基于多因素 TF-IDF 的文本挖掘应用

摘 要

本文旨在利用自然语言处理及文本挖掘技术进行留言分类和热点整理等任务，打造更加智能的“智慧政务”管理系统，取代以往的人工操作，一定程度上提升政府的管理水平和施政效率。

首先，对附件的留言数据进行分词、过滤停用词等预处理之后，对各留言提取融合词性、位置、词长等多个因素的词频-逆文档频率(MF-TF-IDF)特征。经验证，MF-TF-IDF 提取关键词的效果明显高于其余特征。其次，通过比较决策树、贝叶斯网络、支持向量机(SVM)等多个分类器基于关键词对留言的分类结果，选择分类预测结果最优的 SVM 作为本文的留言分类模型，预测精度为 88.6%，F1-score 为 0.871(表 8)。接着，基于 MF-TF-IDF 计算各留言之间的余弦相似度，通过对比相似度搜索热点问题。并以热点问题的留言、点赞和反对总数为热度评价指标，利用熵权法建立热度评价模型，筛选出热度指数排名前五的热点问题，包括车贷案件、人才新政和车位捆绑销售等问题(表 10)。最后，本文从完整性，可解释性，答复信息量三个指标对留言的答复意见质量进行评价。在量化指标后，分别使用 PCA 数据挖掘算法和 AE 自编码器无监督学习算法进行信息压缩，进而对各留言的答复意见评分。经过主观检验，该算法对答复意见质量的评分较为客观。在文章的最后，对答复意见作出了总体评价，并讨论了本文的不足与展望。

关键词：文本挖掘；MF-TF-IDF；支持向量机；AE 自编码器

Text mining application based on Multi Factors TF-IDF

Abstract

The purpose of this paper is to use natural language processing and text mining technology to carry out message classification and hot-spot sorting tasks, to create a more intelligent "smart government" management system, replace the previous manual operation, to a certain extent to improve the level of government management and governance efficiency.

First of all, after the content of message data is parted, filtered and deactivated words and other pre-processing, the word frequency-reverse document frequency (MF-IDF) characteristics of the various messages quims are extracted with fusion word sex, location, word length and so on. It has been proved that the effect of MF-TF-IDF extraction keywords is significantly higher than that of the remaining features. Secondly, by comparing the decision tree, Bayesian network, support vector machine (SVM) and other classifiers based on keywords on the classification results of messages, select the classification prediction results of the best SVM as the message classification model of this paper, the prediction accuracy is 88.6%, F1-score is 0.871(Table 8). Then, based on MF-TF-IDF, the cosine similarity between messages is calculated, and hot issues are searched by contrasting similarity. And the total number of hot issues of the message, likes and opposition to the heat evaluation index, the use of entropy power method to establish a heat evaluation model, screening out the heat index top five hot issues, including car loan cases, talent New Deal and parking bundling and other issues (Table 10). Finally, this paper from the integrity, interpretability, reply information volume three indicators to the message of the response to the quality of comments to evaluate. After the quantitative indicators, the PCA data mining algorithm and The AE autoencoder unsupervised learning algorithm are used respectively to compress the information, and then score the responses to each message. After subjective examination, the algorithm is more objective in the quality of the responses. At the end of the article, the overall evaluation of the reply comments is made, and the shortcomings and prospects of this paper are discussed.

Keywords: Text Mining; MF-TF-IDF; Support Vector Machine; AE autoencoder

目　录

1 绪论

本次“智慧政务”文本挖掘应用的目的是响应党的十八届三中全会上提出的“推进国家治理体系和治理能力现代化”以及党的十九大提出的“加强和创新社会治理”，用科学的、现代的方法治理社会网络舆情问题。本次数据挖掘主要是为了改进智慧政务系统中的几个功能，包括对留言进行分类、搜索系统所有留言中的热点问题以及对留言答复意见质量的评价。利用本次挖掘所提出的模型与算法，一些原本依靠人工根据经验处理的工作能够被人工智能所代替，一定程度上提升了政府的管理水平和施政效率。

1.1 研究背景及意义

根据中国互联网信息中心发表的第 42 次《中国互联网络发展状况统计报告》显示：2018 年我国网民人数为 8.02 亿。如此庞大的网民群体可通过微信、微博、市长信箱、阳光热线等网络问政平台发表对社会事件的观点和评价。近年来，网络舆论逐渐成为政府了解民意、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。

1.2 挖掘目标

1.2.1 群众留言分类

目前，大部分电子政务系统仍然是依靠人工根据经验对网络问政平台的留言进行划分体系。但是随着网络的普及，留言数量的急剧增长，人为划分的方式效率十分低。本文的第一个挖掘目标就是提供智能分类算法。根据附件 1 的分类三级标签以及附件 2 的留言数据，建立关于留言内容的一级标签分类模型，并对该模型进行评价。

1.2.2 热点问题挖掘

热点问题是指某一时间段内反映特定地点或特定人群问题的留言。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。然而，随着留言数量的增长，人为的方式很难分析得出社会存在的热点问题。本文对附件 3 的留言归类于各个存在的热点问题中，定义合理的热度评价指标，建立一个关于社会问题的热度评价模型。通过该模型选择出附件 3 中存在的热点问题，并输出排名前 5 的热点问题，保存为文件“热点问题表. xls”以及相应的热点问题对应的留言信息，保存为文件“热点问题留言明细表. xls”。

1.2.3 答复意见评价

最后我们给有关部门提出一套对网络问政平台的答复意见质量的评价方案。

本文将从附件 4 的留言详情以及答复意见的完整性、可解释性以及答复信息量等角度建立模型，用以评价留言平台答复意见的质量。

1.3 挖掘过程

本文的总体挖掘过程主要包括以下几个步骤：

Step1 文本预处理：

预处理文本的主要目的是通过控制词汇量的大小来减少问题的维数。文本分类的四个常用预处理步骤是分词，标准化，停止词删除和词干分析。标记化将文本分成单词或其他有意义的部分。

(1) 分词：利用第三方库 jieba 进行中文分词；
(2) 标准化：转换和添加或删除文本中的词语，例如丢弃文章中词少于 10 个单词的，丢弃词汇量少于 2 个或超过 6 个字符的词语，删除时间字符等；
(3) 过滤停用词：停止词是指文本中经常遇到的但与分析无关的词，如的、在、是以及语气词、副词等等；

预处理数据集后，我们用方差阈值法进行特征选择，对特征进行筛选，进而剔除了部分携带信息量少的特征。

Step2 文本特征提取：

本文选用逆文档频率 TF-IDF 向量作为文本特征，基于所有留言内容构造的语料库，计算各个留言的 TF-IDF，作为下文的分类及聚类任务中的输入数据。

经过实验表明，基于传统的 TF-IDF 特征计算所得到的结果不够理想，本文参考文献[1]中的方法构造出融合多因素的 TF-IDF 特征，将其称为 MF-TF-IDF。其中多因素包含了：文本的位置、词性、词长等因素。

Step3 支持向量机文本多分类：

传统的分类器有贝叶斯网络（BN）、决策树（DecisionTree）、逻辑斯蒂回归（LR）、K 近邻法（KNN）以及支持向量机（SVM）模型，本文通过比较上述模型对附件 2 留言数据的分类结果，最终决定选择使用 SVM 作为分类器。

Step4 文本相似度热点搜索：

基于 MF-TF-IDF 计算各留言之间的余弦距离，设置阈值，将相似度高的留言归入同一热点问题中，循环遍历所有的留言。根据留言、点赞及反对数量建立热度评价模型，选择处排名前 5 的热点问题。

Step5 答复意见质量评价：

以留言答复意见的完整性、可解释性以及答复信息量作为评价指标。对各指标进行量化，分别用 PCA 数据挖掘算法及 AE 自编码器无监督学习进行信息压缩，从而对各留言答复意见评分。

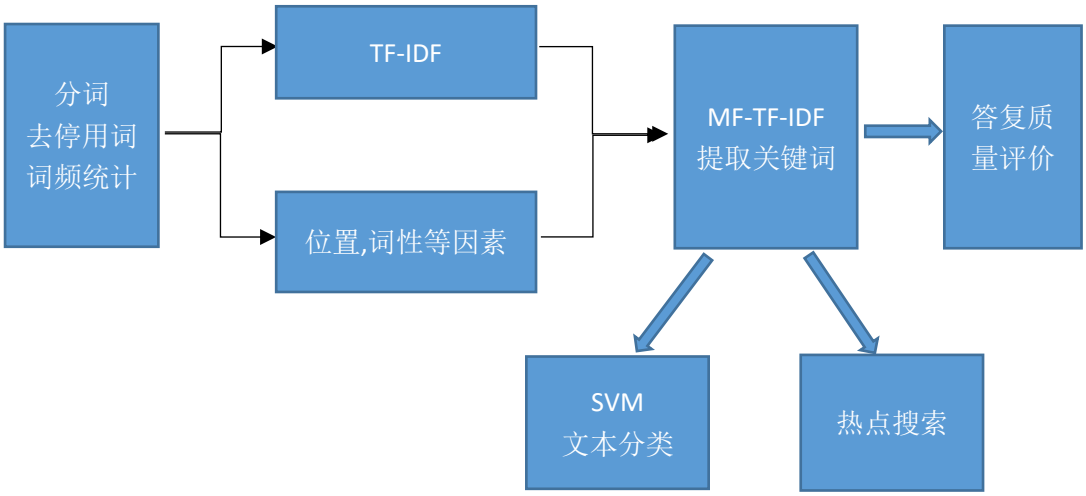

图 1 文本挖掘流程图

2 多因素 TF-IDF 关键词算法

2.1 TF-IDF 算法

TF-IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率)是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

词频(term frequency, TF)指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文档。下面给出 TF 计算公式：

$$\mathrm{TF_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} , \tag{2-1-1}$$

其中$\mathrm{TF_{i,j}}$是 j 文档中第 i 个词的 TF 值，$n_{i,j}$则是 j 文档中第 i 个词的数量。

逆向文件频率(inverse document frequency, IDF)IDF 指的是，如果包含词 w 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。下面给出 IDF 的计算公式：

$$\mathrm{IDF} = \log \frac{|D|}{1+|\{d \in D : w \in d\}|} , \tag{2-1-2}$$

其中$|D|$表示文档总数，$|\{d \in D : w \in d\}|$表示包含词 w 的文档数，这里分母加一是防止包含词 w 的文档数为 0 而导致分母为 0 无法计算。

最终的 TF-IDF 值由以下公式计算

$$\text{TF-IDF} = \mathrm{TF} * \mathrm{IDF} . \tag{2-1-3}$$

2.2 融合多因素 TF-IDF

本文参考文章[1]对 TF-IDF 算法进行改进。考虑到 TF-IDF 权重的计算方法偏向于统计文档中各个词 w 出现的频率，没有考虑词语本身的一些属性对于权重的影响，导致关键词提取效果并不理想。

本文主要针对 TF-IDF 进行研究，综合考虑了文本中其它一些词语的属性，比如词性、词长度、词跨度、词语关联性和位置。考虑到留言数据集中留言详情长度，分句数等一些属性是不规范的，因此我们在本文中不考虑词语关联性和词跨度这两个对文本规范程度要求较高的属性。最终通过一些分析和试验，本文最终选取词性，位置和词长度三个属性作为补充，对 TF-IDF 权重进行改进。本文称这种改进的方法为 Multi Factors TF-IDF 算法，简称 MF-TF-IDF 算法。

2.2.1 位置权重

参考文章[5,6]，文本的标题包含文本主要的中心思想，因此标题中的词语的权重应该相对较大。除此之外，文本的第一段和最后一段往往是对文本中心思想的概括，因此文本第一段和最后一段中出现的词语重要性应大于其他位置出现的词语，即文本第一段和最后一段中的词语权重要应大于其余位置的权重。

本文将留言主题作为留言标题，首句末句分别视为文本中的首段和末段，给出最终的位置权重参数设置如表 1 所示。

表 1 词语位置权重

词语位置	权重名称	权重设置
留言主题	W_title	6
留言详情首句	W_firstsen	5
留言详情末句	W_lastsen	4
留言详情其他	W_other	1

2.2.2 词性权重

中文单词词性可以分为实词和虚词，其中实词包含名词、动词、形容词、数词、量词和代词。虚词包含副词、介词、连词、助词、叹词、代词等。参考文章[8]，中文关键词词性一般是名词或名词性短语，其次是动词，最后是数词、副词等。考虑词性的权重就可以避免一些经常出现却并不能表达一些关键意义的词语作为高权重词语输出。

经过一些实验，在留言数据集中，本文给出最终的词性权重参数设置如表 2 所示。

表 2 词性权重

词性	权重名称	权重设置
名词	W_tagn	6
动词	W_tagv	4
形容词、副词	W_taga	3
其他	W_tago	1

2.2.3 词长权重

据统计，中文文本关键词的词长一般不小于 2，因此将词长小于 2 的词语过滤。关键词词长越长，包含了更多信息，但是关键词词长一般不超过 6，因此将词长大于 6 的词语过滤。本文由公式(2-2-1)计算词长权重，

$$\mathrm{W_{len}} = \frac{len}{len+4}. \tag{2-2-1}$$

其中 *len* 表示词长。

2.3 多因素 TF-IDF 关键词提取方法

在文本挖掘中，因为文本中所含有的词语太多，所以提取文本特征成为了至关重要的一步，本文使用 MF-TF-IDF 算法实现对文本关键词的提取并给出 MF-TF-IDF 算法、TF-IDF 算法及 textrank 算法提取留言中关键词示例作为对比。

2.3.1 数据预处理与 MF-TF-IDF 关键词提取算法步骤

本文基于 PYTHON3.7 的 MF-TF-IDF 关键词提取算法步骤为：

(1) **清洗数据**：清除留言详情中的噪声数据，比如一些多余的空格、符号和乱码等。
(2) **位置标注**：分别对留言主题和留言详情的首句末句进行标注，以便于对后续词语位置属性的分类。
(3) 分词：对文本进行带有词性 tag 的分词，本文中使用北京领馆大学海量语言信息处理与云计算工程研究中心的 PYNLPIR 汉语分词库进行分词。
(4) **过滤停用词**：停用词是文本中一些冗余成分，不具备任何表达能力但有高频的特点，本文使用哈工大停用词表进行停用词的过滤步骤。
(5) **过滤词性**：对词性为介词、连词、助词、时间词、字符串等一些标记词性进行过滤。
(6) **过滤词长**：对词语长度大于 6 小于 2 的词语进行过滤。
(7) **位置、词性分类**：对完成以上过程的词语进行位置和词性的分类，并生成两个键分别是位置名称和词性名称的词典。
(8) **权重计算**：分别计算每个词的 TF-IDF 权重$\mathrm{W_{tfidf}}$、位置权重$\mathrm{W_{pos}}$、词性

权重W_{tag}、词长权重W_{len}，并以公式(2-3-1)计算总权重W_{all}。

$$W_{all} = (\alpha W_{tfidf} + \beta W_{len}) * W_{pos} * W_{tag} \quad (2\text{-}3\text{-}1)$$

2.3.2 MF-TF-IDF 及其他两种关键词提取算法比较

表 3、表 4 和表 5 分别代表了 MF-TF-IDF 算法、TF-IDF 算法和 textrank 算法对附件二中随机选取的 5 条留言的九个关键词提取结果。

表 3 MF-TF-IDF 关键词提取方法

留言编号	关键词	一级分类
1967	自来水 浑浊 百姓 白色 用水水质 乡镇 挑水 食用 常年	城乡建设
2448	养殖场 水产局 动物 畜牧 臭气 青山绿水 粪坑 营业执照 手续	环境保护
3399	城区 行人 三轮车 红绿灯 摩托问题 勇往直前 交通 马路 城市	交通运输
4686	英才 幼儿园学费 家长 学期 子女 价格 班次 中心 关键	教育问题
9211	抚养费 社会 农村 小孩 百姓户口 计划生育 计委 积分 计生办	卫生计生

表 4 TF-IDF 提取关键词方法

留言编号	关键词	一级分类
1967	太桥 自来水 浑浊 H 县国 常年 百姓 原国 希望	城乡建设
2448	养殖场 KO 水产局 臭气熏天 防疫 失效 畜牧 手续	环境保护
3399	严治 城区 行人 G 区桥 桥南 霸站 崭 新 问题	交通运输
4686	幼儿园 英才 家长 学费 预交 子女 涨价 私立 退还	教育问题
9211	抚养费 再交 小孩 交清 老百姓 农村 社会 户口	卫生计生

表 5 textrank 关键词提取方法

留言编号	关键词	一级分类
1967	百姓 希望 太桥 管理 用水 看一看 领导 只能 能够	城乡建设
2448	养殖场 防疫 位于 水产局 动物 井水 畜牧 距离 粪坑	环境保护
3399	城区 问题 行人 霸站 排满 道路 小车 应该 管理	交通运输
4686	家长 学费 涨价 放假 价格 要求 招生 学期 子女	教育问题
9211	需要 抚养费 社会 小孩 没有 农村 办理 缴纳 老百姓	卫生计生

根据提取关键词结果，从主观角度来说，基于 MF-TF-IDF 算法提取关键词的结果是明显优于 TF-IDF 算法和 textrank 算法提取关键词结果的。从数值量化的角度来看，本文使用了 sklearn.feature_extraction.text.CountVectorizer 的方法将关键词文本文档集合转换成词语计数矩阵，统计每种算法提取关键词总数。其中 MF-TF-IDF 算法提取关键词总数为 14308，IF-IDF 算法提取关键词总数为 22745，textrank 算法提取关键词总数为 12917。这说明 MF-TF-IDF 算法相比于 TF-IDF

算法提取出的关键词更加集中，也就是说 MF-TF-IDF 算法在提取文本特征上优于 TF-IDF 算法。虽然 textrank 算法在提取关键词总数上比 MF-TF-IDF 算法更少，但是 textrank 算法提取的文本特征并不能很好地体现文本特征，这一点在第三章基于 MF-TF-IDF 关键词提取算法的留言分类中给出详细实验结果。

3 基于 MF-TF-IDF 关键词提取方法的留言分类

在处理网络问政平台的群众留言时，首先就是需要将留言按照文本进行分类。根据附件 1 中的一级分类标签，我们将留言类别分为七种，分别是城乡建设，环境保护，交通运输，教育文体，劳动和社会保障，商贸旅游和卫生计生。下面根据附件 2 给出的留言数据集建立基于 MF-TF-IDF 关键词提取方法的留言分类模型并作出验证。

3.1 留言分类模型的建立

3.1.1 附件 2 数据预处理

在附件 2 中共有 9210 条留言及其一系列数据，一级标签为城乡建设下的数据为 2009 条，环境保护下的数据为 938 条，交通运输下的数据为 613 条，教育文体下的数据为 1589 条，劳动和社会保障下的数据为 1969 条，商贸旅游下的数据为 877 条，卫生计生下的数据为 1215 条。

因为数据在各个标签下的分布不均，我们选择随机排列 9210 条留言的顺序，使得每一类标签下的样本尽量均匀分布于训练集与验证集中。

使用上文提出的 MF-TF-IDF 算法对打乱的留言数据集进行特征提取，最终得到了每一条留言的关键词，使用 sklearn.feature_extraction.text.TfidfVectorizer 将留言关键词转换成 TF-IDF 矩阵，并以转换完成的数据和原分类标签作为数据集进行留言分类。

3.1.2 分类结果度量指标

本文对分类结果采用三个指标度量，分别是 Precision，Recall 和 F1-measure, 他们的表达式分别为(3-1-1)，(3-1-2)，(3-1-3),

$$\mathrm{P} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FP_i} \tag{3-1-1}$$

其中TP_i代表第 i 类真的正例，FP_i代表第 i 类的假正例，

$$\mathrm{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FN_i} \tag{3-1-2}$$

其中FN_i代表第 i 类的假反例，

$$\mathrm{F_1} = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{3-1-3}$$

其中P_i和R_i分别代表第 i 类的查准率和查全率。

3.1.3 分类器选择

本文最终选择支持向量机模型作为最终的分类模型，支持向量机对于高维数据的分类问题有着非常好的效果，采用 MF-TF-IDF 算法提取的关键词数据集分别对不同的分类器做十折交叉验证，实验结果如表 6，

表 6 分类器性能对比

	SVM	NB	KNN	LR	DecisionTree
Precision	0.88589	0.87000	0.78606	0.87951	0.78595
Recall	0.86078	0.69178	0.76156	0.82535	0.76982
F1	0.87144	0.72514	0.77101	0.84644	0.77487

经过对比试验发现，无论是查准率、查全率还是 F1 度量来看，支持向量机模型都是最优的选择。

3.2 MF-TF-IDF 算法有效性验证和分类结果展示

本节首先对比 MF-TF-IDF 算法提取和未提取留言关键词数据集的分类结果，然后展示三种关键词提取方法的分类结果对比。

3.2.1 MF-TF-IDF 的有效性

在经过同样的数据预处理后，本文使用 MF-TF-IDF 提取的十个关键词数据集，和未提取关键词的直接分词数据集进行分类比较。SVM 分类器参数设置，C=1，kernel=‘Linear’。

对比结果如表 7 所示。

表 7 提取与未提取关键词对比

	MF-TF-IDF 提取特征	**未提取特征**
Precision	0.88589	0.87345
Recall	0.86078	0.84797
F1	0.87144	0.85846
Dimension	(9210,14308)	(9210,34267)

结果表明，MF-TF-IDF 算法提取关键词有效的降低了数据维数，并提升了留言分类的效果。

3.2.2 几种关键词算法的对比

为说明 MF-TF-IDF 提取关键词算法在留言分类模型上的优越性，这里选择了 TF-IDF 提取关键词算法和 textrank 提取关键词算法提取留言数据集中十个关键词，并使用相同的 SVM 分类器模型。

分类结果如表 8 所示。

从表 8 的结果分析，对比 TF-IDF 算法，MF-TF-IDF 算法提取关键词后不仅对原始数据集进行了有效降维而且明显提升了分类效果。而 textrank 算法虽然提取的关键词更为集中只有 12917 维，却是分类效果却是最差的一个，这说明了 textrank 算法提取的关键词集中在文本共有的词语，关键词提取效果并不理想。

表 8 关键词算法分类效果对比

	MF-TF-IDF	TF-IDF	TextRank
Precision	0.88589	0.88498	0.86148
Recall	0.86078	0.83732	0.81626
F1	0.87144	0.85653	0.83413
Dimension	(9210,14308)	(9210,22745)	(9210,12917)

3.3 分类模型总结

针对第一个问题，本文为了有效提取特征，使用了 MF-TF-IDF 算法提取文本关键词。经过实验结果表明 MF-TF-IDF 算法是一种比 TF-IDF 算法和 textrank 算法更好的提取关键词算法。在此基础上，本文选择了适合处理高维数据集的支持向量机模型作为分类器进行留言分类。经过十折交叉验证，模型得到的平均 F1 值为 0.87144，对比其他算法达到了预期的效果。

4 热点问题挖掘

群众留言内容包罗万象，等待解决的问题有轻重之分，自然就导致留言系统需要具备自动搜索出热点问题的功能。热点问题是指某一时间段内反映特定地点或特定人群问题的所有留言，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。

4.1 问题分析

第二问要求挖掘热点问题，本文将其视为一个文本聚类问题，使用聚类方法对附件三的所有留言进行聚类。在聚类结果理想的情况下，每一个类可代表一个热点问题。但实际情况中，存在不属于热点问题的其他留言，聚类算法会将其他留言依据不同的度量计入不同的热点问题中，导致聚类的结果有较大的偏差。本文采用计算文本相似度的方法，逐一计算留言相似度，规定相似度超过阈值的留言属于同一热点问题。

基于上述方法挖掘出各个热点问题，本文建立热度评价模型计算热点问题的热度指数，依据热度指数对热点问题进行排序，从而选择前 5 个热点问题，输出结果。

4.2 基于相似度挖掘热点问题

常用的计算相似度方式有欧几里得距离和余弦距离，公式分别由(4-2-1)和(4-2-2)给出。

$$\mathrm{sim}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4-2-1}$$

$$\mathrm{sim}(\mathrm{x},\mathrm{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \ . \tag{4-2-2}$$

与欧氏距离相比，余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分话题的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题。本文选择余弦距离作为文本相似度的度量，这将更加符合我们的挖掘目标。

基于相似度的热点问题挖掘算法流程如图 2 所示。

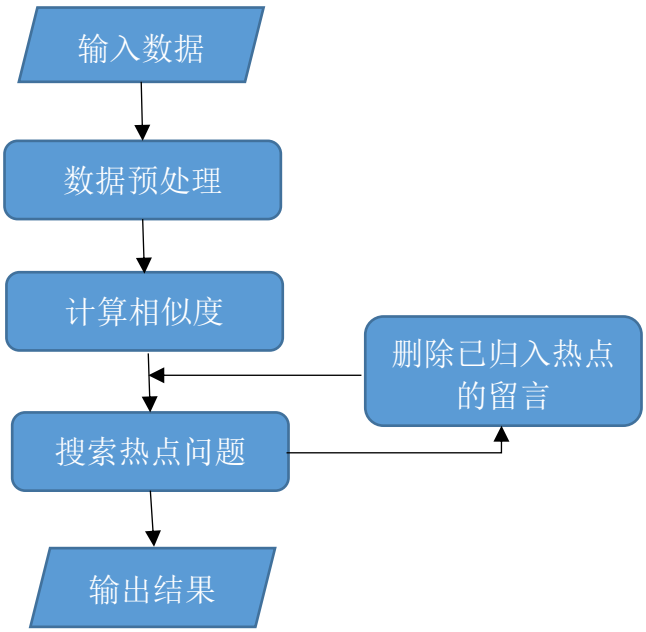

图 2 热点问题挖掘流程

4.2.1 热度评价模型

根据附件三所给出的点赞数以及反对数，我们可以构造一组数据用于描述热点问题：$X = (x, y, z)$，其中 x 为留言数量，y 为点赞总数，z 为反对总数。基于这一数据建模计算热点问题的热度指数 H：

$$\mathrm{H} = \mathrm{W} \cdot \mathrm{X} = \mathrm{w}_1 x + w_2 y + w_3 z \tag{4-2-3}$$

其中$\mathrm{W} = (w_1, w_2, w_3)$为权重，分别对应留言数量、点赞总数和反对总数。

4.2.2 熵权法

下面我们使用熵权法计算W，熵权法的步骤如下：

（1） 数据异常值剔除；

（2） 对数据依据正负向进行归一化处理；

（3） 计算确定熵权值

针对所有热点问题的 3 个变量，得到相关变量矩阵(4-2-4)：

$$.\mathrm{A}=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}. \tag{4-2-4}$$

本问题是讨论热点问题的热度，对于一个问题，参与讨论的主体越多，显然热度应该越高。因此，留言数量、点赞总数和反对总数都应该属于效益型变量，应使用式(4-2-5)对其进行归一化处理：

$$\mathrm{r_{ij}}=\frac{a_{ij}-\min(a_{ij})}{\max(a_{ij})-\min(a_{ij})}\ , \tag{4-2-5}$$

定义第 j 个指标的熵计算公式为(4-2-6)

$$\mathrm{K_j}=-\frac{\sum_{i=1}^{m} f_{ij}\ln(f_{ij})}{\ln(m)}, i=1,2,\ldots,n \tag{4-2-6}$$

上式中$\mathrm{f_{ij}}=r_{ij}/\sum_{i=1}^{m} r_{ij}$，当 $f_{ij}=0$时，令$ln(f_{ij})$=0。

根据式(4-2-7)计算各指标的熵值

$$\mathrm{w_i}=\frac{1-K_i}{n-\sum_{i=1}^{n} K_i}\ . \tag{4-2-7}$$

4.2.3 热点问题结果分析

根据 4.2 的热点挖掘算法，我们得到各个热点问题的详情，下图展示前 5 个热点问题及其包含的具体留言：

```
第1个热点问题为：
包含留言[82, 270, 277, 356, 743, 847, 963, 1003, 1347, 1394, 1519, 1532, 1647,
1695, 1872, 2007, 2127, 2144, 2245, 2409, 2454, 2494, 2769, 2826, 3061, 3088,
3104, 3105, 3444, 3514, 3562, 3753, 3936, 4002, 4202, 4229, 4234] 共37条

第2个热点问题为：
包含留言[89, 319, 690, 703, 731, 787, 890, 1000, 1296, 1298, 1377, 1483, 1545,
1952, 2014, 2122, 2230, 2445, 2659, 2830, 2937, 2951, 3389, 3404, 3417, 3717,
3732, 3869, 3901, 4039, 4138, 4148, 4273, 4276, 4298] 共35条

第3个热点问题为：
包含留言[222, 307, 392, 808, 1085, 1213, 1257, 1384, 1508, 1520, 1646, 1654, 1
736, 1934, 2009, 2195, 2364, 2582, 2712, 2745, 2782, 3011, 3213, 3384, 3567, 3
585, 3675, 3812, 4079, 4217, 4289] 共31条

第4个热点问题为：
包含留言[763, 821, 881, 1209, 1303, 1724, 1841, 1962, 2071, 2143, 2428, 2522,
2591, 2623, 2663, 2676, 2854, 2903, 2958, 3179, 3556, 3773, 3924, 3939, 4035,
4046, 4067] 共27条

第5个热点问题为：
包含留言[24, 145, 371, 598, 877, 1059, 1167, 1361, 1526, 1801, 1877, 2027, 205
8, 2506, 2647, 2723, 2767, 3090, 3291, 3545, 3750, 3853, 3913, 3993, 4257] 共2
5条
```

图 3 留言数量前 6 的热点问题

从图中看到依据包含留言数排名的前 6 个热点问题，例如第一个热点问题，其中包含 37 条留言，分别是[82, 270, 277, 356, 743, 847, 963, 1003, 1347, 1394,

1519, 1532, 1647, 1695, 1872, 2007, 2127, 2144, 2245, 2409, 2454, 2494, 2769, 2826, 3061, 3088, 3104, 3105, 3444, 3514, 3562, 3753, 3936, 4002, 4202, 4229, 4234]号留言。分析上述留言的留言详情，这些留言所反映的是同一事件——“关于伊景园滨河苑捆绑销售车位的投诉”问题。对其余热点问题的分析也得到一致的结果，每个热点问题均是围绕某一事件、场所或政策的留言。据不完全统计，附件三中能够挖掘出的热点问题有：车贷诈骗案件、车位捆绑销售、人才购房新政、公积金、医疗事件、驾校诈骗、地铁运营以及证件办理等等。

根据 4.2.2 熵权法计算出留言数量、点赞总数和反对总数对应的权重如表 9 所示。

表 9 各指标权重

指标	留言数量	点赞总数	反对总数
权重W	0.5936	0.0079	0.3985

由热度评价模型计算上述热点问题的热度指数，依据热度指数排序，选择前 5 个热点问题，并保存为文件“热点问题表.xls”，如表 10 所示。

表 10 热点问题

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	685.5	20190108.0 至 20198023.0	A 市车贷案件	西地省富惠天下商务有限公司涉嫌经济诈骗
2	2	33.18	20190104.0 至 20198030.0	A 市中南大学楚雅附二医院	请督促 A 市中南大学楚雅附二医院退还医保款
3	3	30.20	20181115.0 至 20199027.0	A 市人才新政	在 A 市买公寓能享受人才新政购房补贴吗
4	4	29.94	20190707.0 至 20197028.0	伊景园滨河苑	关于伊景园滨河苑捆绑销售车位的维权投诉
5	5	26.86	20190102.0 至 20198021.0	A 市公积金	咨询 A 市购房公积金贷款的要求

4.2.4 独立热点问题

文本相似度热点问题挖掘算法对于留言数量较多的热点问题的偏好，使得部分留言在挖掘过程中被忽略。通过进一步分析附件三中的留言，发现存在个别留言，其点赞或反对数较多，但除此之外没有与其相似的留言，因此未能被上述算法发现。本文将此类留言称作独立热点问题，如图 4 所示。

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
193091	A00097965	富绿物业丽发新城强行断业主多	019/6/19 23:28:2	是供地摊上买的收	0	242
194343	A00010616	为A市58车贷案警官应跟进关注留	019/3/1 22:12:3	侦并没有跟进市	0	733
208636	A00077171	区汇金路五矿万境K9县存在一系	019/8/19 11:34:0	狗咬人，请问有人	0	2097
217032	A00056543	惩A市58车贷特大集资诈骗案保护	019/2/25 9:58:3	股东、苏纳弟弟	0	790
220711	A00031682	请书记关注A市A4区58车贷案	019/2/21 18:45:1	情消息总是失望	0	821
223297	A00087522	反映A市金毛湾配套入学的问题	019/4/11 21:02:4	纳入配套入学，A	5	1762
263672	A00041448	小区距长赣高铁最近只有30米	019/9/5 13:06:5	到我如下问题：	0	669
203187	A00024716	咨询A9市高铁站选址的问题	019/8/1 13:48:5	区东进的步伐，	53	10

图 4 独立热点问题

显然，表 11 中留言的点赞或反对数是异常地高出其余留言。例如编号 208636 留言，因其描述了 A 市某高端别墅小区种种管理不当行为而获得大量点赞。这表明参与讨论的网民数量庞大，此类留言的热度也不容小觑。利用 3.2.3 的热度评价模型，计算上述留言的热度指数并排序选择前 5 个，保存为文件“独立热点问题表.xls”。如表 11 所示。

表 11 独立热点问题表

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	52	2019/8/1 13:48:57	A 市地铁	咨询 A9 市高铁站选址的问题
2	2	41	2019/8/19 11:34:04	A 市高端别墅小区	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	3	39	2019/4/11 21:02:44	A 市金毛湾	反映 A 市金毛湾配套入学的问题
4	4	16	2019/2/21 18:45:14	A 市车贷案	请书记关注 A 市 A4 区 58 车贷案
5	5	15	2019/2/25 9:58:37	A 市车贷诈骗	严惩 A 市 58 车贷特大集资诈骗案保护伞

5 答复意见评价

科学的评价可以促进工作人员捕捉到咨询群众的需求，意识到日后的工作需要改进的地方，更好地提升群众的满意度。

5.1 问题分析

本问题要求先对其提供多角度分析最终得出一个总体评价。本文先从第 i 个文本中分解得到其 MF-TF-IDF 权重向量，将该向量映射到若干个维度的评价$X = (x_{i1}, x_{i2}, \cdots, x_{ik})$，再将这个向量映射到一个实数上，由于数据中并没有分数的标

签，只有作为输入的文本，因此我们需要找到一个无监督学习算法而非传统的有监督学习算法。

本文采用降维的方法，首先多次尝试不同的评价角度，选出三个合理的不同的角度来评价答复意见。然后用两种方式对这个三维的向量$X=(x_{i1},x_{i2},x_{i3})$进行降维，再根据两种方法的特色和多次尝试的结果对这两种方式的评分做一个平均，确保评分的客观性。

最终通过观察评分靠前的若干答复意见，来验证评价的结果，修改评价的角度和模型。

5.2 答复意见评价指标

对答复意见我们用三个标准去衡量：完整性，可解释性，答复量(如图 5)。D对于相关性，在观察了大量样本后，我们认为答复都能就相关问题进行回应，该项的得分难以拉开差距，所以没有采用这项评价。

图 5 答复意见的评价指标

我们记第 i 个留言文本中关键词向量为$\alpha_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{in})$，答复文本中关键词向量为$\beta_i=(\beta_{i1},\beta_{i2},\cdots,\beta_{im})$，本文在留言文本中选择提取 n 个关键词，答复文本中选择提取 m 个关键词。多次尝试后本文认为 $n=5$，$m=8$ 是比较合理的参数取值。另外，我们在选择词汇的时候，排除了一些出现频率极高但不带有信息的词汇，这些词汇大部分是礼貌用语和官方词汇等。

完整性通过答复文本的词汇向量β_i和留言文本的词汇向量α_i的重合度来体现，即咨询意见中重点的词汇在答复文本中占据重要位置的比例。若α_i中有k_i个元素在β_i中出现，则该文本的完整性得分如式(5-2-1)，其中k_i为不大于 n 的非负整数。

$$x_{i1}=\frac{k_i}{n} \tag{5-2-1}$$

可解释性可以通过动词的多样性来衡量。本文认为，一个事件可能由其他多个事件所导致，答复的过程中就会需要提及很多人和事物的动作。即若第 i 个答复中提及了N_i个动词，我们选出最高频的 5 个词，而不在这 5 个词中的词汇数为p_i，则该文本的可解释性得分如式(5-2-2)。

$$x_{i2}=\frac{p_i}{N_i} \tag{5-2-2}$$

答复量体现在答复文本相对于咨询文本的长度。根据观察，大部分答复的文本是要长于咨询文本的，这是因为解释一个问题往往会比描述一个问题需要更多的信息。而且对于咨询人员而言，更多的答复描述往往意味着官方工作人员的耐心和重视程度，也能收获更高的满意度。记第 i 个咨询文本的长度为L_{i1}，答复文本的长度为L_{i2}，理论上该文本的答复量得分如式(5-2-3)，考虑出现个别咨询文本

$$x_{i3} = \frac{L_{i1}}{L_{i2}} \tag{5-2-3}$$

是长于答复文本的，得分应该控制在 0-1 范围内，而且答复量应该是以多为好。最终答复量得分如式(5-2-4)所示。

$$x_{i3} = 1 - \min\left(\frac{L_{i1}}{L_{i2}}, 1\right) . \tag{5-2-4}$$

至此，我们得到了一个 S×3 的矩阵 D，其中 S 为样本数，并且$D_{ij} \in [0,1], i = 1,2,\dots,S, j = 1,2,3$. 该矩阵 D 作为我们接下来建模的数据。

5.3 数据信息压缩

问题分析中提到，我们需要构造一个三维向量到一维实数的映射，这一个维度应尽量保留更多的信息。一般的思路是取三维向量的 L2 范数作为一维实数，但是这样会极大地丢失信息，并且每个指标的权重相同并不符合客观情况。这里我们选取了两种无监督学习算法作为信息压缩的工具：PCA 数据挖掘算法和 AE 自编码器深度学习模型。

5.3.1 PCA 数据挖掘算法

PCA 是一种传统的降维压缩算法，主要是利用了基的线性变换，然后去掉信息量较小的维度，只留下最主要的维度作为表示。记数据的矩阵为 D，PCA 算法的步骤如下：

（1）将每一列零均值化；
（2）算出 D 的协方差矩阵Σ；
（3）求解Σ的特征值和对应的特征向量；
（4）将特征向量按特征值从大到小选取 K 个，组成矩阵 T；
（5）$D' = TD$ 即为所求最终结果。

5.3.2 AE 深度学习算法

AE 深度学习算法相比于其他神经网络，它的特点在于 AE 自编码器是无监督学习，其目标输出就等于输入。一般的 AE 自编码器的隐层神经元数是少于输入和输出层的，目的是通过编码和解码的过程来压缩信息，用较少的神经元来还原较多的神经元。

一般的AE自编码器分为编码和解码两个部分，编码则输入层到隐层的过程，而解码则为隐层到输出层的过程。

隐层第 i 个神经元的值为：

$$h_i = \sigma(\sum w_{ij} x_j + b_i) . \tag{5-3-1}$$

输出层第 k 个神经元的值为：

$$y_k = \sigma(\sum w'_{kj} h_j + r_k). \tag{5-3-2}$$

其中σ为激活函数，b_i和r_k为阈值。

本次架构的 AE 自编码器网络总共为 5 层，神经元数目分别为 3 个，2 个，

1 个，2 个，3 个。

其中最中间的一层的值即为我们所求的一维数据，它最大程度地保留了原来三维数据的信息量，并可以还原成与原本的数据非常接近的程度。其收敛情况和均方误差如图 6 所示。

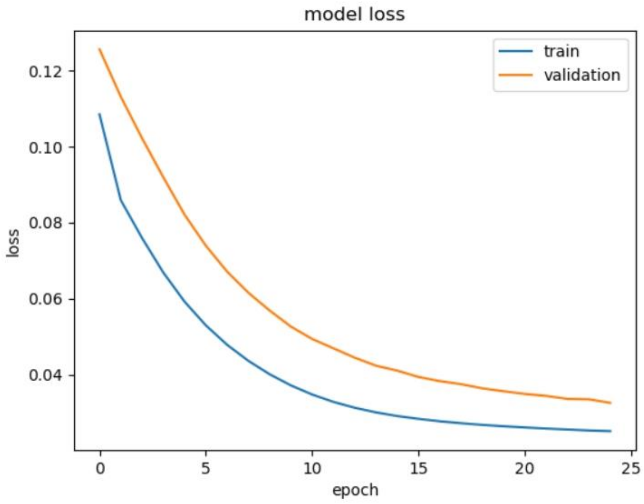

图 6　AE 自编码器的均方误差收敛情况

可以看到建模后无论是训练数据还是测试数据都得到了良好的收敛，说明该隐层神经元可以很好地还原数据矩阵 D 中的信息。

5.3.3 结果展示

两个算法对数据的还原情况接近一致，对评分的偏好比较相同，都对综合排名靠前的答复打出较高的分数。经过检验，我们对排名靠前的回答同样给出了比较高的评价，两种评分算法的部分结果如表 7 所示。

表 12 平均分数前十名明细

留言编号	完整性	解释性	信息量	PCA 评分	AE 评分	平均分数
1743	0.600	0.830	0.950	0.992	0.994	0.993
1918	0.400	0.782	0.992	0.988	0.993	0.991
2035	0.400	0.782	0.992	0.988	0.993	0.991
2037	0.800	0.721	0.899	0.988	0.993	0.991
1320	0.400	0.784	0.959	0.990	0.989	0.989
758	0.600	0.826	0.876	0.986	0.991	0.988
2214	0.400	0.858	0.908	0.988	0.988	0.988
2247	0.400	0.817	0.899	0.986	0.990	0.988
1737	0.800	0.841	0.774	0.983	0.987	0.985
152	0.200	0.832	0.922	0.987	0.982	0.984

5. 3. 4　答复总体评价

我们观察了排名靠前的答复，基本上都满足一些特点：能客观认真地指出咨询问题的根本原因，能详细地向咨询人员描述事实情况，而且能在答复过程中体

现出充分的调研和探究。

这些因素都表现出相关部门的工作人员在答复时不只是空谈和推卸责任，而是有认真地对咨询的问题作深刻的分析和充分的调查取证，最后能依法处理相关责任人，将为人民服务落实到底。而我们的算法也能够发现这些优秀的工作人员和他们的答复，这为日后其他工作人员提供了一个良好的借鉴作用。

6 不足与展望

1. MF-TF-IDF 算法虽然在提取关键词上提升效果非常明显，但是因为需要计算每一个词语的多因素权重所以在遍历词语时所需要的时间复杂度是相对较高的，导致提取关键词时程序运行速度较慢。

2. 在热点问题搜索中，本文给出的基于相似度搜索的算法，显然已经足以挖掘出热点问题，但是经过仔细阅读附件三内的留言，该算法没能将所有相关的留言归入热点问题中。一方面是本文为了获得类间准确度而放弃了类内精度所导致的结果，另一方面也存在部分留言较难识别的因素。

3. 在答复意见评价中，本文给出了两种基于无监督算法的数据压缩工具，对答复意见作出了相似的评价，是基本符合客观情况的。但是这两套算法依赖于对文本的多维度评价矩阵，评价标准的不同会导致最终评价偏离客观。所以应进一步探究各种不同的评价维度，以求最终结果的公正性。

参考文献

[1] NIU Yong-jie, TIAN Cheng-long. 融合多因素的 TFIDF 关键词提取算法研究[J]. 计算机技术与发展, 2019, 029(007):80-83.
[2] 张建娥. 基于 TFIDF 和词语关联度的中文关键词提取方法[J]. 情报科学, 2012(10):110-112+123.
[3] Chen Y H , Lu J L , Tsai M F . Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors[J]. Expert Systems with Applications, 2014, 41(2):663-670.
[4] 李航. 统计学习方法[M]. 清华大学出版社, 2012.
[5] Habibi M , Popescu-Belis A . Keyword Extraction and Clustering for Document Recommendation in Conversations[J]. Audio Speech & Language Processing IEEE/ACM Transactions on, 2015, 23(4):746-759.
[6] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 数据分析与知识发现, 2013, 29(9):30-34.
[7] Wang D , Zhang H . Inverse-Category-Frequency based supervised term weighting scheme for text categorization[J]. 2010.
[8] 张红鹰. 中文文本关键词提取算法[J]. 计算机系统应用, 2009(08):75-78.
[9] 赵胜辉, 李吉月, 徐碧,等. 基于 TFIDF 的社区问答系统问句相似度改进算法[J]. 北京理工大学学报, 2017(09):106-109.