“智慧政务”中的文本挖掘应用

摘要

随着越来越多的网络问政平台的涌现，政府和群众之间也因此增加了不同的交流方式，但同样也给政府部门对于留言的划分和热点整理的工作增加了很大的难度，因此利用网络文本分析和数据挖掘技术对于解放人工劳动力、提升政府的管理水平和施政效率具有极大的推动作用。

本队伍做 C 题的过程中，首先对数据进行基本的预处理工作，包括数据清洗、文本去重、去停用词、分词、情感分析等步骤，为后面问题的处理做好铺垫。

对于问题 1 群众留言分类，此题体系完整，但是数据不全，故采用了两种特征提取的方法：根据使用词的绝对频率作为特征提取标准；使用词和类别的互信息量进行特征项抽取的判断标准。接下来采用 SVM **空间向量模型算法**进行分类。

对于问题 2 热点问题挖掘，分析问题后将其分为了三个子任务，分别是问题识别、文本归类和热度评价。将处理后的标题分词转化成词频向量，转换成 TF-IDF **权重矩阵**，特征提取，构建模型分类，取同类型的留言最早和最晚出现的时间作为时间范围。再**利用** TF-IDF **算法**计算每一个留言详情的相似度后进行留言归类整理，最后根据 Reddit **算法**结合问题二本身进行了改动后得出热度评价指标的公式。

对于问题 3 答复意见的评价，对数据进行相应处理以及词频统计，得出能够反映评价质量的**高频特征词**。随后建立评价模型，确定

相关性、完整性、可解释性对评价的**影响系数**，再通过 TOPSIS 法，得到能输入打分系统的评论得分数据。

最后，通过 python 实现代码运算，实现了对文本的挖掘与分类问题。我们对工作进行修改与完善，希望能对智慧政务的网络问政平台的操作问题起到一定的帮助。

关键词：去重　SVM 空间向量模型算法　TF-IDF 算法 自然语言处理 TOPSIS 法

目录

一、挖掘目标 4

二、爬虫技术简介 4

三、问题重述 5

四、 问题分析 6

4.1 问题一分析 6

4.2 问题二分析 6

4.3 问题三分析 7

五、问题的解决 7

5.1 问题一的解决方案 8

5.1.1 分词 8

5.1.2 特征提取 8

5.1.3 SVM 空间向量模型算法 9

5.1.4 测评方法： 11

5.2 问题二的解决方案 12

5.2.1 数据清洗 12

5.2.2 分词并利用 TF-IDF 算法进行问题识别及归类 13

5.2.3 建立热度评价指标 15

5.3 问题三的解决方案 15

5.3.1 数据预处理 15

5.3.2 建立评价模型 16

5.3.3 模型总结 17

一、挖掘目标

在当今的大数据时代里,伴随着互联网和移动互联网的高速发展，人们产生的数据总 量呈现急剧增长的趋势，当前大约每六个月互联网中产生的数据总量就会翻一番。互联网产生的海量数据中蕴含着大量的信息，已成为政府和企业的一个重要数据来源，互联网数据处理也已成为一个有重大需求的热门行业。借助网络爬虫技术，我们能够快速从互联网中获取海量的公开网页数据，对这些数据进行分析和挖掘，从中提取出有价值的信息，能帮助并指导我们进行商业决策、舆论分析、社会调查，政策制定等工作。

能够实现：对于“智慧政务”中的文本挖掘应用，按照一定的划分体系对留言进行分类，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，定义合理的热度评价指标，并给出评价结果。

二、爬虫技术简介

网络爬虫是一个自动提取网页的程序，它为搜索引擎从网页上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL 开始，获得初始网页上的 URL,在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一-定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。然后，它将根据- -定的搜索策略从队列中选择下- -步要抓取的网页URL,并重复上述过程，直到达到系统的某一条件时停止。 另外，所

有被爬虫抓取的网页将会被系统存贮，进行一一定的分析、 过滤，并建立索引，以便之后的查询和检索;对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

网络爬虫分为以下几种类型通用网络爬虫(Generar Purpose Web Crawler)、聚焦网络爬虫(Focused Web Crawler) 、增量式网络爬虫(Incremental Web Crawler)、深层网络爬虫(Deep Web Crawler).本文所设计的算法即是通用网络爬虫,针对所有留言,可以爬虫到留言的用户,留言时间，留言内容等。

三、问题重述

问题一：在处理网络问政平台的群众留言时，工作人员首先要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。请建立一个合适的文本分类系统,对留言进行分类。并使用 F-Score 对分类方法进行评价。

问题二：某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 3 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，定义合理的热度评价指标，并给出评价结果。

问题三：针对附件 4 相关部门对留言的答复意见，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。

四、问题分析

4.1 问题一分析

我们要对于留言进行分类，以便后续将群众留言分派至相应的职能部门处理。为了建立一个合适的文本分类系统，第一步进行分词，分词就是在文本连续的词条间加入分隔符，将文本从连续字符流形式转化为离散词流形式的过程。然后进行特征提取。并使用 F-Score 对分类方法进行评价。为了提高分类精度，对于每一类，我们应去除那些表现力不强的词汇，筛选出针对该类的特征项集合。特征提取，就是将文本中对表达 文本所属类别有比较强说服力的词汇从文本中抽取出来，形成一个向量。我们的系统采用了两种特征提取的方法：根据使用词的绝对频率作为特征提取标准；使用词和类别的互信息量进行特征项抽取的判断标准。接下来采用 SVM 空间向量模型算法进行分类。思路如下：根据算术平均为每类文本集生成一个代表该类的中心向量，然后在新文本来到时，确定新文本向量，计算该向量与每类中心向量间的距离(相似度)，最后判定文本属于与文本距离最近的类。

4.2 问题二分析

问题二是热点问题挖掘，题目中给出的热点问题的定义是某一时段内群众集中反映的某一问题可称为热点问题。关键词为一段时间集中爆发的问题，多人反映同一问题。同时根据附件 3 将某一时段内反应特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，得出评价结果，按格式给出排名前 5 的热点问题和具体留言信息。所以问题 2 可以分为三个子任务，子任务一是问题识别，从众多留言

中识别出相似的留言，文本相似的问题。进行俩俩留言比较计算是否是相似的。识别出后，子任务二是问题归类，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，打上相应的标签，结果对应表 2。子任务三就是热度评价，问题归类后可能会有上百个上千个标签，这个时候就需要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 1。思路如下：首先利用 TF-IDF 算法进行问题相似度的比较以及归类，随后根据 Reddit 算法，我们依据问题二的现实情况得出了热度评价指标的公式进行排序。

4.3 问题三分析

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。利用自然语言处理和文本挖掘的方法解决针对附件 4 相关部门对留言的答复意见，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。首先进行数据预处理，去除一些相同的数据，重复的回复意见等没有价值的数据。大多数文本去重是基于文本之间的相似度，这些会使得我们去除一些相近的表达，造成错删。故本文采用比较删除法，直接删除完全相同的评论，尽量保留有用的评论。然后通过答复意见文本与星级建立一个评分模型，通过该模型可以得到单条回复的打分情况，从而观察到各回复的综合评价。最后，通过 TOPSIS 法计算出能够直接输入打分模型的评论得分。我们将其依据可信度进行优化。

五、问题的解决

5.1 问题一的解决方案

文本分类是指在给定的分类体系下，根据文本的内容自动地确定文本关联的类别的过程。从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中，该映射可 以是一一映射，也可以是一对多的映射。

f:A→B,A 为待分类的文本集合，B 为分类体系中的类别集合。文本分类的映射规则是系统根据已经掌握的每类若干样本的数据信息，总结出分类的规律性而建立的判别公式和判别规则。然后在遇到新文本时，根据总结出的判别规则，确定文本相关的类别。

5.1.1 分词

分词就是在文本连续的词条间加入分隔符，将文本从连续字符流形式转化为离散词流形式的过程。具体方法如下：

本文使用 WordNet 文本中的非字母的特殊字符进行分割，包含了对单词之间连接符号的处 理，前面的方法作为一个单词来处理，这种方法是将单词分拆成多个单词，然后将得到的单词使用 WordNet 查得词性和词根，得到动词、名词和形容词，丢弃其他词性的单词，并且将得到的词根加入全局单词序列，和本文档的单词序列。而且我们丢弃了长度小于 4 个和长度大于 29 个的单词，我们认为这样的单词基本上代表的文章的信息量太少。

5.1.2 特征提取

由留言文本得到留言文本特征向量，要经历一个特征提取的过程。为了提高分类精度，对于每一类，我们应去除那些表现力不强的词汇，筛选出针对该类的特征项集合。特征提取，就是将文本中对表

达 文本所属类别有比较强说服力的词汇从文本中抽取出来，形成一个向量。通常提取方法都是统计的方法，首先利用不同的方法对特征项进行评分，然后选出分值较高的作为特征构成特征向量空间。存在多种筛选特征项的算法，常用的特征提取方法有如下所列

(1) 根据词和类别的互信息量判断

(2) 根据词熵判断

(3) 文档频率

(4) 信息增益

(5) 卡方统计

在我们的系统中采用了两种特征提取的方法：根据使用词的绝对频率作为特征提取标准；使用词和类别的互信息量进行特征项抽取的判断标准。

5.1.3 SVM 空间向量模型算法

本系统采用了 SVM 空间向量模型算法。支持向量机是建立在计算学习理论的结构风险最小化原则之上，其主要思想针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。

该方法的思路如下：根据算术平均为每类文本集生成一个代表该类的中心向量，然后在新文本来到时，确定新文本向量，计算该向量与每类中心向量间的距离(相似度)，最后判定文本属于与文本距离最近的类，具体步骤如下：

步骤一：计算每类文本集的中心向量，计算方法为所有训练文本

向量简单的算术平均;

步骤二：新文本到来后，分词，将文本表示为特征向量;

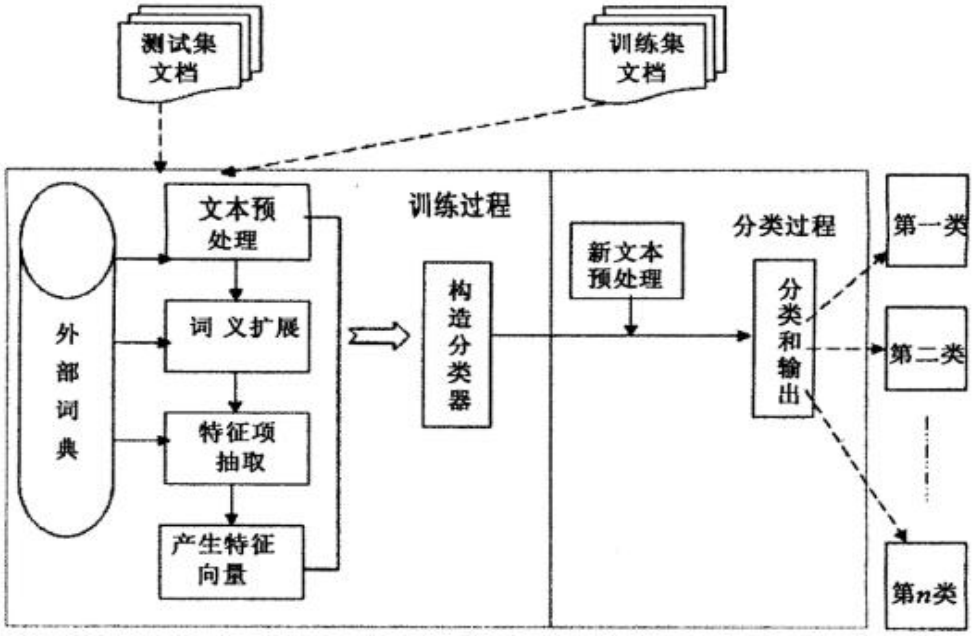
步骤三：计算新文本特征向量和每类中心向量间的相似度，公式为:

$$sim(d_i,d_j)=\frac{\sum_{k=1}^{M} w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{M} w_{ik}^2)(\sum_{k=1}^{M} w_{jk}^2)}}$$

其中，d_i,为新文本的特征向量，d_j为第 j 类的中心向量，M 为特征向量的维数，w_k为向量的第 K 维.

步骤四：比较每类中心向量与新文本的相似度，将文本分到相似度最大的那个类别中。

系统框架:



5.1.4 测评方法：

文本分类从根本上说是一个映射过程，因此评价分类系统的标志是映射的准确程度和映射的速 度。评价分类效果的标准很多，常用的评价标准是准确率和查全率，故系统也采用准确率和查全率 作为评价标准'气

准确率(Precision) p_r 是所有判断的文本中与人工分类结果吻合的文本所占的比率，表 征的是分类的正确性。其数学公式表示如下：准确率=分类正确的文本数/实际分类的文本数。

查全率(Recall) R_e 是人工分类结果应有的文本中分类系统吻合的文本所占的比率，表 征的是分类的完整性。其数学公式表示如下：查全率=分类正确的文本数/应有的文本数。

$$P_{ri}=\frac{TP_i}{TP_i+FP_i} \qquad R_{ei}=\frac{TP_i}{TP_i+FN_i}$$

这里 FP_i 指的是测试集中被错误地分到 C_i 类中的文档数。类似地有 TN_i,TP_i,FN_i，如下表：

类别	专家判断		
C_i		文档属于	文档不属
分类判断	属于	TP_i	FP_i
	不属于	FN_i	TN_i

$$F_\alpha=\frac{1}{\alpha\bullet\frac{1}{p_r}+(1-\alpha)\frac{1}{R_e}}$$

上式中，α 可被看作 P_r 相对 R_e 的重要程度。如果 $\alpha=1$，则 $F_\alpha=P_r$，如果 $\alpha=0$，则 $F_\alpha=R_e$。通常情况下，都取 $\alpha=0.5$。

5.2 问题二的解决方案

问题二总结下来是一个热点问题的挖掘以及分类，关键词是一段时间集中爆发的问题，多人反映同一问题。

分析问题二，可以将其拆分为以下三个任务，

任务 1 问题识别

如何从众多留言中识别出相似的留言，文本相似的问题，俩俩比较，计算是否是相似的。

任务 2 问题归类

把特定地点或人群的数据归并，相似的留言归类，结果对应表 2。

任务 3 热度评价

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。

因此在本问题中，针对以上的任务，我们先选用 TF-IDF 算法进行分类及检验。最后依据 Reddit 算法建立属于本文的热度评价公式，将热点问题进行排序。

5.2.1 数据清洗

数据清洗是将重复、多余的数据筛选清除，将缺失的数据补充完整，将错误的数据纠正或者删除，最后整理成为我们可以进一步加工、使用的数据。

本文中对于数据清洗分为以下几个阶段

1、预处理阶段

此阶段要做的就是将数据导入处理工具进行查看，一是看元数据，二是抽取部分数据进行人工查看，对数据本身也有一个直观的了解。

2、缺失值清洗

此阶段将对每个字端都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定比较策略。其次就是去除不需要的字段即可。

3、格式内容清洗

本文的数据都是由人工收集以及用户填写而来，很大可能在格式和内容上存在一些问题。对于时间日期数值全半角等显示格式不一致时，将其处理成一致的某种格式即可。如果内容中有不该存在的字符时，需要半自动校验半人工方式来找出可能存在的问题，并取出不需要的字符。

4、逻辑错误清洗

这部分的工作室去掉一些使用简单逻辑推理就可以直接发现问题的数据，防止分析结果走偏，在进行了格式内容清洗后，再进行去重处理，有利于得到更准确的数据。

5.2.2 分词并利用 TF-IDF 算法进行问题识别及归类

本文中主要使用的算法是 TF-IDF。TF 指的是 term frequency 词频。IDF 指的是 inverse document frequency 倒文档频率。主要思想是：如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

在本文中我们首先将主题的文本分词后，转化成词频向量，转换成 TF-IDF 权重矩阵，特征提取，构建模型分类，取同类型的留言最早和最晚出现的时间作为时间范围。

接下来将各类主题与留言详情进行相似度比较，将留言详情当作网页文档，已经分过类的留言主题当作用户查询，进行如下操作，构建热点问题留言详情表。

操作 1、文本相似度计算的需求始于搜索引擎。

搜索引擎需要计算“留言主题”和爬下来的众多”留言详情“之间的相似度，从而把最相似的留言详情归类到留言主题下，构建热点问题留言详情表。

操作 2、把每个留言详情分词，成为词包（bag of words）。统计留言的总数 M。统计第一个留言详情中的词数 N，计算第一个留言详情中第一个词在该留言中出现的次数 n，再找出该词在所有留言中出现的次数 m。则该词的 TF-IDF 为：

$$\frac{n}{N} \times \frac{1}{m/M}$$

（还有其它的归一化公式，这里是最基本最直观的公式）

操作 3、重复计算第一个留言详情中的第二个词，第三个词等等，直到计算出一个留言详情中所有词的 TF-IDF 值。其次，重复上述操作，直到计算出所有留言文档中每个词的 TF-IDF 值。

操作 4、相似度的计算

使用余弦相似度来计算处理好的留言主题和留言详情之间的夹角。夹角越小，越相似。

5.2.3 建立热度评价指标

定义一个内容的热度, 其实是一个相对的概念, 面对不同的需求, 表达也是不同的。因此考虑到在一定时间范围内，问题出现的频率, 用户对于问题的点赞数和反对数, 依据 Reddit 算法我们建立了以下热度评价公式

$$H = \frac{W + I}{(T + 1)^G}$$

其中，H 是一个问题的热度，W 是问题的点赞数—反对数，I 是问题的初始数值，即问题共出现的次数，G 是一个衰减的重力参数, T 是问题发布以来的时长。

5.3 问题三的解决方案

我们利用自然语言处理和文本挖掘的方法解决针对附件 4 相关部门对留言的答复意见，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

5.3.1 数据预处理

步骤一：文本去重

去除一些相同的数据，重复的回复意见等没有价值的数据。大多数文本去重是基于文本之间的相似度, 这些会使得我们去除一些相近的表达，造成错删。故本文采用比较删除法，直接删除完全相同的评论，尽量保留有用的评论。

（1）结巴分词

采用 jieba 分词方法对中文文本进行分词处理，是处理中文文本的一个基础性工作。

(2) 去停用词

停用词是指没有什么实际意义，对于整句的句意表达没有影响的词，通常是一些高频词汇、数字和特殊符号，如‘的，了，呢’等。我们需要在预处理阶段就将停用词进行删除操作。

(3) 词频统计 制作词云图

根据高频特征词反应出特点。

5.3.2 建立评价模型

1.答复意见评分模型

我们希望通过答复意见文本与加权系数建立一个评分模型，通过该模型可以得到单条回复的打分情况，从而观察到各回复的综合评价。为达到这个目的，我们将结合完整性、可解释性、相关性这几项数据集，先建立单条评论评分模型。

2.答复意见得分

本部分将计算出能够直接输入打分模型的评论得分。我们将其依据可信度进行优化。通过 TOPSIS 法，得到能输入打分系统的评论得分数据。

方法：Topsis 法得数据

引入变量 n 表示导入的数据集中的答复意见总条数。

步骤一：

将优化后的答复得分 a' ，导入成一个 n 行 1 列的标准化矩阵

步骤二：

计算该矩阵第一列的最小值并定义为

$$Z_{j1}^{-} = \min_{1 \le i \le n} \left| Z_{i1} \right|$$

计算该矩阵第一列的最大值并定义为

$$Z_{j1}^{-} = \max_{1 \le i \le n} \left| Z_{i1} \right|$$

步骤三:

计算第 i 个评价对象与最小值的距离

$$D_{i1}^{-} = Z_{j1} - Z_{j1}^{-}$$

计算最大值与最小值的距离

$$\mathrm{D} = Z_{j1} - Z_{j1}^{-}$$

步骤四:

计算第 i 个评价对象的最终得分

$$a_{i1} = D_{i1}^{-} / D$$

5.3.3 模型总结

这部分主要运用模型上数据的利用花了较大的篇幅，目的是为了建立一个关于回复意见评分模型，最终结果会获得一个百分制的得分。