“智慧政务”中的文本挖掘应用

摘　　要

近年来，随着各类社情民意相关的文本数据量的不断攀升，对于行政部门而言，传统人工进行留言划分和热点整理的弊病愈加明显：工作量大，分类标准不统一，分类效率低，遗漏率、差错率高……同时，随着大数据、云计算、人工智能等技术的飞速发展，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，既能够优化传统的分类方法，提高文本分类的效率，又能够更加个性化、人性化地满足或改善民生需求。对提升政府的管理水平和施政效率具有极大的推动作用。附件给出了收集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。

在数据预处理阶段，我们对附件中留言主题这一属性中的无效主题，根据留言详情人工提炼，然后进行分词和去停用词处理，并且针对不同的问题，我们采用不同的词向量表示方法。

对于问题一，以留言详情作为神经网络模型的数据集，词向量采用 one-hot 编码。实验尝试 TEXTCNN，FSATTEXT，LSTM 等模型进行文本分类，然后在 TEXTCNN 模型的基础上，修改得到 KMAXCNN 模型。使得神经网络模型的 F1-score 及准确度、模型训练效率等均有所提高，在验证集上的准确度可以达到 90%左右。

对于问题二，作为使用单遍聚类(single-pass)模型将留言主题进行自动聚类。使用实体命名识别提取出各个分类问题中的地点，综合问题留言的时间跨度、主题数、赞成总数、反对总数，使用熵值法计算各影响因素的权值，加权后计算得到热点指数，按热点指数降序排序得到热点问题表及热点问题留言明细表。

对于问题三，相关性方面，将留言详情及答复意见以 word2vec 向量表示，以留言详情的句子粒度及答复意见的句子粒度计算余弦相似性得到相关性评价指标。可解释性根据易读性公式量化指标。完整性方面利用留言详情与答复意见的关键字匹配度和答复意见字数进行熵权确立权重，计算得完整性评价指标。及时性通过计算留言时间与答复时间的时间间隔来量化指标。然后对于可解释性、完整性、及时性做数据降维，归一化处理。从留言回复的相关性、可解释性、完整性、及时性方面对留言回复进行综合评价，在此过程中使用熵值法确定四个特性的权重比例，从而计算得到答复质量指数。

关键词：KMAXCNN 模型；SINGLE-PASS 算法；熵值法

Abstract:

In recent years, with the continuous increase in the amount of text data related to various social conditions and public opinion, for administrative departments, the

disadvantages of traditional manual message division and hotspot sorting have become more obvious：Large workload, inconsistent classification standards, low classification efficiency, high omission rate and error rate … At the same time, with the rapid development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government systems based on Natural Language Processing technology (NLP) has become a new trend in the development of social governance innovation. It can not only optimize traditional classification methods and improve the efficiency of text classification, but also meet or improve the needs of people's livelihood more personally and humanely. It greatly promotes the government's management level and governance efficiency. The appendix gives a record of the public's questionnaires collected from public sources on the Internet, as well as the relevant departments' answers to some of the public's messages.

In the data pre-processing stage, we manually refine the invalid subject in the attribute of the message subject in the attachment, and then perform word segmentation and stop word processing, and for different problems, we use different word vector representation methods .

For question one, the message details are used as the data set of the neural network model, and the word vectors are encoded with one-hot. Experiments with TEXTCNN, FSATTEXT, LSTM and other models for text classification, and then modified the KMAXCNN model based on the TEXTCNN model. The F1-score and accuracy of the neural network model and the efficiency of model training have been improved. And the accuracy on the verification set can reach about 90%.

For question two, a single-pass clustering (single-pass) model is used to automatically cluster topics. Use Named Entity Recognition to extract the locations in each classification problem, integrate the time span of the problem message, the number of topics, the total number of approvals, and the total number of oppositions. Use the entropy method to calculate the weight of each influencing factor. Sort by hotspot index in descending order to get hotspot problem list and hotspot problem message list.

For question three, in terms of relevance, the message details and response opinions are expressed as word2vec vectors, and the cosine similarity is calculated based on the sentence granularity of the message details and the response opinion to obtain the correlation evaluation index. Interpretability quantifies indicators based on legibility formulas. In terms of completeness, the entropy weight is established by using the keyword matching degree of the message details and the answer opinion and the word number of the answer opinion, and the integrity evaluation index is calculated. Timeliness quantifies the index by calculating the time interval between message time and reply time. Then perform data reduction and normalization for interpretability, completeness and timeliness. From the relevance, interpretability, completeness, and timeliness of the message reply, the message reply is comprehensively evaluated. In this process, the entropy value method is used to determine the weight ratio of the four characteristics, thereby calculating the reply quality index.

Keywords: KMAXCNN model；SINGLE-PASS；entropy method.

目　　录

一、引言

近年来，随着科学技术的不断进步，为了能够及时、快速地了解到民生民意，政府部门设立了民生服务热线，供群众表达诉求。短短几年，微信、微博、市长信箱、阳光热线等网络问政平台已经逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，随着各类社情民意相关的文本数据量的不断攀升，传统人工进行留言划分和热点整理的弊病愈加明显：工作量大，分类标准不统一，分类效率低，遗漏率、差错率高……同时，随着大数据、云计算、人工智能等技术的飞速发展，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，既能够优化传统的分类方法，提高文本分类的效率，又能够更加个性化、人性化地满足或改善民生需求。对提升政府的管理水平和施政效率具有极大的推动作用。

二、实验目标

本组目的在于利用赛方发布的数据，利用 jieba 中文分词工具对文本进行分词、Hanlp 进行命名实体识别、Pyltp 进行词性标注、利用 KMAXCNN 模型、Single-Pass 聚类算法、熵值法等算法，建立基于自然语言处理技术的智慧政务系统。总目标可分为以下三个目标：

1、建立留言分类系统，通过机器学习技术对用户的留言信息进行自动留言分类，高效、精准划分出问题所在的类别，减少大部分电子政务系统依靠人工经验处理，而产生的工作量大、效率低，且差错率高等问题，提高执政效率。

2、自动将留言问题归类并且对热点问题进行挖掘，重点在于如何实现留言问题自动聚类，怎样将问题反映的特定地点和特定人群从留言中提取处来，提出一种合理的问题热点指数量化方案，使得问题的分类更加智能化、便捷化。发现热点问题便于帮助政府更加迅速地了解民众诉求，并且有侧重地知悉民生民意，集中解决社会大众反映的问题。

3、建立答复意见评价模型，对相关部门的答复意见进行评判，最终以答复评分的形式反馈给社会大众，以便监督相关行政部门能够更加及时、高效、切实地解决民生问题。

三、分析方法与过程

3.1 问题一分析方法与过程

3.1.1 流程概览

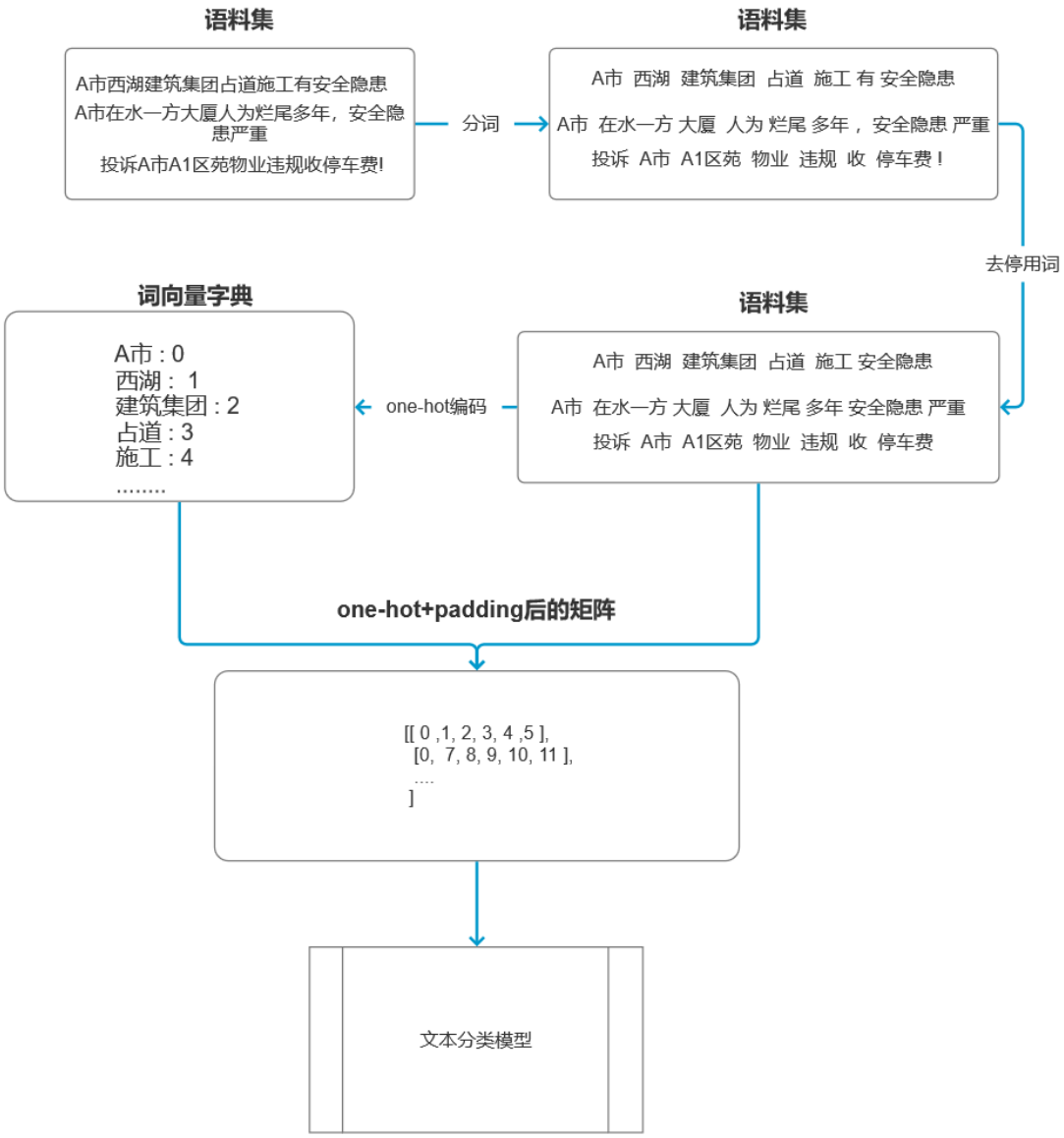

图 1. 问题 1 流程示意图

3.1.2 数据预处理

3.1.2.1 留言信息去重去空

对于附件所给的留言信息，数据几乎不存在空值或重复值等，但是为了提高程序适用的广泛性，防止爬虫在爬取数据时出现数据采集重复或采集失败等错误，我们在数据预处理中仍然有去重去空这一步骤。

3.1.2.2 文本分词

汉语文本与英语文本不同，英语文本单词之间有空格分隔，计算机可以以此来识别单词和句子之间的语义关系；而汉语文本以汉字作为基本单位，词语之间界限模糊，为方便计算机识别，我们必须将句子中的词与词分开，即分词[1]。同时要去除标点符号，以防标点符号对分词结果造成影响。

jieba 分词是 python 的中文分词组件，其算法基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (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。在所有分词处理部分，均使用 jieba 进行分词[2]。但分词结果并不完全准确，我们引入自定义词典，将有歧义或不易切分的词语写入自定义词典，提高了分词效果。

3.1.2.3 去停用词

文本分词处理后，文本由原来的完整句子变为独立的词语，其中包含一些使用广泛但是不具有实际意义的词，例如“我”、“的”等。该类字词被称为“停用词”[3]。筛去停用词对于文本分类具有消除干扰、降低文本维度等重大意义。本文的停用词表由哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表等停用词表进行去重合并后得到[4]。

3.1.2.4 one-hot 编码

one-hot 编码属于词袋模型。使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。

在进行分词及去停用词后。我们对每个词进行编号，再使用 one-hot 对每句话提取特征向量。其优点在于解决了分类器不好处理离散数据的问题，一定程度上也起到了扩充特征的作用。

3.1.3 模型选择

本文使用 FastCNN、TextCNN、LSTM 等三种模型，对留言详情及留言主题分布进行训练分类。留言详情的分类效果明显优于留言主题。对于留言主题的训练效果就不一一展示。在三种模型的训练结果中，TextCNN 在训练速度和训练结果上明显优于另外两种模型。

3.1.3.1 TextCNN 模型

卷积神经网络（CNN Convolutional Neural Network）最初在图像领域取得了巨大成功，其核心在于能够捕捉局部相关信息。TextCNN 是 CNN 模型在 NLP 领域的文本分类模型，来源于 Yoon Kim 在 2014 年“Convolutional Neural Networks for Sentence Classification”论文[5]中提出的利用卷积神经网络进行文本分类的算法。

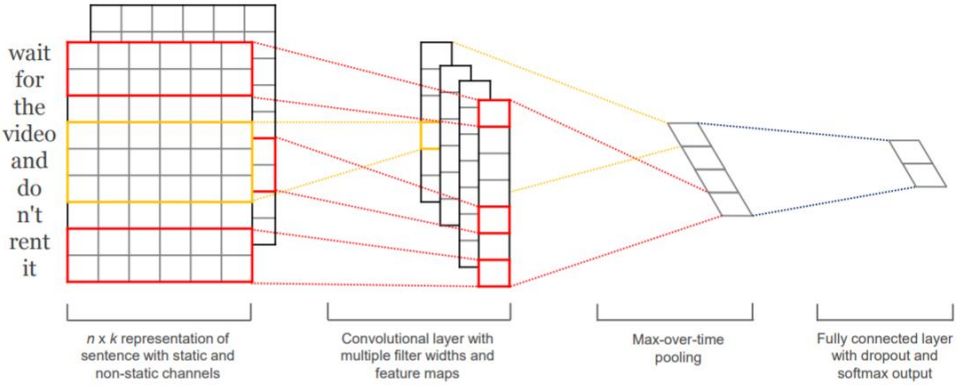

Figure 1: Model architecture with two channels for an example sentence.

图 2. TextCNN 模型框架

3.1.3.1.1 TextCNN 模型具体步骤

1、句子的矩阵化表示

每个单词都由一个 k 维向量来表示，并且每个句子都填充到最大长度 n，因此每个句子都可以表示为一个 n×k 的矩阵，公式如下：

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n$$

句子的长度为 n，x_i表示的是每个单词，$\oplus$指的是将 n 个 1×k 的向量拼接起来形成 n×k 的矩阵，该矩阵即为句子的矩阵表示形式。

2、卷积操作

创建一个大小为 h×k 的卷积核，其中 k 对应为每个词向量的维度 k，h 是一个小于 n 的值。卷积核会在表示句子的矩阵上纵向滑动，形成多个特征值，公式如下：

$$c_i = f(W \cdot x_{i:i+h-1} + b)$$

其中$x_{i:i+h-1}$表示的是句子矩阵的第 i 到 i+h-1 行，随着 i 值不断增大，卷积核会与多个这样的矩阵相乘，最终得到 n-h+1 个特征值：

$$c = [c_1, c_2, \ldots, c_{n-h+1}]$$

3、池化操作

使用最大池化的方法进行池化操作，公式如下：

$$\hat{c} = \max\{c\}$$

4、全连接

当每个句子都经过前三个步骤后会得到多个句子的值，将这些值共同放进全连接层，通过一个 softmax 后进行多分类。

3.1.3.2 KMAXCNN 模型

在 TextCNN 的基础上，为了再对模型进行优化，所以我们选择了 KMAXCNN 模型。其训练速度在原来的基础上提高了 87.65%，验证集准确率达到了 90%左右，下面是对 KMAXCNN 模型的详细介绍。

KMAXCNN 模型[6]与 TextCNN 模型的不同之处主要在 polling 层，池化层没有直接使用 maxpooling 这种粗暴的方式， 因为该方式只提取最明显的特征而舍弃了其余所有的特征，因此我们的模型采用 k-maxpooling 层，k-maxpooling 会取特征值中得分 Top-k 的值，并保留这些特征值的原始先后顺序，以待后续使用，该方法有利于提取更多句子的信息。例如：“我觉得这个比赛确实很难，但是最重要的是我们得到了锻炼”，虽然前半部分是负面情感，但整体情感是偏正面的，这时使用 k-maxpooling 就可以很好的捕捉此类信息。

3.1.3.2.1 KMAXCNN 模型层次结构

1、嵌入层（embedding layer）

通过隐藏层，将 one-hot 编码后的词向量投影到一个指定维度的高维空间中进行特征提取，编码语义特征。经过该层后，语义相近词的欧氏距离或者余弦距离会比较近。

2、Spatial Dropout 层

Spatial Dropout 层会随机的将部分区域置 0，目的是为了防止数据量过大或者模型过大导致的过拟合问题。

3、卷积层（Convolution layer）

用来进行特征抽取，可以设定超参数来指定设立多少个特征抽取器（Filter），对于某个 Filter 来说，可以想象有一个 k×d 大小的移动窗口从输入矩阵的第一个字开始不断往后移动，其中 k 是 Filter 指定的窗口大小，d 是 Word Embedding 长度。对于某个时刻的窗口，通过神经网络的非线性变换，将这个窗口内的输入值转换为某个特征值，随着窗口不断往后移动，这个 Filter 对应的特征值不断产生，形成这个 Filter 的特征向量。每个 Filter 都如此操作，形成了不同的特征抽取器。在本模型中指定了 4 个卷积层分别对 k 取 1 到 4 的 4 种情况进行卷积操作，然后将输出连接在一起。

4、k-maxpooling 层

由于在卷积层中我们使用了不同高度的特征抽取器，导致通过卷积后得到的向量维度不一致，因此使用池化层将特征向量池化为 k 个值，我们的模型中 k 取值为 3，该 k-maxpooling 层会取特征值中得分 Top-3 的值，并保留这些特征值的原始先后顺序。

5、全连接层（Fully connected layer）

全连接层的第一层使用 relu 作为激活函数，第二层使用 softmax 作为激活函数来得到属于每个类的概率。

3.1.3.2.2 KMAXCNN 模型结构图

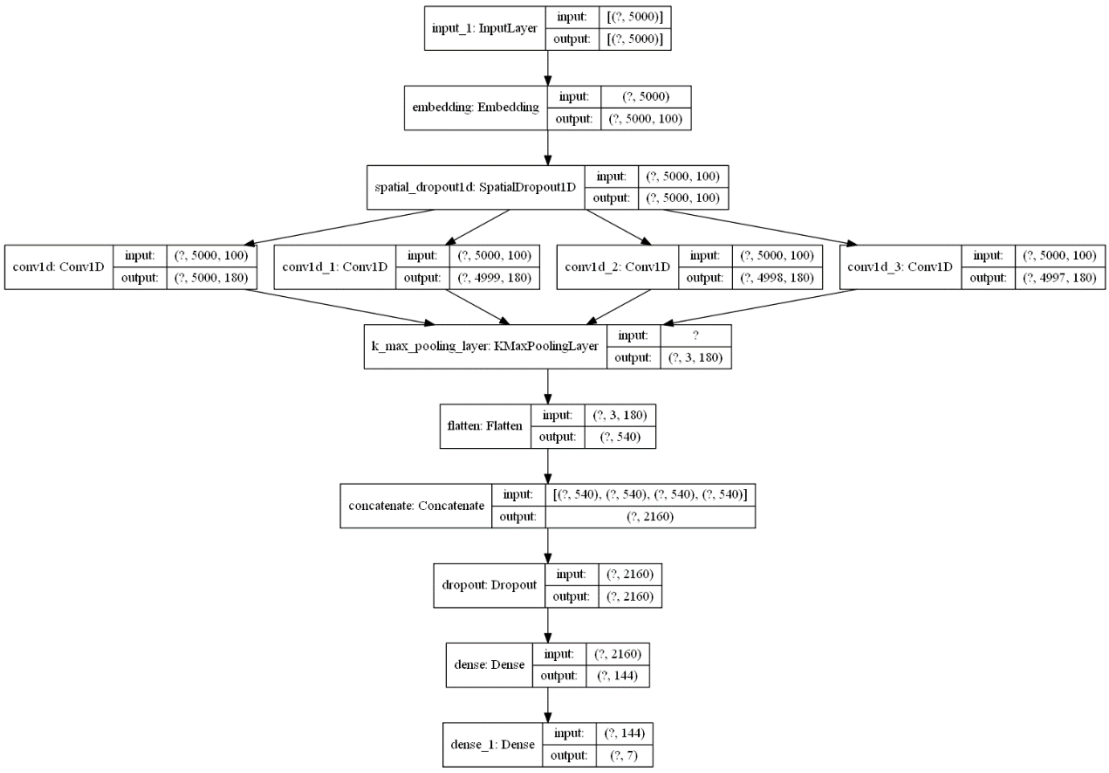

图 3. KMAXCNN 模型结构图

3.1.4 模型评估

1、查准率（Accuracy）

系统检索到的相关文档数 / 系统检索到的全部文档数，即关注预测为正样本的数据(可能包含负样本)中,真实正样本的比例，公式如下：

$$\boldsymbol{P = \frac{TP}{TP + FP}}$$

2、查全率（Recall）

系统检索到的相关文档数 / 系统所有相关的文档总数，即关注真实正样本的数据(不包含任何负样本)中,正确预测的比例，公式如下：

$$\boldsymbol{R = \frac{TP}{TP + FN}}$$

3、F-Score

当模型对查准率和查全率的要求都高时，我们采用$\boldsymbol{F_1}$，即查准率和查全率的调和平均值来评估模型，公式如下：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

3.1.5 训练结果对比及分析

在实验过程中，我们借助以上指标依次对比了 FastText 模型、LSTM 模型等，最后测试出我们的模型——KMAXCNN 模型均优于上述模型。

- FastText 模型

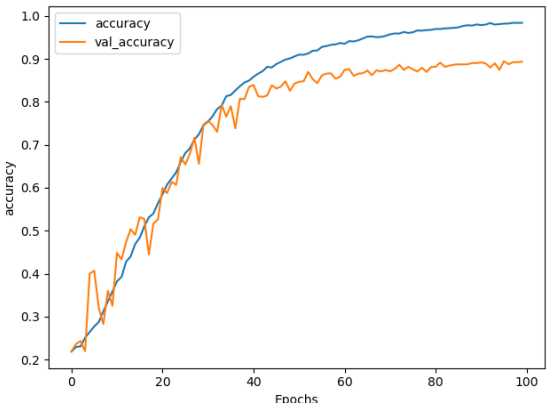

图 4. FastText 模型训练集、验证集准确度

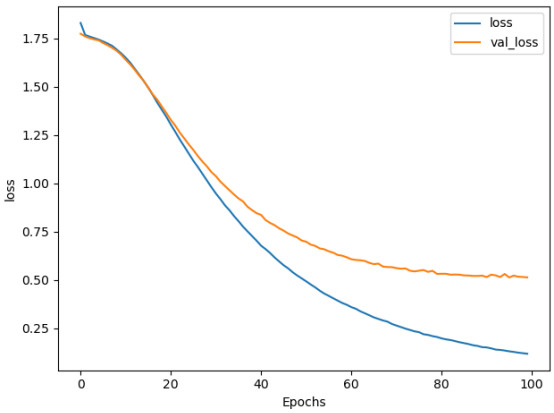

图 5. FastText 模型训练集、验证集损失值

- LSTM 模型

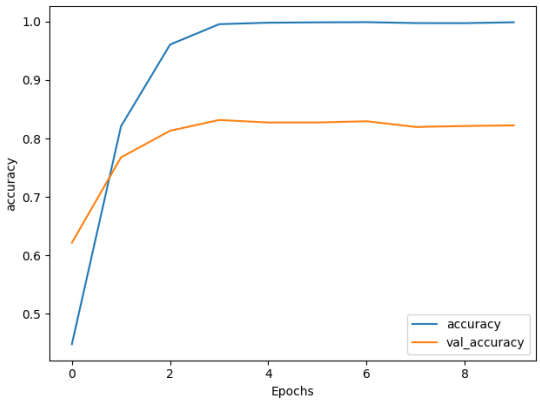

图 6. LSTM 模型训练集、验证集准确度

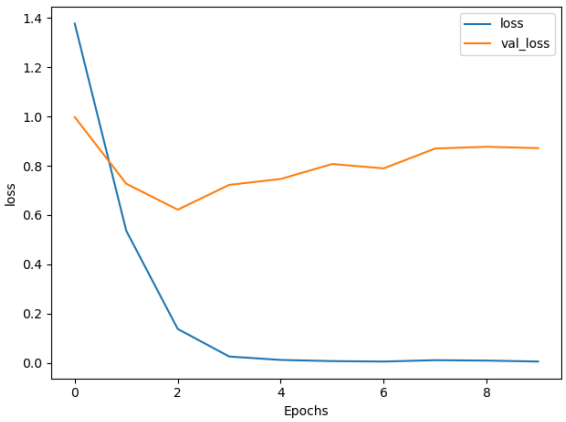

图 7. LSTM 模型训练集、验证集损失值

- TextCNN 模型

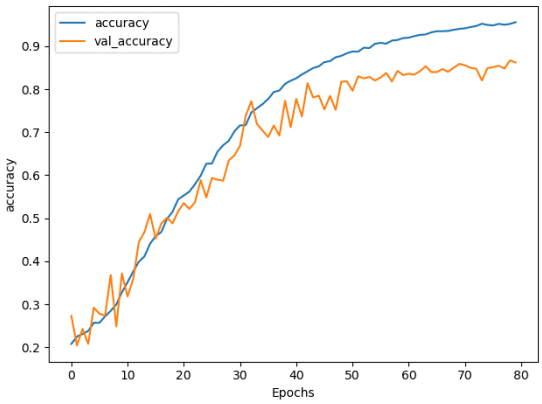

图 8. TextCNN 模型训练集、验证集准确度

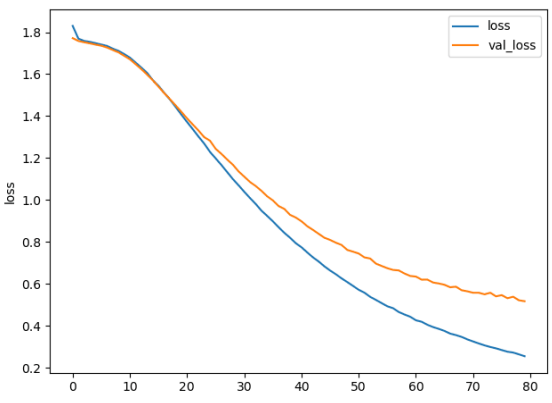

图 9. TextCNN 模型训练集、验证集损失值

- KMAXCNN 模型

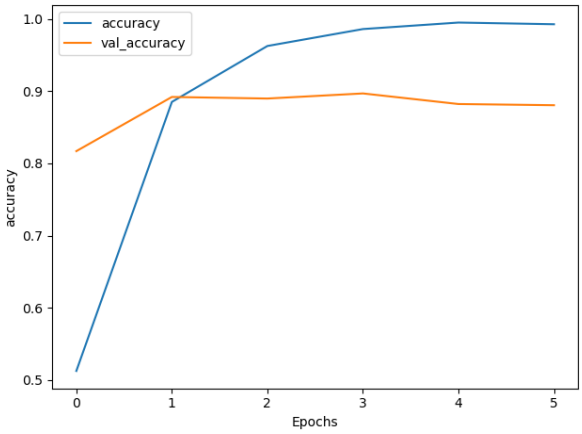

图 10.KMAXCNN 模型训练集、验证集准确度

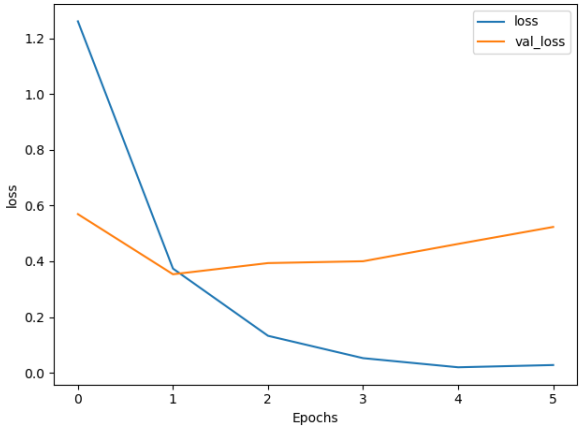

图 11.KMAXCNN 模型训练集、验证集损失值

MODEL	F1-Score	Precision	Recall	Time/s
KMAXCNN	0.900439	0.901504	0.900322	235
TEXTCNN	0.890555	0.887454	0.894591	1903
LSTM	0.83817	0.847946	0.835855	540
FASTTEXT	0.867296	0.874649	0.862118	2080

图 12. 各个模型评价指标对比图

3.2 问题二分析方法与过程

3.2.1 流程概览

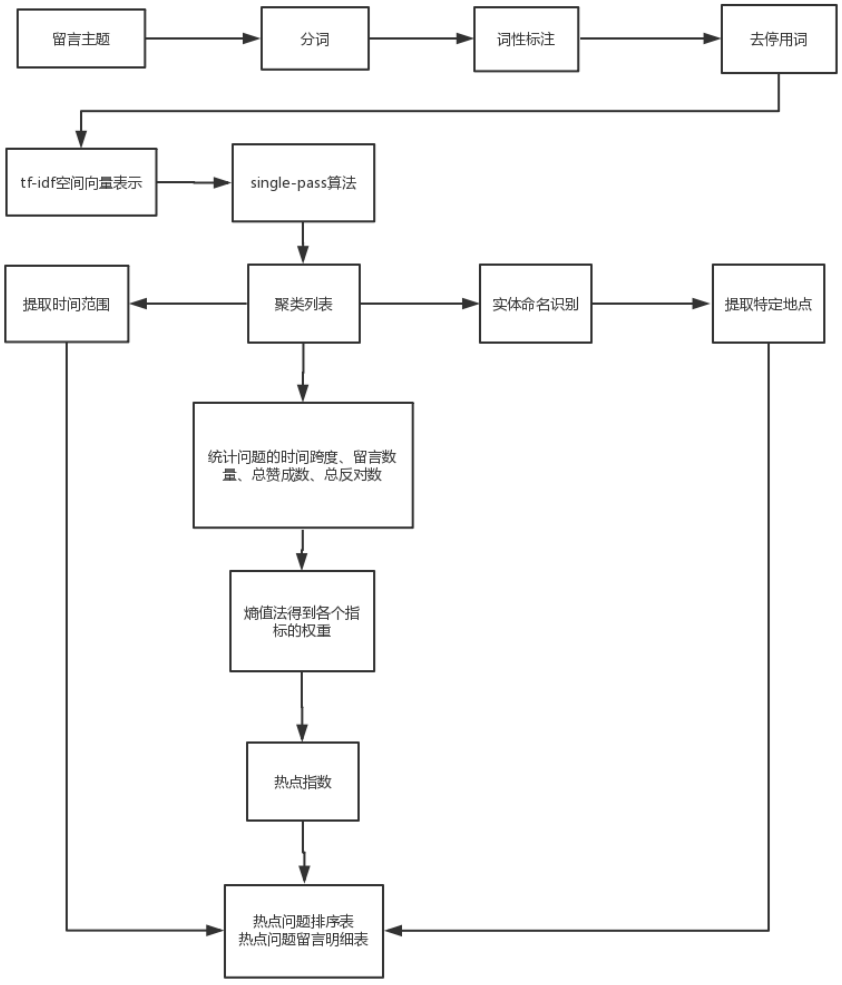

图 13. 问题 2 流程示意图

3.2.2 数据预处理

3.2.2.1 分词

留言主题存在少量无效主题，采取人工总结留言详情摘要的方法替换无效主题，保证留言数据的完整性。针对本题，本组仍然采用 jieba 分词，但是分词结果往往存在切分错误，比如“魅力之城”被切分为“魅力”“之城”。所以我们加入自定义词典以提高切词准确度。

3.2.2.2 词性标注和去停用词

词性标注是指对分词后的词赋予词性的过程。为得到留言主题的关键词，我们只保留有意义的词性的词，去除介词等不必要词；在此题中我们使用的是 pyltp 这个第三方包来进行词性标注[7]；同时我们也引入停用词表与切词结果进行匹配，去除停用词。

3.2.2.3 TF-IDF 算法

TF-IDF 算法是一种基于特征词在文本中出现的频率，计算文本权重的方法，用来评估一个词条对于一个文件集中一份文件的重要程度。TF-IDF 由 TF 及 IDF 两部分组成[8]。

1、TF(词频，Term Frequency)表示词语在文本中出现的频率，词语的词频与该词描述文本的能力呈正比。如果某个词在一篇文章中出现的频率 TF 高，而在其他文件中很少出现，则认为此词条具有很好的类别区分能力，适合用来分类。公式如下：

$$TF_{i,j} = \frac{n_{i,k}}{\sum_k n_{k,j}}$$

分子是词条t_i在文件d_j中出现的次数，分母是文件d_j中所有词条出现的次数之和。

2、IDF(逆文档频率，Inverse Document Frequency)表示文本数据集中文本的总数与包含给定词语的文本数之商，公式如下：

$$IDF_i = log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

对数内的分子是文件总数，分母是包含词条t_i的文件数，如果该词不存在，就会导致分母为零，因此一般使用$1 + |\{j: t_i \in d_j\}|$作为分母。

3.2.3 Single-Pass 算法

Single-Pass 是一种无监督的单遍聚类方法，主要针对流式文件进行单遍聚类，聚类个数具有不确定性。以不同的顺序输入文本，得到的结果各不相同，这里我们控制输入的顺序保持不变，利用余弦距离计算文本间的相似度。

3.2.3.1 Single-Pass 算法步骤

1、接收文本向量列表及留言主题列表；
2、将输入的第一个留言主题加入到话题列表中作为第一个话题 D1；
3、将输入的文本向量与话题列表中的所有文本向量进行相似度计算，并选出相似度最大的话题 Di 的相似度；
4、若相似度大于阈值 theta，则将该向量对应的留言主题归类为话题 Di，若相似度小于阈值 theta，则将文本向量对应的留言主题创建为新话题；

5、以此循环往复，实现所有文本的自动聚类，聚类结束[9]。

3.2.3.2 Single-Pass 算法聚类流程示意图

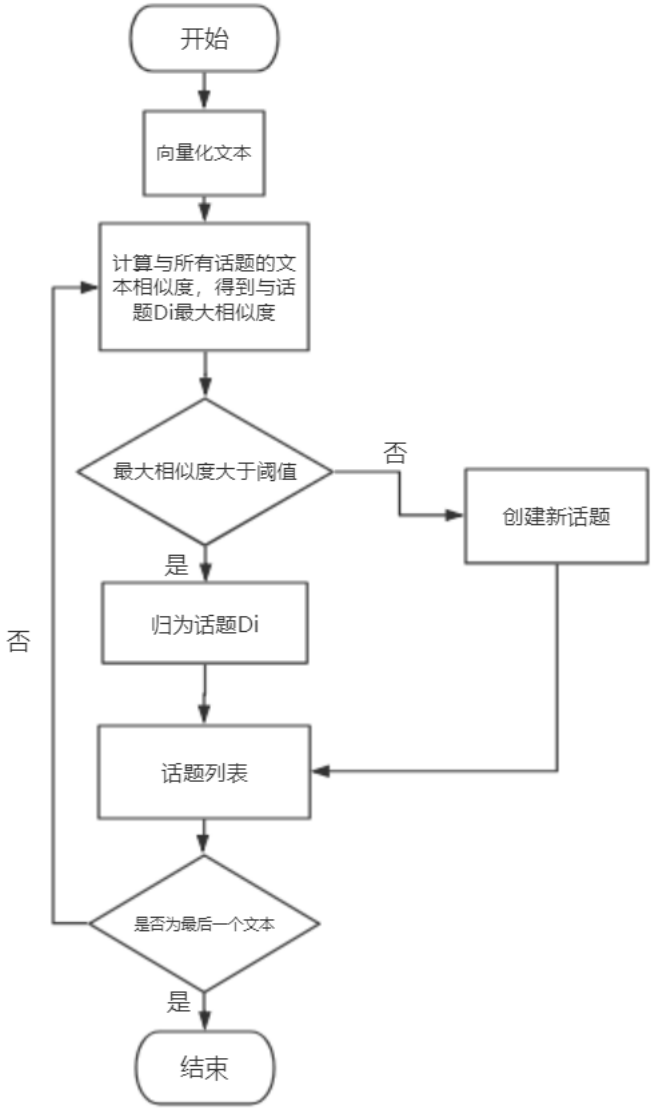

图 14. Single-Pass 算法聚类流程示意图

3.2.4 实体命名识别

实体命名识别（Named Entity Recognition，简称 NER），是指识别出文本中的人名、地名、机构名、专有名词等。我们使用实体命名识别从每个话题中提取出该话题反映的特定地点及特定人群。本文使用 hanlp 封装的实体命名识别方法提取地点[10]。其原理是一种基于层叠隐马尔可夫模型的中文命名实体识别方法。

同一话题中，不同留言主题提取出的地点存在不完整性。因此我们将提取出的地点进行组合得到最终的地点。

3.2.5 问题描述句提取

TextRank 算法是一种文本排序算法，由谷歌的网页重要性排序算法 PageRank 算法改进而来，它能够从一个给定的文本中提取出该文本的关键词、关键词组，并使用抽取式的自动文摘方法提取出该文本的关键句[11]。PageRank 算法根据万维网上页面之间的链接关系计算每个页面的重要性；而 TextRank 算法将词视为“万维网上的节点”，根据词之间的共现关系计算每个词的重要性，并将 PageRank 中的有向边变为无向边。该算法在 3.3.2.2.1 中有具体介绍。

3.2.6 热点指数计算

3.2.6.1 计算公式

问题的热点度与时间跨度 T、留言主题总数 S、反对总数 A、赞成总数 D 呈正相关。下面给出热点指数的计算公式：

$$H = \omega_1 T + \omega_2 S + \omega_3 A + \omega_4 D$$

其中 ω_i 表示各指标权重。

3.2.6.2 权重计算

熵值法是计算指标权重的经典算法之一，它是指用来判断某个指标的离散程度的数学方法[12]。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也就越大。根据熵的特性，我们可以通过计算熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响越大。在问题二中，我们使用该方法来计算各项指标的权重。

3.2.6.2.1 算法步骤

1、数据有 n 行记录，4 个变量，数据可以用一个 n×4 的矩阵 A 表示(n 行 4 列，即 n 行记录数，4 个特征列)[12]

$$A = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ & \cdots\cdots & & \end{bmatrix}$$

2、计算第 j 项指标下第 i 个记录所占比重

$$P_{ij} = \frac{x_{ij}}{\sum_1^n x_{ij}} (j = 1,2,3,4)$$

3、计算第 j 项指标的熵值

$$e_j = -k \times \sum_1^n P_{ij} \times \log P_{ij} \qquad k = \frac{1}{\ln(n)}$$

4、计算第 j 项指标的差异系数

$$g_j = 1 - e_j$$

5、计算第 j 项指标的权重

$$\omega_j = \frac{g_j}{\sum_1^4 g_j}$$

3.2.6.2.2 计算结果

ω_1	ω_2	ω_3	ω_4
0.16098887297264244	0.14623540587399741	0.37718640926412933	0.31558931188923073

3.3 问题三分析方法与过程

针对相关行政部门的答复意见，利用相关性、完整性、可解释性、及时性四个指标，衡量用户的留言主题与答复意见之间的关系，最终以答复评分的形式反映相关行政部门对问题的解决效率、能力和成效。

3.3.1 流程概览

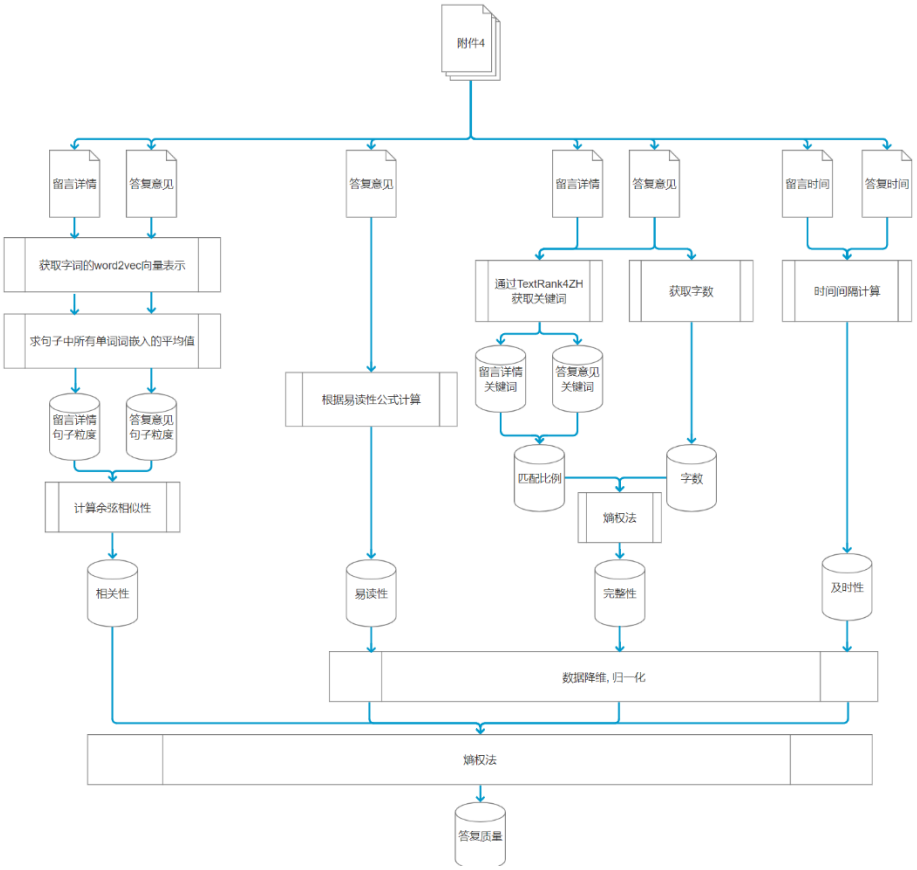

图 15. 问题 3 流程示意图

3.3.2 评价指标

3.3.2.1 相关性

为判断留言详情与答复意见的相关性，首先分别获取留言详情和答复意见的word2vec 向量表示，然后求句子中所有单词词嵌入向量的平均值，用该平均值来表示句子粒度的词嵌入向量，最后我们通过计算留言详情句子粒度词嵌入向量和答复意见的句子粒度词嵌入向量之间的余弦相似性来衡量它们之间的相似程度[13]。

3.3.2.1.1 余弦相似性

对于 n 维向量：假定 A 和 B 是两个 n 维向量，A 是 [a1, a2, ..., an] ，B 是 [b1, b2, ..., bn] ，则 A 与 B 的夹角 θ 的余弦值（即余弦相似性）为：

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

余弦值的范围在[-1, 1]之间，值越趋近于 1，代表两个向量的方向就越接近，两个向量就越相似；越趋近于-1，他们的方向越相反；越趋近于 0，表示两个向量越近乎于正交。

3.3.2.2 完整性

完整性我们理解为有没有完整的回答留言用户的问题。因此，针对该评价指标，我们决定通过关键词匹配率和字数这两个特征，来衡量答复意见的完整性。首先通过 TextRank 算法分别获取留言详情和答复意见的关键词，然后计算关键词的匹配率，再获取答复意见的字数，而后通过熵值法计算两项指标各自的权重，最后计算权值，以衡量答复意见的完整性。

3.3.2.2.1 TextRank 算法

TextRank 算法是一种基于图的用于文本的排序算法,其思想来源于谷歌用于搜索引擎的 PageRank 算法。TextRank 算法通过把文本分割为若干组成单元并建立图模型, 利用投票机制对文本中重要的成分进行排序，从而实现文本关键词的提取。

算法的公式如下：

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

TextRank 中一个单词 i 的权重取决于与在 i 前面的各个点 j 组成的(j,i)这条边的权重，以及 j 这个点到其他边的权重之和。

算法步骤：

1、对给定的文本 T 按照完整的句子进行分割，即$T = [S_1, S_2, \ldots, S_m]$。

2、对每个句子$S_i \in T$进行分词和词性标注，同时过滤停用词，保留指定词性的单词，如名词、动词、即$S_i = [t_{i,1}, t_{i,2}, \dots, t_{i,n}]$，其中$t_{i,j} \in S_j$是保留的候选关键字

3、构建出候选关键词图$G = (V, E)$, V 为结点集合，由步骤二生成的候选关键词组成。然后根据结点间共现词的个数来构造结点之间的边。

4、根据上面的公式进行迭代传播各点的权值，直到收敛。

5、对结点的权值进行排序，从而得到最重要的几个关键字作为候选关键字。

6、将步骤五得到的关键字在原始文本中进行标记，若形成词组则组合为多词关键字， 加入关键字序列[14]。

3.3.2.2.2 熵值法

针对完整性的指标权重计算，我们依然选用熵值法，选择关键词匹配率和字数两项作为评估完整性的指标。其中关键词匹配率计算公式如下：

$$\text{关键词匹配率} = \frac{\text{答复意见中属于留言详情的关键词数}}{\text{留言详情中的关键词数}}$$

算法步骤[12]：

1、数据有 n 行记录，2 个变量，数据可以用一个 n×2 的矩阵 A 表示(n 行 2 列，即 n 行记录数，2 个特征列)

$$A = \begin{bmatrix} x_1 & x_2 \\ \dots & \dots \end{bmatrix}$$

2、计算第 j 项指标下第 i 个记录所占比重

$$P_{ij} = \frac{x_{ij}}{\sum_1^n x_{ij}} (j = 1,2)$$

3、计算第 j 项指标的熵值

$$e_j = -k \times \sum_1^n P_{ij} \times \log P_{ij} \qquad k = \frac{1}{\ln(n)}$$

4、计算第 j 项指标的差异系数

$$g_j = 1 - e_j$$

5、计算第 j 项指标的权重

$$W_j = \frac{g_j}{\sum_1^2 g_j}$$

6、计算权值

$$\text{权值} = \text{关键词匹配率} \times W_1 + \text{字数} \times W_2$$

3.3.2.3 可解释性

可解释性我们理解为易读性。首先我们搜索到汉字常用字表[15]和难字表，然后选择陈世敏的易读性公式[16]作为衡量可解释性的指标。易读性分数越高，答复

意见的可解释性越低。

1、两个指标计算公式如下：

$$\text{每句平均字数} = \frac{\text{答复意见中的总字数}}{\text{答复意见中的句子数量}} \qquad \text{难字半分比} = \frac{\text{答复意见中的难字数}}{\text{答复意见中的总字数}}$$

（备注：句子数量通过计算整体段落中有几个中断语气和标点符号来得出，标点符号包括逗号、句号、冒号、分号、破折号、感叹号和问号。）

2、易读性公式如下：

$$Y = 0.4X_1 + X_2^{[16]}$$

其中 Y 为易读性分数，X_1为每句平均字数，X_2为难字百分比。并且考虑到参考文献中的文本是新闻，而本题文本为答复意见，陈述的语言形式比新闻稍显白话，所以每句平均字数的权重降低到 0.4。

3.3.2.4 及时性

选择间隔时间作为衡量及时性的指标。间隔时间越长，及时性越低。公式如下：

间隔时间=答复时间-留言时间

3.3.3 归一化处理

由于各个指标的数据范围不一样，因此我们通过归一化处理将各个指标的范围压缩到[0,1]区间中。由于完整性、可解释性和及时性这三个指标的最大值和平均值之间的差距过大，因此先采用 10 为底的对数来减少数据间的差距，然后再通过归一化公式进行归一化。

以下展示的是三项指标的分数在进行对数转换之前与之后的结果：

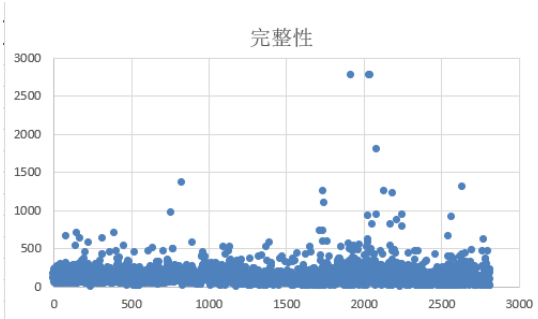

图 16. 完整性得分范围

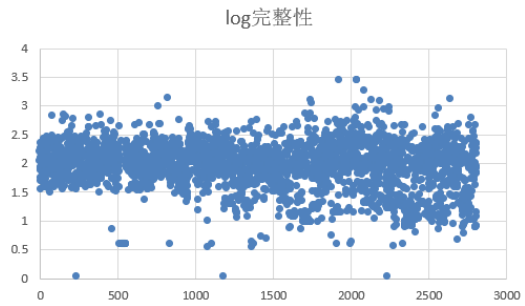

图 17. 对数转换后完整性得分范围

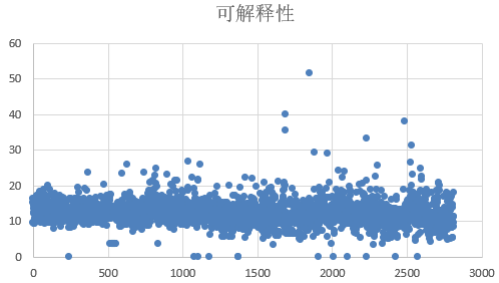

图 18. 可解释性得分范围

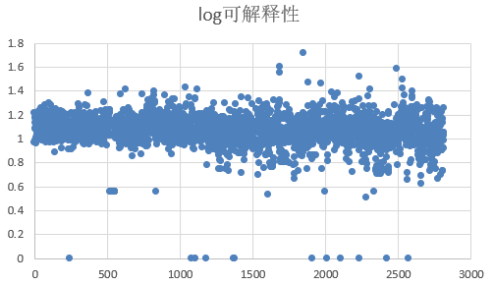

图 19. 对数转换后可解释性得分范围

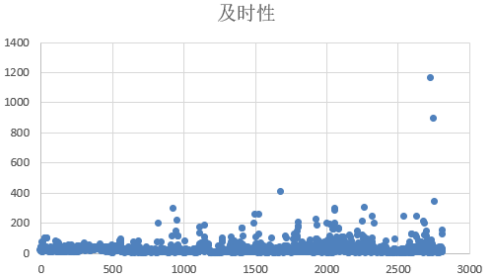

图 20. 及时性得分范围

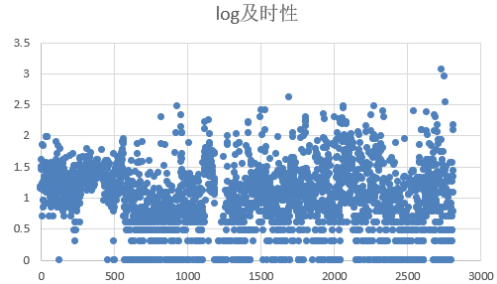

图 21. 对数转换后及时性得分范围

四、实验结果

由于数据量较大，所以本组对这三道题只做了部分数据的截图，结果如下：

4.1 群众留言分类

	A	B	C
1		留言详情	标签
2	0	A3区 大道 西行 便 道 未管 路口 加油站 路段	城乡建设
3	1	位于 书院 路 主干道 在水一方 大厦 一楼 四楼	城乡建设
4	2	尊敬 领导 A1区 苑 小区 位于 A1区 火炬 路 小	城乡建设
5	3	A1区 A2区 华庭 小区 高层 二次 供水 楼顶 水	城乡建设
6	4	A1区 A2区 华庭 小区 高层 二次 供水 楼顶 水	城乡建设
7	5	2015 年 购买 盛世 耀凯 小区 17 栋 3 楼 4 楼	城乡建设
8	6	西地省 地区 常年 阴冷 潮湿 气候 近年 气候	城乡建设
9	7	尊敬 胡书记 您好 家住 A市 A3区 桐梓 坡 西路	城乡建设
10	8	梅家田 社区 辖区 内 小区 居民 每年 都 依法	城乡建设
11	9	尊敬 A市 政府 领导 您们好 A市 A3区 魏家坡巷	城乡建设
12	10	尊敬 A市 政府 领导 您们好 A市 A3区 魏家坡巷	城乡建设
13	11	请求 依法 监督 泰华 一村 小区 第四届 非法 业	城乡建设
14	12	住 梅 溪湖 壹号 御湾 4 楼 2019 年 8 月份 住	城乡建设
15	13	尊敬 领导 你们好 A市 A4区 捞刀河 镇 彭家巷	城乡建设
16	14	地铁 5 号线 施工 导致 万家 丽路 锦楚 国际	城乡建设
17	15	尊敬 领导 你好 A6区 润 紫 郡 业主 今年年初	城乡建设
18	16	A市 A5区 朝晖路 锦楚 国际 新城 三区 6 月份	城乡建设
19	17	肯定 选择 A9 市 西南角 支持 A9 市 西南角 设	城乡建设
20	18	尊敬 领导 A6区 几年 发展 突飞猛进 城市道路	城乡建设
21	19	A5区 楚府 线 包括 森林 雅苑 楚府 十城 天际	城乡建设
22	20	涂 愈 一名 普通 建筑业 从业者 求助 请求 相	城乡建设

Sheet1

图 22. 标签分类模型结果

4.2 热点问题挖掘

热度排名	问题ID	热点指数	地点/人群	时间范围	问题描述	对应id
1	1	845.0989177	A市A5区K9县	2019/01/15 10:29:32至2019/1	A市A5区汇金路五矿万境K9县存在	438, 818, 843, 1150, 1924, 2695, 3150, 3684, 4031
2	2	693.1970234	A市A3区金毛湾	2019/04/11 21:02:44至2019/0	A3区卓越浅水湾南门对面金毛房	1483, 3766
3	3	587.7280618	A市A4区	2019/02/21 18:45:14至2019/0	承办A市58车贷案警官应跟进关	249, 1383
4	4	334.9396715	A市	2019/02/25 09:58:37至2019/0	请A市依法查处“爱玩客ivankr	1212, 1839
5	5	297.6703926	A市A4区绿地海外滩	2019/01/30 10:00:04至2019/0	A市至赣州高铁对绿地海外滩二	154, 596, 1186, 2313, 2965, 3192, 3333
6	6	149.0798683	A市经济学院体育学	2017/06/08 17:31:20至2019/1	A市经济学院强制学生外出实习	4321, 4322, 4323, 4324, 4325
7	7	91.63431779	A市京港澳高速长楚	2019/01/10 15:01:26至2019/1	建议西地省尽快外迁京港澳高速	2723, 4080
8	8	91.42534645	A市富绿物业丽发新	2019/06/19 23:28:27至2019/0	A市富绿物业丽发新城强行断业	205
9	9	87.4823749	A市房云时代小区三	2019/01/04 11:36:26至2019/1	投诉A市长房云时代小区配套幼	26, 1389, 1432, 1629, 1692, 2084, 2494, 2951, 3827,
10	10	69.64488611	A6区月亮岛路	2019/03/26 10:17:31至2019/0	关于A6区月亮岛路沿线架设110	1278, 1824, 1962, 2798, 3122, 3382, 3546
11	11	68.95675186	A市A7县汽车南站	2019/01/21 18:16:03至2020/0	A市汽车南站过渡站售票窗口只	898, 1577, 1632, 2782, 3969, 4134, 4156
12	12	68.10927289	A市A7县灰埠大市场	2019/01/16 17:01:25至2019/0	请加快A市各市场乱摆乱放整治	521, 1287, 3890, 3928
13	13	64.4540957	A市	2018/11/15 16:07:12至2019/1	咨询A市人才购房补贴通知问题	51, 643, 723, 1514, 2373, 3279, 3375, 3467, 3962, 40
14	14	63.56784578	A市地铁1号线地铁7	2019/02/15 15:04:49至2019/1	强烈建议将地铁7号线南延至A市	798, 870, 1697, 2283, 2325
15	15	62.33685905	A市经开区东六线	2019/01/02 20:27:07至2019/0	问问A市经开区东六线以西泉塘	1903, 2156, 2506, 2866
16	16	60.32093315	B市A2区	2019/01/15 13:27:13至2020/0	对远期A市地铁的建议	8, 278, 1223, 1691, 1863, 2188, 2578, 2649, 3077, 32
17	17	60.2106704	A市地铁3号线星沙大	2019/01/03 10:34:29至2019/1	A市地铁3号线在深业睿城小区有	15, 1446, 3018, 4232
18	18	60.14998251	A市美麓阳光小区	2019/01/01 11:43:20至2019/1	A市美麓阳光项目处遗问题久久	471, 1240, 1274, 1395, 3979
19	19	59.35766043	A市A7县交通运输局	2019/01/09 10:07:09至2019/1	A市205路公交车经常不按时发车	9, 579, 820, 1313, 1443, 1586, 2265, 2338, 2521, 259
20	20	59.17469775	E市楚税	2019/01/16 08:30:59至2020/0	咨询A市缴纳社保及购房等问题	1399, 1402, 1643, 2484, 2568, 2831, 3883, 4061
21	21	59.01378616	A市	2019/01/16 09:20:50至2020/0	A市2017年出租车燃油补贴为何	1106, 1513, 2327, 3719, 4175
22	22	56.86525331	A市A7县保利香槟国	2019/03/13 23:53:56至2020/0	投诉A7县保利香槟国际开发商	516, 598, 1068, 2334, 2478, 3773, 3861

图 23. 热点问题排序及聚类结果

热度排名	问题ID	热点指数	地点/人群	时间范围	问题描述
1	1	845.0989177	A市A5区K9县	2019/01/15 10:29:32至2019/11/11 16:30:3	A市A5区汇金路五矿万境K9县存在一系列问题
2	2	693.1970234	A市A3区金毛湾	2019/04/11 21:02:44至2019/09/25 00:17:5	A3区卓越浅水湾南门对面金毛房地产长期半夜施工
3	3	587.7280618	A市A4区	2019/02/21 18:45:14至2019/03/01 22:12:3	承办A市58车贷案警官应跟进关注留言
4	4	334.9396715	A市	2019/02/25 09:58:37至2019/08/21 16:43:2	请A市依法查处“爱玩客ivankr”特大电信诈骗案
5	5	297.6703926	A市A4区绿地海外滩小区	2019/01/30 10:00:04至2019/09/06 18:36:1	A市至赣州高铁对绿地海外滩二期小区影响太大了

图 24. 热点问题表

问题ID	热点指数	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	留言id	主题索引
1	845.098918	198961	A000103957	反映A5区主塘路五矿万	2019/11/11 16:30:39	尊敬的领导：你好！现反映：长期以来，人	0	3	438	391
1	845.098918	208069	A00094436	A5区五矿万境K9县的开	2019/5/5 13:52:50	本人是A5区洞井街道汇金路五矿万境K9县5	0	2	818	391
1	845.098918	208636	A00077171	A市A5区汇金路五矿万	2019/8/19 11:34:04	我是A市A5区汇金路五矿万境K9县24栋的一	0	2097	843	391
1	845.098918	215507	A000103230	A市五矿万境K9县存在	2019/9/12 14:48:07	预交房23栋没有通往负一楼的楼梯，存在	0	1	1150	391
1	845.098918	234086	A00099869	A市五矿万境K9县房子	2019/6/20 9:30:44	五矿万境K9县的房子又出问题了，又是墙	0	6	1924	391
1	845.098918	252650	A00010531	A市五矿万境K9县交房	2019/9/11 15:16:02	尊敬的相关部门，本人家庭于2018年购置人	0	0	2695	391
1	845.098918	262599	A000100428	A市五矿万境K9县房屋	2019/9/19 17:14:49	我是西地省A市五矿万境K9县的业主，201	0	0	3150	391
1	845.098918	275491	A00061339	A市五矿万境K9县负一	2019/9/10 9:10:22	关于五矿万境·K9县负一楼面积缩水的反馈	0	0	3684	391
1	845.098918	283732	A00021495	A市五矿万境水岸三期	2019/1/15 10:29:32	我们是A市五矿万境水岸三期业主。现举报	0	0	4031	391
2	693.197023	223297	A00087522	反映A市金毛湾配套入	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛湾的一	5	1762	1483	1156
2	693.197023	277139	A00024012	A3区卓越浅水湾南门对	2019/9/25 0:17:51	梅溪湖卓越浅水湾南门对面金毛房地产长	0	0	3766	1156
3	587.728062	194343	A000106161	承办A市58车贷案警官	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发受害人	0	733	249	228
3	587.728062	220711	A00031682	请书记关注A市A4区58	2019/2/21 18:45:14	尊敬的胡书记：您好!A4区p2p公司58车贷	0	821	1383	228
4	334.939671	217032	A00056543	严惩A市58车贷特大集	2019/2/25 9:58:37	胡市长：您好！西地省展星投资有限公司	0	790	1212	974
4	334.939671	232063	A00083732	请A市依法查处“爱玩客	2019/8/21 16:43:21	尊敬的市委书记：您好！我们遭遇了电信	2	20	1839	974
5	297.670393	191951	A00041448	A4区绿地海外滩小区二	2019/8/23 14:21:38	尊敬的领导：你好，近日看到了渝长厦高	0	1	154	142
5	297.670393	202575	A00092007	咨询A市绿地海外滩二	2019/9/4 18:32:42	我们是A市A4区绿地海外滩二期5栋居民，	0	17	596	142
5	297.670393	216316	A00097196	A4区绿地海外滩二期业	2019/9/6 10:16:27	我们是A市A4区绿地海外滩二期居民，201	0	2	1186	142
5	297.670393	243551	A00041448	A市至赣州高铁对绿地	2019/9/1 10:18:48	我们是A市A4区绿地海外滩二期居民，201	0	1	2313	142
5	297.670393	258708	A00092008	A市绿地海外滩二期没	2019/1/30 10:00:04	A市绿地海外滩二期还没有交房，已发现重	0	2	2965	142
5	297.670393	263672	A00041448	A4区绿地海外滩小区周	2019/9/5 13:06:55	您好，近日看到了渝长厦高铁最新的红线	0	669	3192	142

图 25. 所有数据热点分析结果

问题ID	热点指数	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	845.098918	198961	A000103957	五矿万境水岸路段拥堵	2019/11/11 16:30:39	尊敬的领导：你好！现反映：长期以来	0	3
1	845.098918	208069	A00094436	县的开发商与施工方建	2019/5/5 13:52:50	本人是A5区洞井街道汇金路五矿万境K9	0	2
1	845.098918	208636	A00077171	金路五矿万境K9县存在	2019/8/19 11:34:04	我是A市A5区汇金路五矿万境K9县24栋	0	2097
1	845.098918	215507	A000103230	境K9县存在严重的消防	2019/9/12 14:48:07	预交房23栋没有通往负一楼的楼梯，存	0	1
1	845.098918	234086	A00099869	万境K9县房子的墙壁又	2019/6/20 9:30:44	五矿万境K9县的房子又出问题了，又是	0	6
1	845.098918	252650	A00010531	万境K9县交房后仍存在	2019/9/11 15:16:02	尊敬的相关部门，本人家庭于2018年购	0	0
1	845.098918	262599	A000100428	矿万境K9县房屋出现质	2019/9/19 17:14:49	我是西地省A市五矿万境K9县的业主，	0	0
1	845.098918	275491	A00061339	矿万境K9县负一楼面积	2019/9/10 9:10:22	关于五矿万境·K9县负一楼面积缩水的	0	0
1	845.098918	283732	A00021495	境水岸三期违规建设	2019/1/15 10:29:32	我们是A市五矿万境水岸三期业主。现	0	0

图 26. 热点问题留言明细表（部分）

4.3 答复意见的评价

	留言主题	留言详情	答复意见	相关性	完整性	可解释性	及时性	综合评分
2	A2区景蓉华苑物业管理有问题	2019年4月以来，位于A市A2区	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2	0.97856	0.63758	0.354389	0.616216	0.594403
3	A3区潇楚南路洋湖段怎么还没修好？	潇楚南路从2018年开始修，到	网友“A0023583”：您好！针对您反映A3区潇楚南路洋湖段怎么	0.917257	0.586958	0.525505	0.625993	0.615044
4	请加快提高A市民营幼儿园老师的待遇	地处省会A市民营幼儿园众多，	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇	0.953455	0.607213	0.379816	0.625993	0.591877
5	在A市买公寓能享受人才新政购房补贴	尊敬的书记：您好！我研究生	网友“A000110735”：您好！您在平台《问政西地省》上的留言	0.946348	0.589023	0.445939	0.625993	0.600269
6	关于A市公交站点名称变更的建议	建议将“白竹坡路口”更名为	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复	0.930292	0.506396	0.516827	0.616216	0.584623
7	A3区含浦镇马路卫生很差	欢迎领导来A市泥泞不堪的小径	网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差	0.919934	0.552415	0.41375	0.513336	0.537301
8	A3区教师村小区盼望早日安装电梯	尊敬的胡书记：您好！过去在	网友“A000100804”：您好！针对您反映A3区教师村小区盼望早	0.949571	0.559519	0.444014	0.477213	0.535443
9	反映A5区东澜湾社区居民的集体民生讠	我做为一东澜湾社区居民，我	网友“UU00812”您好！您的留言已收悉。现将有关情况回复如下	0.906177	0.677944	0.384537	0.527761	0.575358
10	反映A市芙蓉阳光住宅楼无故停工以及	我是芙蓉阳光a栋803业主，现	网友“UU008792”您好！您的留言已收悉。现将有关情况回复如	0.96137	0.65109	0.397692	0.607069	0.603651
11	反映A市洋湖新城和顺路洋湖壹号小区	胡书记好！根据规划，洋湖新	网友“UU008687”您好！您的留言已收悉。现将有关情况回复如	0.964504	0.548174	0.532927	0.607069	0.60113
12	反映A2区大托街道大托新村违建问题	我家住在A市A2区大托街道大托	网友“UU0082204”您好！您的留言已收悉。现将有关情况回复如	0.954935	0.646908	0.431551	0.397904	0.531814
13	A5区鄱阳村D区安置房人防工程的咨询	胡书记：您好，我想请问一下	网友“UU008829”您好！您的留言已收悉。现将有关情况回复如	0.918676	0.629799	0.453464	0.517983	0.572739
14	A4区万国城小区段请求修建一座人行天	尊敬的书记：我是一名居住在	网友“UU00877”您好！您的留言已收悉。现将有关情况回复如下	0.920714	0.487492	0.496948	0.607069	0.569891
15	举报A市芒果金融平台涉嫌诈骗	尊敬的领导：我们是贵省A市芒	网友“UU0081480”您好！您的留言已收悉。现将有关情况回复如	0.926137	0.446598	0.529937	0.771911	0.625701
16	建议增开A市261路公交车	建议增开A市261路公交车趟数	网友“UU0081227”您好！您的留言已收悉。现将有关情况回复如	0.911128	0.539561	0.515227	0.598477	0.586772
17	关于A市新开铺路与披塘路交叉路口通	2016年下半年新开铺路全线开	网友“UU008444”您好！您的留言已收悉。现将有关情况回复如	0.969793	0.683321	0.482837	0.402012	0.557798
18	投诉A3区桐梓坡路益丰大药房以次充好	12月16日上午，我来到A3区桐	网友“UU0081194”您好！您的留言已收悉。现将有关情况回复如	0.911763	0.492523	0.366153	0.724226	0.58486
19	建议在A市梅溪湖开办一个图书馆	梅溪湖至今没有一个图书馆，	网友“UU008706”您好！您的留言已收悉。现将有关情况回复如	0.936393	0.489302	0.492172	0.590377	0.564514
20	希望相关部门治理A3区中海国际社区一	希望相关部门治理一下中海国	网友“UU008201”您好！您的留言已收悉。现将有关情况回复如	0.930822	0.52818	0.451535	0.636496	0.584547
21	希望A市社保卡、医保卡、居民健康卡	看病需要带社保卡、医保卡、	网友“UU0081681”您好！您的留言已收悉。现将有关情况回复如	0.892938	0.481552	0.476768	0.68861	0.591381
22	希望A市潇楚一卡通尽快支持手机nfc虚	希望潇楚一卡通尽快支持手机	网友“UU0081681”您好！您的留言已收悉。现将有关情况回复如	0.929807	0.530798	0.417888	0.68861	0.597079
23	反映A9市北盛镇对泉水村塘下组土地征	A9市北盛镇对泉水村塘下组土	网友“UU0081500”您好！您的留言已收悉。现将有关情况回复如	0.971119	0.629987	0.439526	0.625993	0.614071
24	呼吁A5区交警大队纠正电子交通警察依	尊敬的市委、市纪委驻市公安	网友“UU0081057”您好！您的留言已收悉。现将有关情况回复如	0.970492	0.531445	0.44469	0.561938	0.559709
25	关于A市地铁轨道8号线北段在楚江北路	A市委市政府：根据国家发改委	网友“UU008162”您好！您的留言已收悉。现将有关情况回复如	0.941446	0.548886	0.470815	0.492144	0.542897
26	咨询A市商业住房贷款转公积金贷款问	深圳市缴存住房公积金，是否	网友“UU0081604”您好！您的留言已收悉。现将有关情况回复如	0.866085	0.456279	0.40954	0.636496	0.546611
27	咨询A市东六线（劳动东路-机场高栗）	A市高铁新城管委会负责建设位	网友“UU008694”您好！您的留言已收悉。现将有关情况回复如	0.941311	0.555989	0.472108	0.625993	0.595091
28	咨询A3区西湖街道茶场村公路规划问题	您好，胡书记，我是A3区西湖	网友“UU008765”您好！您的留言已收悉。现将有关情况回复如	0.922602	0.569582	0.493422	0.625993	0.602696
29	反映A3区靳江洋湖集体资产的有关问题	尊敬的全体村民，靳江洋湖所	网友“UU0082119”您好！您的留言已收悉。现将有关情况回复如	0.905952	0.702302	0.368777	0.575445	0.597323

图 27. 答复意见评价结果

参考文献

[1] 师伟. 基于语义相关的在线热点话题发现算法的研究与应用[D]. 西安:西安石油大学, 2014.
[2] 分词所用的库：https://github.com/fxsjy/jieba
[3] 薛彬. 面向社会治理的文本分类方法研究与应用[D]. 杭州:中国计量大学, 2018.
[4] 停用词库参考：
https://github.com/Pirate-Xing/stopwords
[5] Yoon Kim. Convolutional Neural Networks for Sentence Classification 1408.5882 [cs.CL], 2014.
[6] KMAXCNN 模型编写参考的框架：https://kashgari-zh.bmio.net/
[7] 词性标注所用的自然语言处理包：
https://github.com/HIT-SCIR/pyltp
[8] TF-IDF 算法简介：https://blog.csdn.net/leaf_zizi/article/details/82684921
[9] single-pass 算法代码参考：https://zhuanlan.zhihu.com/p/74810166
[10] 提取地点用到的 python 库：
https://github.com/hankcs/HanLP
[11] TextRank 算法简介：
https://blog.csdn.net/wotui1842/article/details/80351386
[12] 熵值法简介：
https://blog.csdn.net/qq_24975309/article/details/82026022
[13] 计算文本相关度所用的库：
https://github.com/shibing624/text2vec
[14]TextRank 算法代码参考，从中文文本中自动提取关键词和摘要：
https://github.com/letiantian/TextRank4ZH
[15] 常用字表来源：
https://wenku.baidu.com/view/eaf27b18777f5acfa1c7aa00b52acfc788eb9f6a.html
[16] 黄敏. 汉语特质与中文新闻易读性公式研究[J]. 新闻与传播研究, 2010(4):93-97