基于NLP的“智慧政务”中的文本挖掘方案

摘要

近年来，随着网络信息技术的发展，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，市民向政府反应问题渠道的多元化也带来为相关部门的数据整理带来了许多问题。传统的人工留言划分在处理大量数据时效率较低，且长时间的工作对准确率也有着较大影响，在面临大量无序不定形数据是也无从下手。因此，结合自然语言处理技术（NLP）与机器学习实现“智慧政务”对提升政府的管理水平和施政效率具有极大的推动作用。

对于问题一：在进行文本分类之前，通过相似度计算以留言主题代替留言详情，大大简化文本数据，然后通过 jieba 分词等自然语言处理技术对文本进一步降维，突出文本的特征词。为了将文本向量化表示，我们使用 TF-IDF 权重策略将文本转化为向量形式同时得到权重矩阵。同时利用高斯朴素贝叶斯算法在处理连续数据时的优点，采用高斯朴素贝叶斯算法对文本进行分类。

对于问题二：基于对留言得到的权重矩阵，采用 K-Means 算法结合欧式距离进行聚类分析。由于文本数量较大且分布十分不均匀，因此我们采用多次聚类将留言中群众广泛反映的问题筛选出来，概括其问题类型，再从中聚类提取出具体的问题。在得到问题的留言数量后，根据层次分析法，计算留言数量、时间范围、反对数、点赞数四个指标的权重，结合各项指标的具体信息，以百分制得到问题的热度指数，从而得到热点问题排名及前五的热点问题。

对于问题三：从答复的相关性、完整性、可解释性三个角度建立对答复意见的评价方案。对于相关性，我们通过将文本之间的相似性转化为特征词项向量之间的相似度。同时引入加权因子以便更好衡量特征词项在文本中的重要程度。并通过词项相似度加权树（TSWT）来计算两个关键词项向量的相似度；对于完整性，按照完整性标准建立新的语料库并导入 BTM 模型，利用 BTM 模型与循环神经网络结合，对留言答复意见内容进行筛选处理，最后利用机器学习处理全部数据；对于可解释性，主要利用相似性与完整性方法结合，某种程度上可以理解为相似性越高，可理解性越强，内容越完整，解释性越高。

关键词：自然语言处理（NLP）、TF-IDF、K-Means、欧式距离、层次分析、机器学习、TSWT、BTM

Abstract

In recent years, with the development of network information technology, network political platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit. The diversification of the channels for citizens to respond to problems to the government has also brought many problems to the data collation of relevant departments. Traditional manual message division is less efficient in processing a large amount of data, and long hours of work also has a great impact on the accuracy rate, in the face of a large number of unordered data is also impossible. Accordingly, combining natural language processing technology (NLP) and machine learning to realize "intelligent government" has a great promotion effect on improving the management level and governance efficiency of the government.

For question one: before the text classification, the message topic is replaced by the message details by similarity calculation, which greatly simplifies the text data, and then further reduces the dimension of the text by jieba word segmentation and other 1 natural language processing techniques to highlight the characteristic words of the text. for representing text to quantization, we use TF-IDF weight strategy to transform text into vector form and get the weight matrix at the same time. At the same time, using the advantages of Gauss simple Bayesian algorithm in processing continuous data, the text is classified by Gauss simple Bayesian algorithm.

For question two: based on the weight matrix obtained from the message, the K-M eans algorithm combined with Euclidean distance is used for cluster analysis. Because the text quantity is large and the distribution is very uneven, we use multiple clustering to screen out the problems widely reflected by the masses in the message, summarize the problem types, and then extract the specific problems from the clustering. After getting the number of messages, according to the analytic hierarchy process, the weight of the four indexes of message number, time range, opposition number and likes number is calculated, combined with the specific information of each index, the heat index of the problem is obtained by percent system, so as to get the hot spot problem ranking and the top five hot spot problems.

For question 3: build responses evaluation programme from three perspectives: relevance, completeness and interpretability For correlations, we transform the similarity between texts into the similarity between feature word item vectors. Weighted factors are also introduced to better measure the importance of feature

terms in the text. The similarity of two key word vectors is calculated by word item similarity weighted tree (TSWT). For integrity, a new corpus is built according to the integrity standard and BTM model is imported. The BTM model is combined with cyclic neural network.

Keywords: Natural Language Processing (NLP)、 TF-IDF 、 K-Means 、 Euclidean distance、Hierarchical analysis、Machine Learning、TSWT 、BTM

目录

一、简介

1.1 研究背景

伴随着互联网的广泛应用和发展，人们的思想意识、价值观念产生了极大变化，民主意识比过去相对封闭的情况下大大增强，党的执政环境已经悄然发生变化，互联网正在改变中国的政治生态环境。我国政府在不断变化的执政环境中，本着必须不断地发展与创新，及时跟上时代发展的步伐的理念，面对挑战，主动适应，加强网络问政的制度建设，促进网络问政常态化，形成了以电子政务为中心的管理链条。电子政务是政府部门利用现代信息科技和网络技术，实现高效、透明，规范的电子化内部办公，协同办公和对外服务的程序、系统、过程和界面。与传统政府的公共服务相比，电子政务除了具有公共物品属性，如广泛性、公开性、非排他性等本质属性外，还具有直接性、便捷性、低成本性以及更好的平等性等特征。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、凝聚民气重要渠道的同时，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。

1.2 研究现状

在传统政务阶段,由于技术等条件的限制,政府主要采取面对面的方式为市民提供公共服务。随着信息通信技术的发展和互联网的出现,电子政务应运而生。政府服务的效率得到极大的提高，但其服务的提供仍受时间和空间的限制。当前阶段政府变得更加“智慧”、效率更高、管理更透明，智慧政务呈现出简便、透明、自治、移动、实时、智能和无缝对接等特征。从某种程度上讲，智慧政务综合体现了“以公众为中心”、“惠及所有人”、“无缝”、“透明的政府”、“回应的政府”、“变革的政府”等理念，是一种先进的、成熟的公共服务范式。当今世界各国都将推进电子政务建设作为带动本国经济发展和社会信息化发展的一项重要举措。我国政府对电子政务的研究和应用进行了大量的投入，中国电子政务学科领域已形成了较为完备的理论研究体系、方法论研究体系以及实证研究体系,如何梳理这个复杂研究系统的结构，并从整体上把握现有的研究热点并进行应用具有重要意义。

二、研究目标及流程

对于第一问，利用 jieba 分词对经过预处理的数据进行分词处理，并通过去

停用词和词云图绘制进行数据清洗，突出高频词汇。考虑到词频对分类结果的作用，通过 TF-IDF 权重策略将文本数据进行向量化表示并放入模型训练。对于模型的构建，由于受到文本数据量的限制以及所得属性为连续值，假定满足高斯分布，以高斯朴素贝叶斯分类算法构建模型。

对于第二问，首先将经过分词预处理的留言主题通过 TF-IDF 权重策略转化为权重向量。由于需要在留言中提取热点问题，并且热点问题类型未知，因此需要将留言进行聚类分析。我们通过 K-Means 算法，根据欧式距离将留言聚类进行相似性度量，通过多次聚类分析提取出集中反映的问题类型。然后根据层次分析法得到确定的 4 项度量指标的权重，并以计算热点问题的热度指数，得出热点问题排名。

对于第三问，建立答复意见的评价体系将通过答复意见与留言之间的相似性、完整性及可解释性三方面构建评价模型。

三、群众留言分类

3.1 数据预处理

在分类前，首先将附件中的一级分类标签进行赋值，结果如表 1 所示：

一级分类标签	赋值
城乡建设	1
党务政务	2
国土资源	3
环境保护	4
纪检监察	5
交通运输	6
经济管理	7
科技与信息产业	8
民政	9
农村农业	10
商贸旅游	11
卫生计生	12
政法	13
教育文体	14
劳动和社会保障	15

表 1

在对数据进行预处理前，首先对留言主题和留言内容进行相似度计算[1]。

(1)句子预处理:对留言详情进行分词处理和句法分析两个部分，对群众提出

问题当中关键词权重的设置做基础。

(2)关键词权重设置:根据留言详情中关键词的词性以及在句子中的依存关系，为每个关键词设置权重。

(3)关键词扩展:由于有些词有同义词和近义词，为了更准确的与留言主题句进行相似度的计算，对留言详情句中的关键词进行了相应的同义词近义词扩展。

(4)二叉树带权路径长度计算:将群众提出的所有关键词对应的权重以最优二叉树的形式表示出来，并计算该二叉树的带权路径长度。

(5)留言详情句与留言主题句的相似度计算:根据留言详情的带权路径长度和留言主题中包含的关键词，对留言详情与留言主题之间的相似度进行计算。得出留言主题的权重可视为其与留言详情之间的句子相似度。

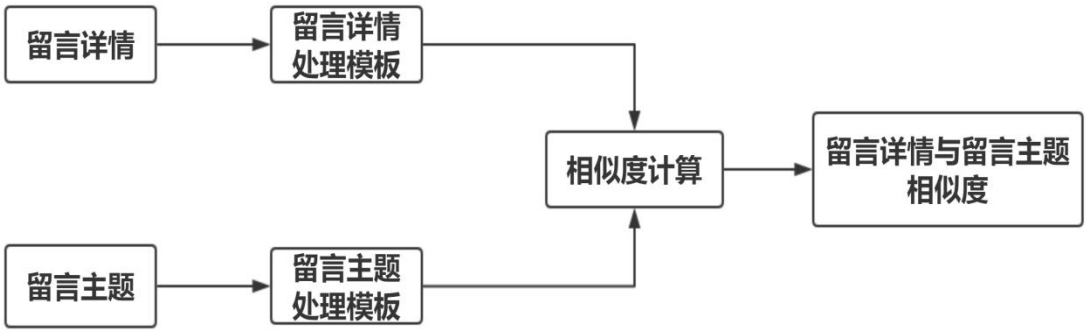

图 1

根据计算结果，可以确定留言主题对留言内容的符合程度，从而用留言主题替代留言内容，简化留言的数据量，根据留言主题对留言进行分类。

由于留言数据中包含敏感数据，在进行分词前，对数据进行脱敏处理，将数据中如留言用户的敏感信息进行脱敏处理，通过 python 去除脱敏后的 X 序列。

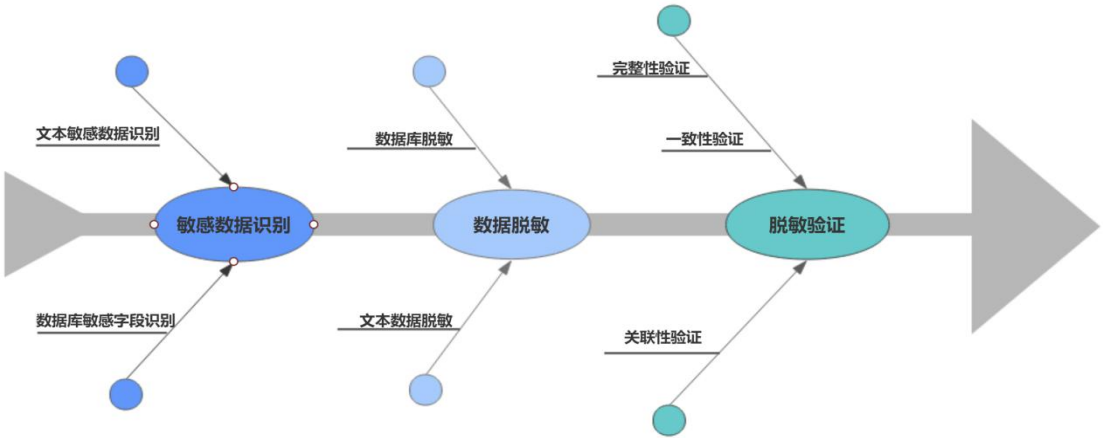

图 2

然后根据对一级分类标签的赋值，将留言数据根据不同的标签分别抽样提取，将分布不均衡的数据通过干预实现均衡，以保证在训练过程中能完整记录样

本特征，防止过拟合或欠拟合的情况。

3.1.1 分词处理

由于中文文本中的词与词之间没有明显界限，因此要对文本中的词语进一步处理首先需要对留言进行 jieba 分词分词处理。

jieba 分词主要是基于统计词典，构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，也就得到了最终的切分形式。

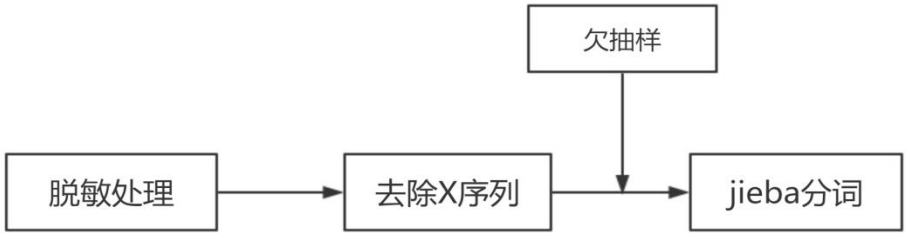

图 3

3.1.2 去停用词

在留言主题中，存在着许多功能及其普遍，没有实际含义，对模型没有实际意义甚至负面影响的数据，因此在对数据预处理的阶段，根据分词处理的结果，对数据进行清洗，主要为去停用词。

在文本处理中，停用词是指那些功能极其普遍，与其他词相比没有什么实际含义的词，这些词往往是语言中一些表意能力很差的辅助性词语，它们通常是一些单字、单字母、标点符号以及高频的单词。在预处理阶段，主要是针对文档中的词组进行处理，对于中文文本而言,因为词与词之间没有明显的切分标志，所以首先需要对中文文本进行分词。分词技术是文本分类首先要解决的问题，分词结果的好坏直接影响到分类系统的性能。另外，预处理阶段停用词的处理也有着粗降维和提高分类精度的作用，提高计算速度，而且对文本分类的准确性也有很大的帮助。在这里，我们结合哈工大停用词表和四川大学机器智能实验室停用词表，并根据文本分词结果自定义去停用词词典。

3.1.3 绘制词云图

词云图是文本结果展示的有利工具，通过词云图的展示可以对短信文本数据分词后的高频词予以视觉上的强烈突出效果，更直接的获取到主旨信息。

以城乡建设类为例，得到词云图如下所示

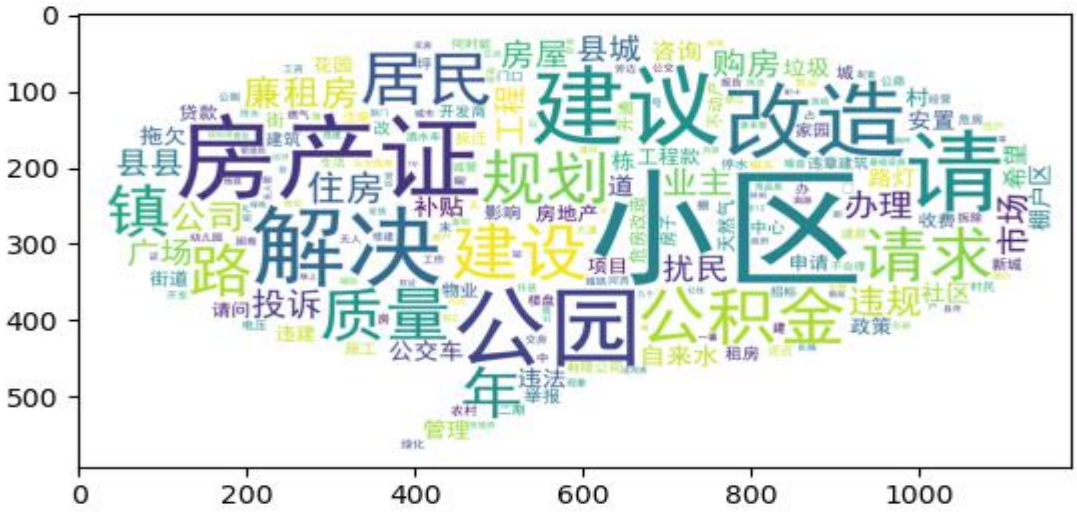

图 4

3.2 模型构建

对信息进行预处理后，需要根据分词结果构建信息识别模型。首先需要对数据进行处理，进行向量化表示。考虑到信息的词频特征，通过 One-Hot 编码将分词所得词袋中的每一个词表示为向量，从而将文本转化为向量形式的方法忽略了文本的词频信息，因此考虑采用 TF-IDF 权重策略对文本进行向量表示。

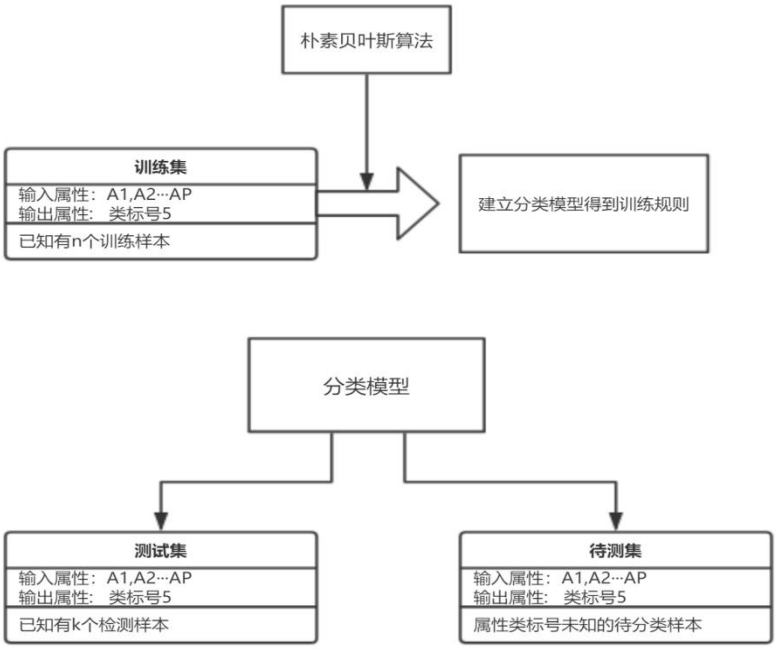

图 5

3.2.1 文本的向量表示

TF-IDF 权重策略是一种用于评价某一字词对于文本重要程度的有效方法。一般来说，我们将文本表示为文本中出现的 N 个加权词项组成的向量，如果一个字或词项在某一文件中出现的次数越多，说明其与文本主题越相关，显得越重

要，且二者呈正比例增加，我们称之为关键词词频（Term Frequency）。

$$TF = \frac{N（词频）}{M(总文本数)}$$

IDF（Inverse document frequency）则是指逆向文本频率，是用于衡量关键词权重的指数。以 D 表示文档总数，D_w 表示某单词在多少个文档出现次数。

$$IDF = \log(\frac{D}{D_W})$$

$$TF - IDF = TF \times IDF$$

通过 TF-IDF 权重策略将分词后的文本转化成词频向量，得到 TF-IDF 权重矩阵，放入模型。

3.2.2 模型训练评估

在数据分析领域，分类算法有很多，其原理千差万别，有基于样本距离的最近邻算法，有基于特征信息熵的决策树，有基于 bagging 的随机森林和梯度提升分类树，但其实现的过程相差不大。在 python 的 sklearn 库中提供了多种分类算法。由于所给文本有限，数据较少，经过对多种算法的比较，经过朴素贝叶斯算法构建的模型所训练的结果准确率最高。

朴素贝叶斯算法对小规模的数据表现很好，能处理多分类任务，并且对缺失数据不太敏感，常用于文本分类。模型描述如图所示：

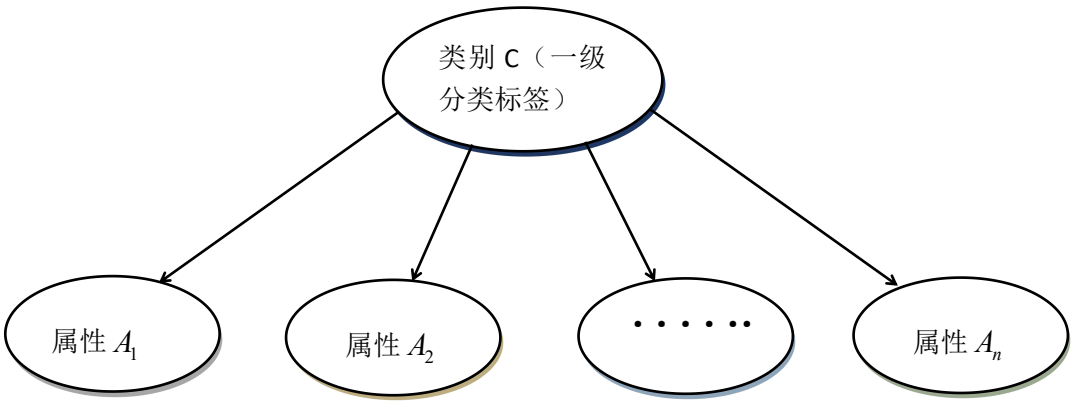

图 6

朴素贝叶斯分类模型[2]假设所有的条件属性 $A_i(i = 1, 2, \cdots, n)$ 都作为类变量 C 的孩子节点，将给定的一个待分类样本 $X = \{a_1, a_2, \cdots, a_n\}$ 分配给类 $C_i(1 \le i \le m)$，当且仅当：$P(C_i \mid X) > P(C_j \mid X)\,(1 \le i, j \le m, j \ne i)$。

根据贝叶斯定理，有：

$$P(C_i \mid X) = \frac{P(C_i)P(X \mid C_i)}{P(X)}$$

如果事先不清楚类在数据集中的概率情况时，可以假设每个类别的概率相等。即有:

$$P(C_i) = P(C_j)\text{，}\ (C_i, C_j \in C,\ \ i \neq j)$$

并根据这个对 $P(C_i \mid X)$ 最大化。否则，最大化 $P(X \mid C_i)P(C_i)$。因 $P(X)$ 对于所有的类别均为常数，故有:

$$P(C_i \mid X) = \frac{P(C_i)P(X \mid C_i)}{P(X)} \propto P(C_i)P(X \mid C_i)$$

由朴素贝叶斯分类算法的条件属性相互独立的假设，有:

$$P(C_i \mid X) \propto P(C_i)\prod_{k=1}^{n} P(a_k \mid C_i)$$

其中 $P(C_i) = \frac{S_i}{S}$，S_i 是 C_i 在训练样本中的实例数，S 是训练样本数。则 NBC 模型的公式表达式为：

$$NB(X) = \arg\max_{C_i \in C} P(C_i)\prod_{k=1}^{n} P(a_k \mid C_i)$$

概率 $P(a_1 \mid C_i), P(a_2 \mid C_i), \cdots, P(a_n \mid C_i)$ 可由训练样本估值。其中，由于经过向量化表示的属性 A_k 是连续值，一般假定它服从高斯分布。因而，

$$P(a_k \mid C_i) = g(a_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(a_k - \mu_{C_i})^2}{2\sigma_{C_i}^2}}$$

其中 $g(a_k, \mu_{C_i}, \sigma_{C_i})$ 是属性 A_k 的高斯密度函数，μ_{C_i}，σ_{C_i} 分别为平均值和标准方差。

以 F-Score 对分类方法进行评价：

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

经过对模型的训练，F1 可达到 0.661

四、热点问题挖掘

4.1 权重向量的生成

在对热点问题进行挖掘分析之前，先要把非结构化的文本信息进行处理。由于留言主题中含有无效的内容，我们先将对留言主题进行检查并完善。为了便于

转换，先要对这些描述信息进行中文分词。这里采用 python 中的 jieba 库进行分词，在 jieba 分词前，我们已经对数据进行预处理，即去停用词（由于第一问已应用，此处不再赘述）。在对热点问题描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。在此，我们运用 TF-IDF 权重策略，生成 TF-IDF 权重矩阵。

4.2 热点问题分类

根据生成的 TF-IDF 权重矩阵，对留言反应的问题进行分类。这里采用 K-Means 算法把类型分成 K 类。

4.2.1 K-Means 聚类分析

K-Means 是一种基于划分方法的典型聚类算法，该算法始于一个簇的中心集合，该集合采用通过 TF-IDF 权重策略所得到的权重矩阵。在每次的迭代过程中，每个样本点根据计算相似度被分配到最近的簇中。然后，重新计算簇的中心，也就是每个簇中所有数据的平均值[3]。

每个簇的中心就是所有这个簇的所有样本点的中心:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q$$

其中 N_k 是属于簇 k 的样本数目，μ_k 是指簇 k 的中心。其具体做法是在欧几里得多维空间里把包含 N 个数据对象的数据集合划分为 K 个划分 $(k << n)$，其中每个划分分别代表一个聚类的簇。首先，我们指定将要聚类簇的个数 K，并用初始聚类中心选择策略选择 K 个数据对象作为初始聚类中心，对集合中除初始聚类中心（初始聚类中心可以不是集合中的对象）以外的其它数据对象，根据其与各个聚类中心的欧式距离将它置于离其最近的簇中。然后，重新计算每个簇中所有数据对象的平均值形成新的聚类中心。这个聚类过程重复迭代进行，直到满足预先设定的聚类终止条件为止。K-Means 聚类算法的步骤的描述如下：

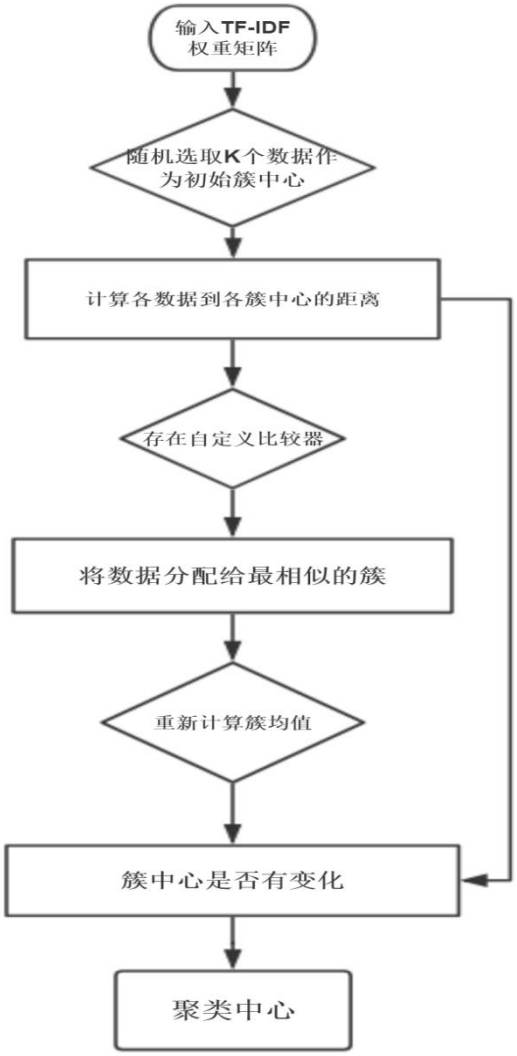

图 7

(1)从 TF-IDF 权重矩阵中随机取 K 个元素，作为 K 个簇的各自的中心。

(2)分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

(3)根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

(4)将 X 中全部元素按照新的中心重新聚类。

(5)重复第 4 步，直到聚类结果不再变化。

(6)输出结果。

选取 K=17，根据得到聚类结果，将 17 个聚类结果中内容紧密相关，相似度大的 5 个聚类簇进行简单概括，如表 2 所示：

聚类簇	留言概括	聚类簇	留言概括
15	购房补贴	10	安全隐患

12	城市中心建设	8	捆绑销售
11	商铺扰民		

表 2

通过 K-Means 聚类算法对留言进行一次聚类分析所得到的结果并不准确，在同一聚类簇有许多互不相关的留言主题，由于其数目较少，分布零散，仅通过一次聚类分析无法将包含在聚类簇中的无关留言分离出来并对热点问题进行准确的聚类。因此，在第一次聚类结果的前提下，将聚类效果良好的聚类簇剔除，进行多次聚类分析，以得到留言中由群众集中反映的问题类别。

第二次选取 K=15，得到三个聚类簇，如下图所示：

聚类簇	留言概括	聚类簇	留言概括
4	施工扰民	5	拖欠工资
10	虚假宣传		

表 3

第三次选取 K=17，得到三个聚类簇，如下图所示：

聚类簇	留言概括	聚类簇	留言概括
6	幼儿园	8	学校违规问题
12	住房质量		

表 4

第四次选取 K=18，得到三个聚类簇，如表 5 所示：

聚类簇	留言概括	聚类簇	留言概括
8	非法诈骗	12	拆迁问题
16	搅拌厂扰民		

表 5

第五次选取 K=23，得到三个聚类簇，如表 6 所示：

聚类簇	留言概括	聚类簇	留言概括
3	商铺及渣土车扰民	7	物业不良行为

表 6

4.2.2 层次分析法量化指标权重

热点问题的衡量指标包括留言数量、时间范围、点赞数、反对数。为了根据不同的四个指标综合衡量留言的热度指数，采用层次分析法确定四类指标的权重。

应用 9 分位标度法，判断矩阵 A 中要素 a 的值按如下方法确定：

取数字 1、3、5、7、9 分别表示两两指标相比为同等重要、稍微重要、明显重要、强烈重要和极端重要；2、4、6、8 为介于每两个等次之间的取值。

标度值	含义
1	两指标对某属性同等重要
3	两指标对某属性，一指标对另一指标稍微重要
5	两指标对某属性，一指标对另一指标明显重要
7	两指标对某属性，一指标对另一指标强烈重要
9	两指标对某属性，一指标对另一指标极端重要
$\frac{1}{9}$	两指标对某属性，一指标对另一指标极端次要
$\frac{1}{7}$	两指标对某属性，一指标对另一指标强烈次要
$\frac{1}{5}$	两指标对某属性，一指标对另一指标明显次要
$\frac{1}{3}$	两指标对某属性，一指标对另一指标稍微次要
2、4、6、8	上述相邻判断的中间值

表 7

1、根据资料文献，结合专家分别对留言数量、时间范围、赞成数、反对数四者进行的打分，取 10 位专家对同一层次内 4 个指标的相对重要性（两两因素之间）的打分结果，得到的平均数建立 4×4 的判断矩阵：

$$\begin{bmatrix} 1 & 3 & 1 & \frac{1}{9} \\ \frac{1}{3} & 1 & \frac{1}{5} & \frac{1}{9} \\ 1 & 5 & 1 & \frac{1}{6} \\ 9 & 9 & 6 & 1 \end{bmatrix}$$

2、计算权重，将矩阵 A 的各行向量进行几何平均（方根法）

（1）计算判断矩阵每一行元素的乘积

（2）计算每一行的 4 次方根

然后进行归一化处理，得到各评价指标权重和特征向量W（利用 Matlab 进行矩阵运算）

$$w_i=\frac{\overline{W_i}}{\sum_{i=1}^{n}\overline{W_i}};\quad W=\begin{Bmatrix}W_1\\W_2\\\vdots\\W_n\end{Bmatrix}$$

得到权重向量$\begin{bmatrix}0.759\\0.293\\0.955\\4.695\end{bmatrix}$，四者权重系数分别为：留言数量：0.7031；时间范围：0.0448；反对数：0.1107；点赞数：0.1414。

3、 判断矩阵的一致性检验，所谓一致性是指判断思维的逻辑一致性。计算最大特征根$\lambda_{\max}$， n为判断矩阵阶数：

$$\lambda_{\max}=\frac{1}{n}\sum_{i=1}^{n}\frac{(AW)_i}{w_i}$$

根据随机一致性指标RI，计算一致性指标CI和一致性比例CR：

阶数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI值	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	0.49	0.52	1.54	1.56	1.58	1.59

表 8

$$RI=0.89;\quad CI=\frac{\lambda_{\max}-n}{n-1};\quad CR=\frac{CI}{RI}$$

经计算$CR<0.1$，符合要求。

4.3 热点问题生成

根据聚类得到的集中留言概括提取出的重要问题，如下表所示：

问题描述	留言数量	时间范围	点赞数	反对数
投诉A市伊景园滨河院捆绑销售车位	49	2019/7/7 至 2019/9/1	25	1
A5区魅力之城小区底层商铺油烟噪音扰民	21	2019/07/21 至 2019/9/25	18	18
A市城市中心城市经济建设	18	2019/1/8 至 2019/7/30	28	1
A市人才购租房政策咨询	23	2019/1/16 至 2019/12/2	39	2
A市地铁噪音扰民	39	2019/1/8 至 2019/12/4	133	3
A市经济学院强制学生实习	10	2019/4/28 至 2019/11/27	15	0
村民询问A7县拆迁问题	17	2019/1/30 至 2019/11/22	16	2
A3区枫林路存在扰民现象	13	2019/3/22 至 2019/9/2	0	0
A3区天顶街道饭店娱乐等对市民带来问题	23	2019/1/8 至 2019/12/12	16	0
A3区金麓小区经营等问题反映	11	2019/1/15 至 2019/7/25	8	0
A7县金科时代物业或生活问题	18	2019/1/2 至 2019/11/2	6	0

表 9

结合各项指标权重，得到各热点问题的热度指数，以此得到排名前 5 的热点问题，如下表所示：

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	72.84	2019/7/7 至 2019/9/1	A市伊景园滨河院	投诉A市伊景园滨河院捆绑销售车位
2	2	66.45	2019/1/8 至 2019/12/4	A市地铁	A市地铁噪音扰民
3	3	41.64	2019/7/21 至 2019/9/25	A5区魅力之城小区	A5区魅力之城小区底层商铺油烟噪音扰民
4	4	33.47	2019/1/16 至 2019/12/2	A市人才	A市人才购租房政策咨询
5	5	29.46	2019/1/8 至 2019/12/12	A3区天顶街道	A3区天顶街道饭店娱乐等对市民带来问题

表 10

五、答复意见的评价

附件四给出的留言意见答复的相关内容深刻表现了政府对民众问题的关心，但是不同的答复内容所反映的结果不同，一些答复深入民心并颁布相关措施，一些答复只是对问题进行了简单的回应却没有实际的解决方法。我们将针对附件 4 相关部门对留言的答复意见进行相关性、完整性、可解释性三方面分类研究，再结合在一起对答复意见的质量给出一套评价方案。

5.1 相关性评价

5.1.1 特征词项向量转化

对于留言回复的相关性评价，一般的文本预处理考虑到人名、地点、组织机构等词语在进行 TF-IDF 计算是具有较高的值[8]，从而容易导致对文本关键词项的错误选择，但结合留言回复的相关性的评价要求，地点人群等词对于文本相关性有一定的要求，因此在对文本进行预处理及向量转化及特征词频的选取时需要考虑这部一分内容。

当每个答复意见和其对应的留言通过文本的预处理和 TF-IDF 算法转化为为特征向量之后，需要通过计算两文本之间相似度从而对他们的相关性做出评价。由于文本中最重要的信息就是特征词项，所以文本相似度可以转化为由特征词项向量之间的相似度进行叙述，原来的文本之间的相似性就转化为特征词项向量之间的相似度。

由文献[4]知，对两个文本之间的相似度做如下定义：

若 $Sim(x,y)$ 代表 x 与 y 之间的相似度，那么应该满足下列条件：

当且仅当 $x=y$ 时，$Sim(x,y)=1(0\le S\le 1)$

根据文本相似度定义：

$$TEXT\ Sim(v_i,v_j)=\omega f\times VECT\ Sim(v_i,v_j)$$

其中，v_i，v_j 分别为两篇不同文本的特征项量。$v_i=(\omega_{i1},\omega_{i2},\cdots,\omega_{im})$，$v_j=(\omega_{j1},\omega_{j2},\cdots,\omega_{jn})$，$\omega f$ 为二者之间的加权因子，$VECT\ Sim(v_i,v_j)$ 为两特征词项向量间的相似度。

从每一段文本中挑选若干重要的词项：当两篇文本中存在很多相似度较高的词项，而且这些词项的 TF-IDF 值在各自所在文本中的比例还很高，此时可以断定上述的词项在文本中具有重要地位，根据这一原理，可以对关键词向量中满足

相似度值条件的关键词项的 TF-IDF 值在全篇文本 TF-IDF 值总和中所占的比例进行加权。

若两篇文本相似度越高，则说明含有相似词项越多，而各文本中各词项的 TF-IDF 值比率越高，则说明它们在文本中更为重要，能更好的反映文本相似的情况。

5.1.2 加权因子

余下的语义相似度较低的词项，通过它们计算文本相似度就没有太大可信程度，但是通过它们在整个文本集中的概率分布情况，也可以反映相似度。所以，可以通过计算加权因子，以便更好衡量特征词项在文本中的重要程度。加权因子通过特征词向量中满足相似度阈值条件的特征词项的 TF-IDF 值在整篇文本 TF-IDF 值的总和中所占的比例进行加权得到。

加权因子公式如下:

$$\omega f = \frac{1}{2}(\sqrt{VECT\ Sim(v_i, v_j)} - VECT\ Sim(v_i, v_j)) \times \left[\frac{\sum_{k \in \Lambda_i} TFIDF(w_{ik})}{\sum_{k=1}^{m} TFIDF(w_{ik})} + \frac{\sum_{l \in \Lambda_j} TFIDF(w_{jl})}{\sum_{l=1}^{n} TFIDF(w_{jl})} \right]$$

$$\Lambda_i = \left\{ k : 1 \le k \le m, \max_{1 \le l \le n} \left\{ Sim(w_{ik}, w_{jk}) \right\} \ge \mu \right\}$$

$$\Lambda_j = \left\{ l : 1 \le l \le n, \max_{1 \le k \le m} \left\{ Sim(w_{ji}, w_{ik}) \right\} \ge \mu \right\}$$

上式中，$TFIDF(w_{ik})$ 为特征词项 w_{ik} 的 TF-IDF 权值。

如果特征项向量 v_i 中的某一个特征项 w_{ik} 和另一特征项向量 v_j 中的某一特征项 $w_{jl}(l = 1,2,\cdots,n)$ 的相似度已经超过设定的相似度阈值 μ，那么就把特征词项 w_{ik} 纳入进集合 Λ_i 之中，同样，依照集合 Λ_i 的处理方法对集合 Λ_j 内包含的元素对特征项 v_j 内的特征项进行筛选。

$$VECT\ Sim(v_i, v_j) = \frac{1}{2}\left[\frac{1}{m} \sum_{k=1}^{m} \max_{1 \le l \le n} \left\{ Sim(w_{ik}, w_{jl}) \right\} + \frac{1}{n} \sum_{l=1}^{n} \max_{1 \le k \le m} \left\{ Sim(w_{jk}, w_{ik}) \right\} \right]$$

向量 v_i，v_j 中所包含的词项相似度决定了 $VECT\ Sim(v_i, v_j)$ 相似度很高的词项一定包含在相似的向量中，相似度低的词项所在的向量也必然不相似。

5.1.3 词项相似度加权树

词项相似度加权树 TSWT（Term Similarity Weight Tree）[5]是计算文本相似度时的一种数据结构，该数据结构用于计算加权因子 ωf 。TSWT 是一个高度为

3 的平衡树，它包含叶结点和非叶结点，词项之间相似度超过某个值 μ 的词项按照相似度从大到小的顺序组织成一个有序队列保存在叶结点中；而非叶结点则保存词项最大、最小、平均相似度以及词项数目等集合信息。下图为 TSWT 的结构图：

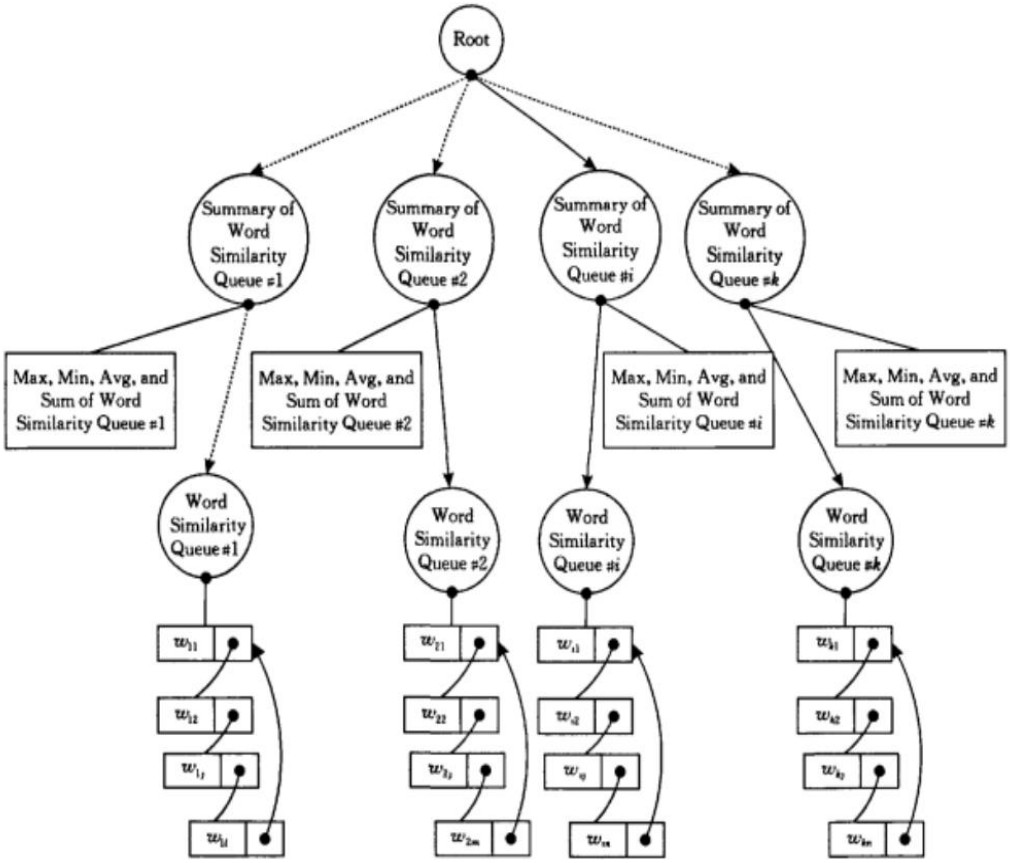

图 8

（1）词项相似度加权树的初始化

根据用户的具体任务项选择特定的特征词项构建 3 层词项相似度加权树（如果没有指定的初始条件，那么根据相应更新的规则，同时根据特征词项向量的相似度计算，系统可以自动建立相应的加权树）。在树的每个叶节点中包含的特征词项，它们彼此之间的相似度均大于某一阈值。同时，依据特征词项和其他词项的相似度的大小，由高到低依次排列特征词项。

（2）词项相似度加权树的加权及更新

在计算特征词项向量 v_i 和 v_j 之间的相似度过程中，如果特征词向量 v_i 和 v_j 中的某一对特征词项 w_{ik}，w_{jl} 满足下列条件之一的话则对关键词项向量 v_i 和 v_j 的相似度结果进行加权处理。（其中，加权的权重由特征词项 w_{ik}， w_{jl} 的 TF-IDF 值在各自特征词项向量的 TF-IDF 值总和里所占百分比确定，且特征词项 w_{ik}，w_{jl} 的相似度满足阈值 μ 的条件。

①w_{ik}， w_{jl}都属于加权树中某一个叶结点的词项有序队列。

②如果w_{jl}属于加权树中某一个叶节点的词项有序队列，而w_{ik}却不属于，且二者之间具有超过阈值μ的高相似度，则根据w_{ik}和所在词项有序队列中其它词项的相似度，在含有w_{jl}的词项有序队列中来确定w_{ik}在词项有序队列中的顺序位置。反之亦然。

③如果w_{ik}，w_{jl}都不属于加权树中某一个叶节点的词项有序队列，二者之间具有超过阈值μ的高相似度，且二者和加权树中某个叶节点的词项有序队列中的具有最大相似度的特征词项以及具有最小相似度的特征词项的相似度值，都小于某一阈值μ时，就应建立一个分支，并且将w_{ik}， w_{jl}插入到这个分支叶节点的特征词项队列里面。

④w_{ik}，w_{jl}如果都不属于加权树中某一个叶节点的词项有序队列，且二者和加权树中某个叶节点的词项有序队列中的具有最大相似度的特征词项以及具有最小相似度的特征词项的相似度值，都大于某一阈值μ时，就应该依据w_{ik}， w_{jl}和其它词项的相似度确定在该词项有序队列中w_{jl}和w_{ik}的顺序位置。

（3）文本相似度计算

根据式$TEXT\ Sim(v_i,v_j)=\omega f\times VECT\ Sim(v_i,v_j)$并利用词项相似度加权树计算两个关键词项向量的相似度。

5.2 完整性评价

由于大部分留言答复意见和文章相比属于短文本，每篇文档的信息量较少，如果通过引用外部语料(比如搜索片段，背景信息等)来将短文本扩充成长文本后应用主题模型进行建模计算,那么符合扩充条件的语料有时并不容易获得，并且最终结果很大程度上依赖用来扩充的语料信息。基于上述出现的问题我们运用BTM 主题模型中的方法对“完整性”进行分析。该方法通过整个语料库建立 BTM 语料库，不直接使用信息少的单个短文本，而是利用丰富的整个语料库的信息来进行建模和推断。基于 BTM 模型的相似度计算很好的解决了语义稀疏问题，并且无需引用外部语料进行扩充，解决了短文本过度依赖外部语料信息的问题。

5.2.1 建立语料库

语料库指经科学取样和加工的大规模电子文本库。借助计算机分析工具，研究者可开展相关的语言理论及应用研究。

首先，经过查询资料得到段落完整性的语言成分要求，比如你好、答复、感谢、回执等格式关键词。运用自然语言处理技术将这些文本进行分类，以得到的关键词词频信息及其权重为指标，建立新的语料库。

5.2.2 BTM 模型构建

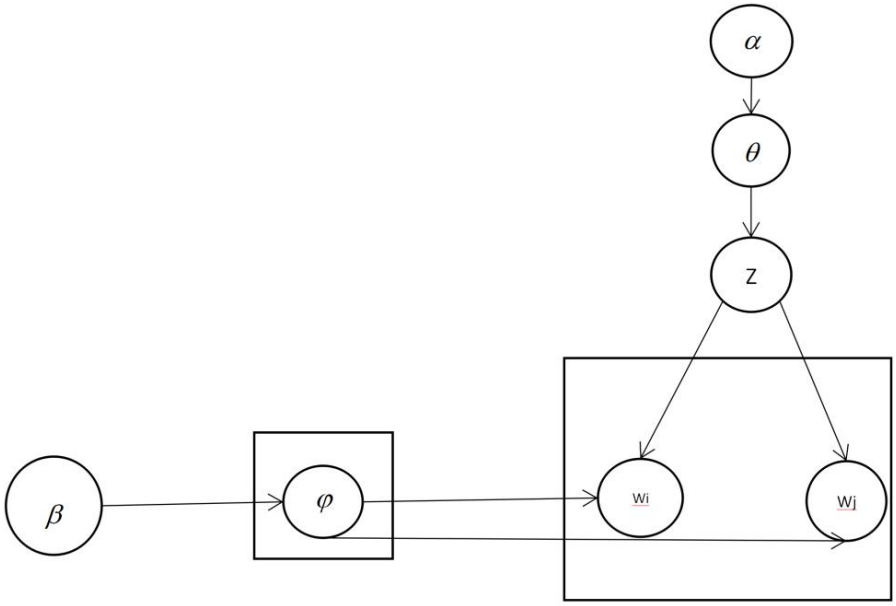

图 9

其中α是 BTM 语料库中的答复意见概率分布，θ是答复意见词对(biterm)的分布，z为词对 biterm 的主题标号，“w_i，w_j”是该 biterm 中的两个词。

第一步：对每个答复意见 Z 得到特定主题词分布$\varphi_z \sim Dir(\beta)$

第二步：得到全部集合的一个主题分布$\theta \sim Dir(\alpha)$

第三步:对集合 B 中的每个词 b：

（a）得到一个主题分布$z \sim Multi(\theta)$

（b）得到两个词的分布：w_i和$w_j \sim Multi(\varphi_z)$

按照上面的生成过程，一个词对$b=(w_i,w_j)$的联合概率表示如下：

$$P(b)=\sum_z P(z)P(w_i \mid z)P(w_j \mid z)=\sum_z \theta_z \varphi_i \left| z\varphi_j \right|$$

经过以上步骤，就可以将答复意见短文本的词向量空间映射为主题向量空间。

利用 BTM 模型[7]可以做到对答复意见的筛选作用，通过对关键词的处理判断答复意见的完整程度。接下来取部分答复意见作为训练集，结合词向量特征的循环神经网络语言模型方法，通过引入词关联信息的方法解决模型中存在长距离信息的学习问题，通过对词的分布式表达巧妙解决数据稀疏对统计建模的影响，同时克服模型参数的维数灾难问题来改善语言模型的性能。

循环神经网络的网络结构包含输入层、隐含层和输出层 3 个部分，输入向量

w 代表 t 时刻时输入词，维数与词汇表大小相同。隐含层 $s(t)$ 为网络的状态，表示 t 时刻时答复意见的信息，输出向量 $y(t)$ 表示待预测词在词汇表上的概率分布 U 、V 、W 分别为各层之间的权值矩阵。

在 t 时刻，网络的输入由当前词向量 $w(t)$ 和前一时刻的隐含层输出 $s(t-1)$ 构成，联合计算下一个隐含层，通过隐含层循环的方式，可利用更长的上下文答复意见，更好地表示自然语言。

在输入层中增加一个特征层，并通过权值矩阵 F 、G 分别与隐含层和输出层相连，此时网络的输入变为

$$x(t) = w(t) + s(t-1) + f(t)$$

在特征层中，输入向量 $f(t)$ 为词的上下文相关向量，其中包含更多的长距离信息，是对输入向量 $w(t)$ 的一个补充，可使词概率的计算更准确。采用随机梯度下降法训练网络，此时输出向量 $y(t)$ 表示待预测词在给定当前词 $w(t)$ 、上下文 $s(t-1)$ 和特征向量 $f(t)$ 下的概率分布，计算公式为：

$$s(t) = f(Uw(t) + Ws(t-1) + Ff(t))$$

$$y(t) = g(Vs(t) + Gf(t))$$

循环神经网络的特点是每层都有时间反馈循环,并且层与层之间是叠加构成的。每次神经网络的更新，新信息通过层次传递,每层神经网络也获得了时间性上下文信息。

将公式和答复意见训练集的内容按上述流程结合循环神经网络进行机器学习，利用循环神经网络强大的处理各种输入和输出类型的能力，可以精准快速的用于对答复意见内容序列的处理,即使语句的结构发生颠倒也可以很好地学习训练集当中的先前知识，来理解这些数据并作出相应的判断。将 BTM 模型与该方法进行结合，可以对每层的答复意见数据进行记忆，而不是分层次处理。通过每次迭代,新的信息被添加到每一层中,RNN 可以通过无限制次数网络更新将信息传递下去,使得 RNN 可以获得无限的记忆深度。使机器产生强大的自主分析效果，有效的对测试集进行处理来确定每一个答复意见内容的“完整性”。

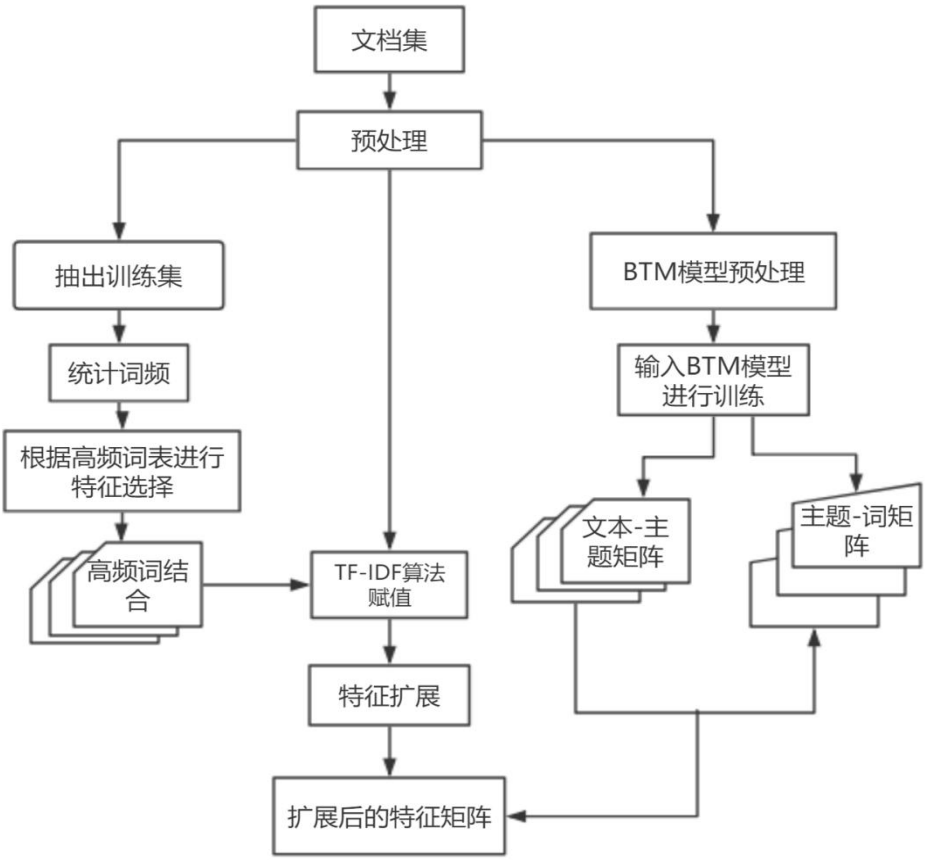

图 10

5.3 可解释性评价

在对可解释性分析评价的过程中，主要方法为前文相似性与完整性所用方法的结合。

首先，可解释性的一个方面包括了回复中是否有相关文件的引用[6]，对于民众提出的各类问题，政府文件或法律政策等具有权威说服力的文件的引用，在回复中会更容易令人理解、信服，从而可解释性大大提高。对此，在解决完整性的流程中，已经介绍过语料库的建立、BTM 模型的使用以及循环神经网络的学习，同样，在本部分，依然借助上述操作，由于相关文件引用时必定会出现书名号或引号，此时，可以将书名号、引号导入语料库，同时，具体数字、年份等同步导入，再进行上述一系列流程，实现可解释性的初步衡量。

操作流程如下：

（1）在语料库中导入书名号、括号、引号、数字、年份等与“可解释性”有关的符号或文字。

（2）将新的语料库与 BTM 模型相互结合，使其以语料库当中的指标为条件对留言答复意见进行判断筛选。

（3）在附件四给出的留言答复意见中选取部分数据作为训练数据导入模型，与上述循环神经网络结合在一起，使机器具有精准的自主判断性。

（4）将全部数据导入体系，可以完成对答复意见“可解释性”的评价。

接下来，运用自然语言处理技术，也即相似性[9]处理时的计算，算出结果后，相似性越高，对于民众理解越有利，可解释性自然提高。例如，在某条留言详情中网友写到“希望潇楚一卡通尽快支持手机 NFC 虚拟公交卡。现在钱包不用带了，钥匙不用带了，但是居然还要带公交卡。小米、华为、苹果等手机都无法开通 NFC 接触刷卡的 A 市虚拟公交卡，希望公交卡发行公司和各手机企业做好对接，尽快支持此功能。为市民提供便利。”回复中写到：“网友“UU0081681”您好！您的留言已收悉。现将有关情况回复如下：目前，潇楚支付公司正在进行“潇楚一卡通”APP 及 NFC 虚拟潇楚卡的开发工作，上线计划已在部署，具体上线时间请关注潇洒支付公司官网、“潇楚一卡通”微信公众号。若您需要了解更多潇洒卡其他内容，可致电 0731-0000-00000000。感谢您对我们工作的支持、理解与监督！2019 年 1 月 3 日”。可以直观地看到，回复与留言中相似的关键词不在少数，相比于回复中只写出“您好，信已收悉”，明显前者可解释性强于后者,符合模型构建的需要。

六、模型评价及优化

6.1 拉普拉斯平滑处理

高斯朴素贝叶斯算法作为朴素贝叶斯算法的一种，在处理具有不同属性特点的数据集时也能保持稳定的分类性能，而不用考虑各个属性间的关联，具有简单、高效优势。

但是朴素贝叶斯算法存在的不足同样存在于高斯朴素贝叶斯算法。在实际应用中基本上不可能满足其属性条件独立性的假设，对那些属性间存在高度相关性的数据，如果直接进行处理，分类效果很难达到实际预期。另外，在处理留言问题中也存在不平衡数据，少量的训练集很难涵盖所有的留言类型从而对测试集造成影响，对此可以结合拉普拉斯平滑处理，当样本增大时，对每个分量的计数加 1 造成的估计概率变化可以忽略不计，但可以方便有效的避免在训练时因文本训练不全面的问题。

6.2 回复可解释性的大众认知

由于可解释性与人的认知能力相关，因此不同文化程度的民众理解起来存在一定差距，改进中可以将可解释性与认知心理学结合[6]，根据认知心理学的一些

文献提出的原型模型和范例模型对人的分类行为进行建模,反映了人的分类学习过程，是容易被人所理解的，并可以参考选择注意理论对回复的可解释性进行全面评价。

七、参考文献

[1] 刘云芳.自然语言处理中相关语义技术的研究.西安交通大学.2011

[2] 阿曼.朴素贝叶斯分类算法的研究与应用.大连理工大学.2014

[3] 陈宝楼.K_Means 算法研究及在文本聚类中的应用.安徽大学.2013

[4] Tan P,Michael S,Vipin K.Introduction to Data Mining.Addison Wesley,US,2005

[5] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法. 计算机学报.2011

[6] 全文君.数据挖掘过程中的可解释性问题研究.重庆大学计算机学院.2018

[7] 沈磊.基于规则与机器学习方法的中文微博情感分析研究.2015

[8] 王子慕.一种利用 TF-IDF 方法结合词汇语义信息的文本相似度量方法研究.吉林大学.2015

[9] 王春柳，杨永辉，邓霏.文本相似度计算方法研究综述.2018