大数据时代下智慧政务

——基于自然语言处理的文本挖掘

大数据时代下智慧政务

——基于自然语言处理的文本挖掘

摘要：在大数据时代，政府部门通过网络问政了解民生民意。随着网络问政的不断发展，网络留言成为中国公民行使知情权、参与权、表达权和监督权的重要渠道。在国家决策层网络问政的重视程度提升的同时，也带来了日渐增长的巨量网络留言数据。传统的人工处理留言存在处理时间长，效率低，错误率高的问题。本文通过自然语言技术处理数据，极大地提高了效率。

我们首先进行数据预处理，通过 nlpir 分词及停用词进行文本向量化，通过计算文本余弦值删除同一个人密集发布的留言，再计算 TF-IDF 向量化文本矩阵。为了提高分类准确率，我们用多元 Logistic 回归对群众留言进行分类。然后，运用决策树模型测算出可以用于确认热点的属性，通过减枝去掉分散数据，减少运算数据量。用正则表达式提取地点等文本属性值，运用可变的优化 DBSCAN 聚类方法找出潜在热点地区，筛选后用改进的 Hacker news 方法计算热度值，前五为热点问题。最后通过层次分析法构建多目标评价模型，计算出完整性、相关性和可解释性的权值，对回复进行打分。

关键词：TF-IDF 向量化,DBSCAN 聚类，层次分析法，多元 Logistic 回归

Abstract: In the era of large data, government departments understand people's livelihood and public opinion through the network to ask people about politics. With the continuous development of online inquiry and administration, online message has become an important channel for Chinese citizens to exercise their right to know, participate, express and supervise. But at the same time, it brings a huge amount of message data.

The traditional manual processing of messages has the problems of long time, low efficiency and high error rate. Through the large data processing method, this article can deal with messages fast and accurately.

First, we do data preprocessing, through nlpir word segmentation and processing stop words, by calculating the text cosine value to screen out the same person densely published messages to delete. Then calculate the TF-IDF to quantify the text matrix. In order to improve

the accuracy of classification, we use multivariate Logistic regression to classify mass messages. Then, the decision tree model is used to calculate the attributes that can be used to confirm the hot spots, remove the scattered data by pruning, reduce the amount of operation data, extract the value of text attributes such as location by regular expression, use variable optimization DBSCAN clustering method to find the potential hot spots, then use the improved Hacker new method to calculate the heat and find the hot spots after screening. Finally, the multi-objective evaluation model is constructed by AHP, and the responses are graded.

Keywords: TF-IDF direction quantification, DBSCAN clustering, analytic hierarchy process, multivariate Logistic regression

目录

一. 简介

1.1. 挖掘背景

我国政务信息化实现了政府结构由物理碎片化到虚拟空间整体性的转变、政府管理由封闭到开放的转变、政府内部治理由部门协调到整体协同的转变、政府运行由传统的手工作业到智能智慧的转变，数字政府初露端倪。在大数据时代，我国智慧政务将迎来新的发展，力求进一步改善施政效率，促进服务发展方式转型，使人工智能得到充分应用。

在大数据时代，微信、微博等便捷、多样化的信息方式极大地解放了民众的话语权，开创政府网络问政的新模式，成为党政部门倾听民声，集中民智的重要渠道。由于各类相关的文本数据过于庞大，给相关部门在处理数据方面的工作带来了极大的挑战。利用大数据技术进行相关的文本处理已经成为了顺应时代发展潮流的选择。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。

构建智能文本挖掘模型，提取问政留言记录中的相关热点，了解群众当前状况下最强烈的需求，及时发现热点问题，并给出针对性地处理，提升政府的服务效率。

1.2. 挖掘目标

本文利用词频-逆向文件频率、改进的决策树等模型构建智能文本挖掘模型。模型可以起到帮助相关部门按照一定的划分体系对群众留言进行分类的作用，提取问政留言记录中的相关热点，定义出合理的热度评价指标，给出评价结果，帮助相关部门提高效率。并且尝试实现针对相关部门对留言的答复意见，从答复相关性、完整性、可解释性等角度对答复意见的质量给出评价。

在当前智慧政务不断发展的大体背景下，本文构建的智能文本挖掘模型希望能够很好的提取出潜在的热点问题，帮助党政部门更好地了解民生民意，并针对相应的热点问题给出答复，有效提高了政府部门的施政效率。

1.3. 挖掘流程

如图 1，挖掘过程主要分成数据预处理、群众留言分类、热点问题挖掘和答复意见评价四个部分。其中数据预处理包括文本分词，去停用词。群众留言分类包括了生成 TF-IDF 矩阵并计算类别权值。热点问题挖掘利用改进的决策树模型提取地点，提取动词，生成聚类并计算热度。答复意见评价利用层次分析法从完整性，相关性，可解释性三个角度对答复意见的质量进行评价。

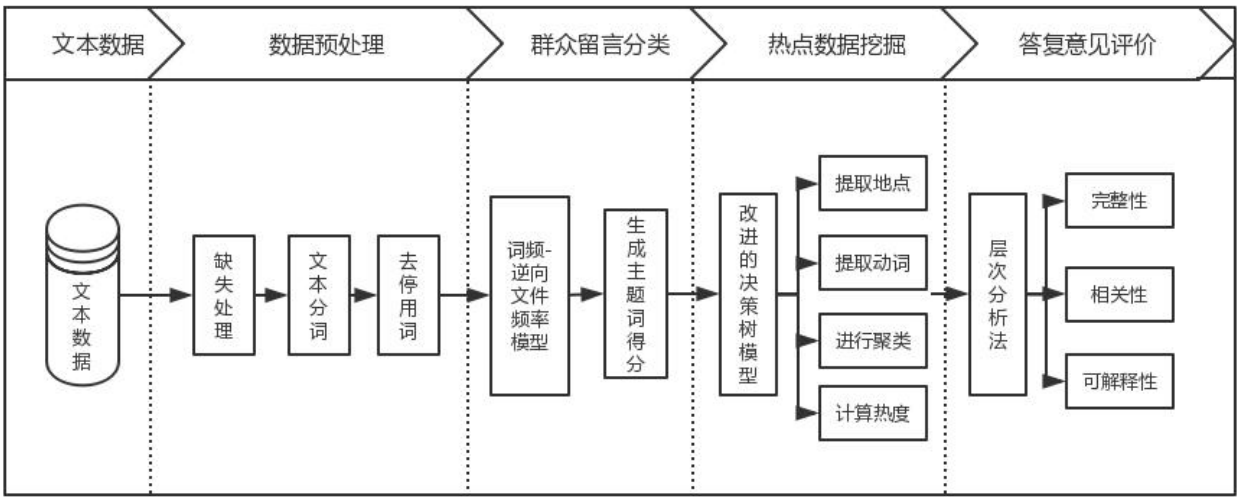

图 1 挖掘流程

二. 数据预处理

2.1. 数据相关性处理

首先用 python 对数据进行缺失值查找，因数据量较大，少量缺失值对结果影响甚微，予以剔除。然后，通过主层次分析法计算各因子对结果的影响程度，将对问题热点影响较小的数因子剔除。

2.2. 文本分词

由于中文文本词与词之间没有明显界限的特点，从文本中提取词语时需要进行分词，本文使用的是中科院汉语分词系统，利用 python 中的 pynlpir 包进行分词。nlpir 分词系统的主要思想是先通过 CHMM（层叠形马尔可夫模型）如图 2 进行分词。

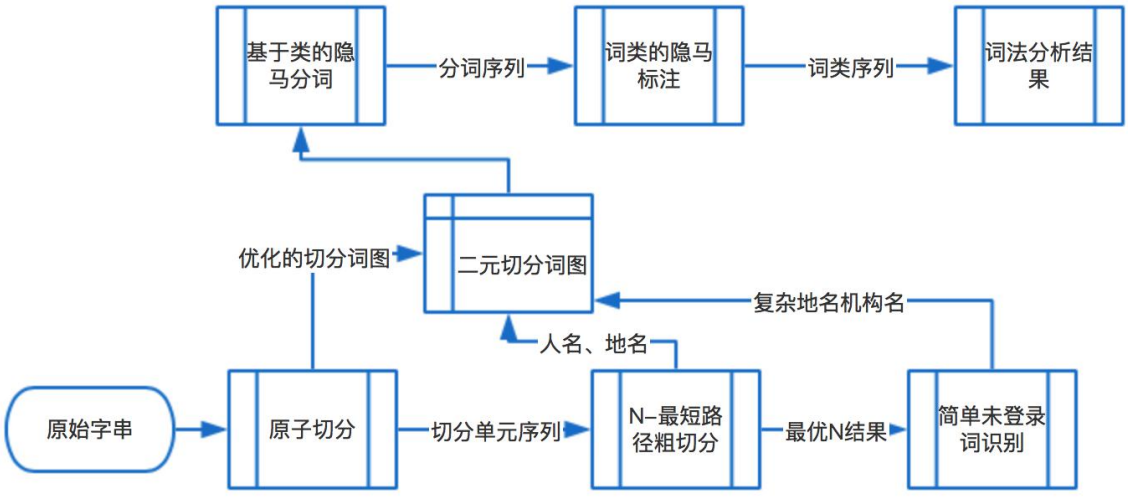

图 2 层叠形马尔可夫模型

步骤 1:提取文档中的句子，对句子进行了原子切割。

步骤 2:导出 coreDict 字典进行第一步的分词处理，将分词结果放在一个储存容器中。

步骤 3:导出 biGramDict 字典，对步骤 2 的结果进行处理，将结果放在一个新的容器中，并计算出两个词的平滑度。

步骤 4:进行初次切分，通过最短路径算法，权重为步骤 3 计算出的平滑值。

步骤 5:进行人名识别，基于角色标注的算法，通过查找 nr 字典，最终匹配出人名。

步骤 6:处理地点等信息.

步骤 7:优化结果，添加词性。

2.3. 去停用词

在文本处理中，包含着一些功能普遍，却没有什么实际含义的词语，比如中文中的“的、了、地”。本文所用的停用词为网络流行语停用词表，以及组员整合的词表。去除停用词，更能突出热点词，有利于后面对热点问题的筛选。

三. 群众留言分类

3.1. 词频-逆向文件频率模型

词频-逆向文件频率模型（TF-IDF）是一种用于文本向量化的方法，通过计算逆文档频率和词频相乘，将出现次数较多但是没有太多意义的词语重要性程度减少，可以得到一个关于文本词语重要性的矩阵，词语对应的数值越大，该词语就越重要。

TF:词频，表示一个词语与文本的相关性。计算时用该词在文本中出现的次数除以文本的总词数。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

IDF:逆向文件频率，表示一个词语的出现的普遍程度。可以表示为 log(总文本数/出现该词语的文本数)。

$$idf_i = \lg \frac{|D|}{|\{j : t_i \in d_i\}|}$$

TF-IDF= TF * IDF ，词频和逆向文件词频的乘积表示词语的重要程度。

3.2. 计算主题词得分

将每个主题的 TF-IDF 矩阵每个词的得分累加,并处以文档数量,生成各个主题的主题词以及其对应的权重，利用如下公式:

$$\mathrm{word_i}=\sum_{\mathrm{k=1}}^{C}\mathrm{Score_{tf\text{-}idf}}$$

计算各个主题的下的各个词语的得分，下面是环境保护主题中得分排名前五的词语以及其得分，发现得分确实与环境保护主题关系较为密切。

环境保护	水	厂	村	环境	污染
得分	0.026195696	0.026656337	0.027818657	0.02796967	0.035835967

此后，将文本中的每个词都带入各个主题中计算其得分，计算公式为:

$$\mathrm{T\text{-}score_k}=\sum_{i=1}^{i\in D}word_i*n$$

其中 n 为文本中第 i 个词出现的频率，wordi 为第 i 个词在某个主题下的得分。得分最高的即为其分类。

此后为了更好的优化模型，我们采用多元 softmax 回归的方法处理 TF-IDF 向量。

我们构造了多元分类器函数如下:

$$\mathrm{h}_\theta(x^{(i)})=\begin{bmatrix}p(y^{(i)}=1\mid x^{(i)},\theta)\\p(y^{(i)}=2\mid x^{(i)},\theta)\\\vdots\\p(y^{(i)}=7\mid x^{(i)},\theta)\end{bmatrix}=\frac{1}{\sum_{c=1}^{k}e^{\theta_c^T x^{(i)}}}\begin{bmatrix}e^{\theta_1^T x^{(i)}}\\e^{\theta_2^T x^{(i)}}\\\vdots\\e^{\theta_7^T x^{(i)}}\end{bmatrix}$$

其中，X(i)为每个文本的词向量，θ是用极大似然估计计算出的参数，使得 h 最大的 y，即为文本的种类。

四. 热点问题挖掘

4.1. 改进的决策树模型

决策树模型是一种在多层次或多阶段决策帮助决策者进行序列决策分析的有效工具。决策树模型是由根节点、叶节点及内部节点构成，通常以最大收益期望值或最低期望成本作为决策准则，通过图解方式求解在不同条件下各类方案的效益值，然后通过比较，做出决策。我们希望利用决策树模型找到一个合适的文本属性决策顺序，可以通过寻找合适的属性决策顺序减少文本数据的运算量，利用改进的决策树模型对决策树进行剪枝，去掉不可能成为热点的数据，减少数据量，加快运算速度。

改进的决策树模型如图 4，提取分类后相似度依然很高的语句作为热点词语句。

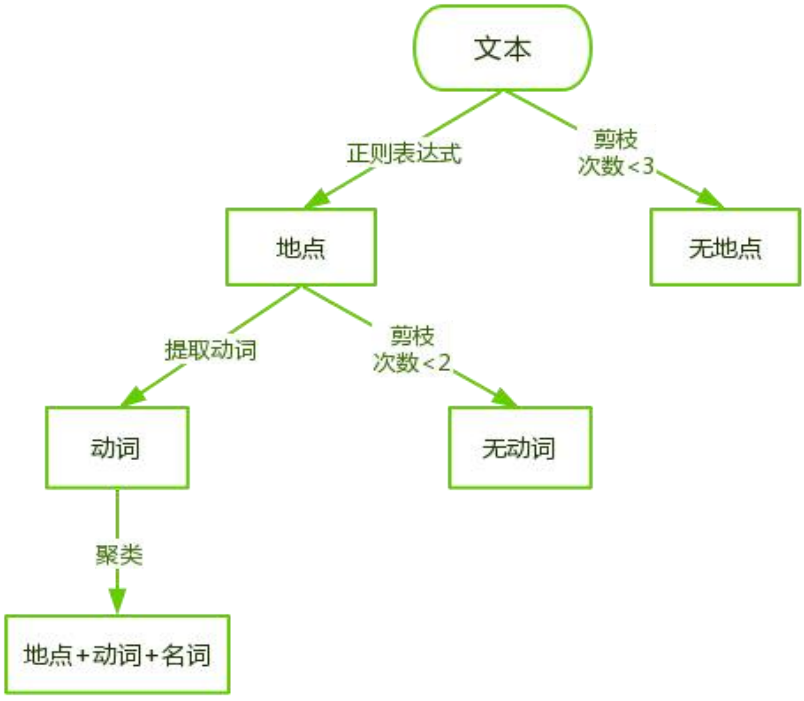

图 4 改进的决策树模型

步骤 1:我们通过分词和去停用词提取出有用的地点词，动词以及名词。

步骤 2:利用信息熵，计算公式分别计算地点词，名词，动词的信息熵，信息熵最大的为地点词，用地点词作为第一步的分类叶节点。信息熵的计算公式如下：

$$H(X) = -\sum_{i \in k} p(x) * \log p(x)$$

步骤 3:分类后计算名词与动词的信息熵，动词的信息熵比名词大，因此将动词作为第二步的分类

4.2. 提取地点

提取文本中的地点和人群，发现其中的热点问题，有助于相关部门进行针对性的处理，提高政府的服务效率。没有清晰地点的反馈缺少客观的事实依据，无法作为热点问题。利用如下公式计算各个留言的余弦值

$$\cos\theta = \frac{X * Y}{|X| * |Y|}$$

如果余弦值大于等于 0.99，并且间隔时期小于一周，说明两文本可能是短时间内同一人重复反映的，将其删去。本文在九千多个数据中，删去两百多个不含有地点的数据，利用正则表达式来提取地点。正则表达式是对字符串操作的一种逻辑公式，我们可以利用正则表达式来对文本进行地点的查找。

本文采用了下列公式

$$\$[A-Z][0-9]\{市，区，县\}\$$

$$\$[A-Z]\{市，区，县\}\$$

筛选出标题和正文中含有的地点名词，如 A5 市等，公式中$[A-Z][0-9]代表的是从第一个字符 A 到 Z，第二个字符 0-9 第三个字符市区县。

地点少于等于两个的说明反映问题的人较少，即使是点赞的人较多,我们也认为它没有办法成为热点,并且存在一个人多次反馈的可能，因此我们删去少于等于两个的地点,共得到 34 个反映数量大于等于 3 个的地区，成为潜在的热点地区。

4.3. 提取动词

我们读取文档后发现,热点事件的关键在于动词，在删去诸如”请”,’谢谢’等动词后，利用前文的文本分词技术分出有词性的词语，在具有相同地点的前提下，判断各个标题和文本的动词是否一致。删去少于等于一的动词，统计动词和地点全部一致的文本数量，得到各个地区前三数量的动词，作为潜在热点问题的动词。

4.4. 提取名词

将特定动词和特定地点的文本筛选出来后，用 TF-IDF 方法向量化。

Dbscan 如图 3 是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。我们可以利用不断的修改 Dbscan 的参数达到使得热点问题进行聚类，非热点问题成为噪音。

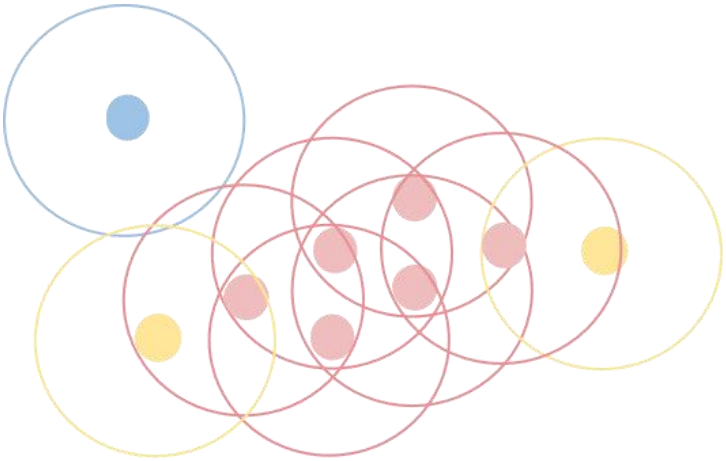

图 3 dbscan 聚类算法

定义:ε邻域

设 $N_{\varepsilon}(x)=\{y\in X:d(y,x)\le\varepsilon\}$ $x\in\varepsilon$,称为 x 的 ε 邻域,显然 $x\in N_{\varepsilon}(x)$

设 $x\in X$ 称 $\rho(x)=\left|N_{\varepsilon}(x)\right|$ 为 x 的密度

设 $x\in X_{nc}$ 若 $\rho(x)>M$ 则称 x 为 X 的核心点,记 X 中所有核心点构成的集合为 X_c 并记 $X_{nc}=X\setminus X_c$ 表示由 x 中所有非核心点构成的集合

若 $x \in X_{nc}$ 且 $\exists y \in X$ 满足 $y \in N_{\varepsilon}(x) \cap X_c$ 即 x 的 ε邻域 中存在核心点,称 x 为 X 的边界点

记 $X_{noi}=X \setminus (X_c \cup X_{bd})$ 若 $x \in X_{noi}$ 则称 x 为噪音点

在我们挖掘热点的模型中,我们可以通过 ε邻域 的值由小变大,在某个区域内,可以找出在空间存在较为接近的核心点(即不是所有的点都为噪音点),这些较为接近的核心点就是可能的热点数据,而离的较远的噪音点就是其他的非热点数据.

受 dbscan 启发,我们将 Eps（扫描半径）由小到大增加,直到出现大量核心点,说明此时出现了大量的相似文本,作为潜在的核心热点.

此后筛选出潜在核心热点的关键词，查找所有与关键词相同的文本,为热点问题

4.5. 计算热度

热度计算公式采用 Hacker news 的计算方法

热度计算公式为

$$\sum \frac{P+1}{(\frac{t}{4}+2)^{1.2}}$$

t 表示距离发帖的时间（单位为月），加上 2 是为了防止最新的帖子导致分母过小

P 为点赞得分,计算方法为点赞数*1

G 表示"重力因子"，即将排名往下拉的力量，我们设置的默认值为 1.2

将每个热点问题的相关帖子热度累加即为该热点问题的热度值。

五. 答复意见评价

5.1. 答复意见

答复意见要解决的问题有:提示问题目前的受理情况,告知办事的法律依据以及给出事情的通俗解释，给出解决方案。本文从答复的相关性、完整性、可解释性三个角度对答复意见的质量进行评价，如图 5 所示。

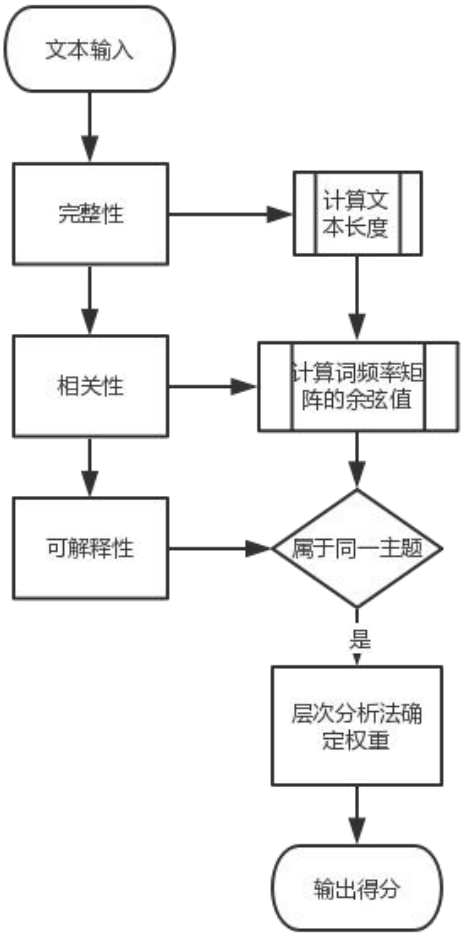

图 5 答复意见评价流程图

5.2. 完整性

完整性就是对问题的解释以及阐述完整。从总体来说，语义完整性和文字的长度成正比，而根据心理学研究显示，留言者也将答复意见的长度作为评判完整性的重要依据，

因此我们将文字的长度作为完整性的评价指标，将根据不同文档长度的排名给予答复的完整性得分

5.3. 相关性

相关性是指答复意见和正文的相关性，可以利用文本的相关性余弦进行计算。

步骤 1:生成各个文档的 TF-IDF 矩阵。

步骤 2:根据各个文档 TF-IDF 矩阵取出答复意见和正文数值最大的 20 个特征值生成特征向量。

步骤 3:计算公式

$$\cos\theta = \frac{X * Y}{|X| * |Y|}$$

cos 的绝对值越小，说明文本的相似度越高。

5.4. 可解释性

可解释性的意思是回复是否能很好的解释留言的内容，我们用第一题做出的逻辑回归分别去预测文本和留言的主题类别，判断文本与留言属于的主题，若属于同一个主题即为有可解释性，不属于同一个主题为主题不明，不具有可解释性。

5.5. 计算得分

根据层次分析法确认可解释性，相关性以及完整性的权重，按照权重乘以各自的得分计算总得分来给出答复意见的评价。

层次分析法 AHP 如图 6，是指将与决策总是有关的元素分解成目标、准则、方案等层次，在此基础之上进行定性和定量分析的决策方法。

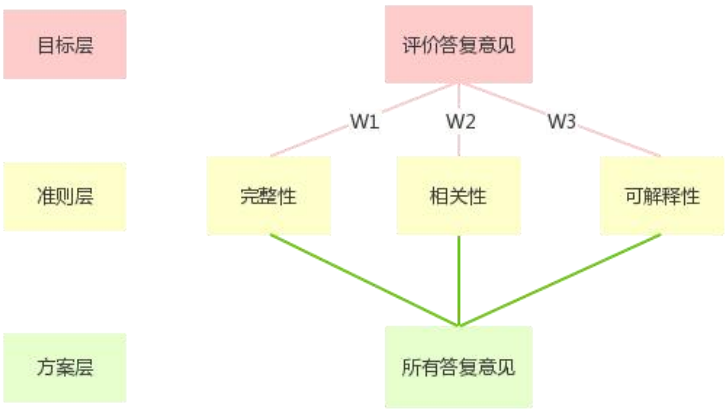

图 6 层次分析法

查找论文统计出对比矩阵如下表 1:

表 1 对比矩阵

	完整性	可解释性	相关性
完整性	1	5	3
可解释性	0.2	1	0.333333
相关性	0.33333	3	1

利用公式

$$CI = \frac{\lambda - n}{n - 1}$$

对矩阵的一致性进行检验，CI 值为 0.0193，表示矩阵的一致性较好。

得到完整性,可解释性,相关性的权值如下：0.2583,0.1047,0.6370。

步骤 1:计算各个文档的长度，然后用最大值减最小值的方法进行归一化处理。

步骤 2:计算词频率矩阵，将答复意见和留言详情进行对比后，求出其余弦值，并用最大值减去最小值的方法进行标准化，得到概率分布。

步骤 3:将答复意见和留言详情利用第一题的分类模型进行分类,比对其主题,如相等则加 0.5 分,如不相等则不加分,得到各项的分数。

六. 实验结果及优化

6.1. 群众留言分类结果

6.1.1 第一次分类结果

第一次的分类结果如下表 2 所示

表 2 第一次分类结果

	商贸旅游	卫生计生	劳动与社会保障	教育文体	交通运输	环境保护	城乡建设
商贸旅游	700	21	23	38	5	17	81
交通运输	282	93	154	200	563	70	481
教育文体	45	9	68	1109	2	10	37
城乡建设	60	24	48	68	9	45	1090
劳动与社会保障	44	42	1494	87	5	25	78
卫生计生	36	649	134	47	10	11	57
环境保护	48	39	48	40	19	760	182
准确率	0.576	0.740	0.758	0.697	0.918	0.810	0.543

第一次分类得到的准确率为 73%，从表中可以观察到，商贸旅游和城乡建设的准确率较低分别为 57.6%和 54.3%，交通运输的准确率达到了 91.8%，较为准确，但是总的来说分类的效果没有达到预期目标。

根据表画出柱状堆积图 7

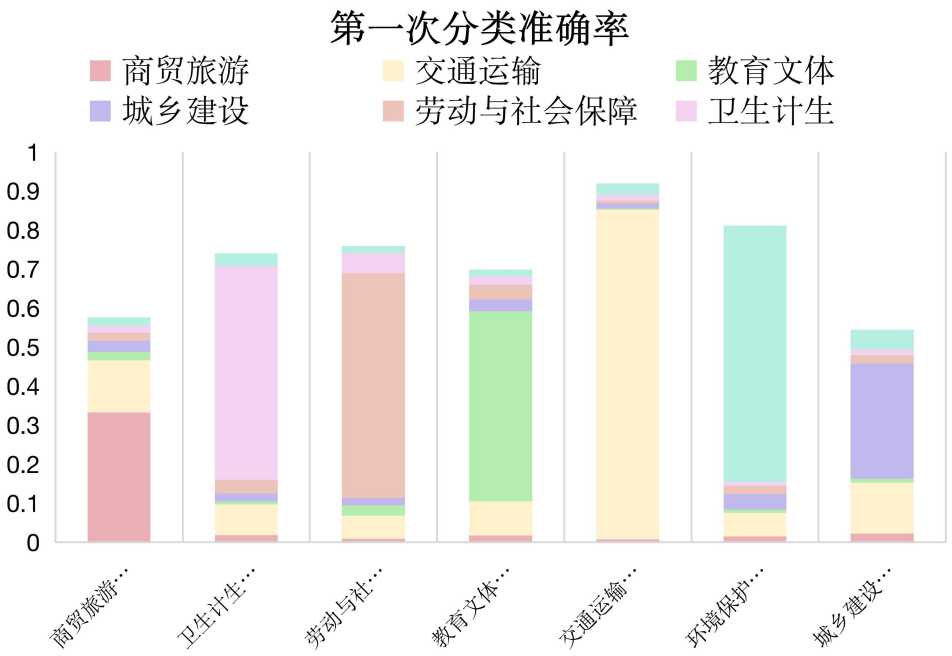

图 7 第一次分类准确率

6.1.2 模型优化

本文根据逻辑回归模型对算法进行了优化。逻辑回归中的多元分类，通常使用 Softmax Regression 模型，如图 8 所示。对每一类别进行估算，将其特征相加，然后将特征转换为判定是这一类的概率。

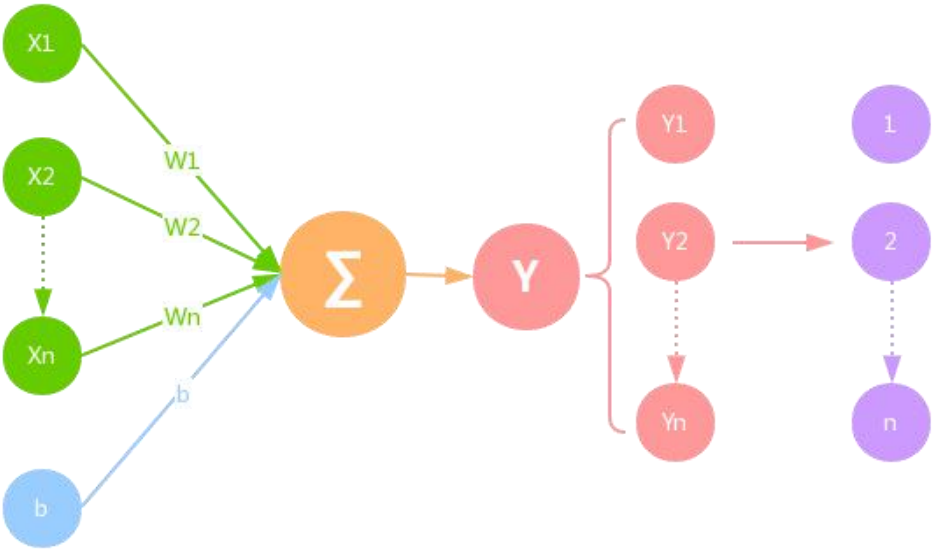

图 8 逻辑回归模型

我们将 TF-IDF 矩阵中的数据用公式

$$p_i = \frac{e^{y_i}}{\sum_{k=1}^{C} e^{y_i}}$$

计算出每一类中每一个词对应的概率值,其次将交叉上公式

$$H(p,q) = -\sum p(x), \log^{q(x)}$$

作为损失函数,进行逻辑回归的计算,结果如下:在抽取了 80%的训练集和 20%的预测集的情况下,10 次测试预测集的正确率如下表 3 所示

表 3 10 次预测正确率

次数	1	2	3	4	5	6	7	8	9	10
准确率	0.858	0.860	0.851	0.863	0.872	0.871	0.865	0.867	0.869	0.862

6.1.3 第二次分类结果

第二次的分类结果如下表 4 所示

表 4 第二次分类结果

	城乡建设	商贸旅游	卫生计生	环境保护	交通运输	教育文体	劳动和社会保障
教育文体	6	4	6	4	0	277	10
交通运输	7	3	0	0	110	0	0
城乡建设	335	32	6	15	14	12	14
环境保护	10	3	0	147	0	2	0
商贸旅游	10	216	7	5	0	4	0
卫生计生	3	3	144	0	0	0	12
劳动和社会保障	3	6	13	2	4	16	378
准确率	0.8957	0.8089	0.8181	0.8497	0.8593	0.8906	0.9130

易看出，优化后的分类模型在准确率上得到了显著的提高，尤其在城乡建设和劳动和社会保障方面有了显著的提高。根据表画出的柱状堆积图如下图 9

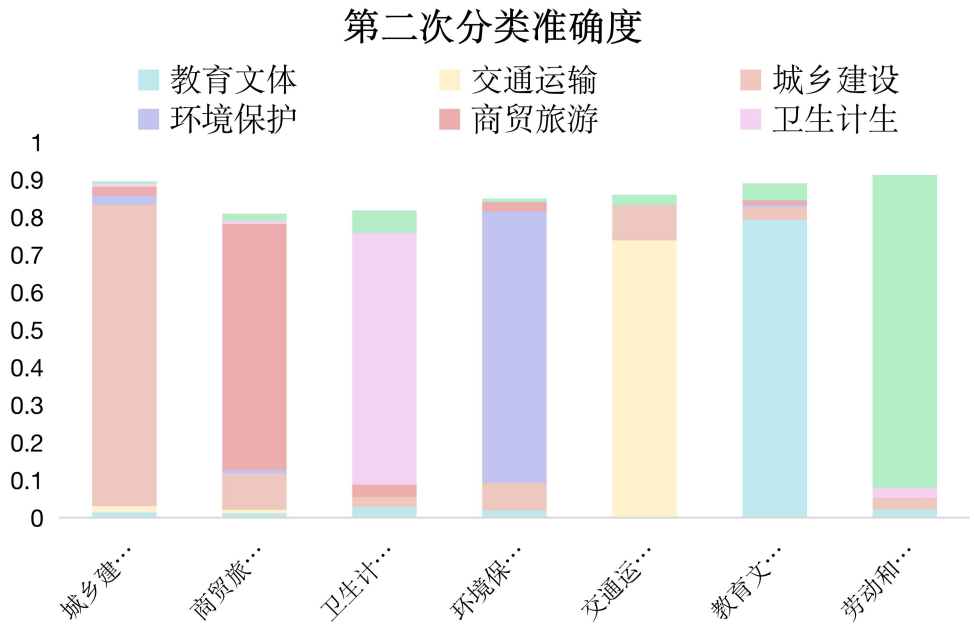

图 9 第二次分类准确度

6.2. 热点问题挖掘结果

6.2.1 提取地点结果

利用正则表达式和改进的决策树模型得到的地点提取结果如下表 5 所示。

表 5 地点提取结果

地点	数量	地点	数量
A 市 A3 区	215	A6 区	99
A 市 A6 区	73	A7 县 A1 区	3
A7 县	623	A1 区	97
A 市 A2 区	136	A 市 A5 区	117
A3 区	252	M 市	3
A 市	1546	A 市 K9 县	4
A 市 A7 县	57	A 市 L6 县	5
A2 区	158	A 市 A8 县	18
A 市 A4 区	129	C 市	4
A 市 A1 区	118	L 市	5
A4 区	131	A5 区	103
A8 县	73	A 市 M3 县	3

观察可得，地点主要集中在 A 市。

根据上表绘制柱状图 10

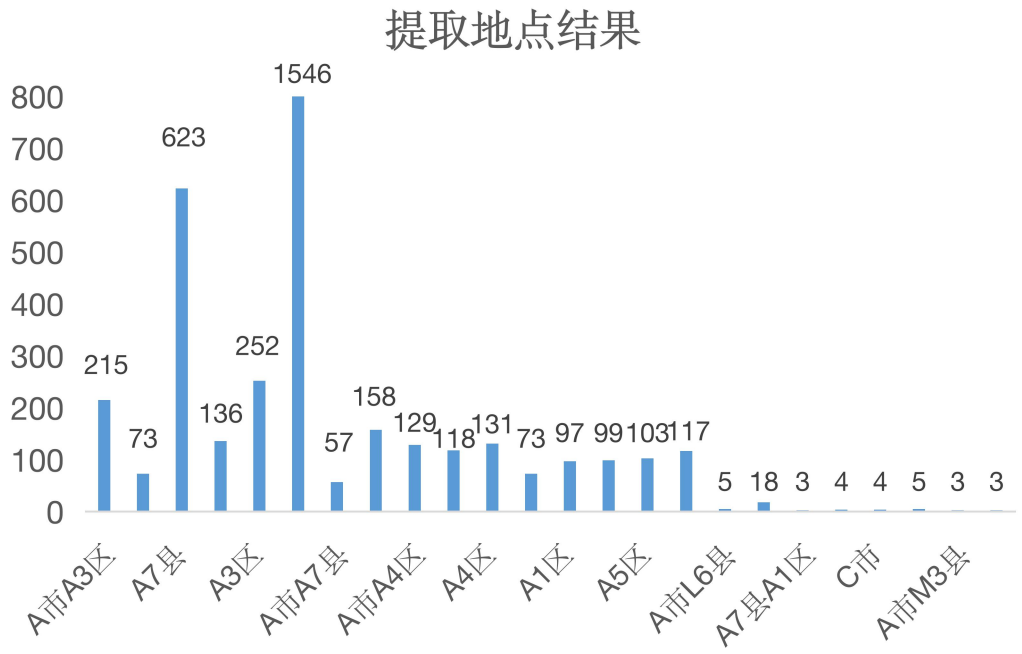

图 10 提取地点结果

6.2.2 计算热度值结果

对热度值进行计算，选取了前五个热点问题如下表 6 所示：

表 6 热点问题

热度排名	热度指数	地点/人群	问题描述
1	1.14440636	A6 月亮岛路	高架高压线安全隐患
2	0.893098599	A6 地铁 7 号线泉塘街道	规划地铁口，延长地铁
3	0.886017635	A7 松雅西地省站	人流量大安全风险
4	0.602736922	A5 魅力之城底层商铺	油烟进入小区噪声扰民
5	0.319126036	A2 福满新城	夜间通宵施工噪音扰民

6.3. 答复意见评价结果

根据数据能够绘制出完整性概率密度图 11

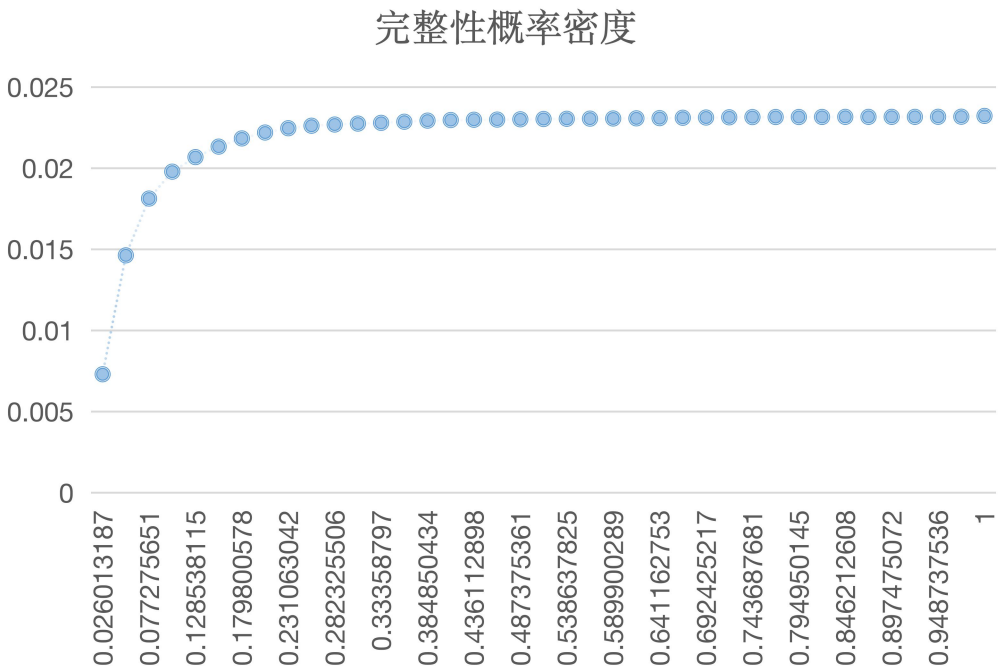

图 11 完整性概率密度

挑选了编号为 2549 的留言答复作为例子，如下表 7

表 7 答复评价例子

留言编号	留言主题	相关性	完整性	可解释性	得分
2549	A2 区景蓉华苑物业管理有问题	0.533419633	0.050573248	0.5	0.474233325

6.4. 实验总结

为了顺应大数据时代潮流的发展，建立基于自然语言处理技术的智慧政务系统，本文基于大数据技术的相关理论和实验，利用了 TF-IDF、改进的决策树、dbscan 等模型构建了智能文本挖掘模型。模型主要分成数据预处理、群众留言分类、热点问题挖掘和答复意见评价四个部分。在对赛题研究的基础上，我们根据研究思路撰写本论文，通过实验验证了本模型的可行性，基本实现本赛题设立的目标。

本文构建的智能文本挖掘模型基本完成了目标，能够提取出潜在的热点问题，帮助党政部门更好地了解民生民意，并针对相应的热点问题给出答复，有效提高了政府部门的施政效率。

参考文献

[1]《信访事项内容分类》说明,国务院,国信发〔2014〕7 号

[2]郝亚洲, 郑庆华, 陈艳平等. 面向网络舆情数据的异常行为识别[J]. 计算机研究与发展, 2016, 第 53 卷(3):611-620.

[3]徐绪堪, 华士祯. “互联网+政务服务”背景下的政务 APP 评价 ——基于直觉模糊层次分析法[J]. 情报杂志, 2020, (3):198-207

[4]刘擎权. 基于改进的 TFIDF 算法在文本分析中的应用[D].南昌大学,2019.

[5]赵胜辉，李吉月，徐碧瑢，孙博研.基于 TFIDF 的社区问答系统问句相似度改进算法[J].北京理工大学学报,2017,第 37 卷(9): 982-985

[6]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,第 39 卷(9): 1450-1461

[7]陈列蕾，方晖.基于 Scopus 检索和 TFIDF 的论文关键词自动提取方法[J].南京大学学报(自然科学版),2018,(3): 604-611

[8]A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester，Hans-Peter Kriegel，Jrg Sander.1996..

[9]Weibing Zuo,Yingli Li.A New Stochastic Restricted Liu Estimator for the Logistic Regression Model.统计学期刊(英文),2018,8(1):25-37.

[10]Zhang Shangli，Zhang Lili.Variable Selection in Logistic Regression Model[J].电子学报,2015,第 24 卷(4): 813-817

[11]Ernest Yeboah Boateng，Daniel A. Abaye.A Review of the Logistic Regression Model with Emphasis on Medical Research[J].数据分析和信息处理(英文),2019,(4): 190-207