

DOI:10.20033/j.1003-7241.(2026)10-0159-04

基于改进模糊聚类的电网在线客服语音文本挖掘方法

王 牧, 梁耀明, 马少华, 林增健, 张 薇

(南方电网供应链集团有限公司, 广州 510000)

摘 要: 电网在线客服语音文本具有流式连续输入、句式结构松散、领域术语密集的特点, 传统基于静态主题建模或单轮次意图识别的方法, 难以捕捉同一用户诉求在不同通话阶段、不同业务语境下的语义漂移与隐性关联, 导致相同问题被归入不同类别、不同问题被混淆为同一类别, 挖掘结果在诉求归并与业务分类两个层面均存在显著偏差。针对该难点, 提出基于改进模糊聚类的挖掘方法。首先构建语义解析机制, 按意图边界切分关键词序列, 避免整句输入的语义模糊与特征混叠; 其次引入词性关联约束与松弛因子修正隶属度函数, 使聚类受词性转移概率与主题共现置信度双重约束, 抑制语义漂移对聚类中心的牵引; 最后设计关联度滤波函数, 以相关性系数归并高度关联文本, 实现诉求相近文本聚合式分层挖掘。实验表明, 该方法查准率明显优于对比方法, 相同时间内挖掘数量提升超过 60%, 有效解决了电网客服语音文本因语义漂移导致的挖掘失准难题。

关键词: 改进模糊聚类; 电网在线客服; 语音文本挖掘; 语义漂移; 关联模式识别; 自然语言处理; 主题聚类

中图分类号: TP311.13; TM73; TP391.42

文献标志码: A

文章编号: 1003-7241(2026)10-0159-04

Speech and text mining method for online customer service in power grid based on improved fuzzy clustering

Wang Mu, Liang Yaoming, Ma Shaohua, Lin Zengjian, Zhang Wei

(Southern Power Grid Supply Chain Group Co., Ltd., Guangzhou 510000, China)

Abstract: Power grid online customer service speech texts feature streaming continuous input, loose sentence structures and dense domain terminology. Traditional static topic modeling or single-turn intent recognition fails to capture semantic drift and implicit relevance of user demands across conversation stages, leading to obvious classification biases. An improved fuzzy clustering mining method is proposed. A semantic parsing mechanism segments keyword sequences by intent boundaries. Part-of-speech constraints and relaxation factors modify membership functions to suppress clustering center deviation caused by semantic drift. A correlation filtering function aggregates highly similar texts to realize hierarchical mining. Experiments verify higher precision and over 60% improvement in mining volume, addressing inaccuracies induced by semantic drift.

Keywords: improved fuzzy clustering; power grid online customer service; speech text mining; semantic drift; association pattern recognition; natural language processing; topic clustering

在信息化时代, 电网客服系统积累了大量的语音数据, 其中包含了丰富的用户需求和问题信息。传统的文本挖掘方法往往难以处理文本的歧义和不确定性, 在应对大规模高维数据的处理时, 通常会面临计算复杂度显著提升以及聚类性能易于波动的问题。因此, 对电网客服语音文本进行挖掘对提高电网服务质量和用户体验具有重要意义。

文献[1]利用注意力机制来处理文本数据, 通过给文本的不同部分分配不同的权重, 从而提取更关键的特征, 采用卷积神经网络(convolutional neural network, CNN)和多层感知机(multilayer perceptron, MLP)进一步从文本中提取深层特征, 从而生成更全面、更准确的推荐结果。该方法能够自适应地关注文本中的重要信息, 提高挖掘的准

确性和个性化。同时能够自动从文本中提取深层特征, 无须手动设计复杂的特征工程。但模型的计算复杂度相对较高, 导致挖掘效率较低。文献[2]通过对收集的数据进行清理和过滤, 利用单词袋模型和词频逆文档频率(term frequency-inverse document frequency, TF-IDF)对数据进行降维, 并建立文本挖掘模型, 以此实现语音文本的挖掘。这种方法不受语言、文化和文档格式的限制, 可以处理任何格式的文本, 并可以灵活地适应不同的任务要求。但该方法的有效性在很大程度上取决于输入数据的质量和数量。如果数据质量差或数量不足, 可能会导致模型性能下降。文献[3]利用双向编码器表示变换器(bidirectional encoder representations from transformers, BERT)模型对文

收稿日期: 2025-06-28; 录用日期: 2025-09-03

基金项目: 广东省科技基金资助项目(ZBKJXM20232162)

作者简介: 王 牧(1989—)男, 经济师, 研究方向为供应链领域、客户服务领域。

引用本文: 王牧, 梁耀明, 马少华, 等. 基于改进模糊聚类的电网在线客服语音文本挖掘方法[J]. 自动化技术与应用, 2026, 45(10): 159-162. (Wang Mu, Liang Yaoming, Ma Shaohua, et al. Speech and text mining method for online customer service in power grid based on improved fuzzy clustering[J]. Techniques of Automation and Applications, 2026, 45(10): 159-162.)

本进行深入的语义理解,并从中获得文本向量,使用隐狄克雷分配(latent dirichlet allocation, LDA)模型对 BERT 模型生成的文本向量进行主题聚类,结合聚类结果深入分析文本情感,由此完成文本深度挖掘。该方法可以自动发现文本中的主题,而无需手动定义,减少了人为干预的影响,提高了文本挖掘的客观性。但在使用 BERT 和 LDA 模型进行文本挖掘之前,需要对文本进行严格的预处理,若是处理不当可能会影响模型的准确性和效率。

因此,为提高电网客服语音文本挖掘的准确性与效率,本文结合改进模糊聚类算法对此展开深入研究,旨在通过对电网客服语音数据的深度挖掘和分析,为电网企业提供决策支持和服务优化。

1 语音文本挖掘方法设计

1.1 语音文本语义特征辨识

在进行语音特征分析之前其首要的任务是将语音转化为文本形式。电网在线客服的语音转文本工作,其核心在于将客户与客服之间的对话由语音形式转化为文字形式,这不仅有助于更好地理解和分析对话内容,同时也为后续的数据处理和分析提供了便利^[4]。这种转换不仅可以高效地处理和分析对话内容,并且便于有效特征的提取。

鉴于电力在线客服语音处理的对象是庞大的文本集合,这些文本内容均源自自然语言,不仅涵盖了语法和语义的层面,还涉及了语用的复杂维度。在处理这些文本时,需要充分考虑其自然语言的特性,从而确保能够准确理解和提取出有价值的信息^[5]。因此,应将其转换为可编辑、可检索的文本形式。基于语音文本的存储形式,建立一个语义解析机制,对语音关键词进行划分,表达式为

$$I_s = \beta_t \sum_{i=1}^n \frac{q_i}{h_0 \times b_0} \quad (1)$$

式中, β_t 表示第 t 个关键词的语义动作系数; q_i 表示第 i 个关键词对应的文档特征维度; 0 表示基准值; h_0 表示文本向量的基准模长; b_0 表示语义解析的基准系数。

根据语音文本的表征状态和语义折中率将原始语音数据转换为文本形式,表达式为

$$X_0 = I_s \cdot \delta_k (c_1 + c_2) \quad (2)$$

式中, k 表示文本表征状态类型编号, δ_k 表示第 k 类文本表征状态的转换系数; c_1, c_2 分别表示语义折中率的最大值和最小值。

为解决文本的结构化表示,采用自然语言处理技术对转换后的文本进行预处理^[6],为后续特征提取奠定基础。

首先按照词汇的词性构建词典以辅助分词,表达式为

$$A_0 = \frac{X_0 \times \tilde{\omega}_s}{\sqrt{|n_s|}} \quad (3)$$

式中, $\tilde{\omega}_s$ 表示第 s 类词性的词频个数; n_s 表示第 s 类词性的偏差度函数值。

通过词向量模型将文本嵌入数值空间,以检测出其中的低频词^[7],并去除,表达式为

$$w_e = \frac{A_0 \times \gamma_b}{\|C_x\|^2} \quad (4)$$

式中, e 表示频段序号; γ_b 表示第 b 个频段的信噪比系数; C_x 表示第 x 个文本向量的协方差矩阵,文中取单位矩阵。

进而对有用的词汇进行词性标注,即

$$T_j = \frac{w_e \times \vartheta_0}{m_c} \cdot h_0 \quad (5)$$

式中, ϑ_0 表示特定词性在句子起始位置出现的概率; T_j 表示状态转移概率矩阵; m_c 表示观测概率矩阵中第 c 个观测值; h_0 表示初始状态转移基矩阵。

根据词性统计结果提取语音文本中的语义特征^[8],计算式为

$$K = \sqrt{\| \frac{T_j \cdot \sigma_c}{Q} \|^2} \quad (6)$$

式中, σ_c 表示第 c 组词汇对应词出现的共现概率比值; Q 表示随机场函数。

1.2 改进模糊聚类的文本关联识别

采用常规的模糊聚类算法对语音数据资源展开分析时不难发现,该算法对初始参数的设定异常敏感,这直接导致了其收敛速度的迟缓。更为关键的是,在迭代的过程中,算法极易陷入局部最优,这无疑对数据分析的精准性和可靠性造成了不利的影响。因此,为了提升语音数据资源分析的效果,文中将模糊聚类算法进行改进,以适应电网在线语音文本数据的特性,从而更好地识别其中的关联关系。

使用 K 均值法计算初始聚类的中心^[9],具体表达式为

$$e_0 = \sum_{j=1}^M |p_x - g_c|^2 \quad (7)$$

式中, M 表示聚类中心的空间维度; x 表示样本编号; p_x 表示第 x 个样本点到聚类中心的距离; g_c 表示第 c 个聚类中心的二维空间映射系数。

引入有效激活函数描述聚类中心中样本数据的耦合性^[10],表达式为

$$B_y = \frac{e_0 \cdot \zeta_1}{x_k} \quad (8)$$

式中, ζ_1 表示残差项; x_k 表示有效激活函数的输入变量(即第 x 个样本的特征值)。

为避免数据噪声对特征模糊矩阵的构造造成影响,利用松弛因子设计隶属度约束函数,即

$$\eta_t = B_y \times d_0 \times f_p \quad (9)$$

式中, d_0 表示模糊指数; f_p 表示第 p 类松弛因子。

定义每个主题中-出现次数大于某一阈值的单词集合为该文档的主题模型,从而生成主题函数^[11],表达式为

$$R(x) = \frac{\sqrt{\eta_t \cdot \sigma_g}}{v_i / f_w} \quad (10)$$

式中, σ_g 表示第 g 个主题标签的激活强度系数; v_i 表示第 i 个词汇的词典阈值参数; f_w 表示第 w 个时间窗口的时间

常数。

将大于阈值主题标签分配给每个文档^[12], 以此得到语音数据的传输特征为

$$g_i = R(x) \times \omega_k \quad (11)$$

式中, ω_k 表示第 k 个关键词的权重系数。

结合滤波函数分析相邻语音文本数据之间的关联模式, 表达式为

$$\mu = g_i \exp\left(-\frac{\tau_0}{c_f \cdot K}\right) \quad (12)$$

式中, τ_0 表示语音文本存储延时的基准值; f 表示滤波函数类型编号; c_f 表示第 f 类滤波函数的滤波系数; K 表示语音文本语义特征; μ 表示相关性系数。该系数的取值区间为 $[0, 1]$, 即越接近于 1, 说明语音文本数据间的相关性越高^[13], 从而可以将具有高度相关性的文本存储在同一文档中, 便于后期深度挖掘。

1.3 语音文本深度挖掘

假设电网在线客服语音文本类别用 N 表示, 待挖掘的文本数量用 D 表示, 则文本挖掘信息的相似度可表示为

$$S_d = \frac{N \cdot D}{\mu \cdot q_s} \quad (13)$$

式中, d 表示文档样本序号; μ 表示文本数据相关系数; q_s 表示待挖掘文本在第 s 个特征维度上的相关系数。

根据词语映射关系, 对每个单词分配权重系数, 从而构建词频矩阵^[14], 即

$$\mathbf{H} = S_d \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{m1} & \varphi_{m2} & \cdots & \varphi_{mn} \end{bmatrix} \quad (14)$$

式中, $\varphi_{m'n'}$ 表示目标词汇在第 m' 次对话中出现的次数为 n' 。

结合合并后的重要邻近词, 层次化文本特征, 以获取隐含空间内节点距离测度^[15], 计算公式为

$$\mathbf{l}_e = (\mathbf{H} \cdot \mathbf{v}_c) \cdot \mathbf{J} \cdot \mathbf{F}_d^{-1} \quad (15)$$

式中, $\mathbf{v}_c$ 表示第 c 个主题的主题词共现矩阵; $\mathbf{F}_d$ 表示第 d 个文档的非对称矩阵; $\mathbf{J}$ 表示一阶滞后变量。

进一步可利用下式完成对语音文本的深度挖掘, 即

$$P_L = \frac{\|\mathbf{l}_e - \mathbf{1}\|}{\|\mathbf{E}\| \cdot \psi_h} \quad (16)$$

式中, $\mathbf{E}$ 表示文本挖掘因子; ψ_h 表示第 h 个文本挖掘节点的激活值; P_L 表示挖掘文本 L 的词条集合。

通过所述计算与分析过程, 即可实现对电网在线客服语音文本的挖掘。

2 实例论证分析

2.1 实验准备

为验证本文方法的有效性, 以某省电网公司数据系统存储的客服语音文本数据为研究对象, 数据集涵盖了客服与用户的交互内容。具体包括对话编号、客服 ID、用户 ID 和问题分类, 数据集描述如表 1 所示。

表 1 数据集描述

Tab. 1 Dataset description

对话编号	客服 ID	用户 ID	问题分类
0001	C001	U001	账单查询
0002	C002	U002	故障报修
0003	C001	U003	投诉建议
0004	C003	U001	政策咨询
0005	C002	U004	业务办理

其中对话 ID 表示每个对话的唯一标识符, 以数字形式表示, 便于数据管理和查询; 客户服务 ID 表示客户服务代表的唯一标识符, 用于跟踪客服的工作和绩效; 用户 ID 表示用户的唯一标识符, 用于识别其身份和查询其历史记录; 问题分类是根据对话内容对问题进行分类, 有助于快速定位问题, 提高处理效率。5 类语音文本的分类情况如图 1 所示。

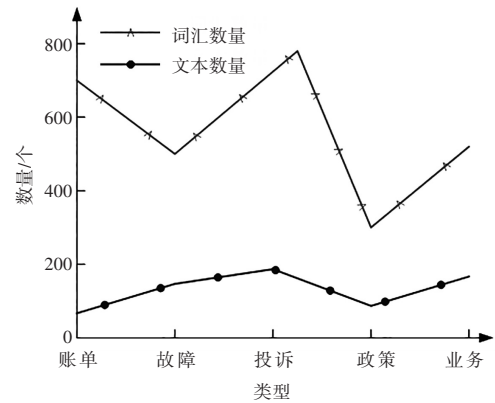


图 1 语音文本词汇分布情况

Fig. 1 Distribution of speech text vocabulary

实验环境硬件为一台配备英特尔酷睿 i7 处理器、32 GB RAM 和 1 TB SSD 的计算机; 软件环境的操作系统采用 Windows 10, 安装 Python 3.8 以及相关的数据分析库。利用语音识别服务, 将电网中在线客服的语音数据转换为文本格式进行后续处理, 并基于 Python 编程环境, 利用已安装的数据分析库和机器学习库, 构建具有文本预处理、特征提取、模糊聚类功能的文本挖掘平台。

2.2 实验说明

为了丰富数据集的属性特征, 本次实验通过企业网站平台和多种移动平台, 共计捕获了近 10 000 条客服语料, 这些语料涵盖了电力运行、电力销售以及电力维修等多个方面的客服语音文本。随后, 将这些新获取的语料与原有的语音文本进行了整合, 并依据 1:3 的比例划分了测试集和实验集, 以确保数据分布的合理性和实验的有效性。

设置语音采样率为 16 kHz, 词语窗口长度为 10, 词频大小为 5。并使用 UTF-8 编码保存转换后的文本数据, 采用自然语言处理技术去除文本中的停止词、标点符号和特殊字符, 进行词干提取与词形恢复。设置语音文本的特征维度为 500, 并保留前 5% 的词汇作为关键特征。

对上述采集的电网客服语音文本数据集采用改进模糊聚类算法进行特征提取与关联模式识别。由于电网在线客服问题类型为账单、故障、投诉、政策和业务5类,因此设置算法的聚类数量为 $K=5$,聚类中心的空间维度 $M=6$,聚类中心到样本的距离 $p_x=2.0$,二维空间的映射系数 $g_c=0.04$,模糊因子 $m=2$,特征隶属度范围 $[1.5,2.5]$,最大迭代次数 $M'=100$,聚类中心变化阈值 $T=0.001$,加权参数 $s=0.02$ 。本研究的文本挖掘实验流程如图2所示。

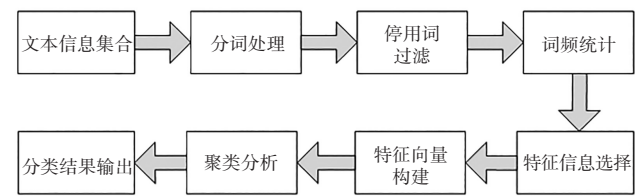


图2 语音文本挖掘实验流程

Fig. 2 Experimental process of speech text mining

根据以上实验过程与相关参数的设定,对本文提出的语音文本挖掘方法的性能进行测试,并依据实验结果分析本文方法的应用效果。

2.3 语音文本挖掘结果与分析

采用混合神经网络算法(方法1)、事件抽取技术(方法2)作为对比组,本文方法为实验组,以便于更加清楚地展示本文提出方法的优势。分别采用3种方法对该电网客服语音文本进行挖掘,统计不同方法挖掘结果的关键词查准率,结果如图3所示。

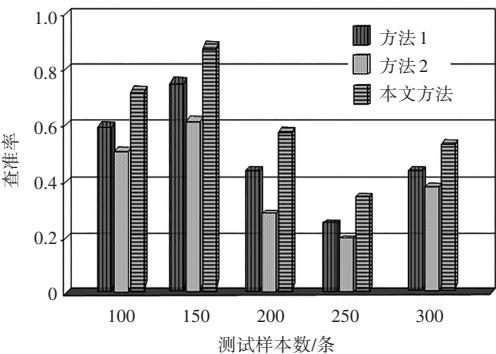


图3 语音文本挖掘查准率对比结果

Fig. 3 Comparison results of precision in speech text mining

与实验对比方法相比,本文方法的文本挖掘查准率显著高于对照组方法,说明本文方法在电网客服语音文本挖掘中具有较高的挖掘准确率。

2.4 对比实验与分析

对于三种方法的挖掘效率进行测试与分析,实验结果如表2所示。

由表2可以看出,在相同挖掘时间条件下,应用本文设计的方法对实验语音文本进行挖掘,得到的挖掘规模远大于方法1和方法2,说明本文方法能够在较短时间内挖掘出更多的文本数量,挖掘的效率较高。

表2 语音文本挖掘效率对比结果

Tab. 2 Comparison results of speech text mining efficiency

仿真时间/s	挖掘数量/条		
	方法 1	方法 2	本文方法
5	1 546	963	2 140
10	1 723	987	3 369
15	1 886	1 025	4 517
20	2 147	1 103	6 682
25	3 365	1 354	7 845
30	3 680	2 247	8 503
35	3 840	2 471	8 745
40	5 523	3 061	9 004

3 结论

本研究基于改进的模糊聚类算法,对电网在线客服的语音文本进行了深入挖掘和分析。通过文本数据预处理和提取特征以及识别数据间的关联模式,从而实现了语音文本的深度挖掘。本研究的意义在于促进模糊聚类算法在文本挖掘领域的应用和发展,为相关领域的研究提供新的思路和方法。未来,将继续深化这一领域的研究,探索更有效的文本挖掘技术和方法,以更好地服务于电网企业的客户服务和运营管理。

参考文献

[1] 张婧,陈增照,段超,等. 基于注意力机制的文本挖掘深度混合推荐模型[J]. 华中师范大学学报(自然科学版), 2023, 57(5): 733-740.

[2] Eldeeb G, Mohamed M. Transit electrification state of the art: a machine-learning based text mining approach[J]. Transportation Research Part D: Transport and Environment, 2023, 23(15): 245-246.

[3] 蔡方中,田宇阳. 基于BERT和LDA模型的酒店评论文本挖掘[J]. 计算机应用与软件, 2023, 40(7): 71-76, 90.

[4] 杨秀璋,武帅,廖文婧,等. 基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘研究[J]. 现代计算机, 2023, 29(8): 12-19.

[5] 耿新青,王正欧. 基于增量式模糊聚类算法的文本挖掘[J]. 南京理工大学学报, 2022, 46(5): 579-585, 593.

[6] 武帅,施奕,杨秀璋,等. 基于社交网络分析和LDA主题挖掘的短文本挖掘研究[J]. 现代电子技术, 2022, 45(20): 124-128.

[7] 麦晓庆,张天湖,王亮,等. 基于NLP的变电主设备缺陷文本自动化挖掘算法[J]. 信息技术, 2022, 46(9): 151-156.

[8] 杨帆,黄河,李海英,等. 基于ERNIE预训练的电力设备缺陷文本挖掘方法[J]. 电气传动, 2022, 52(14): 52-58.

[9] 陈婷,郭亚琪,彭杰,等. 基于相空间重构及模糊聚类的供电服务风险评价模型[J]. 自动化技术与应用, 2024, 43(10): 217-220.

[10] 丁思媛,乔晓东,张运良. 基于事件抽取技术的听证公开文本挖掘方法研究[J]. 情报杂志, 2022, 41(1): 52-59, 30.

[11] 毛泽强. 金融机构海量投诉数据分析与应用——基于LDA-TPA模型文本挖掘[J]. 金融发展评论, 2021(9): 81-95.

[12] 白勇,张占龙,熊隽迪. 基于FP-Growth算法和GRNN的电力知识文本挖掘[J]. 计算机科学, 2021, 48(8): 86-90.

[13] 唐晓彬,董曼茹,徐荣. 大数据背景下CPI预测问题的文本挖掘技术设计与应用[J]. 统计研究, 2021, 38(8): 146-160.

[14] 宋卓远,阚乾超,赵凯,等. 基于共词分析的国内文本挖掘研究[J]. 图书馆学刊, 2021, 43(4): 104-111.

[15] 童沐雨,刘建平,林熠来. 基于遗传算法的APP用户隐私保护文本挖掘系统设计[J]. 自动化技术与应用, 2024, 43(3): 116-119.