基于改进 SO-PMI 算法的电力大数据词典构建方法

隋石妍[1] 董佳林[1] 潘 伟[1]
SUI Shiyan DONG Jialin PAN Wei

摘 要 在大规模文本数据中，许多词汇对的共现情况非常稀少，导致 PMI 值的计算不稳定。传统 SO-PMI 算法在计算 PMI 值时，往往只考虑了词汇对在文本中的共现情况，而忽略了电力大数据特殊的语义关系和上下文关系，导致所计算的词间关联度不够精确，进而影响了电力大数据词典构建的覆盖率、专业性及准确性。为此，文章提出了一种基于改进 SO-PMI 算法的电力大数据词典构建方法。首先，对电力大数据文本进行分词处理，并对分词结果进行词性标注。计算每个词在文本中的词频 - 逆文档频率（TF-IDF）值，以筛选出的特征词作为候选词。利用改进 SO-PMI 算法，计算候选词集中每一对词的 PMI 值，在此过程中充分考虑电力大数据的特殊语义关系，并引入特定上下文窗口进行平滑处理，得到更准确的词间关联度，有效解决未针对电力大数据特殊语义关系而导致关联度不准确的问题。基于准确的关联度筛选出与电力大数据领域相关的强关联词对，构建电力大数据词典。实验结果表明：采用该方法构建的词典平均覆盖率高达 99.45%，其 AUC 值达到 0.95。与传统方法相比，该方法在识别电力大数据领域相关词汇方面表现出更高的全面性和准确性，所构建的词典在覆盖性、专业性和准确性方面均展现出显著优势。

关键词 改进 SO-PMI 算法；电力领域；大数据；分词；词典构建

doi：10.3969/j.issn.1672-9528.2025.04.038

0 引言

电力大数据作为智能电网的核心资源，蕴含着丰富的信息，对于电力系统的运行管理、故障预测、资源优化等方面具有不可替代的作用。在当前的电力大数据应用中，数据来源广泛、数据量巨大、数据类型多样，这些特点为数据的处理和分析带来了极大的挑战[1]。为更好地利用这些数据，需要构建一个全面、准确、易于使用的电力大数据词典，以帮助相关人员更好地理解、应用和交流相关术语和概念。然而，目前电力大数据词典的构建方法尚不完善，存在许多亟待解决的问题。例如文献 [2] 提出方法通过深入理解电力系统的物理本质，建立物理或数学模型，进而对数据和现象进行解释，但建立模型时未充分考虑电力大数据特殊语义关系，可能忽视术语内在联系，解释数据时出现偏差。文献 [3] 提出方法侧重于从大量数据中提取有用信息，发现数据之间的关联性和规律性，但未充分考虑电力大数据特殊语义关系和上下文关系，提取信息时可能误读数据含义，影响词典专业性。

针对这些问题，本文提出了一种基于改进 SO- PMI 算法构建电力大数据词典的方法。该方法利用改进的 SO- PMI 算法，结合电力大数据特定的语义关系和上下文关系精细计算关联度，从而提高了词典构建的准确性，为电力行业的数据处理和分析提供有力支持。

1 电力大数据分词处理

在电力大数据词典构建过程中，原始文本数据往往包含大量无关信息（如特殊字符、标点符号、HTML 标签等），这些信息的存在会干扰后续的分词、词性标注、特征词提取等步骤，从而影响词典构建的质量和效率。通过预处理，可以确保文本数据的准确性和一致性，提高分词和词性标注的准确性，从而为特征词提取和词典构建提供坚实的基础[4]。

预处理后的文本要进行分词，这是自然语言处理的关键步骤。根据电力大数据特点调整分词工具参数（如最大匹配长度、最小词频）优化效果，分词结果为词汇列表，每个词汇有一个或多个可能的分词结果。

对分词结果进行词性标注，遵循实词（名词、动词、形

1. 威海海源电力工程有限公司 山东威海 64200

容词等）和虚词（副词、介词、连词等）的标注规则，按需调整规则和参数优化效果，标注结果以词汇和对应词性形式呈现，如“电力 /n”（名词）、“设备 /n”（名词）、“故障 /n”（名词）等 [5]。

通过以上流程，不仅能对电力大数据文本进行分词处理和词性标注，还能为后续的特征词提取、词典构建等步骤提供坚实的基础。

2 电力大数据特征词提取

电力大数据信息量较大，涵盖众多电力领域的专业术语、设备名称以及运行状态描述等内容，各类信息相互交融。直接处理未提取特征词的电力大数据，会有数据冗余、难以聚焦关键信息等问题。为解决这一问题，本文提出对经分词处理和词性标注后的电力大数据提取特征词。提取特征词可缩小数据处理范围，依据阈值筛选出高频且与电力领域紧密相关的词汇。最终得到高频且与电力领域紧密相关的候选特征词集。这一词集为后续词间关系计算和强关联词对筛选提供高质量数据输入，有助于构建出覆盖率高、专业性和准确性强的词典。

利用上一步分词处理得到的词汇列表作为输入数据，对分词结果进行深入的词频统计，以计算每个词汇在整个文本集中出现的频次。这一统计过程需遵循公式：

$$\delta(w_i)=\sum_{j=1}^{M}\beta(w_i,t_j) \tag{1}$$

式中：δ 表示词频；$\beta(w_i, t_j)$ 表示指示函数，当 w_i 出现在 t_j 中时取值为 1，否则为 0；w_i 表示词汇列表中的第 i 个词汇；t_j 表示文本集中的第 j 个文本；M 表示文本集的总数 [6]。

基于词频统计的结果，筛选出一定数量的高频词汇，将其作为候选特征词。这一步骤旨在从海量词汇中初步筛选出出现频率较高、可能蕴含重要信息的词汇 [7]。筛选公式可以简化为：

$$S=\{w_i|\delta(w_i)\geq\varphi\} \tag{2}$$

式中：S 表示候选特征词集；φ 表示阈值，根据排名前 N% 的设定来确定，以此确保提取出的特征词不仅出现频率高，而且与电力领域紧密相关，从而提高了特征词的专业性和准确性。

3 基于改进 SO-PMI 算法计算词间关联度

在电力大数据领域，专业术语与概念关联复杂，尤其在设备故障描述、运维记录等特定上下文文本中。基于此，本文先通过特征词提取技术，从电力大数据文本筛选出候选特征词集用于后续计算。接着，依据特征词共现情况计算点互信息（PMI）值来量化关联程度。针对传统 SO-PMI 算法处理这类文本时因不能充分理解上下文存在局限，本文引入上下文权重和词性信息改进算法，最终算出特征词关联度矩阵，该矩阵揭示了词汇关联关系，为构建电力大数据词典提供重要依据。

PMI 值是一个衡量两个词汇在文本中共现程度的统计量，其计算公式为：

$$\mathrm{PMI}(S_1,S_2)=\log\frac{P(S_1,S_2)}{P(S_1)P(S_2)} \tag{3}$$

式中：$P(S_1, S_2)$ 表示候选特征词 S_1、S_2 在同一文本中共现的概率；$P(S_1)$、$P(S_2)$ 分别表示特征词 S_1、S_2 在文本集中出现的概率。PMI 值的大小直接反映了电力大数据特征词之间的关联紧密程度，即 PMI 值越大，说明这两个特征词在同一文本中共现的可能性越高，其关联性也就越强 [8]。

然而，传统 SO-PMI 算法在计算 PMI 值时，往往只考虑了词汇对在文本中的共现情况，而忽略了词汇所处的上下文环境。这可能导致算法无法准确捕捉到词汇之间的语义联系和细微差别。为此，本文在传统 SO-PMI 算法的基础上改进，引入上下文权重和词性信息等特征，对 PMI 值进行平滑处理。改进后的 PMI 值计算公式为：

$$\mathrm{PMI}_n(S_1,S_2)=\alpha\cdot\mathrm{PMI}(S_1,S_2)+\gamma\cdot\varpi(S_1,S_2)+\kappa\cdot Q(S_1,S_2) \tag{4}$$

式中：α、γ、κ 均表示调整参数，用于平衡不同特征对 PMI 值的影响；$\varpi(S_1, S_2)$ 表示候选特征词 S_1、S_2 在上下文中的权重，反映了在文本中的相对重要性；$Q(S_1, S_2)$ 表示候选特征词 S_1、S_2 的词性信息，用于捕捉其上下文关系和语义关系 [9-10]。

根据电力大数据的特点和需求调整算法参数 α、γ、κ 的取值，以提高词间关联度计算的准确性和稳定性。使用参数调整后的改进 SO-PMI 算法计算特征词之间的关联度矩阵，其计算公式为：

$$R=\left[\mathrm{PMI}_n(S_i,S_j)\right]_{N\times N} \tag{5}$$

式中：N 表示特征词集的长度；$\mathrm{PMI}_n(S_i, S_j)$ 表示特征词 S_i、S_j 之间的改进 PMI 值。关联度矩阵反映了特征词之间的关联关系，为后续词典构建提供了重要依据。

4 电力大数据词典构建

为构建一部结构清晰、语义精准且能够全面反映电力大数据领域词汇关系的词典，本文采取了通过关联度矩阵筛选关联度较高词对的策略，以解决直接构建词典面临的数据量庞大且信息杂乱的问题。在词典构建过程中，依据事先计算得出的词间关联度矩阵，设定合理阈值，精准控制候选词对的数量和质量，避免将语义关联微弱的词汇纳入词典。经过

词性标注、词义解释等后续处理，最终成功构建出一部既专业又实用的电力大数据词典，为相关领域的文本处理、分析及专业人士的查询和研究提供了有力的支持。

在词典构建过程中，根据之前计算得到的词间关联度矩阵，筛选出关联度较高的词对作为词典的候选词对，为词典构建提供基础。关联度筛选公式可以简化为：

$$S_p = \left\{ \left(S_i, S_j \right) \middle| R[i,j] \geq \psi \right\} \quad (6)$$

式中：S_p 表示候选词对；$R[i,j]$ 表示词间关联度矩阵中第 i 行第 j 列的元素，即特征词 S_i 和 S_j 之间的关联度值；ψ 阈值根据实际需求设定，用于控制候选词对的数量和质量。

在词典结构设计上，本文精心规划了包括词汇表、词性标注、词义解释等关键部分在内的词典框架，以确保词典的清晰性、易用性和可扩展性。电力大数据词典结构示例如表 1 所示。

表 1 电力大数据词典结构示例

序号	词汇表	词性标注	词义解释
1	变压器	名词	一种电力设备，用于改变电压。
2	电流	名词	电荷的定向移动形成的物理量。
3	发电	动词	利用能源产生电能的过程。
...	...	...	...

词典构建过程中，将筛选出的关联词对及其相关信息严格遵循设计结构进行组织和填充，对每个词汇实施词性标注和词义解释，确保词典的完整性和准确性。此外，词典还需定期更新和维护，以适应电力大数据的发展和变化，确保词典的时效性和准确性，使其能够持续为电力大数据文本处理和分析提供有力的支持。

5 实验分析

5.1 实验数据

本实验以某大型电力公司 2023 年的电力运行数据为样本。如：（1）电力负荷数据详细记录了该地区每小时的电力负荷量，总计 8 760 条记录，这些数据在不同季节和天气条件下展现出显著的波动性；（2）电压电流数据集则包含了多个变电站和线路的电压电流信息，每个站点和线路每小时记录 1 次，累计数 10 万个数据点；（3）故障记录数据则收录了数百起故障案例，涉及短路、过载、接地等多种故障类型。

在电力大数据词典的构建环节，实验选取了与电力系统紧密相关的 1 000 个词汇作为候选，这些词汇不仅来源于上述电力运行数据的描述性内容，还涵盖了电力行业的专业术语，如设备名称（变压器、断路器）、故障类型（短路、过载）以及操作指令（启动、停止）等。

5.2 实验环境

本实验采用了 Python 编程语言，并借助 NumPy、Pandas、Scikit-learn 等库来实现数据的处理与分析。对于改进 SO-PMI 算法，其实验参数设置如表 2 所示。

表 2 改进 SO-PMI 算法参数设置

序号	参数	数值
1	基准词列表	包含 100 个与电力系统相关的正面情感词和负面情感词。
2	PMI 计算阈值	0.2
3	SO-PMI 计算阈值	0.1
4	折扣系数	0.8

实验过程中，首先对原始数据进行了归一化处理，以消除不同变量间的量纲差异。随后，从预处理后的数据中提取出与电力系统相关的词汇，这些词汇经过自然语言处理技术的分词和词性标注，确保了其准确性和代表性。

5.3 实验结果

使用本文提出的基于改进 SO-PMI 算法的电力大数据词典构建方法，构建词典 A；同时，采用文献 [2] 和文献 [3] 中提出的两种传统构建方法，分别构建词典 B 和词典 C，以进行对比分析。为了评估 3 种方法在电力大数据领域的表现，选取了一个包含正例（相关词汇）和负例（非相关词汇）的测试词汇集，总数为 1 700 个。通过对测试词汇集进行覆盖测试，记录了每种方法覆盖的词汇数量，并得到了如表 3 所示的对比结果。

表 3 电力大数据词典覆盖程度对比结果

实验标号	本文方法	文献 [2] 方法	文献 [3] 方法
01	1 685	1 520	1 612
02	1 689	1 526	1 624
03	1 690	1 643	1 638
04	1 688	1 538	1 629
05	1 694	1 567	1 634
06	1 695	1 602	1 628

从表 3 的覆盖程度对比结果可以看出，基于改进 SO - PMI 算法构建的电力大数据词典（词典 A）在 6 次实验中的词典覆盖率均最高，平均覆盖率达 99.45%，显著高于文献 [2] 方法的 89.96% 和文献 [3] 方法的 95.74%。这表明该词典在电力大数据领域专业性和覆盖性更强，更符合电力大数据分析需求。本文方法的优势在于应用改进的 SO - PMI 算法，充分考虑电力大数据的上下文关系并引入特定上下文窗口进行平滑处理，得出更准确的词间关联度，有效解决了未针对电力大数据特殊语义关系而导致关联度不准的问题，提高了词

典在该领域的专业性和覆盖性。

为进一步评估每种词典的构建效果，根据每种词典的覆盖结果计算了真正率和假正率。计算公式分别为：

$$TPR = \frac{TP}{TP + FN} \quad (7)$$

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

式中：TPR 表示真正率；FPR 表示假正率；TP 表示被正确识别为相关词汇的数量；FN 表示被错误识别为非相关词汇但实际上为相关词汇的数量；FP 表示被错误识别为相关词汇但实际上为非相关词汇的数量；TN 表示被正确识别为非相关词汇的数量。在得到每种词典的 TPR 和 FPR 后，绘制 ROC 曲线，并计算 ROC 曲线下的面积，即 AUC 值（介于 0 到 1 之间，越接近 1，表示词典构建性能越好）。

由图 1 的对比结果可以看出，词典 A 的 ROC 曲线最接近左上角，这意味着在相同的假阳性率（FPR）下，其真阳性率（TPR）更高，能更精准地识别电力大数据领域相关词汇。词典 A 的 AUC 值为 0.95，接近理想值 1，进一步表明其分类效果佳。相比之下，词典 B 和词典 C 的 ROC 曲线位于词典 A 下方，AUC 值分别为 0.85 和 0.88，均低于词典 A。本文方法的关键优势是利用改进的 SO-PMI 算法计算词间关联度，考虑电力大数据上下文关系进行平滑处理，这提高了关联度准确性，解决了因未针对电力大数据特殊语义关系而导致关联度不准的问题。因此，基于准确关联度筛选强关联词对构建的词典 A，在识别相关词汇时准确性更高，其 ROC 曲线更接近左上角，AUC 值更接近 1。

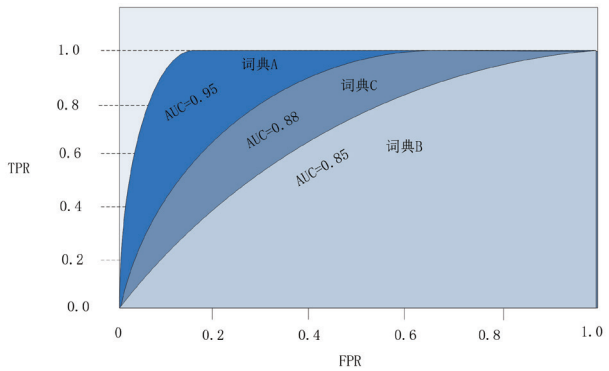

图 1 ROC 曲线及相应的 AUC 值

6 结语

本文针对大规模电力大数据文本中词汇对共现稀少导致的 PMI 值计算不稳定问题，提出了一种基于改进 SO-PMI 算法的电力大数据词典构建方法。该方法通过综合考虑电力大数据的特定语义关系和上下文关系，引入特定上下文窗口进行平滑处理，实现了更准确的词间关联度计算，提高了词典构建的准确性，为词义理解和文本分析奠定基础。实验结果显示，采用该方法构建的词典在覆盖性、专业性和准确性方面均表现出显著优势。此外，词典设计注重实用性和用户友好性，包含词汇表、词性标注和词义解释等清晰结构，便于查阅和使用，并为后续扩展和更新预留了空间。

参考文献：

[1] 刘驹 . 基于数据中台的电力大数据高效挖掘分析技术研究 [J]. 电工技术 , 2024(10): 76-81.

[2] 范海威 , 范士雄 , 李斌 , 等 . 电力调度领域专业词典构建方法研究 [J]. 电力信息与通信技术 , 2021, 19 (1): 57-65.

[3] 王书鸿 , 郑少明 , 刘中硕 , 等 . 面向某地区电网继电保护装置缺陷知识图谱构建的实体关系抽取 [J]. 电网技术 , 2023, 47 (5): 1874-1887.

[4] 袁金斗 , 潘明明 , 张腾 , 等 . 基于规则和词典的用电安全领域命名实体识别 [J]. 电子技术应用 , 2022, 48 (12): 22-27.

[5] 李坚林 , 张晨晨 , 赵昊然 , 等 . 基于多源数据融合的电网设备技术监督知识图谱构建 [J]. 电工电气 , 2021 (9): 60-63.

[6] 要权 . 面向三维可视化场景的电力大数据分析模型构建研究 [J]. 长江信息通信 , 2021, 34 (8): 124-126.

[7] 蒲天骄 , 谈元鹏 , 彭国政 , 等 . 电力领域知识图谱的构建与应用 [J]. 电网技术 , 2021, 45 (6): 2080-2091.

[8] 钱建煜 , 沈利 , 沈纪约 , 等 . 基于人机交互的发电知识图谱动态更新研究与应用 [J]. 电力大数据 , 2023, 26 (10): 58-66.

[9] 刘中硕 , 郑少明 , 陶畅 , 等 . 继电保护装置缺陷文本专业词典构建及其语言特性分析 [J]. 中国电力 , 2023, 56 (7): 146-155.

[10] 吴树芳 , 尹凯 . 基于敏感语义和复合共现的网络敏感词典构建研究 [J]. 情报科学 , 2023, 41 (10): 12-20.

【作者简介】

隋石妍（1990—），女，山东青岛人，硕士，工程师，研究方向：电力工程。

董佳林（1992—），男，山东威海人，本科，工程师，研究方向：电力工程。

潘伟（1995—），男，内蒙古满洲里人，本科，助理工程师，研究方向：电力工程。

（收稿日期：2024-12-10）