“智慧政务”中的文本挖掘应用

摘 要

近年来，政府通过网络平台的方式来了解民意，并对其数据整理和处理，进而给予相对应的负责部门进行解决。但在这一过程中，我们发现相关民意的文本数据量在渐渐增多，因此造成依靠人工的方式来处理这些文本数据所需要的时间和人力物力在不断上升。 而且目前的科技在大数据、人工智能方面都有一定的发展，因此我们可以运用相关的技术去提高对这些文本数据的处理，进而减少对时间的损耗，并令相关部门尽快对事件做出处理。对群众留言的处理可分为三个问题即群众留言分类，热点问题的挖掘与答复意见和评价。

针对问题 1，即对群众流言进行分类，我们可以进行数据清洗，在进行分词和去停用词，并根据文本特征进行分类，并采用一级，二级，三级标签进行分类。

针对问题 2，即对热点问题的挖掘，需要我们将将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度计价指标，并给出评价结果，将结果进行排行并导入具体留言信息。

针对问题 3，需要针对相关部门对留言的答复，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，并尝试实现。

关键字：数据处理　热点词筛选　相似度计算　jieba

目录

问题分析及处理

3.1 问题一的分析及处理

针对问题 1，对群众留言的分类问题可能面临下面的困难：1.文本语义带来的词语交叉。2. 多分类问题带来的难度。3.数据不平衡带来的影响 4.长文本的无意义表达太多。因此可以先进行数据清洗，然后分词与对词性的选择，然后去停用词最后在利用常规方法进行分类，可直接用 excel 进行分类汇总。

数据处理首先要进行预处理，即数据清洗。大体可分为首先要进行缺失值清洗。1.确定缺失值范围，即对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略[1]，2.去除不需要的字段，这一步很简单，直接删掉即可。[2]3.进行格式内容清洗 4.逻辑错误清洗 5.非需求数据清洗有同时这可以解决长文本的无意义表达太多的问题需要重复操作。最后进行关联性验证。[3]、这一步关乎后面的进行，是相当重要的步骤。其关键在于特殊字符处理于正则表达式建立。当出现多分类问题带来的难度时需要将其转化为多个二分类。用多个标签来确定一个留言。

数据清洗之后进行分词与词性的划分，可通过 python 进行分词，并用 jieba 库的停用词表来进行词性划分与去停用词。其中当出现数据不平衡时即数据集各个类别的样本量极不均衡应一般模型都是假定数据平均而建立的，我们需要进行数据增强，通过对不同词性的权重比的改变让数据平衡。

将清洗后的数据用 excel 通过范围的大小与逻辑进行标签分类，可得出诸如城乡建设，党委政务，国土资源，环境保护等一级标签，安全生产，保险证券期贷等二级标签，安全防护，安全隐患等三级分类，标签关系为一级包涵二级，二级包涵三级。用数学方法选取最具分类信息的特证，用 excel 将文本留言进行汇总。这是数据的预处理，下面给出分词的代码：

```python
from collections import Counter
import jieba


# 创建停用词列表
def stopwordslist(filepath):
    stopwords = [line.strip() for line in open(filepath, 'r', encoding='utf-8').readlines()]
    return stopwords


# 对句子进行分词,并去停用词
def seg_sentence(sentence):
    sentence_seged = jieba.cut(sentence.strip())
    stopwords = stopwordslist('stoplist.txt')  # 这里加载停用词的路径
    outstr = ''
    for word in sentence_seged:
        if word not in stopwords:
            if word != '\t' and '\n':
                outstr += word
                outstr += " "
    return outstr
```

图 3-1 代码（上）

```
inputs = open('aaa.TXT', 'r', encoding='utf-8')   # 加载要处理的文件的路径
outputs = open('bbb.TXT', 'w', encoding='utf-8')   # 加载处理后的文件路径
#aaa.txt是待处理的文件，bbb.txt是分词后导出的文件

for line in inputs:
    line_seg = seg_sentence(line)  # 这里的返回值是字符串
    outputs.write(line_seg)
outputs.close()
inputs.close()
```

图 3-2 代码（下）

3.2 问题二的分析及处理

针对问题 2，即热点问题的挖掘可分为三个子任务。分别为子任务 1:问题识别，即如何从众多留言中识别出相似的留言 ；子任务 2:问题归类，即如何把特定地点或人群的数据归并，即把相似的留言归为同一问题；子任务 3:热度评价，即如何进行热度评价指标的定义和计算方法，对指标排名之后得出对应表。

针对子任务 1，在众多留言中找到相似的留言，即便将留言进行了三级标签的划分，留言的数量仍然巨大。解决的关键在于对文本特征的确定与归类。这个问题一般可分为三要素：1.特定的时间；2.特定的地点；3.发生问题。对于这个问题因为已经对数据进行了预处理，所以可直接对文本特征进行选择。其关键在于数据清洗的特殊字符的处理与正则表达式的构建。对于特殊字符的处理，需要将不能输入的特殊字符进行代替，可用 replace all 来处理。因为此文本的要素中时间要素有一定的格式，因此可参照 2018.08.06 这样的格式构建正则表达式来进行筛选。地点则利用正则表达式的贪婪性来处理[4]。具体文本可用查找进行匹配与导入。本问题的其难度在于表达的多样化，对于地点，人群的识别会因表达的多多样化而难于归类与识别。为解决这个问题要多次重复处理。选择好文本特征后，将文本按照三要素的表达进行表示可使用可使用 word2vec，然后进行相似度计算。然后制作语料库。首先用 dictionary 方法获取词袋（bag-of-words)，然后在词袋中用数字对所有词进行了编号。将新语料库通过 tfidfmodel 进行处理，得到 tfidf 通过 token2id 得到特征数 ，稀疏矩阵相似度，从而建立索引，得到最终相似度结果[5]将相似问题汇合于一个表中。

针对子任务 2，把特定地点或人群的数据归并，即把相似的留言归为同一问题。将数据进行归一化处理，通过对特征词在相似留言中的出现频次来确定问题的类别，通过相似问题自身文本特征词出现的频率来区分问题。可通过离差标准化，是对原始数据的线性变换，使结果值映射到[0 - 1]之间[6]。直接使用 excel 中的函数进行归一化。将相似问题通过偏差值进行分类，将相似度高的问题归为同一问题，其标签为其文本特征。其函数为 $x^* = \frac{x - \min}{\max - \min}$。可通过 jieba 来进行词频的统计[7]，通过词频可以直接完成子任务 3，计算方式为单个文本标签数目除去文本标签总数，因数据清洗中已经完成了权重的平衡可直接用 excel 表格涵数进行处理。将问题按照热点进行排行，以五个排名为一级，进行排行。

3.3 问题三的分析及处理

针对问题 3，对于问题的答复，要从问题的答复的相关性，完整性，与可解析性着手。可以先导入对相关问题的标准答复数据库，对不同问题答复进行规范。可先导入范式，再对答复进行填充，保证答复的完整性，并用 excel 中的查阅进行复查，确保完整性。同一类问题下，要有多种不同方面的回答，以保证其问题的可解释性。并将这些问题打上标签，将其与问题留言的标签重复率与差异率作为变量，用皮尔森（pearson）相关系数[8]进行相关性的计算，皮尔森相关系数是衡量线性关联性的程度，p 的一个几何解释是其代表两个变量的取值根据均值集中后构成的向量之间夹角的余弦.公式为

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

$$\rho_{X,Y}=\frac{N\sum XY-\sum X\sum Y}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}$$

http://blog.csdn.net/AlexMerer

$$\rho_{X,Y}=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})(\sum Y^2-\frac{(\sum Y)^2}{N})}}$$

具体要用 matlab 进行处理。两个连续变量(X,Y)的 pearson 相关性系数(Px,y)等于它们之间的协方差 cov(X,Y)除以它们各自标准差的乘积(σX,σY)。系数的取值总是在-1.0 到 1.0 之间，接近 0 的变量被成为无相关性，接近 1 或者-1 被称为具有强相关性。[9] 可能会面临因数据过大而导致相关性计算过于繁多的问题。其指标按照线性指标 0.8-1.0 极强相关，0.6-0.8 强相关，0.4-0.6 中等程度相关， 0.2-0.4 弱相关， 0.0-0.2 极弱相关或无相关。
完成后，将相关性交大的答复导入 excel 中，在实际处理时进行匹配与答复。

总结

政府通过网络平台的方式来了解民意，并对其数据整理和处理，进而给予相对应的负责部门进行解决，对于群众留言的处理，需要进行分类，热点问题挖掘，与答复匹配方法的说明。对于分类是采取先进行预处理，在通过逻辑，相关性来为其分类。挖掘的计算着重于对以预处理的数据中标签的频次。答复则利用了皮尔森系数并以重复率与差异概率为变量。通过上述方法完成对群众留言问题的解决。

参考文献

【1】csdn 博主相国大人原创文章
【2】csdn 同上
【3】csdn 同上
【4】哔哩哔哩 up 酒米家的猫儿
【5】csdn 博主番番要吃肉原创文章
【6】csdn 博主 haoji007 原创文章
【7】百度百科
【8】csdn 博主 ruthy-wei 原创文章
【9】csdn 同上