“智慧政务”中的文本挖掘应用

摘要

近年来，随着微信、微博、阳光热线等网络问政平台的出现，网络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，自然语言处理技术已经是社会治理创新发展的新趋势，对提高政府的工作效率具有极大的推动作用.从网络平台获取的群众问政留言记录数目众多，政府工作人员工作量大，在此本文将利用自然语言处理和文本挖掘的方式解决一系列问题.

对于问题 1，首先将附件 2 当中的留言主题提取，调用 Python 中的 jieba 等库通过自然语言把留言主题进行中文分词，将分词所得到的词储存.然后将所储存的词中所包含的停词，比如“了”、“的”、“吧”等词语删除，并将所剩下的有效词语进行向量化（数值化）处理.其次利用朴素贝叶斯分类的机器学习方法把训练集当中的分离特征集合目标集进行分类处理.最后将所得结果通过 F-store 进行分类评价.

对于问题 2，首先使用 HMM 完成词汇粗切分，在粗切分的基础上再建立一层 HMM，同时依据地名特征建立角色表，并将该角色表作为状态变量取值空间，将粗切分词汇作为观测序列，再次应用 HMM，从而完成对地名命名实体的识别.然后通过 k 均值聚类算法将得到的地名和人群名，进行聚类分类.最后将分类之后，按出现次数进行排序，并通过轮廓系数评价方法对该模型进行评价.

对于问题 3，先将相关文本同问题 1 的方法进行中文分词并去停词，然后将所得到的关键词进行相似度对比，从而得到政府相关部门对留言回复的准确性与完整性等相关评价指标，本文在建立用户相关性指标模型时，充分参考了萨拉塞维克提出的相关性评价指标，在此将相关性、完整性、可解释性和效率性 4 个因素作为评价指标

关键词：朴素贝叶斯分类　中文分词　HMM　粗切分　k 均值聚类算法

目录

1. 挖掘背景

近年来，随着微信、微博、阳光热线等网络问政平台的出现，网络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.伴随着大数据，人工智能、云计算等技术的出现，自然语言处理技术已经是社会治理创新发展的新趋势，对提高政府的工作效率具有极大的推动作用.

从网络平台获取的群众问政留言记录数目众多，政府工作人员工作量大，在此本文将利用自然语言处理和文本挖掘的方式解决一系列问题.

2. 挖掘目标

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.在此，本文利用自然语言处理和文本挖掘方法代替人工，将各种各样的留言进行分类，并通过 F-Score 对分类模型进行评价，得出最佳的标签分类模型.并将同一时间段，特定地点的热点问题进行摘取与热度排行，以便与发放到相应的政府部门，使得政府部门及时回复群众的留言问题.最后将建立综合评价模型对附件 4 的政府回应，对政府的工作效率做出相应的评价.

3. 分析方法与过程

3.1. 问题 1 分析方法与过程

3.1.1. 数据预处理

利用 pandas 包中的函数将附件 2 中的数据读入程序，利用 jieba_cut(text)方法对整个数据表中的自然语言进行分词.通过使用 numpy_iter(arrary)方法将所分的每一个词汇存入数组，平却返回一个数组.利用 bayes()方法去掉所返回数组中每一个分词的空格和空行.

3.1.2. 分离特征集和目标集

使用 feature 数组存放附件 2 中留言主题，留言时间，留言详情三列信息，作为特征集，使用 target 数组存放附件 2 中一级分类列的信息，作为目标集.随后，将数据进行批量分词，将每个留言主题分别对应留言详情中的每个分词，存放到 feature 中.

3.1.3. 分割数据集和目标

利用 sklearn 之 train_test_split()函数将矩阵随机划分为训练子集和测试子集，并返回划分好的训练集测试集样本和训练集测试集标签.train_data：被划分的样本特征集.train_target：被划分的样本标签.test_size：如果是浮点数，在 0-1 之间，表示样本占比.

3.1.4. 对数据集进行抽取

使用 sklearn-TfidfVectorizer( )函数将该类会统计每个词语的 tf-idf 权值.即把词语变成数值方面后面统计.随后利用 fit_transform( )函数对附件 2 中的留言主题和留言详情中的词语进行重要统计.并分别存入训练集和测试集.

3.1.5. 朴素贝叶斯分类

分类是数据分析和机器学习领域的一个基本问题.文本分类已广泛应用于网络信息过滤、信息检索和信息推荐等多个方面.朴素贝叶斯分类算法是学习效率和分类效果较好的分类器.直观的文本分类算法，也是最简单的贝叶斯分类器，具有很好的可解释性，朴素贝叶斯算法特点是假设所有特征的出现相互独立互不影响，每一特征同等重要.但事实上这个假设在现实世界中并不成立：首先，相邻的两个词之间的必然联系，不能独立；所以需要采用合适的方法进行特征选择，这样朴素贝叶斯分类器才能达到更高的分类效率.

朴素贝叶斯分类(NBC)是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y .

运用本模型将附件 2 当中的留言进行一级标签分类.

设有样本数据集 $D=\{d_1,d_2,\cdots,d_n\}$ 、对应样本数据的特征属性集为 $\mathrm{X}=\{x_1,x_2,\cdots,x_d\}$ 、类变量为 $Y=\{y_1,y_2,\cdots,y_m\}$ ，即 D 可以分为 y_m 类别.其中 $x_1,x_2,\cdots,x_d$ 相互独立且随机，则 Y 的先验概率 $P_{prior}=P(Y)$ ，Y 的后验概率 $P_{post}=P(Y|X)$ ，由朴素贝叶斯算法可得，后验概率可以由先验概率 $P_{prior}=P(Y)$ ，证据 $P(X)$ 、类条件概率 $P_{post}=P(X|Y)$ 、计算出：

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \qquad 【1】$$

朴素贝叶斯基于各特征之间相互独立，在给定类别为 y 的情况下，上式可以进一步表示为下式：

$$P(X|Y=y)=\prod_{i=1}^{d}P(x_i|Y=y) \qquad 【2】$$

由以上两式可以计算出后验概率为：

$$P_{post}=P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(x_i|Y)}{P(X)} \quad 【3】$$

由于$P(X)$的大小是固定不变的，因此在比较后验概率时，只比较上式的分子部分即可.因此可以得到一个样本数据属于类别y_i的朴素贝叶斯计算如下图所示：

$$P(y_i|x_1,x_2,\cdots,x_d)=\frac{P(y_i)\prod_{j=1}^{d}P(x_j|y_i)}{\prod_{i=1}^{d}P(x_j)} \quad 【4】$$

利用 MultinomialNB(alpha=1.0)，mlt.fit(x_train,y_train)进行朴素贝叶斯分类.预测出特征集和目标集的准确率.

首先我们可以计算准确率(accuracy),其定义是:对于给定的测试数据集，分类器正确分类的样本数与总样本数之比.也就是损失函数是 0-1 损失时测试数据集上的准确率.需要先需要定义 TP,FN,FP,TN 四种分类情况，具体情况如下表:

	Relevant	NonRelevant
Retrieved	true positives(TP 正类判定为正类)	false positives(FP 负类判定为正类)
NotRetrieved	false negatives(FN 正类判定为负类)	true negatives(TN 负类判定为负类)

表 1-1

精确率(precision)的公式是$P=\frac{TP}{TP+FP}$,它计算的是所有被检索到的留言 (TP+FP) 中,"应该被检索到的留言（TP）"占的比例.召回率(recall)的公式是$R=\frac{TP}{TP+FN}$,它计算的是所有检索到的留言（TP）占所有"应该被检索到的留言（TP+FN）"的比例.

记$\frac{2}{F}=\frac{1}{P}+\frac{1}{R}$,$F_1$定义为$P$和$R$的调和平均数.可得:

$$F_1=\frac{2PR}{P+R}=\frac{2TP}{2TP+FP+FN} \quad 【5】$$

计算出特征集和目标集的查准率和查全率.运行结果如下:

```
f1_micro: 0.7663916630481981
f1_macro: 0.7192977423576735
```

由于“检索策略”并不完美，希望更多相关的文档被检索到时，放宽“检索策略”时，往往也会伴随出现一些不相关的结果，从而使准确率受到影响.分类程序和 F-Store 的程序全部放在程序 1.

3.2. 问题 2 分析方法与过程

3.2.1. 文本地名实体识别

首先使用 HMM 完成词汇粗切分，在粗切分的基础上再建立一层 HMM，同时依据地名特征建立角色表，并将该角色表作为状态变量取值空间，将粗切分词汇作为观测序列，再次应用 HMM，从而完成对地名命名实体的识别.

HMM 是一种著名的有向图模型，主要用于顺序数据的建模.在词汇粗切分任务中，新闻文本中各单字构词情况的求解问题可抽象为已知观测变量序列（分句），需要求取一个与该观测序列最佳匹配的状态变量序列（单字状态）的概率计算问题，即:

$$arg\max\ P(C|B) \qquad 【6】$$

式中，$C=\{y_1,y_2,\cdots,y_n\}$为各个单字状态变量序列$B=\{x_1,x_2,\cdots,x_n\}$为留言文本字串，$arg\max$表示求使得$P(C|B)$取到最大值的序列C.由贝叶斯公式可得:

$$arg\max P(C|B)=arg\max\frac{P(B|C)P(C)}{P(B)} \qquad 【7】$$

依据标点符号可将留言全文切分为多个文本序列，针对每一个文本序列有$P(B)$一致，故可得:

$$arg\max\frac{P(B|C)P(C)}{P(B)}\Longleftrightarrow arg\max P(B|C)P(C) \qquad 【8】$$

又由马尔科夫链的定义可得:

$$P(B|C)P(C)=P(y_i)\prod_{i=1}^{n}P(x_i|y_i)\prod_{i=2}^{n}P(y_i|y_{i-1}) \qquad 【9】$$

式中，$P(x_i|y_i)$为状态观测概率，即单字x_i为状态y_i的概率；$P(y_i|y_{i-1})$为状态转移概率，即由上一单字状态得到当前单字状态的概率.这些参数均可由语料库统计得到.

在文本粗切分中，x_i代表观测变量，y_i代表状态变量，则有观测序列$\{x_1,x_2,\cdots,x_n\}$（字符串）和状态序列$\{y_1,y_2,\cdots,y_n\}$($y_i\in$｛词首，词间，词尾｝)，故【9】式依据各状态链的联合概率可表示为

$$P(y_1,y_2,\cdots,y_n|x_1,x_2,\cdots,x_n)=P(y_1)P(x_1|y_1)\prod_{i=2}^{n}P(y_i|y_{i-1})P(x_i|y_i) \qquad 【10】$$

联立式【6】和式【10】可求解粗切分的结果.将 HMM 进行第 2 次应用完成地名角色标注，即将观测变量变更为粗切分后的字词，状态变量变更为地名角色集合（如地名的首间尾以及后缀等），其余原理和方法与粗切分时相同.

3.2.2. K-means 聚类算法聚类

对于给定的一个包含n的d维数据点的数据集$X=\{x_1,x_2,\cdots,x_n\}$，其中$x_i\in R^d$，以及要生成的数据子集的数目K，K-Means 聚类算法将数据对象组织为K个划分$C=\{c_i,i=1,2,\cdots K\}$.每个划分代表一个类c_k，每个类c_k有一个类别中心μ_i.选取欧式距离作为相似性和距离判断准则，计算该类内各点到聚类中心μ_i的距离平方和

$$J(c_k)=\sum_{x_i\in C_k}\|x_i-\mu_k\|^2 \qquad 【11】$$

聚类目标是使各类总的平方和$J(C)=\sum_{K=1}^{K}J(c_k)$最小.

$$J(C)=\sum_{k=1}^{K}J(c_k)=\sum_{k=1}^{K}\sum_{x_i\in C_k}\|x_i-\mu_k\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}\|x_i-\mu_k\|^2 \qquad 【12】$$

其中，$d_{ki}=\begin{cases}1,若x_i\in c_i\\0,若x_i\in c_i\end{cases}$，显然根据最小二乘法和拉格朗日原理，聚类中心$\mu_k$应该取为类别$c_k$类各数据点的平均值.

K-means 聚类算法从一个初始的 K 类别划分开始，然后将各数据点指派到各个类别中，以减小总的距离平方和.因为 K-means 聚类算法中总的平方和随着类别个数 K 的增加而趋向于减小(当$K=n$时，$J(C)=0$).因此，总的距离平方和只能在某个确定的类别个数K下，取得最小值.

3.2.3. K-means 算法的流程

该算法是一个反复迭代过程，目的是使聚类域中所有的样品到聚类中心距离的平方和$J(C)$最小，算法流程图包括 4 个步骤具体流程图如下所示：

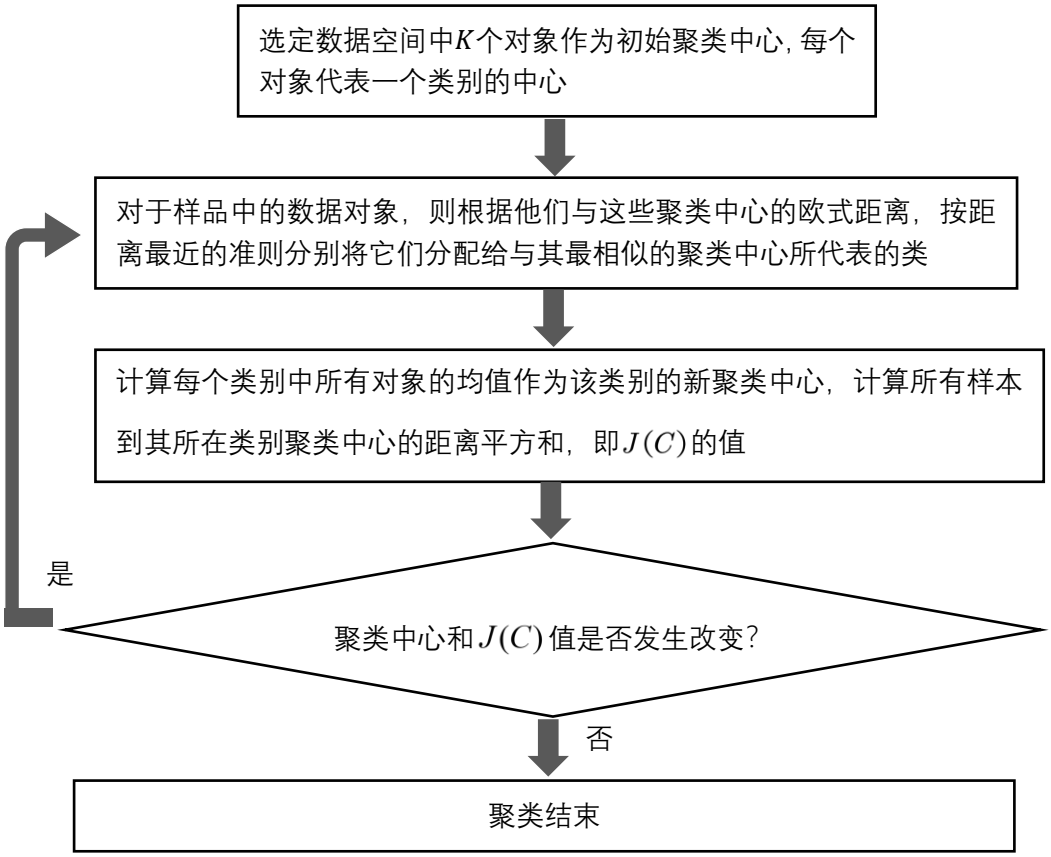

图 2-1

3.2.4. 模型的评价

3.2.4.1. 凝聚度和分离度分析

在对数据集进行聚类时一个依据的标准就是样本之间的 “距离”， 将“ 距离”较近的样本划分到同一簇中， 而“距离”较远的样本尽量划分到不同簇中.样本之间的“距离”实际上 就是对于样本邻近的度量，根据不同的邻近性度量方式会划 分出不同形状的簇.两个样本之间的邻近度是两个样本对应属性之间的邻近度的函数.包含两类邻近性度量方式，分别是基于距离的(或称为相异度)和基于相似度的.两个样本越类似.

在文本聚类应用中， 文档向量的邻近性度量采用余弦相 似度. 假设某一文档特征向量表示为$T=(T_1,T_2,\cdots,T_n)$， 另 一 文档的特征向量表示为$T'=(T_1',T_2',\cdots,T_n')$则两个文 档之间的相似度就可以利用$T$ 和T'的向量之间夹角的余弦值来表示:

$$C(T,T')=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}T_i\times T_j'}{\sqrt{\sum_{i=1}^{n}T_i^2\sum_{j=1}^{m}T_j'^2}} \qquad 【13】$$

簇的凝聚度和分离度并不是独立的，两者之和为一个常数，等于总平方和，即每个样本

到总均值的距离的平方和，由此结论可以推断最小化凝聚度等价于最大化分离度.当前提出的有效性函数大多是基于凝聚度和分离度的组合及其加权改进.XB 函数定义如下：

$$XB=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u^{2}{}_{ij}\|x_j-v_i\|^2}{n\left(\min\limits_{h\neq k}\|v_h-v_k\|^2\right)} \quad 【14】$$

其中：分子通过簇内样本距离求平方和表示凝聚度，分母计算簇问最小距离表示分离度，聚类结果越好，则 XB 取值越小.当样本x_j属于以为质心的簇时，令μ为 1，否则为 0.

3.2.4.2. 样本个体的轮廓系数

针对数据集里样本d_i,假设样本d_i被聚类到簇A,其轮廓系数s_i定义如下

$$s_i=\frac{b_i-a_i}{\max(a_i,b_i)} \quad 【15】$$

其中：a_i＝样本d_i与与其同簇其他样本的平均聚类;对于其他非簇A的簇C而言，则令：$D(i,C)$＝样本d_i与聚C众所有样本的平均距离，则$b_i=\min\limits_{C\neq A}\{D(i,c)\}$，假设样本与簇$B$中所有样本的平均距离取得该最小值$b_i$

3.2.4.3. 聚类的轮廓系数：

对于数据集的某次聚类而言，其轮廓系数定义如下：

$$S_k=\frac{\sum_{i=1}^{n}S_i}{n} \quad 【16】$$

其中：n为数据集中样本个数，k为聚类数.也可称为平均轮廓系数，可以采用S_k进行聚类有效性分析.

3.2.5. 排序及输出

该算法的程序和运行结果分别保存在程序 2、热点问题表.xls 和热点问题留言明细表.xls 中，其中结果部分截图如下图所示：

	A	B	C	D	E	F	G	H	I	J	K	L
1	热度排名	问题ID	热度指数	时间范围	地点	问题描述						
2	2	1	1997			A3区一米阳光婚纱艺术摄影是否合法纳税了？						
3	4	3	418			A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时						
4	3	2	339			关于拆除聚美龙楚在西地省商学院宿舍旁安装变压器的请求						
5	1	0	298			A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民						
6	5	4	10			请A市加快自来水深度净化改造力度						
7												
8												

3.3. 问题 3 分析方法与过程

本文的研究对象为社会化问答服务，根据网络信息服务的特殊性，将评价指标作出适 当的调整.首先，用户对于最佳答案的判定标准是主观且自由的，不存在政府对留言问题问题做出不正确的判定，因此对信度这一指标不做分析信息内容是否能解决用户实际处境下的问题，即效用指标.研究将相关性评价指标分为的相关性、完整性、可解释性和效率性 4 个指标，具体指标见下图

$$\text{指标}\begin{cases}\text{相关性}\\ \text{完整性}\\ \text{可解释性}\\ \text{效率性}\end{cases}$$

3.3.1. 时间的提取

效率性：将附件 4 当中每条留言的留言时间与答复时间提取，保存到附件 5.txt，通过C++编程语言，提取附件 5.txt 中的数据，记t_i为第i条留言的留言时间，记t_i'为对第i条留言的回复时间，则有

$$T=\frac{\sum_{i=1}^{n}(t_i'-t_i)}{n} \qquad 【17】$$

T 为政府的平均留言效率.程序存放到程序 3.cpp.

3.3.2. 相关性与可解释性的对比

对于社会化问答服务而言，用户提出问题就是希望可以通过他人的经验或知识来解决自己遇到的问题，因此是否可用对提问者而言意义重大.实际情况中，提问者往往会从两个方面来考察答案的效用，一种是尝试过答案提供的方法后确实解决了实际问题，另一种是在自己实 际所处的环境下，答案中提出的解决方案最可行 .

相关性指标的整体分布可以反映出留言回复中用户对留言回复的偏好与态度.通过留言回复与用户留言的相似度对比，可以得到政府相关部门的工作的准确性与解释性.

4. 结果与评价

优点：本文用到的算法与模型，原理比较简单，实现也是很容易，收敛速度快.有稳定的分类效率.对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练.对缺失数据不太敏感.

缺点：K 值需要预先给定，很多情况下 K 值的估计是非常困难的.K-Means 算法对初始选取的质心点是敏感的，不同的随机种子点得到的聚类结果完全不同，对结果影响很大.对噪音和异常点比较的敏感.朴素贝叶斯模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好.而在属性相关性较小时，朴素贝叶斯性能最为良好.对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进.

5. 参考文献

[1] 刘晨晖. 中文文章与主题关键短语提取方法研究[D]. 西安理工大学, 2019.

[2] 余本功,张宏梅,曹雨蒙.基于多元特征加权改进的 TextRank 关键词提取方法[J].数字图书馆论坛,2020(03):41-50.

[3] 李志强,潘苏含,戴娟,胡佳佳.一种改进的 TextRank 关键词提取算法[J/OL].计算机技术与发展,2020(03):1-5[2020-05-06].

[4] 刘竹辰.基于层次主题模型的网络热点分析研究与实现[D].北京邮电大学,2019.

[5] 阮光册,夏磊.基于主题模型的检索结果聚类应用研究[J].情报杂志,2017(3):179-184.

[6] 朱连江,马炳先,赵学泉.基于轮廓系数的聚类有效性分析[J].计算机应用,2010(12):139-141,198.

[7] 王千,王成,冯振元,等.K-means 聚类算法研究综述[J].电子设计工程,2012(7):21-24.

[8] 杨莉,万常选,雷刚,等.基于特征词权重的文本分类[J].计算机与现代化,2012(10):8-13.