“智慧政务”中的方法探讨与数据挖掘

摘要

随着时代科技的不断发展与社会经济的进步，以及社会民主化的不断完善与普及，政府部门不断致力于对民意的充分了解，以便更好的为人民服务，因此针对如何完善“智慧政务”系统，让相关部门充分了解民意的研究具有充分的挑战性，针对“智慧政务”中的方法探讨与数据挖掘十分具有研究意义。

针对问题一的群众留言分类，首先进行模型构建对留言初始数据进行预处理，去重、去空，然后利用 jieba 分词对留言详情进行中文分词，对于未登录词，采用了基于汉字成词能力的 HMM 模型，以达到主次分明的情况。然后，在进行分词时对停用词进行过滤，利用 TF-IDF 算法进行对关键词进行提取，计算 TF 词频。之后利用 K-means 算法原理对留言进行具体操作的分类。通过欧氏距离计算进行最邻近规则分类，最后利用 Excel 中的排序和筛选对留言的时间进行排序。

针对问题二的热点问题的挖掘，首先建立民意热点数据采集模型，利用单因素模糊评判综合评定和 Delphi 方法建立综合热度评价指标模型，采用模糊评判，全概率评分法，热度评价指标的分级与主成分分析综合整理得出适当的评价指标。

针对问题三的答复意见评价，依据层次分析法（AHP）和几何平

均值，采用线性相关性质与几何知识结合，构造留言程度判断矩阵对其一致性的检验来构建答复意见的质量评价模型，根据模型建立一套合理的评价方案。

关键词：留言指标 MATLAB 聚类 综合分析 层次分析

目录

$1 数据挖掘目标

随着云计算，大数据，人工智能的高速发展，政府部门处理政务带来极大便利。但仍然存在部分不足，需要不断的加强与完善，以便提高相关部门的工作效率。本次建立模型、挖掘数据的主要目标是利用网络的各个民意信息收入平台系统的数据，利用 TF-IDF 算法，K-means 算法 ，欧氏距离，Delphi 方法，单因素模糊评判综合评定法，全概率评分法，指标分级，层次分析，采用线性相关性质与几何知识结合以达到如下的目标结果：

1.利用 jieba 分词对留言详情进行中文分词，TF-IDF 算法对留言进行关键词提取，计算逆向文件频率 IDF 来区分出不同的留言类别。用 TF-IDF 在过滤掉无用常见的字词后会保留剩下重要的部分。通过权重距离来聚类使文本向量化，用 K-means 算法原理，KNN 算法，以最临近原则达到最邻近规则分类，最后利用 Excel 中的排序和筛选对留言的时间进行排序，以达到最终的留言分类目的。

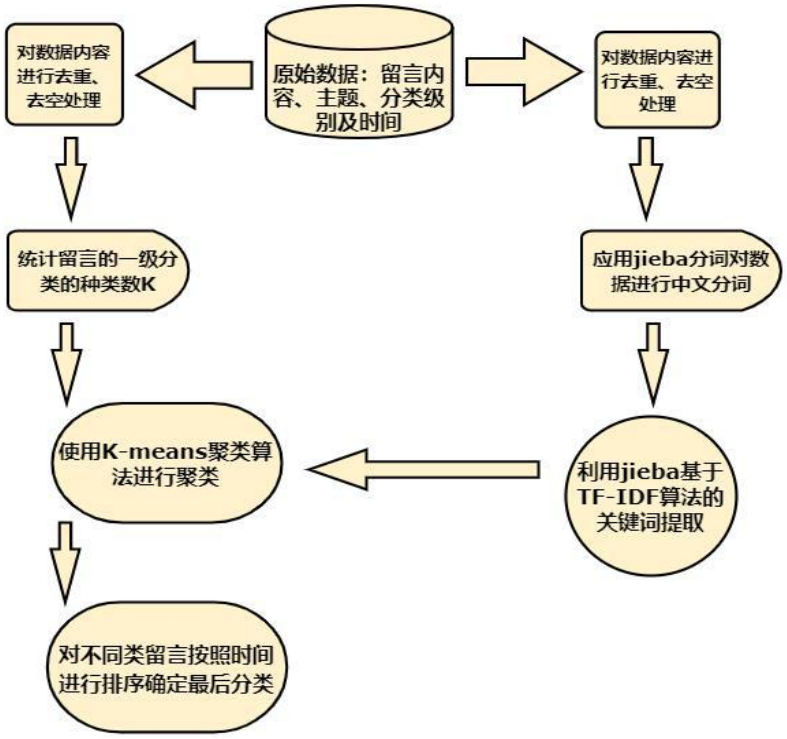
2. 采用模糊评判，全概率评分法，热度评价指标的分级与主成分分析综合整理得出适当的留言评价指标。

3 依据层次分析法（AHP）和几何平均值，采用线性相关性质与几何知识结合，构造留言程度判断矩阵对其一致性的检验来构建答复意见的质量评价模型，根据模型建立一套合理的评价方案

$2 分析方法与建模设计总体流程

$2.1 问题 1 的方法分析

2.1.1 数据处理的总体流程图



2.1.2jieba 分词对留言详情进行中文分词

为使中文分词结果更准确，在此之前，要对数据进行预处理，去重、去空，当留言中有相同内容却属不同分类时，去重只保留一项。去重、去空后保存命名为附件 2 去重。同时要判断数据是否为字符串，把文字化信息转化为计算机能识别出的机器语言。这里对已去重的附件 2 的留言详情采用 python 中的 jieba 中文分词。Jieba 分词所用方法为基于 Trie 树结构实现高效的词图扫描，构造前缀词典；将句子重新拆分，若包含特殊字符，将其分开；生成的句子中汉字重新组成的所有可能

成词情构成的有向无环图（DAG）;采用了动态规划查找最大概率路径,找出基于词频的最大切分组合,并对词性进行分析；在可能会出现新词的条件下，对于未登录词，采用了基于汉字成词能力的 HMM 模型。

2.1.3 分词时对停用词进行过滤

在中文分词过程中会出现大量的无用字词或标点，会对分词结果有一定的干扰，所以为节省空间和提高搜索效率，这里将会使用停用词表进行对分词过滤，停用词表不能自动生成，只能通过人工输入。所以我们可以根据某些词或标点出现的频率，人工输入一些常出现的停用词，形成一个停用词表，并利用停用词表过滤掉无用的部分，减少对中文分词的干扰，提高效率。

2.1.4TF-IDF 算法进行关键词提取

对留言文本进行概括摘要时，可以根据字词所处文本的位置、出现的次数来判断其重要性，可对文本提取关键词。这里将通过 jieba 分词使用 TF-IDF 算法实现关键词提取。TF-IDF 表示为词频和逆向文件频率，通过某些字词在留言文本中出现的次数来评估该字词在文本中的重要程度，若某字词出现频率高，则认为该字词具有区分不同留言分类的能力。

3.1 计算 TF 词频

计算关键词在文本中出现的次数与频率，使用公式：$tf_{ij}=\frac{n_{i,j}}{\sum_k n_{k,j}}$

（$n_{i,j}$是该字词在文件留言中出现的次数）因为每一条留言的内容不同，

长短不一致，所以选取较好计算的对象。

即：$TF=\frac{\text{在文件留言中某字词出现的次数}}{\text{该文件留言中出现次数最多的字词的次数}}$

1.1. 计算逆向文件频率 IDF

因为文件中只有单个文件及留言文本，所以应建立一个语料库，语料库中包含大量字词，将语料库中的总数除以包含某字词的留言条数，若 IDF 值越大，说明该字词能很好地区分出不同的留言类别。

使用公式：$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$

（$|D|$ 是语料库中所含字词的总数，$|\{j:t_i\in d_j\}|$表示包含字词t_i的文件数目（即$n_{i,j}\neq 0$ 的文件数目）即：$IDF=\log(\frac{\text{语料库的文件总数}}{\text{包含某字词的留言条数}+1})$

3.3 计算 TF-IDF

TF-IDF 在过滤掉无用常见的字词后会保留剩下重要的部分。所以 TF-IDF 值可以直观的看出当某些字词出现次数最多的就为文本关键词。

使用公式：TF-IDF=TF*IDF（词频×逆向文件频率）

3.4 文本向量化并计算权重

1. 因为留言为文本形式，计算机不能对文本形式的数据进行处理，所以首先要转换成计算机可以处理的形式，将其转化为向量。每一个字词转化后会被赋予上数字，该数字为向量所占维度，并计

算其权重，通过权重距离来聚类。使用公式：$W = TF * IDF$

若文本中包含该字词的留言条数过多，可以乘上一个常数 N 来增大权重值，不会因权重值过小而难以判断，即为：$W = TF * N * IDF$

2.1.5 留言的分类

1.1K-means 算法原理

在找出关键词后，对数据采用 K-means 聚类算法，根据关键词对留言进行分类。。易燕飞学者[1]研究中说明“K-means 算法是从所有的样本对象中选择出 K 个元素作为最开始的聚类目标，之后按照规则算法要求，对剩下元素和目标中心元素之间距离进行分析，根据计算的数值确定元素和中心元素之间的关系。”

算法步骤：

（1）对已给数据随机选取 k 值（k 可为所分类别数）作为聚类中心，因为附件 1 中已把留言分为 7 类，所以 K=7。首先，假设样本中包含 n 个点，即 S=$\{x_1, x_2, \cdots\cdots, x_n\}$，其中将已给数据分为 K 类，即为 A=$\{a_1, a_2, \cdots\cdots, a_k\}$，每一个类别即为一个$a_k$。

（2）计算每个点分别到 K 个聚类中心的距离，将每一个点分到最近的聚类中心，相近的点和聚类中心可行成了 k 个簇。使用夹角余弦相似度（Cosine）计算距离，当计算出的夹角越小时，说明两个点越相似。通过向量运算，两个 n 维向量$(b_1, b_2, \cdots\cdots, b_n)$和$(c_1, c_2$

，……，c_n)　　可使用的公式：$\cos\frac{b_n c_n}{|b_n||c_n|}(n=1,2,\cdots\cdots,n)$

（3）重新计算每个簇的质心（均值），使用误差的平方和（SSE）作为度量标准来比较那个簇更好。S=$\{x_1，x_2，\cdots\cdots，x_n\}$，$x_n$为样本中所包含点；数据分类 A=$\{a_1，a_2，\cdots\cdots，a_k\}$在每个分类$a_k$均有一个类别中心$\xi_i$；一共有 k 类所以$C_j$ 为第 j 个簇，c_j为每一簇的质心，所以计算两点间的欧氏距离（dist）

即公式为：$\mathrm{dist}(c_i)=\sqrt{\sum_{x_n\in a_k}^{n}(x_n-\xi_k)^2}$

由此可计算 SSE，$\mathrm{SSE}=\sum_{j=1}^{k}\sum_{x\in c_j}\mathrm{dist}(c_i，x)^2$

若 SSE 的值越小，说明该点越接近所属质心，聚类效果也会更好。

（4）最后为保持准确性，重复以上步骤，直到质心的位置不再发生变化或者达到我们所设定的迭代次数。

2.1.6 最邻近规则分类

在文本聚类后，利用 KNN 算法以最临近原则，根据已知聚类来计算，对未知未分类的样本可直接进行处理。首先选择参数 K（在实际中选取较小的值更容易分类），通过欧氏距离计算

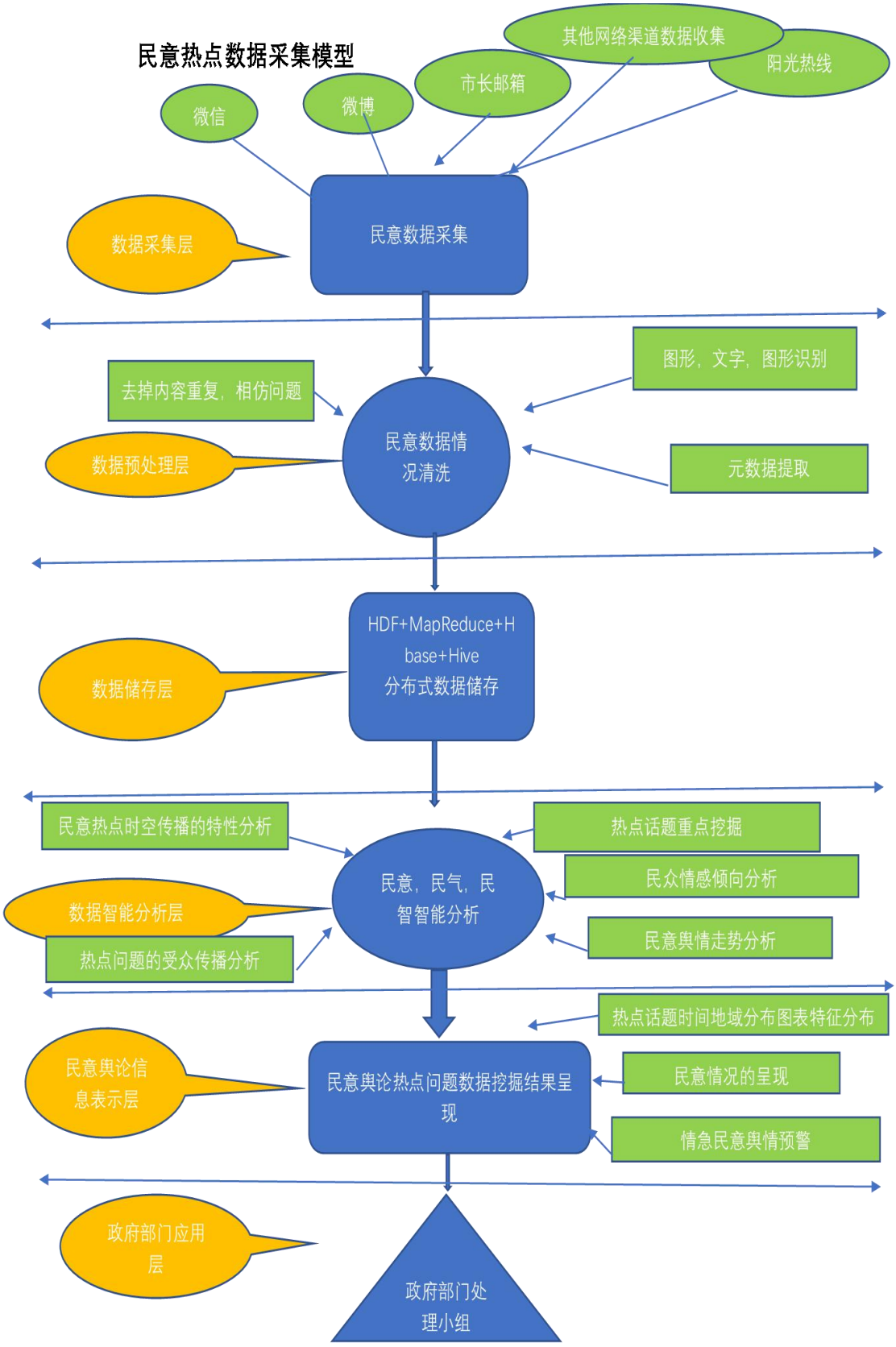
2.1.7 时间排序

在题目所给数据中，有不同类别的一级标签，以及大量的留言文本内容。且需要对留言内容进行分类派至相关部门处理，存在工作量大、效率低，差错率高等问题，因不同时段都可以进行留言，为优先解决

发送时间较长的留言，更快处理大量信息，因此对数据进行时间排序处理。把经过处理分类后的留言按照时间发送的先后进行排序。使用 Excel 中的排序和筛选对留言的时间进行排序，按照时间先后顺序依次解决。

$2.2 问题 2 的方法分析

2.2.1 民意热点数据采集模型


民意热点数据采集模型
其他网络渠道数据收集
阳光热线
市长邮箱
微博
微信
民意数据采集
数据采集层
图形，文字，图形识别
去掉内容重复，相仿问题
民意数据情况清洗
数据预处理层
元数据提取
HDF+MapReduce+Hbase+Hive分布式数据储存
数据储存层
民意热点时空传播的特性分析
热点话题重点挖掘
民众情感倾向分析
民意，民气，民智智能分析
数据智能分析层
民意舆情走势分析
热点问题的受众传播分析
热点话题时间地域分布图表特征分布
民意舆论信息表示层
民意舆论热点问题数据挖掘结果呈现
民意情况的呈现
情急民意舆情预警
政府部门应用层
政府部门处理小组

2.2.2 问题 2 的模型方法与分析

为了加强民意工作的合理化，权威性，顺畅性，我们打算利用数理统计的方法，即数学方法将整个民意回复信息系统进行不断整改与优化提炼，以此让相关部门充分了解到广大群众反馈的问题。为此，我们特地的就有关的多指标综合评价与权重系数的选择进行探讨，深入构建民意留言评价指标模型。

2.2.3 留言合理化的热度评价指标模型

影响民意留言是否成为热度留言的因素，主要有，同一时间内，相同地区的提到指数，留言用户，留言内容的相似性，点赞数，反对数，中立态度，浏览量等，所以该因素影响集为：

W={同一时间内，相同地区的提到指数，留言内容的相似性，点赞数，反对数}

对于留言热点问题的评判，由于相关因素的影响权重不一样，因此对于不同的影响因素，权重也不同，根据相关系数法，即

$$r_{ij}=\frac{\sum_{a=1}^{p}(x_{ia}-\bar{x}_i)(x_{ja}-\bar{x}_j)}{\sqrt{\sum_{a=1}^{p}(x_{ia}-\bar{x}_i)^2\sum_{a=1}^{p}(x_{ja}-\bar{x}_j)^2}}$$

-1 小于等于 r_{ij} 小于等于 1，于是 R=（r_{ij}）,其中 r11=r22=r33=.....=r_{NN}=1,则可以·根据 R 对 N 个样品进行分类。由此模型计算各个因素之间的相关情况，由此来确立模型之间的权重关系。

根据题目提供的附件 3 给予的数据，我们利用 Excel 电子表格进

行排序并进行适当的示范性处理得到如下形式：

	A	B	C	D	E	F	G
1	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
2	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	公司签了合同，并且公司也要和我们签	9	0
3	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	都不知道！学校很小但是这几年只顾扌	3	0
4	286572	A23525	A市地铁2#线在梅溪湖CBD处增设一	2018-10-27 15:13:26	桃花岭和梅溪湖，傍晚坐2#线回家，	3	0
5	289408	A0012413	市人才app上申请购房补贴为什么通不	2018-11-15 16:07:12	也购买了灵活就业人员养老保险和医	0	0
6	304503	A011244	市什么时候能实行独生子女护理假？	2019-02-26 15:22:05	生子女多，照顾四个老人真的力不从心	0	0
7	313964	A108906	申请驾驶证期满换证，一个星期了都	2019-04-26 15:28:42	申请的，都快一个星期了都没人受理	0	0
8	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	求学生必须去学校安排的几个点实习，	0	0
9	316619	A235259	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	全覆盖的城区建成。看消息介绍，5G	0	0
10	319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	，雨雪天气后过道全部是水和雪，而	0	0
11	321736	A9992521	A市能不能提高医疗门诊报销范畴	2019-06-12 08:23:01	又不让，小孩体弱多病各种开支，长此	1	0
12	323034	A012414	业服务收费标准应考虑居民的经济承	2019-06-19 17:46:24	拆墙垃圾清运费按不超过30元/平方米	0	0
13	323149	A1241141	给K3县乡村医生发卫生室执业许可证	2019-06-20 20:38:47	新村的是证件下来啊。有些老村医反映	0	0
14	360107	A0283523	魅力之城小区一楼的夜宵摊严重污染	2019-07-21 10:29:36	觉得要维护社会和谐稳定，合法维权，	3	0
15	360101	A324156	A5区劳动东路魅力之城小区油烟扰民	2019-07-28 12:49:18	清洗也没有。每天油烟直排。熏死树	4	0
16	360108	A0283523	之城小区一楼的夜宵摊严重污染附近	2019-08-01 16:20:02	觉得要维护社会和谐稳定，合法维权，	6	0
17	360104	A012417	力之城商铺无排烟管道，小区内到处	2019-08-18 14:44:00	进出都搞得业主一身油烟味，而且每天	0	0
18	360106	A235367	城小区底层商铺营业到凌晨，各种噪	2019-08-26 01:50:38	闻不说还严重影响健康。大家对此深恶	0	0
19	360105	A120356	之城小区一楼被搞成商业门面，噪音	2019-08-26 08:33:03	住，影响我们的晚年生活。架空层被侵	1	0
20	360109	A0080252	力之城小区底层门店深夜经营，各种噪	2019-09-04 21:00:18	斥着吆喝声、拼酒声、炒菜烧烤的锅铲	0	0
21	360100	A324156	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	人。一天24小时都是烟。请政府关闭	3	0
22	336608	A0005623	希望西地省把抗癌药品纳入医保范围	2019-09-08 21:01:59	贵。明明有药可治，就是买不起药品。	0	0
23	360102	A1234140	劳动东路魅力之城小区底层餐馆油烟	2019-09-10 06:13:27	进屋内，窗户长期不能打开，晚上营	0	0
24	337458	A078325	能否分层单独补交超面积地款？	2019-09-16 18:48:29	土地，得知该房占地属划拨用地而且占	0	0
25	360103	A0012425	劳动东路魅力之城小区临街门面烧烤	2019-09-25 00:31:33	为烧烤夜宵更加扰民，油烟24小时晕列	1	0
26	342119	A090900	咨询移动通信业务问题	2019-10-18 23:52:58	，导致10月份套餐生效，当我要求改回	0	0
27	343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	交区，南塘小学，A市一中城南中子等	0	0
28	360111	A1204455	市经济学院组织学生外出打工合理吗	2019-11-05 10:31:38	十个小时以上，（晚班时间是20：30-	1	0
29	360110	A110021	经济学院寒假过年期间组织学生去工厂	2019-11-22 14:42:14	有多难过！虽说不是强制性的，但不去	0	0
30	351074	A001241415	对A市参保记录的几点疑问	2019-12-19 17:46:04	大字，真是讽刺。以上，我四问A市社	0	0
31	353426	A0098773	4县供销合作社在岗失业职工追缴社保	2020-01-06 10:20:31	中国人民共和国宪法》和《劳动法》相	2	0
32							
33							

以表格部分数据为例同一时间内

我们根据综合评定和 Delphi 方法，得出主要的影响因素有点赞数，反对数，相同地区的提到指数，其他因素如同一时间内，浏览量，留言用户，留言内容的相似性不是重点。

所以对于各个留言热点影响因素的权重可确定为：

H=(0.05 0.4 0.25 0.25 0.05)

1.建立留言备择集

因为综合评判是为了挖掘出热点留言问题的热度指标，总的评判结果是各个留言问题的热点程度，根据热点问题的性质程度，以及题目的目标所求，得到如下的备择集的情况为：

M={留言热度特高 留言热度较高 留言热度一般 留言热度较差 留言热度很差}

2.单因素模糊评判

为了方便性，以及问题本身的简便化，单独地从留言以上的各个因素出发，对留言热度进评判。选取 m 个数据，对涉及到的各个指标的留言热点因素进行打钩，如对相同地区的提到指数，有 m1 个项目达到热度特高，有 m2 个项目达到热度较高，有 m3 个项目达到热度一般，有 m4 个项目达到热度较差，有 m5 个项目达到热度很差，其中 m1+m2+m3+m4+m5=m,则相同地区的提到指数的热度因素评判为

U_{1i} =(m1/m,m2/m,m3/m,m4/m,m5/)

假设用此方法得到了单因素评价矩阵为

S=[0,0.1,0.3,0.1,0;0.2,0.1,0,0,0.5;0.5,0.3,0.1,0.4,0.3;0.3,0.5,0.6,0.5,0.2;0,0.1,0.3,0.1,0]

S =

0.2000	0.1000	0	0	0.5000
0.5000	0.3000	0.1000	0.4000	0.3000
0.3000	0.5000	0.6000	0.5000	0.2000
0	0.1000	0.3000	0.1000	0

3.模糊综合评判

选取模型：M（∧，∨） $bj = \bigcup_{i=1}^{m}$ (ai∧rij) (j=1,2,3,…,n)，这里的“∧”、“∨”表示取大取小运算。

则由留言类型的程度中各个因素数据之间的关系，利用 MATLAB

数学软件进行运算得如下的权重结果

L=H*S= H=(0.05 0.4 0.25 0.25 0.05)*

S=[0,0.1,0.3,0.1,0;0.2,0.1,0,0,0.5;0.5,0.3,0.1,0.4,0.3;0.3,0.5,0.6,0.5,0.2;0,0.1,0.3,0.1,0]

= 0.2800 0.2500 0.2050 0.2350 0.3250

4.留言评判指标的处理

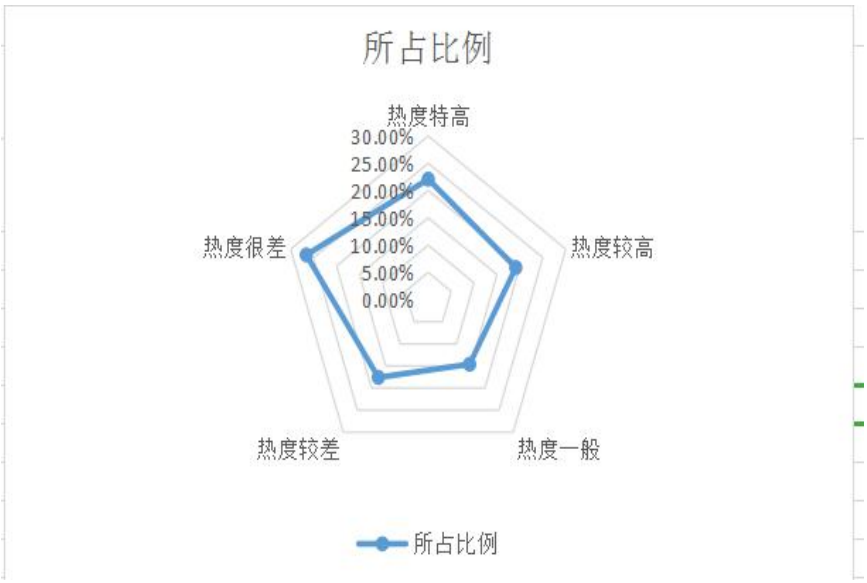
根据整个模型的构想方法，采用模糊分布法，将上述的留言各个因素指标进行归一化处理，由此可得：

0.2800 + 0.2500 + 0.2050 + 0.2350 + 0.3250=1.295

遍除每一个指标得

S&=(0.221,0.191,0.146,0.176,0.266)

根据这一评判结果表明：在这一次热点问题筛选过程中，22.1%的项目“热度特高”，19.1%项目“热度较高”，14.6%的项目“热度一般“，17.6%的项目“热度较差”，26.6%的项目“热度很差“。



2.2.4 留言热度指标的评分

对此，我们采用全概率评分法。设 Q k 为第 K 级留言分类，K=1,2,3,4，.....K，W z 为第 Z 个指标，且 W1,W2,W3,W4,.....,W n,互不相容。又设各个留言指标的重要程度之比分别为 W1:W2:W3.....:WK=h1:h2:h3:....h k,

则各个指标的概率分别为 P(WZ)=h z/N, z=1,2,3,4.....K

假设令 G k z 表示为第 z 个指标下的第 k 个留言测定值，令 B z 表示为第 z 个留言指标下的各个试验结果之和，则有

B z= $\sum ni$ =1G k z k=1,2,3,4....n ,z=1,2,3,4,.....k

容易得知 P（QK/WZ）=G k z/B z

由全概率公式的定义和全概率公式得：

P(Q k k z=1)= $\sum p(Wz)p(Qk/Wz)$ （k=1,2,3,4,.....k;z=1,2,3,4,.....,k）

由概率论的性质与热点民意指标的结合，容易得知，当公式分越小或越大时留言热点的模型体系就越优化。

2.2.5 留言热度评价指标的分级与指标简化

根据上述的留言指标中的热度指标对留言热度进行分级，关键是建立确定留言热度评价的主要指标，建立综合评价模型，从而进行分级。根据上述建立的模型方法，确定对留言热度影响较大的理化指标，将留言数据进行性标准化处理，采用主成分分析法，找出留言数据处理的主成分，以此不断简化评价指标由计算的得到的主成分作为新的评价指标，建立对留言热度分级的综合评价模型，对留言热度情况进行分级，容易知道对于留言指标的选取太多，会对整体信息的整合情况展现出不准确，主观性强，信息重复，不易排序规划等问题，所以采用主成分分析法是为了将这些现象最小化，将多指标问题化为综合指标问题。所以对于留言热度评价主要定了 6 个评价指标，分别为（留言热度，同一时间内，相同地区的提到指数，留言内容的相似性，点赞数，反对数）。

2.2.6 热度评价指标的分级与主成分分析

将上述的留言指标数据用 excel 排序分类取出,然后将 6 个指标的计算值进行标准化处理。令 $x=(x1,x2,x3,x4,x5,x6)^T$,建立留言数据标准化模型

$$X = \frac{x-\bar{x},}{\sigma}$$

其中 x 表示每个评价指标所对应的由全概率评分法对应的每个指标的分值，$\bar{x}$表示其对应的六个指标分值的平均值，σ表示对应指标分值的标准差，利用数学软件 MATLAB 的 X=score（x）命令就可以得

到分值对应的标准化留言分值矩阵，根据上述的数值计算，利用 v 数矩阵。接下来同样利用 MATLAB 命令中的【vac val】=eig(R)计算留言指标相关矩阵 R 的特征值，特征向量，特征值的贡献率。其中贡献率为特征值除以指标个数。接下来计算 6 个主要留言指标的累计计算率，取主成分中的前几个累计贡献率超过百分之八十的主要成分，然后计算这几个留言主成分的载荷。

2.2.7 留言热度评价指标分的评价系统求解

利用留言各个指标主成分分析可以计算出每个留言样品指标的主成分值，将每个留言样品的所有主成分综合加起来就是这个留言样品的综合分数。建立样品关于 6 个留言主成分分值的综合留言评价模型为 $Q = XV = (Q_i\dot{j})$6x6

$$Q_i = \sum_{j=1}^{6} Q_i j, \text{i=1,2,3,4,5,6}$$

其中 X 表示 6 个留言热度指标标准化后的矩阵结果，Q 表示为 6 个留言指标主成分的各个留言指标的载荷，*其中* Q_i表示为第 i 留言指标样品的综合分值。根据上述的分析与计算可以将留言热度分为四个等级：火，较火，一般，凉，进而可以对留言热度指标的分数进行分级比较，不断优化结构，进行对比，合理化分析，观测整体结果的效果。

$2.3 问题 3 的方法分析

影响答复意见的质量可从答复的相关性、完整性、可解释性等角度评价分析，而影响其主要因素有：同一类问题，留言次数；留言用

户；答复意见的相关性、完整性、科学化、可解释性；答复时间等。M={同一类问题，留言次数；留言用户；答复意见的相关性、完整性、科学化、可解释性；答复时间}

对于答复意见的质量评价，其与答复的相关性、完整性、可解释性联系比较紧密，所以其答复的相关性估算法如下：

记答复的相关性、完整性、可解释性分别为 X、Y、Z，其观测数据为记为（x_i，y_i，z_i，）i=1,2,3…n。以上三个影响因素的线性方程的一般式为：ax+by+cz+d=0

将相关方程法式化则有：$\frac{(ax+by+cz+d)}{\pm\sqrt{a^2+b^2+2^2}}=0$

其中 A=cosα=a/$\pm\sqrt{a^2+b^2+c^2}$

B=COSβ=b/$\pm\sqrt{a^2+b^2+c^2}$

C=cosθ=c/$\pm\sqrt{a^2+b^2+c^2}$

则容易得到 AX+BY+CZ+D=0 且 X,Y,Z 的系数满足关系式：A^2+B^2+C^2=1,有数学专业知识知，答复意见的关系性质观测点（xi,yi,zi）到相关平面的距离为 $D_i=|AX_i+BY_i+CZ_i+D|$,i=1,2,3,……n。所有答复意见的关系性质观测点到相关平面的距离平方之和为：

$$Q=\sum_{i=1}^{N} di^2=\sum_{i=1}^{N}(Axi+Byi+Czi+D)^2$$

在线性约束条件下：A^2+B^2+C^2=1，求得 Q 的值最小，从而可以求得 A,B,C,D,由此可以得到答复意见的关系性质估计的相关平面进行直观的观测。

依据附件 4 所给的多组数据可看出，当前社会主要存在高层避重就轻、故受利益；商家缺乏标准、疏于经营；人民安保等问题，从而

可对以后反馈的问题进行归类。我们利用 Excel 表格进行筛选得到如下形式

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
2549	A00045581	A2区景菁华苑物业管理有问题	2019/4/25 9:32:09	业公司却以交20万保证金，不	年管理费，在业主大会结束	2019/5/10 14:56:53
2554	A00023583	A3区潘楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	后面的生意带来很大影响，里面	换填，且换填后还有三趟雨	2019/5/9 9:49:10
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	同时更是加大了教师的工作压	聘任教职工要依法签订劳动	2019/5/9 9:49:14
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗？	2019/4/24 15:07:30	落户A市，想买套公寓，请问男	岁以下（含），首次购房后	2019/5/9 9:49:42
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭坡岭”的问题。公交站点的		2019/5/9 9:51:30
2759	A00077538	A3区含浦镇马路卫生很差	2019/4/8 8:37	每把泥巴冲到右边，越是上下	中没有说明卫生较差的具体	2019/5/9 10:02:08
2849	A000100804	A3区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	为老社区惠民装电梯的规范性	人民政府办公室下发了《关	2019/5/9 10:18:58
33970	A000100240	B7县二中违规乱补课	2018/8/12 10:56:10	什么原因，又开始补课。简直	完成教学任务，但须征得	2018/8/20 9:25:34
33978	A00044584	市601小区钻石新村20栋楼下快乐休闲网吧扰民	2018/8/8 13:15:50	网吧空调污水直接排放到过道	当事人双方以及茨菇塘办	2018/8/17 9:49:43
33984	A00091054	咨询在B市用临时身份证办理出生证明的问题	2018/8/3 21:26:53	少钱，急着办落地险，希望能	您拨打12345市长热线，可	2018/8/7 9:13:42
34239	A00084182	B市泰民米粉厂存在严重安全隐患	2018/2/3 16:27:45	坡安全，而当地政府和铁路局	二、厂房并没有加长，老	2018/4/16 11:33:45
34249	A00057771	B市港口街乱象何时了	2018/1/25 12:01:13	设路口的江天宾馆，把自己的秩序，对违章摆摊当事人进		2018/4/12 10:10:40
34252	A00014375	区栗雨东路和王家坪路的十字路口没有任何信号	2018/1/24 11:12:25	王家坪小学，小孩上学必经之	坪路与响合路交叉口因道	2018/4/12 11:09:34
35462	A00050959	请问B市西环线两侧辅道何时能全线贯通	2019/12/2 19:34:28	全线贯通，城市不断发展，快部分及西环线东辅道目前正		2019/12/10 15:16:08
35467	A00044412	请求农行B市分行公开年收入	2019/11/28 15:41:22	说假话？》等文章赫然登在	被格特殊办理内退或病退	2019/12/2 16:37:17
35479	A00079768	B2区泉中路车辆乱停乱放，占用人行横道	2019/11/19 9:32:36	门，交警部门给的答复是该道	方案，图纸已经出来了。区	2019/12/2 16:43:23
35492	A00057180	三一歌雅君路口实在太危险了，需红绿灯和摄	2019/11/5 21:51:41	，希望交管部门重视，重视行	增长较大。关于该路口的交	2019/11/12 14:53:48
35798	A000105942	咨询B市公摊精装修费用问题	2018/12/5 23:49:06	题，我市是否有给出解决方案	营者应当对每套商品房进	2018/12/21 9:25:42
35801	A00010143	B4区金锄社区的路面硬化高低不平易积水	2018/12/4 10:09:51	水。这边的路面会像4栋那样铺	都会铺设沥青路面。2.关	2018/12/21 9:27:31
35812	A00037449	能否在上下班高峰期增开B市63路公交车班次	2018/11/21 14:43:59	一班，目前，下午5点至6点不均匀，乘客候车时间增长		2018/11/23 10:30:56
35818	A00098677	B市男职工生育险其配偶能报吗？	2018/11/9 14:58:05	孕了，她一直没有买过保险，现	印件（生育证编号1开头的	2018/11/15 9:42:34
37467	A000100395	反映B9市醴官公路超载问题	2019/1/4 16:10:28	好两台车的宽度，而且已经到	、街道办事处，各相关单	2019/1/9 16:01:29
37474	A00031527	请B市农村户口村民可以一次性补缴养老保险吗？	2018/12/25 14:41:15	了，以后想有个保障，想交政	元。2014年至今，每年的缴	2018/12/30 17:41:59

依据层次分析法（AHP）和几何平均值，得出主要因素有答复的相关性、完整性、可解释性，答复时间的快慢程度，其他因素如同一类问题，留言次数；留言用户。

所以影响答复意见的质量的因素所占比例可确定为

P={0.25　0.25　　0.25　　0.1　　0.1　　0.05}

2.3.1 对答复程度进行分类

从回复的程度来看，大部分回复的能做到系统的回复了当时人的疑惑，解释可行度高，但也有一部分没有正面回答当事人的问题。

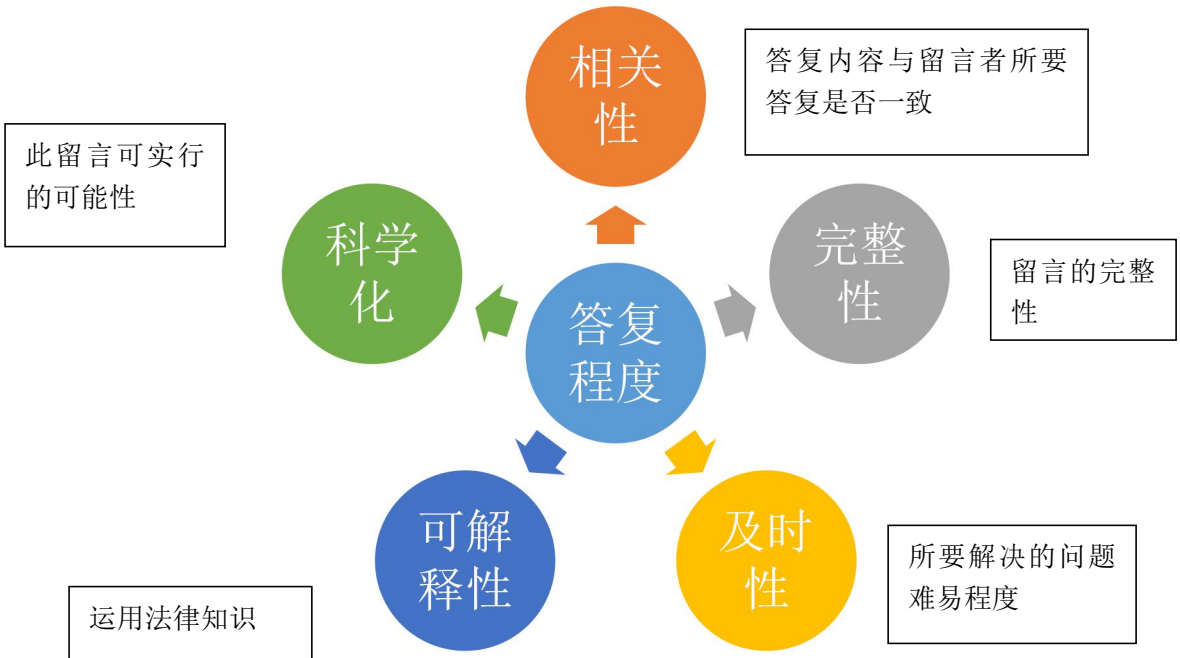
所以，A={相关性紧密、相关性弱、完整性高、完整性低、可解释性强、可解释性弱}

2.3.2 对答复种类进行分类

通过对数据进行统计，可主要分为九大民生问题，主要为学生补课、噪音扰民、就医医保、农民工工资、二胎及独生女、住房拆迁、教职

工、环境问题、交通安全。

对回复程度依据层次分析法（AHP）和几何平均值进行评估

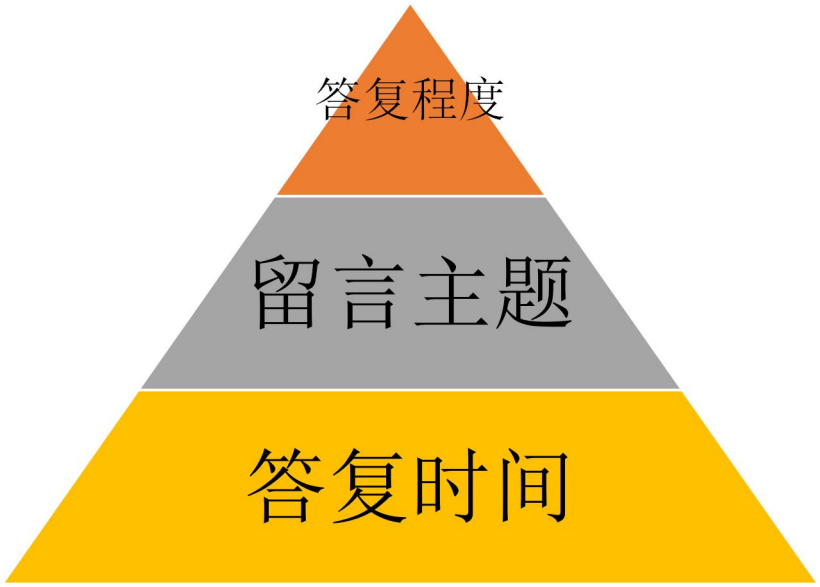


$$J=\frac{1}{m}\sum_{i=1}^{m}\left\|\chi_{-}^{(i)}u_{C}(i)\right\|^{2}$$

2.3.4 层次分析法

1.建立层次结构模型

将答复程度、留言主题、答复时间按它们之间的相互关系分为 TOP、Interlayer 和 Lowest layer，依此绘制出层次结构图。



2.构造留言程度判断矩阵

如果对留言程度堆在一起比较计较困难，此处我们采取两两比较的方法，更加精确的选出其程度偏向，其中 Qmn 为因素 m 与因素 n 重要性比较结果，按表所列出的 9 个重要性等级及其赋值。按两两比较结果构成的矩阵称作判断矩阵。判断矩阵具有如下性质：

Qmn=1/Qmn

判断矩阵元素 Qmn 的标度方法如下：

因素 m 比因素 n	比较值
m=n	λ1
m≈n	λ2
m>n	λ3
m>=n	λ4
M 远大于 n	λ5
邻近值之差	λ2-λ1、 λ3-λ2、 λ4-λ3、 λ

	5-λ4

3.对其一致性的检验

我们知道，想要对留言程度判断是否一致，就要检验其所构造出的矩阵的最大特征根λmax 的特征向量，经过 NormaliZation 后，将此时的值记为 D。D 代表多个程度的因素中某因素相对重要性的排序权值，此时，n 阶一致阵的唯一非零特征根为 n；n 阶正互反阵 A 的最大特征根λ>=n， 当且仅当λ=n 时，A 为一致矩阵。

此题中，λ连续的依赖于 Qmn，

1) 若λ 远大于 n，则 A 的不一致性越严重，此时我们需要用 Consistency index （CI）计算，CI 与一致性成反比，即 CI 越小，说明一致性越大。
2) 用最大特征值对应的特征向量作为被比较因素对上层某因素影响程度的 Weight vector（权向量），其不一致程度越大，引起的判断误差越大。故而可以用 λ-n 数值的大小来衡量 A 的不一致程度。定义一致性指标为：

$$\mathrm{CI}=\frac{\lambda-\mathrm{n}}{\mathrm{n}-1}$$

λ=n，全部包括；λ≈n，有几个程度不包括；λ远大于 n，只包括一个程度指标或者一个都不包括。

为衡量 CI 的大小，引入随机一致性指标 RI：

$$R\bar{I}=\frac{CI_1+CI_2+\ldots+C^{I}n}{n}$$

其中， RI 和判断矩阵的阶数有关，一般情况下，矩阵阶数越大，则

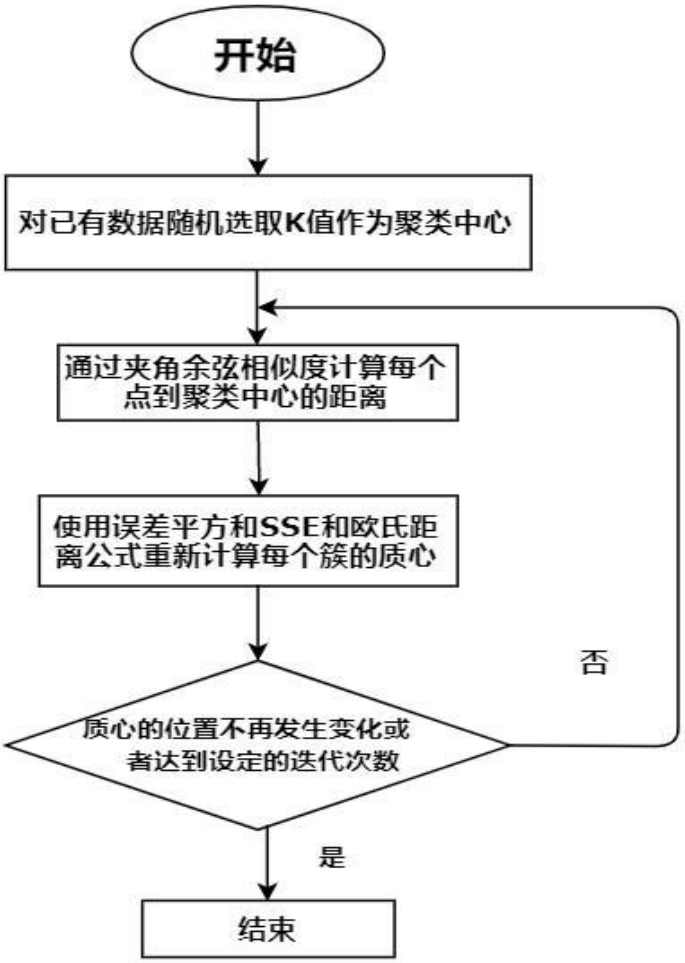
出现一致性随机偏离的可能性也越大，其对应关系如表：

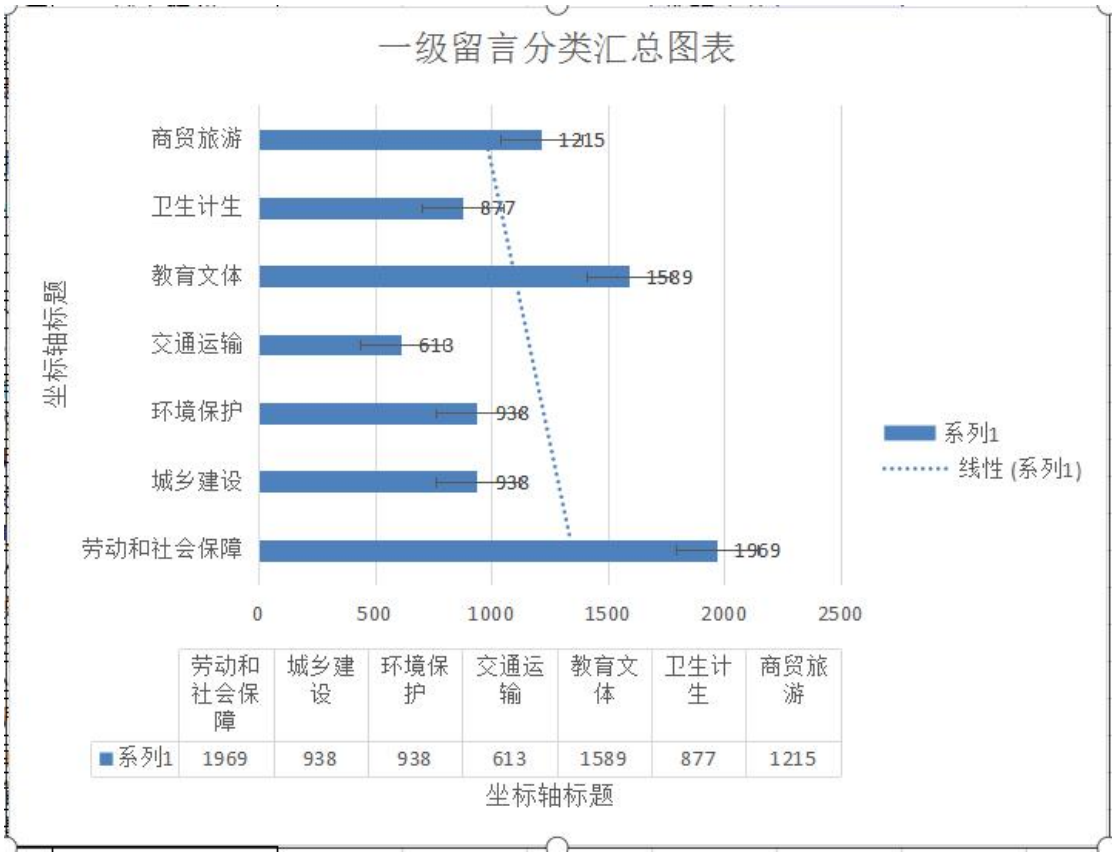
矩阵阶数	1	2
RI	0	0

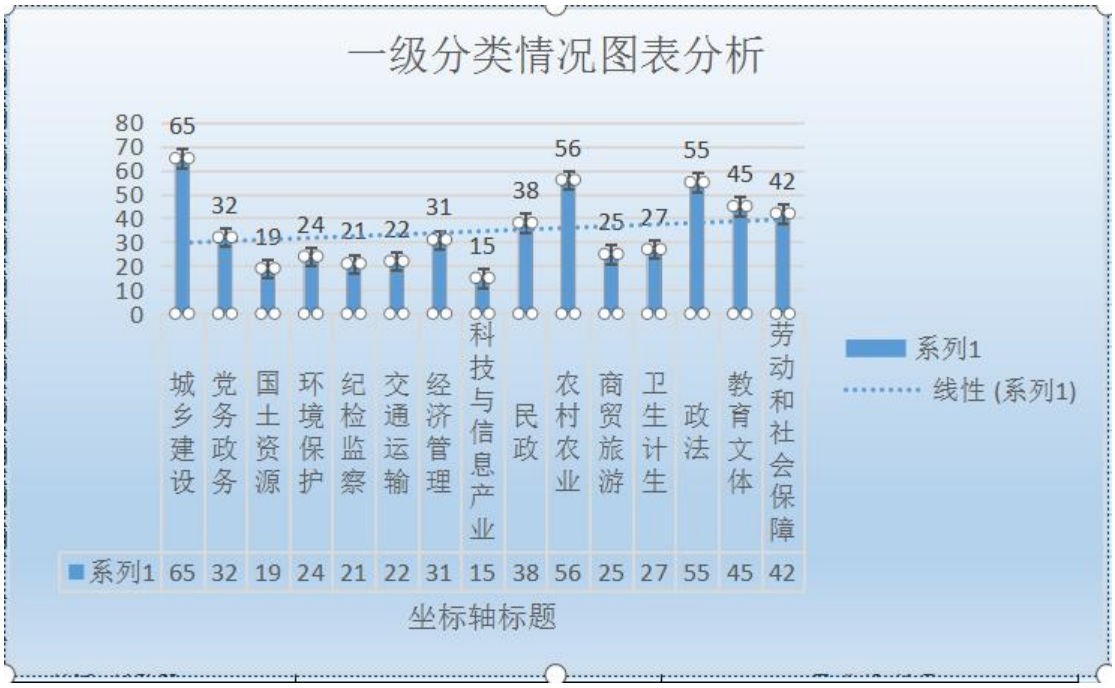
$3 数据模型构建处理结果情况

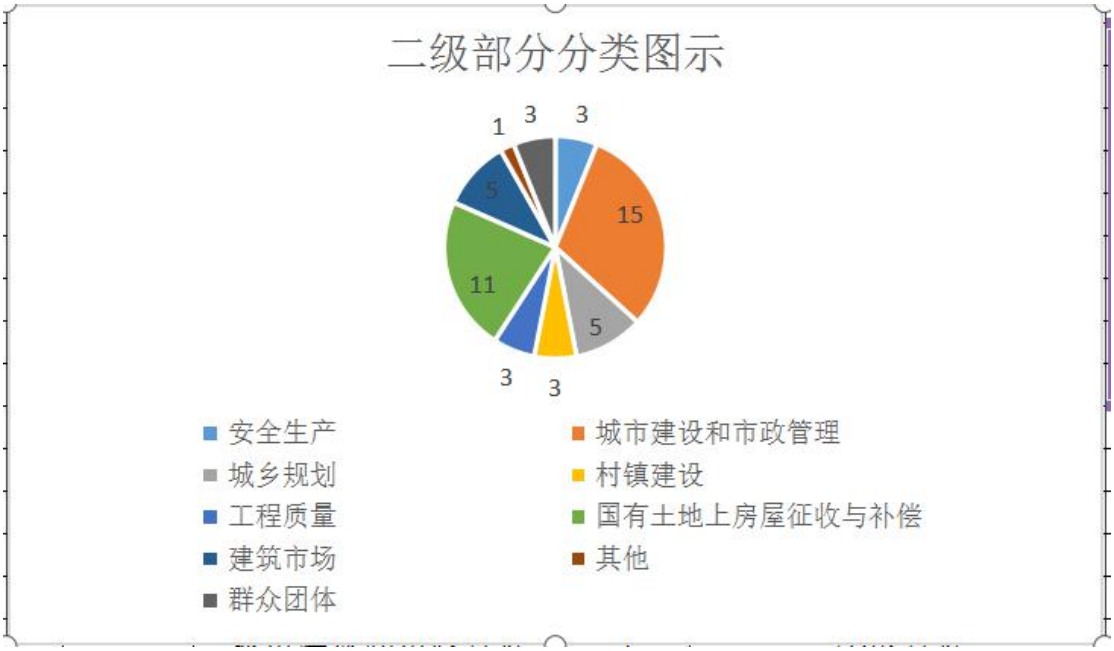
$3.1 问题 1 结果情况

算法流程图



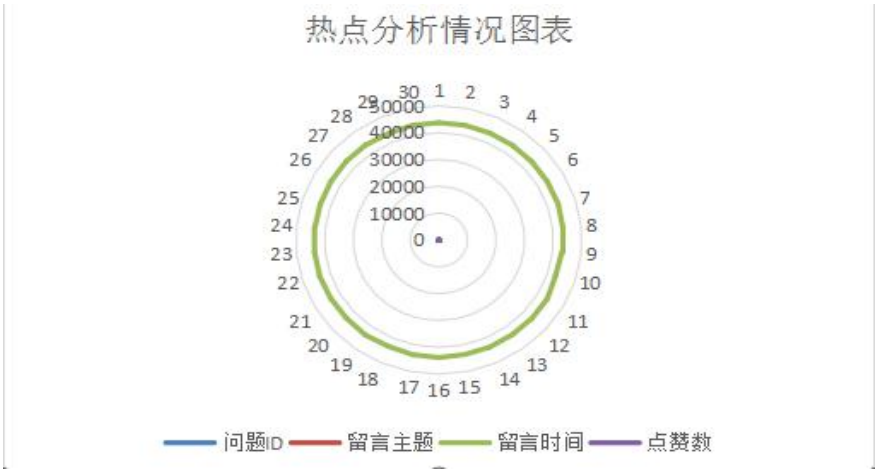

一级留言分类汇总图表
商贸旅游
卫生计生
教育文体
交通运输
环境保护
城乡建设
劳动和社会保障
坐标轴标题
系列1
线性 (系列1)
0
500
1000
1500
2000
2500
劳动和社会保障 城乡建设 环境保护 交通运输 教育文体 卫生计生 商贸旅游
系列1 1969 938 938 613 1589 877 1215
坐标轴标题


一级分类情况图表分析
80
70
60
50
40
30
20
10
0
65 32 19 24 21 22 31 15 38 56 25 27 55 45 42
城乡建设 党务政务 国土资源 环境保护 纪检监察 交通运输 经济管理 科技与信息产业 民政 农村农业 商贸旅游 卫生计生 政法 教育文体 劳动和社会保障
系列1
线性 (系列1)
系列1 65 32 19 24 21 22 31 15 38 56 25 27 55 45 42
坐标轴标题

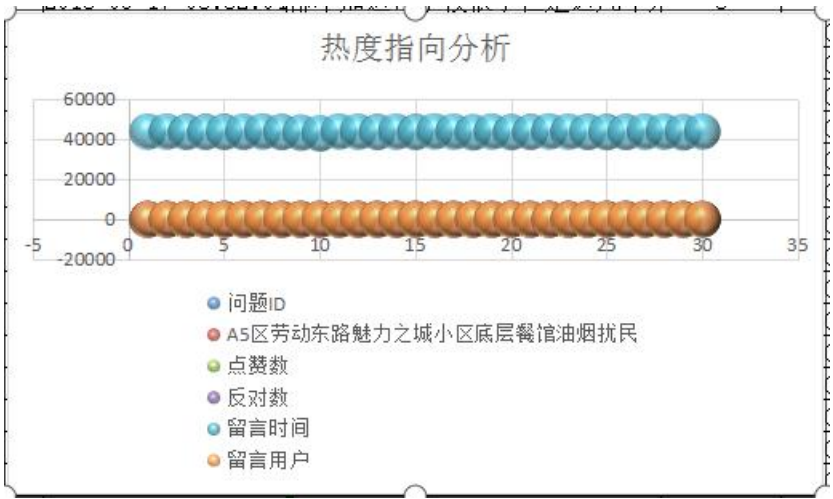


$3.2 问题 2 结果情况

图格简单明了的指出各个相关因素的情况



留言热度指向的关系通过图表也比较充分的表现出来。



热点问题表与热点明细表请参考文件

$3.3 问题 3 结果情况

3.3.1 留言评价方案构成

评价方案的实现

1.评价方法

（1）对于相关性，看其相关系数，越接近 1，表明相关程度越紧密，即我们常说的所问即所答。

（2）对于及时性，对于难题，超过三天未及时回复，则说明不及时，对于简单的问题，超过一天未回答，则说明回复不及时。

（3）对于完整性、可解释性、科学化，看其中 λ 和 n 的值，进行比较，若 λ=n，则说明几个答复意见指标全部包括； λ≈n，有几个指标不包括； λ 远大于 n，只包括一个程度指标或者一个都不包括。

2.评价准备

（1）目前居民投诉最多的问题可以归结为九大问题，主要为学生补课、噪音扰民、就医医保、农民工工资、二胎及独生女、住房拆迁、教职工、环境问题、交通安全。而影响答复意见的质量主要为答复的相关性、完整性、可解释性、科学化、及时性等，所以对于相关性采用 ax+by+cz+d=0 计算，对居民提出的疑难和答复的意见进行判断，

看答复意见是不是随居民疑难的变化而随着变化，从而判断其相关性。

（2）对于及时性我们采用最常规的办法，用 excel 直接进行排序，判断回复的及时性。

（3）最后，对于相关性、完整性、可解释性、科学化、及时性，最后进行检验，我们知道，答复意见条数多且文字多，采用算法不方便，所以此处采用层次分析法（AHP），采取两两分析对比，用，判断答复意见的质量，看其与哪几个更加紧密相连。

（3）采用两种算法，对于其答复意见的质量更能做到精确分析和判断。

留言评价方案的说明

（1）通过采用上诉方法进行判断，全部指标都包括的则说明此答复意见质量高，对于一个指标都不包括的说明此答复意见质量差，可定为不合格。

（2）评价答复意见指标更能凸显出为人民服务的效率，对于不足的地方更应该好好改正，将“智政服务”充分体现为民服务的宗旨，切实解决民生问题。

$4 数据研究探讨结论

对于“智慧政务”中的方法探讨与数据挖掘，是为了政府部门能够充分的了解民意，其中各个网络渠道如市长信箱，阳光热线，微博等问政平台充满大量零散的民意信息，如何实现从零散信息获得关键信息，热点信息，减少政府部门相关的工作量，让政府部门能够根据

民意充分的解决问题以达到为人民服务的宗旨尤为重要，本文就对于“智慧政务”中的方法探讨与数据挖掘采用很多建模方法，数据研究方法如利用 TF-IDF 算法，K-means 算法 ，欧氏距离，Delphi 方法，单因素模糊评判综合评定法，全概率评分法，指标分级等对留言进行归类，构建留言评价指标，制作答复评价方案以达到目标所求。但本着能力有限，知识所备方面不足，在计算方面仍有待加强。但本文研究的方法与建立的模型对于解决留言政务处理起到很大的帮助，减少很多重复性和零散性的特点，对于分析出来的热点问题也比较具有代表性。本文着重对文本进行量化，数字化，权重归类分析化综合分析，减少误差与主观性的干涉，以达到目标的优化系统。

【1】刘学才.线性相关模型及其估计.《中国科技信息》2007

【2】层析分析法.https://zhuanlan.zhihu.com/p/35051786

【3】层次分析法如何确定权重..https://jingyan.baidu.com/article/2f9b480dca546b41cb6cc2df.html 工商银

【4】马忠秀.青海省门源县油菜花观光农业开发问题研究 .《中国海洋大学硕士论文》2012

【5】戚琦.基于 GIS 和目标层次联合分析方法的元阳县地质灾害易发程度评价研究.《中国地质大学(北京)硕士论文》2012

【6】刘寒.大数据环境下数据质量管理、评估与检测关键问题研究.《吉林大学博士论文》2019

【7】马彦 大数据环境下微博舆情热点话题挖掘方法研究

【8】 王晖，陈丽，陈垦，薛漫清，梁庆 《多指标综合评价方法及权重系数的选择》

【9】 赵静 但琦 严尚安 杨秀文 《数学建模与数学实验》

【10】孙海峰 郑中枢 杨岳武 《网络招聘信息的数据挖掘与综合分析》

【11】易燕飞.《基于 K-MEANS 聚类的数据分析》.2017

【12】《现代信息分析与预测》 书籍数据北京理工大学出版社 2011-02-01

【13】《南通地区干线公路桥梁综合技术状况评定方法》学术期刊《交通世界（运输车辆）》 2011 年 5 期

【14】《企业供应链柔性的 FAHP 综合评价研究》学术论文中国海洋大学 2009

【15】《张家界国家森林公园旅游植物资源评价体系研究》 学术期刊《湖南林业科技》2015 年 6 期

【16】《基于层次分析法的儿童游乐场所质量综合评价》学术期刊《建筑工程技术与设计》2017 年 13 期

【17】《构造判断矩阵讲解（层次分析法）-图文-百度文库）