智慧公务中的文本挖掘应用

摘要

随着互联网技术的迅速发展，政府与群众之间的沟通变得越来越便捷。明察暗访不再是政府了解民情的主要方式，人民群众可以通过各个网络问政平台进行留言。人民意愿得以表达的同时，给进行留言分类和热点整理的工作人员带来了极大的挑战。为了减轻政府工作人员的工作压力，现在提出智慧政务的需求，使用数据挖掘的相关方法，对大众提出的建议进行分类，以便进行后续处理。

本篇文章将市民反应的不和谐城市情况收集起来，使用断句处理以及数据挖掘的方法，对这些问题进行集中化处理。第一步，对市民提出的问题进行合理分类；第二步，根据市民反映的问题以及设计的算法，对众多文本数据进行挖掘，只要有某一个问题的相关程度系数超过一定的范围，那么我们就将这样的问题称为热点问题；最后，还要实现对留言问题的解决情况的评价。

针对问题一，使用 python 语言中的 jieba 库；对市民提出的相关问题进行合理的断词处理，这样一来就可以将整体的文本进行细致化处理，建立模型，对留言关键词内容进行训练，增加数据的匹配能力，更好更方便的比较市民提出问题的关键之处，能便利的对这些问题进行合理的划分，这样一来就有效的提升了公务处理的速度以及准确率。

针对问题二，对热点问题进行挖掘；众多市民对某一种问题的反映信息特别多，针对各位市民在进行信息反馈时，使用的语言顺序不一样，表达方式不一样等，使用 jieba 进行断句处理之后，对每组的断句关键词进行对比，只要其中一组的信息以其他众组的其相关程度系数超过一定的范围，就可以认定这样的问题为热点问题。

针对问题三、对留言的答复意见的评价；市民提出问题之后相关部门将进行处理，处理结束之后，需要对市民的反馈处理情况，以完成智慧政务的系统闭环。通过对答复意见的分析处理，对比答复意见中的相关关键词与市民提出问题的吻合情况，即可判断出相关部门的答复是否完整，那也就可以判断该答复的可解释性以及相关性。

关键词：数据挖掘；断句处理；相关问题；相关程度系数；

一、 问题重述

1.1 问题背景

传统的公务处理依赖人工分类，但是人工分类有着工作量巨大、分类效率低下、容错率高等不足之处；计算机的迅猛发展给传统的公务处理带来了希望，只要通过对样本的数据采样以及分析、训练、对比、处理，按照我们提前对相关公务问题的概述，就可以将众多的问题进行合理分类，有效地减少了人工处理时的种种问题与不便，极大的提高了工作效率。

1.2 问题提出

根据上文中背景，以及给出的几个附件内容，有以下几个问题需要解决：

（1）对群众的留言内容进行分类，进行合理的分类，能有效的提升公务处理时的效率与正确率。

（2）对热点问题进行挖掘，热点问题的提前分辨与处理，能够有效的减少公众留言的数量，大幅度提高效率。

（3）对相关问题的答复意见，从问题提出到问题解决需要形成一个闭环的回路，评价答复意见，能判断该答复是否完整可行，进而判断该问题是否得到了合理的处理。

二、问题分析

2.1 题目一的分析

题目要求将群众提出的问题进行合理分类，首先根据附件一以及附件二，可以发现相关的问题分类已经进行了合理的处理，只需要对现在已有的附件二中的数据进行读取，然后对其中每个用户的留言内容进行断句处理，取关键词。分析对比就能获取到每一个问题分类中的关键词，然后对这些关键词进行训练，筛检合格的关键词，并用他们对后续的数据进行对比分析，当即就可以得到现在这些数据的分类结果。

2.2 题目二的分析

题目要求我们挖掘热点问题，这一个问题与题目一的相关内容有部分相似。根据对题目一的分析研究以及抓取关键词，我们同样可以运用到样本三的文本研究以及抓取关键词，这样一来，我们有了大量的数据文本，接下来，只要对比每一类问题中每个用户留言的关键词，我们就能发现该关键词的重复情况，只要我

们对比各个留言之间的关键词相似情况，那么我们大体上就可以找到热点问题。由此题目二完成。

2.3 题目三的分析

题目要求从公务答复开始入手，分辨答复的准确程度。那么首先就要从答复的内容开始做起，第一步，读取答复的具体内容，对其进行断句处理并提取关键词，第二步提取群众反映问题的关键词，最后将两者的关键词进行对比。只要两者的关键词匹配率达到一定的程度，或者极其重要的关键词匹配率很高，那么，我们就可以断定该答复是对群众问题不错的答复。

三、模型假设

1.每个用户每天只能提交不多于三条信息。

2.用户提供的信息数据都是真实且合理的。

3 马尔科夫模型的基本假设：

（1）齐次马尔科夫性假设：即假设隐藏的马尔科夫链在任意时刻 t 的状态只依赖于其那一时刻的状态，与其他时刻的状态及观测无关，也与时刻 t 无关；

（2）观测独立性假设：即假设任意时刻的观测只依赖于该时刻的马尔科夫链的状态，与其他观测即状态无关

四、模型建立与求解

由于题目中给的数据都是使用 excel 表格呈现的，我们很难从中找到分析的途径，所以在进行实际的数据处理之前，我们需要有一个明确的流程图来简化整个步骤：其图示 4-1 所示。

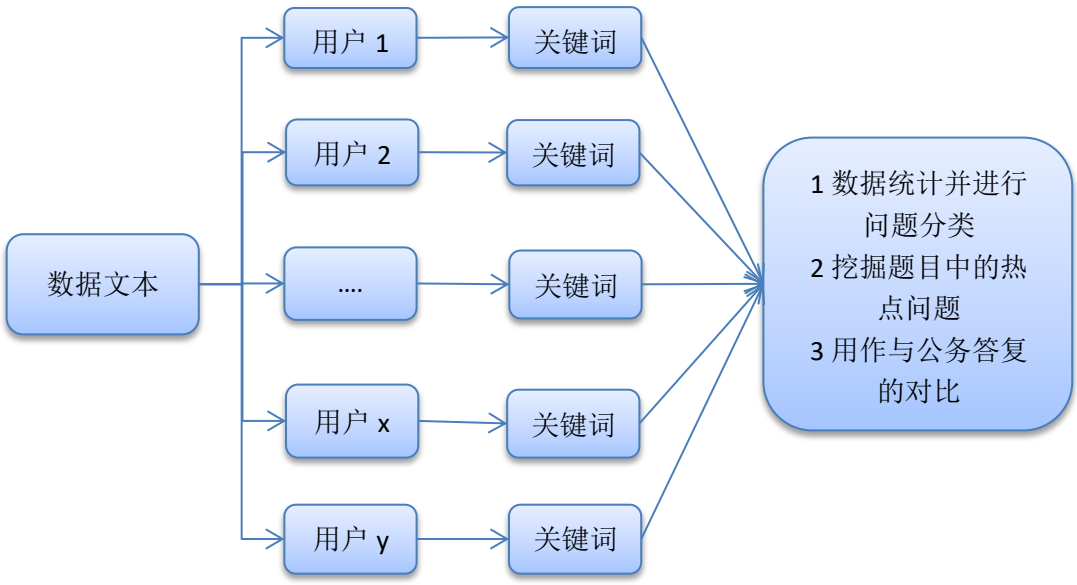

图 4-1　数据样本处理流程图

4.1 问题一

针对第一个问题，首先要分析样本数据。我们要根据比赛中提供的数据一与数据二，借助Anconda软件中的jieba库对一部分用户反映的信息进行断句处理，借助处理后的文件，首先分析这部分用户反映的情况中关键词都是在哪一些大问题分类下，然后我们对整体的样本进行处理，判断各个信息的关键词应该属于那个大问题，进行统计。然后将使用机器语言进行分类的文本与之前的文本信息进行对比，我们就可以得到该算法程序的准确性。

我们将每个用户反映的信息、我们提取的关键词、之前的信息分类以及我们使用机器语言进行的问题分类取几个简单的例子绘制到下面的表格中，如表4-1：

表 4-1 数据文本处理表

A00074011	反映的信息	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。
	之前的分类	城乡建设
	断句处理	A3 区 大道 西行 道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团 燕子 山 安置 房 项目 施工 围墙 上下班 期间 条 路上 人流 车流 安全隐患 请求 文明城市 市 整改 文明 路段
	机器分类	城乡建设
A00027642	反映的信息	C4市出租车过年乱收费，不打表，一口价，过河15元。才一两公里路!! 这股风气谁允许的?
	之前的分类	交通运输
	断句处理	C4 市 出租车 过年 乱收费 打表 一口价 过河 15 元 一两 公里 路 股 风气
	机器分类	交通运输
U0005379	反映的信息	尊敬的环保局领导: 我是C3县樟树村杉山下组组长楚龙石，我从您局拿到的红燕化工征我组林地作重金属废渣堆场的环评报告中有一项民众参与情况是民众坚持支持和条件支持率为100%。那为什么会有反对征山的反对呢？我能提供反对照片。

	之前的分类	环境保护
	断句处理	尊敬 环保局 领导 C3 县 樟树 村杉 山下 组组长 楚龙石 局 拿到 红燕 化工 征 组 林地 作 重金属 废渣 堆场 环评 报告 中有 一项 民众 参与 情况 民众 支持 条件 支持率 100% 反对 征山 反对 我能 提供 反对 照片
	机器分类	环境保护

通过机器语言对留言详情的分段，根据关键词描述自动归类这一操作来看。基本的关键词划分以及问题分类是可行的，最后只要将机器算法的训练比例控制在合理的范围内，机器在解决此类问题时的准确率就会得到保障。

4.2 问题二

根据题目要求，需要挖掘文本数据中的热点问题。我们需要对热点问题有一个定义，什么是热点问题，热点问题就是重要的关键词频繁出现，以及其相关关键词出现次数多的问题。根据这个我们来实现第二个题目的要求，对众多样本文本进行数据分析，对文本中的关键词进行对比，比如上一条的关键词与下一条的关键词进行比较，设置一个从 0 到 1 的相似度（0 代表两个留言之间并无相似，1 代表着两个留言是相似的），如果有一条留言的关键词，对其他所有子样本的相似度最高，那么就可以确定热点问题。

在题目一的分析过程中，我们已经获取了获得众多留言详情的关键词的方法，在这里我们同样可以将这一方法运用到题目二的分析之中，从这里面我们就可以得到重复率高的词，进而我们便可以知道那些问题是热点话题。我们列出 5 组数据的相关程度系数，结果如下表 4-2

表 4-2　相关程度系数

	第一组	第二组	第三组	第四组	第五组
第一组	1	0.0016852	0.000208016	0.00107792	0.00686955
第二组	0.0016852	1	0.00315634	0.000987825	0.00255567
第三组	0.000208016	0.00315634	1	0.00761007	0.00282677
第四组	0.00107792	0.000987825	0.00761007	1	0.00680175
第五组	0.00686955	0.00255567	0.00282677	0.00680175	1

在此表格中我们可以发现每一组与其他组的相似度，相关程度系数越大代表着这个问题被关注的越多。一般情况下，如果两条信息之间的相似度大于等于 0.05，那么我们就认定这个问题就是我们所要挖掘的热点问题。在表中，如果我

们将所有的数据全部汇总出来，做成表格样式，就能发现表格数据按照对角线是对称的，而且每组数据对自己来说相关度都是 1.那么我们就可以简化表格，简化之后的表格内容如下表 4-3：

表 4-2　简化后相关程度系数

	第一组	第二组	第三组	第四组	第五组
第一组	1				
第二组	0.0016852	1			
第三组	0.000208016	0.00315634	1		
第四组	0.00107792	0.000987825	0.00761007	1	
第五组	0.00686955	0.00255567	0.00282677	0.00680175	1

简化后的表格，在数据运算以及存储方面都具有一定的优势。也方便我们对热点问题的挖掘。

4.3 问题三

对于最后一个问题，群众提出问题之后，需要政府相关部门进行公务处理，最终形成一个整体闭环。从问题提出到问题解决，一步都不能缺少，但是对于群众提出的问题处理情况如何，我们需要有政府相关公务部门对群众提出的问题进行合理的答复，所以现在需要对答复的内容与群众的反映情况提取关键词，对比两者的关键词，分析他们的相关度。只要他们之间的相关度大于一定的数值，或者满足一定的条件。那么就可以确定政府的公务处理是满足可行性以及解释性要求的，即答复是合格的。下面列出几组数据来观察一下，见表 4-4：

4-4 答复与反映的相关度系数

组别	相关度系数	组别	相关度系数
第一组	0.2790943	第八组	0.21831232
第二组	0.06706527	第九组	0.1556399
第三组	0.21698096	第十组	0.2983386
第四组	0.11368805	第十一组	0.1575496
第五组	0.17946064	第十二组	0.12698989
第六组	0.05406117	第十三组	0.17849246
第七组	0.22028741	第十四组	0.08606629

经过机器语言的识别，就可以得到群众反映的情况跟政府公务回复之间的相关度系数，我们对其进行辨别，并且规定只要相关系数小于 0.1，那么我们就认定此回复与群众反映的情况不匹配，也就是回复不合格，根据运算的结果，一共有 667 组答复是不满足要求的。

五、模型评价与改进

5.1 训练数集的缺陷

在第一个题目中，我们挑选了一部分的用户留言数据作为训练集，并且由此训练集，来分析剩余部分的分类效果。但是，其中的一个缺点就是训练集的数量大小会影响后续对样本处理时的准确度。如果我们能有一个提前确定好的分组，这样对后续的处理将有一个很好的帮助作用。

5.2 在第二个以及第三个问题中，我们采用了相关程度大小的比较方式。但是相关程度的比较有一部分取决于我们人主观上的确定以及筛选。而我们人为的确定与筛选又会直接影响着最终结果的准确程度。

六、参考文献

[1] 聂笃忠，陈 桦，米承继，彭礼红.马尔科夫链状态概率转移矩阵修正算法[J]. 统计与决策，2013(3):14-17.

[2] 廖普明.基于马尔科夫链状态转移概率矩阵的商品市场状态预测[J].统计与 决策，2015(2):97-99

[3] 谢盛嘉.大数据时代背景下数据挖掘技术的应用研究[J].计算机产品与流通,2020(05):128.

[4] 陈银娣,王三梅.大数据时代装备科技信息研究系统探索——基于高端需求和信息挖掘技术的装备科技信息研究方法[J].情报理论与实践,2020,43(04):14-17.

[5] 臧玉魏,谢连科,张永,张国英,吴健,白晓春.基于电力营销聚类分析的数据挖掘算法研究[J].信息技术,2020,44(04):56-59+64.

[6] 金琳.基于数据挖掘的用户行为分析研究[J].电子商务,2020(04):41-42.

[7] 白萍.计算机数据挖掘技术的开发及应用[J].电子世界,2020(07):160-161.

[8] 李小庆.基于大数据的客户关联风险分析和挖掘[J].金融科技时代,2020,28(04):20-24.

[9] 乔岩.大数据背景下统计数据质量的优化策略[J].中外企业家,2020(12):91.

[10] 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.

[11] 苏悦来.计算机数据挖掘技术的开发及应用[J].计算机与网络,2020,46(06):44.

[12 佚名 使用 python 和 sklearn 的中文文本多分类实战开发] csdn