基于智慧政务中的文本数据挖掘分析

摘要:近年来，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。本文主要围绕三个问题展开,群众留言分类,热点问题挖掘以及答复意见的评价.群众留言分类主要采用python软件进行分词,数据清洗以及去停用词来解决.热点问题挖掘主要采用文本转化矩阵之后通过聚类来解决,答复意见的评价采用模糊综合评价法来解决。

关键词: 智慧政务　数据预处理　聚类分析　模糊综合评价

Abstract: In recent years,WeChat Micro Blog Mayor Mailbox Sunshine Hotline and Other Internet Political Consultation Platforms become gradually an important way which government to understand public opinion,amass the people’s wisdom,and Condense people's spirit.The amount of text data related to various social conditions and public opinions is continuously rising,Work for related departments that used to rely mainly on manpower to divide messages and sort out hot spots bring great challenges.At the same time, with the development of big data cloud computing artificial intelligence and other technologies.The establishment of a smart government system based on natural language processing technology has become a new trend of social governance innovation and development which will greatly promote the government's management level and efficiency.This article focuses on three issues:Classification of mass messages,Mining Hot Issues and Evaluation of responses.Dealing with the first one ,we mainly adopt python software to separate words,clean data ,and remove stop words .Solving the second one,we use Text transformation matrix and Cluster analysis. Settle the third one, we apply FCE .

Key words : Smart government Data Preprocessing Cluster analysis FCE

目录

1.挖掘目标

本次建模主要针对“智慧政务”中的文本挖掘应用，采用python软件进行去重、分词、去停用词以及spss软件进行聚类,还有采用模糊综合评价法达到以下目标：

（1）对文本分类

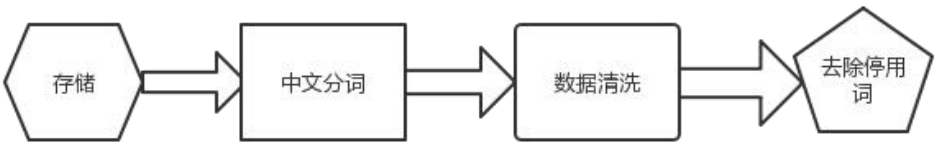
（2）挖掘热点问题

（3）评价答复意见

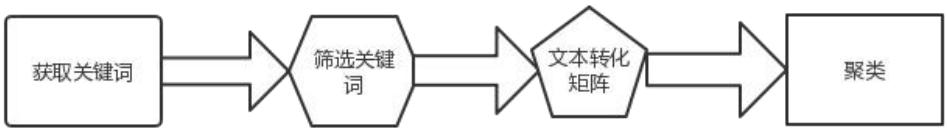
2.分析过程与方法

2.1 流程

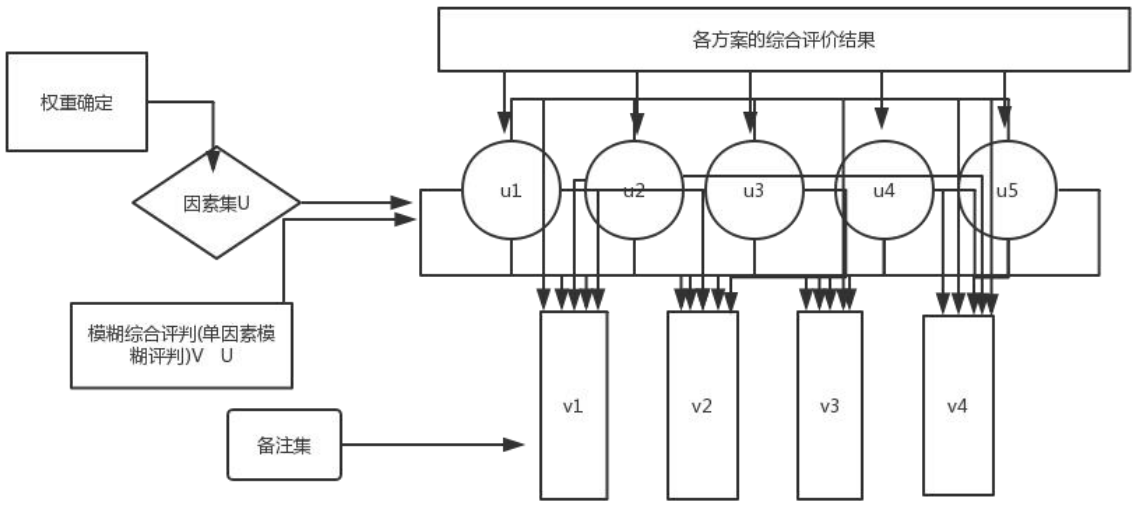
问题一流程图



问题二流程图



问题三流程图



2.2 具体步骤

问题一分析过程与方法

具体方法：先用一级分类,分十五个类,每个类再通过三级分类给关键词,对每个类中的关键词都进行去停用词和去重,对附录二中的留言进行分词,对分词的结果进行去停用词,用每个留言去匹配十五个类中的关键词,统计每一个类中关键词的次数。

存储：excel 中每一个一级标题后面对应的三级标题的内容以 txt 的形式存储下来。部分程序段如下：

```
#!/usr/bin/env python3
# coding: utf-8
import xlrd
import jieba

# 打开 excel 文件，创建一个 workbook 对象，book 对象也就是 fruits.xlsx 文件，表含有 sheet 名
rbook = xlrd.open_workbook('temp1.xlsx')
# sheets 方法返回对象列表，[<xlrd.sheet.Sheet object at 0x103f147f0>]
rbook.sheets()
# xls 默认有 3 个工作簿，Sheet1，Sheet2，Sheet3
rsheet = rbook.sheet_by_index(0)    # 取第一个工作簿

# 循环工作簿的所有行
f1 = open('./word/word1.txt','r+')
f2 = open('./word/word2.txt','r+')
f3 = open('./word/word3.txt','r+')
f4 = open('./word/word4.txt','r+')
f5 = open('./word/word5.txt','r+')
f6 = open('./word/word6.txt','r+')
f7 = open('./word/word7.txt','r+')
f8 = open('./word/word8.txt','r+')
f9 = open('./word/word9.txt','r+')
f10 = open('./word/word10.txt','r+')
f11 = open('./word/word11.txt','r+')
f12 = open('./word/word12.txt','r+')
f13 = open('./word/word13.txt','r+')
f14 = open('./word/word14.txt','r+')
f15 = open('./word/word15.txt','r+')
for row in rsheet.get_rows():
        product_column = row[2]    # 品名所在的列
        product_value = product_column.value    # 项目名
        if product_value != '三级分类':    # 排除第一行
                price_column = row[2]    # 价格所在的列
                price_value = price_column.value
                if row[0].value=='城乡建设':
                        f1.write('  '+price_value)
                elif row[0].value=='党务政务':
                        f2.write('  '+price_value)
                elif row[0].value=='国土资源':
                        f3.write('  '+price_value)
                elif row[0].value=='环境保护':
```

中文分词：无论是在汉语还是在英语中，词一般都是代表最小的语义单位，所以为了便于对语言信息处理，句子一般需要划分成词再进行后续的分析和处理，中文分词需要根据语义进行划分，由于汉语单字可以前后连接成多个单词，汉语将面临更多歧义和未登录词识别等问题，所以分词难度很大，本文采用 jieba 分词，jieba 是一个基于 Python 的中文分词器，可利用隐马尔可夫模型和维特模型比算法解决部分未登录问题。

分词程序段如下：

```
import jieba
import codecs
stopwords_path='./stopwords.txt'
stopwords = []
with open(stopwords_path, 'r') as f:
  for line in f:
    if len(line)>0:
      stopwords.append(line.strip())
for i in range(1,16):
  name = './word/word'+str(i)+'.txt'
  words_save = []
  with open(name, 'rb') as f:
    for line in f:
      seg = jieba.cut(line.strip(), cut_all = False)
      s=' '.join(seg)
      m=list(s)
      op = './use-info/word'+str(i)+'.txt'
      with open(op,'wb')as f:
        temp = ''
        for word in m:
          if word!=' ':
            temp = temp+word
          elif word==' ':
            if temp not in stopwords:
              if temp not in words_save:
                words_save.append(temp)
                temp = ' '+temp
                f.write(temp.encode('utf-8'))
            temp = ''
          #print word
```

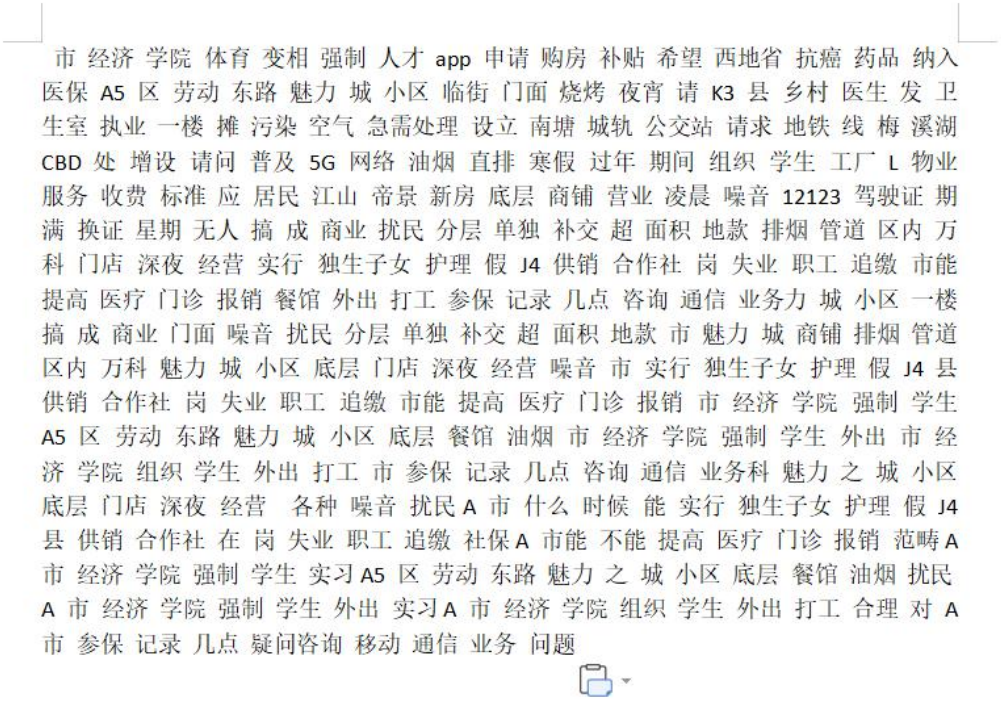
数据清洗：数据清洗是指发现并纠正数据文件中可识别的错误的一个步骤，包括检查数据的一致性，处理无效值和缺失值等。在当今互联网时代，数据质量参差不齐，比如，文本格式可能不一致，存在链接地址，表情符号，大量空格和空行等等，所以针对这些问题需要做出数据预处理，如删除空格空行，删除标点符号等等。

去除停用词：将留言进行 jieba 分词，然后去停用词，去除停用词可以大大减小特征词的数量，进而提高文本分类的准确性，对文本分类来说，像"的""和""在""是"等副词，量词，介词，叹词，数词。这些词汇几乎在所有文本中都会出现，不具有特殊性，没有区分度，反而会稀释那些有区分度的词，所以经常把这些词移除。

问题二分析过程与方法

获取关键词：通过去重和分词，将标题进行分词，获取关键词。

获取的关键词如下图：


市 经济 学院 体育 变相 强制 人才 app 申请 购房 补贴 希望 西地省 抗癌 药品 纳入 医保 A5 区 劳动 东路 魅力 城 小区 临街 门面 烧烤 夜宵 请 K3 县 乡村 医生 发 卫生室 执业 一楼 摊 污染 空气 急需处理 设立 南塘 城轨 公交站 请求 地铁 线 梅 溪湖 CBD 处 增设 请问 普及 5G 网络 油烟 直排 寒假 过年 期间 组织 学生 工厂 L 物业 服务 收费 标准 应 居民 江山 帝景 新房 底层 商铺 营业 凌晨 噪音 12123 驾驶证 期满 换证 星期 无人 搞 成 商业 扰民 分层 单独 补交 超 面积 地款 排烟 管道 区内 万科 门店 深夜 经营 实行 独生子女 护理 假 J4 供销 合作社 岗 失业 职工 追缴 市能 提高 医疗 门诊 报销 餐馆 外出 打工 参保 记录 几点 咨询 通信 业务力 城 小区 一楼 搞 成 商业 门面 噪音 扰民 分层 单独 补交 超 面积 地款 市 魅力 城 商铺 排烟 管道 区内 万科 魅力 城 小区 底层 门店 深夜 经营 噪音 市 实行 独生子女 护理 假 J4 县 供销 合作社 岗 失业 职工 追缴 市能 提高 医疗 门诊 报销 市 经济 学院 强制 学生 A5 区 劳动 东路 魅力 城 小区 底层 餐馆 油烟 市 经济 学院 强制 学生 外出 市 经济 学院 组织 学生 外出 打工 市 参保 记录 几点 咨询 通信 业务科 魅力 之 城 小区 底层 门店 深夜 经营 各种 噪音 扰民 A 市 什么 时候 能 实行 独生子女 护理 假 J4 县 供销 合作社 在 岗 失业 职工 追缴 社保 A 市能 不能 提高 医疗 门诊 报销 范畴 A 市 经济 学院 强制 学生 实习 A5 区 劳动 东路 魅力 之 城 小区 底层 餐馆 油烟 扰民 A 市 经济 学院 强制 学生 外出 实习 A 市 经济 学院 组织 学生 外出 打工 合理 对 A 市 参保 记录 几点 疑问咨询 移动 通信 业务 问题

筛选关键词：通过统计设定一个阈值，之后再筛选出真正有意义的关键词。

文本转化矩阵：本文中利用词向量的方法将文本数据转化为结构化的向量矩阵。

1）词向量化

文本矩阵转化的第一步就是词向量化，顾名思义，词向量化即用空间向量模型表示各个词语，进而提高计算机对自然语言的处理能力。词向量具有良好的语义特性，是表示词语特征的常用方式。把对文本内容的处理简化成对一定长度的向量的处理时，通常使用较低维度的空间向量来表示词语的特征，避免数据维数灾难。词向量中每一维的值代表一个具有一定的语义和语法上解释的特征。

2）向量矩阵

词向量化后便可以将文本数据转化向量矩阵了。通常情况下，我们将词语 w 映射到 n 维空间向量，即 $w \in R^n$ 一个文本或者句子中含有 m 个词语，把这 m 个 n 维空间向量堆放在一起，就得到整个文本或句子的空间向量模型，一个词向量矩阵 $L \in R^{n*n}$.例如给定句子 c 含有 m 个词语，$1 \leqslant i \leqslant n$, w_i 为句子 c 的空间向量矩阵 L 中的第 k_i 列，且除了第 k_i 个分量为 1,其余分量均为 0。

具体操作：以每个关键词为列，判断每个关键词是否出现，如果出现设置为 1，未出现设置为 0，获得一个以留言(标题)和关键词组成的二维数组。

聚类：对该二维数组进行聚类，获得热点问题的聚集范围。本题采用了 K 均值法聚类。

K 均值法的基本步骤

(1)选择 k 个样品作为初始凝聚点，或者将所有样品分成 k 个初始类，然后将这 k 个类的重心(均值)作为初始凝聚点。

(2)对所有的样品逐个归类，将每个样品归入凝聚点离它最近的那个类(通常采用欧氏距离)，该类的凝聚点更新为这一类目前的均值，直至所有样品都归了类。

(3)重复步骤(2)，直至所有的样品都不能再分配为止。

```
examp6.3.3<-read.table("C:/mvdata/examp6.3.3.csv", header=T,
row.names="region", sep=",") #读取文本文件

d<-dist(scale(examp6.3.3), method="euclidean", diag=T, upper=F,
p=2) #method 为距离计算方法，缺省时为"euclidean"（欧氏距离），
还包括："manhattan"（绝对值距离），"minkowski"（明氏距离），
"canberra"（兰氏距离）等

#diag 为是否包括对角线元素（缺省时为 F），upper 为是否包括上
三角距离（缺省时为 F），p 为明氏距离的幂（p=2 即为欧氏距离）

hc<-hclust(d, "ward") #离差平方和法
```

#方法还包括："single"（最短距离法），"complete"（最长距离法），"average"（类平均法），"centroid"（重心法），"median"（中间距离法）等

cbind(hc$merge, round(hc$height,2)) #聚类过程

plot(hc, hang=-1) #聚类树形图，hang 指定标签在图形中所处的高度（负值时挂在 0 下面）

rect.hclust(hc, k=3) #将聚成的三类用边框界定

cutree(hc, k=3) #将聚成三类的结果分别以 1, 2, 3 表示

examp6.3.7<-read.table("C:/mvdata/examp6.3.7.csv", header=T, sep=",") #读取文本文件

d<-as.dist(1-examp6.3.7[-1], diag=T) #转换为距离矩阵

d

hc<-hclust(d, "complete") #最长距离法

plot(hc, hang=-1) #树形图

rect.hclust(hc, k=2) #将聚成的两类用边框界定

cutree(hc, k=2) #将聚成两类的结果分别以 1, 2 表示

examp6.3.3<-read.table("C:/mvdata/examp6.3.3.csv", header=T, row.names="region", sep=",") #读取文本文件

km<-kmeans(scale(examp6.3.3), 3) #k 均值法，聚成 3 类

sort(km$cluster) #对聚类结果进行排序

结果：

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	0.33	2019/08/18至2019/09/04	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	0.17	2017/06/08至2019/11/22	A市经济学院学生	学校强制学生去定点企业实习
3	3	0.067	2019/12/19至2020/01/06	A市人社局及J4县供销合作社失业员工	追缴养老保险及参保记录疑问
4	4	0.065	2019/06/12至2019/09/08	A市及西地省	抗癌药物纳入医保及提高医疗门诊报销范畴
5	5	0.063	2018/10/27至2019/10/31	A市	增设公交和地铁站

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	点赞数	反对数
1	360103	A0012425	卡路魅力之城小区临街门面	2019/9/25 0:31:33	烟。作为烧烤夜宵更加扰民，油烟24小时晕死人，尽	1	0
1	360107	A0283523	力之城小区一楼的夜宵摊严	2019/7/21 10:29:36	居民还是觉得要维护社会和谐稳定，合法维权。为此我	3	0
1	360108	A0283523	小区一楼的夜宵摊严重污染	2019-08-01 16:20:02	居民还是觉得要维护社会和谐稳定，合法维权。为此我	6	0
1	360100	A324156	之城小区临街门面油烟直排	2019-09-05 12:29:01	烧烤熏死人。一天24小时都是烟。请政府关闭处理这个	3	0
1	360101	A324156	劳动东路魅力之城小区油烟	2019-07-28 12:49:18	前油烟机清洗也没有。每天油烟直排。重死树木。对我	4	0
1	360106	A235367	区底层商铺营业到凌晨，省	2019-08-26 01:50:38	十里，难闻不说还严重影响健康。大家对此深感困扰，	0	0
1	360105	A120356	小区一楼被搞成商业门面，	2019-08-26 08:33:03	办法居住，影响我们的晚年生活。架空层被侵占，本	1	0
1	360104	A012417	成商铺无排烟管道，小区内	2019-08-18 14:44:00	，每天进出都搞得业主一身油烟味，而且每天到凌晨	0	0
1	360109	A0080252	魅小区底层门店深夜经营，有	2019-09-04 21:00:18	边都充斥着吆喝声、拼酒声、炒菜烧烤的锅铲炭火声	0	0
1	360102	A1234140	东路魅力之城小区底层餐馆	2019-09-10 06:13:27	有油烟扑进屋内，窗户长期不能打开，晚上营业到零	0	0
2	360114	A0182491	经济学院体育学院变相强制	2017-06-08 17:31:20	说学校与公司签了合同，并且公司也要和我们签合同	9	0
2	360110	A110021	学院寒假过年期间组织学生去	2019-11-22 14:42:14	们的心有多难过！虽说不是强制性的，但不去不给学	0	0
2	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	，学校要求学生必须去学校安排的几个点实习，并且	0	0
2	360113	A3352352	市经济学院强制学生外出实	2018-05-17 08:32:04		3	0
2	360111	A1204455	济学院组织学生外出打工合	2019-11-05 10:31:38	一天工作十个小时以上，（晚班时间是20：30-第二	1	0
3	353426	A0098773	供销合作社在岗失业职工追	2020-01-06 10:20:31	根据《中国人民共和国宪法》和《劳动法》相关规	2	0
3	351074	A001241415	对A市参保记录的几点疑问	2019-12-19 17:46:04		0	0
4	336608	A0005623	西地省把抗癌药品纳入医保	2019-09-08 21:01:59		0	0
4	321736	A9992521	能不能提高医疗门诊报销范	2019-06-12 08:23:01		1	0
5	343985	A108051	市能否设立南塘城轨公交站	2019-10-31 21:19:59	技学院新校区，南塘小学，A市一中城南中子等等，。	0	0
5	286572	A23525	地铁2#线在梅溪湖CBD处增	2018-10-27 15:13:26	下午游玩桃花岭和梅溪湖，傍晚坐2#线回家，度过愉	3	0

问题三分析过程与方法

方法：模糊评价法

因素集 U 的确定：把评判对象的各种因素 u_i 作为元素组成的集合 $U=\{u_1, u_2, \ldots, u_m\}$。

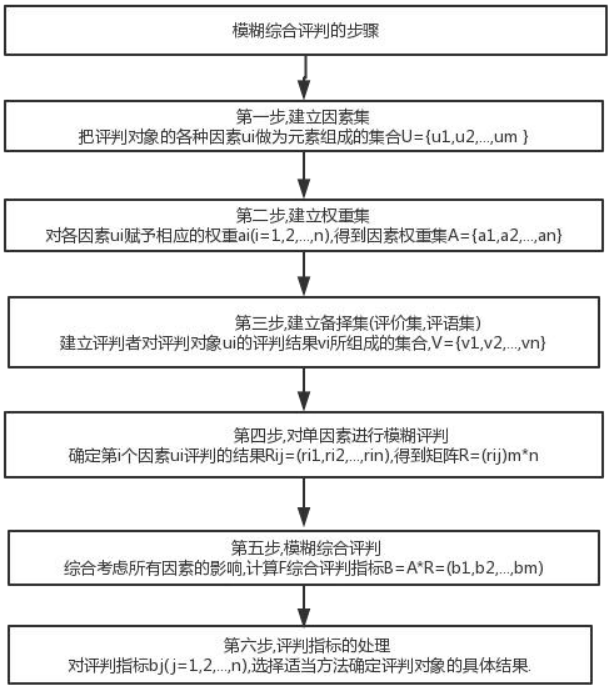
权重集 A 的建立：对各因素 u_i 赋予相应的权重 $a_i(i=1,2,...,n)$，得到因素权重集 $A=\{a_1,a_2,...,a_n\}$。

备择集 V 的建立：建立评判者对评判对象 u_i 的评判结果 v_i 所组成的集合，$V=\{v_1,v_2,...,v_n\}$。

进行单因素的模糊评判：确定第 i 个因素评判的结果 $R_{ij}=(r_{i1},r_{i2},...,r_{in})$，得到矩阵 $R=(r_{ij})_{m*n}$ 。

模糊综合评判：综合考虑所有因素的影响，计算 F 综合评判指标 $B=A*R=(b_1,b_2,...,b_m)$ 。

评判指标的处理：对评判指标 $b_j(j=1,2,...,n)$，选择适当方法确定评判对。



3.参考文献

[1] 高俊波栾翠菊王晓峰　新的关键字提取算法研究 上海海事大学信息工程学院上海海事大学信息工程学院 上海 200135　期刊 2008-02-16

[2]基于词序统计组合的中文文本关键词提取技术[2] 万静,吴凡,何云斌,李松　新的降维标准下的高维数据聚类算法　期刊 2019-05-15 14:54

[3]陈永胜 北京邮电大学　基于降维的聚类分析算法设计与实现 硕士论文 2015-12-262015-12-26

[4]赵玉娟　数据降维的常用方法分析　期刊 2019-11-21 10:30

[5]路明 湘潭大学 硕士 我国政务微博公共信息服务绩效评价研究 2016-06-01

[6]刘志英　基于模糊综合评价的场景应用分析与探讨 江西工程学院江西新余 338000 期刊 2020-04-29