# 智慧政务中的文本挖掘应用

## 摘要

随着互联网的发展，各个网站平台不止是为人民娱乐、八卦，还逐渐发展出各种问政平台，成为政府了解民意、汇聚民智、凝聚民气的重要渠道，同时为我们带来巨大方便的同时，也存在着各种各样的问题，问题一收集到的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。我们可以利用自然语言处理和文本挖掘的方法解决相应的问题。是一个多分类问题。

问题 2 我们可以采用，通过利用 excel 去对重后的附件 3 在某一时段内反映特定地点或特定人群问题的留言按民生生活、社会安定、环境保护分类三个项目进行归类，定义合理的热度评价指标，并给出评价结果，根据计数多的话题去判定热点话题。

留言板是因特网上常见的一种服务，也是一种交互式网页。通常人们可以在留言板上上传一些文章或发起一些话题,或是通过留言板给网站的工作人员留下一些意见和看法，留言板作为网友之间互传消息、相互交流的渠道之一。且工作人员也对留言板进行部分管理，并且对留言板页面进行设计，使页面更雅观，实现的功能有所增强;加上对管理页面的设计,使管理员能更方便清楚的对用户、页面进行管理等。总而言之，留言板将会成为一个功能强大、雅观方便、让人们畅所欲言的“公共场所”。且留言板较全面地利用技术实现留言板的基本功能:增、删、查、改，并增加了一些特色功能。

**关键字：**留言问题、多分类、文本挖掘、聚类。

# 一、问题的重述

1、群众留言分类

在处理网络问政平台的群众留言时，工作人员首先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。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

通常使用 F-Score 对分类方法进行评价：其中$P_i$为第 *i* 类的查准率，$R_i$为第 *i* 类的查全率。

2、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

3、答复意见的评价

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。

## 二、问题的分析

问题一分析方法与过程：在因特网飞速发展的今天，互联网成为人们快速吸收、分享和传递信息的重要渠道之一，它在政治、经济、生活等各个方面发挥着重要的作用。也成为了我们生活中必不可少的一项技术。人们想在网上发布信息是通过网站来实现的，获取信息也是在成千上万的网站中按照一定的检索方法将所需要的信息从网站上获取出来。随着互联网的发展，各个网站平台不止是为人民娱乐、八卦，还衍生出各种问政平台，成为政府了解民意、归纳民智、凝聚民气的重要渠道之一，但这些问政平台为我们带来巨大方便的同时，也存在着各种各样的问题，我们知道对各种留言进行分类依靠人工经验处理，存在工作量大、效率不高，且差错率高等问题。智慧政务中的文本挖掘应用，附件给出了收集到的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。我们可以才用自然语言处理和文本挖掘的方法来解决相应的问题。

问题二分析我们采用，通过利用 excel 去对重后的附件 3 在某一时段内反映特定地点或特定人群问题的留言按民生生活、社会安定、环境保护分类三个项目进行归类，定义合理的热度评价指标，并给出

评价结果，进行计数汇总统计，根据计数多的话题去判定热点话题。

问题三随着大数据时代的发展,人们已经对网络不再感到陌生。在科技飞速发展的今天,网络技术与各行各业都进行了有效的结合。网上购物不再是梦想，网上交友正逐渐形成一种新文化，电子商务、网络营销已成为一种潮流。为了匹配强大的网络功能，我们必须要有一个健全的系统才能使网络技术得到最好的利用。随着时代的进步网站的作用越来越重要，被称之为数字媒体，优势众多，所以现在不少企业都有或正在建设自己的网站。而留言板作为网站重要的一个部分，一直在人类的生活中充当着重要角色。

## 三、模型假设

假设一：文本语义带来的词语交叉,比如:交通局的亲属拖欠我们工资；

假设二：多分类问题带来的难度(转化为多个二分类)；

假设三：数据不平衡带来的影响(数据增强)；

假设四：长文本的无意义表达太多(是否转为短文本、关键句)；

## 四、符号说明

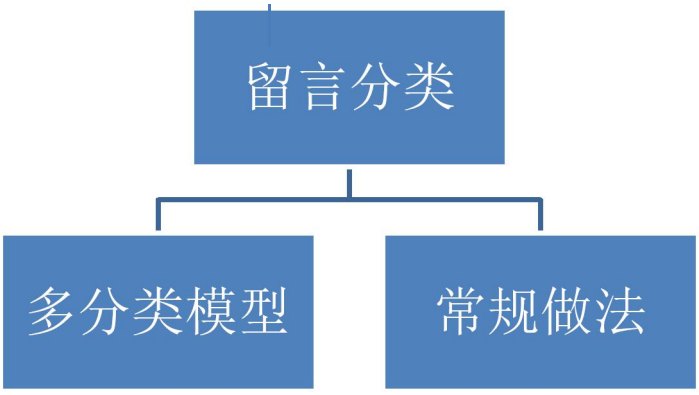
| | |
|---|---|
| P_i | 其中 P_i 为第 i 类的查准率 |
| R_i | R_i 为第 i 类的查全率。 |

# 五、建立模型求解

问题一的求解

由附件一为三级标签体系的关系，体现的是三级标签之间是从属和包含关系，以及三级标签数据中的体现，给出的标签体系是完整的，但是数据中并不一定包含全部。而附件二给出的是具体数据，用 excel 打开查看数据可以看到一共 7 个一级标签，二级标签、三级标签都很多。附件一、附件二体现的是一个多分类问题，我们可以对问题一的留言分类，将二分类模型扩展到多分类的方法。建立一个多分类模型，然后加上常规做法。

问题一解决方法分为两大步：如下图所示




模型的建立：附件一存在三级标签，属于多分类问题，为了解决多分类问题带来的难度我们可以将多分类问题转化为多个二分类，由此达到简化，还要考虑附件二，文本留言详细部分大部分长文本表达并无太大意义，关键词关键字只有寥寥几句或一句，其长文本多而复杂，为了提取有用信息我们应将长文本转为短文本、关键句，好进行信息的分类，我们的文本挖掘技术这就可以上场了。如果有 k 个类

别要预测，训练阶段就只需训练 k 个二分类模型。如，针对第 i 类的训练。我们需要将第 i 类的样本取出存储作为类别 1 的训练数据，再将其余类样本取出作为类别 1 的训练数据，之后依照得到的训练数据来训练这个模型。往往预测阶段需输入一个预测实例，用这 k 个模型来分别预测，得到 k 个预测值，将预测值最大的那个模型所对应的类别标记作为预测结果。

## 多分类——One-vs-Rest

- 对于一个K类问题，OvR将训练K个二分类模型$\{G_1, G_2, \dots, G_K\}$，模型$G_i$会把第$i$类看成一类，把其余类看成另一类并尝试通过训练来区分开第$i$类和剩余类别。若$G_i$有比较大的自信来判定输入样本$x$是第$i$类，那么$G_i(x)$会是（或不是）一个比较大的正（负）数；否则$G_i(x)$会是比较小的正（负）数。
- 训练好K个模型后，直接将输出值最大的模型所对应的类别作为决策即可，亦即：

$$y_{pred} = arg\max_i G_i(x)$$



这个方法存在缺陷，使用这个方法会使本来样本平均的训练集合变得不均衡。比如对于一个有 k 个类的多种分类问题，正样本集合和负样本集合的样本数之比在最初训练集均匀的情况下是 1/(k−1) 。对于该缺陷的解决方法，一种较常见的做法是仅抽取负样本集合中的一部分来进行训练（比如抽取其中的三分之一）。

我们知道 One-vs-On 与 One-vs-Rest 有所不同，我们在这里将训练 k(k-1)/2 个二分类模型。也就是说在练习阶段要随机的从 k 个类别中抽取 2 个类别来练习为一个分类器。预测阶段，先输入一个样本实例，然后统计 k(k-1)/2 分类器的预测类别，在最终结果中出现次数最多的类别，就是得到的最终预测结果。

## 多分类——One-vs-One

- 对于一个K类问题，OvO将直接训练出$\frac{K(K-1)}{2}$个二分类模型$\{G_{12}, G_{13}, \dots, G_{1K}, G_{21}, G_{22}, \dots, G_{2K}, \dots, G_{k-1K}\}$。模型$G_{ij}(i<j)$将接受且仅接受所有第$i$类和第$j$类的样本，并尝试通过训练来区分开第$i$类和第$j$类。输出

$$G_{ij}(x)=\begin{cases}-1, x\in C_j\\ +1, x\in C_i\end{cases}$$

- 训练好$\frac{K(K-1)}{2}$个二分类模型后，OvO将通过投票表决来进行决策，在$\frac{K(K-1)}{2}$次投票中得票最多的类即为模型所预测的结果。
- 如果只有两个类别得票一致，则直接看针对这两个类别的模型结果即可。如果多于两个类别的得票一致，则需要具体问题具体分析。



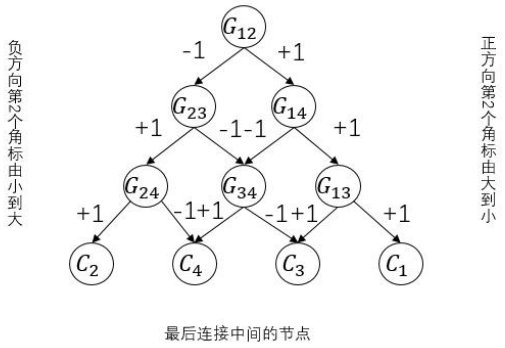
这个模型的缺陷是由于模型的量级是 k^2，所以它的时间开销会相当大。

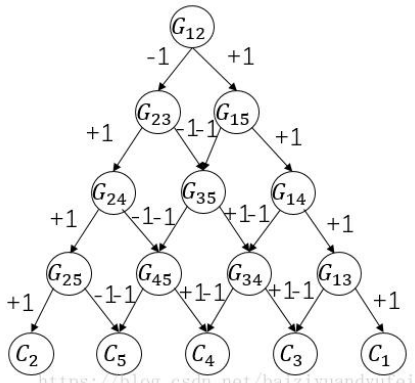
而 Directed Acyclic Graph Method 的训练阶段和 *OvO* 的训练阶段几乎一致，区别只在于最后的预测结果部分有所不同。

构建一个下图所示的分类器图：

## 多分类——Directed Acyclic Graph Method

- 它的训练过程和OvO的训练过程完全一致，区别只在于最后的决策过程






首先输入一个需要预测的实例，首先将它送入根节点 G12 分类器，再按照预测得出的结论，依次送入后续分类器节点，我们用树枝上的数字（即权重）+1 表示预测结果为 $G_{ij}$ 中的 i 类别，-1 表示预测

结果为 $G_{ij}$ 中的 j 类别。以叶节点所表示的类别作为预测结果。

再利用公式

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i}$$

通常使用 F-Score 对分类方法进行评价得到其查全率，查准率。

模型的求解：实际上我们都可以知道这是一个文本分类(多分类)问题要解决这一问题，我们一般常用的就是最经典的 word2vec 工具，该工具在 NLP 领域具有非常重要的意义，还有要了解，常用的文本分类有哪些?然后知道其大体上分为基于传统机器学习的文本分类模型，是基于深度学习的文本分类模型，分类问题的评价方法

结果分析：所以从附件一、附件二的内容我们需要对文本进行数据挖掘，结合附件一、附件二，进行信息分类。对附件二中的留言信息结合附件一，分类，把留言内容根据附件一的二级分类、三级分类，留言中是否包含相关文字信息，然后归类。可以从自然语言处理涉及的应用方向解决问题。问题一可能存在的难点：文本语义带来的词语交叉,比如:交通局的亲属拖欠我们工资、多分类问题带来的难度(转化为多个二分类)、数据不平衡带来的影响(数据增强)、长文本的无意义表达太多(是否转为短文本、关键句)，对其进行数据筛选、数据统计。最后的到结果，根据从留言内容中查找关键词、关键句，进行模型建立，最终得到结果，留言分类成功，大大减少了工作量、解决了效率低，且差错率高等问题。

5.2 问题 2 分析方法与过程.

5.2.1 数据筛选

(1)根据附件 3 对不同话题类别进行分类筛选，得到小区油烟噪音扰民、毕业学生的实习地点、城市环境污染、生活中的违规违纪现象、电信诈骗、传销等多个不同的话题类别。

(2)根据附件 3 对不同话题进行分类，得到更详细的话题。

(3)根据附件 3 对不同的话题所属大类分类，分为民生生活、社会安定、环境保护三大类。

5.2.2 数据统计

(1)对各个话题出现的讨论次数进行计数，通过排序得出排名前 5 的话题，定义为热点话题,并输出（见附件热点问题表.xls）；

(2)统计出现的热点话题，按照热点话题的热度依次排序得出排名前 5 的所有话题，并输出所有留言信息（见附件热点问题明细表.xls）；

(3)依据附件 3 给出的数据进行归类，定义合理的热度评价指标。

5.2 问题 2 结果分析

5.2.1 对热门话题的分析

通过对所有话题排序计数，选取排名前 5 的热点话题进行分析，(详见附件热点问题表)分析得出小区临街餐饮店油烟噪音扰民、A 市的违规违纪现象、搅拌站噪音污染、电信诈骗、学校强制学生去定点企业实习五个热点问题。

（见表 1)可以发现当今社会人们对于与自己息息相关的民生问题颇为关注，尤其在最近几年里，随着经济的增长，人们越来越关注自己的上层建筑了。民生问题是一个永远都不会落后的话题，它无时无

刻不在。关注民生就是在关注我们自己。

表一：热点问题表

| 热度指数 | 问题描述 |
|---|---|
| 26 | A市的违规违纪现象 |
| 24 | 小区临街餐饮店油烟噪音扰民 |
| 22 | 搅拌站噪音污染 |
| 8 | 电信诈骗 |
| 5 | 学校强制学生去定点企业实习 |

5.2.2 定义热门评价指标

随着经济发展，我们的生活日新月异，随着微信、微博、市长信箱、阳光热线等的推广，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。而我们通过对附件 3 的分析可以得出人们普遍关注的都是有关民生生活的话题，其次是环境，最后是社会安定。（如图 1 所示）与民生相关的关注度远远提高。同时，也帮助相关部门及时解决。

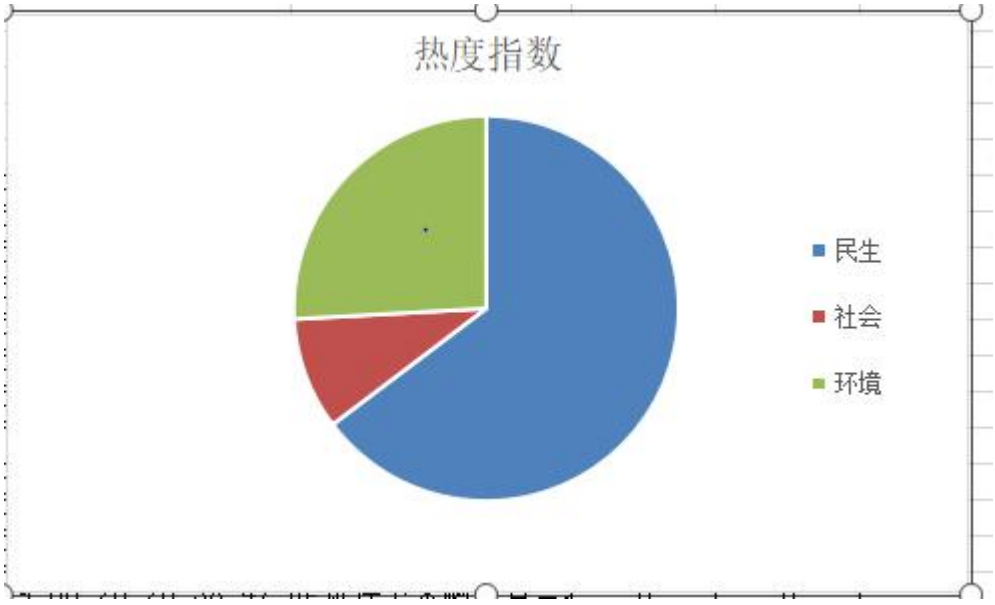



图 1-热门评价指标图

如图所示，在前 5 个热点问题里民生问题所占比例为 65%，环境污染占了 26%，社会安定占了 9%。

二、数据分析

从 2016 年 9 月至 2018 年 10 月，网站留言渠道收到的有效信件为 3458 件，主要可分为咨询类、投诉类、救助类和献策类。其中，投诉类占比较多，占比达到 57%；咨询类占 39%，救助类占 3%，献策类占 1%。如下图 2 所示。

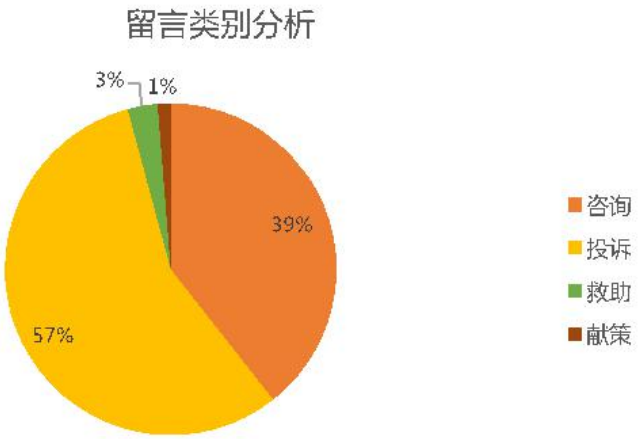



图 2 留言类别

对样本的内容进行分析后，发现主要分布在以下主题：户籍身份、教育培训、医疗卫生、住房租房、交通出行、救助资助、计划生育、小区物业、社会治安、环境污染、劳动仲裁、噪音污染、企业开办、工商税务、人才引进等。各主题的分布情况如下图 3 所示。


图 3 主题分布

“公众留言”渠道已是政府汇聚民意、倾听民声、感受民情的网络窗口，区政府建立了完善的分发、流转、办理、反馈和督办机制，严把时间和质量关，认真对待每一封信件，力争在第一时间回复公众。对于复杂的信件，及时交办相关部门，对相关情况进行核实，沟通协商处理方案，并尽快回复公众，争取得到公众认可。

## 六 参考文献

[1]周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 中国科学院研究生院(计算技术研究所)，2005.

[2]王礼礼. 基于潜在语义索引的文本聚类算法研究[D]. 西南交通大学，2008.

[3]高策理，蔡斌. 使用主成分分析进行综合排名时出现高相关指标的研究[J]. 数学的实践与认识，2004

[4] https//github.com/fxsjy/jieba

[5]玉千， 王成，冯振元，叶金凤.K-means 聚类算法研究综述.2012

[6]朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学硕士学位论文.2013