基于自然语言处理和文本挖掘的智慧政务应用

摘要

近年来，网络问政平台、成为政府了解民意、汇聚民智、凝聚民气的重要渠道，关于社情民意的留言文本数据量一直处于攀升状态，建立基于自然语言处理技术的智慧政务系统是社会治理发展的一种创新，对提升政府的管理水平和施政效率有很大的推进作用。

针对问题一，本文将留言文本进行数据预处理，中文分词和词频统计后利用 $TF-IDF$ 算法生成特征项集合，进行编码标注，确定词义相似度计算方法，用基于 LDA 模型的 K-means 文本聚类的方法，得出了一个更加合理的一级标签分类模型（城乡建设、交通运输与旅游业、资源管理、教育文体与体育、社会保障、卫生计生、政法与民法共七类）。最后采用 F-Score 对分类方法进行评价，结果良好。

针对问题二，本文运用时间序列图进行分析。在建模之前，对于噪声数据的处理时采用随机森林理论对数据进行平滑，并对数据进行转换和消减。根据图像得出 2010 年 2 至 2011 年 1 月存在退休老人工资少发等热点问题。

针对问题三，本文利用层次分析法来求出评价指标的权重，得到答复的快速性、相关性、完整性、可解释性的权重分别为 0.324，0.426，0.1124，0.1376，且通过一致性检验。

关键词：LDA 模型；$TF-IDF$ 算法；K-means 算法；随机森林理论；层次分析法

Summary

In recent years, network stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, based on natural language processing technology the wisdom of the e-government system has is the new trend of development of social management innovation, to enhance the management level of government and governance efficiency has a great role in promoting.

Aiming at problem one, this article will leave a message text data preprocessing, Chinese word segmentation and word frequency statistics, the TF - IDF algorithm generates feature collection, coding, determine the semantic similarity calculation method, use K - means based on the LDA model text clustering method, it is concluded that a more reasonable level 1 label classification model (urban and rural construction, resource management, transportation and tourism, political science and law and civil law, the education style and physical education, social security, the health to family planning, a total of seven classes).Finally, f-score is used to evaluate the classification method, and the result is good.

Aiming at problem two, this paper USES time series graph to analyze.Before modeling, the random forest theory is used to smooth the noise data, and transform and subdue the data.According to the figure, from February 2010 to January 2011, there were some hot issues such as the underpayment of the salary of the retirees.

For question 3, this paper USES analytic hierarchy process (ahp) to analyze the weight of the evaluation index, and obtains the weight of the rapidity, relevance, completeness and interpretability of the response as 0.324, 0.426, 0.1124 and 0.1376, respectively, and passes the consistency test.

Key words: LDA model；*TF – IDF* algorithm；K - means algorithm；Random forest theory；Analytic hierarchy process

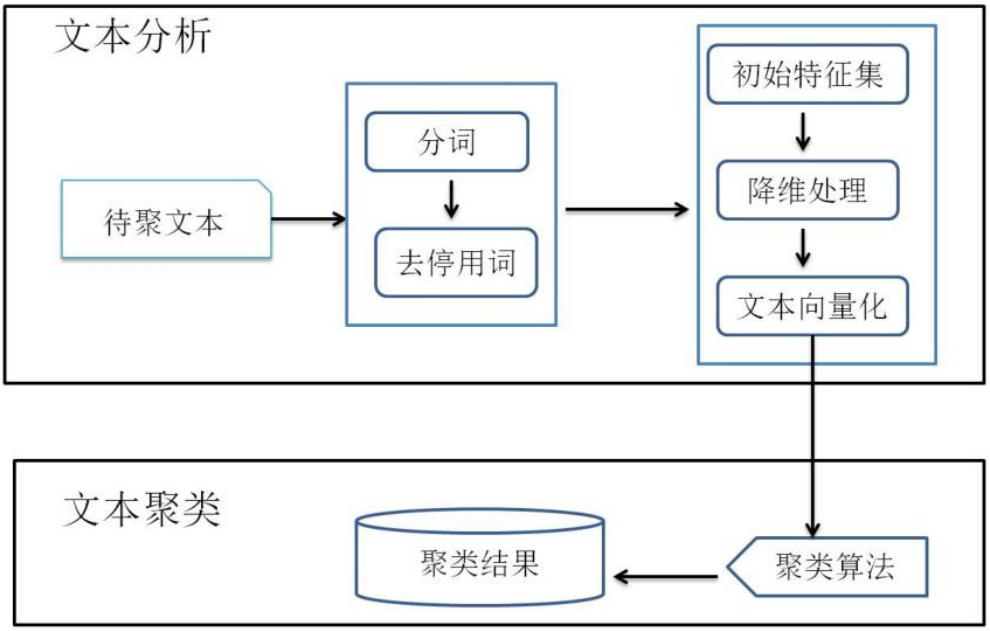
目录

1. 挖掘目标

本文建模目标是利用一万五千多条网络问政平台的群众留言数据，利用中文分词、去停用词等数据预处理方式后，通过 TF-IDF 特征提取文本向量化、建立 LDA 主题模型、K-means 文本聚类等多种自然语言处理和挖掘方法得到群众留言分类指标和热点问题的筛选。

2. 分析方法与过程

2.1 总体流程



2.2 具体步骤

(1)步骤一：数据预处理，对留言内容去除重复项及空行、jieba 中文分词、停用词过滤。

(2)步骤二：文本向量化，生成特征项集合，编码标注，确定词义相似度计算方法。

(3)步骤三：主题分析后文本聚类，根据文本向量，计算文档间的欧式距离，再基于 k-means 聚类算法对留言进行基于划分的聚类。

(4)步骤四：建立评价模型。

3. 文本分析

3.1. 数据预处理

3.1.1 对留言信息进行中文分词

本文数据预处理的第一步就是分词，在对群众留言信息进行分析处理时，由于原始文档中的自然语言是计算机无法识别的非结构化数据，所以要将中文文本信息进行转化处理以便于机器识别。在文档中词语是能够体现文本内容的最小成分，后续要对文本进行聚类需要构造以词为基础的空间向量，就要先将大量的文本内容分解成一个个的词语。在附件中的群众留言表中，我们先要对这些留言信息进行中文分词以便于数据转换。此处采用了 *Python* 的中文分词包 *jieba* 进行分词。 *jieba* 的原理是基于前缀词典实现的高效词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图, 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，对于未登录词，采用了基于汉字成词能力的 *HMM* 模型,以更好实现中文分词效果。

3.1.2 停用词过滤

在经过中文分词处理后，我们就将初始文本处理成了词的集合，但其中仍然含有对文本内容分析没有多大贡献的词，即无意义词语，将其删除可以减少噪音数据以提高文本挖掘的效率，此类词即为停用词，例如中文中的“啊”、“了”、“的”、“地”等等。停用词既可能是表达情绪的功能词，也可能是对文本信息表达没有价值从而难以区分的实际具体词语。原始文本中的停用词介入可能会造成我们选出的特征词会包含大量停用词，这将影响到最终分析结果。因此本文利用现有的哈工大停用词表对文本数据进行停词匹配，若分词结果与停用词表中词语出现重合，则进行删除过滤。

3.2 特征提取与特征权重计算

特征词对文本权重的计算方法即给每个文本特征词赋予数值以计算它的重要程度,高频词特征权重高,低频词特征权重低,权重即为特征词对于文本的重要作用程度。对特征的过滤权重提取就是通过一些复杂的方法将一些特征权重低的特征词给文本过滤掉,抽取一些特征权重高的特征词语组成一个主要用于后续研究实现的文本特征集,这种过滤方法又被称为特征降维。

3.2.1TF-IDF 算法

$TF-IDF\left(Term\ Frequency-Inverse\ Document\ Frequen\ cy\right)$是用来体现词语重要性的文本信息挖掘的一种流行的加权算法。其中 TF 为词频，IDF 为反文档频数。词的词频越高意味着它对文档的贡献程度越大越重要。反文档频数说明了特征词在全部文本中的分布情况，用数理统计学的思想进行表达,就是在词频的基础上对各个语料进行分配词语的重要性和权重,其大小与一个语料词的常见使用程度成很大的正反比。一个词语的 IDF 越大,则说明该词对于文本的准确区分能力越强。明白了词频（TF）和反文档频数（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。

在对留言分词去重后，由于需要把群众留言信息转换为权重向量，以供挖掘分析使用，以下是$TF-IDF$算法的原理与细节：

第一步：计算词频，及 TF 权重。

$$词频(TF)=某个词在文章中的出现次数$$

考虑到留言内容有长有短，为了便于不同留言的比较，进行“词频”标准化。

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

或者

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

第二步：计算 IDF 权重，即反文档频数。

这时需要创建一个语料库来模拟语言环境的使用。一个词语越常见则其分母就越大，反文档频数就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即全部文本都不包含该词）。*log* 表示对得到的值取对数。

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

第三步，计算$TF-IDF$值。

$$\text{TF-IDF} = \text{词频(TF)} \times \text{反文档频数(IDF)}$$

明白了“词频”（TF）和 “反文档频数”（IDF）以后，将这两个值相乘，就得到了一个词的$TF-IDF$值。某个词语的$TF-IDF$值越大则意味着该词对文本的重要程度越高。可以看到，$TF-IDF$与一个词在文本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。所以计算留言中每个词的$TF-IDF$值，进行降序排序，次数最多的即排在最前面的是要提取的留言中文本的关键词。

3.2.2 生成$TF-IDF$向量

生成$TF-IDF$向量的具体步骤如下:

(1)使用$TF-IDF$算法，找出每个留言描述的前 5 个关键词;

(2)对每个留言描述提取的 5 个关键词，合并成一个集合，计算每个留言描述对于这个集合中词的词频，如果没有则记为 0;

(3)生成各个留言信息的$TF-IDF$权重向量，计算公式如下:

$$\text{TF-IDF} = \text{词频(TF)} \times \text{反文档频数(IDF)}$$

3.3 文本的数值表示

除了特征项选择问题，文本-数值转换技术中另外一个重点和难点就是文本的数值表示，经过之前的特征项选择步骤，我们得到了特征项集合。之后，我们所面临的主要问题就是如何将特征项用机器能识别的方式表示。“数值表示”这一词语并不 能准确表达本次研究中这一步骤的具体内容，更准确的描述应该是“码表示”，不过为了能明确这一步骤在聚类分析中的地位，我们在这里仍然使用了传统的名称“数值表示”。

3.4 词云图可视化

本文利用 ROST 系列人文社研究大数据计算工具绘制出词云图如下，小区问题和住房问题频现。

4. 文本聚类

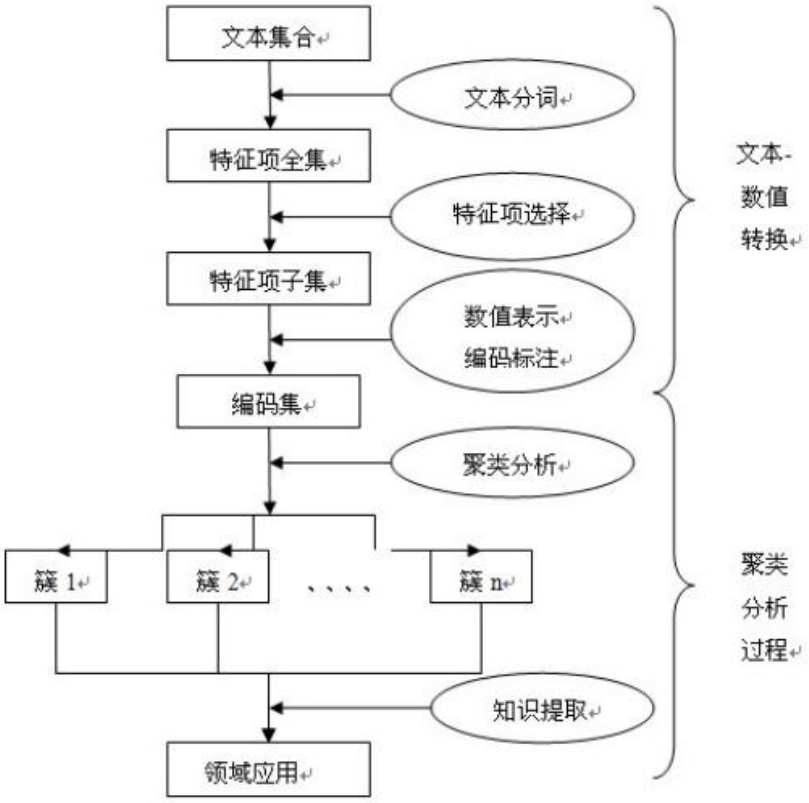


图 1 文本聚类过程图

4.1 群众留言的分类

生成群众留言的 TF-IDF 权重向量后，根据每个留言的权重向量，对留言进行分类。这里采用基于 LED 模型的文本聚类。用优化 K-means 算法把留言内容分成七类。

4.1.1 基于 LDA 模型的 K-means 文本聚类

本文将 LDA 主题模型应用于文本聚类领域，从统计学角度挖掘出文本集的内部语义信息，利用 LDA 模型生成的文本-潜在主题模型结合传统的 TF-IDF 特征词模型，将潜在主题知识融入特征词空间，进而提高文本聚类质量。

基于 LDA 模型的 K-means 文本聚类，是在 LDA 主题向量空间上，以文档在每一主题上的概率分布情况作为文档的特征值，然后采用 K-means 聚类算法，对文本特征向量聚类。

4.1.2 LDA 主题模型介绍

LDA(*Latent Dirichlet Allocation*)模型是 2003 年 *Blei* 提出的基于贝叶斯模型的生成式主题模型，挖掘文档中所隐含的主题信息，使用户或者读者快速的了解文档的信息。生成模型，即认为每一篇文档的每一个词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语"。

LDA 的主题特征分析模型有三层文本的结构，从上到下分别为文档、主题、词，实质就是通过利用数据集中文本的关注点和特征词的共同出现点和特征信息来有效挖掘数据集中文本的关注点和主题，能够有效对整个文本主题进行概率建模，和市场上传统的数据空间概率向量特征分析模型相比，增加了文本与概率的相关信息。通过利用 LDA 的主题特征分析模型，能够有效挖掘文本数据集中的潜在文本主题，进而有效分析文本数据集的集中关注点及其与文本相关的特征词。其层次非常的清晰，上中下依次是文档集层、主题层、和特征词层，其结构图如图所示：

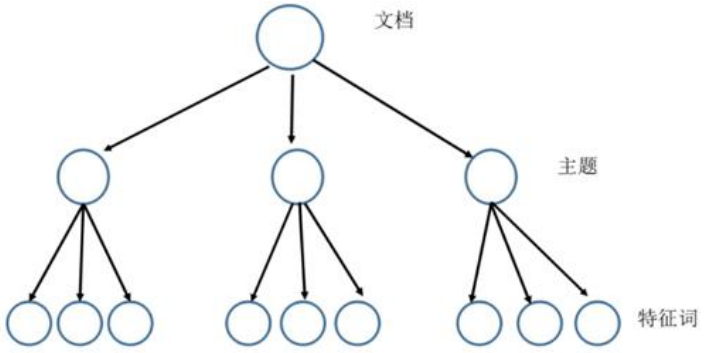


图 2LDA 层次结构示意图

LDA 的思想也可以用矩阵的形式通俗易懂表现出来，将整个文档看作是文档词条矩阵，可以分解成文档主题矩阵和主题词条矩阵，在下图中展示了三者之间的关系。

图 3 关系图

其中，“文档词项”矩阵表示每篇文档关于词项的概率分布，“文档主题”矩阵表示每篇文档关于主题的概率分布，“主题词项”矩阵表示每个主题关于词

项的概率分布。对于已知的文档，“文档词项”矩阵是可以根据前文所提到的 TF 或者 TF-IDF 得到。

LDA 模型使用词袋模型将每则文本视为一个词频向量，从而将文本信息转化为易于建模的数字信息。定义词表大小为 V，一个 V 维向量$(1,0,0,..,0,0)$表示一个词。由 N 个词构成的评论记为$d=(w_1,\ w_2,\ldots,\ w_N)$。假设某一群众留言集 D 由 M 篇留言构成，记为$D=(d_1,\ d_2,\ldots,\ d_M)$。M 篇留言分布着 K 个主题，记为$Z_i(i\ =1,2,\ldots,\ K)$。记α和β为 *Dirichlet* 函数的先验参数，θ为主题在文档中的多项分布的参数，其服从超参数为α的 *Dirichlet* 先验分布，ϕ为词在主题中的多项分布的参数，其服从超参数β的 *Dirichlet* 先验分布。LDA 模型图示见下图:

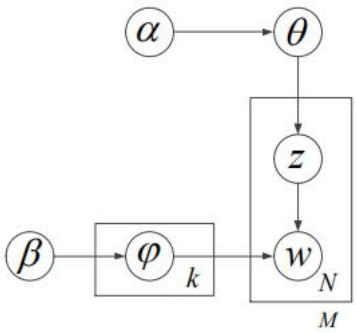


图 4 LDA 模型

LDA 模型假定每篇留言由各个主题按一定比例随机混合而成，混合比例服从多项分布，记为:

$$Z\,|\,\theta\ =\ Multinomial\,(\theta)$$

而每个主题由词汇表中的各个词语按一定比例混合而成，混合比例也服从多项分布，记为:

$$W\,|\,Z,\ \varphi\ =\ Multinomial(\phi)$$

在留言d_j 条件下生成词w_i 的概率表示为:

$$\mathrm{P}(w_i\,|\,d_j)=\sum_{s=1}^{K}\mathrm{P}(\mathrm{w}_i|\ \mathrm{z}=\mathrm{s})\times\mathrm{P}(\mathrm{z}=\mathrm{s}\,|\,\mathrm{d}_j)$$

其中，$\mathrm{P}(w_i\,|\,z=s)$ 表示词w_i属于第 s 个主题的概率，$\mathrm{P}(z=s\,|\,d_j)$ 表示第 s 个主题在留言d_j 中的概率。

4.1.3 LDA 主题模型估计

LDA 模型对参数θ、ϕ的近似估计通常使用马尔科夫链蒙特卡洛$\left(Markov\ Chain\ Monte\ Carlo,\ MCMC\right)$算法中的一个特例*Gibbs*抽样。利用*Gibbs*抽样对 LDA 模型进行参数估计，依据下式:

$$P(z_i = s \mid Z_{-i},\ \mathrm{W}) \propto (\mathrm{n}_{s,-i} + \beta_i) \times (n_{s,-j} + \alpha_s)$$

其中，$z_i = s$ 表示词w_i 属于第 s 个主题的概率，Z_{-i} 表示其他所有词的概率，$n_{s,-i}$ 表示不包含当前词w_i的被分配到当前主题z_s 下的个数，$n_{-j,s}$ 表示不包含当前文档d_j 的被分配到当前主题z_s 下的个数。

通过对上式的推导，可以推导得到词w_i在主题z_s中的分布的参数估计ϕ_{sj}，主题z_s在评论d_j 中的多项分布的参数估计$\theta_{j,s}$ ：

$$\phi_{s,j} = (n_{s,j} + \beta_i) \Big/ (\sum_{i=1}^{V} n_{s,i} + \beta_i)$$

$$\theta_{j,s} = (n_{j,s} + \alpha_s) / (\sum_{s=1}^{K} n_{j,s} + \alpha_s)$$

其中，$n_{s,j}$ 表示词w_i 在主题z_s 中出现的次数，$n_{j,s}$ 表示文档d_j 中包含主题z_s的个数。

LDA主题模型在文本聚类、主题数据挖掘、相似度分析计算等相关技术方面都已经有广泛的研究以及应用,相对于其他的主题模型,其引入了狄利克雷先验的知识。因此，模型的泛化分析能力较强，不易在其中出现过拟合概率现象。其次，它采用的是一种完全无需人工监督的训练模式，只需要提供一份用于训练的文档，它就已经可以自动通过训练计算出各种拟合概率，无需任何人工进行标注的过程，节省大量的人力及训练时间。

再者，LDA 词汇或主题的模型同样可以很好地解决多种特征词指代的问题。例如在本文附件二的媒体和群众留言中,根据了分词的一般使用规则，经过了分词的群众留言语句往往会将 “费用”一词单独地分割了出来，而“费用”一词无论是指医疗的费用，还是指建筑施工的费用等其他的情况，如果简单地对其进

行词频概率统计及其情感分析，是完全无法准确识别的，从而也可能无法准确地了解媒体和群众留言中所反映的具体情况。通过运用上述 LDA 词汇或主题的模型，可以准确地求得同一个词汇指代在同一主题群众留言中的概率分布，进而可以准确判断“费用”一词指代属于哪个词汇或主题，并准确地求得属于这一词汇或主题的群众留言概率和同一词汇或主题下的其他特征词，从而可以解决多种特征词指代的问题。

4.2 运用 LDA 模型进行主题分析的实现过程

4.2.1 随机森林理论

机器学习中，随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数来决定。 *Leo Breiman* 与 *Adele Cutler* 发展出推论出随机森林的算法， "*Random Forests*" 是其商标。随机森林理论术语是 1995 年由贝尔实验室的 *Tin Kam Ho* 所提出的随机决策森林$(random\ decision\ forests)$而来的。这个方法则是结合 *Breimans* 的 "*Bootstrap aggregating*" 想法和 *Tin Kam Ho* 的"*random subspace method*" 以建造决策树的集合。

根据下列算法而建造每棵树：

1. 用 N 来表示训练例子的个数，M 表示变量的数目。

2. 我们会被告知一个数 m ，被用来决定当在一个节点上做决定时，会使用到多少个变量。m 应小于 M

3. 从 N 个训练案例中以可重复取样的方式，取样 N 次，形成一组训练集（即 *bootstrap* 取样）。并使用这棵树来对剩余预测其类别，并评估其误差。

4. 对于每一个节点，随机选择 m 个基于此点上的变量。根据这 m 个变量，计算其最佳的分割方式。

5. 每棵树都会完整成长而不会剪枝$(Pruning)$（这有可能在建完一棵正常树状分类器后会被采用）

4.2.2 热点问题

在本次对群众留言热点问题的留言模型研究中，对群众留言问题集中的潜在

主题和浏览量进行了挖掘,其中的留言特征词浏览量是留言模型分析中的一个重要可观测变量。一般来说,每条群众留言问题都必定存在一个中心思想,即潜在主题。有时某个潜在的主题同时也可能是多条群众留言问题集中的潜在主题,即这一潜在的主题能方便地被认为是整个群众留言问题集中的一个热点问题。

首先，为了更加精确主题分析在不同方向的留言内容热点问题反映的情况，本文在已划分的分类结果上对不同部门问题的潜在主题进行了挖掘分析，并考虑了不同的情况。例如，选取一则留言“目前，在城管局等部门和各单位的支持和努力下，A 市在全国公交都市创建工作已经达标，但由于城市扩容和城市道路发展非常迅速，公交线路的优化变动同步加快，公交站房须随之变更设置，因此有如下工作亟需贵局完善：一、优化公交站房审批流程：1、公交公司持公交线路批文报告城管局。2、由城管局数字化信息中心核定具体公交站房设置位置。3、通报区城管局及所属城管队免采集数字化处罚和免拆。二、由于常规公交的公益性和亏损严重情况，明确对公交站房占用城市空地免收相关费用和免定期审批。专此建议，敬请支持为感。”在这则留言中，“城管局”和“公交”出现频率较高，可作为此文本潜在主题，同时还能得基于潜在主题特征词的概率分布情况，反映潜在主题“公交”的特征词包括“线路”、“站房”,联系“城管局”的特征词包括“核定”、“数字化”。

接着，分别统计整个留言语料库中具有特点的主题分布情况，对不同情况下，各个主题热点出现的次数从高到低进行排序，根据分析需要，选择排在前若干位的主题作为留言集中的热点问题，然后根据潜在主题上的特征词的概率分布情况，得到所对应的热门关注点的特定词。

4.2.2.1 基于划分的 K-means 聚类原理

聚类算法我们采用基于划分的 k-means 方法，为保证实验环境的公正性,我们人工拟定初始聚类中心。

基于划分的聚类算法的原理可以简单概括为数据集是众多零散的点，它们需要划分成几个类，最终聚类结果可以达到同一个类内的样本点足够近，不同类间的距离足够的远。首先需要根据划分目的设定出划分的个数 K，即需要聚成 K 个类，再随机的挑选 K 个点作为初始的类中心，得到初始的划分结果。然后在初始

划分结果的基础上迭代重置，重新寻找新的类中心，重新分配每个类的样本点，直到能够达到类内相似度高，类间相似度低的聚类原则为止。

具体 K-mean 聚类的原理如下:

假设有一个包含 n 个 d 维数据点的数据集 $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$，其中 $x_i \in R^d$，K-means 聚类将数据集 X 组织为 K 个划分 $C = \{c_k, i = 1,2,\ldots,K\}$。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，计算该类内个点到聚类中心 μ_i 的距离平方和

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

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，

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

其中，$d_{ki} = \begin{cases} 1, \text{若} x_i \in c_i \\ 0, \text{若} x_i \notin c_i \end{cases}$，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。

4.2.2.2 基于 LDE 模型的 K-means 算法总结

基于 LDA 模型的 K-means 算法步骤总结如下所示:

①初始化随机选取 K 个对象作为聚类的中心质点。

②遍历整个数据集的各个对象，计算 LDA 模型下的对象与中心质点的相似度，将相似的对象归类到一个簇中。

③替换中心质点，以每个簇的对象的均值中心点取代，然后进行迭代操作，直到满足终止条件为止，算法描述如下所示。

算法:基于 LDA 模型的 K-means 算法
输入:初始化聚类的个数 K，包含 n 个文本的数据集 D
输出:分为 K 个集群的数据集元素
1:从数据集 D 中人工指定 K 个聚类中心点($s_1, s_2, s_3, \cdots, s_k$)
2:重复
3:计算相似度，计算剩下的对象到这 K 个中心点的相似度，记做 d,然后更新 d_{min},然后将每个对象归类到值最小的中心点所在的簇

4:对上个步骤得到的新的簇进行计算，求出新的簇的中心点
5:直到准则函数不再发生变化
6:终止，聚类结束

4.2.2.3 聚类之后的时间序列图

聚类结束，以时间为横坐标，事件数值为纵坐标建立时间序列图，如下。

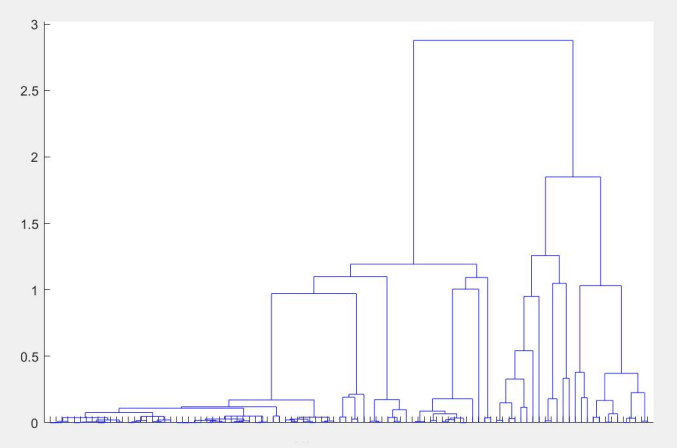


图 5 时间序列图

4.3 建立评价指标体系

针对附件四中相关部门对群众留言的答复意见，我们需要对答复意见的质量给出一套评价方案。评价指标体系的构建是关键，是第一步需要解决的问题。对于意见答复评价指标的选择问题，我们根据评价指标体系筛选原则筛选了一些选址评价指标，如下：

1) 答复快速性；
2) 相关性；
3) 可解释性；
4) 完整性。

首先先构造两两比较判断矩阵 P：

$$P=\begin{pmatrix} 1 & \frac{1}{2} & 3 & 5 \\ 2 & 1 & 3 & 4 \\ \frac{1}{3} & \frac{1}{3} & 1 & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{4} & 5 & 1 \end{pmatrix}$$

表 1 矩阵的比例标度及含义

标度	含义
1	表示因素 u_i 与 u_j 比较，具有同等重要性
3	表示因素 u_i 与 u_j 比较，u_i 比 u_j 稍微重要
5	表示因素 u_i 与 u_j 比较，u_i 比 u_j 明显重要
7	表示因素 u_i 与 u_j 比较，u_i 比 u_j 强烈重要
9	表示因素 u_i 与 u_j 比较，u_i 比 u_j 极端重要
2，4，6，8	表示分别介于 1~3，3~5，5~7，7~9 的中值
倒数	由因素 u_j 与 u_i 比较，$a_{ij}=1/a_{ji}$

根据判断矩阵 P，求出最大特征方根 $\lambda_{\max}$ 所对应的特征向量 ω，所求的特征向量即为各评价因子的权重分配，对其进行归一化，得到准则层对目标层的权向量为

$$\omega = (0.324, 0.426, 0.1124, 0.1376)^T$$

然后检验权重分配是否合理，这里对判断矩阵进行一致性检验，检验使用公式为：

$$CR = \frac{CI}{RI} \tag{1}$$

式中：CR 为判断矩阵的随机一致性比率;CI 为判断矩阵的一般一致性指；RI 为判断矩阵的平均随机一致性指标。

它由下式给出：

$$CI = \frac{\lambda_i - n}{n-1} \tag{2}$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n} A_{ij}\cdot\omega_j}{\omega_i} \tag{3}$$

$$RI = \frac{1}{m}\sum_{k=1}^{m} CI_k \tag{4}$$

接下来计算平均随机一致性，进行矩阵一致性检验，参考文献得到 1 阶到 16 阶重复计算 1000 次的平均随机一致性指标如下：

表 3 平均随机一致性指标 RI 值

1	2	3	4	5	6	7	8	9	10
0	0	0.5104	0.8845	1.1108	1.2493	1.3416	1.4046	1.453	1.4851
11	**12**	**13**	**14**	**15**	**16**				
1.5152	1.536	1.5542	1.57	1.5851	1.5947				

最后计算一致性比率为得一致性比率 CR=0.0619<0.1。故层次总排序通过一致性检验。

4.4 结果评价

4.4.1 F 值评价法

查准率$(precision)$定义如下，表示被分为正例的示例中实际为正例的比例。

$$P=\frac{TP}{TP+FP}=\frac{TP}{P'}$$

查全率(召回率)$(recall)$是覆盖面的度量，度量有多个正例被分为正例，

$recall=TP/(TP+FN)=TP/P=sensitive$，可以看到召回率与灵敏度是一样的。

$$R=\frac{TP}{TP+FN}=\frac{TP}{P}$$

P 和 R 指标有时候会出现的矛盾的情况，这样就需要综合考虑他们，最常见的方法就是 $F-Measure$（又称为 $F-Score$）。$F-Measure$ 是 $Precision$ 和 $Recall$ 加权调和平均：

$$F=\frac{(a^2+1)P*R}{a^2(P+R)}$$

当参数 α=1 时，就是最常见的 F1，也即

$$F1=\frac{2*P*R}{P+R}$$

可知 F1 综合了 P 和 R 的结果，当 F1 较高时则能说明试验方法比较有效。

4.4.2 热度评价指标法

基于附件三的留言，在对某一时段内反应特定地点或特定人群问题的留言进行归类后，我们需要定义合理的热度评价指标。我们设计了关键词权重计算公式提取关键词，依据关键词归纳总结热点问题，结合群众点赞反对参与量（我们称之为浏览量）提出如下热度评价指标：

$$\begin{cases} p = \dfrac{s_i}{\sqrt{\sum s_i^2}}, s = \text{浏览量} \\ \text{热度=score} * p * 1000, \text{score=关键词权重} \end{cases}$$

我们认为留言的热度与从其文本提取出来的特征关键词权重成正比关系，并且浏览量也是一个影响热度的指标。关于关键词抽取的算法，目前主要有 TF-IDF 算法、初代 KEA 算法、*TextRank* 算法和 *ICTCLAS* 等。

本文研究采用的技术是一种基于 *ICTCLAS* 的 *Ansj* 群众留言关键词权重提取分析技术，其提取的原理主要为通过依据不同的词性来确定词语的初始权重，其中主要是将标题中每个词的权重加倍，再通过结合留言特征词在文中经常出现的时间位置和出现频率进行调整后，得到每个关键词的初始权重。通过结合本次群众留言的特征词浏览量可对留言关键词的权重作进一步的分析和改进，此处的浏览量相当于民众参与程度，可用留言表中的点赞数与反对数代表浏览量。

5. 结论

5.1 分类指标

我们根据附件二给出的数据通过聚类建立了关于留言内容的新一级标签分类模型，并结合国家统计局的分类标准，最终得出其分类指标有城乡建设、交通运输与旅游业、资源管理、教育文体与体育、社会保障、卫生计生、政法与民法七大类。分类图如下：

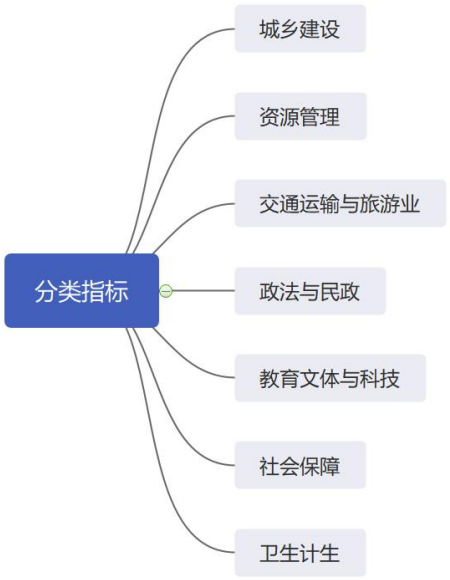


图 6 一级分类指标图

5.2 LDA 模型构造主题词结果分析

本文运用 LDA 主题模型的算法，并采用 *Gibbs* 抽样方法对其中参数进行近似估计，模型中的 *Dirichlet* 先验参数 α 和 β 、主题个数 T 这三个可变量是需要确定最佳取值的。我们使 K 采用统计语言模型中常用的评价标准困惑度 50，选取 $\alpha=50/T$，$\beta=0.1$。

因为附件中的现有主题留言模块占有很多问题具体地点等主题判断无关词，所以我们采用的是具体留言模块与主题相结合的文本信息进行分析计算特征权重。运用 LED 主题模型提取了群众留言文档集中的 7 个主题，利用聚类来进行总体划分。主题词列举如表所示：

城乡建设	资源管理	交通运输与旅游	政法与民政	教育文体与科技	劳动与社会保障	卫生计生
施工	环保	交通	政策	教育	保险	医疗
建设	环境	旅游	计划	校园	社保	医院
设施	污染	收费	管理	学生	工资	医务
规划	监测	出租车	法律	教师	福利	卫生
故障	排放	旅客	质量	文化	劳动	安全
街道	场地	运输	费用	补助	医保	药品
小区	生态	货物	优惠	工资	员工	医患
修理	破坏	规划	出台	运动	养老	计生
建筑	饮用	事故	罚款	培训	就业	违规

6. 参考文献

[1]Mohamed Atef Mosa. A novel hybrid particle swarm optimization and gravitational search algorithm for multi-objective optimization of text mining[J]. Elsevier B.V.,2020,90.

[2]Syaamantak Das,Shyamal Kumar Das Mandal,Anupam Basu. Mining multiple informational text structure from text data[J]. Elsevier B.V.,2020,167.

[3]余传明,王曼怡,林虹君,朱星宇,黄婷婷,安璐.基于深度学习的词汇表示模型对比研究[J/OL].数据分析与知识发现:1-19[2020-05-06].http://kns.cnki.net/kcms/detail/10.1478.G2.20200423.1336.012.html.

[4]包清临,柴华奇,赵嵩正,王吉林.采用机器学习算法的技术机会挖掘模型及应用[J/OL].上海交通大学学报:1-22[2020-05-06].https://doi.org/10.16183/j.cnki.jsjtu.2020.99.007.

[5]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.

[6]Hüseyin Fidan,Mehmet Erkan Yuksel. A Novel Short Text Clustering Model Based on Grey System Theory[J]. Springer Berlin Heidelberg,2020,45(1).

[7]王新新.面向模式的文本数据描述模型[J].科技创新与应用,2020(10):28-30.

[8]王惠.基于 LDA 主题模型的文本聚类研究[D].甘肃:兰州大学,2018.DOI:10.7666/d.D01449208.

[9]董婧灵. 基于 LDA 模型的文本聚类研究[D].华中师范大学,2012.

[10]李少温. 基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学,2019.

[11]涂慧明. 互联网舆情监控系统的关键技术研究与实现[D].东华理工大学,2016.

[12]吴柳,程恺,胡琪.基于文本挖掘的论坛热点问题时变分析[J].软件,2017,38(04):47-51.

[13]泰迪杯优秀论文 https://max.book118.com/html/2018/0331/159451850.shtm

[14]泰迪杯优秀论文 https://wenku.baidu.com/view/04c1a20ecdbff121dd36a32d7375a417866fc1c0?qq-pf-to=pcqq.group

[15]刘建平博客 https://www.cnblogs.com/pinard/p/6164214.html

[16]mAP 准确率（precision）召回率（recall） f-meansure 之间的区别 - CSDN博客
[17]陈俊宇,郑列.基于R语言的商品评论情感可视化分析[J].湖北工业大学学报,2020,35(01):110-113.
[18]成永坤. 滑雪游客的产品认知、情感表达及满意度研究——基于网络评价文本分析[C]. 中国体育科学学会.第十一届全国体育科学大会论文摘要汇编.中国体育科学学会:中国体育科学学会,2019:1261-1262.
[19]唐涵. 基于文本情感特征的信息隐藏及其分析[D].广州大学,2019.
[20]韩慧. 基于深度森林学习的评论文本情感分类研究[D].郑州大学,2019.
[21]王占忠.基于多文本数据的联通工单系统辅助研判技术[J].通讯世界,2020,27(02):73-75.
[22]王涛,李明.基于LDA模型与语义网络对评论文本挖掘研究[J].重庆工商大学学报(自然科学版),2019,36(04):9-16.
[23]马慧芳,刘文,李志欣,蔺想红.融合耦合距离区分度和强类别特征的短文本相似度计算方法[J].电子学报,2019,47(06):1331-1336.
[24]范宁. 基于文本挖掘在民宿满意度中的研究[D].广西师范大学,2019.
[25]童昱强. 基于数据挖掘的网络新闻热点发现系统设计与实现[D].北京邮电大学,2019.