并同时要保证它原本的真实意义没有发生改变。因为在文档信息中，有些标点符号和数据以及英语都是比较容易识别的，所以我们在做预处理的时候就可以利用到这些书面符号，把他们进行简单的处理，得到一系列短语或者就是单句，这样就可以减轻整个系统分词的压力。而且这些标点符号对文章的本身也没有太大的实际意义，所以可以同时去掉这些符号，除此之外对待英语上，虽然格式不同，大小写不同，但意义还是相同的，所以在预处理过程

中可以将英文文本全部转化成小写的。

(2)增强型最大匹配分词方法（Improved Matching Method）

最大分配分词方法的原理可总结为对一些文本信息中的字符串和词典中的词条进行比对的一个循环过程。这样虽然很容易出现歧义词的现象，但它的优点是时间短和比较简单，在此基础上，我们做了一些改进，也就是增强型最大匹配分词方法（IMM法）。而IMM分词法的核心思想概括为：预先设置$Smain[m,n]$和$Sbuff[i,j]$两个缓冲区，用这两个缓冲区存储字段的内部结构信息，然后再对文本进行分词。对于从句子中的第M字符开始收集N字符可记为$Smain[m,n]$，而$Sbuff[i,j]$表示为从句子中的第I字符开始收集J字符。我们从预处理后的文本中抽取一个句子作为实例，假设词典中最大词长为$MAXLEN$（通常$MAXLEN$值取4，因为组成汉语词的字符数大多小于等于4，N为读入的句子长度，接着对抽取的这个句子进行一次整体扫描，把该句子放入到缓冲区$Smain[1,N]$中储存，依据输入不同的N值，我们可以从两个方向来处理这种情况：假设当系统最大词长$MAXLEN$等于4时，在中文自动分词中，产生的分词歧义主要是三字长交集型分词歧义字段，因为对$N < MAXLEN$或者$N >= MAXLEN$这两种情况，就要分开来讨论：

当$N < MAXLEN$时，文本句子中最多只含有$MAXLEN-1$个字的字串，由于本系统的最大词长是设为4，所以该字串中最多包含3个字。如果这个字串不是词，那么则从第二个字开始选择判断$Sbuff[2,2]$是不是词，如果找到了能够相匹配的，则分词成功；否则就要按照最大匹配法，不用考虑字串最右边的一个字，判断字串的前两个字是否能成词，如果是，则分词成功；否则可以看作由一个字构成一个单字词，分词结束。

当$N >= MAXLEN$时，就从该句子的首个字开始截取一个长度为k（$k <= MAXLEN$）的字串保存在$Sbuff[1,k]$(即句子的开头k个字)，令$Sbuff[1,k]$同词表中的词条逐个依次匹配。如果在词表中找不到一个词条可以与$Sbuff[1,k]$匹配，就需要从句子第2个字开始截取一个长度为k的字串$Sbuff[2,k]$重复以上过程。如果还时没有找到，则依次从第3，4，…，$N-i$字开始分别依次循环进行匹配。假设所有的匹配都不成功，那么句子中没有长度为k的词，则长度k就要递减1，然后就不停的重复这个过程，直到整个句子都被切分掉；如果在某一次匹配中查到一个长度为k字词匹配成功，就把这个字串句子中切分出去，然后再搜索停用词表，如果有停用词，那么就直接删除，否则在写入新的文本时，就把原句中位于这个字串前后的部分视为两个新的句子，然后对每个句子再进行递归

调用这一过程。

2.1.2 基于 KL-Divergence 的特征选取算法

群众留言文本分类过程的一个重要环节是特征选取，它已经成为了一个研究热点。实际上它是文本样本和分类器中间的桥梁，主要是根据语料库的信息，去除掉文档中不能提供区别类别的词条，构造出一个新的特征集。为了能够有效地实现文本的自动分类识别，我们就要对原始数据进行变换，从而获得更准确的特征，进行特征选择和提取。但是关于特征选取技术，目前一直被认为是文本分类中的一项瓶颈技术。在特征选取的整个过程中，特征项都是在某一类的训练样本中出现的次数最多的信息代表该类，同样在整个过程中贡献肯定也是最大的，当然，假设特征项中所有类的训练样本中出 现的次数越多，它就越不具有代表性，在类别区分上的贡献就越小。

在分类的过程中，应合适的选择特征提取算法，必须对将要使用的算法的复杂度和特征提取效果进行综合考虑。这里实现了一种基于 KL 散度的特征提取算法而且结合了特征项的 TFIDF 权值，大大的提高了其效率。

2.1.2.1 改进 KL-Divergence 特征选取法

KL-divergence 的特征选取法都是主要通过“距离”这一测度进行特征选取，有点类似于期望交叉熵，基于 KL-divergence 的特征选取方法的思想是：这里表示两个概率分布的区别性，主要是从距离测量度的角度上，对文本的表示方法是采用向量空间模型，用一些特征项所构成的向量来表示文本，也可以认为这个概率分布是由特征词组成的，首先假设这两个要是已知的，也就是将要分类文本中所有的特征项在文本中的概率分布，还有在每个类中的概率分布，那么这个文本应该被分配到与概率分布最为相近的一个类中，从“距离”的角度上去看，就应该将文本分配到“距离”最小的这一个类中。所以 KL-divergence 的特征选取法就是文本类别的概率分布和在出现了特定的词的条件下文本类别的概率分布之间的距离，距离值越大，词与文本类别分布的影响也越大。根据 KL-Divergence，这两个分布之间的距离可以表示为：$KL(p(w|d),p(w|cj))$。

特征项 W 就是在基于 KL-divergence 的特征选取评估函数上能够选出的，能正确的表达上式的距离最短。具体表现在：对于给予较高的评估值的特征都是那些在同一类的训练文本集中具有相似分布的特征，而对那些较低的评估值都是在某一类的个别训练文本中的分布与其在类中的分布表现出特别大的差异的特征项，然后根据一定的阈值进行选取分值较高的特征项保留，从而可以形成所谓的特征空间。

对于文本中的每一个类而言，根据特征词的分布，通过词表而去掉那些分值较低的特征项，这样才能使类中的训练文本和该类以及这些训练文本之间都将会变得更加相似，使得每一个类的特性更加明显。假设训练文本和测试文本具有相同的分布，那么测试文本和其对应的类相似性就会加大，所以就可以获得一个比较精确的分类结果。

在本文中，针对 KL-divergence 我们做了一定的近似简化，主要是对文本集和文本所属类间的平均距离进行平均化了，得出的平均距离更加客观，切合实际，误差会相对比较小，下面是关于对其特征选取评估函数进行详细的解释。

训练文本集: $S = \{d1,d2,\ldots,ds\}$, 文本 di 的类别：用 $c(di)$表示。那么根据特征词 wt，训练文本集和各文本所属类之间的平均 KL-divergence 表示为：

$$KL_t(S) = \frac{1}{|S|}\sum_{d_i \in S} KL(p(\omega_t|d_i),p(\omega_t|c(d_i))) \tag{2.1}$$

关于 KL-Divergence 的特征选取法的测试指标主要从特征提取的分散度入手，所选取的特征都是一些能够在某一类中均匀分布出现的词。在现实中发现一旦结合文本特征权重，在从事进行特征提取前，首先必须要先剔除一部分权值极低的词，然后统计一下特征项的频度，分散度和集中度这三种因素，最后选取出的特征项就是更能代表类的特征。

2.1.2.2 特征权重的计算－TFIDF 法

对于整体的群众留言文本自动分类而言，不同的特征项对应的文本的重要程度和区分度也不同，所以进行处理的时候，需要对其特征项进行赋权。当前文档的相对词频和在训练文档库中的词频共同决定了特征项的权重。对于开根号权重和布尔权重计算等方法而言，都存在较大的不足点，就是不能更好的体现特征项的区分性和其对文本的贡献。而目前由很多实验证明更简单直观、处理速度快的方法是 TFIDF 方法，并且它还是一种非常有效的赋权方法。

通常对于 TFIDF 方法我们只考虑两个因素：一个就是 TF 词语频率，就是词语在文档中出现的次数；另一个就是 IDF 词语倒排文档频率 IDF，它的意思大概是该词语在文档集合中分布情况的一种量化，通常的计算方法是 $log(N/n_k + 0.01)$,其中文档集合中的文档数目为 N，而出现该词语的文本数用n_k来表示。依据上面的两个因素，可以得出公式：

$$w_{ik} = tf_{ik} \times \log\left(\frac{N}{n_k} + 0.01\right) \tag{2.2}$$

因为要考虑到文本长度对权值的影响，所以要对项的权值公式做归一化的处理，将各项权值规范到[0,1]之间，公式如 2.3：

$$\omega_{ik} = \frac{\mathrm{t}f_{ik}\log(\mathrm{N}/n_k + 0.01)}{\sqrt{\sum_{k=1}^{n} {tf_{ik}}^2 \cdot log^2(\mathrm{N}/n_k + 0.01)}} \tag{2.3}$$

除此之外，对于那些特征比较明显的文本类别，常常有少数项的出现频率大于其它项，那么就根据上述公式计算出权值，这个计算出的权值会很高，如果个别权值很高的话，那么在分类过程中就会抑制其它项对文本的影响。所以通过统计出的词频做适当的均衡处理，才能再进行计算各项权重，比较简单的方法是对统计出的权值进行开平方。经过词频均衡处理的 TFIDF 权值计算公式为：

$$w_{ik} = \frac{\sqrt{\mathrm{t}f_{ik}\log(\mathrm{N}/n_k + 0.01)}}{\sum_{k=1}^{n} \mathrm{t}f_{ik}\log(\mathrm{N}/n_k + 0.01)} \tag{2.4}$$

这个公式是建立在一个假设，区分文档最有意义的特征词就是在那些指定文档中出现频率比较大，而且在其它文档中出现频率相对比较小的词。用 TFIDF 权重计算法可以体现特征项表达文本内容属性的能力，TF 越大，特征项出现的分布范围在文档集中越广泛，那么它的重要程度也越高；IDF 越大，当特征项在区分文档内容属性方面的能力于文档中的分布范围成正比,所以那些被赋予较高的权值的特征项只会在文档集合中较少的文档中出现[2]。

2.1.2.3 构建类模型

构建类模型的构建原理：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对留言进行分类后的留言集合文档，通过特征选取来建立类模型，这些都是自动分类的前期基础。在构造类模型的算法中,一般每个模型的向量表示包括这类的特征词和相对应的权值。

在本文的算法中，特征词提取就是将每一个文档类中的所有训练文档都归并为一个类文档来进行的。本文的算法在构建类的模型上注重从权值的角度考虑为了进一步提高模型的代表性，而且是与基于 KL 的特征选取算法结合在一起来实现类模型的建造。并且对于每一类而言，主要的构建算法步骤如下：

输入：类中各训练文本；

输出：类对应的类模型 ClsModel；

数据结构：向量 docvec 和 clsvec，分别存放类信息和单文档信息，定义如下：

```
typedef vector<TermStruct>docvec;
typedef vector<TermStruct>clsvec;
```

假设类模型 ClsModel 为空，这样一来每一类的 clsvec 也为空。向量 docvec 和 clsvec 中每个元素为一结构体变量 TermStruct，定义为：

```
struct TermStruct
{
int wordid;//在特征词典中的词号
CString word;//特征词
int wordfreq;//单词频数，初始值为 0；
int docfreq;//文档频数，初始值为 0；
double weight;//特征权重，初始值为 0.0；
double dKL;//特征评估值，初始值为 0.0；
};
```

定义了两个向量用于存放中间计算的概率值：PcjVec 和 PwcjVec；

```
typedef vector<StructPcj>PcjVec;
typedef vector<StructPwcj>PwcjVec;
```

其中结构体变量定义为：

```
struct StructPcj
{
int classid;//类号
double dPcj;//每类的概率分布 p(cj)，初始值为 0.0；
```

```
};
struct StructPwcj
{
int wordid;//特征词
int wordfreq;//词频，初始值为 0;
double dPwcj;//该特征词在类 cj 中的概率分布 p(w|cj)，初始值为 0.0;
};
```

如果所有训练文本都已没有了还要被剔除的停用词，那么对类中的训练文本再次进行处理。因为系统程序量比较大，所用到的函数太多，所以这里简单的介绍下步骤，具体如下：

Step1：docvec.clear()；//清空向量 docvec，用来存放单文档信息；

Step2：使一篇文档按照顺序读入，再把所有特征项写入 docvec，然后就统计每个特征项的词频信息，最后令所有词的 docfreq=1，所有向量 docvec 中的特征项不允许重复；

Step3：从 docvec 中的顺序取出特征项，假如 clsvec=empty，那么就为空，所以令 clsvec=docvec；不然，如果要查找 clsvec：万一 clsvec 中找不到想要的某一特征项 wt，那么就将特征项 wt 写入 clsvec，而且使 clsvec 中 wt 的词频等于 docvec 中 wt 的词频，cv[wt].wordfreq=dv[wt].wordfreq；然后令 cv[wt].docfreq=1；不然，在 clsvec 中找到特定的某一特征项 wt ，然后将 cv[wt].wordfreq+=dv[wt].wordfreq ，cv[wt].docfreq+=1；

Step4：判断一个样本是否被全部读入训练，假设没有读完，那么就转如 Step 1 继续执行；否则，就转入 Step 5；

Step5：首先对于已知的特征频数和文档频数值，按照权值计算公式，再根据基于 KL-divergence 的特征提取公式，通过这个公式来计算 clsvec 中每一个特征项 wt 的权值及特征评估值 dKL，再分别存放于 clsvec[wt].weight 和 clsvec[wt].dKL 中，然后再把中间结果 P(cj)及 p(w|cj)的值分别保存在向量 PcjVec 和 PwcjVec 中；

Step6：按照从大到小顺序对这些特征项权值进行排序，同时删掉小于阈值的特征项；然后对其它的特征项进行判断：如果已知特征阈值大于这个特征项的 dKL 值，就可以删除；否则就应保留该特征项，保留下来的特征项和权值一起加入到类模型 ClsModel 中；

Step7：然后把输出 ClsModel 存入到相应的 txt 格式的文件中；

对于阈值的确定一般采用预定初始值的方法，这里当然也是采用这种方法。例如，特征项权重的阈值初始值设为 0.25，dKL 的初始阈值设为 0.15，然后再对阈值进行调整，这个调整的过程中是根据分类测试的精度。

2.1.3 实验模型测试

2.1.3.1 实验难点

从上面的介绍中可以了解到在当今自然语言分类中阈值的确定是一个比较棘手的问题，单从理论上来说，并没有很好的解决方法，一般都是工作人员先按照一定的划分体系对留言文本进行分类，然后根据分类的准确程度对初始值进行调整，往往这样就存在着两个缺点，首先，采用人工预定的初始值，完全是人工凭借自己的经验假设确定的，并不是很科学的，其次，初始值设定好，需要调节的时候，调节的幅度无法确定，调整的过高过低都没有依据，只能反复测试和调整，这样增加了额外的时耗。而且，由于测试样本不同，任意一个分类系统中的阈值也不会相同，也不能够用于到其他的分类系统中。

通过长期的研究总结出来一个方法来确定阈值，那就是用百分比来确定，首先根据训练算法和分类算法构造相应的分类器，接着对那些类别事先确定好阈值，使用分类器对所有的样本进行分类，使得每个文本都可以获得一个相关的值，然后根据样本训练值按递减顺序排列起来，假设该类中有 n 篇文本，则这些文本的值分别表示$d_1,d_2,\ldots,d_n$，那么该类阈值 y 确定方法如下：$y = d_{sn\%}$

其中，s 假设为初始值，根据训练文本的质量程度，预先设定了本类的初始阈值，可以确定为 85 或者更高，s 越小，查全率就越低，准确率就越高，相反地，s 越大时，此分类器的查全率越高，准确度越低，我们可以通过实际测试对其进行进一步的调整。

相应地，s 值的调整可以通过调整阈值来转化，当对查全率不满意或者对准确率不满意时，则可以增加 s 值，否则就减少 s 值。

2.1.3.2 实验中用到的算法模型

(1) 贝叶斯算法模型

贝叶斯算法的基本思路即计算文本有多大的概率属于某一类别，实际上就是计算每个词属于这个类别的概率，简单的说就是计算文本属于这个类别的概率，具体如下：

第一步：计算所有类别中所有特征词的概率，$(\omega_1,\omega_2,\omega_3,...,\omega_n)$；

其中，$\omega_k = P(W_k|C_j) = \frac{1+\sum_{i=1}^{|D|} N(W_k,d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(W_s,d_i)}$与该公式相同的是对互信息量进行计算的公式。

第二步：当接收到新文本时，经常通过一些特征词分词，按照公式计算该文本d_i属于类别C_j的概率：

$$P(C_j|d_i;\widehat{\theta}) = \frac{P(C_j|\widehat{\theta})\prod_{k=1}^{n} P(W_k|C_j;\widehat{\theta})^{N(W_k,d_i)}}{\sum_{r=1}^{|C|} P(C_r|\widehat{\theta})\prod_{k=1}^{n} P(W_k|C_r;\widehat{\theta})^{N(W_k,d_i)}} \tag{2.5}$$

其中，$P(C_j|\widehat{\theta}) = \frac{C_j\text{训练文档数}}{\text{总训练文档数}}$，$P(C_r|\widehat{\theta})$用来表示相似含义，$|C|$用来表示类的总数，$N(W_k,d_i)$表示 W_k 在 d_i 中的词频，N 用来表示特征词总数。

第三步：比较所有类包含新文本的几率，然后根据比较后的结果将文本分配到概率最大的那个类别中。

(2)KNN（K 最近邻居）算法模型

这种算法的主要思路是：当接收到新文本后，首先比较新文本与训练集距离相近的 K 个文本，根据这 K 篇文本的类别决定了新的文本所属的类别，其具体的算法步骤归纳如下：

第一步：根据特征项集合重新描述形成训练文本向量。

第二步：在接收到新文本的时候，通过特征词分词来确定新文本的向量。

第三步：在训练集中寻找与新文本最为相似的文本，在比较的过程中选取出其中最相似的 K 个文本，计算公式为：

$$Sim(d_i,d_j) = \frac{\sum_{k=1}^{M} W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}} \tag{2.6}$$

其中，K 值还是很难确定，基本都是人工首先确定一个值，然后根据实验测试的结果调整 K 值，一般人工假设的数值都在几百或几千之间。

第四步：依次在新文本的 K 个邻居中计算每类的权重，计算公式如下：

$$p(\vec{x},C_j)=\sum_{\vec{d}_i\in KNN} Sim(\vec{x},\vec{y}_i)y(\vec{d}_i,C_j) \tag{2.7}$$

其中，$Sim(\vec{x},\vec{d}_i)$为相似度计算公式，新文本的特征向量表示为$\vec{x}$，计算方法等同于上一个公式，而 $y(\vec{d}_i,C_j)$表示为类别属性函数，其含义是：当$\vec{d}_i$属于类 C_j 时，函数值为 1，否则为 0。

第五步：最后对类的权重进行比较，并将文本分配到权重最大的那个类别中。

(3)简单向量距离分类法模型

简单向量距离分类法的主要思想即可描述为对文本集的每个类分别生成一个代表类的中心向量，而这些都是根据算术平均来确定的，因此当接收到新文本时，先确定其向量，判断一个文本是否属于那个与其距离最近的类，主要是计算其向量与每一个类的中心向量的距离来确定的，其具体步骤如下：

第一步：首先对有训练文本向量简单的算术平均，计算出每类文本集的中心向量。

第二步：然后对新到来的文本进行分词，把新到来的文本表示为特征向量。

第三步：计算每类中心向量间和新文本特征向量的相似度，公式为：

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M} W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}} \tag{2.8}$$

其中，第 j 类的中心向量是 d_j，新文本的特征向量是 d_i，向量的第 K 维是 W_k，特征向量的维数是 M。

第四步:比较新文本与每类中心向量的相似度，选出相似度最大的那个，然后把相应的文本分配到相应的类别中。

2.1.3.3 实验过程流程图

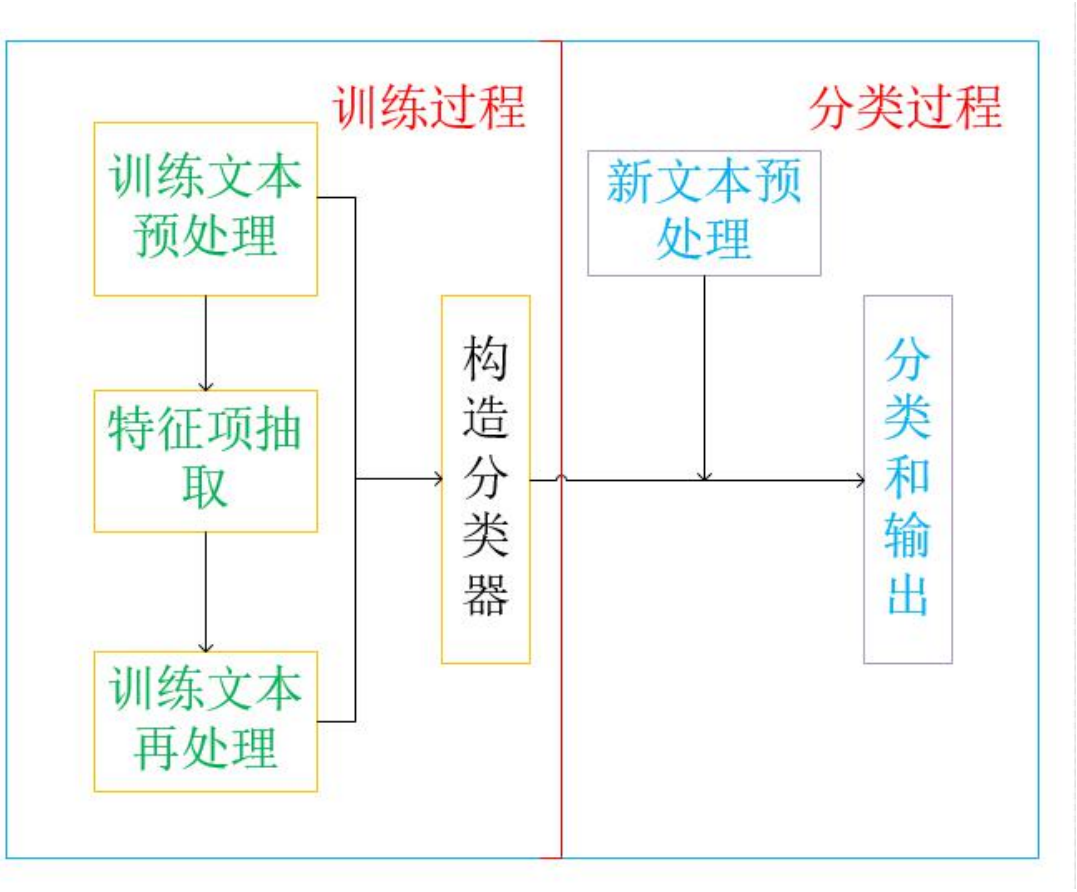


图 2-3 文本分类流程图

2.2 问题 2 分析方法与过程

过去旧中国的老百姓在衙门前击鼓鸣冤，这时的“鼓”的声音就成为了反映问题的渠道。而如今我们可以在政府门户网站上的留言版进行问题反馈，也有各种形式的便民平台。但平台多了、渠道多了，不可避免会产生信息多余冗杂、关键信息无法及时传达到位的问题。通过近年来互联网的实践来看，舆论关心的问题往往热度和浏览量也越大。对于公开性征求意见、每个人的意见都能被他人反馈的信息类型，“热度”具体地反映在浏览量、点赞数量、评论数量、转发数量等数据上。而对于政府与民众这样点对面的信息类型，每个人反映的问题中会存在一定的关键字、关键词，这些关键字词的重复次数就成为了我们需要进行比对的数据的重点。因此，对于“智慧政务”这样一个概念来说，如何“智慧”地发现反映人数较多、群众意见较大的“热点”问题，是实现“政务”的前提。而有了上述的几种数据，我们就可以从这些数据中挑选出“热点问题”，进而进一步地解决。及时发现热点问题，有助于

有关部门进行有针对性地处理，提升服务效率。该部分内容重点介绍热点问题发现系统的详细设计过程。首先从总体上对系统的流程图进行分析，然后对系统各个关键功能组做详细描述，根据特定时间、特定地点、以及发生的问题建立模型，再将问题归并，定义合理的热度评价指标，给出评价结果。

2.2.1 数据清洗

数据清洗(Data cleaning)——对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

数据清洗从名字上也看的出就是把“脏”的“洗掉”，指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。数据清洗是与问卷审核不同，录入后的数据清理一般是由计算机而不是人工完成。

一致性检查(consistency check)是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。例如，用 1-7 级量表测量的变量出现了 0 值，体重出现了负数，都应视为超出正常值域范围。SPSS、SAS、和 Excel 等计算机软件都能够根据定义的取值范围，自动识别每个超出范围的变量值。具有逻辑上不一致性的答案可能以多种形式出现：例如，许多调查对象说自己开车上班，又报告没有汽车；或者调查对象报告自己是某品牌的重度购买者和使用者，但同时又在熟悉程度量表上给了很低的分值。发现不一致时，要列出问卷序号、记录序号、变量名称、错误类别等，便于进一步核对和纠正。

由于调查、编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。常用的处理方法有：估算，整例删除，变量删除和成对删除。

估算(estimation)，最简单的办法就是用某个变量的样本均值、中位数或众数代替无效值

和缺失值。这种办法简单，但没有充分考虑数据中已有的信息，误差可能较大。另一种办法就是根据调查对象对其他问题的答案，通过变量之间的相关分析或逻辑推论进行估计。例如，某一产品的拥有情况可能与家庭收入有关，可以根据调查对象的家庭收入推算拥有这一产品的可能性。

整例删除(casewise deletion)是剔除含有缺失值的样本。由于很多问卷都可能存在缺失值，这种做法的结果可能导致有效样本量大大减少，无法充分利用已经收集到的数据。因此，只适合关键变量缺失，或者含有无效值或缺失值的样本比重很小的情况。

变量删除(variable deletion)。如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要，则可以考虑将该变量删除。这种做法减少了供分析用的变量数目，但没有改变样本量。

成对删除(pairwise deletion)是用一个特殊码(通常是 9、99、999 等)代表无效值和缺失值，同时保留数据集中的全部变量和样本。但是，在具体计算时只采用有完整答案的样本，因而不同的分析因涉及的变量不同，其有效样本量也会有所不同。这是一种保守的处理方法，最大限度地保留了数据集中的可用信息。

采用不同的处理方法可能对分析结果产生影响，尤其是当缺失值的出现并非随机且变量之间明显相关时。因此，在调查中应当尽量避免出现无效值和缺失值，保证数据的完整性。

对于数据值缺失的处理，通常使用的方法有下面几种：

（1）删除缺失值

当样本数很多的时候，并且出现缺失值的样本在整个的样本的比例相对较小，这种情况下，我们可以使用最简单有效的方法处理缺失值的情况。那就是将出现有缺失值的样本直接丢弃。这是一种很常用的策略。

（2）均值填补法

根据缺失值的属性相关系数最大的那个属性把数据分成几个组，然后分别计算每个组的均值，把这些均值放入到缺失的数值里面就可以了。

（3）热卡填补法

对于一个包含缺失值的变量，热卡填充法的做法是：在数据库中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。最常见的是使用相关系数矩阵来确定哪个变量（如变量 Y）与缺失值所在变量（如

变量 X）最相关。然后把所有变量按 Y 的取值大小进行排序。那么变量 X 的缺失值就可以用排在缺失值前的那个个案的数据来代替了。

还有类似于最近距离决定填补法、回归填补法、多重填补方法、K-最近邻法、有序最近邻法、基于贝叶斯的方法等。

2.2.2 分词和词性

文本处理过程中通常有一个先行环节——分词，即将文本切分为以词为单位组成的序列。一般认为，词是最小的、能够独立活动的、有意义的语言成分。在文本分析的过程中，如果没有分词环节，就难以在词汇基础上对文本进行句法以及语义分析。英文文本中的单词之间存在空格，因此文本分词时只需把单词的形态变化还原为原型即可。而中文与英文不同，中文里的词基本上没有形态变化，词汇之间也没有固定的分隔符，因此需要特殊的面向中文的文本分词方法来处理。中文分词，就是在没有固定分隔标志的中文文本中建立词汇之间的边界，将中文文本转换为符合语言实际的词汇序列。目前中文分词算法有很多，通常可以分为三种类型：基于统计的分词方法、基于字符串匹配的分词方法以及基于理解的分词方法[3]。

基于统计的分词方法认为词汇是字与字之间的一个稳定组合，如果相邻的字一起出现的频率越高，则它们组合成一个词的概率就越大。在基于统计的分词方法中，通常对这种相邻字的组合同时出现的频次进行统计，继而评估它们之间的紧密程度，如果达到预设的标准，则认为它们组成了一个词汇。由于基于统计的分词方法只需要统计相邻字组合的出现频度，不需要依赖分词词典，故而它又常被称作无字典分词方法。基于统计的分词方法常用的统计模型包括隐马尔科夫模型、最大熵模型以及 N 元文法模型等等。

基于字符串匹配的分词方法将待切分的中文字段与分词字典中的词汇以某种给定的策略进行字符匹配，如果能够匹配成功，则将该词汇切分出来。故而，这种分词方法也常被称为基于字典的分词方法。基于字符串匹配的分词方法有三个要素：分词词典、扫描顺序以及匹配原则。其中，扫描顺序包括正向（自左至右）、逆向（自右至左）、双向三种方式；匹配原则包括最大（长）匹配、最小（短）匹配、最优匹配等[4]。

基于理解的分词方法在对中文字段进行分词的过程中，同时还进行句法和语义分析，通过获取待切分字段中的句法、语义信息来解决歧义切分问题。这种方法让计算机来模拟人类对中文字段的理解，故而也被称为基于人工智能的分词方法。基于理解的分词方法依赖于大

量的语言知识，由于中文的特殊性，目前很难将汉语中的语言信息转换成计算机能够直接利用的形式，故而这种分词方法现在仍处于试验阶段。基于理解的分词方法主要有专家系统分词法以及神经网络分词法等。

经过一系列方法的分析，本系统使用基于词典的最大匹配算法对文本进行分词。在具体的分词模块使用了中文分词第三方库——jieba。Jieba 分词依靠中文词库，利用一个中文词库，确定汉字之间的关联概率，汉字间概率大的组成词组，形成分词结果。除了分词，用户还可以添加自定义的词组。jieba 分词的三种模式分别为精确模式、全模式、搜索引擎模式。精确模式就是把文本精确的切分开，不存在冗余单词，而全模式则是把文本中所有可能的词语都扫描出来，有冗余，搜索引擎模式是在精确模式基础上，对长词再次切分。基于前缀词典实现高效的词图扫描，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)；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；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。应用 jieba 分词器的测试代码如图 2-4 所示：

```
#encoding = utf-8
import jieba
seg_list = jieba.cut("A市经济学院体育学院变相强制实习",cut_all = True)
print("Full Mode:"+"/".join(seg_list))#全模式
seg_list = jieba.cut("A市经济学院体育学院变相强制实习",cut_all = False)
print("Default Mode:"+"/".join(seg_list))#精确模式
seg_list = jieba.cut("A5区劳动东路魅力之城小区临街门面烧烤夜宵摊")#默认为精确模式
print(",".join(seg_list))
seg_list = jieba.cut_for_search("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，急需处理！")#搜索引擎模式
print(",".join(seg_list))
```

图 2-4 Jieba 分词器测试代码

Jeiba 分词器测试结果如图 2-5 所示：

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\hp\AppData\Local\Temp\jieba.cache
Full Mode:A/市/经济/经济学/济学/学院/体育/学院/变相/强制/实习
Default Mode:A/市/经济/学院/体育/学院/变相/强制/实习
A5,区,劳动,东路,魅力,之,城,小区,临街,门面,烧烤,夜宵,摊
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,, ,急需,处理,急需处理,!
Loading model cost 0.739 seconds.
Prefix dict has been built successfully.
```

图 2-5 jieba 分词器测试结果

用户自定义词库的构建：针对各类社情民意相关的文本中出现的专用术语，添加了专业词库，存储与部分专业问题密切相关的词，并整理与热点主题相关的词保存到数据库中，将可定义的字典文件以 dictionary 作为拓展名保存到 jieba 的字典安装目录下；其次新建无用词库，专门存储那些已停用或与主题关联度不大的词条。

2.2.3 文体特征分析

一个容易想到的思路，就是找到出现次数最多的词。如果某个词很重要，它应该在这篇文章中多次出现。于是，我们进行"词频"（Term Frequency，缩写为 TF）统计。出现次数最多的词是"的"、"是"、"在"这一类最常用的词。它们叫做“停用词”（stop words）。停用词，这个概念最早由 Hans Peter Luhn 提出，是指在文本处理前后将被选择为自动过滤掉的字或词汇。在文本分类技术中，停用词通常指没有实际含义的虚词，主要包括了语气词、副词、拟声词、介词、连词、叹词等等。这些类型的词汇通常没有明确的词义，只有在完整的语境之中才具备一定的语法意义。一般认为，这些类型的词所附着的文本信息较少，对于文本分类的贡献不大，可选择将其从中文分词的结果中过滤掉。假设我们把它们都过滤掉了，只考虑剩下的有实际意义的词。这样又会遇到了另一个问题，我们可能发现"中国"、"蜜蜂"、"养殖"这三个词的出现次数一样多。显然不是这样。因为"中国"是很常见的词，相对而言，"蜜蜂"和"养殖"不那么常见。如果这三个词在一篇文章的出现次数一样多，有理由认为，"蜜蜂"和"养殖"的重要程度要大于"中国"，也就是说，在关键词排序上面，"蜜蜂"和"养殖"应该排在"中国"的前面。

所以，需要一个重要性调整系数，衡量一个词是不是常见词。如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关

键词。

用统计学语言表达，就是在词频的基础上，要对每个词分配一个"重要性"权重。最常见的词（"的"、"是"、"在"）给予最小的权重，较常见的词（"中国"）给予较小的权重，较少见的词（"蜜蜂"、"养殖"）给予较大的权重。这个权重叫做"逆文档频率"（Inverse Document Frequency，缩写为 IDF），它的大小与一个词的常见程度成反比。知道了"词频"（TF）和"逆文档频率"（IDF）以后，将这两个值相乘，就得到了一个词的 TF-IDF 值。某个词对文章的重要性越高，它的 TF-IDF 值就越大。所以，排在最前面的几个词，就是这篇文章的关键词。

第一步，计算词频：

$$词频(TF) = 某个词在文章中的出现次数 \tag{2.9}$$

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化：

$$词频(TF) = \frac{某个词在文章中的出现次数}{文章的总词数} \tag{2.10}$$

或者 $$词频(TF) = \frac{某个词在文章中的出现次数}{该文出现次数最多的词的次数} \tag{2.11}$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越*接近*0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对得到的值取对数。

第二步，计算逆文档频率。这时，需要一个语料库（corpus），用来模拟语言的使用环境。

$$逆文档频率(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数+1}\right) \tag{2.12}$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对得到的值取对数。

第三步，计算 TF-IDF：

$$TF - IDF = 词频(TF) \times 逆文档频率(IDF) \tag{2.13}$$

可以看到，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。所以，自动提取关键词的算法就很清楚了，就是计算出文档的每个词的 TF-IDF 值，然后按降序排列，取排在最前面的几个词。

就附件 3 中的留言主题而言，假定该文长度为 1000 个词，"小区"、"业主"、"物业"各出

现 20 次，则这三个词的"词频"（TF）都为 0.02，共有 4325 个留言主题。"小区"该词出现的频数为 4577，"业主"该词出现的频数为 4220，"物业"该词出现的频数为 2060。则它们的逆文档频率（IDF）和 TF-IDF 如表 2-1 所示：

表 2-1 文档词频统计

	该词出现的频数	IDK	TF-IDF
小区	4577	0.603	0.0121
业主	4220	2.713	0.0543
物业	2060	2.140	0.0482

从上表可见，"业主"的 TF-IDF 值最高，"物业"其次，"小区"最低。（如果还计算"的"字的 TF-IDF，那将是一个极其接近 0 的值。）所以，如果只选择一个词，"业主"就是关键词。除了自动提取关键词，TF-IDF 算法还可以用于许多别的地方。比如，信息检索时，对于每个文档，都可以分别计算一组搜索词（"小区"、"业主"、"物业"）的 TF-IDF，将它们相加，就可以得到整个文档的 TF-IDF。这个值最高的文档就是与搜索词最相关的文档。TF-IDF 算法的优点是简单快速，结果比较符合实际情况。缺点是，单纯以"词频"衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。而且，这种算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词，都被视为重要性相同，这是不正确的。（一种解决方法是，对全文的第一段和每一段的第一句话，给予较大的权重。）TF-IDF 算法可以用于无监督学习，不需要知道文档的类别，但是对同一个词来说，它在不同的文档中有不同的 TF-IDF 值，我这里处理的策略是每个主题取 top K，然后做一个去重。

2.2.4 构建模型

2.2.4.1 原理分析

通过对投诉文本对象的特点和项目需求的分析，得到系统工作流程图如图 2-6 所示：

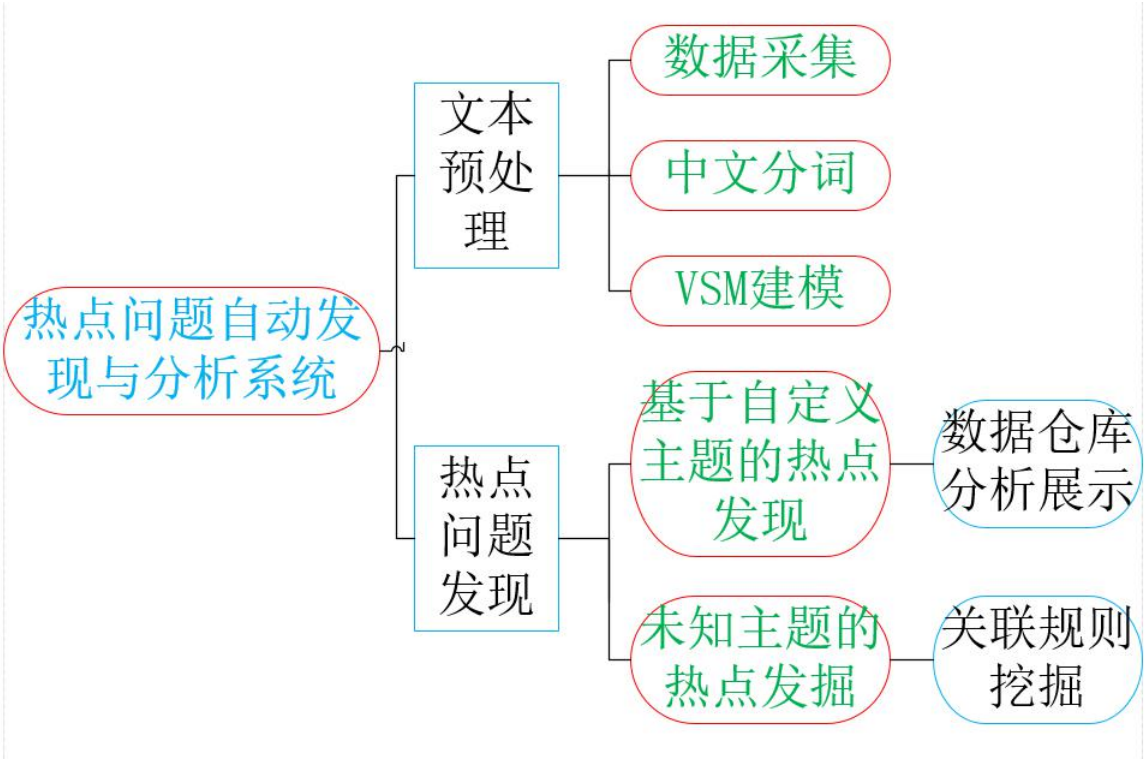


图 2-6 系统工作流程图

概括描述如下：从门户网站的留言板块获取一定格式的投诉源数据，对采集到的数据进行预处理，在过程中去除噪声文本，使用基于词典的分词方法对投诉文本进行分词处理，使用词频分析等方法计算权重，按权值排列顺序抽取特征词，建立向量空间模型，对于用户自定义主题的热点问题使用数据仓库技术进行分析展示。由于英文等拉丁语系语言以空格作为分隔符，所以不用进行分词处理。而中文的特殊之处在于词间没有分隔，因此分词就显得有必要了。而且，后续进行的文本向量化的前提也是要先对文本进行分词。其次，分词的效果直接影响着聚类分析的质量。经过上一章的分析，本系统使用基于词典的最大正向匹配算法对文本进行分词，在具体的分词模块使用了开源的中文分词器——jieba。它是一个基于 python 的中文分词系统，该分词器具有高效率和高扩展性，运用了面向对象的设计思想。

2.2.4.2 文本预处理技术

从字面意义上来看，文档包含词、短语、句子和段落等要素，在多数文本分类方法中，都将文本中出现的这些要素作为文本特征，而且随着要素级别的增高，其表达的语义越清晰，附带的信息也越丰富，但是特征组合的数目也会越大，因此，很少使用句子和段落作为特征。根据研究人员的实验，目前常见的特征项表示方法有：词、短语（Phrase）和 N-gram 项等。

词袋模型将一段文本作为一个个分离的词进行处理，通过不同类文本中可能出现词的差异对文本进行分类。必须指出的是对于文本分类，上下文对于其真正的类别有着强相关性。此类方法可能只是为了算法的简易性选择性地放弃了文本中的上下文信息，或者说只关注词频信息也能够获得符合要求的文本分类效果。词袋模型的三部曲包括分词（tokenizing），统计修订词特征值（counting）与标准化（normalizing）。在词袋模型统计词频的时候，可以使用 sklearn 中的 CountVectorizer 来完成。

2.2.5 问题归类

在进行中文分词后，关键字词的出现频次和数量就可以通过一定的方法轻松地统计出来。此处以 python 应用统计附件三中重点词语的出现频率为例，结果如图 2-7 所示。

```
#排除排前的不需要的词
excludes = {"2019","10","没有","没有","一个","我们","12","现在",
            "A7","11","A3","15","2018","是否","18","20","17","14"}
txt = open("C:/Users/hp/Desktop/附件3数据.txt", "r", encoding="utf-16LE").read()
words = jieba.lcut(txt)
counts = {}
for word in words:
    if len(word) == 1:#若为单个字的词，则过滤掉
        continue
    else:
        rword = word
    counts[rword] = counts.get(rword, 0) + 1
for word in excludes:
    del(counts[word])
items = list(counts.items())
items.sort(key=lambda x:x[1], reverse=True)
for i in range(5000):#输出拍前的关键字
    word, count = items[i]
    print("{0:<10}{1:>5}".format(word, count))#输出词频
```

	标签词	词频	词性
2	的	35222	其他
3	A	13132	名词
4	是	9995	动词
5	了	9037	其他
6	在	8122	其他
7	市	6874	名词
8	不	6782	副词
9	我	6131	代词
10	我们	5712	代词
11	有	5207	动词
12	年	5135	名词
13	和	4785	其他
14	小区	4577	名词
15	一	4329	其他
16	业主	4220	名词
17	也	3560	副词
18	月	3548	名词
19	区	3534	名词
20	就	3466	副词
21	没有	3291	动词
22	到	3236	动词
23	都	3087	副词
24	问题	2972	名词
25	3	2858	其他
26	为	2829	其他
27	多	2625	形容词

	A	B	C
34	个	2172	其他
35	7	2166	其他
36	对	2154	其他
37	领导	2140	名词
38	物业	2060	名词
39	与	2053	其他
40	县	1972	名词
41	这	1968	代词
42	要	1951	动词
43	4	1924	其他
44	日	1900	其他
45	给	1893	动词
46	能	1889	动词
47	公司	1883	名词
48	部门	1880	名词
49	严重	1875	形容词
50	相关	1862	动词
51	省	1734	名词
52	等	1727	动词
53	后	1727	其他
54	5	1703	其他
55	政府	1686	名词
56	居民	1686	名词
57	栋	1669	名词
58	大	1641	形容词
59	被	1640	其他

	A	B	C
67	将	1450	其他
68	进行	1447	动词
69	希望	1420	动词
70	房	1416	名词
71	会	1411	动词
72	上	1375	其他
73	可以	1373	动词
74	但	1367	其他
75	及	1345	其他
76	说	1343	动词
77	6	1334	其他
78	三	1328	其他
79	并	1323	其他
80	情况	1297	名词
81	影响	1291	动词
82	下	1272	其他
83	要求	1266	名词
84	元	1243	其他
85	未	1218	副词
86	已经	1200	副词
87	几	1184	其他
88	中	1180	其他
89	两	1150	其他
90	让	1145	动词
91	8	1139	其他
92	规划	1139	名词

图 2-7 词频统计示例

从例子中可以看出，仅需要简单的几行代码就可以将一个文本中出现最多的词频数统计出来，可见该例句中提到较多的就是在小区、物业、幼儿园等几个方面，与我们直观感受相一致。但这种方式仍然存在如重复统计、分词不当的情况，可以通过将特定地点或人群的数据相归并，将留言中出现的相似词作为同一问题归类等方法解决，需要进一步改进。同样地，

对于政府部门所接收的群众意见来说，很多问题存在一定的时效性，在一定的时间段后，一些词语的即时性及其反映热点的效用就不足，因此我们必须进行热度评价来提高词频统计效率。

2.2.6 热度指标定义

在进行热度指标定义时，可以考虑将留言版块作为公开形式，让居民对某一个已经有人反馈的问题进行反馈，类似于社交软件的“点赞”功能。热度指标考虑采用点赞数量结合地点的形式，对于某一块地区的全部问题按点赞数量进行排序。同时还要注意热度衰减问题，因为有可能存在问题已经解决但点赞数量仍然会增加的情况。所以在量化某个问题反馈的热度指标时，还应该与时间挂钩，并且结合上述停用词的方式，以“单位时间有效词增加量”来剔除一部分的事件噪音。此外，根据社交媒体的实践经验，往往存在同一个问题多人反馈的现象，因此还需要根据上述的中文分词方式进行关键字词的提取。对于单位时间内重复出现频率高的关键字词的数量与点赞数量相结合，就可以在四个维度上定义出一个“热点问题”，从而提高判断的准确性。

2.3 问题 3 分析方法与过程

随着科技的进步，网络问政平台逐渐成为群众行使监督权、知情权的主要方式。了解分析政府相关部门对群众反馈意见进行答复的质量好坏，不仅有利于群众反映社会问题，而且还对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题三，对答复意见的质量从可解释性、完整性、相关性给出一套评价方案，总体思路是运用模糊综合评价方法和 BP 神经网络结合起来，先进行数据进行模糊处理再根据建立的基于模糊神经网络的答复信息质量评估模型对质量进行指标性评估。

2.3.1 答复意见的评价指标

答复意见的评价是指对信息资源的内在质量进行判断和估算，是最根本、适用性最广的评估方式，主要包括可解释性、完整性、相关性、等三个指标（表 2-2）。

表 2-2 答复意见质量评价体系

目标	指标	指标解释
答复意见质量	可解释性	理解为什么某些事情发生的机制
	完整性	资源内容广度和深度
	相关性	反映了信息资源与用户需求的匹配程度

可解释性是关于理解为什么某些事情发生的机制。在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息，答复内容必须能够根据问题给出可以解释的回答。

完整性包括信息资源内容广度和深度两个层面。从广度上看，信息的非同一性决定了既定信息资源是由内容互不相同的信息单元组成的集合，要准确地表达一种思想或描述一个事物，不能缺少任何一个信息内容单元，否则意思将会不完整或者造成歧义；从深度上看，有价值的信息，无论其粒度粗细，都是对海量数据进行深度分析的结果，其隐含知识越多，价值越大，完整性越好。

相关性反映了信息资源与用户需求的匹配程度。强相关性意味着人们在需要时能够及时获得信息资源，而且它们与用户当前的工作任务或决策需求紧密相关

2.3.2 答复意见的评价方案

对于信息质量的评估方法有很多，随着信息技术的快速发展和信息领域研究的不断深入，对信息资源质量的研究开始出现信息资源的质量、语法和语用、传输和使用相统一的趋势。然而，目前还没有成熟的适用于所有信息资源评估的通用方法，已有方法分别适用于一种或少数几种信息资源。基于本题将模糊综合评价方法和 BP 神经网络结合起来实现对的答复质量评估的模糊神经网络模型构建。模糊综合评价方法将数据进行模糊处理,为 Bp 神经网络确定训练目标函数值;通过上述思想的描述结合 Bp 神经网络的网络结构(见图 2-8)和算法流程(见图 2-9)可得出该模糊神经网络与 BP 神经网络各层的对应关系为:模糊化层与 BP 神经网络的输入层相对应,模糊推理层、隐含层对应 BP 神经网络，去模糊层对应 BP 神经网络的输出层。

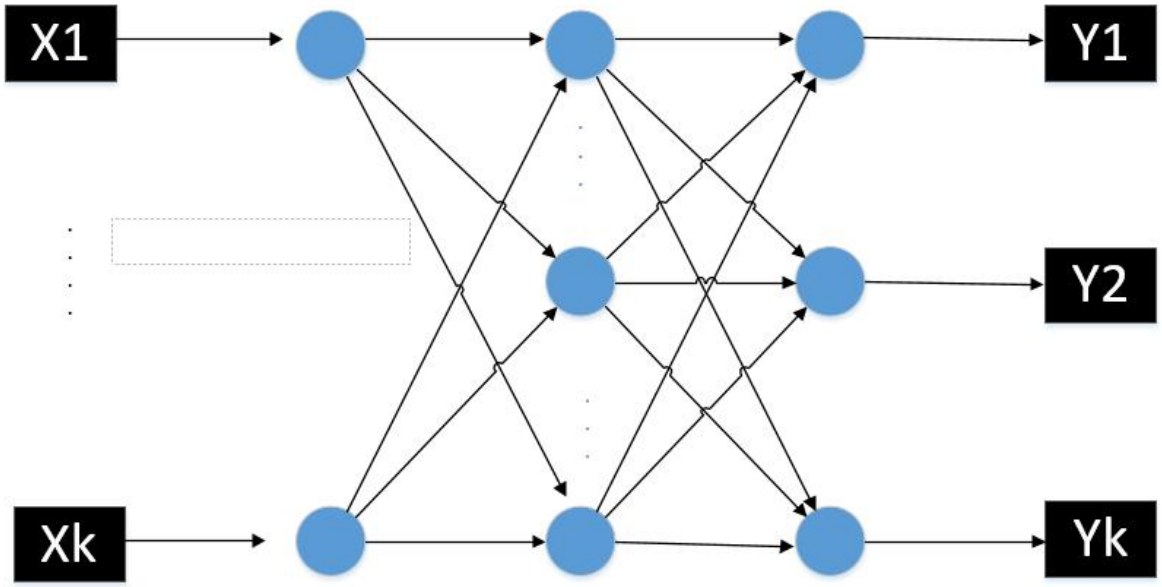


图 2-8 三层 BP 神经网络结构

图 2-9 BP 神经网络算法

基于前面对模糊神经网络及 BP 神经网络的描述，建立基于模糊神经网络的答复信息质

量评估模型，如图 2-10 所示：

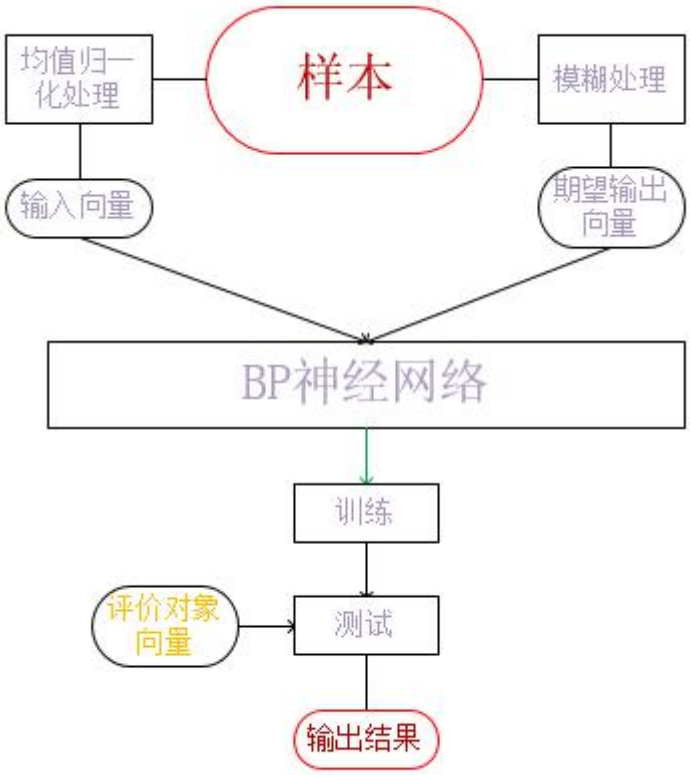


图 2-10 答复信息质量评估模型

根据评价指标设计针对答复意见质量的调查问卷，对与收到的有效数据进行整理，一方面,求出各指标的平均得分,然后对均值归一化处理作为评估模型的输入向量;另一方面,将搜集到的数据进行模糊化处理,求出模糊值作为社交媒体信息质量评价模型的期望输出向量。将输入数据和期望输出数据按要求输入到 B 神经网络中,设置好各项训练参数,模型通过图 4.3 的 B 算法进行自行的计算,最终达到预定目标时完成训练,整个评价模型构建完成。对于待评价的对象,只需将各指标的数据进行均值归一化,作为输入向量输入到训练好的模型即可得到评价结果。

2.3.3 数据模糊处理

在进行神经网络计算之前,要对所有数据进行模糊处理。如上节所述,本文主要采用模糊

综合评价的方法对数据进行模糊处理,以确定 BP 神经网络的训练目标输出向量。具体步骤如下:

（1）建立模糊集和评判集

分别构造社交媒体信息质量因素集$U=\{u_1,u_2\ldots u_m\}$和评判集$V=\{v_1,v_2\ldots v_m\}$.影响答复信息质量的因素集来源于信息质量指标体系中的一级是指:$U=\{u_1,u_2,u_3,u_4\}=\{$ 信息内容质量，信息表达质量，信息效用质量，信息来源质量 $\}$。在本文中u_1来源于指标体系中一级指标“信息内容质量”下的四个二级

标,$u_1=\{u_{11},u_{12}u_{13},u_{14}\}=\{$客观性，完整性,相关性,可解释性$\}$ u_2来源于指标体系中一级指标“信息表达质量”下的四个两个级指标,即$u_2=\{u_{21,}u_{22}\}=\{$易理解性，语言规 范性$\}$, u_3来源于指标体系中一级指标“信息效用质量”下的两个二级指标,即

$u_3=\{u_{31},u_{32}\}=\{$时效性，有用性$\}$, u_4来源于指标体系中一级指标“信息来源质量”下的四个二级指标,即$u_4=\{u_{41},u_{42},u_{43},u_{44}\}=\{$传播力，活跃度，权威性，信誉度$\}$。本文的评价集

$V=\{v_1,v_2\ldots v_m\}$分别对应优、良、中、及格、差。

（2）建立评估指标权重集——熵值法

熵值在信息论中代表了信息量大小,信息量越大,不确定性就越小,熵也就越小;反之,信息量越小,不确定性越大,熵也越大。根据熵的这一特性,我们用熵值来判断某项指标的离散程度,指标的离散程度越大,则该指标对综合评价的影响越大。所以我们用熵值法来对指标权重进行修正,从而使确定的权重更具有合理性。具体的步骤如下:

首先,对原始矩阵$A=(x_{ij})_{m\times n}$，将其进行归-化处理得到判断矩阵$R=(r_{ij})_{m\times n}$,在本文对微博信息质量进行评价的指标是属于大者为优,故归一化式如(2.9)$\max x_{ij}$及$\min x_{ij}$分别为不同评价对象对同一评价指标最满意的打分和最不满意打分。

$$r_{ij}=\left(x_{ij}-\min x_{ij}\right)/\left(\max x_{ij}-\min x_{ij}\right) \tag{2.14}$$

其次，计算各项指标下第 i 个评价指标值的比重：

$$f_{ij}=r_{ij}/\sum_{j=1}^{m}r_{ij} \tag{2.15}$$

再次计算第 i 项指标的熵值：

$$H_{\mathrm{i}} = -k\sum_{i=1}^{m} f_{ij} \ln f_{ij}.k = 1/\ln m \tag{2.16}$$

最后计算第 i 项指标是熵权值

$$\beta = (1 - H_i)/(n - \sum_{i}^{n} H_i) \tag{2.17}$$

求出综合隶属度向量

$$S_j = \sum_{i=1}^{n} r_{ij} w_j \tag{2.18}$$

（3）反模糊处理

为了能够更清楚的表示出评价对象的等级,需要对评价结果的隶属向量进行精确具体化,也可称之为反模糊化。本文采用常用的重心法对结果向量进行反模糊处理,5 个等级对应 5 个分值, $D=(d_1,d_2,d_3,d_4,d_5)=(1,2,3,4,5)$, 利用

$$L = S \cdot D^T \tag{2.19}$$

评价得出的具体数值则可以作为训绫模糊神经网络评价模型的期望输出值。

（4）BP 神经网络建模

<1>网络层数的确定

根据 Kolmogrov 理论,对于任意一个既定的函数(连续的),都可由三层神经网络实现,所以本文模糊神经网络评价模型的隐含层网络层数确定为 1 层。

<2>确定各层神经元的个数

输入层神经元个数是指作为输入指标的个数,根据对社交媒体整体影响因素的分析及归纳,确定本文输入神经元个数为 12 个。

通常情况下,隐含层神经元的个数是由研究对象的复杂程度及网络收敛效果来决定的。按照一般的经验判断,确定隐含层神经元个数一般用 $s=\sqrt{m+n}+a$, 其中,m 为输入神经元个数,n 为输出神经元个数,a 为 1-10 之间的常数,得到隐含层神经元个数为 7,这样按照经验公式计算出来的神经元个数只是估计值,是作为初始设置值,在网络训练过程中可以对其进行调整,在调整过程中按照误差尽可能小的原则。

我们对社交媒体信息质量进行评价,得到的结果作为神经网络的输出,则输出神经元的个数为 1,所以本文针对社交媒体信息质量的评价建立的神经网络模型的结构为 12×7×1。

<3>确定转换函数及训练函数

一般选择 Sigmoid 函数: $f(x)=\frac{1}{1+e^{-\lambda x}}$ 作为 Bp 神经网络模型各层神经元之间 I+e 的转换函数。各层网络之间采用动量梯度算法的训练函数,输入层和隐含层之间的激励函数采用 transie,而隐含层到输出层则设置为 purelin。

<4>期望输出向量的确定

本文采用模糊综合评价方法来计算出 Bp 神经网络的期望输出向量。通过调查问卷,得到关于社交媒体信息质量的评估指标体系标的 12 个二级指标评价得分,综合 AHP 法和熵值法确定社交媒体信息质量评估体系的权重,两者之积(式 4.5)就是所求的模糊综合评价值最后运用式 47 对评价结果的隶属向量进行反模糊处理,得到最终的模糊值。

<5>运用 BP 算法对神经网络进行训练

根据收集到的样本量,确定所建Bp神经网络的学习速率、最大训练步长日标误差等训练参数,将遴选出的样本数据进行两次处理,一个是求出 12 个指标的均值,神经网络要求输入的数据具有一致性,因此,我们要对原始数据进行归一化处理,可以使所有的分量在同一区间内变化,从而可以提高网络的效率。将上述表中的原始数据采用如下的公式：

$$T = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2.20}$$

归一化处理后作为训练样本的输入向量:另一个模糊综合的方法求出模糊评价值,反模糊处理后得到的值作为训练样本的期望输出值。运用 Matlab 软件对神经网络进行训练,检查网络实际输出与期望向量之间的误差是否在可允许范围内,如在该范围内则训练完成,可以通过该网络模型对目标对象进行评价[5]。

三. 结果分析

3.1 问题 1 结果分析

利用本赛题提供的测试数据来测试系统所实现的分类算法，通过实验对该算法的效率和结果进行比较分析。我们选取一级标签分类中的留言数据作为训练集和封闭测试集。将本赛题提供的测试数据为开放测试集，运行分类算法，计算出执行 20 次分类操作的平均值，其实验结果如表 3-1 所示：

表 3-1 三种算法实验结果对比

算法	封闭测试查全率	封闭测试准确率	封闭测试 F1 值	开放测试查全率	开放测试准确率	开放测试 F1 值
贝叶斯	83.28%	85.02%	83.06%	77.19%	78.28%	77.81%
KNN	90.11%	92.41%	91.24%	82.23%	82.23%	82.23%
简单向量距离	88.09%	88.09%	88.09%	84.19%	86.12%	86.30%

除此之外，根据许多学者已经对这三个算法进行了研究，对其时间复杂度也得出了结论，从各个算法的耗时上考虑，预先假设系统的训练文本集是由 m 篇文本（向量）组成的，同时这些训练文本集分别属于 K 类，从这些文本集当中抽取的特征项为 n 维，则该三种算法的时间花费见表 3-2：

表 3-2 三种算法时间花费

算法	训练算法	分类过程
贝叶斯	$O(mn)$	$O(kn)$
KNN	无	$O(km+nm)$
简单向量距离	$O(mn)$	$O(kn)$

因此，从上述两个表格所示的结果看来，KNN 算法在分类效果上更接近于真实值，而且在训练过程中时耗也最少，但是其在分类过程中耗时最多，这一点使总的耗时增加并不利于文本的实时处理，实用性大大降低；而简单向量距离算法和贝叶斯算法在耗时上近似，两者分类效果也近似，但从实验结果分析简单向量距离算法模型的效果略好。

3.2 问题 2 结果分析

根据关键词数量、时间跨度、点赞数量和地区类别四个维度，热点问题的前五名列表如表 3-3 所示：

表 3-3 排名前五的热点问题

热度排名	问题 ID	时间跨度	地点/人群	问题描述
1	A00077171	2019/8/19-9/17	A 市/小区	小区安全问题，物业是否需要提供保障?
2	A00087522	2019/1/10-6/30	A 市/小区	反映 A 市金毛湾配套入学的问题
3	A00031682	2019/4/15-7/8	A 市/停车场	请书记关注 A 市 A4 区 58 车贷案
4	A00056543	2019/9/6-11/5	A4 区/小区	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗?
5	A000106161	2019/8/6-11/9	A7 县/幼儿园	A7 县智慧桥幼儿园拒绝按照普惠性幼儿园标准收费

分别针对关键词的出现频率和时间跨度，我们可以列出一下两个表格。其中小区、物业、扰民、幼儿园、车位是出现频率前五的关键词，小区的出现频率最高。根据该关键词与其他四个关键词的联系，可以看出大部分的居民问题集中在小区管理上。而其中教育和生活又是居民关心的最多的两个点。针对物业的关键词有 1946 条，无疑是居民对小区物业管理能力的质疑，可见这方面需要加强改进。

针对反映问题的时间跨度，可以看出在 9 月份大家反映问题的热情相对较高。结合关键词的频率可以得知，9 月份的入学时期，教育问题是大家都较为关心的一个点热点。这与综合分析第二点的配套入学问题及第五点的幼儿园收费标准是吻合的，因此在教育方面应加强监管和整治。

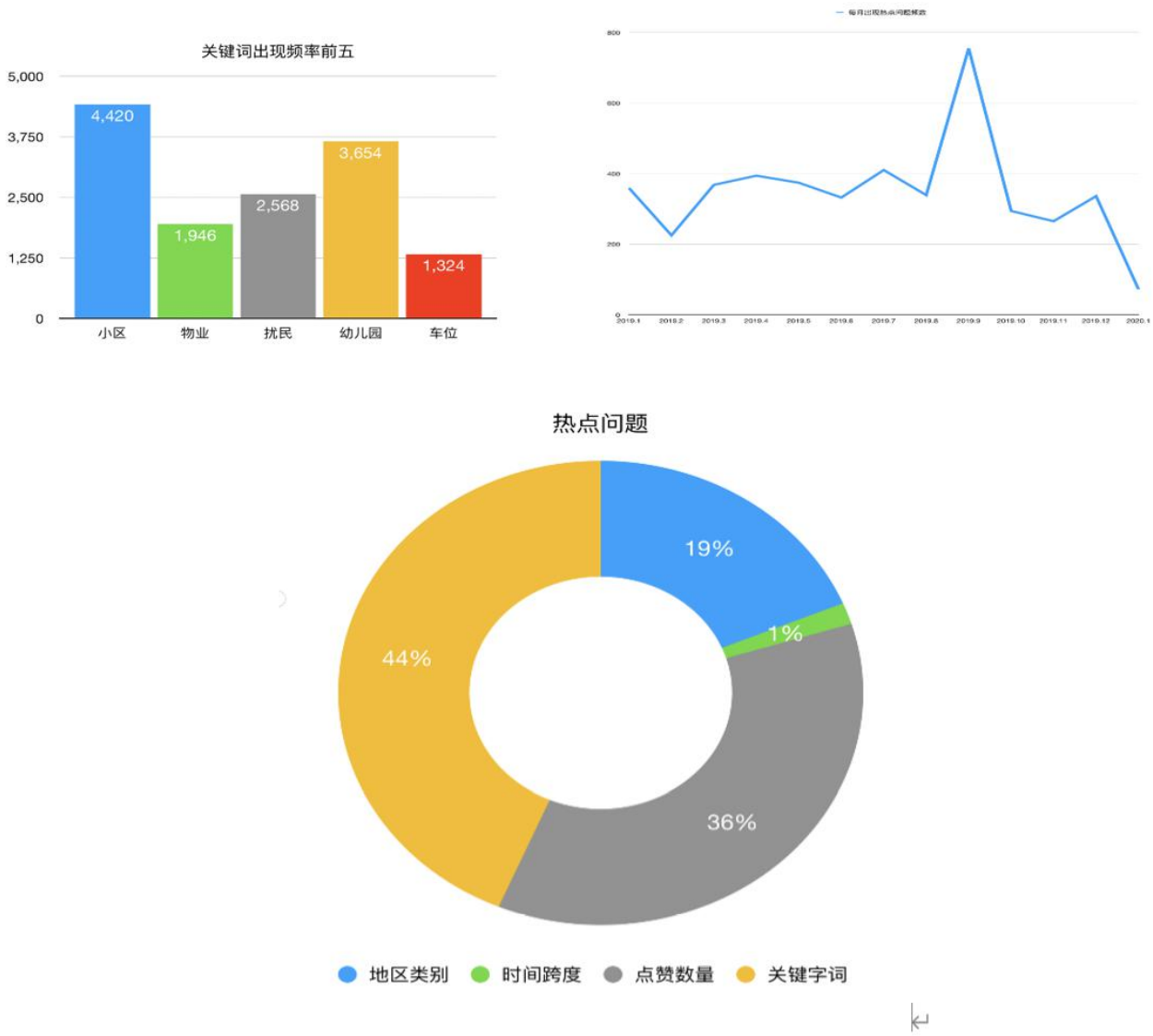
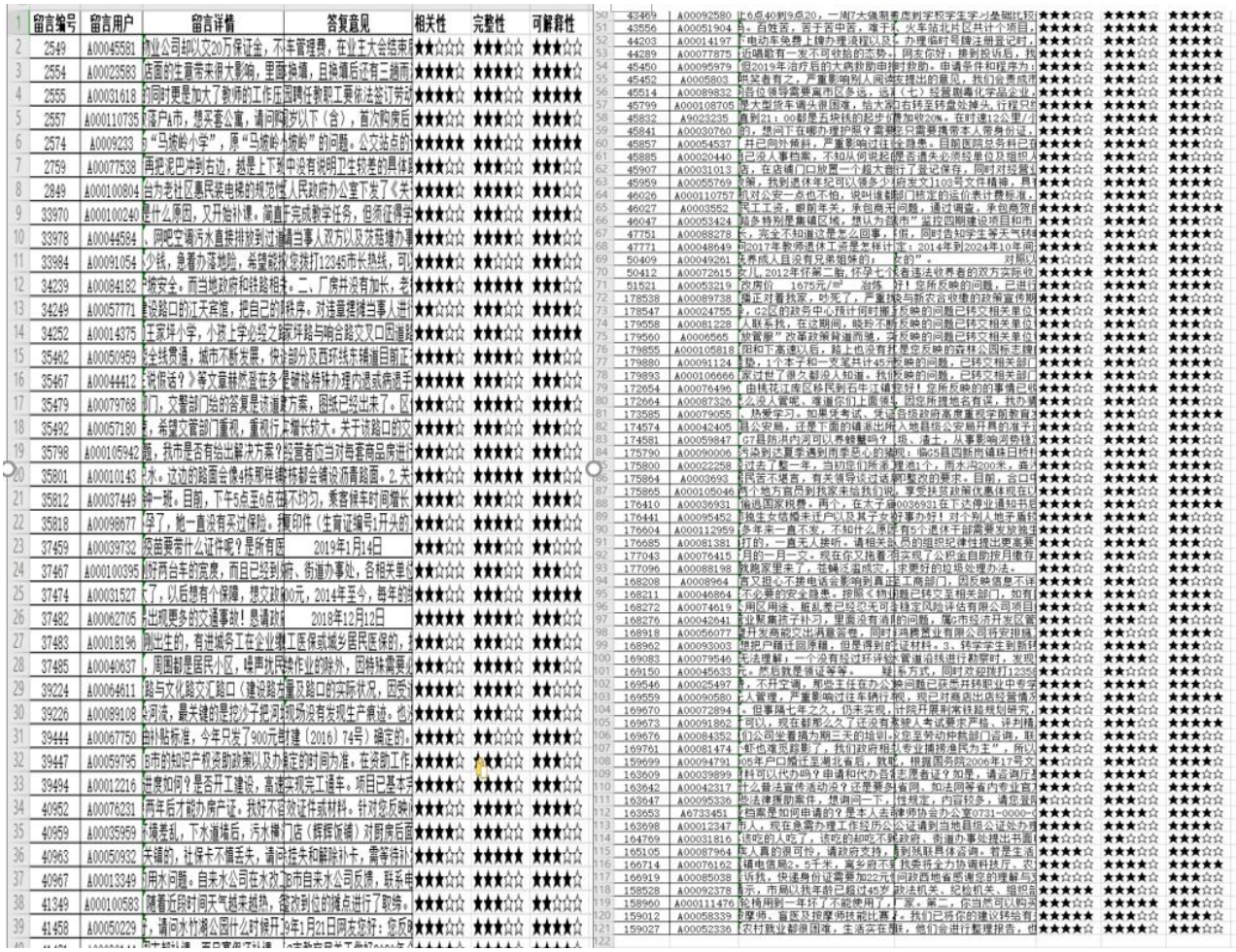


图 3-1 热点问题的探究

热点问题被喻为社会的“皮肤”“温度计”，是社会环境的动态表达式，是多元民意的集中式体现。热点问题提取模块从文本聚类结果中提取话题信息并进行热度评估与排序。在某一时段内将反映特定地点或特定人群问题的留言进行归类，定义出地区类别、时间跨度、点赞数量、关键字词四个热度评价指标，由图可知，关键字词和点赞数量占比更重，更能决定问题的热度。统计热点问题数据采集时间段内平台上的热门话题，将热点结果进行分析，发现热度排序结果上存在一定的差异。通过对热点原始数据的分析，认为原因可能是在热点的热度计算方案上存在不足，导致排序结果有所偏差。进一步而言，对热点问题建模的研究还有所欠缺，需要后续改进。

3.3 问题 3 结果分析

根据上述方案对本题目建立 BP 神经网络算法，带入附件四部分的数据得到如图 3-2 所示：



留言编号	留言用户	留言详情	答复意见	相关性	完整性	可解释性
2549	A00045581	物业公司却以交20万保证金，不车管理费，在业主大会结束后		★★☆☆☆	★★★☆☆	★★★☆☆
2554	A00023583	店面的生意带来很大影响，里面施墙，且施墙后还有三施雨		★★★★☆	★★★★☆	★★★★☆
2555	A00031618	的同时更是加大了教师的工作压因聘任教职工要依法签订劳动		★★★★☆	★★★☆☆	★★★★★
2557	A000110735	客户a市，想买套公寓，请问购房以下（合），首次购房后		★★★☆☆	★★★☆☆	★★★☆☆
2574	A0009233	5“马坡岭小学”，原“马坡岭小坡岭”的问题。公交站点的		★★★★★	★★★☆☆	★★★★★
2759	A00077538	再把泥巴冲到右边，越是上下雨中没有说明卫生状差的具体		★★★★☆	★★★☆☆	★★★☆☆
2849	A000100804	台为老社区惠民装电梯的规范性（人民政府办公室下发了《关		★★★★☆	★★★☆☆	★★★☆☆
33970	A000100240	是什么原因，又开始补课。简直扩完成教学任务，但须征得学		★★★☆☆	★★★★☆	★★★★☆
33978	A000445584	，网吧空调污水直接排放到过道路当事人双方以及放茹塘办事		★★★☆☆	★★★★☆	★★★★☆
33984	A00091054	少钱，急着办落地的，希望能找你投打12345市长热线，可		★★★★★	★★☆☆☆	★★★☆☆
34239	A00084182	质安全。而当地政府和铁路相关。二、厂房并没有加长，老		★★★★★	★★★★★	★★★☆☆
34249	A000577711	设路口的江天宾馆，把自己的厨房。对违章建筑当事人进行		★★☆☆☆	★★★★☆	★★★★☆
34252	A000114375	王家坪小学，小路上学必经之路，环路与响合路交叉口因建路		★★★☆☆	★★★★☆	★★★★☆
35462	A000509569	全线贯通，城市不断发展，你社部办及西环线东辅道目前正		★★★★☆	★★★★☆	★★★★☆
35467	A000444412	说假话？》等文章赫然登在多《金融格特殊办理户退或病退手		★★★★★	★★★☆☆	★★★☆☆
35479	A000797681	门，文暨部门始的答复是该道建方案，图纸已经出来了。区		★★★☆☆	★★★★☆	★★★★☆
35492	A000571801	建，希望交管部门重视，重视行人增长较大。关于该路口的交		★★★★☆	★★★★☆	★★★★☆
35798	A000105942	题，我市是否有给出解决方案？经营者应当对每套商品房进行		★★★☆☆	★★★★☆	★★★★☆
35801	A00010143	水。这边的路面会像4栋那样铺特都会铺设沥青路面。2.关		★★★★☆	★★★★☆	★★★★☆
35812	A000037449	种一班。目前，下午5点至6点车均不均匀，乘客候车时间增长		★★★★☆	★★★★☆	★★★★☆
35818	A00098677	孕了，她一直没有买过保险。i复印件（生育证编号1.开头的		★★★☆☆	★★★★☆	★★★★☆
37459	A000039732	疫苗要带什么证件呢？是所有国 2019年1月14日		★★★★☆	★★★★☆	★★★★☆
37467	A000100395	好两台车的宽度，而且已经到府、街道办事处，各相关单位		★★★★☆	★★★★☆	★★★★☆
37474	A000031527	了，以后想有个保障，想交政和0元，2014年至今，每年的缴		★★★☆☆	★★★★☆	★★★★★
37482	A000662705	出现更多的交通事故！恳请政府 2018年12月12日		★★★★★	★★★☆☆	★★★☆☆
37483	A00018196	刚出生的，有进城务工在企业缴工医保或城乡居民医保的，才		★★★★☆	★★★★☆	★★★★☆
37495	A000406371	，周围都是居民小区，噪声扰民特作业的除外，因特殊需要必		★★★★☆	★★★★☆	★★★★☆
39224	A000646111	路与文化路交汇路口（建设路加重及路口的实际状况，因受		★★★★☆	★★★★☆	★★★★☆
39226	A000891081	河流，最关键的是挖沙子把河和坝场没有充分发生产痕迹。也		★★★★☆	★★★★☆	★★★★☆
39444	A000677501	由补贴标准，今年只发了900元就建（2016）74号）确定的。		★★★★☆	★★★★☆	★★★★☆
39447	A000597951	B市的知识产权资助政策以及办确定的时间为准。在资助工作		★★★★☆	★★☆☆☆	★★★★☆
39494	A00012216	进度如何？是否开工建设，高速实现完工通车。项目已基本完		★★★★☆	★★★★☆	★★★★☆
40952	A000762311	两年后才能办房产证。我好不容放证件或材料。针对您反映的		★★★☆☆	★★★★☆	★★★★☆
40959	A000359591	铺差乱，下水道堵后，污水槽门店（辉辉饭铺）对厨房后面		★★★★★	★★★☆☆	★★★☆☆
40963	A000509321	关铺的，社保卡不慎丢失，请问挂失和解除补卡，需等待补		★★★★★	★★★★☆	★★★★☆
40967	A00013349	用水问题。自来水公司在水改工作中自来水公司反馈，联系电		★★★★☆	★★★★☆	★★★★☆
41349	A000100583	随着近段时间天气越来越热，室改到位的楼点进行了取暖。		★★★☆☆	★★★★☆	★★★★☆
41458	A000502291	予，请问水竹湖公园什么时候开为年1月21日网友您好：您反		★★★☆☆	★★★★☆	★★★★☆

图 3-2 答复质量结果（部分）

根据结果，综合来看出该部门对居民提出问题的答复质量等级为良。答复内容的相关性、完整性、可解释性分别为良、中、良。（表 3-4）说明该政府部门对群众反映问题的效率、与用户需求的匹配程度还是比较良好的，唯一可以加强的是对群众意见内容的广度和深度进行探究。

表 3-4 答复意见质量综合图

答复意见质量	相关性	完整性	可解释性
	★★★★☆	★★★☆☆	★★★★☆

四. 结论

总结本次比赛，我们根据群众留言的分类及其特点，利用增强型最大匹配分词方法（IMM 法）对留言进行预处理，利用改进的 KL-Divergence 特征选取算法和特征权重的计算 TF-IDF 法对特征项赋权值，进而结合 KL 特征选取算法得出群众留言的分类模型；然后又根据特定时间、特定地点、以及发生的问题建立模型，利用 jieba 分词器对热点问题进行探究；最后又利用 BP 神经网络算法对群众留言答复质量进行评价。最终实现了本次的挖掘目标：对群众留言进行分类，群众反映强烈的热点问题进行文本挖掘以及答复质量的评价。

本次评论数据挖掘分析的过程中，每一步都通过程序实现，进行了大量的数据挖掘分析工作，实验中的每一步都有理有据，各个步骤之间联系密切，条理清晰且系统地完成了本次数据挖掘分析工作。但是在实验过程中依旧遇到了很多瓶颈问题，例如在寻找热点问题上，怎样对数据进行更高效的清洗，还有在建立群众答复质量评估问题上，对 BP 神经网络中提及的答复意见质量的调查问卷如何获取以及其问卷的可靠性如何最优化。在之后的研究学习过程中，我们将继续针对热点问题的寻找方法进行优化探究。

五. 参考文献

[1]杨贵军,徐雪,凤丽洲,徐玉慧.基于最大匹配算法的似然导向中文分词方法[J].统计与信息论坛,2019,34(03):18-23.

[2]赵金楼,朱辉,刘馨.基于改进 TFIDF 的图书馆知识群体特征提取研究[J].系统科学与数学,2019,39(09):1450-1461.

[3]鲁芳.多重文本数字水印技术研究[D].长沙:湖南大学,2005.

[4]张启宇,朱玲,张雅萍.中文分词算法研究综述[J].情报探索,2008(11):53-56.

[5]石莉,陈诚,邵艺.基于 BP 神经网络的大学生实践教学效果评价研究[J/OL].扬州大学学报(高教研究版),2020(02):1-7[2020-05-06].https://doi.org/10.19411/j.cnki.1007-8606.2020.02.018.