“智慧政务”中的文本挖掘应用

摘要

近年来，随着互联网的飞速发展，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。如何在享受互联网的好处的同时解决这因便利而带来的巨大的工作量问题成为我们当前需要思考的问题。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，能有效的解决留言划分和热点整理等问题，减少相关部门的工作量。

对于问题 1，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。可以转化为多类分类问题，利用 SVM 分类器进行一步一步求解。

对于问题 2，热点问题挖掘。利用 Python 进行自然语言处理。第一步：利用已有的语料进行语料预处理。首先将不感兴趣的，无关紧要的内容进行清洗删除，然后将剩下的文本全部进行分词，得到文本挖掘分析所需的最小单位粒度的词或词语。对每个词或者词语打词类标签，如形容词、动词、名词等。这样做可以让文本在后面的处理中融入更多有用的语言信息。最后将对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词去除。第二步：特征工程。做完语料预处理之后，把分词之后的字和词语表示成计算机能够计算的类型。

关键词：SVM 多分类 K-means 聚类原理 关联规则挖掘

In recent years, with the rapid development of Internet, WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence,

condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. How to enjoy the benefits of the Internet and at the same time to solve the problem of the huge workload caused by convenience has become a problem we need to think about. As the big data, cloud computing, artificial intelligence, such as the development of technology, based on natural language processing technology the wisdom of the e-government system has is the new trend of development of social management innovation, to enhance the management level of government and governance efficiency has a great role, can effectively solve the problem of message classification and hot finishing, reduce the workload related department.

For question 1, according to the data given in annex 2, a first-level label classification model for message content is established. It can be transformed into multi-class classification problem and solved step by step by using SVM classifier.

For question 2, hot issue mining. Use Python for natural language processing. The first step: using the existing corpus for corpus pretreatment. First, the content that is not interested in is cleaned and deleted, and then all the remaining text is segmented to get the word or word with the minimum unit granularity required for text mining analysis. Label each word or word as an adjective, verb, noun, etc. Doing so allows the text to incorporate more useful language information into later processing. Finally, words that do not contribute to the characteristics of the text, such as punctuation, mood, person, and so on, are removed. Step 2: feature engineering. After the corpus preprocessing, the words and words after word segmentation are expressed as the types that the computer can calculate.

Key words: SVM multi-classification k-means clustering principle association rule mining

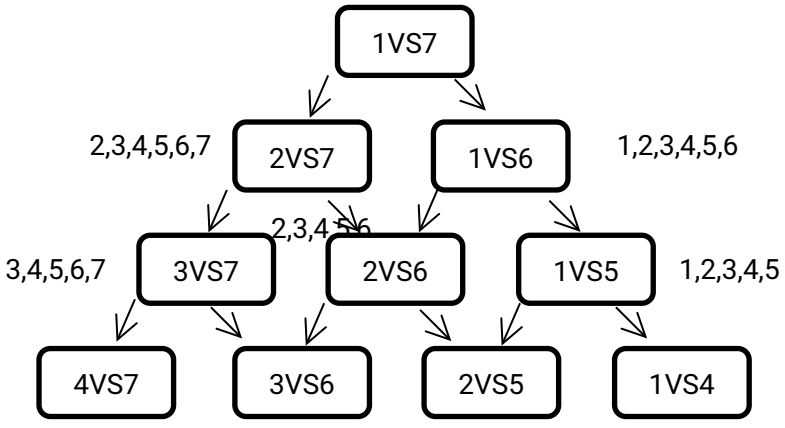
目录

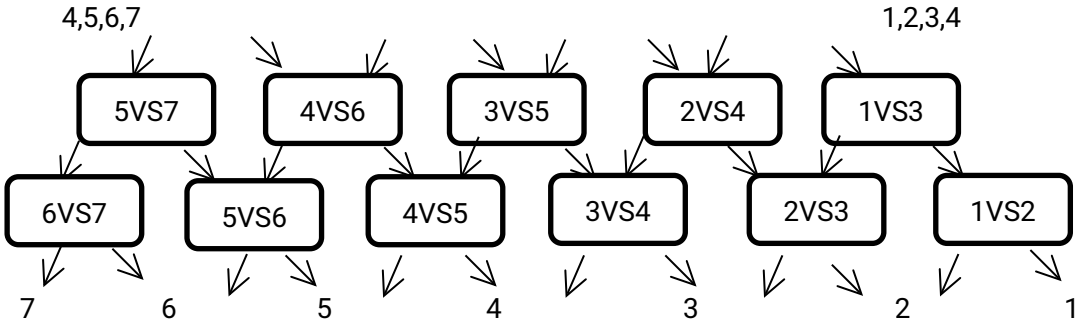
1.挖掘目标

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，为了让相关的职能部门更加高效快速地获取群众留言信息,减少人工处理的工作量。本次建模目标是利用自然语言处理和文本挖掘的方法解决问政平台引出的问题。在对群众留言数据进行属性数值化，去重、去空，中文分词，停用词过滤后。一方面，利用 SVM 分类器根据分类标签对群众留言内容进行合理分

类，建立一级标签分类模型，减小人工根据经验处理数据的弊端，提高工作效率，降低差错率。另一方面，给某一时间段内反映特定地点或特定人群问题的留言进行归类，经过成分分析，对热点问题进行综合排名，以便相关的职能部门快速有效地对地方热点问题进行处理，给群众一个满意的答复。能够从答复的相关性、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，对他工作人员有一定的借鉴作用，更好地解决地方社会问题。

问题一的分析方法与过程：建立关于留言内容的一级标签分类模型，从附件 2 给出的数据中，我们可以知道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，卫生计生共七大类。一条留言内容只对应一个一级分类。要想一次性考虑所有分类，并求解一个多目标函数的优化问题，一次性得到多个分类面，这个求解计算量实在太大，无法实行。这时，我们考虑一类对其余的方法，并将它转化为我们熟悉的解一个两类分类（SUM 是一种典型的两类分类器）的问题。我们一级分类总共有 7 个类别，为解题方便，我们将它依次标记为 1 到 7。每次选一个类的样本作正类样本，而负类样本则变成只选一个类（称为“一对一”的方法，），第一个只回答“是第 1 类还是第 2 类”，第二个只回答“是第 1 类还是第 3 类”，第三个只回答“是第 1 类还是第 4 类”，如此下去，我们可以马上得出，这样的分类器应该有 7 X 6/2=21 个（通式是,如果有 k 个类别,则总的两类分类器数目为 k(k−1)/2）。从通式我们可以知道 k 越大，分类器就越多，当 k 是 1000，要调用的分类器数目会上升至约 500,000 个（类别数的平方量级）。这个方法存在巨大的缺陷，我们可以在此基础上加以改进。同样用一对一的方法，但对一条留言进行分类之前，我们先按照下面图的样子来组织分类器。（这是一个有向无环图，因此这种方法也叫做 DAG SVM）





分析结果

这样在分类时,我们就可以先问分类器“1 对 7”（意思是它能够回答“是第 1 类还是第 7 类”），如果它回答 7，我们就往左走，再问“2 对 7”这个分类器，如果它还说是“7”，我们就继续往左走，这样一直问下去，就可以得到分类结果。好处在哪？我们其实只调用了 6 个分类器（如果类别数是 k，则只调用 k-1 个），分类速度飞快，且没有分类重叠和不可分类现象！缺点在哪？假如最一开始的分类器回答错误(明明是类别 1 的留言内容,它说成了 7),那么后面的分类器是无论如何也无法纠正它的错误的。（因为后面的分类器压根没有出现“1”这个类别标签）。

2.总体流程与步骤

2.1 总体流程：

步骤一:对数据进行预处理，对附件 1 结构化文本数据数值化处理，对附件 3 非结构化文本去除重复项及空行，中文文本分词、停用词过滤，以便后续分析；

步骤二:文本向量化,基于 TFIDP 权重法提取关键词,构造词汇-文本矩阵,进而利用奇异值分解算法进行语义空间降维，去除同义词的影响，简化计算。

步骤三:文本聚类,根据文本向量,计算文档间的欧式距离,再基于 k-means 聚类算法对各个留言主题（留言内容）描述进行聚类。

步骤四:需求分析与预测，构造排名算法判断热点问题，并引入特定地点、特定人群、时间、群众点赞因素预测短期相应热点问题，对热点问题深入分析其热度指数。

步骤五:利用上述结果分析热点问题及对应的留言信息，并定义合理的热度评价指标。

3.文本聚类

3.1 数据预处理

3.1.1 数据描述

通过观察所给数据，可以发现数据量比较大(共四千多条记录)，且附件 1:内容分类三级标签体系中的字段大多为文本格式,需要将其量化成数值形式才能对其进行分析。而附件 3 留言数据中有大量标点符号、语气、人称等一些词以及重复的情况，如果不做处理会对后续分析造成影响，如果把这些数据也引入进行分词、词频统计乃至文本聚类等，则必然会对聚类结果的质量造成很大的影响，于是本文首先更对数据进行预处理，去除对文本特征没有任何贡献作用的字词。

3.1.2 文本预处理

我们把这些文本数据的预处理分为四个部分:

属性数值化

对于附件 1：内容分类三级标签体系,一级分类、二级分类、三级分类等指标,需要将其数值化，例如:城乡建设、党务政务、国土资源、 环境保护、 纪检监察 、交通运输、 经济管理、科技与信息产业、 民政 、农村农业、 商贸旅游 、卫生计生、 政法 、教育文体、 劳动和社会保障编码转换为 B1....B15。

去重、去空

对于附件 3 留言数据存在大量标点符号、语气、人称等一些词以及重复的样本。

中文分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要分词，本文采用 Python 开发的一个中文分词模块—jiebe 分词[1]，对附件 3 中每一个留言主题、留言详情进行中文分词，jieba 分词用到的算法:

◆基于 Trie 树结构[2]实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)

◆采用了动态规划查找最大概率路径，找出基于词频的最大切分组合

◆对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

jieba 分词系统提供分词，词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

图 3.1 的分词结果是没有停用词过滤的结果，从图中可以看到，其中有大量标点及表达无意义的字词，对后续分析会造成很大影响，因此接下来需要进行停用词过滤。

停用词过滤

为提高搜索效率节省存储空间和节省存储空间，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为 Stop Words(停用词)。停用词有两个特征:一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。

为了找出这些停用词，需要–些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应枝看作是噪声词。因此用以下指标衡量词语的有效性：

词频(T)

TF 是一种简单的评估函数，其值为训练集合中此单词发生的词频数。TP 评估函数的理论假设是当一个词在大量出现时，通常被认为是噪声词。

文档频数(DF)

DF 同样是一种简单的评估函数，其值为训练集合中包含此单词的文本数。DF 评估函数的理论假设是当一个词在大量文档中出现时，这个词通常被认为是嗓声词。

本文选用 DF 方法筛选出如下停用词：我，有，的，了，是等。将筛选出的停用词加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则进行删除处理。

3.4.文本聚类

3.4.1 文本相似度计算

相似度是用来衡量文本间相似程度的一一个标准。在文本聚类中，需要研究文本个体的差异大小。也就是需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。目前相似度计算方法分为距离度量和相似度度量。本文采用的是基于距离变量的欧几里得距离计算留言详情文本见差异。

欧氏距离[12](Eucliden Distance)：

令 $i=(x_1, x_2,\ldots, x_n)$ 和 $j=(y_1, y_2,\ldots, y_n)$ 是两个被 p 个数值属性标记的对象 i 和 j 之间的欧式距离定义为 $dis(i, j)=\sqrt{\sum_{k=1}^{p}(x_k-y_k)}$ 。

3.4.2 文本聚类

所谓文本聚类就是将无类别标记的文本信息根据不同的特征,将有着各自特征的文本进行分类,使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法根据不同的留言主题、留言详情进行分类。通过聚类方法，可以对每一条留言进行分类。

3.4.2.1 K-means 聚类原理

means 算法[13]是很典型的基 f 划分的聚类算法。采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似性就越大。

K-mears 算法的基本思想是:以空间中 k 个点为中心进行聚类,对最靠近他们的对象归类。通过迭代的方法。逐次更新各聚类中心的值，直至得到最好的聚类结果。

假设要把样本集分为 k 个类别，算法描述如下:

适当选择 k 个类的初始中心；

⑵在第 k 次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类；

⑶利用均值等方法更新该类的中心值；

对于所有的 k 个聚类中心，如果利用（2）⑶的选代法更新后，值保持不变，则选代结束，否则继续迭代。

该算法要求在计算之前给定 k 值。本文通过计算附件 1 给出的内容分类三级标签体系中的一级分类，并以此作 A 的值，这里令 k=15，即城乡建设、党务政务、国土资源、 环境保护、 纪检监察 、交通运输、 经济管理、科技与信息产业、 民政 、农村农业、 商贸旅游 、卫生计生、 政法 、教育文体、劳动和社会保障。

3.4.2.2 K-means 聚类结果

4.热点问题挖掘

4.1.1 热点问题的定义

1.某一时段内群众集中反映

2.受广大群众的关注

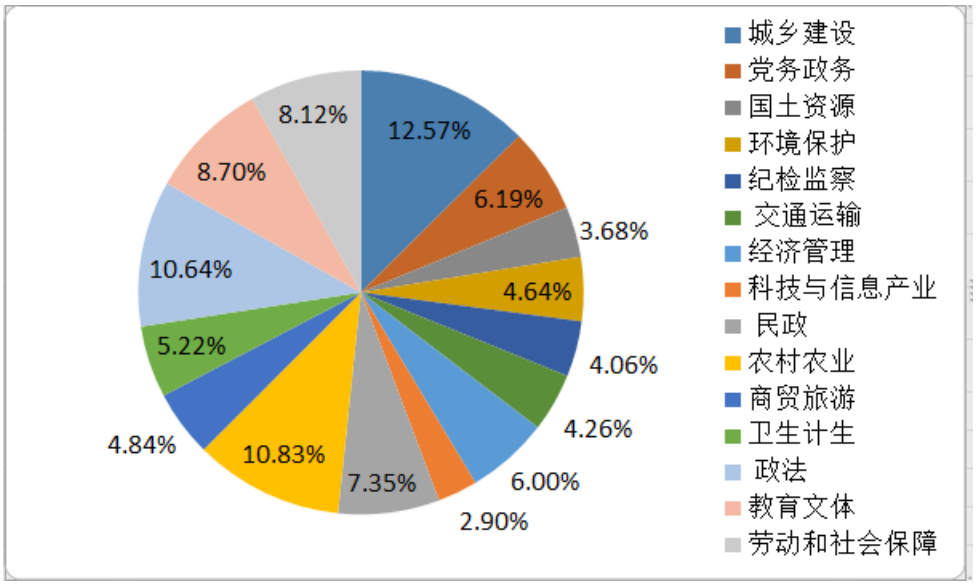
3.关注度高，引起轰动

针对上述的分析，结合所给的数据，本文用一定时间、一定范围内公众最为关注的问题描述热点问题。

4.1.2 数据预处理

由内容三级标签体系（附件一），城乡建设占 12.57%，其中城乡建设和市政管理、住房保障与房地产占 24.65%，农村农业占 10.83%，其中林业管理、水利水电占 15.50%， 政法占 10.64%，其中社会治安占 18.18%。

表 2.1



对于附件一，文本已经转化为数值型数据，对于留言主题，根据需求将其划分等级分类。

4.2.主成分分析综合排名算法

4.2.1 基本原理

主成分分析[15]又称为主分量分析或主轴分析，将多指标转化为少数几个综合指标的一种同计分方法。在实际问题中，研究很多变量的问题是经常遇到的，并且彼此之间有一定的相关性，因而使得所观测到的数据在一定程度上反映的信息有所重叠。而且在变量较多时，在高维空间中研究样本的分布规律比较复杂，势必增加分析问题的复杂性。人们自然希望用较少的综合变量来代替

原来较多的变量，而这几个综合变量又能尽可能多的反映原来变量的信息，并且彼此之间互不相关。

4.2.2 热门话题排名

据统计，附件一中有 15 类一级分类，114 个二级分类，在一级分类中首先筛选出排名前 5 的话题占 50.87%，利用 SPSS 对其进行综合排名。

一级分类	数量	比重
城乡建设	65	12.57%
农村农业	56	10.83%
政法	55	10.64%
教育文体	45	8.70%
劳动和社会保障	42	8.12%
民政	38	7.35%
党务政务	32	6.19%
经济管理	31	6.00%
卫生计生	27	5.22%
商贸旅游	25	4.84%
环境保护	24	4.64%
交通运输	22	4.26%
纪检监察	21	4.06%
国土资源	19	3.68%
科技与信息产业	15	2.90%
合计	517	100.00%

热门话题排名前五的是：城乡建设、政法、农村农业、教育文体、 劳动和社会保障。

5.大数据留言详情

5.1 数据筛选

5.2 关联规则挖掘

5.3 留言分布情况（话题分布情况、地域分布情况、领域分布情况）

5.4 大数据留言特征

6 评价体系

7.民生民意、政策、满意度、了解民意、改善民生

7.1 数据

7.2 现状

7.3 民生需求

参考文献

⑴SVM 学习(六)： 将 SVM 用于多类分类

⑵孙海锋，郑中枢，杨武岳，网络招聘信息的数据挖掘与综合分析，北京林业大学，2017

⑶贾园园，蔡黎，饶希，网络招聘信息的分析与挖掘，湖北工程学院，2019

⑷张彬城，陈杰彬，林越，一种基于潜在语义索引和卷积神经网络的智能阅读模型，暨南大学，2019