“智慧政务”中的文本挖掘应用

摘要

“智慧政务”的发展大致经历了三个阶段：电子政务、互联网+政务、智慧政务，政务服务进一步向数字化和智能化的方向发展，“让信息多跑路，让群众少跑腿”,已经越来越多的政务服务领域变为现实。利用大数据人工智能等技术，能够提升我国政府的管理水平和服务效率。通过群众所给的问政留言记录，解决一些关于留言分类的问题，建立基于自然语言处理技术的智慧政务系统，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 1，运用 R 软件中的 gsub 函数及 segmentCN 函数对文本进行中文分词处理,运用 emoveStopWords 算法以及 lapply 函数的全文本扫描功能对文本进行停用词去除，得到新的群众留言信息。 运用 R 语言内 Knn 算法对其进行模型建立，根据生成的文本矩阵，选取原文本集数据的 75%作为训练数据，25%作为测试数据，根据训练数据内文本矩阵及已知分类进行训练，将测试文本矩阵作为分类依据进行分类。后根据预测结果与真实结果计算查准率及查全率，根据 F-Score 分类评价公式计算模型准确率。

对于问题 2，根据附件 3 所给文本数据提取摘要，提取文本数据中的关键词进行聚类，将信息内容相近的文本数据聚为一类，对各类别文本数据提取摘要并通过点赞数与反对数计算热度指数生成热点问题表，并据提取出的热点信息主要内容生成热点问题留言明细表。

对于问题 3，主要提取其中的留言详情及答复意见进行计算研究，从相关性进行分析，对照留言详情与答复意见在关键词之间的相关性计算，计算出相关性数据；设置计算各留言详情与答复意见之间的完整性数据及可解释性数据，建立层次分析评价模型，对所有文本的得分情况进行计算。

关键字：R 软件 Knn 算法 F-score 层次分析模型 热点问题

目录

1. 问题提出

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台已经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上升，给以往主要依靠人工进行留言划分和热点整理的相关部门的工作带来了极大的挑战，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。在营造政务新常态的今天，网络和政务服务已经连为一体，建设了高效、亲民、智能的信息平台，提供大众贴心满意的服务。从人工服务到智能化服务，运用大数据等科技技术，实现智能感知的政务服务，将社情民意工作做到最好。

在全面深化改革的持续发展中，我国政府全方位为群众提供优质、高效、便捷的公共服务理念，不断完善我国政务服务，加强政务服务体系建设。通过互联网收集得到的群众问政留言记录，及相关部门对部分群众留言，解决群众或相关部门的一些民政问题，及时给群众关于问政方面的答复，能够最大限度便企利民提供强有力保障。

1.2 智慧政务现状

智慧政务是利用现代信息技术提高政府智能化水平的一种形态，提高政府办公、服务、监管、决策的智能化水平，从而形成高效、集约、便民的服务型政府运营模式。智慧政务的建设是实现电子政务升级发展的突破口，是政府从管理型走向服务型，智慧型的产物。近年来，越来越多地区使用智慧政务规划编制工作，便企利民。随着我国电子政务发展已经过渡到智慧政务，大规模的硬件投入已经开始降低，软件及服务的需求逐步加大。在这互联网的时代中，信息技术和信息经济领域持续增长，为智慧政务发展提供了有效的技术支撑，推动了智慧政务向纵深发展。在国家的大力支持和推动下，当

前的电子政务正向智慧政务转型，未来的智慧政务必将有很大的成果。

1.3 问题重述

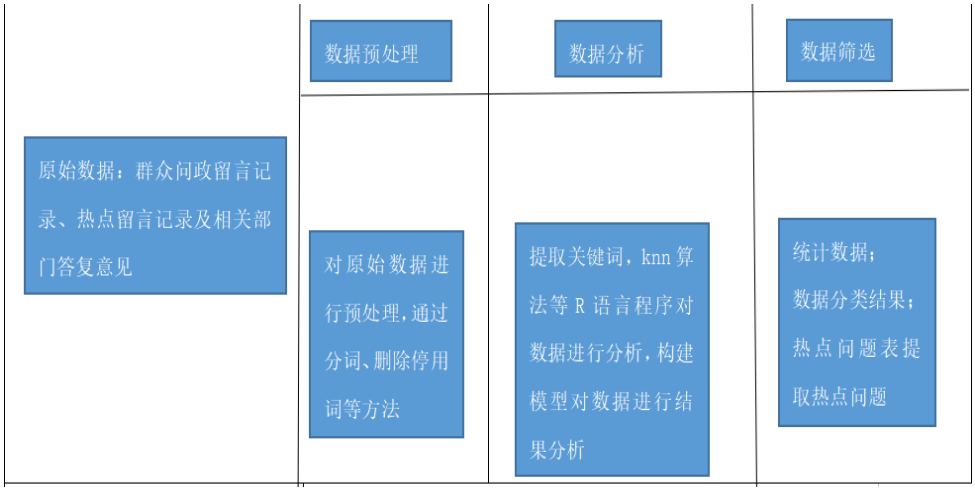
我国群众通过互联网的问政留言记录不断上升，通过利用自然语言处理和文本挖掘的方法，能够快速整理、解决这些问题。

问题一：群众留言分类。工作人员先按照一定的划分体系（三级标签体系），将网络问政平台的群众留言进行分类处理，方便派放至相应职能的部门处理，但是在处理过程中，也是通过人工处理，存在着工作量大，效率低，更可能还会存在差错率高等问题。根据所给的数据，建立关于留言内容的一级标签分类模型，对分类方法进行评价。

问题二：热点问题挖掘。群众在某一时段内集中反映的某一问题为热点问题，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。根据所给数据，将某一时段内反映特定地点或特定人群问题的留言进行分类，定义合理的热度评价指标。

问题三：答复意见评价。针对所给数据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，并尝试实现。

1.4 总体流程图

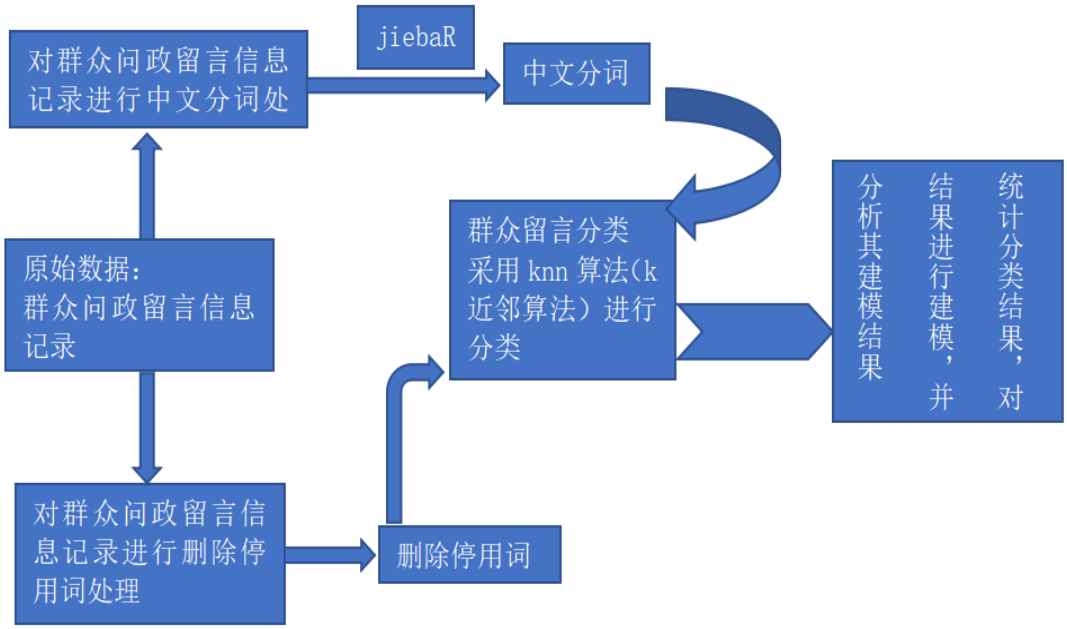


图一 总体流程图

2. 分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 流程图



图二 问题一流程图

2.1.2 数据预处理

在对群众问政留言信息进行挖掘分析之前，先把非结构化的文本信息转化为计算机能够识别的结构化信息。为了便于文本转换，将题目所给出的数据，在 R 软件中利用中文分词包 jiebaR，其中有四种模式，支持祖达概率法、隐式马尔科夫模型、索引模型、混合模型四种，同时又词性标注，关键词提取，文本 Simhash 相似度比较功能。运用 gsub 函数及 segmentCN 函数对文本进行中文分词处理，gsub 可以用于字段的删减、增补、替换和切割，可以处理一个字段，也可以处理由字段组成的向量。

在题目给出的数据中，有这一类词并不包含任何信息，如“的”、“了”等词，干扰了问题的分析，从而利用删除停用词的方法，将不需要的词从文本删除。先行设置停用词，停用词字典来源于网络，共涉及 1893 个词汇符号，停

用词存储在 txt 文本内，方便进行挺用词的删除；运用 emoveStopWords 算法以及 lapply 函数的全文本扫描功能对文本进行停用词去除。

将分词及去除停用词后文本数据集转换为向量矩阵形式，运用 R 软件安装包 tm 软件包中的 VCorpus 函数生成临时预料集，DocumentTermMatrix 函数基于 VCorpus 语料集对文本集进行矩阵化转换，将文本集转换为普通矩阵

2.1.3Knn 最邻近分类算法

由 K-Means 分类得到聚类中心，利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。根据向量空间模型，将每一类 别文本训练后得到该类别的中心向量记为$C_j(W_1,W_2,\cdots,n)$，将待分类文本 T 表示成 n 维向量的形式$T(W_1,W_2,\cdots,W_n)$，则文本内容被形式化为特征空间中的加权特征向量，即$D=D(T_1,W_1;T_2,W_2;\cdots,T_nW_n)$。对于一个测试文本，计算它与训练样本集中每个文本的相似度，找出 K 个最相似的文本，根据加权距离和判断测试文本所属的类别。具体算法步骤如下:

(1) 对于一个测试文本，根据特征词形成测试文本向量。

(2)计算该测试文本与训练集中每个文本的文本相似度，计算公式为:

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{\sum_{k=1}^{M}W_{ik}^2}\sqrt{\sum_{k=1}^{M}W_{jk}^2}}$$

式中，d_i为测试文本的特征向量，d_j为 j 类的中心向量; M 为特征向量维数: W_k为向量的第 k 维。k 值的确定一般先采用一个初始值，然后根据实验测试 K 的结果来调整 K 值。

(3) 按照文本相似度，在训练文本集中选出与测试文本最相似的 k 个文本。

(4)在测试文本的 k 个近邻中，以此计算每类的权重，计算公式如下:

$$\mathrm{P}(\mathrm{X},C_j)=\begin{cases}1，\ 若\sum_{d\neq Knn}Sim(x，\ d_i)y(d_i，\ C_j)-b\geq 0\\ 0，\ 其它\end{cases}$$

式中，x 为测试文本的特征向量; Sim(x, d_j)为相似度计算公式: b 为阈值，有待于优化选择:而 y(d_i, C_j)的值为 1 或 0,如果d_i属于C_j，则函数值为 1,

否则为 0.

(5)比较类的权重， 将文本分到权重最大的那个类别中。

2.1.4 群众留言分类

对问题一所提到的群众问题分类，重点对已收集到的数据表中的详细数据内容进行识别，依据详细内容主要内容进行分类，以附件 2 中所给数据分为训练数据与测试数据，进行机器学习并进行检测，计算模型 F-Score 分类指标值，以此观察模型分类效果，F-Score 值计算公式：

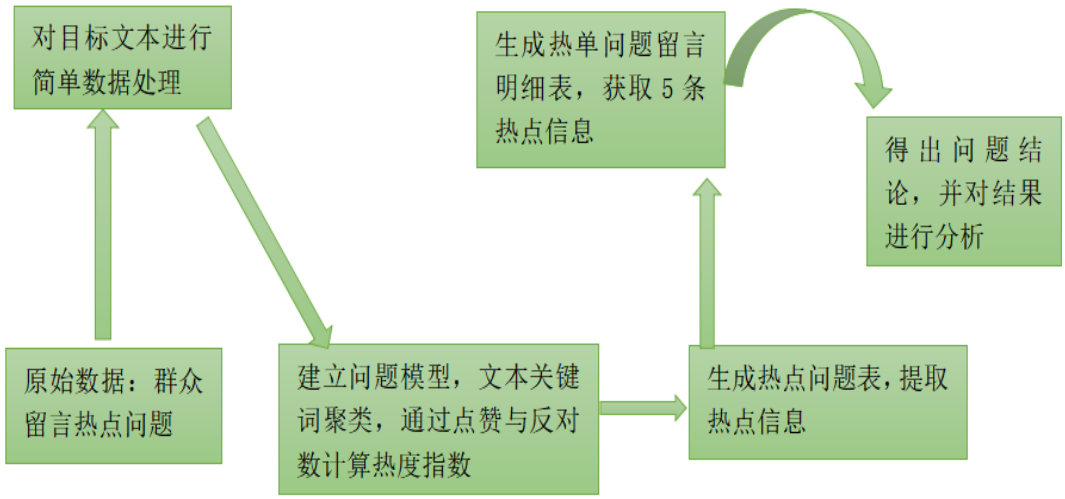
$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

分类模型采用 knn 算法（k 近邻算法），运用 R 语言内 knn 算法进行模型建立，根据生成的文本矩阵，选取原文本集数据的 75%作为训练数据，25%作为测试数据，根据训练数据内文本矩阵及已知分类进行训练，将测试文本矩阵作为分类依据进行分类，后根据预测结果与真实结果计算查准率及查全率，根据 F-Score 分类评价公式计算模型准确率。

所给附件数据共 9210 组数据，提取数据的详细内容及一级标签作为后续分类属性，所有数据中，一级标签分为 7 类，每类提取 75%数据作为训练数据，25%数据作为测试数据，分别生成“附件实验训练数据”与“附件实验测试数据”，并利用模型进行训练与准确率计算；建模过程中，R 软件并不能处理全 9210 组数据，程序无法运行，将训练数据及测试数据进行缩减，共保留训练数据 1908 组，测试数据 700 组。

2.2 问题 2 分析方法与过程

2.2.1 流程图



图三 问题二流程图

2.2.2 问题分析

为了能够及时发现热点问题，及时让相关部门进行针对性处理，提升服务效率，将群众的问政留言记录进行归类，并定义出合理的热度评价指标，对目标文本需要进行简单数据处理，目标获得文本的前五条热点信息，根据附件 3 所给文本数据提取摘要，提取文本数据中的关键词进行聚类，将信息内容相近的文本数据聚为一类，对各类别文本数据提取摘要并通过点赞数与反对数计算热度指数生成热点问题表，并据提取出的热点信息主要内容生成热点问题留言明细表。

2.2.3 问题建模

依据问题一所用到的文本数据处理方法，对附件 3 文本数据的留言主题进行分词及去停用词操作，程序提取附件 3 内变量留言主题的所有数据，对数据使用自编程的 get_zh_summary 算法摘要函数，该函数参考自网上论坛；根据算法内设置的 get_sentences 算法对数据进行单句分割，得到对数据划分为单句的结果；依据 filter_sentences 算法对分割单句后的数据进行关键句筛选，筛选出

过短的无意义句，简化运算；依据 get_words_list 算法获取每句中的关键词，以方便计算相关度；依据 get_similarity 算法计算句子之间的相似度，相似度计算原则为以两个句子相同单词的数量除以每个句子单词数量取 log 的乘积，以此生成一个相似度的矩阵，并做标准化处理；依据 textRank 算法函数计算最终的权重向量；最终依据主算法函数 get_zh_summary 算法，通过设置提取数值，迭代次数，并对以上各算法函数分别调用，最终获得目标摘要结果，计划提取十个关键句，并依照该十类计算点赞数及反对数，数据之和做为热点度。

2.3 问题 3 分析方法与过程

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，并尝试实现。对于附件 4 文本数据，主要提取其中的留言详情及答复意见进行计算研究，对相关性，主要分析对照留言详情与答复意见在关键词之间的相关性计算，相关于留言详情，以留言详情与答复意见之间相同的关键词数量除以留言详情总关键词数，计算出相关性数据；设置计算各留言详情与答复意见之间的完整性数据及可解释性数据，建立层次分析评价模型，对所有文本的得分情况进行计算；

3．结果分析

3.1 问题 1 结果分析

对提取出的 1908 组训练数据根据文本分类 R 程序经过 R 语言软件进行数据处理及机器学习，对提取出的 700 组测试数据根据文本分类 R 程序经过 R 语言软件进行数据处理及学习，机器学习方法为 knn 函数基础的 K 近邻算法，通过机器学习测试，得到如下表的测试结果，已计算查全率及查准率，根据 F-Score 值计算公式，运用 matlab 软件计算得该分类方式 F-Score 值=0.83557，及准确率为 83.557%，认为该模型准确率较好；

表一 分类表

编号	1	2	3	4	5	6	7	
类别	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	总计
100%	2009	938	613	1589	1969	1215	877	9210
训练提取	300	235	153	397	300	304	219	1908
测试提取	100	100	100	100	100	100	100	700
模拟分类	103	86	118	109	86	92	106	700
正确分类	86	79	90	88	75	80	87	
错误分类	17	7	28	21	11	12	19	
P	0.83	0.92	0.76	0.81	0.87	0.87	0.82	
R	0.86	0.79	0.9	0.88	0.75	0.8	0.87	

3.2 问题 2 结果分析

对附录 3 数据进行文本分析，运算程序为摘要 R 程序，运用 R 语言软件进行运行模拟，对数据进行分词摘要得到关键词结果有：“车贷”，“房子质量”，“问题”，“高铁”，“搅拌场”，“高压电线”，“诈骗”等，依据这些关键词通过 Excel 表格软件进行查找并做标记，将各相关留言及相关信息进行提取另存，对内容大体相同的留言保存，去除极个别存在关键词但与其他大多数内容不一的留言内容，计算所有另存出的类别的热点指数，热点指数依照总相似留言数、点赞数和反对数进行计算，对各值设置权重，总相似留言数为 0.7，点赞数为 0.2，反对数为 0.1，据此通过 Excel 表格进行计算，得到排名前五的热点问题，并按照要求生成热点问题表与热点问题留言明细表；热点问题主要为 A4 区发生的特大诈骗按无进展问题；五矿万境 K9 县房屋质量出现问题；小区与高铁过近造成的噪音问题；违规搅拌场扰民问题；高压电施工环评造假问题。可以看出，热点问题主要是影响范围广，影响程度大，治理不力等原因造成的问题，比如：使该地区居民产生了巨额财产损失的大型诈骗案一直无进展，造成人民不满；比如：该地区出

现房屋质量问题，严重影响居民生活质量及安全；此类情况的问题如不尽早整治，会引起居民很大的不满。

3.3 问题 3 结果分析

提取其中的留言详情及答复意见进行计算研究，在面对相关部门对群众问政留言记录进行答复意见的同时，既要解决群众的留言问题，也要对相关问题答复使得群众满意。主要分析对照留言详情与答复意见在关键词之间的相关性计算，以留言详情与答复意见之间相同的关键词数量除以留言详情总关键词数，计算出相关性数据。通过对相关性数据的分析与整理，设置计算各留言详情与答复意见之间的完整性数据及可解释性数据，建立层次分析评价模型，对所有文本的得分情况进行计算，最后得到评价方案。

4. 结论

对群众问政留言以及相关部门对部分留言的答复意见进行分析研究，了解各类社情民意，对群众的生活满意度的提高有重大意义。同时也是文本分析的一个难题，给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。本文采用 knn 算法（k 近邻算法）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理；采用 get_words_list 算法获取关键词，计算相关度，发现热点问题，通过对这些热点问题的提取，有助于相关部门进行有针对性地处理，提升服务效率。

由分析结果可以看出，群众留言中大概有“车贷”，“房子质量”，“问题”，“高铁”，“搅拌场”，“高压电线”，“诈骗”等关键字，通过分析处理，热点问题为 A4 区发生的特大诈骗按无进展问题；五矿万境 K9 县房屋质量出现问题；小区与高铁过近造成的噪音问题；违规搅拌场扰民问题；高压电施工环评造假问题。

由此可见， 群众所反映的问题主要是影响范围广，影响程度大，治理不力的因素造成的，因此，政府相关部门的人们应该要加大管理力度，增加有关方面的措施，让群众的生活质量显著提高。

5. 参考文献

[1]王涛.“新型智慧政务”将迎来跨越式进步[J].中外管理,2020(Z1):30-32.

[2]吴宏胜.基于可信计算和 UEBA 的智慧政务系统[J].信息网络安全,2020,20(01):89-93.

[3]方刚.微信让城市生活更美好—智慧政务建设案例分析[J].传媒,2015,(5):27-28. DOI:10.3969/j.issn.1009-9263.2015.05.009.

[4]车洪莹.大数据时代智慧政务建设与发展路径研究[J].财会学习，2017（15）.

[5]李思思.基于政府治理能力现代化的智慧政务建设探析[J].中国管理信息化,2019,22(24):142-143.

[6]吴明,于欢.“互联网+政务”打造智慧政府的新路径研究[J].吉林工程技术师范学院学报,2019,35(08):35-37.