创新点/亮点：采用自然语言处理方法提高修井施工方案知识提取效率，解决了修井施工方案文本知识难以提取的难题，使得知识抽取模型精度提高 3.26%，召回率提高 1.69%，关联规则计算效率提高 34.38%。

引用格式：杨希军，孔红芳，赵东，易春飚，于国起. 自然语言方法提取油井修井施工信息提高智能化效率［J］. 石油钻采工艺，2024，46(4)：492-508. // YANG Xijun, KONG Hongfang, ZHAO Dong, YI Chunbiao, YU Guoqi. Extracting oil workover construction information using natural language methods to improve intelligent efficiency［J］. Oil Drilling & Production Technology, 2024, 46(4): 492-508.

自然语言方法提取油井修井施工信息提高智能化效率

杨希军*，孔红芳，赵东，易春飚，于国起

中国石油天然气股份有限公司大港油田分公司，天津滨海新区

*通信作者：杨希军，电子邮箱：dg_yangxj@163.com

基金项目：海上稠油超临界多源多元热流体发生机理及在储层中的作用机制研究 (编号：U22B2074)。

摘要：（目的意义）传统修井知识提取方法，存在人工提供效率低下且无法处理大规模数据的不足，导致修井措施制定水平缺乏科学性。**（方法过程）**为此设计一种基于标签权重的改进注意力机制，与预训练权重模型和 BiLSTM 共同构成修井知识实体提取模型，同时提出一种融合贝叶斯方法和 Hash 树的改进 Apriori 算法，形成了面向施工方案文本的两阶段修井知识智能分析与挖掘方法。**（结果现象）**该方法在大港油田开展应用，结果表明：修井知识实体提取模型的识别精度可达 81.83%，改进的 Apriori 模型挖掘频繁项集数量为 814 条，具有强关联实体组合 515 条，关联规则计算效率提高 34.38%。**（结论建议）**文章提出的修井知识智能分析与挖掘方法可提高修井知识提取效率，为石油工程领域数据提取、数字化建设提供思路。

关键词/主题词：人工智能；大数据；算法；修井；数字经济；室内测试；新质生产力；油气改革

中图分类号：P74；TE28 **文献标识码：**A **DOI:** 10.13639/j.odpt.202411056

收稿日期：2024−03−25；**修回日期：**2024−04−15；**录用日期：**2024−05−20；**编辑：**杨春莉

Extracting oil workover construction information using natural language methods to improve intelligent efficiency

YANG Xijun*, KONG Hongfang, ZHAO Dong, YI Chunbiao, YU Guoqi

PetroChina Dagang Oilfield Company, Binhai New Area, Tianjin, 300200, China

*Corresponding author. YANG Xijun, E-mail addresses: dg_yangxj@163.com

Abstract: Traditional methods for extracting workover knowledge have the shortcomings of low efficiency provided by manpower and inability to handle large-scale data, resulting in a lack of scientificity in the formulation level of workover measures. For this reason, in the entity extraction stage, an improved attention mechanism based on label weights is designed, which, together with the pre-trained weight model and Bidirectional Long Short-Term Memory(BiLSTM), forms a workover knowledge entity extraction model. In the association rule mining stage, an improved Apriori algorithm that integrates the Bayesian method and the Hash tree is proposed, thus forming a two-stage intelligent analysis and mining method for workover knowledge oriented to the texts of construction plans. Upon being applied in Dagang Oilfield, The results indicate that the recognition accuracy of the workover knowledge entity extraction model can reach 81.83%. The number of frequent itemsets mined by the improved Apriori model is 814, with 515 strongly associated entity combinations, and the computational efficiency of the association rules is increased by 34.38%. The intelligent analysis and mining method for workover knowledge proposed in this article can enhance the efficiency of workover knowledge extraction, providing ideas for data extraction and digital construction in the field of petroleum engineering.

Key words: Artificial intelligence; Big data; Algorithm; Workover; Digital economy; Laboratory test; New quality productivity; Oil and gas reform

https://doi.org/10.13639/j.odpt.202411056
Received 25 March 2024; Revised in revised form 15 April 2024; Accepted 20 May 2024

0 引言

修井作业作为在油气开发复杂且漫长的过程中的关键环节，是确保油气井正常运行、提高油气产量的重要手段和方法[1]。当前，修井作业施工方案的制定大多依赖于人工经验和历史数据。这种依赖经验的方式虽然在一定程度上保证了作业的顺利进行，但由于人工经验的局限性和应对复杂问题的能力不足，方案的制定效率低，且无法全面考虑到修井过程中潜在的各种事故隐患及其复杂的关联关系[2]。

此外，许多成功的修井作业方案在作业完成后并未被充分利用，这些方案文本中涵盖诸多复杂工序，例如洗井、压井、冲砂、检泵等作业环节的施工内容与技术细节，蕴含了大量宝贵的专家知识与经验，其中包含了大量的修井施工技术细节、操作注意事项、处理应急情况的方法等，有助于后续作业方案的优化。有效地挖掘和利用这些专家知识，不仅能够显著提高修井作业的效率，还能降低作业过程中的风险，提升油气田的开发效率[3]。修井知识挖掘的目的是通过分析和提取修井过程中的关键数据、操作经验和技术要点，帮助提升修井作业的效率和安全性。通过对历史修井记录、设备性能、地质信息以及施工环境等多维数据的挖掘，可以发现潜在的规律和优化点，从而指导未来修井工作中的决策制定，减少故障发生，降低成本，并提高井下作业的成功率[4]。

目前深度学习、机器学习等人工智能技术在全球油气工业领域内被广泛运用，智慧地质、智慧钻井以及智慧油田的概念相继被提出，但是针对对石油领域的知识挖掘研究较少，与其他热门行业相比，其发展程度尚显不足[5]。这导致石油行业在面对复杂的勘探、开采与生产等环节时，难以充分利用数据资源挖掘潜在价值。例如，在油藏描述方面，若缺乏高效知识挖掘，可能导致对储层特征与流体分布的认识不够精准，从而影响开采方案制定[6]。从技术层面看，石油领域数据具有多源性、复杂性和专业性等特点，传统知识挖掘方法适应性较差。而深度学习等新兴技术虽有潜力，但因起步晚、专业人才匮乏等，应用尚未普及[7]。同时，石油企业对知识挖掘重视程度不一，部分企业因投入成本与回报周期等考量，在相关技术研发与应用上积极性不高，这也在一定程度上制约了石油领域知识挖掘工作的推进与发展，不利于行业的长远创新与竞争力提升[8]。因此，提供一个高效可靠的修井知识挖掘与分析方法显得尤为迫切。

修井知识挖掘方法的本质是寻找具有一定关联的文本对，例如清水-洗井和压井-压井液，在这些具有关联关系的文本对中总结出修井方案中所存在的知识[9]。为了解现有油气领域内容的修井知识挖掘方法研究现状，文章对现有的修井知识挖掘方法进行了调查研究。调研发现修井知识挖掘的方法通常可划分为两类，一类是基于统计规则的方法，通过细致统计数据里的各种规律、频次以及关联关系，挖掘出潜在的知识要点[10]。另一类则是基于深度学习的方法，借助深度神经网络架构强大的学习能力，使模型自主学习数据复杂的特征与模式，从大量数据中深度探寻知识[11]，二者各有侧重。

基于统计学的知识挖掘方法主要是从海量的数据文本里，通过统计各个文本对出现的次数，进而将其作为判断是否为知识的重要依据，关联规则挖掘是一种被广泛应用且成效显著的知识挖掘途径，具备较高的计算准确性和可靠性[12]。Apriori 算法在关联规则挖掘领域占据着重要地位。例如，Kim 等学者提出利用 Apriori 文本挖掘方法，专门针对健康大数据进行深度挖掘，成功地提取出关联特征信息，极大地推动了健康产业大数据的应用与发展，为健康领域的数据化研究提供了新的思路与方法[13]。陈碧云等研究者则采用关联规则挖掘技术，借助 Apriori 算法深入剖析规则，进而提出诱因的诱发度计算方法,他们通过对某区域近 5 年的事故实例展开全面的关联分析，有效地验证了该方法的科学性与有效性，为事故预防与分析提供了有力的技术支撑[14]。于溪芮等以 827 起氢系统安全事故为数据集，通过构建事故致因网络、交互分析致因与后果及 Apriori 算法，得出多种事故致因组合及事故与相关因素强关联结论，为氢安全领域的事故防范做出

了贡献[15]。随着研究的深入推进，Apriori 算法也暴露出一些局限性，例如算法效率低，同时知识挖掘片面。尽管后续出现了 Apriori 的变种算法，但依然难以彻底解决其固有问题。例如，国汉军等在动火作业事故原因的知识关联分析中，运用内外理论与 Apriori 算法虽取得了一定成果，推动了动火作业事故预防领域的进步，但是 Apriori 算法由于其计算开销较大，并且只能进行关联规则的静态挖掘，使得算法执行效率较低[16]。这不仅导致知识挖掘的过程耗时较长，而且在挖掘结果上难以做到全面、系统，往往会遗漏一些潜在的关联信息，具有片面性。

基于深度学习的知识挖掘方法主要以深度学习网络为核心，并结合回归等有监督方法展开，这些有监督方法能够依据已标记的数据进行学习与分析，从而挖掘出潜在的知识模式与规律。在油气田领域，钟仪华等研究者运用深度卷积网络和循环神经网络，致力于对油气田知识进行深度挖掘与精准预测。然而，油气田开发是一个极为复杂的过程，具有显著的时空动态变化特性以及复杂的多变量相互依赖关系，在这种情境下，卷积神经网络和循环神经网络暴露出了一定的局限性[17]。例如，面对油田开发中长时序依赖的数据，它们难以精准地捕捉到远距离数据之间的关联关系。在处理多模态数据时，由于数据来源的多样性和复杂性，模型的泛化能力较弱。此外，尽管目前提出了借助模型库和知识库来筛选最佳预测模型的思路，但在实际操作中，如何高效且自动地从海量的模型中准确甄别出最为合适的那一款模型，依旧是横亘在研究者面前的一道难题[18]。海量的模型意味着庞大的搜索空间，要在其中快速定位到最优解难度较大。并且，当前深度学习方法在实际应用场景中，对计算资源的消耗量极大，这无疑增加了研究与应用的成本。同时，调参过程也极为复杂，不同的参数设置可能会导致模型性能的巨大差异，需要耗费大量的时间与精力去反复调试，这些都在一定程度上制约了基于深度学习的知识挖掘方法在油气田开发等领域的广泛应用与深入发展[19]。

在基于深度学习的知识挖掘方法中，能够以较高精度识别文本数据中的知识实体是开展知识挖掘工作的重要前提条件。命名实体识别作为自然语言处理领域的一项核心任务，其主要目标是从繁杂的文本当中自动且精准地识别并分类特定的实体类型，涵盖了人名、地点、组织机构名等丰富类别[20]，可以胜任修井知识的提取识别。在油气开发这一特定领域，命名实体识别已经得到初步的应用实践。为有效挖掘石油生产数据，鉴于传统统计模型及部分现有深度学习方法在命名实体识别方面存在不足，如传统统计模型依赖大量注释语料库、部分深度学习方法信息传递或特征提取有缺陷，任伟建等提出基于 XLBIC（XLNet-BiGRU-IDCNN-CRF）的命名实体识别模型，该模型在自建石油开采数据集上对多数实体识别效果良好[21]。然而，XLBIC 模型未考虑不同油田或更广泛石油文本场景下的适应性，在遭遇非标准化或者超出油气开发领域范围的文本时，其适应能力略显不足，在处理这类文本时可能会出现识别错误或遗漏等情况。钟源等研究者则聚焦于油气勘探领域知识构建，提出了 BiLSTM + CRF 模型。该模型能够出色地完成实体、属性抽取模型的训练任务，在油气勘探知识挖掘方面展现出一定的潜力然而，其训练过程不可避免地会受到标注数据量和质量的双重限制[22]。例如标注数据量过少，模型可能无法充分学习到各类实体和属性的特征，标注数据质量不高，例如存在标注错误或模糊不清的情况，也会对模型的训练效果产生负面影响，进而制约模型在实际应用中的准确性和可靠性[23]。

综上所述，无论是基于统计学的方法还是有深度学习的方法，目前均存在算法效率较低，挖掘知识不全面的问题，同时命名实体识别作为有监督挖掘方法的前提，对修井知识的识别精度无法保证，缺乏较强的泛化能力[24]。因此，如何结合现代先进的人工智能技术，尤其是利用深度学习和关联规则分析方法系统地提取和利用修井知识，成为了当前研究的热点。

通过对上述各类方法的优缺点对比分析，可得出单一的方法难以高效解决修井知识挖掘的问题的结论。因此，提出了一种面向施工方案文本的两阶段修井知识智能分析与挖掘方法。该方法融合了深度学习与关联规则挖掘方法，为高效理解修井施工文本开辟了行之有效的新路径。在施工方案文本的实体提取阶

段，文章提出一种基于标签权重的改进注意力机制 IAMILW，以明确不同位置标签的语义角色，并与预训练权重模型 Bert [25] 和时序模型 BiLSTM 共同构成修井知识实体提取模型 Bert-BIIAM。在关联规则挖掘阶段，设计一种融合贝叶斯方法和 Hash 树的改进的 BH-Apriori 算法，以提高大量设计施工设计方案文本中修井知识的利用率，为后续修井过程的智能化提供指导。主要创新点如下：

(1) 针对当前修井作业施工记录中知识未得到充分利用的问题，提出了一种融合深度学习与关联规则挖掘的面向施工方案文本的两阶段修井知识智能分析与挖掘方法。

(2) 针对当前 BiLSTM+CRF、Bert 等模型在修井文本实体识别的精度不高，设计了一种基于融合标签加权的改进注意力机制，并组成 Bert-BIIAM 模型。

(3) 针对传统 Apriori 算法在规则挖掘的时效性低以及计算复杂度高，提出了 BH-Apriori 算法以提高算法时效性并保证关联规则的质量。

1 方法过程

1.1 研究方法过程

修井施工文本包含工程质量控制、进度控制与安全管理等相关内容，诸如施工周报、安全隐患排查报告等，其为修井工程建设的具体呈现。修井工程施工关联多元管理主体与不同工程单元，施工文本信息记录有差异，所以从多源文本提取有效信息是文本挖掘的重要前提 [26]。

考虑到单一的统计学方法与深度学习方法无法处理此类复杂文本，因此针对修井施工设计中的文本内容，结合命名实体识别和关联规则的方法，提出融合深度学习与关联规则相结合的两阶段修井实体知识挖掘及分析方法。设计方案文本的修井知识挖掘方法及流程为：

(1) 收集修井施工文档，分析施工方案文档中需要做出知识实体识别的类别数量，建立分类为措施、指标、现象以及物质的实体知识标签体系，同时，对施工方案进行知识识别和知识挖掘的数据预处理，剔除冗余文本项；

(2) 基于 Bert-BIIAM 模型，对经预处理的施工方案文本数据进行修井实体知识的识别提取，建立措施、物质、现象以及指标的知识实体集合，完成关联规则挖掘的数据准备；

(3) 对识别提取到的四类知识实体集合，做出基于 BH-Apriori 算法的关联规则挖掘，分析不同知识集合内的知识实体之间的关联程度，并总结出相关的修井知识；

(4) 以具体的修井施工方案文本为例，按照 (1)(2)(3) 所述步骤总结提取施工文本中所蕴含的知识，并分析不同知识类别之间的关联程度。

1.1.1 修井知识实体识别方法

在施工方案设计的修井知识提取过程中，存在一条文本中可能出现长短不一的实体，同一个实体在不同的文本扮演不同的语义角色，容易误抽取出指定实体之外的关键词或短语的三个知识抽取难题。具体如下：

(1) 一条文本中可能出现长短不一的实体，如在“暂堵剂 24 m^3+清水 30 m^3 正冲砂，全部漏失，下完井”文本中“暂堵剂”、“全部漏失”、“冲砂”等；

(2) 同一个实体在不同的文本中扮演不同的角色，如在“清水 30 m^3，正冲砂，井段：明油三组_1 遇阻”和“清水 30 m^3，正洗井，漏失清水 10 m^3”中，“清水”在前者的角色为物质，而在后者中则是物质和现象均有；

(3) 容易误抽取出指定实体之外的关键词或短语，如在“活动捞砂井段明油二组_3 至明油三组_1 遇阻，起出井内全部管柱”中，容易提取出“井段”等不相关实体。

为了解决在修井知识抽取过程中出现的上述难点，提出了 Bert-BIIAM 模型。Bert-BIIAM 整合了预训练模型 Bert、双向长短时记忆模型 BiLSTM 以及融合标签加权的改进注意力机制 IAMILW 模型，其中 Bert 凭借其强大的语言表征能力，负责生成输入文本的词向量嵌入表示，为文本的初步数字化编码提供高维且语义

丰富的向量形式[27]；BiLSTM 在文本训练流程中承担着关键任务，通过其独特的双向结构特性，有效地提取文本的上下文信息，增强对文本语义连贯性与逻辑性的理解与把握[28]；IAMILW 模型则聚焦于实体角色分析，运用融合标签加权的改进注意力机制，精准计算不同实体在不同位置所扮演角色的概率，从而为深入挖掘文本中的实体关系与语义结构提供有力支持，使得该模型在知识挖掘与文本分析领域具备独特的优势与应用潜力。

一段待执行修井实体识别任务的文本，首先在输入模型时，会被划分成一段 Token 序列，序列中会添加表示文本开始的 [CLS] 标志位以及 [SEP] 语句划分标志位,最终，输入的 Token 序列会经过段落嵌入、编码嵌入和位置嵌入的三种嵌入方式，最终输出一个可以用来计算的向量[29]。

对于具有时序特性的文本数据而言，其在 LSTM 单元中的信息流转遵循特定规律，即当前 LSTM 单元的输出由上一单元的外部状态予以决定[30]。

多头注意力机制是一种在自然语言处理及其他深度学习领域广泛应用且极具影响力的技术，其核心原理是基于注意力机制，通过多个并行的注意力头来计算不同子空间的注意力分布，使它能够从不同的表示子空间中捕捉输入序列的特征信息，例如，在处理文本序列时，一个头可能侧重于词汇语义层面的关联，另一个头则聚焦于语法结构或句子间逻辑关系等方面[31]。这些不同头的输出结果在最后会被拼接或进行其他融合操作，从而得到一个综合了多方面信息、更丰富且全面的序列表示。这种机制有效地增加了模型对输入信息的感知能力和表达能力，使其能够更好地处理长序列数据中的复杂依赖关系和多种特征信息，IAMILW 机制基于多头注意力机制改进[32]，如图 1 所示。

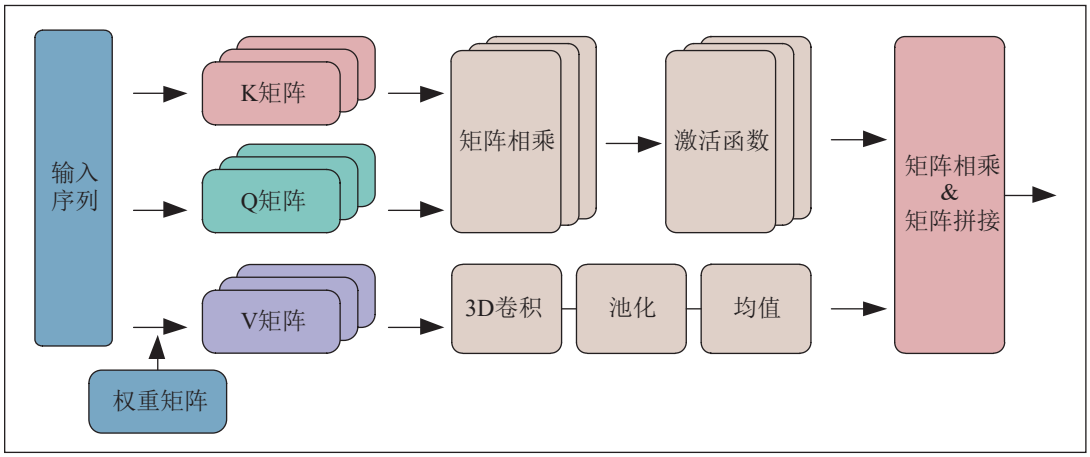

图 1 融合标签加权的改进注意力机制 (IAMILW)

Fig. 1 The improved attention mechanism incorporating label weights

从图 1 中可以看出，输入序列首先被分别处理生成 K、Q、V 三个矩阵，其中 K 和 Q 通过矩阵乘法操作后经过激活函数，再与经过 3D 卷积、池化和求均值操作的 V 矩阵进行相乘和拼接操作，得到注意力权重。IAMILW 架构通过多个处理路径，充分利用了多头注意力机制，能够从不同角度捕捉输入序列的特征，融合多方面信息，提升对长序列数据处理的能力。设一段经过文本信息嵌入表示层输出的文本序列为 P，见式 (1)。

$$\begin{cases} K_P = W^K P \\ Q_P = W^Q P \\ V_P = W^{label} P \\ ASM = \dfrac{Q_P K_P^T}{\sqrt{d_K}} \\ \mathrm{Att}(Q_P, K_P, V_P') = softmax(ASM) \cdot \mathrm{mean}(V_P') \end{cases} \tag{1}$$

式中: W^K, W^Q, W^{label}分别表示与K_P, Q_P, V_P对应的权重矩阵, 特别地, W^{label}是文本预处理过程中标注的实体标签的权重矩阵。Att(Q_P, K_P, V_P')表示 IAMILW 输出的注意力得分矩阵, K_P^T表示对向量K_P的转置, d_K表示目标向量Q_P与上下文信息向量K_P的维度。

对于向量V_P, 为了可以更好的提取文本中的特征, 对这个向量进行了一次卷积和池化操作[33]。为了保持向量的维度信息, 在卷积和池化操作中分别选用 5×5 的卷积核和池化核。

将注意力得分矩阵转化为一维的向量, 向量中的元素代表了实体的标签类型, 输出这些符合标签类型的实体, 即可得到修井文本中的知识实体。将提取的修井知识实体按照措施、物质、指标和现象组成四类实体集合, 为后续关联规则分析做好数据准备。

1.1.2 修井实体知识关联规则挖掘

在修井施工文本理解分析环节, 面对大量涵盖各类施工信息的文本资料, 要运用关联分析的方法, 厘清文本中所阐述的具体工程细节, 精确掌握相应工程所涉及的施工操作步骤以及所需的材料、设备等, 以此提炼出工程施工方面的关键知识要点。关联规则, 旨在从海量数据中挖掘出频繁项集以及对应的关联关系, 从而揭示数据项之间的潜在关系, 这些规则通常以“前提-结果”的形式表达[34]。Apriori 算法在关联规则挖掘中比较常见, 主要用于从大量的事务数据中发现频繁项集和关联规则, 但其存在明显的缺点, 主要是时间消耗较大, 尤其是在处理大规模数据集时, 计算频繁项集所需的多次数据库扫描导致了算法效率低下[35]。此外, 算法依赖于用户人为设定的最小支持度和最小置信度, 这可能导致修井知识挖掘结果的不稳定性和不准确性。为了克服上述不足, 提出了 BH-Apriori 算法: 引入自适应支持度和置信度计算方法, 通过贝叶斯方法更新在每次 N 项频繁项集时最小支持度和最小置信度[36]。使得最小支持度以及置信度在关联规则挖掘过程动态变化, 保证挖掘的规则的质量; 利用 Hash 树优化算法频繁项集的计算效率, Hash 树可通过将项集进行分组和存储, 有效减少候选集的数量, 从而降低数据库的扫描次数, 提高运算效率[37]。

BH-Apriori 算法基于贝叶斯方法更新挖掘过程中的最小支持度和最小置信度, 设$\{S_{\min}^{(0)}, S_{\min}^{(1)}, \cdots, S_{\min}^{(n)}\}$和$\{C_{\min}^{(0)}, C_{\min}^{(1)}, \cdots, C_{\min}^{(n)}\}$为知识挖掘过程中的最小支持度和最小置信度集合, 且最小支持度 S 和最小置信度 C 服从分布 Beta 分布[38], 对于挖掘过程中的最小支持度的计算方法, 见式 (2)。

$$\begin{cases} S_{\min}^{k} = E\left(\dfrac{P(S_{\min}^{k}) \cdot N(S_{\min}^{k-1})(1-S_{\min}^{k-1})^{N(X)-N(S_{\min}^{k-1})}}{P(L_n)}\right) \\ C_{\min}^{k} = E\left(\dfrac{P(C_{\min}^{k}) \cdot N(C_{\min}^{k-1})(1-C_{\min}^{k-1})^{N(X)-N(C_{\min}^{k-1})}}{P(L_n)}\right) \end{cases} \tag{2}$$

式中: $N(S_{\min}^{k-1})$表示L_n在过程中满足最小支持度$S_{\min}^{k-1}$的项集数量, $N(X)$表示频繁项集的数量。$E(\cdot)$表示期望, $P(S_{\min}^{k})$和$P(C_{\min}^{k})$表示最小支持度$S_{\min}^{k}$和最小置信度$C_{\min}^{k}$在先验分布中的概率, $P(L_n)$表示 n 项频繁项集在 n-1 项频繁项集中的概率。

在得到最小支持度$S_{\min}$和最小置信度$C_{\min}$后, 修井知识的关联规则的计算方式, 见式 (3)。

$$Q = \left\{ (V_k, V_g) \left| \begin{array}{l} \sup(V_k, V_g) = \dfrac{\text{count}(V_k \cup V_g)}{|C|} > S_{\min} \\ \text{con}(V_k, V_g) = \dfrac{\text{count}(V_k \cup V_g)}{|C|} > C_{\min} \end{array} \right., V_k, V_g \in R, k \neq g \right\} \tag{3}$$

式中: Q 表示修井施工文档中修井知识实体的关联规则挖掘所得的频繁项集, (V_k, V_g)表示实体V_k与V_g之间的关联规则, $\sup(V_k, V_g)$和$\text{con}(V_k, V_g)$分别表示候选项集 C 在挖掘知识实体V_k和V_g之间知识的支持度和置信度, $|C|$表示候选项集 C 的数量, $\text{count}(V_k \cup V_g)$表示候选项集 C 中$V_k$和$V_g$在同一项集中出现的次数。

BH-Apriori 算法通过初始化的最小支持度与最小置信度分布, 扫描整个修井知识数据集生成可用于知识挖掘的候选项集 C_1, 随后利用 hash 和贝叶斯定理动态更新最小支持度与最小置信度的阈值, 实现多项频繁项集的挖掘, BH-Apriori 算法流程图, 如图 2 所示。

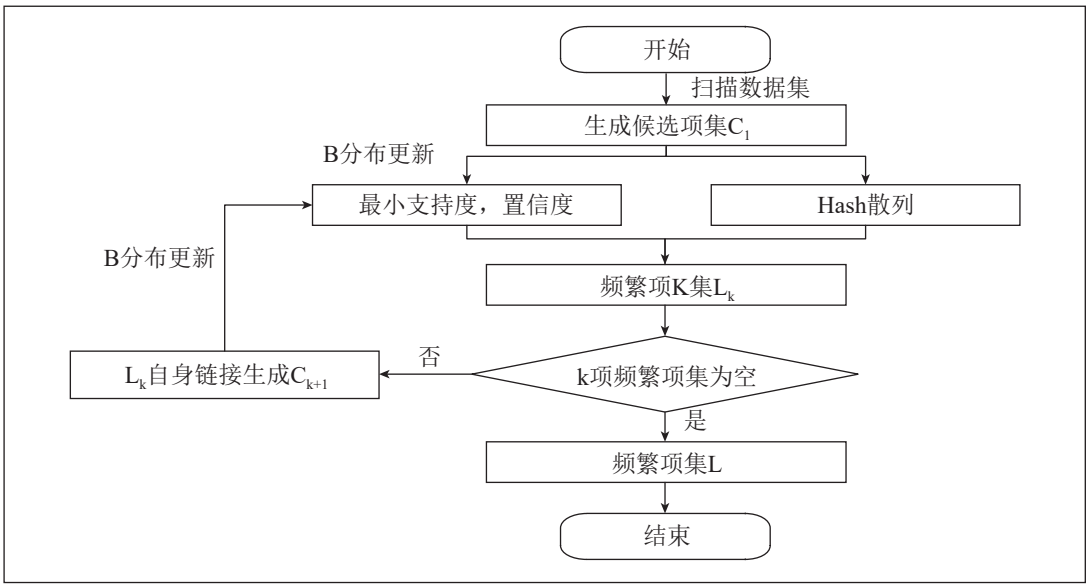

图 2 BH-Apriori 算法流程

Fig. 2 the algorithm process of BH-Apriori

从图 2 中可以看出，对于一个待挖掘的修井知识实体集合 D，BH-Apriori 算法在挖掘过程中通过比对当前挖掘步骤的支持度、置信度与当前步骤下的最小支持度、最小置信度的大小，输出频繁项集，具体流程如下：

(1) 扫描修井知识实体集合 D，形成 1 项-频繁候选集C_1并定义最小支持度S_{min}和最小置信度C_{min}的初始分布 B；

(2) 从筛选出的候选项集C_1，通过散列函数，形成散列表[39]，将散列表中的频繁项的支持度与置信度与最小支持度和最小置信度进行对比，形成 1 项-频繁项集L_1；

(3) 根据 1 项-频繁项集L_1，挖掘 2 项候选项集C_2，通过散列函数形成散列表，同时依据散列表,更新最小支持度和最小置信度的 B 分布，计算 B 分布的期望作为新的最小支持度和最小置信度，形成 2 项-频繁项集L_2；

(4) 借助 (2)(3) 思想推广至挖掘 n 项-频繁项集，判断是否为空，不为空，则继续进行挖掘，为空，则输出频繁项集，算法结束。

1.2 应用过程

以大港油田 2300 条修井文本数据进行实验，这些文本数据中包括历次修井作业的方案设计、地质设计和举升设计。每份报告都全面记录了包括施工进度、资源管理以及工序等可以全面描述修井施工工艺的技术细节，蕴含了大量的宝贵知识及经验。通过对这些数据进行深入挖掘和分析，可以发现隐藏在其中的规律和关联。例如，在修井方案设计中，不同地质条件下的修井策略往往存在差异，而这些差异在文本数据中会有所体现[40]。

进一步分析能揭示出特定施工工序与资源投入之间的关系，以及施工进度受地质因素影响的程度。这有助于精准制定未来修井作业计划，提高施工效率，降低成本与风险，为大港油田乃至石油领域内的修井工程优化提供极具价值的决策参考依据，推动油田开采作业的可持续发展。

1.2.1 修井文本数据预处理

在进行相应的实体识别以及关联规则挖掘任务时，施工记录中的高度距离和工具型号等因素会对修井

知识识别精度产生影响。施工记录中的高度距离往往存在多种单位和精度表述，而工具型号则可能因型号众多且复杂而干扰实体识别。这些因素的存在会导致在识别实体过程中出现误判或遗漏，进而影响到后续关联规则挖掘的准确性。为了避免这种情况，需要对高度距离和工具型号进行模糊处理。具体而言，采用特殊字符来替代这些高度距离和工具类型，这样可以减少不必要的干扰，提高实体识别和关联规则挖掘的精准度，保障整个分析流程的有效性和可靠性。

在修井施工方案设计中，采用特定字符代替施工高度，例如“明油一组”等。在“明油一组”所代表的高度范围内，依据施工文档中的高度分布，将其进一步划分，划分时以 50 m 或 100 m 为间隔形成不同的高度区间，并用在“明油一组”后加下划线和数字的形式来表示这些细分的高度范围，如“明油一组_2”，对应 700 m 到 1400 m 的深度区间。这样在进行知识实体提取时，模型可以根据字符规则快速识别和提取出相关的高度范围数据，有助于从施工文档中挖掘有价值的信息。同样，对于在修井施工方案设计文档出现的具有多种型号的工具进行了模糊处理，例如对 (平式) 油管、通井规、平式接箍等具有不同规格的工具做了编号。对于文档中出现的其他型号的工具，由于其型号相对固定，因此对其不做模糊处理。

1.2.2 施工设计文本实体类型确定

收集的修井施工方案设计文档包含了修井施工中的地质设计、工艺设计和施工设计。其中，地质设计文档中包含了对油井的历次施工作业记录简述，记录了历次作业的目的、采用的工具和方法、出现的问题以及具体的解决措施。工艺设计文档则侧重于阐述具体的修井工艺流程，包括各个环节的操作规范、技术参数设定以及质量控制要点等。施工设计文档则着眼于整个施工项目的组织与安排，涵盖了施工队伍的组建、施工进度的规划、施工设备与材料的调配等方面。

将修井施工方案历次作业记录中出现的实体界定为四个实体知识类型，分别为措施、物质、指标和现象四类，其中措施代表在历次作业中采取的各项操作措施，包括井下作业、压裂、酸化、堵水等工艺；物质类则表示在作业过程中使用的各种材料与化学药剂，如泥浆、封堵剂、卤水等；指标涉及在作业过程中设备探入井下的高度范围，物质材料的用料多少；现象类则记录了作业过程中观察到的井下及地表现象，包括异常现象的出现、设备运行状态的变化等。

在后续的修井施工方案设计文档实体标注过程中，分别将措施、物质、指标以及现象四类实体的标签命名为 Me、Sub、Ti 和 Phe，采用 BME 三段集标注方式，特别地，对文档中出现非实体采用 O 表示。通过这样一套标注规则与方法，能够有效提升修井施工方案设计文档实体标注的准确性、一致性以及可操作性，为基于这些标注数据开展的深入研究与模型开发奠定坚实的基础，确保在修井知识挖掘、关联规则分析以及实体识别模型构建等相关工作中能够获取高质量的数据支持。

1.2.3 实验评价指标建立

1) 修井知识实体提取模型 Bert-BIIAM 实验评价指标说明。将标注好的大港油田施工文本数据，按照固定比例分类，分为训练集、测试集输入到 Bert-BIIAM 模型中，可以得到实体提取模型对 4 类的实体的识别精度。将标注完成且划分为训练集和验证集的文本序列输入 Bert-BIIAM 模型进行训练，优化网络提高文本识别精度，对于修井知识文本识别的效果的评估，使用传统的精确度 (Precision)、召回率 (Recall)、F1 值作为评判[41]，其中精确度能够反映 Bert-BIIAM 模型预测为正类时的准确程度；召回率体现模型对修井知识实体的识别能力；F1 值则可以能够更全面地评估模型的性能。

针对已经标注的文本序列，按照 4∶2∶1 的比例划分成训练集、验证集和测试集，用于模型的训练以及精度的测试。定义文本维度为 768 维，将划分的训练集送入 Bert-BIIAM 模型进行训练。将整个修井施工方案设计实体知识识别网络的训练超参数为：输入的文本维度为 768，总迭代次数为 100 次，学习率为 0.001，隐含的卷积层数为 96。每次批处理的文本序列设置为 4。

2) 修井知识关联规则挖掘 BH-Apriori 实验评价指标说明。在修井施工文本的解析流程中，存有大量涵盖多元施工要点的资料文档。需运用关联分析手段，深度梳理文本所记载的专项工程细节，精准洞察工程所涉及的具体施工工序流程、适配材料以及设备规格等关键信息，进而有效提取工程施工领域的核心知识要点，为后续施工实践提供精准的理论支撑。将从修井施工方案设计知识实体识别任务中的 2300 个句子中提取出的实体作为关联规则挖掘的输入事务集。设置最小置信度和最小支持度的初始 B 分布参数分别为α_s为 0.1，β_s为 0.2，α_c为 0.1，β_c为 0.3。扫描整个事务数据集，开始挖掘事务集中存在的关联关系。

在关联规则实验挖掘的评价指标部分，使用 BH-Apriori 算法的消耗时间以及关联规则的挖掘条数作为评价指标，以此判断改进 BH-Apriori 算法是否符合预期。关于修井知识挖掘与关联规则分析的算法精度、事件效率以及挖掘条数的分析，将在后文讨论。

1.2.4 实验流程设计

1) 完成对修井方案文本的预处理后，依据上文所设置的 Bert-BIIAM 模型参数，开始修井知识实体提取实验，主要步骤安排如下：

第 1 步：收集大港油田港东港西地区共 20 口油井近 3 年的施工方案，按照前文所述预处理方法，对方案文本进行数据处理及标注，形成 2300 条文本数据。

第 2 步：将收集并处理好的 2300 条文本数据，划分为训练集、验证集输入到 Bert-BIIAM 模型，得到模型的实体识别结果。

第 3 步：将测试集输入到 Bert-BIIAM 模型，得到模型对修井实体的识别效果，记录模型对不同实体类别的识别精度、召回率以及 F1 值，以评价模型在修井知识提取领域的表现。

2) 将修井实体提取阶段所提取的不同实体作为关联规则分析的输入数据集，开展关联规则挖掘实验，主要步骤安排如下：

第 1 步：将从 Bert-BIIAM 模型输出的修井知识实体，组织成可用于关联规则分析的事务集，开展修井知识实体之间的关联规则挖掘。

第 2 步，初始化最小支持度和最小置信度的 Beta 分布参数分别为α_s为 0.1，β_s为 0.2，α_c为 0.1，β_c为 0.3。其中α_s，β_s为最小支持度的 Beta 分布的参数，α_c，β_c为最小置信度的 Beta 分布的参数。

第 3 步，将修井知识事务集送入至 BH-Apriori 算法入口，挖掘不同实体之间关联规则。记录算法运行时间以及挖掘的条数，将挖掘出的关联规则按照实体类别统计，分析不同类别的频繁项集组合中所蕴含的知识。

2 结果现象讨论

2.1 知识抽取模型精度及召回率讨论

对于修井知识提取的评价指标，选取修井知识抽取时的模型精度作为模型评价的指标。同样。为了更加全面的评价模型在修井文本识别领域的泛化能力，选择召回率以及 F1 值佐证。

在形成了含有措施、物质、指标、现象的四类实体的修井文本数据集后，将修井数据集输入至 Bert-BIIAM 模型进行训练，记录训练过程中每一类实体的识别精度、召回率以及 F1 值。在模型的训练过程中，训练的超参数为：Bert-BIIAM 模型的总迭代次数为 100 次，每次批处理的文本序列数量为 4。在 Bert-BIIAM 模型的训练过程中，发现当迭代次数达到 80 次时，同时在后续的训练过程中每类实体的精度曲线收敛，说明此时模型训练完成。记录此时模型对各类实体的识别精度收敛时的精度、召回率以及 F1 值作为模型的评判指标。在修井施工方案设计中实体识别实验中，Bert-BIIAM 模型对各类知识实体的识别的平均精度、平均召回率以及平均 F1 值分别为 81.83%、80.68%、81.23%，其中 Bert-BIIAM 模型对 4 类修井知识实体的三类表现数据，如图 3 所示。

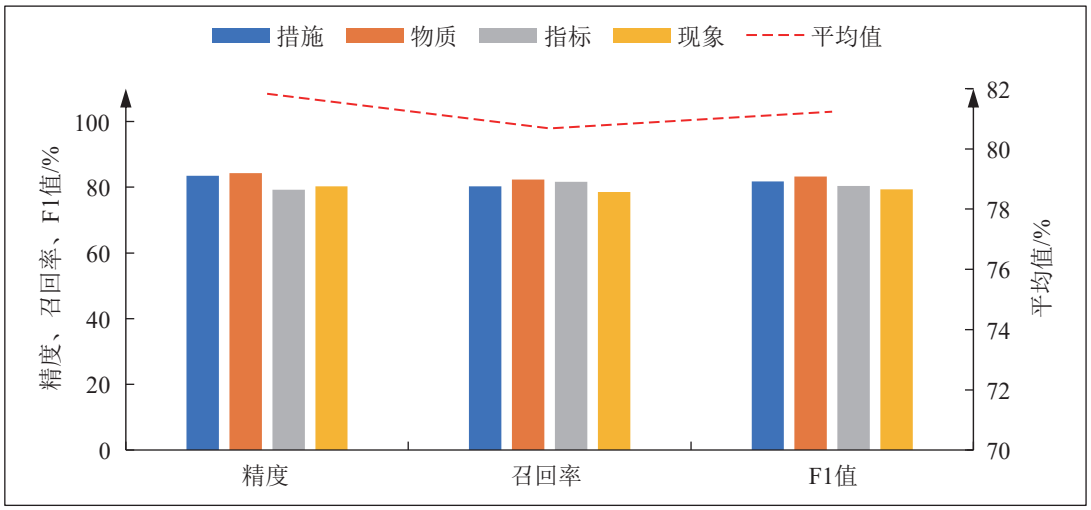

图 3　实体识别效果评价指标

Fig. 3　Evaluation indicators for knowledge entity recognition effectiveness

从图 3 中可以看出，在进行实体识别任务时，Bert-BIIAM 模型的精度、召回率和 F1 值均达到 80%，进一步说明模型在修井过程中的实体识别任务中可靠性和精确性较好。在进行实体识别任务时 Bert-BIIAM 模型的精度达到 81.83%，这意味着在所有被模型判定为正类的样本中，真正属于正类的样本比例较高，模型对实体识别的准确程度较好；召回率为 80.68%，表明模型能够成功识别出的实际正类样本在所有正类样本中的占比较大，漏判的情况较少；F1 值达到 81.23%，F1 值综合考虑了精度和召回率，进一步验证了模型在实体识别任务中的整体性能优良。这些数据充分表明，Bert-BIIAM 模型在修井相关的实体识别场景下的表现可靠且精确。

2.1.1　Bert-BIIAM 模型与传统模型对比

为了验证改进模型的效果，设置了 LSTM+CRF、Bert_BiLSTMCRF、BertLinear、Bert_CNN 和 Bert-BITWA 模型的对比实验。对比实验结果，如图 4 所示。

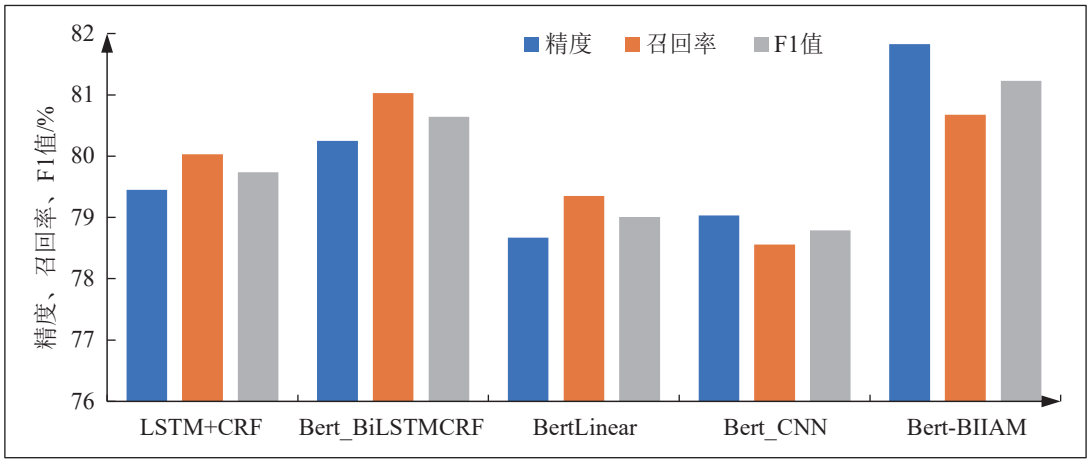

图 4　不同模型精度、召回率、F1 值指标对比

Fig. 4　Comparison of accuracy, recall rate and F1 value indicators of different models

从图 4 中可以看出，从精度、召回率以 F1 值的指标出发，对比模型的在三个指标上的均值分别 79.30%，79.04% 以及 79.54%，Bert-BIIAM 均优于诸如 LSTM+CRF、Bert_BiLSTMCRF 等模型。在精度方面，Bert -BIIAM 模型能够更精准地对修井知识进行判别，减少误判情况。召回率上，该模型对修井知识实体的捕捉能力更强，遗漏更少。F1 值则体现了其在精度和召回率综合性能上的优势，这充分表明 Bert-

BIIAM 模型在修井知识的识别能力上更为出色，在修井知识实体识别等相关任务中具有更高的可靠性，能够为修井相关工作提供更准确的数据支持。

更高的精度和 F1 值指标，意味着模型在修井文本的知识识别提取的漏检误检的情况越少、较对比模型更适合修井文本知识的提取任务。例如在 LSTM+CRF 模型中，“软探砂面值明油二组_3”作为指标的修井知识实体可能未被模型正确识别，同样，在 BertLinear 模型中，“完井”未作为措施类知识实体也可能被模型漏检。从前文可知虽然如 LSTM+CRF、BertLinear 等对比模型的模型性能较 Bert-BIIAM 的差距不大，但是在实际的修井知识识别过程中，对知识实体的漏检及误检现象却比 Bert-BIIAM 模型更为明显。

2.1.2 Bert-BIIAM 模型精度及召回率提高原因

从 Bert-BIIAM 模型的实验能够观察到，Bert-BIIAM 模型在精度与召回率方面相对于其他传统模型分别展现出约 3% 和 1% 的提升。这一提升表明 Bert-BIIAM 模型在修井知识挖掘领域具备更优的适配性。对其精度、召回率以及 F1 值提升的原因进行深入剖析，主要可归纳为以下几个关键方面：

(1) Bert-BIIAM 模型充分发挥了 Bert 预训练语言模型所具备的强大语义表示能力。Bert 在大规模语料库上进行预训练的过程中，运用了先进的神经网络架构与训练算法，能够深度挖掘并捕捉词汇之间复杂而深层次的语义关联信息。这种在大规模数据中学习到的语义知识，为其在下游修井知识挖掘任务中的应用提供了极为丰富且精准的上下文信息基础。当模型处理修井相关文本时，能够依据这些丰富的语义信息，更敏锐地感知和理解文本中各种概念、实体以及它们之间的相互关系，从而极大地提高了对文本进行分类处理的准确性[42]。例如，在识别修井方案中的技术措施与相关物质的关联时，Bert 预训练所获得的语义理解能力能够帮助模型更精准地判断其关系类型，避免误判。

(2) BERT 模型所采用的双向 Transformer 结构是其性能优势的重要保障。与传统的单向处理方式 (如仅从左到右或从右到左) 不同，双向 Transformer 结构能够同时兼顾文本的前后上下文信息[43]。在修井知识文本中，往往存在诸多前后呼应、相互关联的信息表述，例如对某种地质现象的描述可能与后续的修井措施存在因果关联，双向编码方式能够全面地整合这些信息，使得模型在处理复杂语义关系时拥有更强的判别力。Bert-BIIAM 模型通过多层的 Transformer 结构，能够稳定且高效地处理长期依赖关系，深入挖掘文本中的复杂信息脉络，为精准分类提供有力支持

(3) Bert 模块的输出传递给 BiLSTM 模块。BiLSTM 利用其双向结构，以前向和后向两个方向对 Bert 生成的词向量序列进行处理。Bert 提供的语义丰富的向量使得 BiLSTM 能够更有效地提取文本的上下文信息。BiLSTM 能够在 Bert 初步编码的基础上，进一步挖掘文本的语义连贯性和逻辑性。例如，当处理包含因果关系的修井知识文本时，BiLSTM 可以利用 Bert 提供的词汇语义，更好地捕捉到前后文之间的因果关联，这种关联信息的挖掘有助于更准确地理解文本的整体意图[44]，提高 Bert-BIIAM 模型对文本语义理解的深度和准确性。

(4) IAMILW 机制在模型中发挥了独特的作用。通过对输入的标签进行加权，并引入多头注意力机制，进一步优化了模型的注意力分配策略。在修井知识挖掘中，不同的实体标签具有不同的重要性和信息价值，IAMILW 机制能够根据这些差异对标签进行合理加权，使得模型在训练过程中更加聚焦于关键实体的特征学习与识别分类。多头注意力机制则允许模型从多个角度同时关注输入信息，如同多个“视角”协同工作，能够捕捉到更多的细粒度信息，从而在面对复杂多样的修井知识输入时，具备更强的区分能力，有效提升了分类精度，为修井知识挖掘任务提供了更为可靠和高效的模型解决方案。

综上所述，Bert-BIIAM 模型在精度、召回率以及 F1 值方面的提升是多个关键因素协同作用的结果。这种多模块的有机结合与优势互补，使得 Bert-BIIAM 模型在修井知识挖掘领域展现出相较于传统模型更为出色的性能表现，为该领域的知识挖掘工作提供了更高效、更精准的工具与方法，在未来的相关研究与实践应用中有着广阔的发展前景与巨大的应用潜力，有望推动修井知识挖掘技术向更高水平发展，助力修井工程相

关决策与操作更加科学、合理、高效。

2.2 关联规则计算效率讨论

BH-Apriori 算法可过动态调整挖掘频繁项集中的最小支持度以及最小置信度来全面挖掘存在于修井实体之间的关联关系。BH-Apriori 算法在对修井知识的挖掘过程中，最小支持度和最小置信度曲线整体呈递减趋势，变化区间为 0.6 至 0.8 之间，这使 BH-Apriori 算法所引入的自适应最小支持度和置信度可根据 N 项频繁项集情况灵活变动，在确保挖掘规则质量的同时，也让在挖掘多种类别的关联规则时更具准度，有助于提升挖掘多种修井知识关联规则时的准确性，能够更精准地挖掘出修井知识中不同类别间的潜在联系与规律，为修井工作的深入分析与优化提供有力的数据支持。

通过对修井施工方案设计报告中的识别出的共计 800 个实体之间的关联规则提取，可快速理解在修井作业中不同作业操作、工序以及指标之间的潜在关系，同时帮助修井人员在后续的修井施工中更加智能化决策。BH-Apriori 算法在性能上相较于传统 Apriori 算法有较大提升，以挖掘 2 项-频繁项集为例，二者在算法时效性、频繁项集以及强关联组合的挖掘条数上的对比，如图 5 所示。

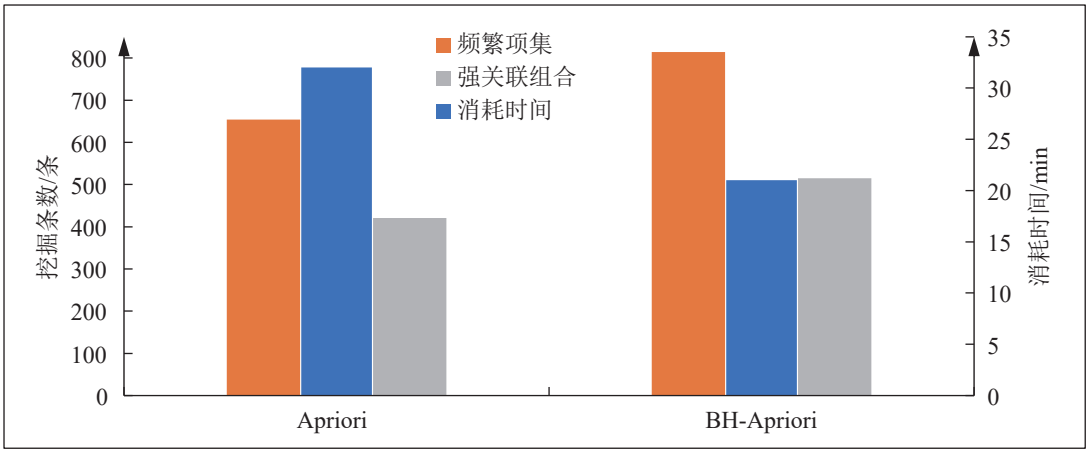

图 5 算法性能对比

Fig. 5 Comparison of Algorithm Performance

从图 5 中可以看出，传统的 Apriori 算法执行此任务需要花费时间共计 32 min，而 BH-Apriori 算法仅需要 21 min，计算效率提升了 0.5 倍。同时，在修井施工方案设计报告中有关修井知识实体的关联规则挖掘的数量上，BH-Apriori 算法挖掘的频繁项集数量为 814 条，具有强关联实体组合的数量为 515 条，而传统 Apriori 算法挖掘的频繁项集和具有强关联实体组合数量分别为 654 和 421 条。综上，BH-Apriori 算法无论在算法效率以及挖掘数量相较于传统的 Apriori 算法有较大提升，可以胜任在修井知识挖掘领域内的关联规则分析，为高效理解修井施工方案文本提供行之有效的途径。

2.2.1 BH-Apriori 算法时效提升原因

相较于传统的 Apriori 算法，BH-Apriori 算法在挖掘频繁项集上所需时间降低 0.5 倍，表明 BH-Apriori 在时效上改进的成功。BH-Apriori 算法在用时上的减少，归功于 Hash 树的引入。对于首次的频繁项集挖掘，BH-Apriori 算法会将符合最小支持度的候选项集映射到散列表中，下次挖掘时，BH-Apriori 算法会直接从散列表中找寻数据，而不是再次扫描整个数据库，从而降低了算法在时间上的消耗从而提高算法效率。

2.2.2 关联规则挖掘条数提高的原因

传统 Apriori 算法在频繁项集的认定中，是以当前频繁项集的支持度、置信度与算法的最小支持度、最小置信度的大小为依据，由此静态的最小支持度以及最小置信度容易造成频繁项挖掘不充分，特别是在修井

施工方案中出现次数相对较少，但事实上已然专家较为认可的知识。因此提出了最小支持度和最小置信度动态变化的 BH-Apriori 算法。

BH-Apriori 算法使用动态变化的最小支持度以及最小置信度作为频繁项集筛选的指标，使得对修井方案中的频繁项集挖掘更加充分、全面，避免了传统 Apriori 在此方面的不足。通过这种动态调整机制，能更好地适应修井数据的多样性与复杂性，捕捉到那些被传统算法忽视的潜在关联。它能够在不同规模和特征的数据集中灵活运作，不仅提升了挖掘效率，还为修井方案的精细化制定提供了更丰富、准确的依据，有力推动修井工程领域的数据分析与决策优化进程。

根据实体组合间的关联规则，提取不同实体类型组合间的前 60%，可以直观理解不同实体组合间的关联规则。以组合 (措施，物质)(措施，现象)(措施，指标)(物质，现象)(措施，物质，指标)(物质，指标，现象) 为例，(措施，物质) 表明在修井施工作业中，不同的修井措施皆与物质类中清水之间存在强关联关系，在这些强关联关系中，可以总结出以下知识：

例如在反压井、正冲砂等操作过程中，清水被十分广泛的添加，在其他诸如洗井、固沙等操作中，除清水外，碱水、卤水、暂堵剂等会被添加以实现特定目的；(措施，现象) 中展示了不同修井措施与施工过程中的现象之间的关联关系，例如抽油杆、扶正防磨块的布设位置直接与井筒偏磨位置存在强相关关系；陶瓷凡尔球座的使用，往往代表该井存在较为严重井筒腐蚀、结垢、磁化现象；(措施，指标) 展示了措施与指标之间的关联关系。其中，“软探砂面明油二组_3”、“软探砂面明油二组_2”，“硬探沙面至明油二组_3”三个指标在整个指标中占比最多，表明在压井、洗井等修井操作中有关高度的指标，这三者占比最多，即整个施工过程多位于这些高度范围内。(物质，现象) 展示了物质和现象之间的关联规则。其中物质清水与各种现象之间均存在关联关系，特别是无漏失，遇阻、合格等现象之间的关联规则较强，因此容易得出清水多与措施的成功与否存在关系，压井、洗井等措施中若仅使用清水物质表明这些措施很大概率不会出现异常现象。

同时，从现象的角度出发，当出现出口反油气水、漏失、遇阻等异常现象时，可作为解决异常现象措施的依据。例如，当出现遇阻的异常现象时，可添加的物质有投球和清水。(措施，物质，指标)，(物质，指标，现象) 分别展示了措施-物质-指标和物质-指标-现象之间的关联规则，容易看出集合 (下完井，清水，软探砂面值明油二组_3)，(反压井，清水，下捞砂工具及油管硬探砂面明油二组_3) 和 (正冲砂，卤水，硬探砂面至明油二组_3) 等在整个强关联集合中占比较高，因此，在这些强关联集中可总结出以下知识，例如在反压井操作中，多在高度范围为明油二组，所添加的物质为清水和卤水。同时，在物质-指标-现象的强关联集合中，(清水、软探砂面明油二组_3，无漏失) 和 (卤水，硬探砂面至明油二组_3，不漏) 的占比较多，可总结出以下知识，例如在操作“软探砂面明油二组_3”中，所添加的物质多为卤水与清水，出现的现象多为“无漏失”。

通过不同类别实体之间的关联规则分析，可以直观的总结出具有强关联关系的频繁项集，从而获得出可靠的修井知识。例如在知识“修井过程中，清水被广泛添加，随后卤水、前置液使用率靠后，且多用于压井、完井、洗井等操作中”中，揭示了材料在修井施工中的重要地位，因此可以快速理解修井施工设计方案文本内容，便于优化在修井施工中不同施工内容的操作流程，为后续修井施工提供更加智能的操作指导，促进油气开发的智能化发展。

3 结论建议

(1) 根据施工方案文本中的修井知识，提取操作、物质、指标和现象四类实体知识类型，提出了融合标签注意力机制的 Bert-BIIAM 模型，构建修井实体知识识别模型 Bert-BIIAM 从而识别施工方案文本中的修井知识。经过对具体的施工设计方案文本进行实体识别，Bert-BIIAM 模型平均精度和平均 F1 值达到 81.83% 和 81.23%，验证了模型的可靠性与适用性。在修井实体知识的关联规则挖掘上，针对传统关联规则挖掘算法在时效性和精确性不足的基础上，提出了引入贝叶斯方法和 Hash 树的 BH-Apriori 算法。基于大港

油田提供的修井施工文本数据的实验证明，BH-Apriori 算法无论是在时效性和精确度上相较于传统的 Apriori 算法均有提升，从而验证了 BH-Apriori 算法在修井知识的关联规则挖掘工作中的优势，表明结合深度学习和关联规则分析的两阶段修井知识智能分析与挖掘方法，符合对修井知识高效挖掘的预期，加深了管理人员对修井施工过程的理解效率，对提高修井施工质量具有重要意义。

(2) 提出的修井知识智能分析与挖掘方法在实体识别和关联规则挖掘方面取得了较好的效果，但仍存在一些不足。首先，Bert-BIIAM 模型在处理复杂、模糊或不规范的文本时可能存在识别准确性不足的问题，且训练数据中的实体类别不均衡可能影响低频实体的识别效果。其次，改进的 Apriori 算法虽然提高了效率，但仍可能受到噪声数据的干扰，且规则的数量过多可能导致挖掘出大量冗余或无意义的规则。

(3) 未来可以尝试结合多模态数据提升实体识别的准确性，并利用数据增强技术平衡数据集；同时，可采用更精细的过滤和优化策略来提升关联规则挖掘的精度，并在算法中引入更强的噪声处理能力，从而提高整体智能化水平。

致谢

感谢中国石油天然气股份有限公司大港油田分公司为本篇论文提供的修井施工文本数据以及实验场所，使得实验工作得以顺利开展，推动本文的研究结果的实现。同时，感谢“海上稠油超临界多源多元热流体发生机理及在储层中的作用机制研究项目”的支持。

稿件申明

1. 论文所有作者都同意文章中的作者排名，不存在争议，对文章中的学术观点一致，对文章中涉及的结论建议意见一致。

2. 文章涉及的内容不存在学术不端的问题。

署名贡献声明

杨希军：提出了文章的创新点，设计了融合深度学习与关联规则挖掘的修井知识挖掘方法，设计了 Bert-BIIAM 与 BH-Apriori 算法，主导实验的开展进行并撰写论文。

孔红芳：负责修井知识实体提取模型 (Bert-BIIAM) 与实现，优化了基于标签权重的改进注意力机制 (IAMILW)，在模型训练与评估过程中提供了重要的理论支持。

赵东：主导了关联规则挖掘阶段的算法设计，实现了融合贝叶斯方法与 Hash 树的改进 Apriori 算法 (BH-Apriori)，并负责相关算法的调优与测试。

易春飚：负责大港油田修井文本数据的收集与处理，参与了实验的实施与结果分析，为验证算法效果提供了实际数据支持。

于国起：参与了研究的初步构思与技术路线设计，协助进行实验结果的分析。

利益冲突说明

1. 论文不涉及企业技术商业秘密。

2. 专家审稿时没有需要避讳的问题。

数据可用性声明

文章所运用的数据系源自大港油田港东港西地区特定的 20 口井的历次修井作业记录，涉及油田生产作业的诸多关键细节、技术参数以及地质信息等。在未获相关授权许可的情形下，无法向外部予以提供，以保障油田数据安全。

参考文献

[1] 王金龙, 盛磊祥, 李婷婷, 等. 深水中型修井系统设计与整体强度分析［J］. 石油机械, 2024, 52(3): 61-66. //WANG Jinlong, SHENG Leixiang, LI Tingting, et al. Design and overall strength analysis of a deepwater medium-sized workover sys-

tem［J］. China Petroleum Machinery, 2024, 52(3): 61-66.

［2］ 蔡萌. 大庆油田采油工程主体技术现状及展望［J］. 石油钻采工艺, 2022, 44(5): 546-555. //CAI Meng. Current situation and prospect of main technology of oil production engineering in Daqing Oilfield［J］. Oil Drilling & Production Technology, 2022, 44(5): 546-555.

［3］ 雷群, 李益良, 李涛, 等. 中国石油修井作业技术现状及发展方向［J］. 石油勘探与开发, 2020, 47(1): 155-162. //LEI Qun, LI Yiliang, LI Tao, et al. Technical status and development direction of workover operation of PetroChina［J］. Petroleum Exploration and Development, 2020, 47(1): 155-162.

［4］ 邓正强, 兰太华, 林阳升, 等. 川渝地区防漏堵漏智能辅助决策平台研究与应用［J］. 石油钻采工艺, 2021, 43(4): 461-466. //DENG Zhengqiang, LAN Taihua, LIN Yangsheng, et al. Research and application of intelligent assistant decision making platform of lost circulation prevention and control in Sichuan-Chongqing area［J］. Oil Drilling & Production Technology, 2021, 43(4): 461-466.

［5］ 窦宏恩, 张蕾, 米兰, 等. 人工智能在全球油气工业领域的应用现状与前景展望［J］. 石油钻采工艺, 2021, 43(4): 405-419+441. //DOU Hongen , ZhANG Lei , MI Lan, et al. The application status and prospect of artificial intelligence in the global oil and gas industry［J］. Oil Drilling & Production Technology, 2021, 43(4): 405-419+441.

［6］ 孙涛, 孟祥娟, 王静. 注水水质对裂缝性油藏储层的影响［J］. 石油钻采工艺, 2022, 44(2): 199-203,210. //SUN Tao, MENG Xiangjuan, WANG Jing. Effects of injected water quality on fractured oil reservoirs［J］. Oil Drilling & Production Technology, 2022, 44(2): 199-203,210.

［7］ 耿黎东. 大数据技术在石油工程中的应用现状与发展建议［J］. 石油钻探技术, 2021, 49(2): 72-78. //GENG Lidong. Application status and development suggestions of big data technology in petroleum engineering［J］. Etroleum Drilling Techniques, 2021, 49(2): 72-78.

［8］ 杨丽丽. 新质生产力理念下中国油气高质量发展战略思考［J］. 中国矿业, 2024, 33(5): 32-38. //YANG Lili. Study on high-quality development strategy of oil and gas industry in China under the concept of new quality productive forces［J］. CHINA MINING MAGAZINE, 2024, 33(5): 32-38.

［9］ 王仁超, 张毅伟, 毛三军. 水电工程施工安全隐患文本智能分类与知识挖掘［J］. 水力发电学报, 2022, 41(11): 96-106. //WANG Renchao, ZHANG Yiwei, MAO Sanjun. Intelligent text classification and knowledge mining of hidden safety hazards in hydropower engineering construction［J］. Journal of Hydroelectric Engineering, 2022, 41(11): 96-106.

［10］ 汪祺能, 宋立明, 郭振东, 等. 采用动态交互作用分析的叶轮机械优化算法研究［J］. 西安交通大学学报, 2023, 57(7): 139-150. //WANG Qineng, SONG Liming, GUO Zhendong, et al. Study on turbomachinery optimization algorithm based on dynamic interaction analysis［J］. Journal of Xi'an Jiaotong University, 2023, 57(7): 139-150.

［11］ 刘爽, 丁哲, 吕超, 等. 基于文本分类和知识挖掘的远洋渔船安全问题分析［J］. 农业工程学报, 2023, 39(24): 215-223. //LIU Shuang, DING Zhe, LYV Chao, et al. Evaluating the safety of distant-water fishing vessels using text classification and knowledge mining［J］. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2023, 39(24): 215-223.

［12］ 田丹, 沈扬, 李明超, 等. 混凝土坝施工文档实体知识智能挖掘方法［J］. 水力发电学报, 2021, 40(6): 139-151. //TIAN Dan, SHEN Yang, LI Mingchao, et al. Intelligent data mining approach of text entity knowledge from construction documents of concrete dams［J］. Journal of Hydroelectric Engineering, 2021, 40(6): 139-151.

［13］ KIM J C, CHUNG K. Associative feature information extraction using text mining from health big data［J］. Wireless Personal Communications, 2019, 105(2): 691-707.

［14］ 陈碧云, 丁晋, 陈绍南. 基于关联规则挖掘的电力生产安全事故事件关键诱因筛选［J］. 电力自动化设备, 2018, 38(4): 68-74. //CHEN Biyun, DING Jin, CHEN Shaonan. Selection of key incentives for power production safety accidents based on association rule mining［J］. Electric Power Automation Equipment, 2018, 38(4): 68-74.

［15］ 于溪芮, 何旭, 孔得朋. 基于 Apriori 算法的氢安全事故统计分析［J］. 中国安全科学学报, 2024, 34(4): 128-134. //YU Xirui, HE Xu, KONG Depeng. Apriori algorithm-based statistical analysis of Hydrogen accidents［J］. China Safety Science Journal, 2024, 34(4): 128-134.

［16］ 国汉君, 江益, 姚勇征. 基于内-外因理论和 Apriori 算法的动火作业事故分析［J］. 中国安全生产科学技术, 2024, 20(11): 101-109. //JIANG Yi, YAO Yongzheng. Analysis of hot work accidents based on internal-external causes and Apriori

algorithm［J］. Journal of Safety Science and Technology, 2024, 20(11): 101-109.

［17］ 钟仪华, 王淑宁, 罗兰, 等. 用深度学习挖掘油田开发指标预测模型的知识［J］. 西南石油大学学报 (自然科学版), 2020, 42(6): 63-74. //ZHONG Yihua, WANG Shuning, LUO Lan, et al. Knowledge mining for oilfield development index prediction model using deep learning［J］. Journal of Southwest Petroleum University(Science & Technology Edition), 2020, 42(6): 63-74.

［18］ 谢坤, 吴湛奇, 李彦阅, 等. 机器学习在油气开发领域的应用及展望［J］. 西安石油大学学报 (自然科学版), 2023, 38(5): 58-67. //XIE Kun, WU Zhanqi, LI Yanyue, et al. Application of machine learning in oil and gas development field and its prospect［J］. Journal of Xi'an Shiyou University(Natural Science Edition), 2023, 38(5): 58-67.

［19］ 王敏生, 光新军, 耿黎东. 人工智能在钻井工程中的应用现状与发展建议［J］. 石油钻采工艺, 2021, 43(4): 420-427. //WANG Minsheng, GUANG Xinjun, GENG Lidong. Application status and development suggestions of artificial intelligence in drilling engineering［J］. Oil Drilling & Production Technology, 2021, 43(4): 420-427.

［20］ 陈雪松, 朱鑫海, 王浩畅. 基于 PMV-LSTM 的中文医学命名实体识别［J］. 计算机工程与设计, 2022, 43(11): 3257-3263. //CHEN Xuesong, ZHU Xinhai, WANG Haochang. Chinese medical named entity recognition based on PMV-LSTM［J］. Computer Engineering and Design, 2022, 43(11): 3257-3263.

［21］ 任伟建, 计妍, 康朝海. 基于 XLBIC 的石油开采数据命名实体识别研究［J］. 计算机仿真, 2024, 41(6): 390-395. //REN Weijian , JI Yan , KANG Chaohai. Research on named entity recognition of petroleum exploitation data based on XLBIC［J］. Computer Simulation, 2024, 41(6): 390-395.

［22］ 钟原, 刘小溶, 王杰, 等. 基于 NER 的石油非结构化信息抽取研究［J］. 西南石油大学学报 (自然科学版), 2020, 42(6): 165-173. //ZHONG Yuan, LIU Xiaorong, WANG Jie, et al. Research of extraction on petroleum unstructured information based on Named Entity Recognition［J］. Journal of Southwest Petroleum University(Science & Technology Edition), 2020, 42(6): 165-173.

［23］ 肖立志. 机器学习数据驱动与机理模型融合及可解释性问题［J］. 石油物探, 2022, 61(2): 205-212. //XIAO Lizhi. The fusion of data-driven machine learning with mechanism models and interpretability issues［J］. Geophysical Prospecting for Petroleum, 2022, 61(2): 205-212.

［24］ 黄灿, 田冷, 王恒力, 等. 基于条件生成式对抗网络的油藏单井产量预测模型［J］. 计算物理, 2022, 39(4): 465-478. //HUANG Can , TIAN Leng , WANG Hengli , et al. A single well production forecasting model of reservoir based on conditional generative adversarial net［J］. Chinese Journal of Computational Physics, 2022, 39(4): 465-478.

［25］ DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding［C］//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, Minneapolis, Minnesota: 4171-4186.

［26］ 刘怀兰, 刘盛, 周源, 等. 基于多源文本挖掘的技术演化路径识别［J］. 情报理论与实践, 2022, 45(11): 178-187. //LIU Huailan, LIU Sheng, ZHOU Yuan, et al. Technology evolution path recognition based on multi-source text mining［J］. Information Studies:Theory & Application, 2022, 45(11): 178-187.

［27］ 孙凯丽, 罗旭东, 罗有容. 预训练语言模型的应用综述［J］. 计算机科学, 2023, 50(1): 176-184. //SUN Kaili, LUO Xudong, Michael Y.LUO. Survey of applications of pretrained language models［J］. Computer Science, 2023, 50(1): 176-184.

［28］ 罗仁泽, 周洋, 康丽侠. 基于 DMC-BiLSTM 的沉积微相智能识别方法［J］. 石油物探, 2022, 61(2): 253-261,338. //LUO Renze, ZHOU Yang, KANG Lixia. Intelligent identification of sedimentary microfacies based on DMC-BiLSTM［J］. Geophysical Prospecting For Petroleum, 2022, 61(2): 253-261,338.

［29］ 靳嵩, 朱艳, 吴可嘉, 等. 基于 BERT 的海上船舶安全隐患分类［J］. 船舶工程, 2023, 45(S1): 381-384. //JIN Song, ZHU Yan, WU Kejia, et al. Classification of ship safety hazards of Marine vessel based on BERT［J］. Ship Engineering, 2023, 45(S1): 381-384.

［30］ 王博乔, 张彬, 林叶锦. 基于 RAdam Bi-LSTM 的 LNG 动力船舶上甲板储罐泄漏后果预测方法［J］. 中国航海, 2023, 46(2): 60-66,73. //WANG Boqiao, ZHANG Bin, LIN Yejin, et al. Consequence prediction of deck tank leakage on LNG powered ship using RAdam Bi-LSTM［J］. China Shipping, 2023, 46(2): 60-66,73.

［31］ 叶月明, 曹晓初, 任浩然, 等. 应用自注意力机制对抗网络进行海洋多次波压制方法研究［J］. 石油地球物理勘探, 2024, 59(3): 454-464. //YE Yueming, CAO Xiaochu, REN Haoran, et al. Marine multiple attenuation method based on SA-

GAN［J］. Oil Geophysical Prospecting, 2024, 59(3): 454-464.

［32］ 文靖杰, 王勇, 李金龙, 等. 多头自注意力机制的 Faster R-CNN 目标检测算法［J］. 现代电子技术, 2024, 47(7): 8-16. //WEN Jingjie, WANG Yong, LI Jinlong, et al. Faster R-CNN object detection algorithm based on multi-head self-attention mechanism［J］. Modern Electronics Technique, 2024, 47(7): 8-16.

［33］ 郁荣荣, 赵飞, 李吉广, 等. 基于神经网络的渣油浆态床加氢产物分布预测模型［J］. 石油炼制与化工, 2023, 54(11): 86-95. //XI Rongrong , ZHAO Fei , LI Jiguang , et al. Prediction model of residue hydrocracking product distribution in slurry bed based on neural network［J］. Petroleum Processing and Petrochemicals, 2023, 54(11): 86-95.

［34］ WANG H B, GAO Y J. Research on parallelization of Apriori algorithm in association rule mining［J］. Procedia Computer Science, 2021, 183: 641-647.

［35］ 王兵, 黄丹, 李文璟. 基于支持度矩阵 Apriori 算法的钻井隐患关联挖掘［J］. 西南石油大学学报 (自然科学版), 2022, 44(2): 113-122. //WANG Bing, HUANG Dan, LI Wenjing. Correlation mining of hidden hazards in drilling based on support matrix apriori algorithm［J］. Journal of Southwest Petroleum University(Science & Technology Edition), 2022, 44(2): 113-122.

［36］ 高良军, 唐义新, 陈亮, 等. 原油船海上航行升沉运动 Bayes-LSTM 预测方法［J］. 油气储运, 2023, 42(11): 1291-1296. //GAO Liangjun, TANG Yixin, CHEN Liang, et al. Bayes-LSTM method for predicting heave movement of crude oil vessels during maritime navigation［J］. Oil & Gas Storage and Transportation, 2023, 42(11): 1291-1296.

［37］ MAR Z, OO K K. An improvement of Apriori mining algorithm using linked list based hash table［C］//2020 International Conference on Advanced Information Technologies(ICAIT), 2020, Yangon, Myanmar: 165-169.

［38］ 孙来平, 楚彭子, 虞翊, 等. 基于 Beta 分布和三角模糊函数的轨道交通信号系统故障检测［J］. 铁道科学与工程学报, 2023, 20(12): 4823-4834. //SUN Laiping, CHU Pengzi, YU Yi, et al. Fault propagation detection for signal system of rail Transit based on beta distribution and triangular ambiguity function［J］. Journal of Railway Science and Engineering, 2023, 20(12): 4823-4834.

［39］ 郭倩, 殷丽凤. 基于散列技术的多层关联规则算法的改进［J］. 计算机工程与设计, 2021, 42(9): 2485-2491. //GUO Qian , YIN Lifeng. Improvement of multi-level association rule algorithm based on hashing technology［J］. Computer Engineering and Design, 2021, 42(9): 2485-2491.

［40］ 郭素杰, 李景卫, 于伟高, 等. 基于知识驱动数据挖掘技术在复杂储层评价中的应用［J］. 石油钻采工艺, 2022, 44(2): 247-252. //Guo Sujie, Li Jingwei, Yu Weigao, et al. Application of knowledge-driven data mining in the complex reservoir evaluation［J］. Oil Drilling & Production Technology, 2022, 44(2): 247-252.

［41］ 姚爽, 徐佳美, 连向伟, 等. 融合额外实体信息的层级联合抽取［J］. 计算机工程与设计, 2024, 45(6): 1698-1704. //YAO Shuang, XU Jiamei, LIAN Xiangwei, et al. Hierarchical federated extraction of additional entity information［J］. Computer Engineering and Design, 2024, 45(6): 1698-1704.

［42］ 李瑜泽, 栾馨, 柯尊旺, 等. 知识感知的预训练语言模型综述［J］. 计算机工程, 2021, 47(9): 18-33. //LI Yuze, LUAN Xin, KE Zunwang, et al. Survey of knowledge-aware pre-trained language models［J］. ComputerEngineering, 2021, 47(9): 18-33.

［43］ Han K, Wang Y, Chen H, et al. A survey on vision transformer［J］. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

［44］ 何玉洁, 杜方, 史英杰, 等. 基于深度学习的命名实体识别研究综述［J］. 计算机工程与应用, 2021, 57(11): 21-36. //He Yujie, Du Fang, Shi Yingjie, et al. Survey of named entity recognition based on deep learning［J］. Computer Engineering and Applications, 2021, 57(11): 21-36.