摘要

智慧政务是利用数据挖掘、云数据计算、移动互联网等技术，旨在为政府办公、监管、服务、决策的智能化水平提高，从而创造形高效、敏捷、便民的新型政府。但随着移动互联网的不断发展进步，各种网络问政平台（例如微信、微博、市长信箱、阳关热线等）的推广，拉近了市民与政府之间的距离，但随之而来的不断攀升的市民对政府的决策、部署的意见的留言，给各部门以往通过人工处理这些市民留言和热点整理、划分工作带来了极大挑战。

本文我们采用 python 和 R 对附件给出的文件进行文本挖掘应用，首先根据附件 2 的结果使用 F-score 方法倒推留言 1 级大类分类方法；接着使用 R 软件对附件 3 的数据进行数据处理与挖掘应用，提取指标反对数、赞成数与留言时间，建立合理评价模型得到排名前 5 的热点问题及其对应的留言信息；最后，通过 python 软件对附件 4 的数据进行提取，并通过算法对其相关性、完整性和可解释性做评价并得到评价方案。

关键词： 排序　时间间隔　文本相似度　统计词频

Abstract

Smart government is to use data mining, cloud data computing, mobile Internet and other technologies to improve the intelligent level of government office, supervision, service and decision-making, so as to create a new government with high efficiency, agility and convenience for the people.However, with the continuous development and progress of the mobile Internet, the promotion of various online political platforms (such as wechat, Weibo, mayor's mailbox, Yangguan hotline, etc.) has narrowed the distance between the citizens and the government. However, with the continuous increase of the opinions of the citizens on the decision-making and deployment of the government, these messages and hot spots have been sorted out and planned for all departments in the past through manual processing The division of work brings great challenges.

In this paper, we use Python and R to do text mining application for the files given in the attachment. First, we use F-score method to backward push the class 1 Classification Method of message according to the results of attachment 2. Then

we use R software to process and mine the data of attachment 3, extract the number of objection, approval and message time of indicators, establish a reasonable evaluation model to get the top 5 Finally, we extract the data of attachment 4 through Python software, and evaluate its relevance, integrity and interpretability through algorithm and get the evaluation scheme.

目录

二、挖掘目标

本次挖掘的目标是智慧政务的文本挖掘应用，运用 R、Python 对所有数据提取出其中的有效数据，排除干扰项，通过有效数据得到关于留言信息的 1 级分类的模型。然后通过编写 R 代码，并根据留言时间、点赞数、反对数建立模型，得到最合理热度评价指标。最后，通过 python 软件运行算法，对有效数据进行文本相似度和完整性及可解释性性做分析，得到评价方案。

三、方法分析及过程

1.问题一求解

1.1 方法分析

根据附件 1，我们得到了每个市民的留言信息的 1 级分类，然后我们使用 F-score 算法，F-score 是衡量分类问题的指标。本题我们使用该算法，对所有数据进行计算查准率和查全率，然后对比发现规律，从而得到

1.2 求解过程

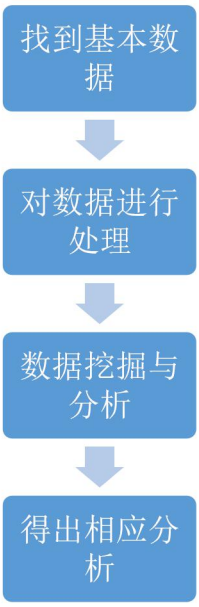
1.2.1 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提 供的内容分类三级标签体系）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。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且 差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。 通常使用 F-Score 对分类方法进行评价：

其中 iP 为第 i 类的查准率， iR 为第 i 类的查全率。

1.2.2 相应的流程与图像



首先分类把一级二级三级分类找出来并且找到他们的关系然后根据公式找出最显著的问题所在。

通常使用 F-Score 对分类方法进行评价：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

1.3 结果分析

①根据公式在城乡建设中出现最多的问题是城乡建设带给老百姓的困惑与不便的问题。

②根据公式在环境保护中出现最多的问题是河水的污染与周围工厂的噪音问题。

③根据公式在交通运输中出现最多的问题是道路与车辆的随便停放带来的相关问题。

④根据公式在教育文体中出现最多的问题是乱收费与教师资源的不足。

⑤根据公式在劳动和社会保障中出现最多的问题是工资的保障问题。

⑥根据公式在商贸旅游中出现最多的问题是价格是否标准是否欺骗老年人。

⑦根据公式在卫生计生中出现最多的问题是保健品的安全与医疗问题。

2.问题二求解

2.1 方法分析

本题为热点问题挖掘应用，主要是建立合理模型挖掘前五的热点问题及其对应的留言信息。我们根据附件 3 给出的热点问题（群众一段时间内集中反映的问题）建立模型给出合理热度评价指标，并给

出包含热度前五的问题的文件，命名为“热点问题表”；同时给出包含对应留言信息的文件，命名为“热点问题留言明细表”。

2.2 符号说明

符号	说明
S	点赞数的时间比率
U	点赞数
D	反对数
T	发布时间间隔
t	当前时刻与某一时间的留言时间差
T_A	当前时刻 A 的时间间隔
T_B	时刻 B 的时间间隔
x	点赞数与反对数之差
z	平滑反对数
y	平滑赞成数

2.3 求解过程

过对问题的分析，我们根据留言时间、反对数、点赞数建立评价指标使模型合理，并得出前五的热点问题和对应的留言信息。模型建立如下：

$$S=(U-1)\div(T+2)^{10}$$

通过上述建模我们得到，当S越大时，说明点赞数的时间比率越大，问题越好。

$$t=T_A-T_B$$

$$x=U-D$$

问题的留言时间差越大，问题的排名越靠后；同时，点赞数与反对数之差越大，说明点赞数越大，即问题越好。

$$z=\begin{cases}|x|,|x|\neq 1,x>0\\ 1,x<0\end{cases}$$

$$y=\begin{cases}1,x>0\\ 0,x=0\\ -1,x<0\end{cases}$$

$$f(t,y,z)=\lg^{z}+\frac{yt}{4}$$

2.4 结果分析

通过对反对数与点赞数的$\lg$平滑，我们遍历得到所有问题的f值，得到排名前五的f值：

表 1：排名前 5 的 f 值

6.5	6.3	4.5	3.8	1.9

对应达到前五的点赞数和反对数：

表 2：排名前 5 的点赞数和反对数

	1	2	3	4	5
点赞数	2097	1762	821	699	242
反对数	0	5	0	0	0

通过 R 软件检索得到排名前五的热点问题表：

表 3：热点问题表

热度排名	热度 ID	热度指数	留言时间	地点人群	问题描述
1	1	6.5	2019-8-19 11:34	A 市 A5 区汇金路五矿万境 K9 县	小区存在安全隐患，物业管理工作漏洞百出。
2	2	6.3	2019/4/11 21:02	A 市金毛湾	小区居民孩子上学问题。
3	3	4.5	2019/2/21/ 18:45-2019 /3/1/22:12	A 市 A4 区 58 车贷案	A4 区 p2 公司 58 车贷案，受害者和市民高度关注后续。
4	4	3.8	2019/9/5/1 3:06:55	A4 区绿地海外滩小区	小区距长赣高铁不到 400 米，电磁辐射不符合防护规定。
5	5	1.9	2019/9/16/ 23:28:27	A 市富绿物业丽发新城	强行断业主家饮用水。

通过编写 R 代码，得到热点问题留言明细表为：

表 4：热点问题留言明细表

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
208636	A00077171	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019-8-19 11:34	我是 A 市 A5 区汇金路五矿万境 K9 县 24 栋的一名业主...	0	2097
223297	A00087522	反映 A 市金毛湾配套入学的问题	2019-4-11 21:02	书记先生：您好！我是梅溪湖金毛湾的一名业主...	5	1762
220711	A00031682	请书记关注 A 市 A4 区 58 车贷案	2019-2-21 18:45	尊敬的胡书记：您好!A4 区 p2p 公司 58 车贷，...	0	821
217032	A00056543	严惩 A 市 58 车贷特大集资诈骗案保护伞	2019-2-25 9:58	胡市长：您好！西地省展星投资有限公司设立 58 车贷...	0	790

3.问题三求解

3.1 方法分析

我们根基附件 4 给出的数据，通过留言主题、留言详情、答复意见进行相似度分析来推测其的相关性；根据答复内容结构是否具有问候，问题针对性分析，和致谢三大要素来判断答复意见完整性；根据答复意见是否针对留言详情进行一一解答来反映答复意见可解释性，

综合三种特性来评价答复意见的质量。

3.2 求解过程

本题文本问题挖掘应用，主要是建立合理模型挖掘答复意见与留言主题、留言详情，相关性大小、完整性百分比、可解释性高低三者的综合评价来评价答复意见的质量。

（1）相关性

我们可以利用文本的相似度来分析其相关性，通过答复意见与留言主题、留言详情两者相似度来分析相关性。利用 python 进行数据处理和结果求解。

由结果显示的数据可知，因为其相似度低，我们通过逆向思维可以判断答复意见与留言主题、留言详情两者具有强相关性。

(2) 完整性

通过答复内容结构是否具有问候，问题针对性分析，和致谢三大要素来判断答复意见完整性。

根据文本中答复内容结构某些一定具有的词出现的词频数的次数来判断答复意见完整性。

(3) 可解释性

通过答复意见与留言主题、留言详情中针对问题，和解决问题是否两两相对应。

参考文献

[1] 基于统计相关性与 K-means 的区分基因子集选择算法[J].

谢娟英,高红超. 软件学报. 2014(09)

[2] 基于特征子集区分度与支持向量机的特征选择算法[J].

谢娟英,谢维信. 计算机学报. 2014(08)

[3] 基于改进的 F-score 与支持向量机的特征选择方法[J].

谢娟英,王春霞,蒋帅,张琰. 计算机应用. 2010(04)

[4] 一种基于 F-Score 的特征选择方法[J].

秦彩杰,管强. 宜宾学院学报. 2018(06)

[5] 一种融合蚁群算法和随机森林的特征选择方法[J].

李光华,李俊清,张亮,辛衍森,邓华伟. 计算机科学. 2019(S2)

[6] 采用机器学习的聚类模型特征选择方法比较[J].

赵玮. 华侨大学学报(自然科学版). 2017(01)