“智慧政务”中的文本挖掘应用

摘要

随着网络信息技术的快速发展和互联网的广泛应用，人们通过微博、微信、阳光热线等网络问政平台进行留言，以此将自己和社区存在的问题发出去，从而政府能够与百姓实现信息的时事交互。如果能够精准快速的对人们的留言进行划分，以及能够及时挖掘热点信息，对政府的管理水平以及施政效率会起到很大的帮助。

针对问题一：我们首先将一级标签的类别进行汇总，在附件 2 中计算出一级标签每一个类别的概率。采用贝叶斯分类的基本思想利用留言的词组与类别的联合概率来估计给定留言的类别概率。假设未标注的留言为一级标签的某一类，类别集合为一级标签的类型。将与假设的留言条件概率最大的那个类别作为该留言的类别输出。建立模型，用 Matlab 软件实现分类结果。最后我们采用 F-Score 对我们的分类方法进行评价。

针对问题二：基于问题一的留言分类的基础上，问题二我们仍然可以采用问题一的模型和 Matlab 软件将类似的一类留言归类。这个模型可以将词组相似度最高的留言归为一类。即将不同的人在不同得时间反映的类似问题汇总在一起。之后比较每一类问题发生的概率占比，即可排出前五名热度最高的留言问题，找出该留言发生的最开始的时间和最后的时间，即可确定其时间段。

针对问题三：我们从相关性、完整性、可解释性三个角度对答复意见的质量进行评价。

关键词：Matlab；贝叶斯分类算法；一级标签；相似度；F-Score；

Application of Text Mining in "Intellige nt Government Affairs"Summary

Abstract

With the rapid development of network information technology and the wide application of the Internet, people leave messages th rough micro-blog, micro-letter, sunshine hotline and other network political platforms, in order to send out their own and community pr oblems, so that the government can realize the current affairs inter action of information with the people.If people's messages c an be divided accurately and quickly, and hot information can be e xcavated in time, it will be of great help to the management level a nd efficiency of the government.

In response to question 1: We first summarize the categories of the first-level tags, and calculate the probability of each category of the first-level tags in Annex 2. The basic idea of Bayesian clas-sification is to use the joint probability of the phrase and category o f a message to estimate the category probability of a given messag e. Assume that the unmarked message is a category of first-level t ags, and the category set is the type of first-level tags. The categor y with the highest probability of the assumed message condition is output as the category of the message.Establish the model and us e Matlab software to realize the classification results. Finally, we us

e F-Score to evaluate our classification method.

Aiming at question two: Based on the message classification o f question one, we can still use the model of question one and Matl ab software to classify similar messages. This model can classify t he messages with the highest phrase similarity. It is to put together similar problems reflected by different people at different times. Aft er comparing the probability of occurrence of each type of problem , you can sort out the top five most popular message questions, fin d out the initial time and the last time of the message, and you can determine the time period.

In response to question three: We evaluate the quality of the r esponses from three perspectives: relevance, completeness, and i nterpretability.

Keyword: Matlab;Bayesian classification algorithm;A primary label; Similarity;F-Score;

目录

一、问题重述

问题一：群众留言分类。首先根据附件 1 所给的一级分类对留言进行分类。由于电子政务系统存在较多问题，其次根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

问题二：热点问题挖掘。做两个表，第一个表是热度指数排前五的留言信息表，第二个表是第一个表里的留言明细表。

问题三：答复意见的评价。从答复的相关性、完整性、可解释性等角度对附件 4 答复意见的质量给出一套评价方案，并尝试实现。

二、符号说明

d	假设某留言为一级标签的某类
c_1	城乡建设
c_2	环境保护
c_3	交通运输
c_4	教育文本
c_5	劳动和社会保障
c_6	商贸旅游
c_7	卫生计生
$N(C_i)$	一级标签集中类别C_i的样本数量
N	一级标签样本总数
v_i	一级标签中的词组表

t_k	v_i中的一个词组

三、问题分析

3.1 问题一的模型建立

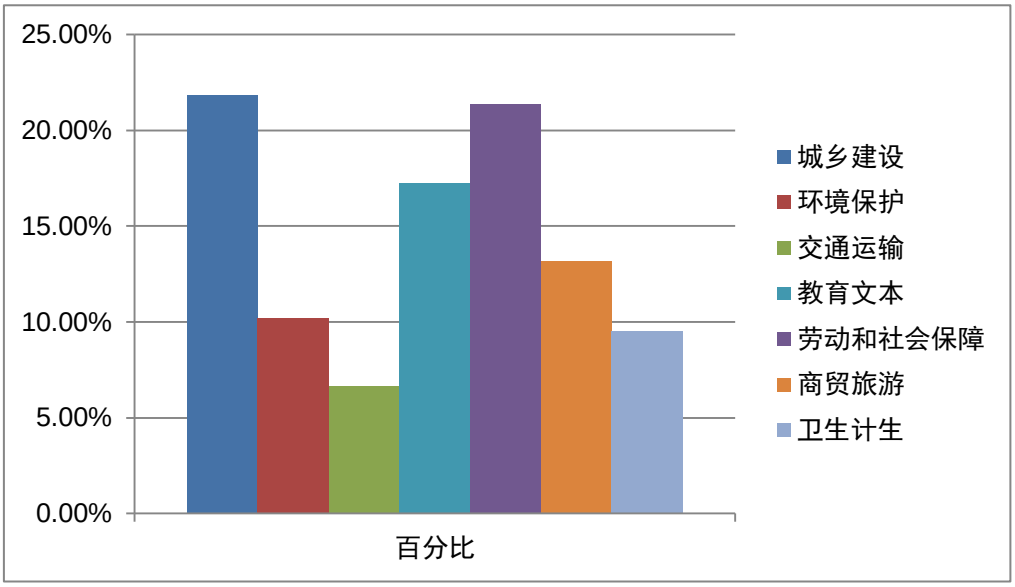
解决问题一主要分为四个部分，第一部分：根据附件 2 分类汇总出一级标签的几个类型并计算每种类型的占比概率；第二部分：采用贝叶斯分类算法建立模型；第三部分：采用 Matlab 软件实现留言分类；第四部：用 F-Score 对分类方法进行评价。

3.1.1 计算一级标签每种类型的占比概率

根据附件 2 可以分类汇总出一级标签中有城乡建设、环境保护、交通运输、教育文本、劳动和社会保障、商贸旅游、卫生计生这七种类型如表一，之后我们采用计算机计算出每一种类型的概率并得出统计图表。

表 3.1.1

一级标签	数量	百分比
城乡建设	2009	21.81%
环境保护	938	10.18%
交通运输	613	6.66%
教育文本	1589	17.25%
劳动和社会保障	1969	21.38%
商贸旅游	1215	13.19%
卫生计生	877	9.52%



图一

3.1.2 模型的建立

此留言分类我们理解为寻找这样一个函数φ：$\mathrm{D}\times\mathrm{C}\to\{T\to F\}$,

其中$D=\{d_1,d_2,\cdots,d_j\}$

表示需要进行分类的留言。C=$\{c_1,c_2,c_3,c_5,c_6,c_7\}$表示一级标签中的七个类型，$T$表示对于$(d_j,c_i)$来说，留言$d_j$属于类$c_i$；$F$表示对于$(d_j,c_i)$来说留言$d_j$不属于类$c_i$，这里就需要找一个函数映射，准确完成$\mathrm{D}\times\mathrm{C}\to\{T\to F\}$的函数映射，这个映射过程就是我们建立的贝叶斯分类模型。

假设未标注的留言为d，一级标签下的类别集合为

C=$\{c_1,c_2,c_3,c_5,c_6,c_7\}$,

概率模型分类是对$1\le i\le 7$求条件概率$P(C_i/d)$,将与留言d条件概率最大的那个类别作为该留言的输出类别。而要将留言d准确的归类到C中，我们需要通过贝叶斯分类算法计算出，在假设留言d发生下C发生的概率。即

$$P(C|D)=\frac{P(D|C)P(C)}{P(D)}=\frac{P(D|C)P(C)}{\sum_{c_i\in C}P(D|C=c_i)P(C=c_i)}$$

贝叶斯规则计算留言d属于每一个类别的可能性$P(C_i/d)$，然后将留言d标注为概率最大的那一类。即

$$Class(d)=arg\,max_{c_i\in C}\frac{P(d|c_i)P(c_i)}{\sum_{c_i\in C}P(d|c_i)P(c_i)}$$

而

$$P(c_i)=\frac{N(c_i)}{N}$$

其中

$$P(d|c_i)=\prod_{k=1}^{n}P(t_k|c_i)\,(n=1\cdots 7)$$

这样我们即可将留言d进行准确的分类。

3.1.3 用 Matlab 软件实现留言的分类

在利用 Matlab 软件实现群众留言分类成一级标签的七个类型时我们分三步进行。首先，将附件 2 的留言导入，将这些留言分为训练集以及测试集。其次，构建分类器，进行留言的数据训练，计算一级标签的每个类别在留言的训练样本中的出现频率，估计每个留言的假设类别与一级标签的七个类别的条件概率，并将结果记录下来。输入的是一级标签中七个类型以及群众留言，输出的是分类器。最后用我们建立出来的分类器对群众留言按照一级标签的分类项进行分类。这一部分见附件的代码。

3.1.4 用 F-Score 对分类方法进行评价。

在模型建立之后我们采用 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$P_i = \frac{检索出的相关信息量}{检索出的信息总量} \times 100\%$，$R_i = \frac{检索出的相关信息量}{系统中的相关信息总量} \times 100\%$。

3.2 找出排名前五的热点问题

解决问题二我们主要完成这两个部分，第一个部分：利用问题一的模型将附件 3 的留言先按照附件 1 的一级分类中的七个类型分类，其次再按照附件 1 的二级分类对留言进行精准分类，最后按照附件 1 的三级分类对留言做最后一次分类，这样我们分类出来的留言就精准到字词之间，就可以将不同的人在不同的时间的反映类似问题归类在一起。第二部分：利用 Matlab 软件实现三次分类后的结果；第三部分：观察计算第一部分分类到三级标签的留言，计算出类似留言的占比概率，排出前五名热度问题。

3.2.1 利用贝叶斯分类算法对附件 3 进行 3 次分类

在附件 3 中我们可以观察到数据是杂乱无章的，我们要排出前五名的热度问题，首先要对 留言的问题进行分类处理，将不同的人在不同的时间的反映类似问题归在一起，利于概率计算排序出热度最高的前五名问题。这里我们利用问题一建立的模型先将附件 3 的留言按一级分类归类，这样初步粗糙的将类似留言归类在一起；再次利用问

题一建立的模型按照附件 1 中二级分类对留言进行再次分类，当进行第二次分类后，我们得出的结果是比较精准了的。之后经过第三次分类，即按照三级分类再次分类就精准的将类似的留言分在一起了。我们这里进行了三次分类，主要是能够将类似的问题精准的分在一起，如果直接用问题一的模型按照三级分类直接分就可能不会那么精准，误差就会比较大，这样可能导致排序出错误的前五名热度问题。

$$其中D = \{d_1, d_2, \cdots, d_m\}$$

需要进行分类的留言。

$$C_{1j}=\{c_{11}, c_{12}, \cdots, c_{1j}\}(i = 1,2,3; j = 1,2, \cdots n)$$

$$C_{2j}=\{c_{21}, c_{22}, \cdots, c_{2j}\}$$

$$C_{3j}=\{c_{31}, c_{32}, \cdots, c_{3j}\}$$

表示一级分类，二级分类，三级分类中的各级类型。

利用模型：

$$Class(d) = arg\,max_{c_{ij} \in C} \frac{P(d|c_{ij})P(c_{ij})}{\sum_{c_i \in C} P(d|c_{ij})P(c_{ij})}$$

进行三次分类。

其中

$$P(c_{ij}) = \frac{N(c_{ij})}{N}$$

$$P(d|c_{ij}) = \prod_{k=1}^{n} P(t_k|c_{ij})\,(n = 1 \cdots i)$$

3.2.2 利用 Matlab 实现附件 3 的留言分类整理

这里的方法与程序与问题一类似，但是这里要对 3.1.3 中的方法进行三次循环，第一次是进行一级分类，第二次是进行二级分类，第

三次进行三级分类。由此我们便可以将附件 3 中类似的留言划分在一起。详细内容见附件。

3.2.3 排序出前五名热点问题

在 3.1.1 中，我们已经将杂乱无章的附件 3 的留言按照附件 1 的三级分类将不同的人在不同的时间的反映类似问题分在一起，这样我们通过观察，统计出类似留言的计数，降序排出热点前五名的问题。之后找出一类问题最开始的留言时间和最后的留言时间便得出时间段。如下表：

表 1—热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	33	2019/11/18 至 2019/12/15	投诉小区附近搅拌站	投诉小区附近搅拌站噪音扰民
2	2	5	2019/3/19 至 2019/5/30	A 市江山帝景新房	A 市江山帝景新房有严重安全隐患
3	3	10	2019/7/18 至 2019/7/23	A7 县新国道 107	国道 107 距我家仅 3 米，相关政府部门为何不同意拆迁？
4	4	16	2019/7/8 至 2019/8/16	A7 县诺亚山林小区门口	坚决反对在 A7 县诺亚山林小区门口设置医院
5	5	94	2019/1/2 至 2019/1/11	A 市经开区东六线以西泉塘昌和商业中心以南	问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划

表 2—热点问题留言明细表

问题ID	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
1	272447	A909206	投诉小区附近建设搅拌站	2019-11-20 19:12:22	，连晚上都不能好	0	0
1	208285	A909205	投诉小区附近搅拌站噪音扰民	2019-12-15 12:32:11	还是有很大噪音，	0	24
1	255008	A909208	投诉小区附近搅拌站噪音扰民	2019-11-18 12:23:22	过来的，体会最	0	0
1	261072	A909207	投诉小区附近搅拌站噪音扰民	2019-11-23 23:12:22	居住区建立搅拌站	2	9
1	266665	A00096279	投诉小区附近搅拌站噪音扰民	2019-12-04 17:23:22	的正常生活。想问	0	0
2	190812	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:34:02	天气后过道全部是	0	0
2	289893	A00095451	A市江山帝景新房有严重安全隐患	2019/5/30 17:20	天气后过道全部是	0	0
2	319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	天气后过道全部是	0	0
2	227671	A0008477	A市江山帝景新房施乱差，有安全隐患	2019/3/20 9:37:45	，日晒雨淋且此处	0	0
2	239937	A0008477	A市江山帝景新房施乱差，有安全隐患	2019/3/19 19:37	，日晒雨淋且此处	0	10
3	226753	A00020543	A7县新国道107要我家仅3米，相关部门为何不同意拆迁？	2019/7/23 11:03	全统计，整栋房屋开	0	0
3	201447	A00020543	7县新国道107要我家仅3米，相关政府部门为何不同意拆迁	2019/7/22 17:04:08	全统计，整栋房屋开	0	2
3	229903	A00020543	7县新国道107要我家仅3米，相关政府部门为何不同意拆迁	2019/7/22 17:05:04	全统计，整栋房屋开	0	0
3	273925	A00020543	7县新国道107要我家仅3米，相关政府部门为何不同意拆迁	2019/7/18 10:47:31	全统计，整栋房屋开	0	3
3	278907	A00020543	7县新国道107要我家仅3米，相关政府部门为何不同意拆迁	2019/7/18 10:48	全统计，整栋房屋开	0	0
4	194022	A00042107	坚决反对在A7县诺亚山林小区门口设置医院	2019/7/8 10:39	我们的安宁居住环	0	1
4	210107	A00042107	坚决反对在A7县诺亚山林小区门口设置医院	2019/7/8 10:38:38	我们的安宁居住环	0	14
4	226871	A000726	坚决反对在A7县诺亚山林小区门口设置医院	2019/8/16 8:37	小孩的身心健康	0	1
4	252413	A000726	坚决反对在A7县诺亚山林小区门口设置医院	2019/8/16 8:36:38	出现的人身误伤；	0	0
5	233542	A00080329	问问A市经开区东六线以西息塘昌和商业中心以南的有关规划	2019/1/2 20:27	重工原厂房和闲置	0	24
5	239670	A00080329	问问A市经开区东六线以西息塘昌和商业中心以南的有关规划	2019/1/11 15:46	置业有限公司厂房	0	41
5	256358	A00080329	问问A市经开区东六线以西息塘昌和商业中心以南的有关规划	2019/1/2 20:27:07	重工原厂房和闲置	0	29

注：详细的热点问题留言明细表见附件

3．3 评价留言的答复意见

针对问题三，我们主要从三步对留言的答复意见进行评价。第一步我们从相关性的角度进行评价；第二步我们从完整性的角度对留言的回复进行评价；第三步我们从可解释性的角度对留言回复进行评价。

3.3.1 相关性方面

是指两个变量的关联程度，即问题与答复之间的关联程度。一般地，从几何图形可以观察到两个变量之间分别有三种关系：即是正相关、负相关、不相关。两个变量的正相关就是一个变量高的值对应于另一个变量高的值，相似地，低的值对应低的值。负相关就是指一个变量高的值对应于另一个变量低的值。如果两个变量间没有关系，即一个变量的变化对另一变量没有明显影响，那么这两个变量不相关。我们根据附件 4 分类汇总得到了网上对问政留言与政府对留言的答复意见是成一一对应关系，政府对留言都在一定时间内一一做出了答复。

3.3.2 完整性方面

是指存储在数据库中的所有数据值均正确的状态，也就是说如果数据库中存在有与这个数据库不符合的数据值，那么可以说这个数据库称为已丧失数据完整性。问题三中具体表现为每一个答复意见对每一个问题所包含内容的概括全面程度，就是在对其进行评价时我们可以根据答复意见的内容判断其对相应问题内容回答的完整性，就可以得到答复意见的完整程度。

3.3.3 可解释性方面

指的是对于某种事物的出现、评价有依据，有条理做出的一种解释。对于附件四网上留言政府恢复等等问题、我们基于问题一、问题

二建立的模型给出了明确的解释，政府给出的答复意见是经过实体调查、讨论、有依据、能落实到实际，能很好的解决问题。所以，对于上述模型建立的分类方法是可行的、有质量的。

四、参考文献

[1]xsdjj,自然语言处理——文本分类概述，2018—11—05

[2]宗成庆，统计自然语言处理，清华大学，2008-5

[3]靖红芳、王斌、杨雅辉、徐燕，基于类别分布的特征选择框架，2009

[4]李国华，贝叶斯公式的应用，牡丹大学学报，2011

五、附件

程序：

```
    data  =  importdata(&apos;F:/新建文件夹/附件 2.xlsx&apos;);
     wholeData=data.data;
    cv=cvpartition(size(wholeData,1),&apos;holdout&apos;,0.01);
trainData=wholeData(training(cv),:);
     testData=wholeData(test(cv),:);
     label=&apos;F:/新建文件夹/附件 2.xlsx&apos;;
attributeNumber=size(trainData,2);
     attributeValueNumber=5;
     sampleNumber=size(label,1); labelData=zeros(sampleNumber,1);
     for  i=1:sampleNumber
            if  label{i,1}==&apos;R&apos;
                    labelData(i,1)=1;
            elseif  label{i,1}==&apos;B&apos;
                    labelData(i,1)=2;
            else
                    labelData(i,1)=3;
            end
    end
    trainLabel=labelData(training(cv),:);
     trainSampleNumber=size(trainLabel,1);
     testLabel=labelData(test(cv),:);
     labelProbability=tabulate(trainLabel);
P_d1=labelProbability(1,3)/200;
P_d2=labelProbability(2,3)/200; P_d3=labelProbability(3,3)/200;
count_1=zeros(attributeNumber,attributeValueNumber);
count_2=zeros(attributeNumber,attributeValueNumber);
count_3=zeros(attributeNumber,attributeValueNumber); for  jj=1:3
            for  j=1:trainSampleNumber
                    for  ii=1:attributeNumber
                            for  k=1:attributeValueNumber
                                    if  jj==1
                                            if  trainLabel(j,1)==1&&t
rainData(j,ii)==k
                                                    count_1(ii,k)=cou
nt_1(ii,k)+1;
                                            end
                                    elseif  jj==2
                                            if  trainLabel(j,1)==2&&t
rainData(j,ii)==k                                                   coun
```

```
t_2(ii,k)=count_2(ii,k)+1;
                                                    end
                                                else
                                                    if  trainLabel(j,1)==3&&t
rainData(j,ii)==k
                                                        count_3(ii,k)=cou
nt_3(ii,k)+1;
                                                    end
                                                end
                                            end
                                        end
                                    end
                                end
    end
    P_a_d1=count_1./labelProbability(1,3);
     P_a_d2=count_2./labelProbability(2,3);
     P_a_d3=count_3./labelProbability(3,3);
     labelPredictNumber=zeros(2,1);
     predictLabel=zeros(size(testData,1),1);
     for  kk=1:size(testData,1)
            testDataTemp=testData(kk,:);
            Pcd1=1;
            Pcd2=1;
            Pcd3=1;
            for  iii=1:attributeNumber
                    Pcd1=Pcd1*P_a_d1(iii,testDataTemp(iii));
                    Pcd2=Pcd2*P_a_d2(iii,testDataTemp(iii));
      Pcd3=Pcd3*P_a_d3(iii,testDataTemp(iii));        end
            PcdPd1=P_d1*Pcd1;
            PcdPd2=P_d2*Pcd2;
            PcdPd3=P_d3*Pcd3;
            if  PcdPd1>PcdPd2&&PcdPd1>PcdPd3
                    predictLabel(kk,1)=1;
                    disp([&apos;this  item  belongs  to  No.&apos;,nu
m2str(1),&apos;  label  or  the  R  label&apos;])                labe
lPredictNumber(1,1)=labelPredictNumber(1,1)+1;
            elseif  PcdPd2>PcdPd1&&PcdPd2>PcdPd3
                    predictLabel(kk,1)=2;
                      labelPredictNumber(2,1)=labelPredictNumber(2,1)
+1;
                    disp([&apos;this  item  belongs  to  No.&apos;,nu
m2str(2),&apos;  label  or  the  B  label&apos;])
            elseif      PcdPd3>PcdPd2&&PcdPd3>PcdPd1                p
redictLabel(kk,1)=3;
                      labelPredictNumber(3,1)=labelPredictNumber(3,1)
```

```
+1;
                    disp([&apos;this  item  belongs  to  No.&apos;,nu
m2str(3),&apos;  label  or  the  L  label&apos;])
            end
    end
```