基于智慧政务的文本挖掘

摘要

随着网络问政平台的不断更新，从大量的留言文本数据中挖掘出有用信息，对文本数据进行合理划分、热点问题整理，对提升政府的管理水平和施政效率具有极大的推动作用。本文旨在通过朴素贝叶斯算法确定留言的一级标签，借鉴 TDT 技术以及 K-means 算法筛选一段时间的热点问题，建立一套完整的答复意见评价体系。

针对任务 1，基于高斯朴素贝叶斯算法、二分类与朴素贝叶斯算法结合分别构建模型。在数据预处理阶段，为了获取较为准确的一级分类标签，首先确定特征属性（提取关键字），获取训练集和测试集，通过 Jieba 中文分词，将训练集分类，创建词典汇总词汇，使用 setofword2vec 获得训练样本的特征向量，生成训练矩阵，求词条属于分类的概率，为了获得更加准确的分类结果，使用拉普拉斯平滑处理数据，其次使用 classifyNB 测试分类结果，最终通过计算 F-score 对模型进行评价。

针对任务 2，主要借鉴 TDT 技术，结合智慧政务留言的特点，设计出一套智慧政务热点问题综合评价指标体系。首先对留言内容进行分词、去停词，词性过滤，然后使用 TF-IDF 计算特征词权重，转化成向量空间模型，K-means 聚类代替分类，其次提取关键词使用 textrank 形成话题，最后计算热度，实现热点问题排序。

针对任务 3，从相关性、完整性、及时性三个方面建立一套完整的关于答复意见的评价体系。通过给定每一个性质一个权重占比，计算出每一条答复意见评分。

关键字：朴素贝叶斯算法　　K-means 算法　　评价体系

Abstract

With the continuous update of the network political platform, useful information is mined from a large number of message text data, and the reasonable division of the text data and the sorting of hot issues have a great role in promoting the management level and efficiency of the government. The paper aims to determine the first-level label of the message through bayes algorithm, use TDT technology and K-means algorithm for reference to select hot issues for a period of time, and establish a complete response comment evaluation system.

For task 1, the model was constructed based on gauss naïve bayes algorithm and through the combination of dichotomy and naïve bayes algorithm. In data preprocessing phase, in order to obtain a more accurate level category labels, first determine the characteristics attributes(to extract the keywords), access to training set and test set, use setofword2vec training sample feature vector, to generate the training matrix, the probability of entry belongs to the category, in order to obtain more accurate classification results, using Laplacian smoothing data, finally use classifyNB test classification results. F-score was used to evaluate the model.

In view of task 2, a set of comprehensive evaluation index system for hot issures of intelligent affairs is designed based on TDT technology and the characteristics of intelligent government affairs message. Firstly, word segmentation, stop word removal and part of speech filtering were carried out for the message content. Then, TF-IDF was used to calculate the weight of feature words, which was converted into vector space model, and K-means clustering was used to replace classification. Secondly, keywords were extracted and textrank was used to form topics.

In view of task 3, a complete evaluation system was established from three aspects: relevance, completeness and timeliness. By giving a weight ratio for each property, the score of each reply is calculated.

Keywords: naïve Bayesian algorithm K-means algorithm evaluation system

目录

1 问题分析

近年来，随着微信、微博、市长信箱、阳光热线等等一系列了解民生的网络问政平台的蓬勃发展，各类与民意相关的文本数据量不断攀升，给人工带来了极大的挑战。依靠人工、根据经验处理各类网络问政平台的群众留言，存在工作量大、效率低、以及差错率高等问题。但是，近年来，大数据、云计算、人工智能等技术不断发展，建立基于自然语言处理技术的智慧政务系统大势所趋，这对提高政府的管理水平和施政效率具有极大的推动作用。

问题给出四个附件数据，内容分别为：附件 1，包括分类三级标签体系；附件 2 是各种详细的留言信息；附件 3 包括详细的留言信息以及点赞与反对数目；附件 4，政府针对具体留言信息给出的相关答复意见以及答复时间等信息。

要求完成的 3 个任务分别为：

（1）任务 1：要求根据分类三级标签体系（附件 1）对附件 2 给出的数据建立关于留言内容的一级分类标签模型，并对分类方法进行评价。

（2）任务 2：要求根据合理的热度评议指标，将某一时间段内反应特定地点或者特定人群问题的留言进行归类，并给出评价结果。

（3）任务 3：要求针对相关部门对留言的答复意见（附件 4）质量给出一套完整的评价方案。

2 任务 1

2.1 两两分类算法

在一般的思路下，我们首先想到的是二分类，将一种分类作为正样本，其余所有的分类作为负样本，这样就能够得到相应分类的特征向量，在这样的想法下我们建立了两两分类模型，它的实现过程如下所示：

第一步：首先建立城市建设模型。

（1）文本预处理部分：对示例数据进行处理，将城市建设作为正样本，选取其他适量样本作为负样本，组成训练集。对训练集中的正负样本进行人工标注，正样本前面标注 1，负样本前标注 0。对训练集进行切词和去除停用词处理，将文本所有词汇取并集汇总，对于每个训练样本，得到其特征向量。

（2）训练函数部分：得到样本数据，样本特征数和正样本概率。拉普拉斯平滑处理得到正概率下每个样本的概率，负样本概率下每个样本的概率。取对数，并将乘法改为加法，防止数值下溢损失精度。

（3）测试部分：比较正样本输入所有出现的词在样本中的概率之积*样本中正样本的概率和负样本输入所有出现的词在样本中的概率之积*样本中负样本的概率的大小，取数值较大的作为最终标签。

（4）测试方法：输入测试文本，进行切词以及去停用词处理；带入测试部分得到结果。

第二步：进行循环处理

（1）依照城市建设模型，建立其余分类模型。

（2）测试输入文本是否属于城市建设模型，是则输出，不是则继续进行测试下一类。依次循环，得到最终的分类结果。

在测试过程中由于数据量大，而人工分类也有不可靠的因素，最终的测试结果仅仅只能达到 30%。

2.2 高斯朴素贝叶斯模型

高斯朴素贝叶斯算法假设样本特征之间是相互独立的，算法流程如下所示：

(1)设$x = \{a_1 \quad a_2 \quad \dots \quad a_n\}$为待分类项,其中 a 为 x 的一个特征属性

(2)类别集合为:$Y = \{y_1 \quad y_2 \quad \dots \quad y_n\}$

(3)根据贝叶斯公式,计算$P(y_i|x)$

(4)如果$P(y_k|x) = max\{P(y_1|x) \quad P(y_2|x) \quad \dots \quad P(y_n|x)\}$,则 x 属于$y_k$这一类。

朴素贝叶斯算法的主要优点有：（1）朴素贝叶斯模型有稳定的分类效率。（2）对小规模的数据表现很好，能处理多分类任务，适合增量式训练，尤其是

数据量超出内存时，可以一批批的去增量训练。（3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

在任务 2 中我们通过对数据的预处理：提取附件 2 中的留言主题与留言详情进行 jieba 分词，使用停用词将数据进行拆分，将分词后的数据转化为字符串，进行向量化处理。同时我们将一级分类标签数据化，定义城市建设为 0，环境保护为 1，交通运输为 2，教育文体为 3，劳动和社会保障为 4，商贸旅游为 5，卫生计生为 6，方便后续数据处理。在测试中，选取全部数据的 20%作为测试数据作为模型的训练。在反复的测试中我们发现使用高斯朴素贝叶斯模型的准确率相较于两两分类的准确率有大大的提高。整体的准确率已经超过了 70%，而每一个一级标签的评价报告如表 2-1 所示。

表 2-1 高斯朴素贝叶斯算法评价报告

	Precision	Recall	F1_score	support
0	0.64	0.74	0.69	396
1	0.78	0.73	0.75	181
2	0.87	0.40	0.55	130
3	0.68	0.80	0.74	323
4	0.73	0.75	0.74	398
5	0.65	0.64	0.64	237
6	0.82	0.64	0.72	177

3 任务 2

3.1 解题思路

智慧政务已经成为公众反应实际生活问题得到快速解决的便捷渠道，智慧政务上的热点问题也代表了众多公民关注的重点。掌握智慧政务热点问题，对提高政府的管理水平和施政效率具有极大的推动作用。

目前，利用网络信息进行的研究主要包括两部分：一类是 Web 数据挖掘研究，另一类是利用话题检测和追踪（Topic Detection and Tracking, TDT）技术进行热点话题识别与追踪研究。TDT 技术已逐步成为当前信息处理领域的研究热点，该项技术中涉及许多算法与模型的运用。

任务 2 主要借鉴 TDT 技术，结合智慧政务留言的特点，设计出一套智慧政务热点问题综合评价指标体系。首先分词去停词，词性过滤，然后使用 TF-IDF 计算特征词权重，转化成向量空间模型，K-means 聚类代替分类，其次提取关键词使用 textrank 形成话题，最后计算影响力，判断热度，对热点问题进行排序。其具体实现步骤如下图所示。

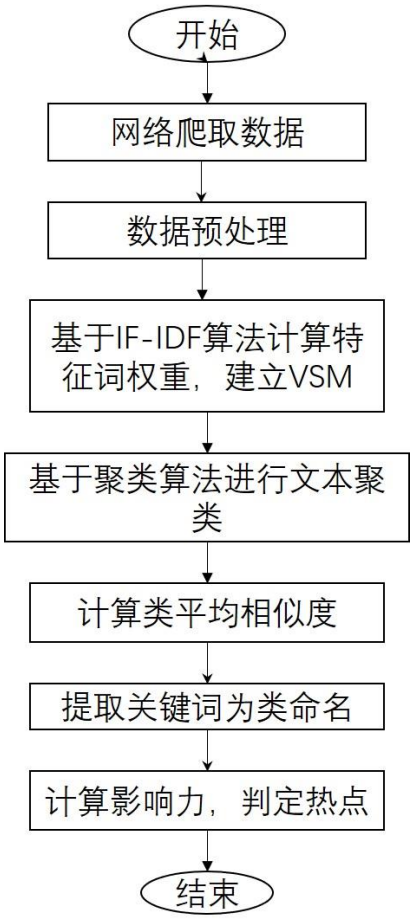

图 3-1 留言内容热点问题识别流程图

3.2 相关理论

（1）中分分词及词性标注

中文分词就是将汉字序列切分成有意义的词，以字为单位，句和段落则是通过标点等分隔符来划界。本文中是用的中文分词算法是 jieba 分词。它的基本原理如下所示：

第一条：基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况的有向无环图（DAG）；

第二条：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；

第三条：对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法

（2）K-means 文本聚类

K-means 算法以欧式距离作为相似性的评价指标，即认为两个对象的距离越

近，其相似度就越大，得到紧凑且独立的簇是聚类的最终目标。K-means 算法中距离的计算公式如下[1]：

$$v = \sum_{i=1}^{k} \sum_{x_i \epsilon S_I} (x_j - u_i)^2$$

K-means 算法流程：

第一步，从数据对象中任意选择 K 个对象（K 值需要预先设定）作为初始聚类中心。第二步，计算剩下的对象与这些聚类中心的相似度（距离），并分别将它们分配给最相似的（聚类中心所代表的）类。第三步，重新计算每个新类的聚类中心（该聚类中心所有对象的均值）。第四步，不断重复第二、三步，直到标准测度函数开始收敛为止，一般采用均方差作为标准测度函数。

该算法在处理大数据集时是相对高效和可伸缩的，计算复杂度为$O(N_{kt})$，其中 N 是数据对象的数目，t 是迭代的次数（一般 K≤N，t≤N，同时算法对顺序不太敏感）。文本聚类效果的验证采用类平均相似度，公式为：

$$AVG_T(SIM) = \frac{\sum C_{T_{i=1}} f_t(avg(sim))}{C_T}$$

其中$AVG_T(SIM)$表示类 T 的平均相似度；C_T表示类 T 所包含的留言条数；$f_t\big(avg(sim)\big)$表示类 T 中单条留言 t 的个体平均相似度，即 t 与类 T 中其余留言的相似程度之和取平均值。将类中所有留言的个体平均相似度之和取一次平均值，从而得到类的平均相似度。

（3）话题影响力相关计算公式

本文基于留言内容和话题本身，提出了热点的判定因素——话题影响力。留言热点话题影响力为该话题中单条相关留言内容的影响力总和，单条留言内容的又分为直接影响力和间接影响力。由于用户发表的留言内容直接呈现给在智慧政务平台上所有看到的其他用户以及相关的部门后端，因此单条留言内容的直接影响力与该条留言的影响人数（受众数）相关。本文此处只考虑留言内容与相应的点赞数与反对数。定义话题影响力计算公式如下：

$$\text{Inf(T)} = \sum_{i=1}^{n} Inf(t), i = 1,2,\dots,n$$

其中 Inf(T)为话题 T 的影响力；n 为该类中与话题相关的留言条数；Inf（t）为单条相关留言内容 t 的影响力。一个话题的影响力为话题中所包含的所有相关

留言内容之和。

$$Inf(t) = Inf_{D(t)} + Inf_{I(t)}$$

其中$Inf_D(t)$为单条留言内容 t 的直接影响力；$Inf_I(t)$为单条相关留言内容 t 的间接影响力。单条留言内容的影响力为直接影响力与间接影响力之和。话题 T 的影响力为：

$$Inf(T) = \sum_{i=1}^{n} Inf_{D(t)} + \sum_{i=1}^{n} Inf_{I(t)} \quad i = 1,2,\dots,n$$

$$Inf_{D(t)} = |followed| \times \frac{1}{N(\Delta h)}$$

其中 followed 为发布留言内容 t 的用户的关注人数（受众数）；N(Δh)为时间段Δh内该用户发布的留言总条数。实际情况中，并非所有受众都随时关注智慧政务平台上 的新动态，因此假设每个受众查看到该条留言内容的概率为$\frac{1}{N(\Delta h)}$，即一段时间内，该平台发布的留言内容越多，那么所发布的留言 t 就越容易被淹没。该用户发布的留言内容 t 被平台所有注册用户接收的次数等于概率与受众数的乘积，也就是 t 的直接影响力。

$$Inf_I(t) = \alpha \times |comments| + \beta \times |retweents|$$

为了将数据中的点赞数与反对数都应用到模型中，默认将所有点赞数与反对数+1 做预处理，其中 comments 代表处理过的点赞数，retweents 代表处理过的反对数，系数α>0,$\beta > 0$，$\alpha + \beta = 1$。而具体参数值可运用经验或者专家打分等手段来确定。

3.3 模型实现

在以上理论的支撑下，我们通过代码实现，最终的热点问题提取结果如下表所示。

表 3-1 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	2285	2019/1/8 至 2019/7/8	A 市 58 车贷案受害人	P2P 平台恶性退出导致出借人受损
2	2	2098	2019/8/1 至 2019/8/19	A 市 A5 区汇金路五矿万境 K9 县小区业主	小区租房混乱造成隐患
3	3	1765	2019/4/7 至 2019/4/11	A 市梅溪湖配套入学政策	小区配套入学被政策除外
4	4	683	2019/8/2 至 2019/9/6	A 市绿地海外滩二期小区业主	小区业主被渝长厦噪音影响
5	5	243	2019/6/1 至 2019/6/19	A 市富绿物业丽发新城业主	物业不作为，断业主水

4 任务 3

智慧政务是人们希望通过网络的手段更方便地处理自己生活中所面临的问题的一个有效平台。任务三是建立一比较完整的针对答复意见的评价体系，通过给出的数据，我们从以下几个方面来考虑问题：留言主题与留言内容作为一个部分，答复意见的详细内容作为一个部分，这两者之间的相关程度我们视为答复意见大相关性，而政府答复意见的时间与留言时间的差值我们视为答复意见的及时性，我们将 J 做过处理的 Jaccard 系数作为可解释性的指标。我们将每一部分都赋予一定的权值占比：相关性占比 50%，及时性占比 30%，可解释性占比 20%，最终给出每一条答复意见的评分，根据评分的多少就能知道某一条答复意见的群

众满意度。我们将最终的评分存为“答复意见评价.xlsx”。答复意见的评价体系如下图所示：

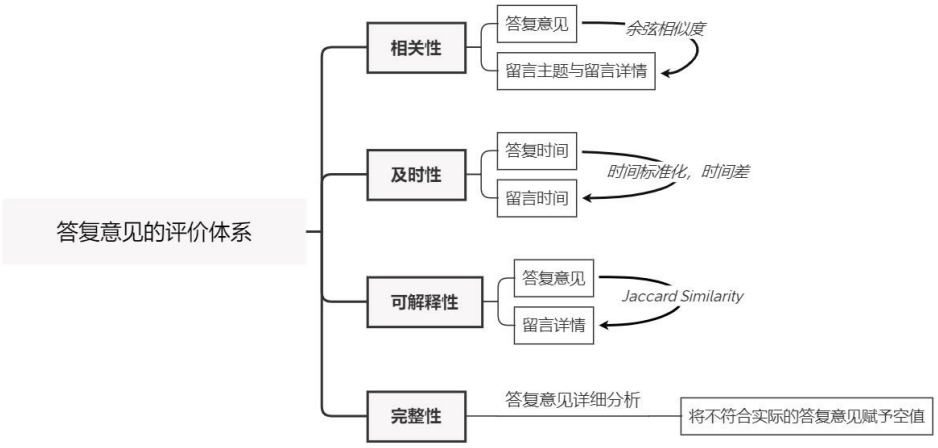

图 4-1 答复意见评价体系实现图

以下就是针对答复意见的每一条性质进行分析的具体内容。

4.1 答复意见的相关性

答复意见的相关性可以作为评价答复意见的指标之一，即相关性的强弱（相似程度的高低），一定程度上可以反映答复意见与留言主题和留言详情的贴合度；可以通过计算答复意见的相关程度来判断答复意见是否针对留言主题和留言详情提出。相似度越高，答复意见便越可靠。

而计算两者之间的相似度将采用以下方式：

(1)使用 jieba 包分别对给定的两个 xlsx 文件进行分词，得到如[‘我’，‘今天’，‘很开心’]的两个字符串数组；

(2)对得到的分词后的数组通过进行词袋模型统计，得到他们每个词在文中出现的次数向量；

(3)对得到的两个次数的矩阵进行余弦相似度计算，得到余弦相似度作为他们的文本相似度。

4.1.1 相关理论

（1）词袋模型

文档内容(以下均简称为文档)中出现频率越高的词项，越能描述该文档（不考虑停用词）。因此可以统计每个词项在每篇文档中出现的次数，即词项频率，记为 $tf_{t,d}$ ，其中 t 为词项，d 为文档。获得文档中每个词的 tf 权重，一篇文档则转换成了词-权重的集合，通常称为词袋模型（bag of words model）。我们用词袋模型来描述一篇文档 。词袋的含义就是说，像是把一篇文档拆分成一个一个的词条，然后将它们扔进一个袋子里。在袋子里的词与词之间是没有关系的。因此词袋模型中，词项在文档中出现的次序被忽略，出现的次数被统计。例如，“a good book” 和“book good a” 具有同样的意义。将词项在每篇文档中出现的次数保存在向量中，这就是这篇文档的文档向量。

（2）余弦相似度计算

前面提出了文档向量的概念。其中每个分量代表词项在文档中的相对重要性。一系列文档在同一向量空间的表示称为 VSM（Vector Space Model）。VSM 是词袋模型。向量空间模型是信息检索、文本分析中基本的模型。通过该模型，可以进行有序文档检索、文档聚类、文档分类等。当然，现在的研究有新发展。出现了很多模型代替 VSM。每篇文档在 VSM 中用向量表示，那么计算两篇文档的相似度自然的想到用两个向量的差值。但是，可能存在的情况是。如果两篇相似的文档，由于文档长度不一样。他们的向量的差值会很大。余弦相似度是使用的非常广泛的计算两个向量相似度的公式，它可以去除文档长度的影响。

$$\mathrm{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

公式的分母是两个向量的欧几里得长度之积，分子是两个向量的内积。这样计算得到的 sim 实际上就是两个欧式归一化的向量之间的夹角的余弦。如下图：

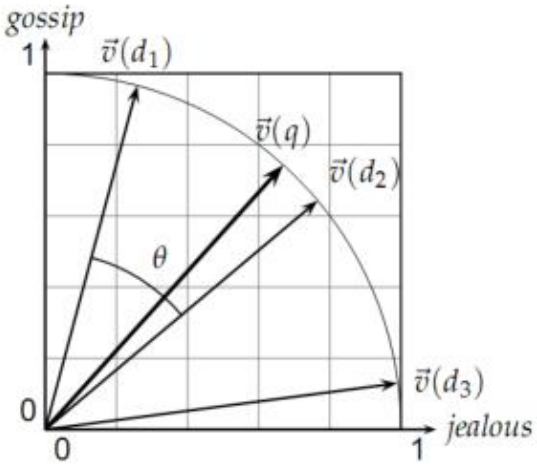

图 4-2 余弦相似度原理图

4.2 答复意见的完整性

数据库中的数据是从外界输入的，而数据的输入由于种种原因，会发生输入无效或错误信息。保证输入的数据符合规定，成为了数据库系统，尤其是多用户的关系数据库系统首要关注的问题。数据完整性因此而提出。本章将讲述数据完整性的概念及其在 SQL Server 中的实现方法。

数据完整性（Data Integrity）是指数据的精确性（Accuracy） 和可靠性（Reliability）。它是应防止数据库中存在不符合语义规定的数据和防止因错误信息的输入输出造成无效操作或错误信息而提出的。数据完整性分为四类：实体完整性（Entity Integrity）、域完整性（Domain Integrity）、参照完整性（Referential Integrity）、用户自定义完整性（User-definedIntegrity）。

数据库采用多种方法来保证数据完整性，包括外键、约束、规则和触发器。系统很好地处理了这四者的关系，并针对不同的具体情况用不同的方法进行，相互交叉使用，相补缺点。

在附件 4 答复意见这一列数据中，我们发现部分数据只有日期，或者政府的答复意见为：网友您好，您的留言已收悉。并没有给出留言人实际需要的答复意见，我们将这一部分数据做处理，认为这一部分数据为空值，即这一部分数据是不满足数据完整性要求的。

4.3 答复意见的及时性

答复意见的及时性同样也可以作为评价答复意见的指标之一，及时性指的是留言与答复时间差，时间差越小说明答复的更及时；时间差越大说明答复出现了迟滞。通过对答复意见及时性的计算无疑可以看出政府对群众反映问题的关心程度。

通过查询相关的资料我们发现，一般政府部分给定的答复时间范围为 15 个工作日，我们为了能够体现答复意见的及时性能，设定了一系列的等级：0-14 天为最高等级——第七级，15-28 天为第六级，以此类推，我们给时间设定了七个不同等级，等级越高说明政府部门答复意见越及时。给出相应的等级之后，做相应处理：

$$等级数 \times 5$$

处理之后的数据即为及时性的得分。

4.4 答复意见的可解释性

答复意见的可解释性就是答复意见能否对群众在留言主题以及留言详情中的问题做出合理的解释与答复。在这一部分中我们采用了 Jaccard similarity。

在生活中，我们可以利用集合的思想对文档的相似度进行分析，进而将文档表示成一个集合。将文档表示成集合的最有效的方法是构建文档中的段字符串集合，如果温塘采用这样的集合表示，那么有相同句子甚至短语的文档之间将会拥有很多公共的集合元素，而在答复意见的可解释性方面，我们可以利用答复意见与留言详情的相似度作为答复意见可解释性的指标。同时将计算的 Jaccard 系数做相应处理：

$$Jaccard \times 1000$$

处理之后的数据作为答复意见可解释性的指标。

4.5 总结

通过建立一个完整的答复意见评价体系，可以明显地减少人力，同时政府方面也能够根据不同方面的问题及时改正，同时加速政府部门的建设与优化，促进

社会更好的进步。

5 总结与展望

5.1 总结

为了减少智慧政务中留言内容所属一级标签人工分类的错误率，某一时间段内的热点问题的提取以及一套完整的答复意见的评价体系，本文基于机器学习的相关理论，构建了一个智能分类模型，让政府部门的工作人员能够减轻一定的人力；构建了一个热点问题提取模型，集中反映某段时间，某个地点或者某个人群的问题，有利于相关部门针对性地解决相关问题，大大提高工作效率；构建了一个完整的答复意见评价体系，基于答复意见的不同性质给出一定的评价标准。在对赛题研究的基础上，我们根据研究思路撰写论文，基本实现本赛题设立的目标。

5.2 展望

5.2.1 数据预处理

机器学习的好坏往往取决于数据集的质量，目前属于不同的一级标签的数据都是来源于不同平台爬取的数据，存在数量有限和参差不齐的现象。同时 K-means 算法要求最终的簇紧凑且不再发生变化，这中间对于数据的处理极其重要。针对这些不同的问题数据预处理方便可能还有待进一步的优化。

5.2.2 热点问题提取的具体情况

根据最终分类的数量，每一个类中都会包含不止一个地点或者人群，如果采用机器来提取的话，中间不免会出现不是一个地点或者一个人群的问题，这样的处理不仅不能节省人力，反而会出现更多的错误，所以在经过聚类后提取得到的关键词，我们选择了人为来提取，但是这样的做法不免会带有一定的人为主观因素，因此在这一方面仍然需要改进。

5.2.3 答复意见评价体系每一部分的占比

不同的政府部门对于答复意见的每个性质的关注程度是不同的，而根据不同性质的不同占比，都会形成最终关于每一条答复意见的评分，所以在真正的应用过程中，应该根据每个不同部门的需要合理分配占比。

参考文献

[1] 何跃,帅马恋,冯韵.中文微博热点话题挖掘研究[J].统计与信息论坛,2014.

[2] 黄贤英,陈红阳,刘英涛,熊李媛.一种细腻的微博短文本特征词选择算法[J].计算机工程与科学,2015.

[3] 崔伟.一种基于朴素贝叶斯算法的中文文本分类系统[J].信息科技与信息化,2015.

[4] 邓志远.基于自然语言处理的电信系统热点问题的提取[J].信息科技与信息化,2020.

[5] 梁昌明,李冬强.基于新浪热门平台的微博热度评价指标体系实证研究[J].情报学报,2015.