“智慧政务”中的文本挖掘应用

摘要

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。本文将基于数据挖掘技术对群众问政留言记录的数据进行挖掘与分析。

对于问题一，我们通过调用百度 AI 开放平台对群众问政留言记录进行分词，根据分词结果进行词频统计，对文本特征进行向量化处理，并通过 TF-IDF 算法进行词特征值修订，最后根据每段文本的 TF-IDF 特征向量数据，建立分类模型。

对于问题二，我们利用 Excel 对群众问政留言记录进行归类，并以相似主题留言出现的频率为热度评价指标，筛选出排名前五的热点问题，得出“热点问题表.xls”，并对应得出“热点问题留言明细表”。

对于问题三，我们通过对附件 4 中相关部门对留言的答复意见进行数据处理，并对其进行分词，筛选出答复中的要点词语，将其与群众问政记录进行对应，进而对答复的相关性、完整性、可解释性进行评价，以进一步完善基于自然语言处理技术的智慧政务系统，推进群众问政与政府执政管理水平的提升。

关键词：数据预处理 中文分词 词频统计 TF-IDF 算法 群众问政结果

Application of text mining
in "intelligent government affairs"

abstract

In recent years, with 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. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. This paper will mine and analyze the data of people's political message records based on the data mining technology.

As for question one, we use baidu AI open platform to segment the message records of people's political questions, conduct word frequency statistics according to the segmentation results, conduct vector-quantization processing of the text features, and revise the word feature values through tf-idf algorithm. Finally, a classification model is established based on the tf-idf feature vector data of each paragraph of text.

For question two, we used Excel to categorize the message records of people asking about politics, and took the frequency of similar topic messages as the heat evaluation index, screened the top five hot issues, and obtained the "Hot issues table.xls", and the corresponding "Hot issues message list.xls".

For question three, we through to the appendix 4 related departments to leave a message in response to data processing, and word segmentation, the key points out a reply words, its corresponding with the crowds asked government records, and then to reply the relevance, integrity, interpretability of evaluation, based on natural language processing technology to further improve the wisdom of the e-government system, promote the crowd asked politics and administration management level of ascension.

Keywords: data preprocessing Chinese words segmentation word frequency statistics

TF-IDF weighted mass participation

目录

一、挖掘目标

本次建模针对收集自互联网公开来源的群众问政留言记录数据及相关部门对应的答复意见数据，对文本进行基本的数据预处理、中文分词后，利用 TF-IDF 算法，通过建立文本分类模型，实现对群众问政内容的热点挖掘及部门答复的质量评价，以对智慧政务系统的实行提供切实有效的建议，促进政府管理水平的提升。

二、数据分析

（一）总体流程

图 1 总体流程图

本论文的挖掘流程大致可以分为如下步骤：

第一步：对所提供的原始数据进行预处理、中文分词等操作；

第二步：对处理过的群众问政记录进行数据分析，在本文中主要是采用 TF-IDF 算法进行特征处理，并以此建立一级标签分类模型；

第三步：根据分类模型筛选热点问题；

第四步：对相关部门的答复意见进行数据处理、中文分词等操作，将提取出的关键词与群众问政记录进行匹配，统计匹配情况。

（二）分析过程

1.问题一分析

（1）数据去重处理

在附件所给的数据中，出现了很多重复的数据和前后文匹配程度不高的数据，考虑到这些数据会影响分词结果，进而影响到最终建立的模型，因此我们在进行中文分词之前分别使用了 python 和 Excel 工具对数据进行了筛选和去重。

（2）对群众问政记录进行中文分词

文本数据与数字数据在建模过程存在很大的差别，由于计算机无法识别连串的中文语句，因此在正式建模之前，我们需要将文本数据进行转换，为了简便转换过程，我们对群众问政记录进行中文分词。在本文中，我们使用 python 语言调用百度 AI 开放平台接口进行分词。

（3） TF-IDF 算法向量化处理

即便对群众问政记录进行了分词处理，计算机仍无法对其进行机器挖掘，因此我们还需要对这些词组进行向量化，在本文中，采用 TF-IDF 算法进行分析。TF-IDF 算法的主要思想是：如果某个单词在一篇文章中出现的 TF（频率）高，且在语料库中其他文档很少出现——IDF（逆文档频率）高，则认为这个单词有很好的类别区分能力，在本文中，我们使用 TF-IDF 算法来建立一级标签分类模型，建立具有强区分能力的一级标签。具体原理如下：

①词频计算：

$$TF = \frac{\text{某个词在文档中出现的次数}}{\text{文档的总词数}}$$

②逆文档频率计算：

$$IDF = \log \frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}$$

③TF-IDF 计算：

$$TF - IDF = TF \times IDF$$

TF-IDF 值越大，表明该一级标签的重要程度越高，因此，在计算出文档中每个词组的 TF-IDF 值，并且将其降序排列，取排在最前面的几个词作为进行输出。在本论文中，我们调用了 python 软件中的 TD-IDF 程序包，得到包含关键词频次信息的特征矩阵。

（4）留言内容一级标签分类模型

在本文中，我们采用多种算法（naive bayes classifier、 KNN Classifier、Logistic Regression Classifier、Random Forest classifier、Decision Tree Classifier、Gradient Boosting Decision Tree、SVM Classifier）对群众留言记录进行一级标签分类，我们将这些算法应用于相同的数据集，并对每种分类结果进行评价，进而得到最优的算法及分类结果。

① naive bayes classifier（多项式贝叶斯算法）

朴素贝叶斯分类器是一组基于贝叶斯定理的监督学习算法，它假设：给定类别变量的每一对特征之间条件独立，即根据朴素条件独立假设：

$$P\left(x_i \middle| y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\right) = P\left(x_i \middle| y\right)$$

对 i 进行遍历，上式可以化为：

$$P\left(y \middle| x_1, ..., x_n\right) = \frac{P(y)\prod_{i=1}^{n} P\left(x_i \middle| y\right)}{P\left(x_1, ..., x_n\right)}$$

由于输入的 $P\left(x_1, ..., x_n\right)$ 是给定的常数值，我们可以得到以下式子：

$$P\left(y \middle| x_1, ..., x_n\right) \propto P(y)\prod_{i=1}^{n} P\left(x_i \middle| y\right)$$

$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y)$

我们随后可以极大化后验估计，来估计 $P(y)$和$P(x_i|y)$，而前者既是训练集中类别 y 的相对频率。

在此理论基础下，多项式贝叶斯算法即实现了多项式分布数据的朴素贝叶斯算法，在这个分布中，每个类别 y 的特征向量 $\theta_y = (\theta_{y1},...,\theta_{yn})$，其中 n 是词典大小，$\theta_{yi}$ 是特征 i 在类别 y 的一个样本中出现的频率 $P(x_i|y)$。

② KNN Classifier

KNN 算法的核心思想是：如果一个一个样本在特征空间中的 K 个最相邻的样本中的大多数都属于某一个类别，则该样本也属于这个类别，并且拥有这个类别上样本的特性。算法步骤如下：

1’ 计算已知词典中数据与当前数据直接按的距离，并他们按照距离递增次序排序；

2’ 选取与当前点距离最小的 K 个点；

3’ 确定前 K 个电所在类别的出现频率；

4’ 返回前 K 个电出现频率最高的类别作为当前点的预测类别。

③Logistic Regression Classifier（逻辑回归分类器）

逻辑回归分类的主要思想就是用最大似然概率方法构建出方程，为最大化方程，并利用牛顿梯度上升求解方程参数。实际上就是对一组权值 $\omega_0, \omega_1,...,\omega_m$ 按照与测试数据线性加和的方式，求出一个 Z 值 $Z = \omega_0 + \omega_1 * x_1 + ... + \omega_m * x_m$，随后按照 sigmoid 函数的形式求出：$\sigma(Z) = \frac{1}{1+\exp(Z)}$，此时将 sigmoid 函数看作样本数据的概率密度函数，对每一个样本点，我们都可以计算出它的概率密度，最终便能求出可能性的大小。

④Random Forest classifier（随机森林分类）

随机森林即是通过集成学习的思想将多棵树集成的一种算法，它的基本单元使决策树，本质属于集成学习。从直观角度解释，每一棵决策树都是一个分类器，则对于每一个输入样本，N 棵树具有 N 个分类结果，随机森林集成了所有的分类结果，出现次数最多的类别则为最终的输出。

⑤Decision Tree Classifier（分类决策树）

决策树的核心思想是：相似的输入必会产生相似的输出，分类步骤如下：

1’从训练样本矩阵中选择第一个特征进行子表的划分，使每个子表中该特征的值完全相同；

2’再每个子表中选择下一个特征按照同样的规则划分更小的子表；

3’不断重复上面步骤直至所有特征全部使用完位置，此时得到所有特征值完全相同的子表；

4’对于待分类样本，根据每一个特征值选择对应的字表进行逐一匹配，用该子表中样本的输出，为待分类样本提供输出。

⑥Gradient Boosting Decision Tree（梯度提升决策树）

梯度提升决策树实际上是一种回归数，基本思想是串行地生成多个弱学习器，每个弱学习器的目标是拟合先前累加模型的损失函数的负梯度，使加上该弱学习器后的累积模型损失往负梯度的方向减少，即系统误差函数最小。

⑦SVM Classifier

对于分类任务，实际上可以看作是在特征空间中，找到了一个超平面，将不同类的样本分开，一个超平面对应一个预测函数，因此找到合适参数的预测函数实际上就是找到了合适的超平面。

SVM 是一种二类分类模型，它要求预测函数不仅完成样本分类，而且还要满足最大间隔条件，因此，它的分类标准比一般的分类多了一个条件，分类标准更加严格。

在本文中，我们使用 F-Score 对分类方法进行评价：

$F=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$ 其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。F 值越高证明该分类方法越优。

对附件二中群众留言记录进行多种分类模型尝试后，我们得到以下结果：

```
Multinomial Naive Bayes Classifier:0.7262021589793916
KNN Classifier:0.6889106967615309
Logistic Regression Classifier:0.7860647693817469
Random Forest Classifier:0.7860647693817469
Decision Tree Classifier:0.7281648675171737
Gradient Boosting Decision Tree:0.7644749754661433
SVM Classifier:0.7782139352306182
```

图 2 F-Score 计算结果

由图 2 结果显示，使用 Logistic Regression Classifier 和 Random Forest classifier 所得的 F 值最高，因此我们选取其中一种方法建立一级标签分类模型便能得到很好的分类效果。

2.问题二分析

具体流程

①将群众留言记录按照一级标签分为七大类，按照七大类别对数据分别进行关键词筛选；

②对各大类别根据不同时间段（由远及近）、不同点赞数（由大到小）进行排序；

③根据排序结果结合各个一级标签下的关键词得出热度前五的热点问题，得到“热点问题表.xls”；

④根据“热点问题表.xls”结果，利用 excel 的筛选工具进行分类，得到“热点问题留言明细表.xls”。

针对大量数据，为了更直观地展示和快速感知最突出的文字，我们根据问题一数据预处理后得到的 AllWord.xlsx 中各一级标签类对应的留言主题中的分词频率进行了词云的绘制得到如下结果：

图 3　‘交通运输’类词云

图 4　‘城市建设’类词云

图 5　‘环境保护’类词云

图 6　‘教育文体’类词云

图 7　‘卫生计生’类词云

图 8　‘商贸旅游’类词云

图 9　‘劳动和社会保障’类词云

词云图是通过形成“关键词云层”或“关键词渲染”，对文本中出现频率较高的关键词作出视觉上的突出，以上图为例，词的出现频次越高，其字号就越大，我们可以根据对词云图像的观察进行对热点问题的简单提取。

由图 3 可见，在属于“交通运输”类的群众意见词云图中，醒目的“出租车”三个字以强烈的视觉冲击效果呈现在我们眼前，反映频次高居榜首的城市为“A 市”，反映的主要问题为“乱收费”，由此可见，在“交通运输”这一类，我们

可以对热点问题作出如下描述：A 市出租车乱收费现象严重。

由图 4 可见，在属于“城市建设”类的群众意见词云图中，“小区”二字视觉效果明显，反映频次高居榜首的城市为“A 市”，反映的主要问题为“房产证”，由此可见，在“城市建设”这一类，A 市小区房产证存在较为严重的问题，相关部门需要尽快处理。

图 5 可见，在属于“环境保护”类的群众意见词云图中，反映频次高居榜首的城市仍然为“A 市”，在问题关键词方面，“污染”、“污水”、“毒气”这些对居民健康造成严重影响的词语十分明显，由此可见，在“环境保护”这一类，“污水”、“毒气”的排放是一个重要的热点问题，有关部门亟需对做出有效的管理。

由图 6 可见，在属于“教育文体”类的群众意见词云图中，“教师”二字视觉效果明显，反映频次高居榜首的城市为“C 市”，主要反应的问题为“补课”，由此可见，在“教育文体”这一类，学生和家长对学校招生补课问题十分关注，教育部门应该重点解决这类问题。

由图 7 可见，在属于“卫生计生”类的群众意见词云图中，“政策”二字备受热议，“西地省”反应频次高居榜首，主要反应的问题为“二胎”、“咨询”，由此可见，在“卫生计生”这一类，群众对政策仍存在较大的疑虑，相关部门应尽快做出有针对性的回答。

由图 8 可见，在属于“商贸旅游”类的群众意见词云图中，“传销”二字争议颇多，反映频次高居榜首的城市为“A 市”，由此可见，在“商贸旅游”这一类，“传销”这一问题严重，而这一问题对于该地旅游业发展影响严重，相关部门必须尽快作出反应。

由图 9 可见，在属于“劳动和社会保障”类的群众意见词云图中，“社保”、“医保”视觉效果明显，“西地省”反应频次再次高居榜首，主要反应的人员为“职工”，由此可见，在“劳动和社会保障”这一类，各单位职工对于“社保”、“医保”十分重视，相关部门应尽快作出反应。

结合词云图与 TF-IDF 得到的高频词列表，我们筛选出各一级标签对应的特征关键词如下：

一级标签	*关键词*				
城乡建设	房产证	A7 县	违建	管理	规划
环境保护	污染	养猪场	垃圾	排放	噪音
交通运输	出租车	快递	A 市	乱收费	交通
教育文体	补课	教师	教育局	C 市	中学
商贸旅游	传销	电梯	垄断	A 市	乱收费
卫生计生	二胎	医生	医院	西地省	K 市
劳动和社会保障	社保	工资	医保	西地省	A 市

表 1 特征关键词

3.问题三分析

（1）数据去重处理

由于我们在对问题一的处理过程中对附件二数据进行了去重等数据预处理，因此在附件 4 所给的数据中，出现了多余且无法进行匹配的数据，考虑到这些数据会影响分词结果，进而影响到最终对部门答复的评价，因此我们在进行中文分词之前分别使用了 python 和 Excel 工具对数据进行了筛选和去重。

（2）对部门答复意见进行中文分词

冗长的语句加大了匹配评价的工作量，为更方便地评价部门答复意见的相关性、完整性、可解释性等方面，我们在正式评价之前，将文本数据进行中文分词，以便于对关键词进行提取。在本文中，我们仍然使用 python 语言调用百度 AI 开放平台接口进行分词。

三、结论

利用云计算、人工智能等技术处理网络问政平台的群众留言，了解民意和挖掘社会亟待处理的热点问题，对提升政府的管理水平和施政效率具有极大的推动作用，同时也是文本分析的一个课题、难题。传统的依靠人工进行留言划分和热点整理已经不能满足数据量日渐攀升的社情民意相关文本信息。本文根据 Logistic 回归和随机森林法针对群众留言建立分类模型，统计目前社会最关注的热点问题，深入分析问题答复和留言反馈的管理现状。

由分析结果可以看出，群众针对不同方面存在许多高热度的问题与建议，需要相关部门进行管理与解决。因此，相关部门答复的相关性、完整性和可解释性对群众问题的解决有着十分重要的作用，相关部门在进行答复的时候需要更加注重民意所向，这样才能提升政府再社会管理和施政方面的能力，推动智慧政务系统的发展。