“智慧政务”中的文本挖掘应用研究

摘要

近年来，随着大数据、云计算、人工智能的迅速发展，基于自然语言处理技术的智慧政务系统极大地提高了政府的管理水平和施政效率，解决了民意相关的文本数据量增多给带来的人工划分的困难。本文旨在运用自然语言处理和文本挖掘的方法对群众问政留言记录进行文本分类和热点问题挖掘。

针对问题一，利用提取附件 1 里不同一级分类对应的三级分类的所有词组，分别汇成一个文本文件，得到 15 个文本文件，将文件里的每一个词组看成是一个特征项。对附件 2 的留言主题信息利用 jieba 中文分词工具进行分词，再利用定义好的停用词表进行去停用词操作，剔除无用的虚词，语气词等词组，得到含只有主要信息的留言主题。接着，利用布尔模型和用布尔表达式的方法来表示主题留言的信息，对一级标签的分类。最后用 F-Score 对分类方法进行评价。

针对问题二，我们利用改进 k-means 聚类算法，通过聚类分析出某一时段内、反映特定地点或特定 人群的问题通过数据分析，数据筛选，数据比较从而得到对应的结果。并且通过 python 数据爬虫处理，合成分析了我们所对应的热点问题。最后通过热点问题的点赞数和反对数的数据合成。得到相应的结果，并且得到很好的验证。

针对问题 3，首先对附件四中的数据进行去重处理、归类，使用统计语言模型来将现实问题抽象化。在问题 2 的基础上，对热点数据进行统计并计算其概率，计算出热度问题的频数，对其进行有效的答复。将数据分为三级标签后再特点地点、特点时间进行归类，利用自然语言的语句分析数据。然后编写好答复格式与基本文本，以及各式各样的答复的分词，把重要的分词与留言信息做匹配，从而高质有效的对留言信息进行答复。最后，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。

关键词：中文分词、布尔权重计算方法、F-Score 方法、TF-IDF 权重

一、 挖掘背景

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。

二、 问题分析

问题 1，要求根据附件 1 提供的内容分类三级标签体系依据附件 2 的留言内容进行文本分类，建立相应的一级标签分类模型，以便高效率地将群众留言分派至相应的职能部门处理，最后使用 F-Score 方法对分类的方法进行评价。

问题 2，我们利用数据清洗，数据合成以及数数据处理分析等手段进行对应的处理和分析，然后将问题进行识别、归类以及热度评价，最后根据数据的特点运用 k-means 聚类算法进行分合成比较。从众多留言中识别出相似的留言按照三级标签进行归类后再详细归类，比如把特定地点或人群或特定时间的数据归并，即把相似的留言归为统一问题，再根据用户对某个问题评论进行热度的评判和分析，即点赞数以及反对数的和，从而得到对应的热度问题。

问题 3，通过以上对数据的分析与归类，建立相对应的模型，对每一类信息以及特定信息进行高质有效的答复，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。利用信息评价模型对答复进行评价。

三、 模型假设

1.假设所给的数据真实，可靠。

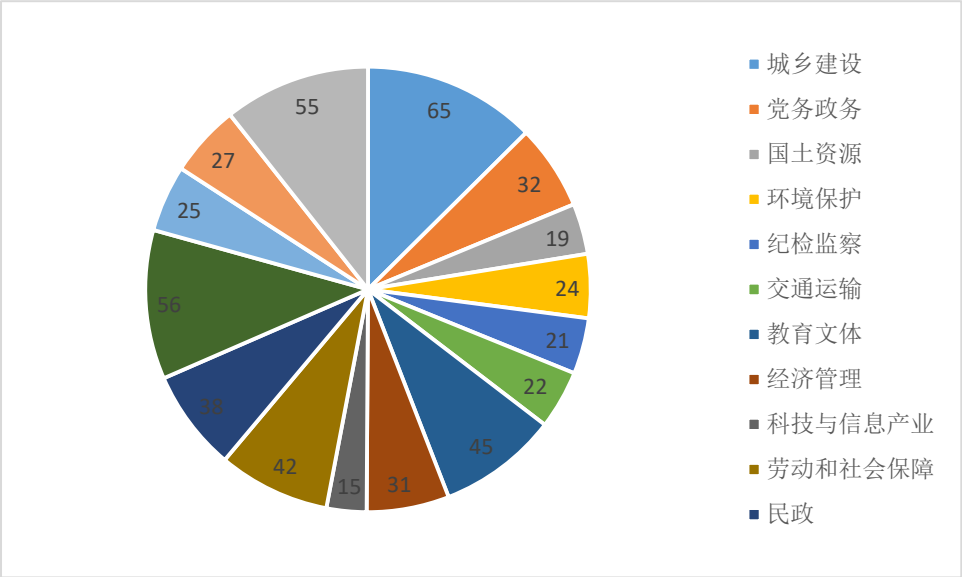
2.假设附件 3 的每一行留言记录的点赞数和反对数都是民众在认真阅读留言后发表的。

四、 问题求解

4.1 任务一

4.1.1 获取各个一级分类的特征文本

在附件 1 给出的三级标签体系中，一级分类、二级分类、三级分类呈层层递进的关系。通过 python 和 excel 将每一个一级分类对应的三级分类所在列的所有词组分别提取出来，合成不同的文本文件。这些文本文件里的词组作为对应的一级分类标签的特征词组。生成的文本文件分别保存为：城乡建设.txt、文本 2 党务政务.txt、国土资源.txt、环境保护.txt、纪检监察.txt、交通运输.txt、教育文体.txt、经济管理.txt、科技与信息产业.txt、劳动和社会保障.txt、民政.txt、农村农业.txt、商贸旅游.txt、卫生计生.txt、政法.txt。各个一级分类的占比如图：



4.1.2 留言内容的进行中文分词、去停留词

留言内容是文本数据，属于非结构化数据，而计算机只能识别结构化数据，所以需要先将文本数据转化为计算机可以识别的数据类型以方便后续操作。为了方便转换，需要先对附件 2 里的留言主题进行中分分词。这里采用 python 的中文分词库包 jieba 进行分词。而停用词通常是作为句子的辅助成分而存在，本身的意思对整个句子的影响甚微，过度的增加停用词不会增加分类性能反而会降低分类性能。因此，在对留言主题分词后去停用词过程是有必要的。这里采用基于定义好的文本文件 stopword.txt 利用 python 对与 stopword.txt 内容相同的文本数据进行剔除操作，得到只保留主要内容的留言主题文本内容。

4.1.3 基于布尔权重计算方法的一级分类

布尔权重计算方法如下：

需要处理的文本用布尔表达式的方法来表示，而且转变后的文本会比原来的文本意思更加明确、呈现的形式也更加简洁明了，更便于计算，每一个留言主题内容是由由多个待查找的特征项 t 来组成，而且我们认为特征项 t 会在步骤 1.1 生成的其中一个文档 di (i=1,2,3...15)中存在，要么不存在，因此其权重大小只有 0 和 1 两种取值。以此对留言主题内容进行分类。

4.1.4 F-Score 对分类方法进行评价

依据分类结果对附件 2 的留言内容的方法实验评价指标采用了 F1 值，查准率和查全率是文本分类中不同的衡量准侧，F1 值是综合考虑准确率和召回率的评价指标。公式为：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

4.2 任务二

4.2.1 划分时间段

用 excel 对附件 3 的留言时间进行排序，可得留言时间总的时间跨度为 2017/6/8 17:31:20 至 2020/1/8 9:32:33，将这个时间范围每隔 3 个月划分成为一个时间区间如下表:

时间类别	时间段
1	2017/6/8 至 2017/9/8
2	2017/9/8 至 2018/12/8
3	2017/12/8 至 2018/3/8
4	2018/3/8 至 2018/6/8
5	2018/6/8 至 2018/9/8
6	2018/9/8 至 2018/12/8
7	2018/12/8 至 2019/3/8
8	2019/3/8 至 2019/6/8

9	2019/6/8 至 2019/9/8
10	2019/9/8 至 2019/12/8
11	2019/12/8 至 2020/1/8

将留言记录按照此表划分成 11 个类别。

4.2.2 留言主题的预处理

首先对留言主题的文本用 jieba 库的中文分词进行分词操作，经过分词处理后，留言主题的文本内容被切割成一个个用“ ”隔开的字、词。再依据文件 stopword.txt 去除掉停用词，即无意义的虚词，其余的就作为文本的特征项。这些特征项能够反映出留言的真实含义，是说明留言内容的主要元素，也是区别于其它留言内容的特征。

4.2.3 基于 TF-IDF 权重算法的留言主题的特征项向量化

要想让计算机理解处理好的留言内容，需要对文本内容进行的数学化表示，依据 TF-IDF 权重算法对能显著体现文本内容特征的特征项赋予高权重，而对不能可以体现文本内容特征的特征项赋予低权重。从效率方面来说，特征项权重的计算对下面文本相似度计算的整体效率具有重要的影响。

TF-IDF 权重是向量空间模型中应用最多的一种权重计算方法，它以词语作为文本的特征项，每个特征项的权重由 TF 权值和 IDF 权值两个部分构成。对于文本 d_i 中的第 k 个特征项 $t_{i,k}$ 其对应权重计算方法为：

$$w_{i,k} = TF_{i,k} * IDF_{i,k}$$

其中 ：

1. TF（Term Frequency）权值:特征项在文本中出现的次数，即如果$t_{i,k}$在文本d_i中出现$n_{i,k}$词，那么

$$TF_{i,k} = n_{i,k}$$

2. 在实际应用中，通常需要对 TF 值进行标准化处理，以避免文本太长所导的统计差：

$$TF_{i,k} = \frac{n_{i,k}}{\sum_m n_{m,k}}$$

3. IDF（Inverse Document Frequency）权值：特征项在全局文本集 D 中的出现频率，即:

$$IDF_{i,k} = \log \frac{|D|}{|\{d : t_{i,k} \in d\}|}$$

假设全局文本集共有 M 篇文本，特征项$t_{i,k}$，共在$m_{i,k}$篇文章中出现过，那么

$$IDF_{i,k} = \log\left(M / \left(m_{i,k} + \alpha\right)\right)$$

其中α为经验常数，一般取 0.01。

4.2.4 留言内容相似度计算

得到留言内容的特征向量以后，文本d_i和d_j之间的相似度$S\left(d_i,d_j\right)$可以通过它们的特征向量之间的关系来衡量。如果两文本的特征向量为

$$V_{d_i}=\left(w_{i,1},w_{i,2},\cdots,w_{i,n}\right),V_{d_j}=\left(w_{j,1},w_{j,2},\cdots,w_{j,n}\right)$$

并且两特征向量在空间中的夹角为θ，根据余弦系数作为它们的之间相似度衡量方法，公式为：

$$S\left(d_i,d_j\right)=\cos\theta=\frac{\sum_{k=1}^{n}w_{ik}*w_{jk}}{\sqrt{\left(\sum_{k=1}^{n}{w_{ik}}^2\right)\left(\sum_{k=1}^{n}{w_{jk}}^2\right)}}$$

即用两个向量的夹角余弦值来衡量不同留言记录的留言主题之间相似度。S越大，表示两个留言主题之间的相似程度越高。将文本相似度较高的文本归为一类并把同一类的文本类标上阿拉伯数字编号即“问题ID”以示区别（1.2.....m,假设m为分成的类别总数），同时统计每一个文本类里包含的个数。

4.2.5 定义热度评价指标

用R表示热度指数，y表示某一行数据的点赞数，n表示某一行数据的反对数，X_i（i=1,2，....m）表示该行数据所在的文本类的个数的总数。

将每一行数据的热度指数设为：$R_i=X_i+y+n$

统计每一行数据的热度指数R_i（j=1,2,....M,M为表的总的数据行数）

4.2.6 热点问题的获取

根据定义好的热度评价指标，依据热度指数对每一行的留言记录进行排列，取热度指数最高前五位作为热点问题。

4.3 任务三

对附件 4 中的数据进行获取后，将数据信息进行分词、去停用词后利用统计语言模型，将数据信息中相似的数据进行统计，然后把相同的问题进行去重处理，再将得出的数据处理进行存储与归类。按照问题 1 中的三级标签进行归类后，再根据特定时间、特定地点、特定人群与事物进行详细归类。编辑好答复的格式与基本文本，将详细归类信息中重点分词进行输出，针对该分词所对应的问题提出解决方案或是其余的答复。然后对此方案中的重点分词进行拆分，与留言信息中的重点分词相比较后，再利用分词重组对留言信息进行答复。

答复中三大重要角度为相关性、完整性、可解释性，相关性意为答复意见的内容是否与问题相关，完整性意为是否满足某种规范，可解释性意为答复意见中内容的相关解释。建立一个信息评价模型，针对答复是否符合三大重要角度进行评价，若评价低于某数值，则该答复失效，对于失效的答复进行删除处理后重新进行答复；若评价数值高于某数值，则答复起效，并将该答复输出。对数据进行一系列操作后，确定信息评价数值。最后将所有答复对应留言信息一一输出，从而达到留言的答复意见具有高质有效性。

五、总结

本文主要目的利用文本挖掘与自然语言处理建立留言答复评价模型。首先，对留言信息进行可视化，分析并提取留言有效信息，通过问题识别、归类、热度评价对问题深入分析，再对留言进行有答复以及对答复的热度评价，为建立热度批评家模型的建立做准备。

首先，利用 jieba 中文分词工具进行分词，再利用定义好的停用词表进行去停用词操作，剔除无用的虚词，语气词等词组，得到含只有主要信息的留言主题。接着，利用布尔模型和用布尔表达式的方法来表示主题留言的信息，对一级标签的分类，再用 F-Score 对分类方法进行评价。其次，识别出相似的留言以及留言重点分词，把特定时间、地点或人群的数据归并，即把相似的留言归为同一问题，利用热度评价指标的定义和计算方法，对指标排名之后得出热点问题表。最后，针对每一特殊留言进行有效答复，为保证答复的相关性、完整性、可解释性，建立答复信息评价模型，确保答复具有高质有效性。

六、参考文献

[1] 李惠富.文本分类中特征提取及分类算法的研究[D].东北林业大学，2018.

[2] 康东.中文文本挖掘基本理论与应用[D].苏州大学,2014

[3] 吴多坚.基于 word2vec 的中文文本相似度研究与实现[D].西安电子科技大学，2016

[4]陈飞宏 .基于向量空间模型的中文文本相似度算法研究[D].电子科技大学，2011

[5]陈纯柱.樊锐.网络问政平台建设研究[A].重庆邮电大学，2015 年第 3 期

[6]叶吉祥，王聪慧 . 改进的 F_score 算法在语音情感识别中的应用[A] . 长沙理工大学，2011

[7]马倩 . 智慧政务的服务流程优化研究[D] . 燕山大学，2016

[8]吴钧鸿 .基于智慧政务的社会治理创新方法[J] . 浙江万里学院，2019（42）

[9]张腾，张建光，尚进 . 基于 DPSIR 模型的智慧政务信息生态评价研究[J] . 中国科技论坛，2017（2）

[10]孙润志 .基于语义理解的文本相似度计算研究与实现 [D] . 中国科学院研究生院（沈阳计算技术研究所） ，2015