基于文本挖掘的政务系统研究

摘要

本文利用互联网公开来源的群众问政留言记录以及相关部门对部分群众留言的答复意见，利用自然语言处理和文本挖掘的方法对数据进行分析。

针对问题一，首先观察群众留言文本数据集，对于每个类别，随机选取 1000 条留言文本进行文本挖掘，提取留言文本信息及留言类别。然后对留言文本进行分词处理，分类别词频统计，构建词袋模型及 PLDA 主题模型降维，转化为以主题为自变量和留言类别为因变量的分类问题。再运用逻辑回归算法对训练集进行建模训练，并在测试集上进行预测。根据测试集正确率、召回率、F 值等模型评价指标对模型进行评价。最终得出模型整体正确率为 80.67%，F-Score 的 F 值为 0.7267，模型预测效果还是不错的，且有高度的一致性。

针对问题二，本文对采集到的留言问题数据进行去噪声文本的处理，接着利用 Python 中 THULAC 中文分词工具包对留言问题文本进行分词处理，然后使用 TF-IDF 方法计算权重，根据权值抽取特征词，建立 VSM 向量空间模型，最后对各个主题的热点问题使用数据仓库技术分析，定义合理热度值指标以及评价结果。

针对问题三，本文从答复的相关性、完整性、时效性、可解释性四个方面入手，为直观的刻画四个指标以得到最终评价，对四个指标采取量化处理的方式，量化指标相关性为相关度，完整性为相关词，时效性为答复时间长短，可解释性为事实说明性分词，将所有量化指标构建为最终评价矩阵，经矩阵运算后刻画其最终评价结果。

关键词：文本挖掘　PLDA 主题模型　逻辑回归算法　分词处理　评价矩阵

目录

1 挖掘目标

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。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

2 分析方法与过程

2.1 群众留言分类

2.1.1 留言文本数据预处理

（一）群众留言相关数据介绍

本文数据来自附件二，由于本文是对群众的留言内容进行自动分类建模，已知有不同的分类体系，根据已有文献，在留言类别代码表中的基于留言主题的分类，其中一级分类类目如下表所示：

表 1 基于留言主题分类中的一级分类类别

主题一级类别名称			
城乡建设	环境保护	交通运输	教育文体
劳动和社会保障	商贸旅游	卫生计生	

由表 1 可知，一级分类包括城乡建设、环境保护、交通运输、教育文体、劳

动和社会保障、商贸旅游和卫生计生 7 个分类，类别比较多，不利于进行有监督的留言内容自动分类。

（二）留言内容文本获取

为了研究留言内容自动分类问题，只运用所得的留言分类一级类目，目标类别变量为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生，为了保持训练集目标变量的平衡，有好的模型训练效果，本文在附件二的留言内容中抽取 3500 条留言信息，其中每个目标变量类别 500 条。由于留言问题来源不同，导致留言文本格式有差异。留言内容列举如下：

表 2 不同类别留言内容文本展示

Prfsn_Lvl_Al_Name	留言描述
城乡建设	反映 C4 市收取城市垃圾处理费不平等的问题
环境保护	抵制新城新世界小区附近乱建基站，重视居住环境与健康
交通运输	A 市暮云公共交通严重滞后
教育文体	A 市民办培训机构乱象丛生，温斯顿英语培训强设霸王条款
劳动和社会保障	A8 县畜牧系统退休工人何时才能领取退休金？
商贸旅游	A 市传销团队猖獗，请求政府整治
卫生计生	K3 县潘市镇卫生院为什么到现在还是私人承包经营？

2.1.2 留言内容文本分词处理与词频统计

（一）中文文本分词方法与结果展示

对获取的群众留言内容的文本做分词处理，首先进行文本分词，文本分词选用 Python 分词工具，其中包括 Jieba、SnowNLP、THULAC 以及 NLPIR 等。接下来，

对这些分词库做简单介绍。

Jieba 分词是最常用简单的 Python 分词工具。结巴分词是基于词典对文本进行扫描，生成所有词的有向无环图，找到基于词的最大切分组合，对于词典中没有的词，运用 HMM 模型，采用 Viterbi 算法计算，结巴分词可以完成词性标注以及关键词的抽取。SnowNLP 库，可以进行中文分词、词性标注、文本关键词提取、情感分析、基于贝叶斯模型的文本分类，以及文本相似度计算等多种功能，通过函数调用可以进行文本的基础分析，但是分析功能简单，需要完善。THULAC 库，可以进行中文分词和词性标注，该库是通过大量的中文语料库训练而成，准确率高，分词速度快。NLPIR 库，可以进行包含中文分词、词性标注、用户词典与新词发现和关键词提取的功能，可以实现文本可视化展示及数据处理加工。

在对本文的电信投诉文本进行分词处理时，运用目前通用的 Python 分析工具中的 Jieba 分词包对文本进行分词处理，Jieba 分词后的部分结果如下：

表 3 留言内容文本分词结果展示

Prfsn_Lvl_Al_Name	留言描述
城乡建设	反映 C4 市 收取 城市垃圾 处理费 不平等 的 问题
环境保护	抵制 新城新世界小区 附近 乱建 基站 重视 居住 环境 与 健康
交通运输	A 市 暮云 公共交通 严重 滞后
教育文体	A 市 民办 培训机构 乱象丛生 温斯顿 英语 培训 强设 霸王条款
劳动和社会保障	A8 县 畜牧系统 退休工人 何时 才 能 领取 退休金
商贸旅游	A 市 传销团队 猖獗 请求 政府 整治
卫生计生	K3 县 潘市镇 卫生院 为什么 到 现在 还是 私人 承包经营

对于分词后的结果，可以看到有些是对文本分析没有用处的词语，比如“问题”、“为什么”、“的”等，另外还有部分标点符号，所以，接下来去除停用

词，停用词在传统的停用词表上添加“问题”、“为什么”等词，以适用于本文数据。去除停用词后的部分分词结果如下所示：

表 4 分词结果去停用词展示

Prfsn_Lvl_Al_Name	留言描述
城乡建设	C4 市 收取 城市垃圾 处理费 不平等
环境保护	抵制 新城新世界小区 附近 乱建 基站 重视 居住 环境 与 健康
交通运输	A 市 暮云 公共交通 严重 滞后
教育文体	A 市 民办 培训机构 乱象丛生 温斯顿 英语 培训 强设 霸王条款
劳动和社会保障	A8 县 畜牧系统 退休工人 领取 退休金
商贸旅游	A 市 传销团队 猖獗 请求 政府 整治
卫生计生	K3 县 潘市镇 卫生院 现在 还是 私人承包经营

（二）留言内容文本分类别词频统计

本节根据城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生的不同类型的留言文本分词后的结果，每个类别分别做词频统计，按照不同类别统计词出现的频数，选取频率高的前 10 个词，结果如下：

表 5 城乡建设类词频统计 TOP10

城乡建设（词频 TOP10）							
改造	1561	清理	913	城管	742	解决	591
规划	1274	处理	903	绿化	719		
拆迁	1189	违章	792	建设	672		

从表 5 可以看出，对于城乡建设类的留言问题，改造、规划、拆迁、清理等词出现频率较高，可以看出在这些方面，群众的呼声较高，政府要重视这方面的

工作。

表 6 环境保护类词频统计 TOP10

环境保护（词频统计 TOP10）							
污染	891	噪音	718	环保局	548	生态	342
垃圾	813	污水	692	工厂	494		
排放	732	环境	673	超标	452		

从表 6 词频统计结果可以看出，对于环境保护类留言问题，污染、垃圾、排放等词，出现频率较高。政府要重视环境的保护，对于生态要友好，才能保证人与自然的和谐相处。

表 7 交通运输类词频统计 TOP10

交通运输（词频统计 TOP10）							
超载	591	交通	383	滴滴	302	出行	252
收费	528	快递	372	车辆	299		
道路	482	出租车	329	司机	281		

表 8 教育文体类词频统计 TOP10

教育文体（词频统计 TOP10）							
招生	1452	补贴	1173	补课	768	文化	671
教师	1358	职称	1098	教育	729		
转学	1280	机构	903	学校	701		

表 9 劳动和社会保障类词频统计 TOP10

劳动和社会保障（词频统计 TOP10）							
保险	1832	报销	1568	劳动	1278	就业	713
待遇	1768	合同	1561	津贴	1098		
医保	1709	退休	1429	养老	897		

表 10 商贸旅游类词频统计 TOP10

商贸旅游（词频统计 TOP10）							
传销	1145	旅游	908	市场	765	经营	509
故障	1089	消费	891	收费	729		
诈骗	1012	电梯	782	价格	671		

表 11 卫生计生类词频统计 TOP10

卫生计生（词频统计 TOP10）							
二胎	801	医生	602	超生	489	行医	289
生育	781	卫生	562	公共	475		
计生	671	医院	503	医疗	398		

同理，我们可以一并得出交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生的词频统计结果。这些不同类别之间的差异比较明显，属性各不相同，自动分类难度较大。

2.1.3 基于 PLDA 主题模型的特征处理

（一）PLDA 主题模型的主题数确定

PLDA 主题模型是无监督的算法，需要确定主题数目，按照主题模型中的困惑度指标进行主题数的确定，图 1 反映了文本在选择不同数目的主题后的不确定程度。根据困惑度的下降程度，设置该值为 50，多于基于规则的一级分类数目 7，足以表示群众留言文本的语义信息。主题模型表示文本信息，即：文本以一定概率选择若干个主题，每个主题以一定的概率选择词。即：

$$P(\text{单词}|\text{文本}) = P(\text{主题}|\text{文本}) \times P(\text{单词}|\text{主题})$$

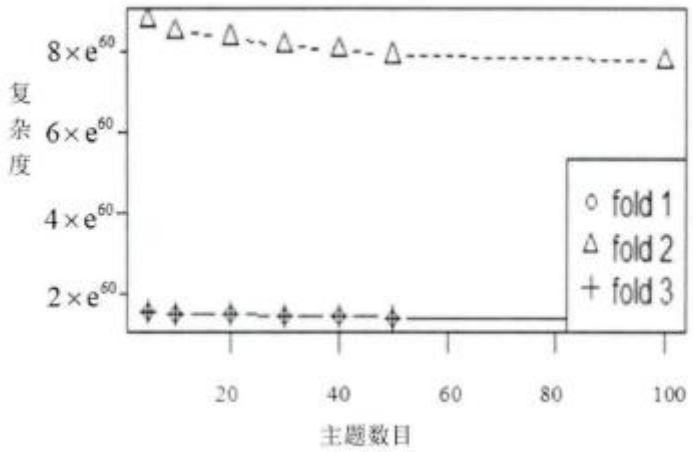

图 1 复杂度与主题数目趋势图

（二） PLDA 主题模型的结果解释

本节首先介绍各个主题词频较高的前 5 词，并根据部分主题词结合留言类别进行解释，另外介绍主题模型结果中的 P（主题|文本）及 P（单词|主题），最后根据主题模型各主题出现的权重，得到目前城乡建设是群众关注的主要问题。

（1）留言文本主题模型不同主题的词频较高的前 3 词如表 12 所示：

表 12 留言文本各主题高频词展示

Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
改造	污染	超载	招生	保险	传销	二胎
规划	垃圾	收费	教师	待遇	故障	生育
拆迁	排放	道路	转学	医保	诈骗	计生

表 12 展示主题模型的词频前 3 词可以看出，部分主题语义特征明显，如主题 1，污染、垃圾等词，说明与环境保护密切相关；主题 6，二胎、生育等词，说明与卫生计生相关；可见主题模型可以很好的概括文本的语义，但较多主题语义模糊，单从主题高频词并不能直接的将主题总结为留言类别，但对于后续自动留言文本分类过程可以有效的进行词表降维特征处理。

（2）表 13 为文本“A8 县灰汤镇枫木桥村全倒户可以申请危房改造资金吗”的 P（主题|文本），反映了该文本对应不同主题的概率。

表 13 文本的 P（主题|文本）

Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
0.7612	0.0398	0.0398	0.0398	0.0398	0.0398	0.0398

从表 13 可以看出，该文本以 76.12%的概率选择主题 0，主题 0 词频包括改造、规划、拆迁等，由此可见该文本所表达的语义是与形成主题 0 的一系列的词的语义一致。

（3）表 14 为词“危房”的 P（单词|主题），对应了该词在 7 个主题下的概率值，如下所示：

表 14 “危房”的 P（单词|主题）

Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
3.38E-08	0.0011	2.75E-08	3.65E-08	3.26E-08	3.75E-08	3.54E-08

表 14 显示了词“危房”在不同主题下的概率分别，在主题 0 的条件下，概率为 3.38×10^{-8}；在主题 1 的条件下，概率为 0.0011；在主题 2 的条件下，概率为 2.75×10^{-8}等等。P（单词|主题）反映了主题模型中的每个主题以多大的概率选择该词，从表 14 可以看出，若出现“危房”，则有较大概率为主题 1，且若文本是反映主题 1 的，则“危房”在其词表范围内。

从表 13 和表 14 可以得到，文本以 76.12%的概率选择主题 0，主题 0 以 3.38E-08 的概率选择词“危房”，文本通过选择若干主题，主题选择词语而形成。

（4）表 15 为 P（主题），如下所示：

表 15 主题模型概率展示

Topic	概率 P	Topic	概率 P
0	0.2181	4	0.2138
1	0.1018	5	0.1319
2	0.0666	6	0.0952
3	0.1725		

表 15 显示了 7 个主题的概率分布，概率之和为 1，显示了各个主题在整个留言文本中的权重。主题 0 在留言文本中出现的概率为 0.2181，总体来看，主题 0 在文本中的概率最大，权重较大。另外，主题 2 在留言文本中出现的概率为 0.0666，主题 2 在文本中的权重较小。说明城乡建设问题是目前群众不满的主要问题，急需得到解决。

2.1.4 基于 PLDA 主题模型的文本分类研究

（一）基于类别的评估指标

对于每个分类类别而言，选用敏感性/召回率、特异性、精确率、F 值、F-Score 对分类算法进行评估，定义如下:

（1）敏感性：TP/（TP＋FN），正确预测的正样本占全部正样本比例；

（2）特异性：TN/（TN+FP），正确预测的负样本占全部负样本比例；

（3）精确率：TP/（TP＋FP），表示正样本预测的准确率；

（4）F 值：$\frac{2\times precision\times recall}{precision+recall}$，综合度量模型精确率和召回率的评价指标；

（5）F-Score: $F=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$，其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

（二）逻辑回归分类算法预测效果展示

运用 Python 的 Scikit-Learn 中的逻辑回归类别的库 LogisiticRegression 对文本进行训练，本文的留言文本每个类别共收集了 9210 条数据，现在按照 7：3 的比例随机地将数据集分为训练集和测试集。

对训练集的 PLDA 主题模型的 P（主题|文本）进行训练，主题作为特征变量，留言类别为因变量，训练逻辑回归模型，逻辑回归的回归系数可表示留言文本类别与主题变量的相关关系，列举各个类别系数较高所对应的主题变量，表示如下：

表 16 各类别变量对应系数较高主题变量表示

类别变量	主题名称	回归系数	类别变量	主题名称	回归系数
城乡建设	Topic0	24.6734	劳动和社会保障	Topic4	20.6053
	Topic2	9.9865		Topic0	13.8730
环境保护	Topic1	18.6235	商贸旅游	Topic5	15.6902
	Topic6	10.9865		Topic2	10.0629
交通运输	Topic2	23.8529	卫生计生	Topic6	12.8043
	Topic5	18.8954		Topic4	7.0946
教育文体	Topic3	17.0943			
	Topic4	15.4509			

由表 16 可知，城乡建设与主题 0 相关关系较高，与表 13 主题词展示效果一致。本文运用主题模型计算所得的文本中不同主题的概率作为特征变量是适合留言文本的分类处理的，主题模型效果可以很好的展示各类别的文本语义信息。另外，从逻辑回归系数可以得到，城乡建设、交通运输分类正确率更高，其与主题模型相关性更显著。

运用逻辑回归模型在测试集上进行预测，将测试集上预测结果与真实类别进行对比，得到的比例矩阵如下表所示：

表 17 逻辑回归算法比例矩阵

	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
城乡建设	0.526	0.286	0.05	0.013	0.065	0.068	0.04
环境保护	0.258	0.435	0.027	0.024	0.069	0.068	0.069
交通运输	0.019	0.042	0.858	0.008	0.013	0.014	0.047
教育文体	0.036	0.044	0.023	0.97	0.076	0.018	0.093
劳动和社会保障	0.098	0.104	0.059	0.034	0.743	0.047	0.039
商贸旅游	0.08	0.098	0.013	0.007	0.027	0.813	0.047

卫生计生	0.056	0.087	0.065	0.075	0.028	0.501	0.069

根据逻辑回归比例矩阵计算分类模型发各项评价指标，如下所示：

表 18 逻辑回归算法各类评估指标

	Sensitivity/Recall	Specificity	Precision	F	F-Score
城乡建设	0.8267	0.9167	0.8556	0.9498	0.8689
环境保护	0.4323	0.93	0.493	0.4609	0.3809
交通运输	0.8578	0.9803	0.899	0.8798	0.8787
教育文体	0.9104	0.9619	0.8289	0.8667	0.7989
劳动和社会保障	0.7457	0.5308	0.6894	0.4117	0.6509
商贸旅游	0.6165	0.5589	0.7934	0.6078	0.7598
卫生计生	0.8974	0.9408	0.8609	0.8056	0.749

通过模拟训练评估可以看出，类别中的城乡建设、交通运输、教育文体等类别，预测效果较好，在 80%左右，而环境保护、劳动和社会保障预测效果较差。

模型整体正确率为 80.67%，F-Score 的 F 值为 0.7267，模型预测效果还是不错的，且有高度的一致性。

2.2 热点问题挖掘

本次方案中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了分析,然后对系统各个关键功能组件做详细描述,从建模到文本处理,再到聚类算法设计,最后是热点发现和分析。在此过程中对用到的技术和问题逐一进行了分析,定义合理的评价指标，并提出了对应的评价结果。

2.2.1 系统设计原理

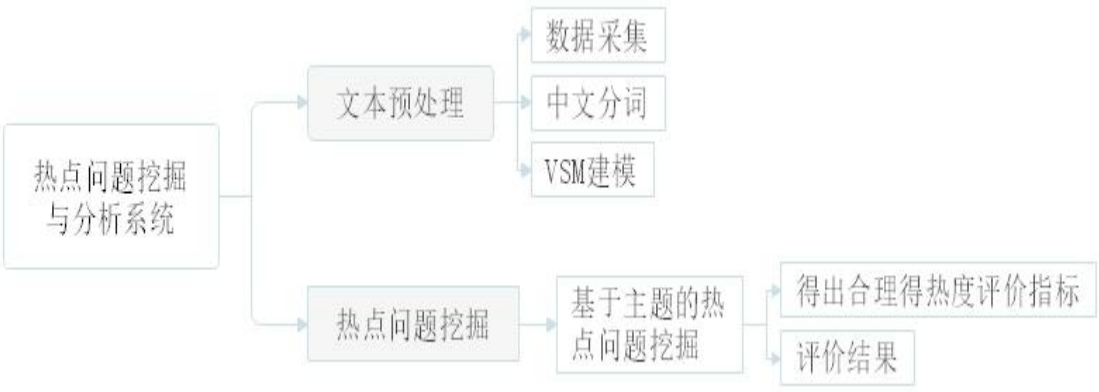

图 2 系统流程图

通过对问题文本对象的特点和问题需求的分析,得到了如图 3-1 所示的系统工作流程图,概括描述如下:

（1）从群众留言反映中获取 Excl 格式的相应的问题数据；

（2）对采集到的问题数据进行预处理,在过程中去除噪声文本；

（3）使用基于词典的分词方法对留言问题文本进行分词处理；

（4)使用 TF-IDF 方法计算权重,按权值排列顺序抽取特征词,建立 VSM 向量空间模型；

（5）对各个主题的热点问题使用数据仓库技术进行分析，定义合理热度值指标,并得出其评价结果。

2.2.2 中文分词模块

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,对于中文来说,其的特殊之处在于词间没有分隔,因此分词就显得很有必要了,而且后续进行的文本向量化的前提也是要先对文本分词,其次,分词的效果直接影响着聚类分析的质量。

经过相关资料分析,本次方案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工具包--THULAC。THULAC（THU Lexical Analyzer for Chinese）由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。THULAC 具有如下几个特点：能力强。利用我们集成的目前世界上规模最大的人工分词和词性标注中文语料库（约含 5800 万字）训练而成，模型标注能力强大。准确率

高。该工具包在标准数据集 Chinese Treebank(CTB5)上分词的 F1 值可达 97.3%，词性标注的 F1 值可达到 92.9%，与该数据集上最好方法效果相当。速度较快。同时进行分词和词性标注速度为 300KB/s，每秒可处理约 15 万字。只进行分词速度可达到 1.3MB/s。

对于本次方案,我们使用了基于 Python 版本的中文分词方案。在使用 THULAC 之前，引入相应需要的包后，应用 THULAC 工具包的测试代码如图 3 所示：

图 3

分词测试的结果如图 4 所示:

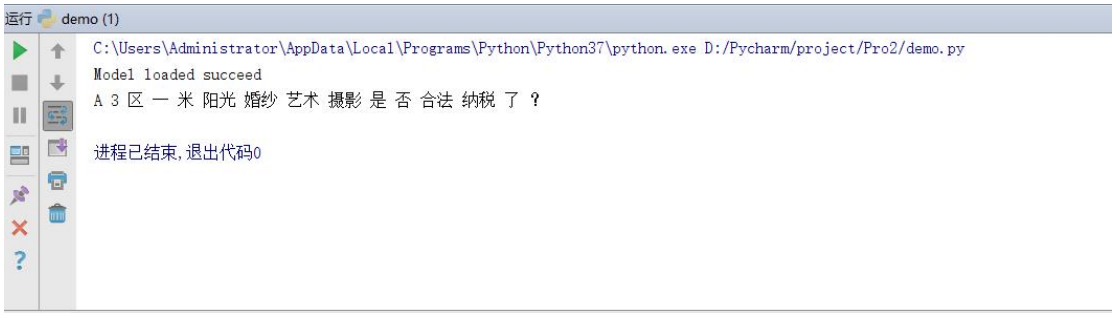

图 4

2.2.3 留言问题建模

（一）数据预处理

文本预处理一般包括:去除特定格式标记、滤除停用词等数据清洗步骤。

经过文本预处理,不仅能够去除噪声文档或垃圾数据,还能滤除对聚类无明

显作用的特征词,达到减少特征向量维数的目的。对于本方案的数据对象投诉文本来说,在对这类文本分析的过程中发现,单条留言问题常包含些无实际意义的词,一般不包含有价值的信息,比如“了”、“问题”、“的”、“现象”等。对此,在分析问题的基础上对这些价值量少的问题进行特征总结,生成一定的过滤规则,最后对留言问题基于这些规则进行预处理。

留言问题作为一种短文本对象,具有向量稀疏性、不规范性等特点,传统的文本处理方法在短文本表示方面遇到很大的困难,随后介绍的特征降维方法能在一定程度避免特征向量稀疏带来的问题。

由于经过文本分词得到的特征向量经常包括庞大的维数,过多的特征项包含的“噪声词”可能降低文本向量的区分能力,而且过多的特征项也会导致向量稀疏的问题,在分析处理中增加计算的复杂性,不仅浪费运行时间和存储空间,效果也不太理想。因此迫切需要降低特征空间的维数,只保留特征最突出的一些特征维度来降低计算量。降维之前,首先要去掉出现频率过小的词,即要先根据待处理文本的总量及分布情况,设定一个可以过滤文本词的最小阈值,大于该阈值的词将被保留作为特征词,接下来进行降维的处理。

目前常用的特征降维方法有:主成分分析法、潜在语义分析法、概念索引法、自组织特征映射法和多维标度分析法(Multidimensional scaling,MDS)等,这些方法进行降维的原理大都是利用词间的依赖关系进行合并,但是随之产生问题是计算量比较大,时间复杂度可以表示为O(m*k),其中m等于降维之前的原始特征数目,第二项k表示最终降维后得到的特征数目,可见其计算量与文本集合的规模成正比。

在降维形成特征向量之后,一般需要评价函数,对特征向量集中的每一个项进行独立评价,然后再对所有的特征向量根据评价值进行降序排序,从中选取前N个特征项组成最终的特征项集。在文本处理中,信息增益、文档频数、互信息、期望交叉熵等是几种常用的评估函数形式。本方案的初始特征向量来自文本分词后的词语,为了提高运算速度与准确度,本方案使用人工方法及 TF-IDF 算法,去掉在统计上对聚类影响不明显的词,对向量进行了降维输出。

（二）文本相似性度量

为了表示不同类别对象间的相似程度,需要先定义一些统计量来度量,常用

的相似度度量标准有距离和相似系数两种。

首先,距离是文本聚类中常用的相似度衡量标准。对于有 P 维特征向量的文本集合来说,N 条文本记录可看作是 P 维特征空间中的 N 个点,一般可用点间距离来度量文本记录间的相似度。第个 x 文本和第 y 个文本间的距离定义为 $L_{x,y}$,并且距离 $L_{x,y}$,应满足如下的条件:

(1) 距离非负,也就是对于任意下标 x 和 y,恒有 $L_{xy} \geqslant 0$,而且等号仅在两个文本 P 个维度对应值都相等时才成立。

(2) 距离对称,对所有的下标 x, y 有 $L_{xy} = L_{yx}$。

(3) 满足不等式:对所有下标 x, y, z 不等式 $L_{xy} \leqslant L_{xz} + L_{zy}$ 恒成立。

由上述几条性质可以得出,两个文本的距离必定在 $0 \rightarrow \infty$ 范围内,并且距离值越小,表示两个文本就越接近。在文本聚类过程中,采用常用马氏距离如下:

马氏距离:明氏距离一般适用于欧式空间,通常认为欧式空间中各维度之间完全独立。但考虑到文本中各特征向量的观测值常常是随机变量,那么随机向量会呈现出一定的分布规律,文本向量的各分量之间是相关的,当需要考虑分量之间的相关性时,一般使用马氏距离,定义为:

$$L_{xy}(M) = \left((C_x - C_y)^T \sum{}^{-1} (C_x - C_y)\right)^{\frac{1}{2}}$$

其中 Σ 是协方差矩阵。

其次,对于 VSM 空间中的两个向量,还可以用相似系数来表示其相似程度。假设第 x 个和第 y 个向量的相似系数定义为 F_{xy},则 F_{xy} 有以下的性质:

(1) 绝对值应小于 1,对于所有下标 x, y,必须满足 $|F_{xy}| \leq 1$,且等号仅在两向量存在线性关系时,也就是可以表示为 $C_x = kC_y$ (其中 k 是不为零常数)时才成立。

(2) 对称性质,对任意不同的 x, y,必然有 $F_{xy} = F_{yx}$。

两文本向量之间的相似系数一般有两种表示形式:

(1) 夹角余弦形式:P 维空间中两个向量基于相对位置会形成一个夹角,相似系数可以用这个角的余弦值来衡量,据此第 x 个和第 y 个文本的相似系数可定

义为:

$$\sin(x,y)=\frac{\sum_{k=1}^{P}V_{zx}V_{zy}}{\sqrt{\sum_{k=1}^{P}V^{2}_{zx}\sum_{k=1}^{P}V^{2}_{zy}}}$$

其中第 i 个文本向量表示为:

$$v_i=(V_{1i},\cdots,V_{zi},\cdots,V_{Pi})$$

（2）相关系数:第x个向量和第y个向量之间的相关系数可表示为

$$r_{xy}=\frac{\sum_{k=1}^{P}\left(V_{zx}-\bar{v}_x\right)\left(V_{zy}-\bar{v}_y\right)}{\sqrt{\sum_{k=1}^{P}\left(V_{zx}-\bar{v}_x\right)^2\sum_{k=1}^{P}\left(V_{zy}-\bar{v}_y\right)^2}}$$

对空间中向量进行两两组合计算距离或相似系数,用矩阵的形式进行排列和组织就形成了 P 维系数矩阵,它详细地描述了空间中对象的不同接近程度,,而且由距离与相似系数的定义可知,这个系数矩阵必然是对称的,这就为文本聚类运算带来了很大的便利,将抽象的问题转换成了对具体矩阵数值的运算。

2.2.4 基于主题的热点聚类

聚类分析是文本挖掘的一个重要分支,经过聚类可以发现隐藏在大量文本中的有用的数据分布模式,其优点在于聚类分析能够不依赖训练知识而直接得到有用的类别或聚簇。因此,聚类又被称为无监督学习,也就是指在输入的文本对象没有初始的类别标记,完全要由聚类算法来自动计算生成。

文本聚类是将文本数据的整个集合按照某种规则分成几个类别或簇,每个类或簇中的数据都在一定程度上相似。文本聚类分析是通过研究文本集合自身的特性,对于待处理的数据对象进行类别划分的方法。总的原则是要使得同类文本最大程度地相似,同时属于不同类的文本要在最大程度上相互区分。

本节主要讨论如何通过具体的聚类算法来发现问题反映的分类聚集情况,以便下一步对特定主题的热点进行深入分析。在上一章对各种常用聚类算法比较分析中可知,各种算法都有各自的优缺点,结合已经主题的投诉对象数据的特点,本文提出了一种基于主题关键词的 K- means 聚类算法,期望通过该算法准确地将 w

问题文本归类到给出的主题类别当中,接下来对该算法的设计做详细的说明。

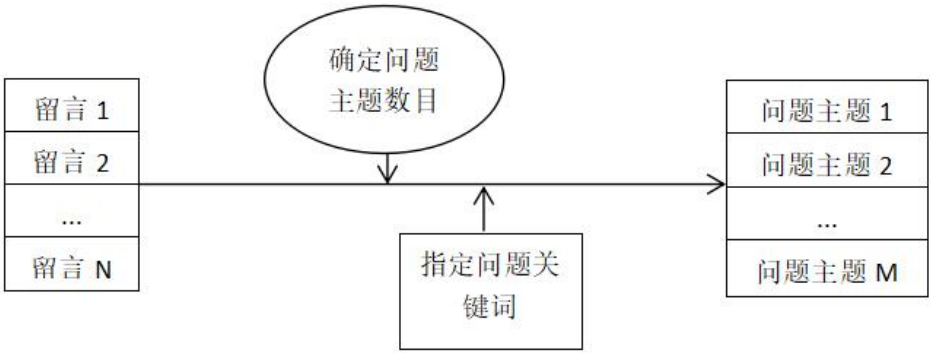

图 5 基于主题的热点类聚

（一）聚类算法设计与优化

K-means 聚类分析算法通常也叫快速聚类算法，是现在聚类分析中最常使用的算法之一。这种聚类算法的思想是先对数据对象进行统计分析，再将它们划分到确定的类别中去。 K-means 算法对于大规模数据的类别划分比较有效，但采用此方法的不足是需要预先给定聚类的数量以及聚类的初始中心位置，这样，此算法就在很大程度上依赖于人为的经验和判断。

在目前的聚类研究中，怎样估计聚类的数目还是一个未解决的问题。一些算法中普遍使用模型选择法来得到聚类数目，对于基于划分的 k- means 聚类算法，可以根据贝叶斯信息准则(Bayesian Information Criterions,BIC)估计类的数目。据此，本文提出按照贝叶斯信息准则函数判断聚类数目的标准，可减少人为决定聚类数目带来的主观影响，使聚类的结果更能客观地反映数据内部特征，而且实验证明，BIC 准则估计聚类数目类的误差很小，本文将通过 BIC 准则来优化 K- means 算法。

具体优化算法如下：

(1) 根据给定的候选模型集，从中选择 BC 值最大时对应的那个函数模型 $f(M_i|\omega)|$。

(2) 执行普通的 K-means 算法，只是在确定样本类别归属时进行以下的判断:计算当前函数模型对应的 BC 来测量未问题主题的数目是否可以再进行优化，即当两个类不能确定是否需要合并时，就比较合并和不合并两种情况下整个样本

集的 BC 值，哪种情况下的 BLC 值更大，就确定那种情况下划分类的方法。

(3) 当(2)中 BC 值达到最大时，表示目前的问题主题数目是在 BIC 准则下最优的。

(4) 根据得到的问题主题数目 k，由用户指定 k 个问题主题的关键词，并对每个主题选择出一个初始的聚类中心，执行 K-means 算法，由于聚类数目和初始中心都不再是随机指定的，因此该算法的聚类效果优于一般的未改进的 K-means 聚类算法。

实验结果：在聚类质量评价中经常使用的评估聚类效果的指标有查准率(precision)、查全率(recall)和 F-measure，其中 F-measure 是由前两个指标运算得到的，是结合两者进行评估的综合指标，如果 F-measure 值越大表明聚类结果越好。分别定义为:

$$precison(i,j)=\frac{N_{ij}}{N_i}$$

$$recall(i,j)=\frac{N_{ij}}{N_j}$$

$$F(i)=\frac{2\times precision(i,j)\times recall(i,j)}{precision(i,j)+recall(i,j)}$$

其中，N_{ij} 代表第 j 个结果簇中属于第 i 个原始分类的样本个数，N_j 为第 j 个结果簇的样本总数，N_i 表示第 i 个原始分类的样本总数。

（二）热点问题挖掘与分析

采用基于密度的聚类后,就对向量空间中的留言问题进行了初步主题的划分,下一步是根据算法对可以代表热点问题的特征词进行挖掘,再通过热度计算公式来识别这些热点问题，进而得出热度评价指标为查准率、查全率和 F-measure。最后，我们根据这三个热度评价指标，得出评价结果由大到小依次为：①A 市各县区和西地省的公司企业存在欺诈、拖欠工资等违法犯罪行为；②A 市各县区的地铁建设存在漏洞和违规；③A 市各县区的小区物管违法、不作为现象；④A 市各县区的小区出现扰民现象；⑤A 市各县区幼儿园出现违规、收费合理

性以及招生问题。

表 19 热点问题表

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	179	2019/1/11 至 2020/1/7	A 市各县区和西地省的公司、企业	公司企业存在欺诈、拖欠工资等违法犯罪行为
2	2	113	2019/1/15 至 2020/1/7	A 市各县区的地铁	地铁建设存在漏洞和违规
3	3	91	2019/1/11 至 2020/1/6	A 市各县区的小区物管	物业违法、不作为现象
4	4	87	2019/1/4 至 2020/1/4	A 市各县区的小区	小区出现扰民现象
5	5	78	2019/1/12 至 2019/9/28	A 市各县区幼儿园	幼儿园出现违规、收费合理性以及招生问题

表 20 热点问题留言明细表

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详细	点赞数	反对数
1	188895	A00063289	票牛 A 市分公司不肯退我草莓音乐节的	2019/6/28 9:31:25	22 日 A 市草莓音乐节的票因为主办方的问题，短信通知我 21 号下午 6 点前把票寄回原公司可以	0	0
1	190462	A00027173	A 市发顺驾驶员培训有限公司教练欺诈	2019/3/5 0:31:47	您好！关于驾校教练和学员的纠纷对于您来说，确实是一件微不足道的小事目三科目四，所以这个	0	1
1	190957	A909140	西地省晨鹭互联网科技有限公司涉嫌	2019/11/12 14:58:30	西地省晨鹭互联网科技有限公司涉嫌长期专业从事电话窃听，网络黑客非法窃取公民信息和盗	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5	286562	A00013299	A3 区梅溪湖卓越浅水湾小区的幼儿园	2019/11/22 16:07:05	本人是 A3 区梅溪湖卓越浅水湾业主，家中小孩已经三岁多。到了上幼儿园的年纪，但是小区的幼儿	0	1

2.3 答复意见的评价

对于问题 3，我们针对附件 4 相关部门的答复意见，决定从答复相关性、完整性、答复时效性、可解释性四个方面入手，对答复意见进行指标量化处理以得到评分，以评分高低排序得出评价，量化指标的思路如下：

2.3.1 数据的预处理

在量化指标前，我们先通过 jieba 分词库，分别对附件 4 中的留言详情和答复意见进行分词筛选，得到数组。

筛选掉所有非法字符，去除停用词，得出与本文相关性强的分词整合成特征向量，并将同一留言编号下的留言及答复整合为同一维度的向量量级：

其中，我们设留言向量为 A_i{a1,a2,...,an},答复向量为 B_i{b1,b2,...,bn}

2.3.2 答复完整性的量化

我们将完整性理解为对留言意见的涉及程度，同一关键词下，当答复分词相对于留言分词的占比越大，其完整度越高，对完整性的量化方式如下（其中 CY 为完整性量化评分）：

$$C_j = 1 - \frac{A_i + 1}{B_i + 1}$$

$$C_Y = \frac{\sum C_j}{j}$$

2.3.3 答复相关性的量化

我我们将相关性理解为对问题要点的相同程度，当两个向量共有的越多，其相关性越强，因此，我们决定将两个向量所共有的分词进行公式计算得到量化指标，对向量内部取 0 或 1，表示是否有该分词存在于向量中，例如，在留言编号为 2549 中，我们得到的向量分别为:

留言向量:A1 {0,1,0,1,1,0,1,1}

答复向量: B1 {1,1,0,0,1,0,1,1}

通过以下方程（其中 CX 为相关性量化评分），

C_i = A_i and B_i

$$C_X = \frac{\sum C_i}{i}$$

2.3.4 答复时效性的量化

我们将时效性理解为答复用时，使用 datetime 模块对日期进行计算，通过差值计算答复所花费的时间，公式如下（其中 T_X 为所花费的时间，T_A 为答复时间，T_B 为留言时间）：

$$T_X = T_A - T_B$$

我们将所求出的时间的量级进行同化，将时间以分钟单位折算，进行下一步计算：

我们希望通过一个指标来反映答复时效性，为解决这个问题，我们引入如下公式（其中 S_i 为时效化指标，T_{MAX} 为最迟答复时间）：

$$S_i = 2 - \frac{2T_X}{\overline{T} + T_{MAX}}$$

2.3.5 答复可解释性的量化

对于答复可解释性的理解，我们认为，可解释即是对留言所作出的实际调查、走访、记录、相关部门问责等一系列以事实依据作为内容的答复，我们将这一类关键词称为解释性分词，对于这一类的信息，量化如下（其中 C_T 为量化指标，C_Z 为解释性分词，S_Z 为去停用词后的分词总数）：

$$C_T = \frac{C_Z}{S_Z}$$

2.3.6 评价量化矩阵

对以上四个方面，我们构建了评价矩阵 K_i，将以上的完整性、相关性、时效性、可解释性的四个维度指标置于该矩阵中：

$$K_i = [C_Y, C_X, T_X, S_i] \Big/ 4$$

通过矩阵运算：

$$K = K_i * K^T$$

我们计算出了每个答复的评价分数。

表 21 答复意见的评价表

留言编号	留言用户	相关性	完整性	时效性	可解释性	最终评分
2549	A00045581	0.5	0.613095238	0.99896154	0.018162937	0.406034956
2554	A00023583	0.375	0.575	0.999026066	0.018164111	0.367408254
2555	A00031618	0.125	0.382440476	0.999352223	0.01817004	0.290230183
⋮	⋮	⋮	⋮	⋮	⋮	⋮
185986	UU008363	0.25	0.541071429	0.999083534	0.018165155	0.338439043

3 结论

本文通过观察群众留言文本数据集，对于每个类别，随机选取 1000 条留言文本进行文本挖掘，提取留言文本信息及留言类别。然后对留言文本进行分词处理，分类别词频统计，构建词袋模型及 PLDA 主题模型降维，转化为以主题为自变量和留言类别为因变量的分类问题。再运用逻辑回归算法对训练集进行建模训练，并在测试集上进行预测。根据测试集正确率、召回率、F 值等模型评价指标对模型进行评价。最终得出模型整体正确率为 80.67%,F-Score 的 F 值为 0.7267，模型预测效果还是不错的，且有高度的一致性。

利用 Python 中 THULAC 中文分词工具包对留言问题文本进行分词处理，然后使用 TF-IDF 方法计算权重，对各个主题的热点问题使用数据仓库技术分析，定义了三个合理热度值指标，得出评价结果由大到小依次为：①A 市各县区和西地省的公司企业存在欺诈、拖欠工资等违法犯罪行为；②A 市各县区的地铁建设存在漏洞和违规；③A 市各县区的小区物管违法、不作为现象；④A 市各县区的小区出现扰民现象；⑤A 市各县区幼儿园出现违规、收费合理性以及招生问题。

从答复的相关性、完整性、时效性、可解释性四个方面入手，对四个指标采取量化处理的方式，将所有量化指标构建为最终评价矩阵，经矩阵运算后刻画其最终评价结果。

参考文献

[1]时志芳. 移动投诉信息中热点问题的自动发现与分析[D].北京邮电大学,2013.

[2]廉素洁. 基于文本分类和情感评分的电信投诉文本挖掘研究[D].浙江工商大学,2018.

[3]刘宁,陈凌云,熊文涛.基于文本挖掘的网络热点舆情分析——以问题疫苗事件为例[J].湖北工程学院学报,2019,39(06):60-64.

[4]姜玉坤. 舆情热点信息挖掘技术的研究与应用[D].天津大学,2017.

[5]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.

[6]张媛,胡庆武.社交网络时空大数据聚类挖掘有效选择分析[J].测绘地理信息,2020,45(02):45-50.

[7]曹春萍,黄伟.基于用户权威度与热度分配聚类的微博热点发现[J].计算机工程与设计,2020,41(03):664-669.

[8]魏利梅.微博社交网络数据挖掘与用户权重分析[J].网络安全技术与应用,2019(12):72-74.

[9]阮光册.基于文本挖掘的网络媒体报道研究[J].图书情报工作网刊,2011(06):24-31.

[10]黄情. 基于文本挖掘的网络舆情分析应用研究[D].大连海事大学,2016.