基于自然语言处理技术的文本挖掘应用

摘　要

随着社会的不断发展，网络问政已成为调查民意、民意交流的重要渠道. 各类民意文本数据量不断攀升，建立基于自然语言处理技术的数据处理方法成为新趋势，对政府管理和施政水平有极大的推动作用.

针对问题一，对附件 2 中的数据进行去噪处理，将留言主题与留言详情按一级标签进行整理后拼合，利用结巴分词算法对结果分割，找出关键词及其频率，筛选关键词，进行权重处理，得到各个一级标签下的核心关键词表，将核心关键词与留言转化成空间向量，建立 VSM 向量空间分类模型，使用 KMP 模式匹配算法求解每一条留言所归属的一级标签. 通过 F-Score 方法对模型分类结果进行评价，得出查全率均为 100%，不同一级标签的查准率不同，其中教育文体的查准率高达 91.88%，最终的查准率指标数为 0.7738.

针对问题二，基于问题一的模型对附件 3 中的数据进行分类，并对数据去噪处理，对不同留言进行余弦文本相似度匹配，将匹配度高的两个留言通过并查集算法归类，用并查集算法查询，得到热点问题的留言集表. 以留言人次、点赞反对数、留言种类、留言时间以及区域影响为标准层，建立留言热度的层次分析结构模型，计算各因素的权重指标，进行一致性检验. 定义热度评价指标，求解得出热度指标较高的前五个热点问题依次是： A 市 58 车贷诈骗、A 市丽发新城社区搅拌站扰民、A 市伊景园滨河苑捆绑销售车位、A 市金星北片高压线规划错误和 A 市五矿万境房屋出现严重质量问题.

针对问题三，分析影响留言答复评价的指标：相关性、及时性、有效性、完整性及可解释性，制定一套关于留言答复质量评价方案的等级评分系统，使用不同的等级 A，B，C，D 表示不同的满意度，最终对相关部门的留言答复进行评价.

关键词：VSM 向量空间模型；KMP 算法；并查集算法；层次分析结构模型

Abstract

With the continuous development of the society, Internet politics has become an important channel for the survey of public opinions and the exchange of public opinions. The data volume of various public opinion texts is continuously increasing. The establishment of data processing methods based on natural language processing technology has become a new trend, which has greatly promoted the level of government management and administration.

In view of the first problem, the data in attachment 2 is de-noised, the message subject and message details are sorted and combined according to the first-level labels, the result is segmented by using stuttering word segmentation algorithm, keywords and their frequencies are found, keywords are screened, the weight processing is carried out, core keyword tables under each first-level label are obtained, core keywords and messages are converted into space vectors, and VSM vector space classification model is established. The KMP pattern matching algorithm is used to solve the first-level labels to which each message belongs. The F-Score method is used to evaluate the classification results of the model. The recall rate is 100%, and the precision rate of different first-level labels is different. The precision rate of educational style is as high as 91.88%, and the final precision index number is 0.7738.

In solve the second problem, the model based on question 1 classifies the data in attachment 3, the data denoising processing, matches cosine text similarity for different messages, classifies the two messages with high matching degree through union search algorithm, and uses union search algorithm to query to obtain a message set table of hot issues. Based on the standard layer of message number, number of comments, message type, message time and regional influence, a hierarchical analysis structure model of message heat is established. The weight index of each factor is calculated and the consistency test is carried out. The heat evaluation index is defined, and the top five hot issues with high heat index are: 58 car loan fraud in A city, disturbance of mixing station in Lifa New Town Community in A city, bundling of parking spaces in Binhe Garden of Yijing Garden in A city, planning error of high-voltage line in North of Venus in A city, and serious quality problems of houses in five mines in A city.

Aiming at the third problem, this paper analyzes the indicators that affect the evaluation of message response: relevance, timeliness, effectiveness,

completeness and interpretability, and formulates a rating system for the quality evaluation scheme of message response. Different grades A, B, C and D are used to express different satisfaction. Finally, the message responses of relevant departments are evaluated.

Key words: VSM vector space model；KMP algorithm； Union search algorithm；Hierarchical analysis structure model

目 录

一、挖掘目标 ---------- 1
1.1 挖掘背景----------1
1.2 挖掘目标----------1
二、问题分析 ---------- 2
2.1 问题一的分析----------2
2.2 问题二的分析----------2
2.3 问题三的分析----------2
三、模型假设与符号说明 ---------- 2
3.1 模型假设----------2
3.2 符号说明----------3
四、数据预处理 ---------- 3
4.1 重复数据的查找与删除----------3
4.2 数据的清洗处理----------4
4.2.1 非文本数据的清洗 ----------4
4.2.2 长串数字或字母数据的清洗 ----------4
4.2.3 无意义文本、标点符号的数据清洗 ----------5
4.3 问题二的数据预处理----------6
4.3.1 时间跨度大且数据量少的数据的剔除工作 ----------6
4.3.2 留言详情以及留言用户两者完全相同数据的删除与标记 ----------6
五、模型的建立与求解 ---------- 7
5.1 一级标签的分类模型 ----------7
5.1.1 结巴分词算法（关键词提取算法）----------7
5.1.2 KMP 算法 ----------8
5.1.3 向量空间模型的基本概念 ----------9
5.1.4 改进的 VSM 向量空间模型 ----------10
5.1.5 模型的求解结果 ----------13
5.2 热点问题的挖掘----------15
5.2.1 留言并查集算法 ----------15
5.2.2 余弦文本相似度算法 ----------16
5.2.3 基于灰色关联度的多目标决策模型----------17
5.2.4 热度评价指标 ----------22
5.2.5 模型的求解过程 ----------22
5.2.6 模型的求解结果 ----------23
5.3 针对答复意见质量的评价方案----------25
5.3.1 答复意见质量的评价方案的指标 ----------25
5.3.2 答复意见质量的评价方案的标准 ----------25
5.3.3 答复意见质量的评价方案的应用 ----------26
5.3.4 答复意见质量的评价方案的意义 ----------27
六、总结与展望 ---------- 27
6.1 全文总结 ----------27
6.2 未来展望----------28
参考文献 ---------- 28

一、挖掘目标

1.1 挖掘背景

随着大数据时代的到来，对数据挖掘的要求也在不断提升，在繁多的数据中找到有用的信息，合理适当的分析出一定的规律，并对未来的发展趋势进行预测都是数据挖掘的重要作用. 网络不断发展，网络平台的数据信息种类繁多，其中文本信息数据更是呈爆发式增长，在处理文本信息数据的过程中，如何在繁多的文字中提取关键的文本信息，并对关键的信息进行分析成为分析数据的关键.

网络问政是社会不断发展进步的产物，网络平台是当今反映社情民意、加强交流的重要平台，是中国公民行使各项权力的重要渠道. 当今社会关于网络问政平台的使用日益广泛，智慧政务正在成为互联网时代政府治理发展的新形态，微信、微博、市长信箱、阳光热线等网络问政平台已经渐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意的相关的文本数据量在不断攀升，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. 因此在大数据、云计算、人工智能等科学技术如此高速发展的背景下，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.

建立完善好基于自然语言处理技术的智慧政务系统，通过分析网络上的数据可以创造出更大的公共价值，通过对海量数据的深度挖掘与多维度剖析，可以较为准确地发现公众新需求，掌握政府服务和管理的变化动态，并且可以合理的评估政府工作人员的工作态度，使政府的管理水平和治理能力得到有效的提升.

1.2 挖掘目标

基于自然语言处理技术的智慧政务系统需要做到的是对群众留言的分类，根据附件 1 熟悉群众留言可能存在的一级、二级以及三级分类，利用附件 2 中所给数据建立文本分类模型，将群众留言根据留言详情进行一级分类，并且使用 F-Score 方法对得到的分类结果进行评价；在众多的群众留言中精确寻找留言热点问题，定义合理的热度评价指标，根据附件 3 的数据对热点问题进行归类以及热度评价，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；针对附件 4 留言问题的答复意见从多种角度制定一套合理的评价方案.

二、问题分析

2.1 问题一的分析

针对问题一，对附件 2 的数据进行分析，将数据进行筛选，建立一个一级标签分类模型，需要找到各个一级分类标签的关键词，建立向量空间模型，将找到的关键词与需要匹配的留言进行匹配，得到留言详情的一级标签的分类. 并计算每一个一级分类的查准率和查全率，通过 F-Score 方法对模型的结果进行评价.

2.2 问题二的分析

针对问题二，对附件 3 中的数据进行处理，建立相应的算法得出留言相近的留言集，将这些留言集作为热点问题，并找到影响留言集热度的因素，通过建立层次分析模型，对一些因素做加权处理，得到影响热度问题的热度评价指标，随后进行定量计算，得到留言热度的结果，根据留言的热度评价指标对留言进行排序，找到热度排名前 5 的留言问题.

2.3 问题三的分析

针对问题三，研究并实现一套对留言答复意见质量的评价方案，就是对相关部门的留言答复做一个相对性的评比方案，用方案去判断有关部门是否切实地服务于人民，为人民答疑解惑. 将所有的影响留言答复的因素进行定量打分，得出留言答复的质量的评价方案.

三、模型假设与符号说明

3.1 模型假设

1．假设居民所给的每一个留言均只属于一个级分类，即不存在一个留言属于多个一级分类的情况；

2．假设所给的数据真实可靠；

3．假设附件 3 中存在的点赞数与反对数均为用户经过认真思考后所做的动作，不存在误点、错点等现象.

3.2 符号说明

符号	说明
Q_q	代表关键词向量
D_j	留言详情向量
R_i	一级标签的查全率
P_i	一级标签的查准率
P	热度评价指标
$sim(d_j, q)$	余弦相似度

四、数据预处理

本文的数据预处理包括问题一和问题二的数据预处理，问题一的数据预处理主要分为两个部分，分别是重复数据的查找与删除以及数据的清洗工作. 由于留言详情中存在部分数据重复，因此先进行重复数据的查找与删除工作. 随后对数据进行三个方向的清洗处理，分别是非文本数据的清洗、长串数字或字母数据的清洗以及无意义文本、标点符号的数据清洗. 问题二的数据预处理同样分为两个部分，分别是对时间跨度较大数据量好的数据进行剔除以及对留言详情以及留言用户两者完全相同的数据进行删除并标记.

4.1 重复数据的查找与删除

附件 2 中所给数据共 9210 条，使用 Excel 对表中数据进行整理，将留言主题与留言详情列中的重复数据使用条件格式操作找到重复的数据，重复数据包括留言主题完全相同数据、留言内容完全相同数据以及留言主题与留言内容均相同的数据，重复数据如表 1 所示.

表 1 附件 2 重复数据表

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
303	U0007137	A1 区蔡锷…	2019/12/6…	A1 区 A2 区…	城乡建设
319	U0007137	A1 区 A2 区…	2019/12/5…	A1 区 A2 区…	城乡建设
476	U0003167	反映 C4 市…	2019/11/15…	我们是梅…	城乡建设
532	U0008488	A 市魏家…	2019/11/10…	尊敬的 A…	城乡建设

10903	U000784	A8 县扶…	2015/11/9…	我叫王良…	城乡建设
15792	U0002483	A7 县北…	2019/11/27…	屋面材料…	城乡建设

重复数据共 636 条，占总数据的 0.069%，占比小，对结果的影响较小，故可以剔除这些数据，余下的数据作为后续计算建模的依据.

4.2 数据的清洗处理

4.2.1 非文本数据的清洗

在留言详情中存在部分数据附带有 HTML 标签、URL 地址等非文本内容，所以需要清除这部分内容对分类没有什么帮助的内容. 例如附件 2 中存在留言详情数据如下：

我是望仙路 187 号景上鑫苑 D 栋的住户，今年开始，我小区东侧开建的“圆梦康乐城”项目，其中一栋楼距离我居住的 https://baidu.com/米，原来阳光灿烂的日子现在不单 24 小时见不到阳光，白天都得开灯照明！我的采光权被生生剥夺. 经网上查询该项目五证齐全（不知道真假）. 请问杨书记，楼间距 14.8 米合法吗？

以上数据中存在“https://baidu.com/”属于非文本数据，可以将该数据进行清洗，得到新的数据如下：

我是望仙路 187 号景上鑫苑 D 栋的住户，今年开始，我小区东侧开建的“圆梦康乐城”项目，其中一栋楼距离我居住的米，原来阳光灿烂的日子现在不单 24 小时见不到阳光，白天都得开灯照明！我的采光权被生生剥夺. 经网上查询该项目五证齐全（不知道真假）. 请问杨书记，楼间距 14.8 米合法吗？

以上类似的非文本数据均需进行数据清洗得到更优的留言详情数据进行后续拆分以及匹配工作.

4.2.2 长串数字或字母数据的清洗

文本中长串的数字通常代表手机号、车牌号、用户名 ID 等文本内容，在非特定的文本分类情境下是可以去除的. 在附件 2 中留言详情的数据中存在多处长串数字，例如下面的留言详情数据：

局长大人

我曾多次打你们客运热线电话 0000-00000000 反映 A1 区南路新姚北路处因为公交车 370 路，115 路，152 路三趟公交车设置不科学（都在这个路

口左转进新姚北路）造成交通拥堵，木莲西路多个小区居民出行不便．但是问题一直得不到解决．希望能不能分一条线路走 A1 区南路木莲西路左转行驶，既解决拥堵问题，也方便我们几个小区居民出行．谢谢！

以上数据中存在“0000-00000000”这样的长串数字，可以将该数据进行清洗，得到新的数据如下：

局长大人

我曾多次打你们客运热线电话 0000-00000000 反映 A1 区南路新姚北路处因为公交车 370 路，115 路，152 路三趟公交车设置不科学（都在这个路口左转进新姚北路）造成交通拥堵，木莲西路多个小区居民出行不便．但是问题一直得不到解决．希望能不能分一条线路走 A1 区南路木莲西路左转行驶，既解决拥堵问题，也方便我们几个小区居民出行．谢谢！

以上类似的长串数字或字母数据都需要进行清洗得到更优的留言详情数据进行后续拆分以及匹配工作．

4.2.3 无意义文本、标点符号的数据清洗

其中无意义文本包含广告内容、版权信息以及个性签名的部分，这些无意义文本对于留言详情数据的拆分与匹配工作毫无意义，标点符号是一种停用词，包括‘，’，‘．’，‘《》’等等常用的标点符号都需要进行数据清洗．将留言详情中的数据进行无意义文本、标点符号的清洗，例如留言详情中存在文本如下：

A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内．每天尤其上下班期间这条路上人流车流极多，安全隐患非常大．强烈请求文明城市 A 市，尽快整改这个极不文明的路段．

以上数据存在‘，’‘．’这样的标点符号，对上述留言详情数据进行数据清洗后得到新的留言详情文本如下：

A3 区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工围墙内每天尤其上下班期间这条路上人流车流极多安全隐患非常大强烈请求文明城市 A 市尽快整改这个极不文明的路段

以上类似的无意义文本、标点符号的数据都需要进行清洗，得到更优的

留言详情数据进行后续拆分以及匹配工作.

4.3 问题二的数据预处理

4.3.1 时间跨度大且数据量少的数据的剔除工作

观察附件 3 的数据，得到群众留言信息中留言时间包含了 2017、2018、2019、2020 四年之内的数据，其中包括 2017 年的数据 1 条，2018 年的数据 4 条，2020 年的数据 79 条，2019 年的数据有 4242 条，全部留言信息共 4326 条，考虑 2017、2018 年的数据总和少，占总数据比值 0.11%，剔除该部分数据对热点问题的查找影响十分小，因此剔除此类数据，仅保留 2019 年的数据做后续计算，剔除的部分数据见表 2.

表 2 剔除的部分留言信息数据表

留言用户	留言主题	留言时间	留言详情	反对数	点赞数
A0182491	A 市经济…	2017-06-08 17:31:20	书记您好..	9	0
A3352352	A 市经济…	2018-05-17 08:32:04	A 市经济…	3	0
A23525	请求 A 市…	2018-10-27 15:13:26	领导好！…	3	0
A0012413	在 A 市人…	2018-11-15 16:07:12	我叫朱琦…	0	0

4.3.2 留言详情以及留言用户两者完全相同数据的删除与标记

附件 3 中存在留言详情数据以及留言用户数据两者完全相同的数据，找到这样的数据并列出表格，这样的数据一共 56 个，见表 3.

表 3 留言详情及留言用户完全相同部分数据表

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
192652	A000100792	A1 区辉…	2019/1/13 9:41:05	A1 区远…	0	0
279702	A000100792	A1 区远…	2019/1/7 21:35:39	A1 区远…	0	1
214746	A00013457	关于“2…	2019/1/5 15:50:14	2018 年…	0	0
280588	A00013457	A 市“2…	2019/1/7 11:09:15	2018 年…	0	0
215247	A00014071	A 市泉塘…	2019/12/20 8:59:58	尊敬的领…	0	1
230571	A00014071	A 市泉塘…	2019/11/22 9:31:44	尊敬的领…	0	2
227100	A00018469	A 市能不…	2019/6/12 8:07:49	尊敬的领…	0	0
282746	A00018469	A 市能不…	2019/06/12 08:23:01	尊敬的领…	0	1

将上诉表格中数据的重复值删掉一个，保留一项数据作为代表放回原数据表，并将保留的数据做上标记，考虑出现相同用户名以及相同留言详情的情况是这一留言问题并没有得到合理的解决，或者出现该用户在第一条

留言发出后，误以为自己发送失败而重新发送留言问题的情况，因此此类留言问题在后续模型的建立与求解中赋予特殊的权重来进行计算.

五、模型的建立与求解

5.1 一级标签的分类模型

处理网络问政平台的群众留言时，要对群众留言按照附加 1 中的划分体系进行分类，建立一级标签的分类模型需要将附件 2 中的数据进行处理，根据各个一级分类下的留言详情利用结巴分词算法对一级分类下的所有留言详情进行关键词的提取，并叠加得到关键词的出现频率汇总成表；随后利用 KMP 字符串匹配算法将附件 2 中所有的留言详情进行一级标签的分类.

5.1.1 结巴分词算法（关键词提取算法）

结巴中文分词：就是将连续的字序列按照一定的规范重新组合成词序列的过程. 结巴中文分词支持三种分词模式包括精确模式、全模式和搜索引擎模式. 这三个模式其中精确模式可以将句子最精确的切开，适合文本分析；全模式可以把句子中所有的可以成词的词语都扫描出来并且运行速度很快；搜索引擎模式是指在精确模式的基础上，对场次再次切分，提高召回率，综合分析，问题一中适合使用全模式的结巴分词算法对留言详情进行分词处理.

全模式的分词算法可以将句子中所有的可以成词的词语都扫描出来，例如将句子“我来到北京清华大学”划分为“我/来到/北京/清华/清华大学/华大/大学”，可以看到所有可能性均被列出，提高了与一级标签相关性高的关键词被提取出的概率. 将分词后统计的一级标签的关键词列出表格. 得到的各个一级标签的关键词表格，其中以一级标签“城乡建设”为例，得到关键词表见表 4．所有一级标签关键词的表格详见附件（一级标签关键词汇总表.xlsx）.

表 4 城乡建设关键词部分表

关键词	词频
业主	2051
小区	1893
政府	1265
领导	1048

开发商	996
建设	970
房屋	899
规划	891
物业	862
A 市	773

5.1.2 KMP 算法

（1）KMP 算法

KMP 算法是一种改进的字符串匹配算法，因为在普通的匹配算法中，时间复杂度为$o(n*m)$，但是在 KMP 算法中时间复杂度为$o(n+m)$，有效的减少了匹配过程中的时间复杂度，因此在数据量很大时，KMP 算法就会显得十分快速，留言详情的数据量很大，因此在关键词匹配时选用 KMP 算法.

（2）KMP 算法过程

Step1：寻求文本中的每个词前缀和后缀最长公共元素；

文本$P=\{p_0,p_1,\dots,p_i\}$，寻找文本 P 中词长度最长且相等的前缀和后缀．即在文本 P 中，如果存在$p_0,p_1,\dots,p_j=p_{i-j},p_{i-j+1},\dots,p_i$，那么在包含$p_i$的文本中有最大长度为 j+1 的相同前缀后缀.

例如，对于文本“我喜欢整理整理我的房间．”，那么该文本中各个词的前缀后缀的公共元素的最大长度 L 见表 5.

表 5 文中词语的前缀后缀公共元素的最大长度表

词	我	喜	欢	整	理	整	理	我	的	房	间
L	0	0	0	0	0	1	2	1	0	0	0

在文本“我喜欢整理整理我的房间．”中，“整理整”有长度为 1 的相同前缀“整”，而“整理整理”有长度为 2 的相同前缀“整理”，因此找到最长公共元素“整理”.

Step2：求$next$文本

其中$next$文本考虑的是除当前词之外的最长相同的前缀和后缀，故通过上一步求得各个前缀的公共词的最大长度后，作出一些变化，使得上一步中求得的值整体右移一个单位，而第一个词对应的值赋值为-1，得到新的最大长度表见表 6.

表 6 公共元素的最大长度表

词	我	喜	欢	整	理	整	理	我	的	房	间
L	-1	0	0	0	0	0	1	2	1	0	0

Step3：根据$next$文本进行匹配

进行匹配时，$i = next[i]$,文本词向右移动的位数为：$i - next[i]$. 如此将文本词继续后移，直至文本之间完全进行匹配，得到最终结果.

将 KMP 算法的算法步骤画出流程图，KMP 算法的流程图见图 1.

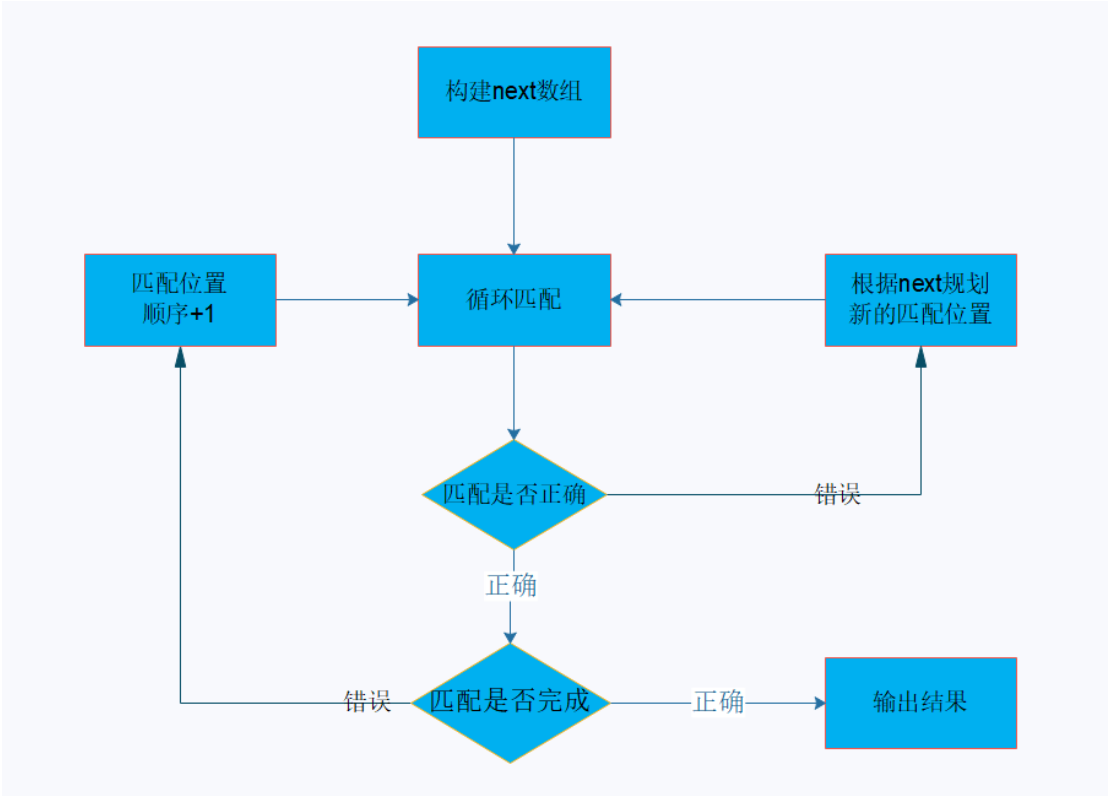

图 1　KMP 算法流程图

5.1.3 向量空间模型的基本概念

将文本内容进行处理，化作向量空间中的向量来进行相关计算，这就是向量空间模型的要素，使用向量空间的相似度来表达文本语义的相似度是向量空间模型的基本思想.

定义文档以及查询都使用向量表示，文档向量表示需要进行匹配的向量，查询向量表示经过关键词提取后得到的关键词组向量，使用n维向量d_j表示文档向量，因此

$$d_j = d(T_{1,j}, T_{2,j}, \ldots, T_{n,j}),$$

其中$T_{i,j}(i = 1,2,\cdots,n)$为文档中所拆分出的关键词，$n$为文档所拆出的关键

词的个数；

使用n维向量q表示查询向量，因此

$$q = q(T_{1,q}，T_{2,q}，\ \ldots, T_{n,q})，$$

其中$T_{i,q}(i = 1,2,\cdots,n)$为固定被匹配的关键词，$n$为在固定文档中找到的关键词的个数.

在文本处理过程中需要用余弦的方式来度量相似性的大小，其中常用的余弦公式为

$$cos\theta = \frac{d_j \cdot q}{\| d_j \| \cdot \| q \|}，$$

其中$d_j \cdot q$是文档向量以及查询向量的点乘，$\| d_j \|$是向量d_j的模，$\| q \|$是向量q的模.

得到模的计算公式为$\| d_j \| = \sqrt{\sum_{i=1}^{n} {T_{i,j}}^2}$，$\| q \| = \sqrt{\sum_{i=1}^{n} {T_{i,q}}^2}$；向量的点乘公式为$d_j \cdot q = \sum_{i=1}^{n} T_{i,j} T_{i,q}$.

将文档向量d_j与查询向量q之间的相似度定义为余弦相似度$sim(d_j, q)$，

$$sim(d_j, q) = \frac{d_j \cdot q}{\| d_j \| \cdot \| q \|} = \frac{\sum_{i=1}^{n} T_{i,j} T_{i,q}}{\sqrt{\sum_{i=1}^{n} {T_{i,j}}^2} \sqrt{\sum_{i=1}^{n} {T_{i,q}}^2}}.$$

5.1.4 改进的 VSM 向量空间模型

（1）一级标签关键词的提取

对附件 2 中的数据进行拆分，将每一个一级标签下的留言详情进行整合，得到一个个文档，使用 Python 软件利用结巴分词算法对文档进行关键词的提取，每一个一级标签得到 150 多的关键词，并且汇总一级标签关键词出的次数.

（2）一级标签关键词的筛选

每一个一级标签所提取关键词都有 150 多个，其中有许多关键词与其他一级标签重复或有歧义，无法精确定义到该一级标签，需要对一级标签关键词进行人为的筛选以及检查关键词的代表性，得到极具代表性且易区分于其他标签的关键词，每个一级标签选取 20 个这样的代表关键词，得到的一级标签的代表关键词见表 7.

表 7 一级标签代表关键词表

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
开发商	环保	出租车	学生	社保	旅游	医生
房屋	环评	的士	老师	工伤	电梯	患者
城管	排放	司机	教师	劳动	传销	手术
规划	保护	邮政	家长	退休	经营	病
工程	噪声	滴滴	考	养老	消费者	治疗
公积金	危害	快递	招生	岗位	导游	住院
领导	噪音	拥堵	补课	下岗	景区	计生
项目	生态	营运	成绩	加班	价格	诊
办理	健康	驾驶	校长	待遇	检验	疫苗
改造	开采	物流	机构	就业	监管	抢救
拆迁	破坏	客运	年级	职业	主管	药
物业	影响	拒载	民办	缴纳	产品	检查
业主	废	卧铺	录取	公务员	屠宰	生育
街道	周边	网约车	文化	考	宣传	护士
政府	土壤	旅客	公办	医保	景点	再婚
建设	污染	路面	幼儿园	医生	物业	过敏
建筑	污水	黑车	体育	保险	业主	输液
征收	排污	拥堵	教育	低保	开发商	结扎
施工	治理	左转	学习	参保	垄断	身体
招标	辐射	客货	学校	职工	销售	接种

（3）模型的准备

根据一级标签的代表关键词表，引入向量Q来表示代表关键词向量，得到公式

$$Q_q = Q(T_{1,q}，T_{2,q}，\ \ldots, T_{n,q})，$$

其中$T_{i,q}(i = 1,2,\cdots,n, q = 1,2,\cdots,m)$为一级标签的代表关键词，$n$表示一级标签的关键词的个数,这里$n = 20$；$m$表示一级标签的种类,这里$m = 7$.

将需要进行匹配的留言详情使用向量D_j来表示，留言详情向量D_j的公式表达为

$$D_j = D(T_{1,j}，T_{2,j}，\ \ldots, T_{n,j})$$

其中$T_{i,j}(i = 1,2,\cdots,n, j = 1,2,\cdots,m)$为留言详情拆分得到的关键词，$n$表示

留言详情所拆的关键词的个数,m表示需要匹配的留言详情的个数.

引入变量$K_{q,j}$，$K_{q,j}$值的大小决定留言详情向量与代表关键词向量之间的相关度，$K_{q,j}$值越大，相关度也大，因此每一个留言详情均需要计算七个$K_{q,j}(q=1,2,\cdots,7)$的值，在其中选择最大的值，作为该留言详情的一级标签分类.

（4）模型的建立

向量空间模型求解结果依靠的是余弦相似度，改进的向量空间模型求解的结果是依靠$K_{q,j}$这个指标变量．$K_{q,j}$的求解是根据Q_q与D_j之间的匹配，当留言详情D_j中所拆分出的关键词$T_{i,j}$在代表关键词Q_q出现，每出现一次，所对应的$K_{q,j}$的值加一，直至留言详情D_j中所拆分出的关键词$T_{i,j}$均匹配完成，得到一个$K_{q,j}$值，最终留言详情D_j与不同一级分类的代表关键词Q_q匹配的$K_{q,j}$值最大，决定留言详情归属的一级分类.

（5）模型的求解步骤

模型的求解思路如下：

Step 1：数据预处理，剔除重复的数据以及对数据进行清中，以得到更加完整和可靠的留言样本集；

Step 2：将得到的样本集分别用结巴分词算法分解成有效的关键词组合，并统计其中每个关键词出现的次数，得到一级标签关键词表；

Step 3：分析和整理关键词表，得到各个一级标签下最为有效的代表关键词组合，并根据关键词的出现频率以及覆盖范围，制定评估方案给每一个代表关键词一个恰当的权重；

Step 4：使用代表关键词去匹配文本，将每一个文本进行分类．将分类后的结果再重新产生关键词集合；

Step 5：多次迭代，不断完善关键词，直至关键词识别能力达到满足条件则输出结果，否则回到 Step 4 去循环迭代，找出更佳的结果，得到的所有留言详情的一级分类，并在得到结果后求解查全率以及查准率，计算所求结果的误差.

画出求解程序流程图，可以更加直观的看出求解的步骤与思路，求解的程序流程图见图 2.

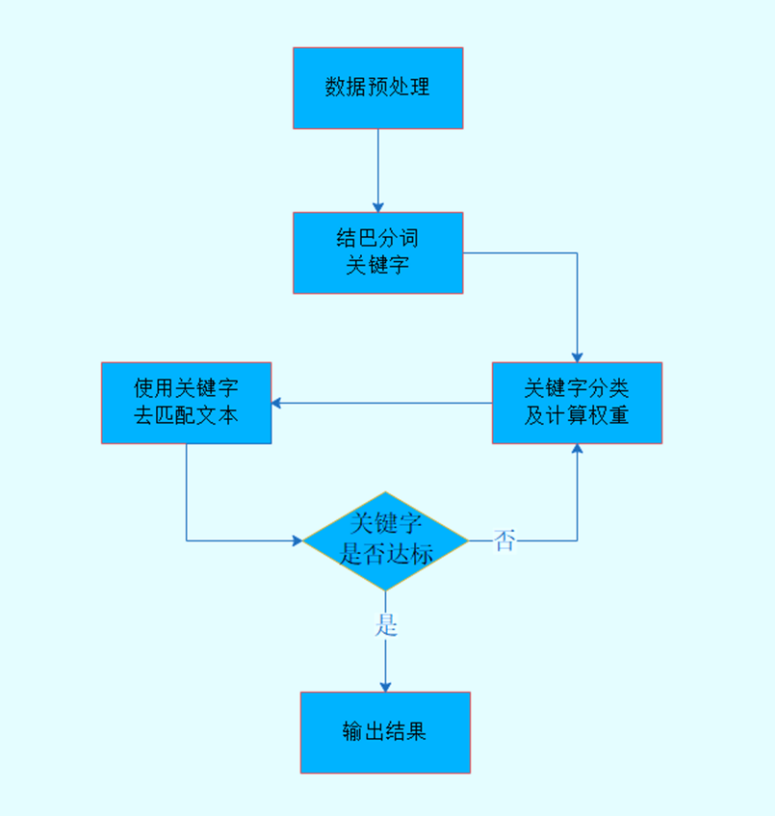

图 2 模型算法程序流程图

5.1.5 模型的求解结果

经过结巴算法的分词处理，提取出关键次，利用 KMP 算法实现向量空间模型的计算得到附件 2 中所有留言详情的分类，将得到的结果用 Excel 导出，其中挑选部分数据列出表格见表 8，全部数据的表格见附件（一级标签分类表.xls）.

表 8 附件 2 对留言详情一级分类结果表

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签	模型分类结果
24	A00074011	A 市西湖…	2020/1/6…	A3 区大…	城乡建设	城乡建设
37	U0008473	A 市在水…	2020/1/4…	位于书院…	城乡建设	城乡建设
4445	U0004752	A 市坪塘…	2018/10/11…	坪塘大道…	环境保护	环境保护
5017	U0006285	A8 县市…	2018/8/17…	位于花明…	环境保护	环境保护
12473	U0007386	A 市地面…	2014/2/28…	现在诸如…	交通运输	交通运输
12484	U0007386	京港澳高…	2014/2/14…	京港澳高…	交通运输	交通运输
165	U0005008	I 市重点…	2010/11/2…	文物是国…	教育文体	教育文体
403	U0004488	E5 县三…	2011/2/25…	E5 县三…	教育文体	教育文体
265	U000872	省、市社…	2010/12/2…	今日接到…	劳动和社会保障	劳动和社会保障

284	U000388	退休工人…	2011/1/9…	希望加退…	劳动和社会保障	劳动和社会保障
0	U0003773	A 市月湖…	2020/1/8…	听闻 A 市…	商贸旅游	商贸旅游
59	A00059726	A3 区山…	2020/1/1…	去年 12…	商贸旅游	商贸旅游
1247	U0008422	黑诊所呀…	2016/3/29…	我们是…	卫生计生	卫生计生
1720	U0005971	F 市 F4…	2013/4/18…	我是西地…	卫生计生	卫生计生

观察表格可以看出根据留言详情的匹配结果是较好的，这里使用 F-Score 对分类方法模型的结果进行评价，在留言一级标签的分类模型中用到的 F-Score 公式为

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}(i = 1,2,\cdots,7).$$

其中P_i为第i类的查准率，它是指检出的相关文献量与检出文献总量的比率，是衡量信息检索系统检出文献准确度的尺度，R_i为第i类的查全率，它是指检出的相关文献量与检索系统中相关文献总量的比率，是衡量信息检索系统检出相关文献能力的尺度.

统计全部数据的匹配结果发现共 9210 条留言详情，其中每一个一级标签的匹配正确个数与需要匹配的留言总条数都可以通过 Excel 计算得到，将留言的条数制作成 Excel 表格，得到*查准率*P_i的计算公式为

$$P_i = \frac{\text{留言正确个数}}{\text{一级标签的留言总个数}},$$

同时表格中所有的留言信息均得到了一个匹配结果，故所有的$R_i = 1(i = 1,2,\cdots,7)$．因此计算出$F_1$的值，计算结果见表 9.

表 9 模型一的结果表

	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
正确数	1634	796	439	1459	1460	648	738
总数	2010	937	612	1588	1969	1214	876
P_i	0.812935	0.84952	0.71732	0.918766	0.741493	0.533773	0.842466
R_i	1	1	1	1	1	1	1
F_1	0.773753232						

5.2 热点问题的挖掘

热点问题的挖掘需要定义一个热度评价指标，影响热点问题的因素有留言人次、点赞-反对数、留言种类、留言时间以及区域影响，因此热度评价指标的计算需要借助层次分析法，对以上因素进行权重比值，进而求得热度评价指标．在热度评价指标计算之前，求解分析不同因素的计算过程，利用留言并查集算法对留言人次进行计算，并且使用余弦相似度匹配算法对留言信息进行处理．

5.2.1 留言并查集算法

并查集的英文名称是：Disjoint set．即使用“不相交集合”将编号分别为$1,2,\cdots,n$的n个对象划分为不相交的留言集，并在每一个留言集中，选择其中的某一个留言作为代表留在留言集中．在留言并查集算法中有两个重要过程，其中一个是“并”，将两个留言集合并成同一个留言集，使得两个集合的最小集指向两个留言集中较小的一个．画出合并的示意图，见图 3．

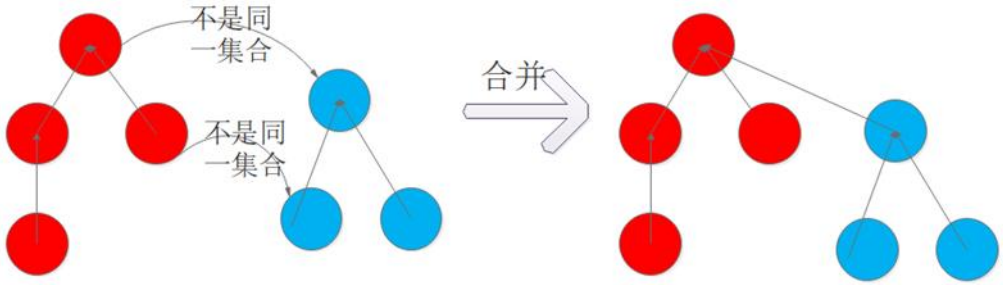

图 3 留言集合并示意图

在留言并查集算法中另一个过程叫做“查”，这一过程的目标是确定留言属于哪一个留言集，可以确定两个留言是否属于同一留言集，如果出现图 4 的情况，图中留言 5 和留言 6 的 set 值均最终指向留言 1 的值，因此留言 5 和留言 6 在同一留言集中．

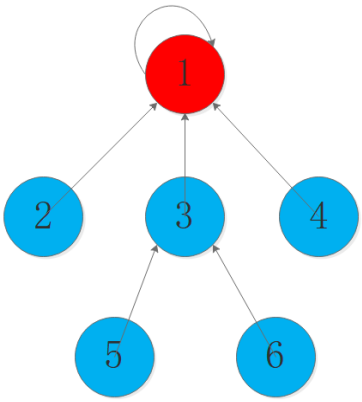

图 4 两个留言集属于同一留言集的示意图

将留言并查集中做进一步优化，即对路径进行压缩，当每一次查找时，路径较长时，则修改部分信息以便下次查找的时候速度更快，首先需要找到留言集的根结点，然后修改查找路径上的所有结点，将这些结点均指向根结点，最终对路径进行压缩，进一步提高查找时的速度，降低算法的时间复杂度，路径压缩的示意图见图 5.

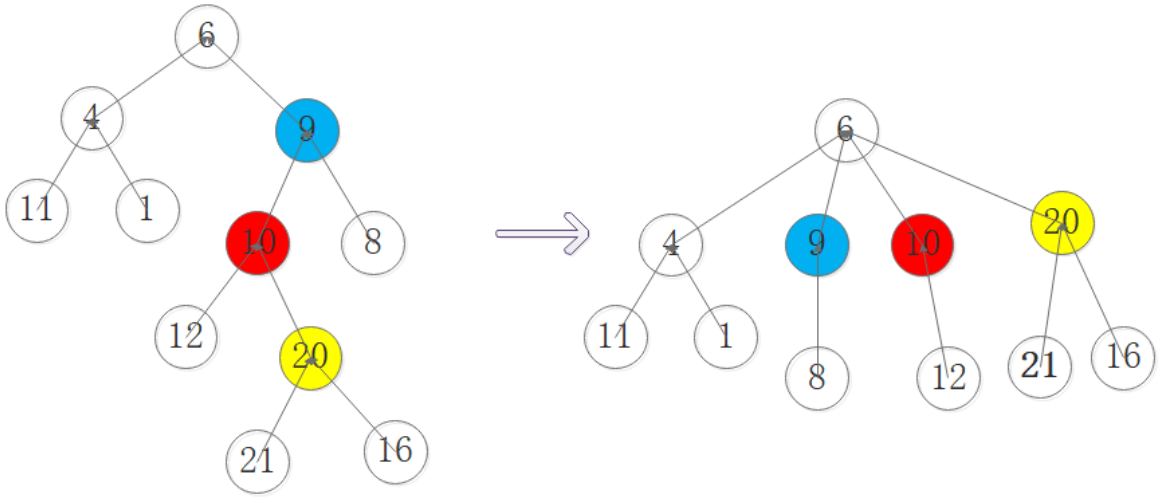

图 5 路径压缩的示意图

5.2.2 余弦文本相似度算法

文本相似度指的是通过一定的策略比较两个或多个词语、短文本或文档之间的相似度，得到一个具体量化的相似度数值，在热点问题挖掘中，需要对比居民留言中的留言详情数据，将文本相似度较高的留言详情归为一类，那么将这一类留言详情作为热点问题.

文本相似度算法的计算中，有表面文本相似度计算以及语义相似度计算，但是语义相似度计算需要对句子进行细致分析，找到依存关系，并且在

依存分析结果的基础上惊醒相似度阶段，过程较为复杂，本文解决热点问题的挖掘可仅通过表面文本相似度计算来完成，达到对居民的留言详情进行热点问题分类的目标. 表面文本相似度计算是针对文本信息的计算，作用在词、词组、字符或字符串组合上，通过比较文本信息与文本信息之间词、词组、字符或字符串的匹配程度作为相似度的衡量标准.

在热点问题的挖掘中，使用余弦相似度对居民的留言详情进行分类. 余弦相似度是通过测量两个向量的夹角的余弦值来度量文本之间的相似性. 余弦值的范围在[-1,1]，故在计算表面文本相似度时，两两文本进行相似度计算所得的余弦值为 1 或接近 1 时，代表文本相似度高，余弦值为-1 或接近-1 时，代表两文本之间的相似度微乎其微.

模型建立之初，需要统计每个字或词在每篇留言中出现的次数，以此作为词向量的值，现作如下假设：

有留言α以及留言β，留言α中出现的词为：$\alpha_{c1},\alpha_{c2},\dots,\alpha_{cn}$，这些词在留言$\alpha$中出现的个数分别为$\alpha_{s1},\alpha_{s2},\dots,\alpha_{sn}$；留言$\beta$中出现的词为：$\beta_{c1},\beta_{c2},\dots,\beta_{cn}$，这些词在留言$\beta$中出现的个数分别为$\beta_{s1},\beta_{s2},\dots,\beta_{sn}$.

α_{c1}和β_{c1}表示在留言α和留言β里的同一个词，α_{s1}和β_{s1}分别表示对应的个数，两向量间的余弦值可以通过欧几里得点积公式求出，这与问题一中余弦相似度的求解一致，公式如下：

$$\alpha \cdot \beta = \|\alpha\|\|\beta\| cos\theta .$$

整理得

$$cos\theta = \frac{\alpha \cdot \beta}{\|\alpha\|\|\beta\|} = \frac{\sum_{i=1}^{n} \alpha_{si} \times \beta_{si}}{\sqrt{\sum_{i=1}^{n}(\alpha_{si})^2}\sqrt{\sum_{i=1}^{n}(\beta_{si})^2}} .$$

5.2.3 基于灰色关联度的多目标决策模型

定义某一时间段内群众集中反映的某一问题称为热点问题，热点问题通常由多项因素共同决定，对热点问题的挖掘最先定义热点问题的热度指数H，热度指数的影响因素应该考虑相同留言的留言人次、留言的点赞数和反对数、留言时间、留言的种类以及地区影响力等多方面的影响因素，因此热点问题的挖掘需要多种标准来衡量确定.

影响热点问题挖掘的因素有留言人次、点赞-反对数、留言种类、留言时间以及区域影响，建立 AHP 层次分析模型将这些因素分解为目标、准则、

方案等层次，在此基础上对影响因素进行定性以及定量的分析.

（1）热点问题的影响因素

留言人次：相同留言指的是留言的内容大约一致，所表达的事件是相同的留言，留言人数也可记作相同留言的留言次数，通过使用余弦文本相似度算法计算得到相同留言的留言人数.

点赞-反对数：留言的点赞数和反对数在附件 3 中存在详细数据，而人为的把该因素定义为点赞数与反对数之差作为该影响因素的定量分析结果.

留言种类：留言种类是指将附加 3 中的留言详情使用问题一中建立的模型求解出留言详情所属的以及分类，将附加 3 中的留言信息根据留言详情的分类划分为不同的一级分类下的留言问题，这里是考虑不同种类的留言所有的热度指数可能存在较大差别，放在一起进行比较模型求解会产生大误差.

留言时间：留言时间是留言相同内容的留言信息所发布的时间的广度，从某个时间节点到另一个时间节点内类似的留言内容都存在，而在这两个时间节点以外从未提到此类留言，故这两个时间节点所表示的就是留言相同内容的留言信息所发布的时间的广度.

区域影响：存在城市、郊区等不同地区的群众留言是有不一样的影响力的，可能存在城市的留言信息比乡村的留言信息更能得到网友的认可，获得更高点赞数的可能性，因此将留言的地区影响力也作为影响热点问题的因素之一.

（2）建立层次分析结构模型

使用层次分析法对热点问题的求解进行定性定量的分析，首先需要建立目标层、标准层和方案层，其中目标层又叫最高层，表示的是建立层次分析法需要求解的最终问题是什么，这里目标层是求解热点问题；准则层又叫中间层，表示得到目标所需要考虑的因素、决策的准则，在这个热点问题的挖掘中，中间层为留言人次、点赞-反对数、留言种类、留言时间以及区域影响；对于方案层，也叫最底层，是在热点问题挖掘决策备选的方案，是最终被选择的留言类，此题中方案层为可供选择的热点问题. 根据描述可以画出层次分析法的结构图见图 6.

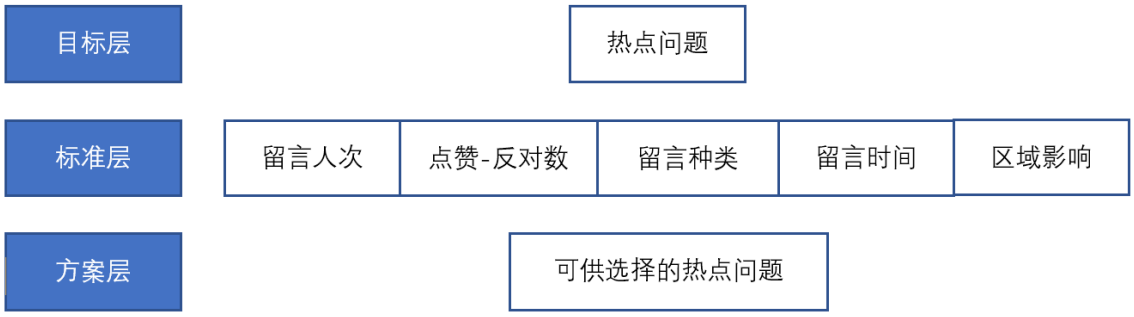

图 6 层次分析法结构图

（3）构造成对比较矩阵（判断矩阵）

比较某个元素与另一个元素相对于上一层某个元素的重要性时需要用到成对比矩阵，是一种使用数量化的相对权重来表示的方法．在确定各层次各元素之间的权重时，如果单纯的只是定性的对各元素之间的权重进行比较，计算难度大，令人难以接受，因此构建成对比较矩阵是一个提高比较准确度的良好方法，成对比较矩阵通过使用量两两元素进行比较，减少由于性质不同而带来的比较困难，达到提高准确度的标准．

这里，比较留言人次、点赞-反对数、留言种类、留言时间以及区域影响 5 个方面的因子对某个成分Z的影响大小，设因子为X，其中X的表示为

$$X = \{x_1, x_2, x_3, x_4, x_5\},$$

其中$x_i(i = 1,2,3,4,5)$分别为影响因素留言人次、点赞-反对数、留言种类、留言时间以及区域影响．

将各个因子进行两两比较，并按其重要性程度评定等级，根据比较结果建立成对比较矩阵，对每个因子x_i和x_j，以a_{ij}表示x_i对$x_j Z$的影响大小之比，比较结果用矩阵A表示． A的表达式为

$$A = (a_{ij})_{5\times5}.$$

且根据定义容易得到两个因子之间的影响大小是倒数的关系，故得到

$$a_{ij} = \frac{1}{a_{ji}}.$$

定义对成对比较矩阵元素a_{ij}的标度方法见表 10：

表 10 成对比较矩阵元素a_{ij}的标度方法表

因素i比因素j	量化值
同等重要	1
稍微重要	3

较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2,4,6,8

因此将标准层 A 的各个影响因素对目标层 Z 的影响两两进行比较，得到的比较结果见表 11.

表 11 标准层各因素对目标层影响相互比较的结果表

Z	A_1	A_2	A_3	A_4	A_5
A_1	1	2	3	5	7
A_2	1/2	1	3	3	5
A_3	1/3	1/3	1	3	3
A_4	1/5	1/3	1/3	1	2
A_5	1/7	1/5	1/3	1/2	1

其中A_1, A_2, A_3, A_4, A_5分别表示为留言人次、点赞-反对数、留言种类、留言时间以及区域影响．根据上表所示，可得到成对比较矩阵A为

$$A = \begin{bmatrix} 1 & 2 & 3 & 5 & 7 \\ 1/2 & 1 & 3 & 3 & 3 \\ 1/3 & 1/3 & 1 & 3 & 3 \\ 1/5 & 1/3 & 1/3 & 1 & 2 \\ 1/7 & 1/5 & 1/3 & 1/2 & 1 \end{bmatrix}.$$

（4）层次单排序及其一致性检验

层次单排序是确定下层各因素对上层某因素影响程度的过程．使用权重表示影响程度，对于成对比较矩阵的最大特征根λ_{max}的特征向量，经过归一化后记为W．得到向量W的元素为同一层次元素对于上一层次元素中某因素的相对重要性的权值的排序．若成对比较矩阵是一致矩阵，则取对应的最大特征根λ_{max}的归一化特征向量$W = \{w_1, w_2, w_3, w_4, w_5\}$，且$\sum_{i=1}^{5} w_i = 1$.

根据上述成对比较矩阵A，计算出最大特征根

$$\lambda_{max} = 5.0590,$$

计算得到其对应的特征向量为$\{0.8026, 0.4919, 0.2871, 0.1501, 0.0946\}$，对其

进行归一化处理后得到

$$W=\{w_1,w_2,w_3,w_4,w_5\}=\{0.4395,0.2693,0.1572,0.0822,0.0518\}.$$

判断得到的结果能否确认层次单排序，需要进行一致性检验，故首先计算一致性指标CI，由于特征根λ连续依赖于元素a_{ij}，则得到结论：特征根λ比标准层元素越大，成对比矩阵A的不一致性越严重，故用一致性指标CI来计算，成对比较矩阵的一致性与一致性指标呈负相关，则当CI越小，一致性越大，CI越大，一致性越小．一致性指标CI的计算公式如下：

$$CI=\frac{\lambda-n}{n-1},\ n=5,$$

当CI=0 时，有完全的一致性；CI接近于 0 时，有满意的一致性；CI越大时，结果的不一致性越严重，通过计算得到

$$CI=\frac{5.0590-5}{5-1}=0.01475,$$

已知得到的CI值接近于 0，因此有满意的一致性，为了降低CI的影响，引入随机一致性指标RI，RI的计算公式如下：

$$RI=\frac{CI_1+CI_2+\cdots+CI_n}{n},\ n=5,$$

已知随机一致性指标RI的数值见表 12．

表 12 随机一致性指标 RI 的数值表

矩阵阶数	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

在该模型中已知矩阵的阶数$n=5$，故随机一致性指标$RI=1.12$．

考虑一致性偏离可能是由随机原因造成，因此在检验判断矩阵是否具有满意的一致性时，还需将一致性指标CI和随机一致性指标RI进行比较，得出检验系数CR，检验系数CR的计算式如下：

$$CR=\frac{CI}{RI}.$$

如果检验系数$CR<0.1$，则认为该成对比较矩阵通过一致性检验，否则就不具有满意一致性．在该模型中计算检验系数CR，

$$CR=\frac{0.01475}{1.12}=0.0132,$$

得到$CR=0.0132<0.1$，因此该成对比较矩阵通过一致性检验．因此建立

层次分析法的模型是合理的，得到的结果是较好的.

5.2.4 热度评价指标

通过层次分析法的计算，得到以留言人次、点赞反对数、留言种类、留言时间以及区域影响为指标计算居民留言的热度通过一致性检验，是合理的．因此可以使用层次分析发的模型求解热度评价指标.

定义热度评价指标为P，计算热度评价指标是利用层次分析法中对各因素所设置的权重，并且因为留言人次与点赞反对数作为决定热度的主要因素，因此定义每一个热点问题所留言人次占总留言数的比例为l_1，定义某条留言的点赞数与反对数的差比上所有留言的点赞数与反对数的差记作l_2；从层次分析法中取得影响留言热度的因素留言人次的权重为α_1，点赞反对数的权重为α_2．故定义每一个热点问题的热度评价指标的计算公式为

$$P = l_1\alpha_1 + l_2\alpha_2.$$

5.2.5 模型的求解过程

热点问题的挖掘需要考虑多个因素，因此将因素的影响力数值化是一个严谨的过程，模型的求解过程中应该对因素进行逐一分析，模型的求解过程如下：

Step 1：对附件 3 中的数据进行预处理，删除时间跨度大且数据量少的数据，并且查找到留言详情以及留言用户完全重复的数据进行标记和删除工作.

Step 2：将不同留言所描述的对象和地点进行余弦定理相似度匹配，并将相似度匹配度较高的留言通过并查集算法归类于一类，得到可能成为热点问题的留言问题.

Step 3：利用留言并查集分类算法将热点问题的留言进行分类，得到多种类别的热点问题.

Step 4：分析影响热点问题的五个因素：留言人次、点赞-反对数、留言种类、留言时间以及区域影响，对这些影响因素进行定量定性分析.

Step 5：根据影响热点问题的五个因素建立层次分析模型，分别赋予不同因素以不同的权重，根据层次分析法求解出热度评价指标，得到留言的影响力.

Step 6：将留言的影响力与留言内容连接起来，建立表格得到热度排名

前五的热点问题，求解出结果，得到热点问题留言明细表.

将上诉求解过程绘制成模型求解流程图，见图 7.

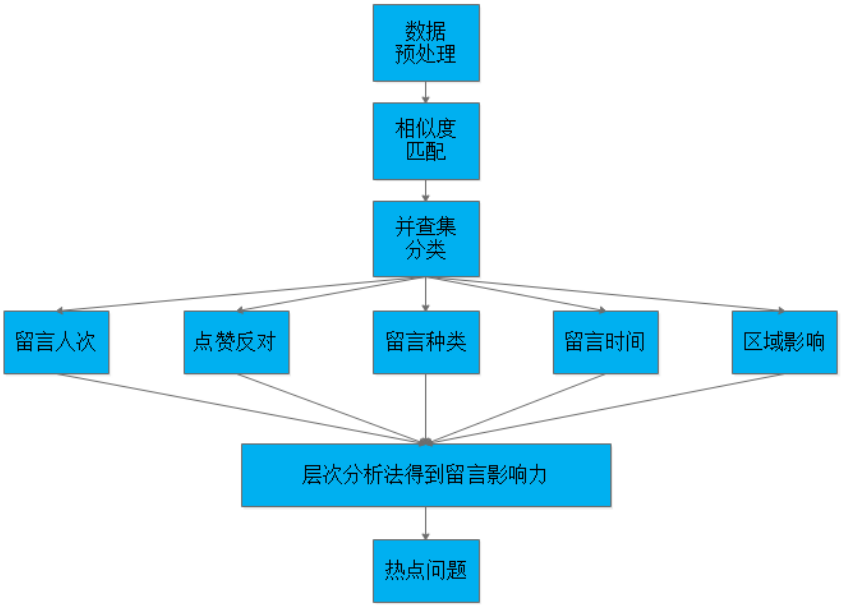

图 7 模型算法求解流程图

5.2.6 模型的求解结果

利用留言并查集算法对留言详情并查集的分类后，找到相同留言的留言信息，将所有热点问题的留言信息制作成 Excel 表格，表格中包含该热点问题属于的留言集以及留言集中包含的留言信息数即留言集数．热点问题的留言集部分表见表 13，完整的表见附件（热点问题的留言集表.xls）.

表 13 热点问题的留言集部分表

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	一级标签	属于的集	留言集数
281851	A000112932	咨询…	2019/3/18…	本人…	0	0	城乡建设	22	34
263429	A000108760	咨询…	2019/1/15…	母亲…	0	0	卫生计生	22	34
259594	A00011365	咨询…	2019/12/15…	我是...	0	0	城乡建设	22	34
220015	A00073641	咨询…	2019/8/9…	本人…	0	0	城乡建设	22	34
253923	A00019917	咨询…	2019/7/26…	领导…	0	0	城乡建设	22	34
265551	A00076292	咨询…	2019/5/23..	去年...	0	0	城乡建设	22	34

根据层次分析模型以及上述得到的热点问题留言集，综合计算得到各

热点问题的热度评价指标即热度指数.列出热度指数排名前 5 的热点问题，结果见表 14.

表 14 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	0.5479	2019/01/08 至 2019/07/08	A 市 A4 区 58 车贷	西地省 58 车贷案诈骗案
2	2	0.4726	2019/11/13 至 2020/02/25	A 市暮云街道丽发新城社区	A 市暮云街道丽发新城社区搅拌站污染大，噪音扰民
3	3	0.3955	2019/07/07 至 2019/09/01	A 市伊景园滨河苑	A 市伊景园滨河苑捆绑销售车位
4	4	0.2391	2019/03/26 至 2019/04/15	A 市金星北片 110kv 及以上高压线	A 市金星北片 110kv 及以上高压线规划错误
5	5	0.1961	2019/05/05 至 2019/11/11	A 市五矿万境 K9 县房屋	A 市五矿万境 K9 县房屋质量问题严重

将热点问题对应的所有留言信息按照所给格式列出表格，表格名为“热点问题留言明细表”，列出部分数据表格见表 15.

表 15 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	留言详情	点赞数	反对数
1	220711	A00031682	请书记…	2019/2/21…	尊敬的...	821	0
1	217032	A00056543	严惩 A…	2019/2/25…	胡市长...	790	0
1	194343	A000106161	承办 A…	2019/3/1…	胡书记...	733	0
1	268251	A000106090	西地省…	2019/2/2…	关于 A...	25	0
1	240554	A00029163	A 市 5…	2019/2/10…	A4 区..	6	0
1	226265	A000106448	恳请 A…	2019/5/28…	唐局长...	3	0
1	254532	A000106062	A 市 5…	2019/1/14…	背景：...	3	0
1	272413	A000106062	西地省…	2019/1/14…	西地省...	2	0

5.3 针对答复意见质量的评价方案

5.3.1 答复意见质量的评价方案的指标

制定一套针对政府部门对群众留言答复意见的评价方案，定义评价方案的指标分别是相关性、完整性、可解释性、及时性以及有效性.

（1）相关性：判断留言答复意见的相关性，是通过居民的留言与相关部门所给的答复意见进行对比，判断谈论的对象是否一致，事件是否为同一件事，描述事件的一致程度.

（2）完整性：判断相关部门对群众留言做出的答复意见是否完成，通过检验答复意见的逻辑性，对答复意见进行完整性检验.

（3）可解释性：判断留言答复意见中，是否存在相关法律条件，是否符合逻辑，如果留言答复意见有依据，则可说明该留言答复意见有较好的可解释性.

（4）及时性：留言答复意见的及时性表示从居民留言时间开始，到相关部门做出留言答复意见结束，这一段时间的长短决定留言答复意见的及时性.

（5）有效性：有效性即表示为相关部门的留言答复意见是否实用，使得居民在实际生活中能够解决问题.

5.3.2 答复意见质量的评价方案的标准

影响答复意见质量的五个指标分别为相关性、完整性、可解释性、及时性以及有效性，针对五个指标制定一套等级评分系统对答复意见做出评价，等级评分系统中制定 A，B，C，D 四个等级，不同的等级代表不同的满意度，如果该答复意见在某指标下，满意度最高，则可以给该答复意见在该指标下给以等级 A；满意度比较高时，给予等级 B；满意度一般时，给予等级 C，当等级获得 D 时，认为该答复意见在某指标下完全不符合，在不同情况中看作不同意思.

当在相关性的评分中，获得 D 等级意味着留言答复意见与群众留言内容毫无关系；在完整性的评级中，获得 D 等级意味着问题的答复不完整，毫无逻辑，在一定程度上体现相关部门工作的不负责；在有效性的评级中，获得 D 等级意味着群众收到的答复意见完全没有实用意义，难以实施；在

及时性的评级中，获得 D 等级意味着相关部门在居民留言之后的很长一段时间没有答复，表示的是工作人员在职工作效率低下，且对待群众所叙述的事情十分敷衍.

相反地，如果留言答复在各指标中的评级为 A，则说明留言答复十分及时，具有较高的有效性，良好的完整性以及高的相关性，表示相关部门的工作人员认真工作，踏实能干，关系群众，可以做到及时的给出有效的解决方案. 值得一提的是关于指标“可解释性”的评级，认为只有两个等级，A 为有依据，B 为无依据. 将答复意见质量评价方案的标准绘制成表格显示见表 16.

表 16 答复意见质量评价方案的标准表

相关性	问题的答复与所述问题完全一致	A
	问题的答复与所述问题比较一致	B
	问题的答复与所述问题一般一致	C
	问题的答复与所述问题不一致	D
及时性	对居民的留言答复相当及时	A
	对居民的留言答复比较及时	B
	对居民的留言答复一般及时	C
	对居民的留言答复不及时	D
有效性	给予居民的答复完全有效	A
	给予居民的答复比较有效	B
	给予居民的答复一般有效	C
	给予居民的答复完全没有效	D
完整性	问题的答复十分完整	A
	问题的回叙比较完整	B
	文图的答复一般完整	C
	问题的答复不完整	D
可解释性	对问题的答复有依据	A
	对问题的答复无依据	B

5.3.3 答复意见质量的评价方案的应用

根据上述制定出的评价方案的指标与标准，对附加 4 中的数据进行合理的分析，首先选取附件 4 中的部分数据，选取的部分数据表见表 17.

表 17 附件 4 的部分数据表

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
2549	A00045581	A2 区…	2019/4/25…	2019 年…	现将网…	2019/5/10…
2555	A00031618	请加快…	2019/4/24…	地处省…	市民同…	2019/5/9…
2557	A000110735	在 A 市…	2019/4/24…	尊敬的...	网友“…	2019/5/9…

对上述表格中的留言信息与答复信息做等级评分系统的评价，得到对应的评价结果见表 18.

表 18 对应的评价结果表

留言编号	留言用户	相关性	及时性	有效性	完整性	可解释性
2549	A00045581	B	B	A	A	A
2555	A00031618	A	B	B	A	A
2557	A000110735	A	B	A	A	A

5.3.4 答复意见质量的评价方案的意义

从相关部门对群众的留言进行答复的情况可以看出各对应部门的工作情况，尤其是工作态度可以很好地展现，根据评价方案的指标与标准的规划，获得等级为 D 的留言答复意见为表示某工作单位的员工工作毫不上心，无法为居民解决实际问题，对待群众的留言敷衍了事，这样的员工应该接受批评，不断去改正，端正态度，真正做到全心全意为人民服务．当然对于多项指标均可获得等级 A 的群众留言答复意见则说明相关部门针对居民提出的问题可以做到有实效的解决办法，在各方面有着较高的评价，因此应该给予嘉奖．鼓励更多的部门向这样的工作者学习．

六、总结与展望

6.1 全文总结

本文的主要利用自然语言处理和文本挖掘的方法建立群众留言分类模型和答复意见详细评价模型．首先，数据去噪处理，通过结巴分词得出关键词，然后经过对关键词的筛选及权重处理，建立 VSM 向量空间分类模型，使用 KMP 算法求解分类结果，最后用 F-Score 方法对模型分类结果进行评价．基于分类模型，对附件 3 留言进行分类，对不同留言进行余弦文本相似度匹配，用并查集算法归类，得到热点问题的留言集表．以留言人次、点赞反对数、留言种类、留言时间以及区域影响为标准层，建立留言热度的层次

分析结构模型，求解出热度指标较高的前五个热点问题. 分析影响留言答复评价的指标：相关性、及时性、有效性、完整性及可解释性，制定一套关于留言答复质量评价方案的等级评分系统对相关部门的留言答复进行评价.

6.2 未来展望

信息智能技术正在爆发式增长，很多智能产品、平台不断生产更新，智能化逐渐将触角伸向各个行业. 将互联网+、云计算和大数据等理念融入智能政务中，为政府部门解决部分相关人员对于居民的留言处理问题存在工作量大，效率低等问题. 且随着数据量的不断增加，人工处理数据已经不能满足人民的需要，据此进行设计，实现全程智能收集，智能分类，智能评判，并将每个环节用信息技术相互关联。该智能、自动、高效率的模式可引领“智慧政务”未来的发展方向.

参考文献

[1]王春柳,杨永辉,邓霏,赖辉源.文本相似度计算方法研究综述[J].情报科学,2019,37(03):158-168.

[2]曾小芹.基于 Python 的中文结巴分词技术实现[J].信息与电脑(理论版),2019,31(18):38-39+42.

[3]李静.字符串的模式匹配算法——基于 KMP 算法的讨论[J].青岛化工学院学报(自然科学版),2002(02):78-80.

[4]陶跃华,王锡钢,王云爱.信息检索向量空间模型中特征提取的研究[J].云南师范大学学报(自然科学版),2000(06):18-20.

[5]朱立军,苑玮琦.基于并查集和边缘检测模板的非理想虹膜定位[J].计算机应用研究,2018,35(06):1879-1882.

[6]许浩,周亚萍,赵亚慧.基于余弦文本相似度计算的英语作文评分算法的应用研究[J].教育教学论坛,2018(06):255-256.

[7]楚存坤,孙思琴,韩丰谈.基于层次分析法的高校图书馆学科服务评价模式[J].大学图书馆学报,2014,32(06):86-90.