智慧政务信息的分析与挖掘

摘要

近年来，伴随着互联网的快速兴起，网络问政平台逐渐成为了政府了解各类社情民意的主要渠道。因此，通过文本分析和数据挖掘技术对政务信息研究具有重大的意义。

对于问题一，通过 Excel 对政务信息进行分类处理，得到各类别的留言内容。利用 jieba 分词对留言内容进行分词，并通过停用词表对分词结果去无意义词汇。利用 TF-IDF 算法构建各类别留言内容的权重词向量空间。采用多项式贝叶斯算法建立 NB（朴素贝叶斯）分类模型。再对数据集进行训练分析，最终得到 F-Score 值为 0.81143。

对于问题二，通过利用 pynlpir 分词完成对留言内容的分词和词性标志，并筛选出分词文本中词性为人名和地名的词汇。再利用余弦距离计算文本相似度，得到相似度矩阵。最后采用 k-means 算法对文本进行聚类，得到热点问题表，完成热点问题的挖掘。

对于问题三，通过计算文本相似度作为答复质量评价的重要指标。利用 jieba 分词对答复、留言内容进行分词，并通过停用词表对分词结果去停用词。再利用余弦相似度和 TF-IDF 相似度算法进行文本相似度的计算。最后通过对比两种相似度计算方式的效果，综合评价答复内容质量。

关键词：TF-IDF 算法；朴素贝叶斯分类模型；文本相似度；K-means 文本聚类

目 录

1.问题背景

1.1 挖掘背景

在互联网发展迅速的时代，生活的方方面面都需要网络信息的参与，“互联网＋”也成为热门的词汇。近年来伴随着网络问政平台的兴起，电子政务[1]逐渐取代了传统纸质政务，政府了解民意、听取民生的方式从原来的走访闻讯工作台逐渐倾向从网络问政平台来听取意见。但如何从增长得越来越快的文本数据量中提取到真正有用的信息成为了新的挑战。因此，建立一个完善的基于自然语言处理技术的智慧政务系统，帮助政府相关部门提取信息，从而更加高效准确的做出部署和回应，解决群众问题，具有重大的意义。

1.2 建模目标

本次建模目标是利用网络问政平台收集到的政务数据，使用 jieba 分词对政务数据进行分词、使用停用词表对分词结果去除无意义词汇、使用 TF-IDF 算法[2-3]完成关键词的提取和权重词向量空间的建立[4]、利用多项式朴素贝叶斯算法[5]以及 K-means 聚类算法[6]来解决三个问题目标：

1.利用文本分词和文本分类的自然语言处理方法对政务数据进行一级标签分类，并使用 F-Score 对分类结果进行评价。

2.利用文本分词、词性标识、相似度计算、文本聚类的方法对政务数据进行热点问题的挖掘。通过词性标注提取出文本中的人名和地名，并对相似度高的文本数据并进行聚类，完成问题归并。

3.利用文本分词和文本相似度的计算，对政务数据中答复内容进行评价，通过答复内容和留言内容之间的相似度，判断答复内容的质量高低。

1.3 分析方法与过程

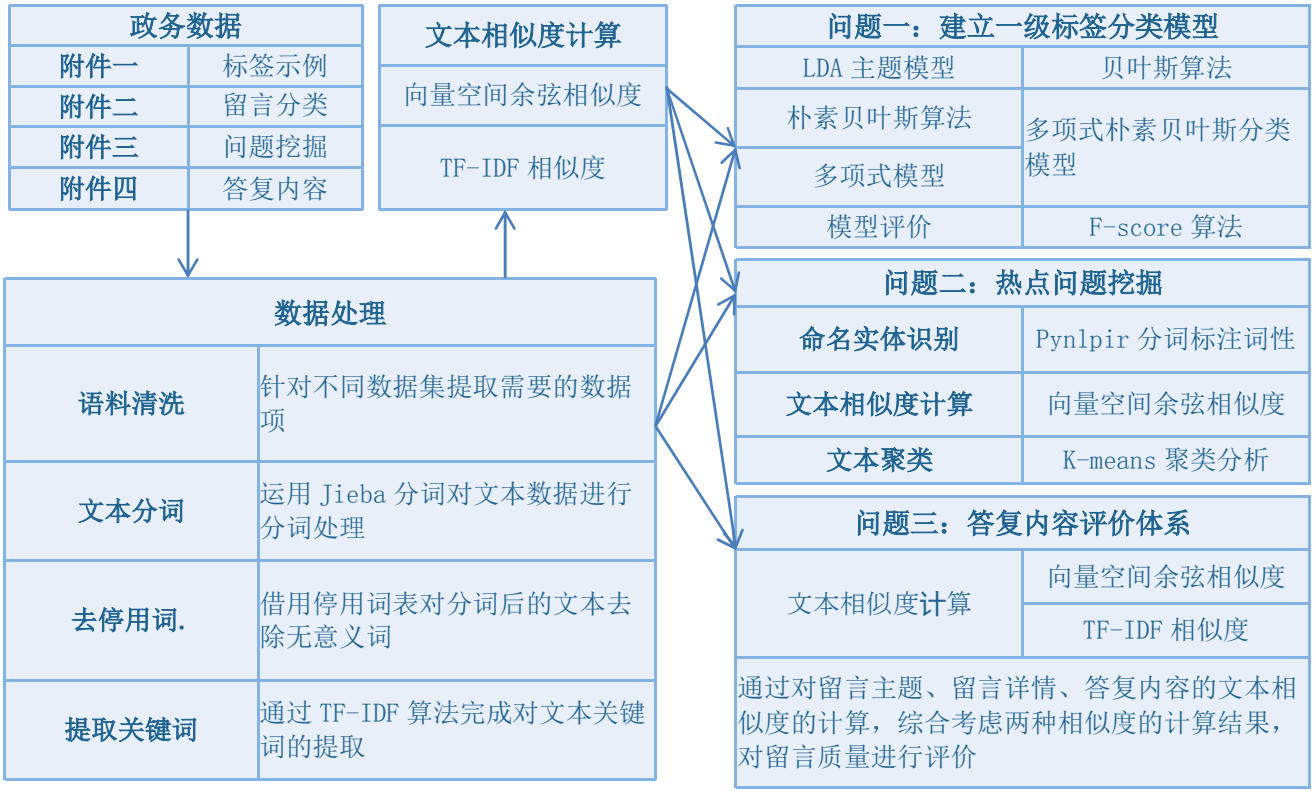

图 1 总体流程图

在挖掘过程中，首先对政务数据中原始的四个数据集进行可用数据的提取，分别为：附件一（标签示例）、附件二（留言详情）、附件三（留言详情）、附件四（留言主题、留言详情、答复内容），针对不同数据集要解决的问题来分析关键数据。提取出关键数据后，需要对数据进行如图 1 中数据处理的四个步骤，并对处理过后的数据进行文本相似的计算。至此，开始针对不同的问题进行相应的解题步骤，具体方法见图 1。

1.4 研究现状

自然语言处理是人工智能领域所研究的重要课题之一，同时也是目前最前沿的科技研究热点之一。在网络快速发展，信息逐渐碎片化，数据量逐渐变大的如今，越来越多的政府部门重视并利用新的互联网平台，强化宣传和互动效果。

智慧政务即通过“互联网+政务服务”构建智慧型政府，利用云计算、移动物联网、人工智能、数据挖掘、知识管理等技术，提高政府在办公、监管、服务、决策的智能水平，形成高效、敏捷、公开、便民的新型政府，实现由“电子政务”向“智慧政务”的转变。

在本篇论文中，通过研究自然语言处理技术中文本挖掘[7-10]中：分词、去停用词、文本特征选择、文本相似度计算、文本聚类、文本分类等技术来完成挖掘目标。

2.符号说明

参数	描述
n	特征维数
N	总的样本个数
k	总的类别个数
α	平滑系数
w_i	第i个单词
d_s	特定文档s
t_j	第i个单词对应的主题
N_{y_k}	类别为y_k的样本个数
N_{y_k,x_i}	类别为y_k的样本中，第i维特征的值是x_i的样本个数
P_i	第i类的精确率
R_i	第i类的召回率
C_i	第i类簇
$avg(C_i)$	第i类簇内样本的平均距离
$d_{\min}(C_i,C_j)$	簇C_i和簇C_j最近样本之间的距离
$d_{cen}(C_i,C_j)$	簇C_i和簇C_j中心点之间的距离
$diam(C_l)$	簇C_l内样本的最远距离

3.数据处理

3.1 分词与词性

在对中文语料进行自然语言处理时，介于中文语料一般为一批短文本或长文本。且一般段落、语句之间的字、词、符号都是连续的，而中文语句的含义跟随词语之间匹配方式的不同有不一样的含义。如果不对中文语料进行分词处理，将语句处理为最小单位词或词语的话，对语料进行数据的挖掘分析的难度将大幅度上升。

在分词中，常见的分词算法有：NLTK、Jieba、Hanlp、Pyltp 等分词算法，在本篇论文的代码实现中，主要使用 Jieba 分词来完成我们的分词处理。同时自然语言处理的经过不断发展、不断完善，在上述算法都自带有函数能在对语料分词的基础上，对分词结果进行了词性的标注。

3.2 去停用词

停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词。所以在一般性的文本处理中，分词之后，接下来一步就是去停用词。但是对于中文来说，去停用词操作不是一成不变的，停用词词典是根据具体场景来决定的，在不同的场景下词语所表达的情感也是不尽相同的。在本篇论文中使用的是哈工大的停用词表并根据政务数据中的文本语料进行了一定程度的修改。

4.问题解答

4.1 基于 TF-IDF 的关键词提取

在自然语言处理技术中，TF-IDF 通过对文本中词语权重计算，以权重为标准进行文本中词语的排序，权重越大词语在文本中的重要程度越强。在本论文中，采取 TF-IDF 算法对文本数据进行关键词的提取。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。其中 TF 代表词频，IDF 代表逆向文件概率。

1.TF：关键字在文本中出现的概率即词频，公式为

$$TF_{\mathrm{w}}=\frac{\text{在某一类中词条}w\text{出现的次数}}{\text{在该类中所有的词条数目}} \tag{1}$$

2.IDF：关键字出现的文件在总文件中的对数概率， 越大，则说明词条具有很好的类别区分能力，公式为

$$IDF=\log\left(\frac{\text{语料库的文档总数}}{\text{包含词条}w\text{的文档数}+1}\right) \tag{2}$$

3.TF-IDF：某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF，公式为

$$TF-IDF=TF*\ IDF \tag{3}$$

在对 tfidf 值计算的过程中，可以发现词频差异会对关键词的提取造成影响，考虑能否通过某种方式来提取结果，由此考虑引入互信息量及信息增益来优化了提取效果。得到 TF-IDF 算法的理论基础后，在自然语言处理技术中，可以通过引用 NLTK、Sklearn、jieba 等完善的第三方库来实现 TF-IDF 算法的 Python 代码实现，在论文中主要采用了 jieba 和 Sklearn 实现了 TF-IDF 算法，完成了对关键词的提取。

4.2 问题一 群众留言分类模型

4.2.1 思路构建

在对群众留言分类的过程中，重点对政务数据中附件 2 的“留言详情”数据进行了挖掘。在挖掘过程中，由于留言数据中无意义词汇偏多，且一些词语在分词结束后存在一些误差。我们预设了两种方式：

1.对留言数据建立 LDA 主题模型[11]并采用朴素贝叶斯对数据建立分类模型；

2.根据留言数据本身的类别标签将数据进行分类并将每个留言数据导出为一个单独的 TXT 文本数据，在对分解过后的数据文件夹进行遍历，建立起各类别的词频矩阵、词向量空间，并采用多项式贝叶斯算法来建立 NB（朴素贝叶斯）分类器。

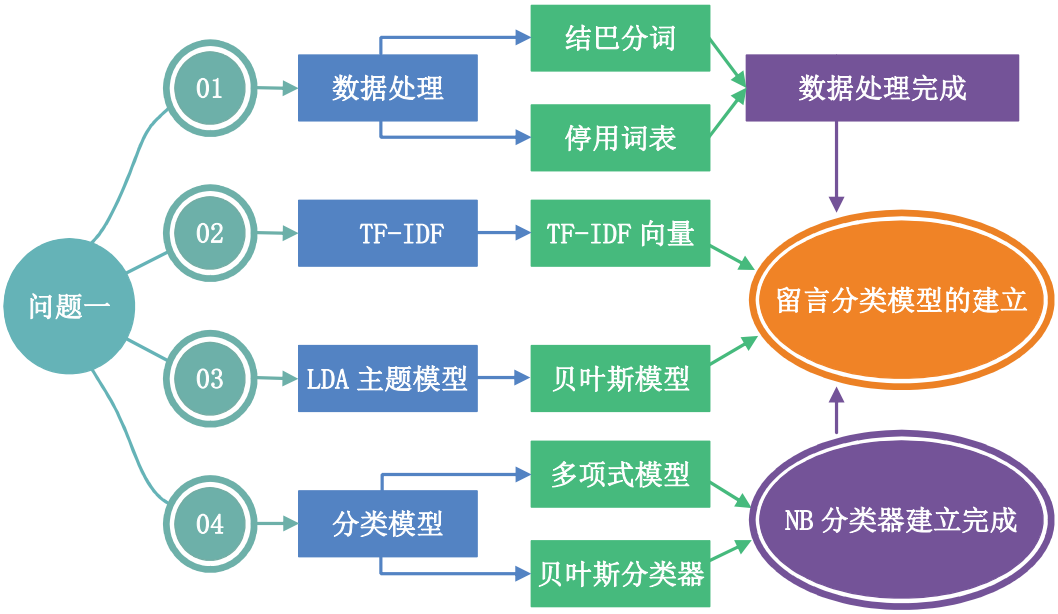

图 2 问题一思路流程图

4.2.2 LDA 主题模型

LDA 主题模型是一种无监督的贝叶斯模型，他可以通过基础的贝叶斯定理将文档集中的每篇文档按照概率分布的形式给出。LDA 的核心公式为：

$$P_j\left(w_i \middle| d_s\right) = P\left(w_i \middle| t_j\right) * P\left(t_j \middle| d_s\right) \quad (4)$$

式中，w_i 代表第 i 个单词，d_s 代表特定文档 s，t_j 代表第 i 个单词对应的主题。

4.2.3 基于多项式朴素贝叶斯算法的留言分类模型

朴素贝叶斯算法是一种简单的分类算法，在文本分类中适用范围比较广泛，朴素贝叶斯算法拥有三种较为常见的模型，分别为：多项式模型、高斯模型、伯努利模型。在这里我们使用多项式模型来进行分类模型的构建。

朴素贝叶斯算法的理论基础源于贝叶斯定理：

$$P\left(B|A\right)=\frac{P\left(A|B\right)P\left(B\right)}{P\left(A\right)} \tag{5}$$

$$P\left(A\right)=\sum_{i=1}^{n}P\left(B_i\right)P\left(A|B_i\right) \tag{6}$$

在贝叶斯定理的基础上，将样本由一维向量的变为n维的特征向量，即建立数据集$\left(X,Y\right)$，其中每个样本x都包含n维特征，即$x=\left(x_1,x_2,x_3,...,x_n\right),x\in X$，类标记集合含有$k$种类别，即$y=\left(y_1,y_2,y_3,...,y_k\right),y\in Y$，由此可以进行公式推导

$$P\left(y_k|x\right)=\frac{P\left(x_k|y\right)P\left(y_k\right)}{P\left(x\right)}=\frac{P\left(x_k|y\right)P\left(y_k\right)}{\sum_k P\left(x_k|y\right)P\left(y_k\right)} \tag{7}$$

$$P\left(x|y_k\right)=P\left(x_1,x_2,x_3,...,x_n|y_k\right)=\prod_{i=1}^{n}P\left(x_i|y_k\right) \tag{8}$$

将公式（8）代入到公式（7）中可以得到朴素贝叶斯分类器表达式

$$f\left(x\right)=\arg\max_{y_k}P\left(y_k|x\right)=\arg\max_{y_k}\frac{P\left(y_k\right)\prod_{i=1}^{n}P\left(x_i|y_k\right)}{\sum_k P\left(y_k\right)\prod_{i=1}^{n}P\left(x_i|y_k\right)} \tag{9}$$

$$f\left(x\right)=arg\max P\left(y_{\mathrm{k}}\right)\prod_{i=1}^{n}P\left(x_i|y_k\right) \tag{10}$$

在得到朴素贝叶斯分类器表达式公式（10）后，通过多项式模型来计算先验概率$P\left(yk\right)$和条件概率$P\left(x_i|y_k\right)$，由此可以得到多项式贝叶斯模型的计算公式

$$P\left(y_k\right)=\frac{N_{y_k}+\alpha}{N+k\alpha} \tag{11}$$

$$P\left(x_i|y_k\right)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n\alpha} \tag{12}$$

式中，n是特征维数，N是总的样本个数，k是总的类别个数，α是平滑系数，N_{y_k}是类别为y_k的样本个数，N_{y_k,x_i}是类别为y_k的样本中，第i维特征的值是x_i的样本个数。

4.2.4 数据分析

在已知理论基础上，通过 Python 编程语言完成了对理论模型的编译，在编译过程中引用了第三方库 gensim 完成了对 LDA 主题模型的搭建，通过第三方库 sklearn 完成了对多项式朴素贝叶斯分类器的搭建。

（1）LDA 主题模型

在对数据进行挖掘的过程中，针对第一种方法首先对数据集进行了处理，对文本完成了分词、去停用词、抽取关键词等过滤操作。在对处理好的数据进行 LDA 主题模型建立时，发现由于数据集样本数不大且其中“留言详情”的文本长度不稳定各有差别，导致在分词和提取文本关键词时，存在一些误差。

表 1 LDA 主题模型运行结果

	主题词汇									
主题一	医院	公司	工作	情况	生育	政策	人员	学校	职工	办理
	0.008	0.006	0.005	0.005	0.004	0.004	0.004	0.004	0.003	0.003
主题二	医院	患者	医生	学校	孩子	工作	医疗	希望	部门	公司
	0.011	0.009	0.008	0.006	0.005	0.004	0.004	0.004	0.003	0.003
主题三	医保	医师	病人	药	医生	手术	药物	医院	用药	银行
	0.019	0.018	0.013	0.012	0.01	0.006	0.005	0.005	0.004	0.003
主题四	屠宰场	药房	景区	居民	部门	项目	电梯	D5	旅游	建设
	0.008	0.007	0.006	0.005	0.004	0.003	0.003	0.003	0.003	0.003
主题五	疫苗	补课	退休金	公司	出租车	业主	老百姓	钱	职工	独生子女证
	0.009	0.009	0.005	0.004	0.004	0.004	0.004	0.003	0.003	0.003
主题六	医院	政策	社保	工作	退休	工资	人员	部门	享受	户口
	0.011	0.008	0.007	0.006	0.005	0.004	0.004	0.004	0.004	0.004
主题七	电梯	业主	旅游	部门	文化	物业	希望	开发商	投诉	发展
	0.016	0.008	0.007	0.006	0.004	0.004	0.003	0.003	0.003	0.003

从表中可以看到，通过 LDA 主题模型建立的主题和原主题存在较大的偏差。运行结果中预测的 7 个主题中有 6 个主题中存在和卫生计生相关的词汇，有 3 个主题中存在教育文体相关的词汇，有 2 个主题中存在商贸旅游相关的词汇，有 2 个主题中存在劳动和社会保障先关的词汇，而与交通运输、环境保护、城乡建设相关的词汇几乎没有在 7 个预测主题中出现。且在数据处理的过程中，通过对关键词提取数量的控制也无法很有效的减少这种偏差，由此选择采用第二种方式进行建模运算。

（2）多项式朴素贝叶斯分类算法

第二种方法在对数据集的处理上相比第一问有了一定的不同，根据政务数据中附件 2 数据中每个样本自带的一级标签，通过 Excel 对样本进行了分类。将每个类别的样本集合在一起，同时通过 Python 代码的实现，将每个样本中的“留言详情”文本，从 EXCEL 表格中提取出来转换为 TXT 文本格式。则原始附件 2 数据集从 Excel 格式数据转变为树形文件夹的形式，对此通过对文件夹的遍历来完成我们对于数据的挖掘。

在处理好数据后，对数据进行了分词、去停用词。相比第一种方法，对处理好的分词数据通过 TF-IDF 权重向量构建出了词向量空间，完善了词库量，增加了分类的准确性。

在建立模型的过程中，为增加分类的准确性，将训练集指定为附件 2 的完整数据，测试集指定为附件 2 的示例数据。运行代码块，对数据进行分析，可以得到分类结果表 2。

表 2 多项式朴素贝叶斯分类器分类结果

类别	测试集样本数量	分类效果	
		成功	失败
交通运输	55	32	23
卫生计生	58	41	17
商贸旅游	48	38	10
教育文体	96	85	11
城乡建设	101	92	9
环境保护	33	30	3
劳动和社会保障	104	93	21

观察表 2，可以看到在各类别对应的基数下，分类结果均较为理想。由此使用 F-Score 算法来对分类结果进行评价。F-Score 算法公式如下：

$$\mathrm{Pr}ecision\left(精确率\right)=\frac{TP}{TP+FP} \tag{13}$$

$$\mathrm{Re}call\left(召回率\right)=\frac{TP}{TP+FN} \tag{14}$$

$$F_1=\frac{2\,\mathrm{Pr}ecision\times\mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call}=\frac{2TP}{2TP+FP+FN} \tag{15}$$

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{16}$$

式中，TP 为真阳性因子，TN 为真阴性因子，FP 为伪阳性因子，FN 为伪阴性因子，P_i 为第 i 类的精确率，R_i 为第 i 类的召回率。将表中数据代入到公式（16）中可以得到最终 F-score 值为：0.81143。

4.3 问题二 热点问题挖掘

4.3.1 思路构建

针对热点问题挖掘，首先对数据进行基本的数据处理环节：分词、去停用词，再通过 pynlpir 分词对数据进行词性标识将数据中的地址、人物名词筛选出来。同时根据分词结果对数据集进行 K-means 文本聚类完成对热点问题的挖掘。

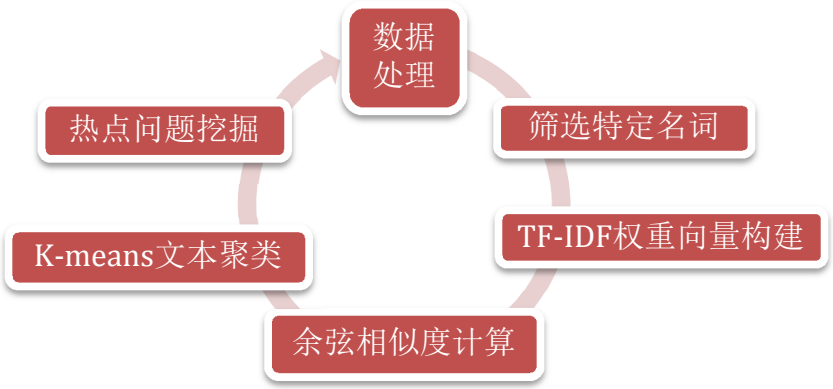

图 3 问题二思路流程图

4.3.2 基于 K-means 聚类分析的热点问题归并

在对相似文本进行聚类归并的过程中，采用了 K-means 聚类算法，K-means 聚类算法的思想比较简单假设将数据分成 k 个类，大概可以分为以下几个步骤：

1.随机选取 k 个点，作为聚类中心；

2.计算每个点分别到 k 个聚类中心的聚类，然后将该点分到最近的聚类中心，行成了 k 个簇；

3.再重新计算每个簇的质心（均值）；

4.重复以上 2~4 步，直到质心的位置不再发生变化或者达到设定的迭代次数。

针对问题要求，在对热点问题进行聚类时无法提前的得知存在多少类的热点问题，即就是说在聚类的过程中 k 值是无法确定的，于是需要对聚类性能度量内部指标的计算。

1.DB 指数（Davies-Bouldin Index）

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg\left(C_i\right)+avg\left(C_j\right)}{d_{cen}\left(C_i,C_j\right)}\right) \tag{17}$$

2.Dumn 指数（Dumn Index）

$$DI=\min_{1\le i\le k}\left\{\min_{j\neq i}\left(\frac{d_{\min}\left(C_i,C_j\right)}{\max_{1\le l\le k}diam\left(C_l\right)}\right)\right\} \tag{18}$$

式中，$avg\left(C_i\right)$ 代表第 i 类簇内样本的平均距离，$avg\left(C_j\right)$ 代表第 j 类簇内样本的平均距离，$d_{cen}\left(C_i,C_j\right)$ 代表簇 C_i 和簇 C_j 中心点之间的距离，$d_{\min}\left(C_i,C_j\right)$ 代表簇 C_i 和簇 C_j 最近样本之间的距离，$diam\left(C_l\right)$ 代表簇 C_l 内样本的最远距离。

DB 指数的计算方法是对任意两个簇内样本的平均距离之和除以两个簇的中心点距离，并取最大值，DBI 的值越小，意味着簇内距离越小，同时簇间的距离越大；Dumn 指数的计算方法是任意两个簇的最近样本间的距离除以簇内样本的最远距离的最大值，并取最小值，DI 的值越大，意味着簇间距离大而簇内距离小。因此，DBI 的值越小，同时 DI 的值越大，意味着聚类的效果越好。

4.3.3 数据分析

在理论基础上，对附件 2 数据进行热点问题的挖掘。在代码实现上，分别进行了一下步骤：

1.应用 pynlpir 分词和停用词表对数据集进行分词、标注词性、去停用词，完成对数据的初步处理并筛选出其中的人物、地点名词；

2.应用 TF-IDF 算法在分词数据的基础上建立词带模型和 tf-idf 词向量矩阵；

3.根据分词数据计算余弦相似度并进行 K-means 聚类。

4.3.4 挖掘结果

对数据的经过分析，最终得到了挖掘出来的热点问题（完整数据见附件）。

表 3 热点问题表

问题 ID	时间范围	地点/人群	问题描述
1	2019/11/29-2019/12/6	A1 区 A2 区华庭	A1 区 A2 区华庭物业管理不善，居民生活环境恶劣
2	2017/6/8-2019/11/26	A 市经济学院	A 市经济学院强制学生实习
3	2015/3/14-208/8/13	A7 县安沙镇	A7 县安沙镇民生问题严重
4	2013/9/6-2019/4/23	A7 县北山镇	A7 县北山镇百姓出行道路需要的到改善
5	2016/12/23-2019/3/8	A5 区幼儿园	A5 区幼儿园普遍出现管理不善的问题

4.4 问题三 答复意见的评价体系

4.4.1 思路构建

在对答复意见质量的评价体系构建过程中，重点对政务数据中附件 4 的“留言主题”、“留言详情”、“答复内容”三类文本数据进行了分析。在对答复意见质量的评定中，通过计算出答复内容与留言主题、答复内容与留言详情两两之间的文本相似度，作为评价体系的核心量化标准。在计算相似度的过程中，采用了两张计算方式，分别是基于向量空间余弦相似度算法和基于 TF-IDF 相似度算法，并两种算法呈现的效果进行比较，最终优化的评价结果。

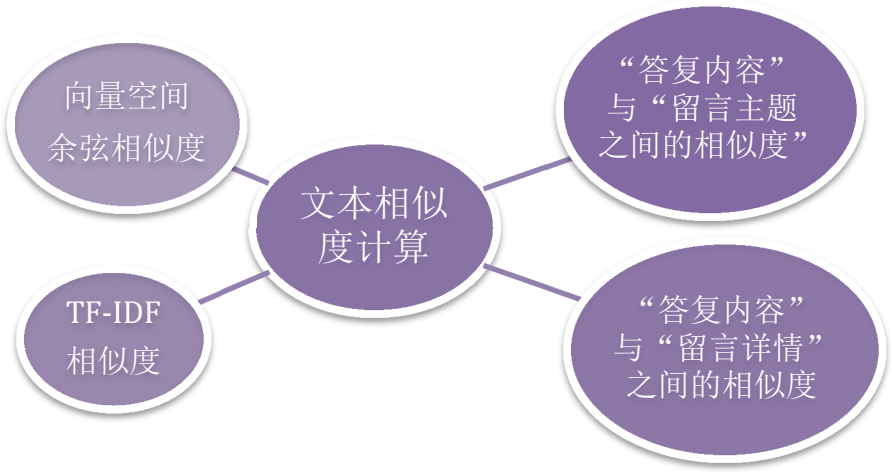

图 4 问题三思路流程图

在对于不同的文本计算他们之间的相似度时，将文本中词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同的向量的差异的大小，来计算文本的相似度。

4.4.2 基于向量空间余弦相似度的文本相似度计算

余弦相似度通过向量空间对两个向量夹角的余弦值作为衡量两个个体间差异的大小。因此，可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。

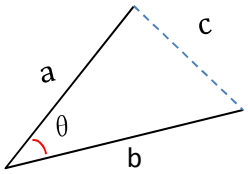

图 5 余弦三角形示意图

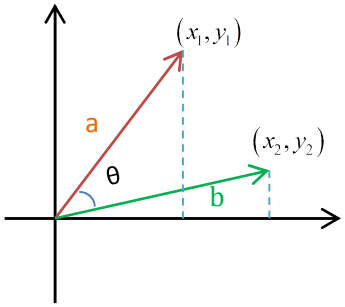

图 6 直角坐标系示意图

在余弦定理公式的基础上，由图 6 令 a 向量是(x_1,y_1)，b 向量是(x_2,y_2)，带入到余弦公式中我们可以得到一个新的余弦公式

$$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab}$$

$$=\frac{\left(\sqrt{x_1^2+y_1^2}\right)^2+\left(\sqrt{x_2^2+y_2^2}\right)^2-\left(\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}\right)^2}{2\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{19}$$

$$=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}$$

当 a 向量和 b 向量均为 n 维向量时可以得到余弦公式

$$\cos(\theta)=\frac{\sum_{i=1}^{n}x_i\times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}=\frac{a\bullet b}{|a|\times|b|} \tag{20}$$

通过对公式（20）运用，可以计算出文本的余弦相似度。

4.4.3 基于 TF-IDF 的文本相似度计算

TF-IDF 是一种用于信息检索与文本挖掘的常用加权技术。在文本 3.1 中已经对 TF-IDF 的算法核心进行了介绍。在使用 TF-IDF 对答复内容和留言内容之间进行文本相似度的计算过程中，我们通过构建留言内容的 tfidf 权重向量并建立词库与通过答复内容建立的文本稀疏向量之间来计算相似度。

4.4.4 数据处理

通过对向量空间余弦相似度算法和 TF-IDF 算法的运用，可以完成文本相似度的计算，但在计算的过程中，由于附件 4 数据集中“答复内容”和“留言详情”的文本数据长短不一，且存在大量无意义的词语。对附件 4 文本数据进行了数据处理流程。依次进行了 jieba 分词、去停用词、提取关键词等过滤操作，完成对数据进行了优化。

表 4 文本处理结果展示

留言编号	留言主题	留言详情	答复意见
2549	景蓉华苑、物业管理……	物业公司、投票、业主、A2、业委会、高昂、水电……	业主大会、业委会、停车、A2、景蓉花苑、建设局、物业管理……
2554	潇楚南路、洋湖段、还没、修好……	一段、方便、潇楚、2018、挖机、这路、一个圈、围栏、很大、稀烂、店面、修好……	施工、坪塘、排水、A3、换填、土方、管线、集镇、道路……
2555	加快、提高、民营、幼儿园、老师……	幼儿园、教师、民营、工作、普惠型、工资待遇、超负荷、压力、民办、雪上加霜……	民办、幼儿园、待遇、教师、学前教育、教职工、依法、提高、保险、保障、工资待遇……
2557	买、公寓、享受、人才、新政……	公寓、研究生、您好落户、新政、尊敬、购房、请问、补贴、人才、享受、毕业……	购房、房屋交易、补贴、管理中心、首次、公寓、在编人员、住房、收悉、限购……
…	…	…	…
185799	燃油、税费、改革、政策、咨询……	公路局、中央财政、咨询、地州 、养路费、递增、财政局、税费……	燃油税、市州、资金、支付、转移、养护、公路局、交通运输、西地省、中央、农村公路……
185986	强烈呼吁、宁朱、公路、拓宽、提质……	公路、朱良桥、好路、宁朱、提质、强烈呼吁、建设、集镇、政府、破烂、动工……	交通运输、前期工作、项目、公路、朱良桥、提质、西地省、规划、建设、实施方案、基本建设……

4.4.5 相似度计算

在数据处理的基础上，对处理过后的分词数据，分别进行了余弦相似度计算和 TF-IDF 相似度计算。在计算结果中我们发现，两种算法的计算出来的相似度会有比较大的差别，而留言详情和答复内容、留言主题和答复主题之间的相似度也存在一些差异。

为验证相似度的计算结果是否准确，我们经过人工判断，抽取出一部分留言内容，并人工定义高、中、低三个质量等级。在抽取过程中，我们依次每 500 份数据抽取高、中、低各 3 份样本，共抽取出 54 份数据。（完整数据见附件）

表 5 文本相似度结果展示

留言编号			留言主题/答复内容		留言详情/答复内容	
			余弦	TF-IDF	余弦	TF-IDF
0-500	高	2557	0.28	0	0.32	0.4743
		2549	0	0	0.31	0.4666
		4105	0.31	0	0.52	0.6804
	中	2554	0.17	0.6249	0.03	0.0452
		2555	0.35	0.2712	0.41	0.371
		3681	0.26	0.3318	0.24	0.3583
	低	4157	0.08	0	0.03	0.1027
		2759	0.36	0.746	0.06	0.0859
		5180	0.18	0.1402	0.34	0.5029
501-1000	高	12614	0.15	0.0392	0.22	0.3072
		12649	0.31	0	0.27	0.1593
		12847	0.18	0	0.35	0.3343
	中	12768	0.27	0.0365	0.45	0.45537
		16708	0.21	0	0.21	0.2041
		17009	0.09	0.0745	0.22	0.3229

	低	11927	0	0	0	0
		11955	0	0	0.03	0.0745
		11974	0	0	0	0
…	…	…	…	…	…	…
2500-2816	高	137047	0.29	0.3405	0.26	0.3694
		139893	0.29	0.4082	0.25	0
		175830	0.07	0.2672	0.07	0.2886
	中	132959	0.05	0	0.25	0.38
		136733	0.21	0.2626	0.11	0.1593
		139579	0.12	0	0.13	0.2696
	低	132461	0.23	0	0.06	0
		140192	0.21	0.2499	0.16	0.136
		139964	0.13	0	0.24	0.2461

对抽取出来的数据进行观察，发现不管是通过余弦相似度还是 TF-IDF 相似度的计算。对于通过“留言主题”和“答复内容”之间相似度的计算结果，普遍不稳定，存在多个 0 相似度的出现，且出现的规律比较混乱，可参考性不高。而“留言详情”与“答复内容”之间计算出来的相似度，得到的结果更加准确。

在余弦相似度中，可以发现高质量“答复内容”相似度集中在 0.3-0.5 之间浮动；中等质量“答复内容”相似度集中在 0.1-0.3 之间浮动；低等质量“答复内容”相似度集中在 0-0.1 之间浮动。

在 TF-IDF 相似度中，高质量“答复内容”相似度集中出现在 0.3-0.7 之间，多在 0.4-0.5 之间浮动，会出现一些低相似度的情况，但频率较低；中等质量“答复内容”相似度集中在 0.1-0.4 之间，多在 0.3 之间浮动；低等质量的“答复内容”相似度集中在 0-0.1 之间，会出现一些高相似度的情况。通过对高相似度的低等质量“答复内容文本”进行观察，发现主要原因为“答复内容”和“留言详情”均属于较短文本，且“答复内容”是重新换了一个方式将

“留言详细”提出的问题进行了无意义的回复。例如留言编号为 5180 的留言语句和答复语句：

表 6 示例语句展示

留言详情	…我们学校要求 8 月 15 号要到校强制自习…强制自习与上课有什么区别，求 A 市教育局明查…
答复语句	…由于新高二年级有很多家长和学生要求暑假能在学校自习，年级组拟自 8 月 15 日起开放部分教室供有需求的学生自主自习，完全自愿，不会强制执行…

通过对余弦相似度和 TF-IDF 相似度的对比，可以发现对人工选出的三个等级的数据来说，余弦相似度的计算结果均偏低，但是对于低等质量“留言内容”计算的结果较为稳定，比较能够体现出结果，而高等和中等“留言内容”的相似度在 0.5 上下浮动，稳定性一般。对于 TF-IDF 相似度的计算结果来说，高、中、低三个等级的取值之间还是拥有较为清晰的分界线，取值范围没有过多的重合。

表 7 相似度取值范围

	余弦	TF-IDF
高	0.3-0.5	0.3-0.7
中	0.1-0.3	0.4-0.5
低	0-0.1	0-0.1

通过计算结果的对比，我们最终得出通过 TF-IDF 计算出的相似度要更加理想和精准，而余弦相似度计算出的结果存在一定的偏差但对于相似度低的文本数据计算出来的结果更加准确。

由此，在对答复内容质量的评价体系的建立中，通过对“答复内容”与“留言详情”之间文本相似度的计算来设定答复内容的质量高低。针对余弦相似度和 TF-IDF 相似度两种算法，将 TF-IDF 算法计算出的相似度作为一级参考标准，将余弦相似度计算的相似度作为二级参考标准。通过两种计算结果的综合考虑，完成对单个答复内容质量的评价。

5.总结

政务数据信息的分析研究，对政府部门了解民意、汇聚民意、实现民意、汇聚民心具有重大意义，同时对政务数据的挖掘分析也是基于自然语言处理技术的全新运用和延伸。在传统线下问政逐渐被线上网络问政取代的今天，庞大的电子政务数据为国家带来了更多社会问题、民生问题、教育问题、环境问题等等，在逐步拉动国家制度完善、社会进步、提高人民幸福感的同时，也带给了我们新的技术挑战。在本篇论文中，我们通过对政务数据进行文本分词（jieba、pynlpir）、词性标识（pynlpir）、去停用词（停用词表）、提取文本关键词（TF-IDF）、LDA 主题模型、多项式贝叶斯算法、K-means 文本聚类等自然语言处理技术，完成了对政务数据的挖掘，得到了留言内容的分类模型、政务数据的热点问题挖掘方法、答复内容的评价体系构建。

从挖掘的结果中可以看出，对于问题一的分类模型的使用，有一定的限制条件，但分类正确率达到了 80%以上，同时当数据量继续增加时可以尝试使用 LDA 主题模型来进行挖掘。对于问题二，使用了 K-means 文本聚类的方法来实现热点问题的挖掘。同时在对特定词汇的提取上，除了使用分词器的词性标识以外，还可以通过其他的方式，例如 LSTM、NLTK 等算法来对特定词汇进行提取。对于问题三，通过对文本相似度的计算得到对答复内容的评价体系的建立，最终得到了较理想的成果。

通过对政务数据的挖掘，可以看出人民对于美好生活的向往程度越来越高，社会进步的脚步也越发快速。网络问政平台的兴起预示着社会问题的逐渐公开透明，在带给国家更大挑战的同时，也使得社会的加速前行。通过对“智慧政务”的信息文本挖掘和模型验证，能够成为促进社会进步的催化剂之一，最终帮助到政府和国家更加便捷有效的解决民生问题、汇聚民心、加快建设步伐。

6.参考文献

[1] 张庆文, 王武魁. 电子政务与数据挖掘技术[J]. 电子政务, 2006, 000(006):82-84.

[2] 武永亮,赵书良,李长镜,魏娜娣,王子晏.基于 TF-IDF 和余弦相似度的文本分类方法[J].中文信息学报,2017,31(05):138-145.

[3] 罗欣, 夏德麟, 晏蒲柳. 基于词频差异的特征选取及改进的 TF-IDF 公式 [J]. 计算机应用, 2005, 025(009):2031-2033.

[4] 谭静. 基于向量空间模型的文本相似度算法研究[D]. 西南石油大学, 2015.

[5] 刘正, 黄震华. 基于多项式贝叶斯分类模型的短文本多情感倾向分析及实现 [J]. 现代计算机(专业版), 2016, 000(014):39-42,47.

[6] 陈宝楼. K-Means 算法研究及在文本聚类中的应用[D]. 安徽大学, 2013.

[7] 李实, 叶强, 李一军, 等. 中文网络客户评论的产品特征挖掘方法研究[J]. 管理科学学报, 2009(02):146-156.

[8] 周胜臣, 瞿文婷, 石英子, 等. 中文微博情感分析研究综述[J]. 计算机应用与软件, 2013, 30(3):161-164

[9] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3):18-24.

[10] 陈晓云. 文本挖掘若干关键技术研究[D]. 复旦大学.

[11] 石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报, 2009, 35(12):1586-1592.