arXiv:2302.08905v1 [cs.DL] 15 Feb 2023

GraphLED：一种用于处理和可视化关联工程文档的基于图的方法

Vanessa T. da Silva, Lucas de A. M. Ribeiro, Willian B.de Lemos, Sílvia S. da C. Botelho，**Nelson L. D. Filho**，**Marcelo R. Pias**Universidade Feral do Rio Grande（FURG）南里奥格兰德，RS，巴西{vanessa.telles,lucasribeiro，willianlemos.b，silviacb，dmtnldf，mpias}@furg.br

摘要

建筑、工程与施工（AEC）行业广泛使用支持产品与流程开发的文档。为此，各组织需要处理包含数百甚至数千份技术文档的大数据，这些文档彼此紧密关联，包括工业厂房的CAD设计、设备采购订单、质量证书和零部件材料分析。然而，用户要分析此类记录并非易事，因为要从数百份文档中梳理信息并建立有价值的关联十分复杂。本文探讨了从关联工程文档中提取的知识如何在IT/OT融合背景下推动工业数字化。本文提出的GraphLED是一套用于数据处理、基于图的建模以及相关文档彩色可视化的系统。基于图的方法有助于更好地理解关联信息，因为图结构为工程文档底层数据属性的建模提供了有前景的工具。初步系统验证表明，基于OCR的数据质量有望得到提升（消除了85.9%的模糊文本数据）。这项工作有望通过对大量文档进行自动摘要并建立文档间的关联，提高工业生产系统的可靠性和韧性，从而为行业带来裨益。

关键词 基于图的处理、文档处理、工业数字化转型、工业IO、图大数据

1 引言

数据的创建无处不在，且增长迅速；大数据采集为多个行业带来了切实的数字化转型效益 [1]。然而，在包括建筑、工程和施工（AEC）在内的传统行业中，数据仍以文档为中心，形式兼有模拟格式和数字格式。组织通常需要处理数百甚至数千份技术文档，例如工业厂房的CAD设计、舰队设计、设备采购订单、质量证书和材料分析 [2]。

借助高效工具管理技术文档，有助于在AEC流程生命周期内高效协调关键的日常活动 [3]。为此，持续的数据处理与可视化可为AEC终端用户带来显著益处：(i) 支持及时检查工程项目；(ii) 根据现有标准和行业协会指南开发自动化文档合规模型；(iii) 促进关键利益相关方（供应商、客户、标准协会）之间的信息交流。然而，建筑、工程与施工文档本质上十分复杂，因为各种非结构化主题通常会关联多份文档来传递信息。处理结构化数据的传统先进方法（例如在关系数据库中按关键词搜索）在这种情况下并不可行 [4]。

表格之间的关系是预先定义好的，因此用户只能获取有限的相关知识 [5]。纸质文档则使问题更加复杂，扫描手写报告、技术规格、工程图纸、示意图、印章、签名等 [6] 都让人难以高效检索信息，即使是最简单的信息也不例外。例如，数据手册——由用于说明项目事件的工程文档组成的集合——包含许多数据（即主题），这些数据与其他数据手册中的数据紧密关联。这种互联

由此形成了一个复杂而强大的相关工程文档网络，并具备多项功能。例如，最终用户可以直观地检查设备部件的材料可追溯性，以及其是否符合现行技术规格（例如 ASTM——美国材料与试验协会）。如今，如果没有大数据工具的辅助，用户在可视化和检查大型文档数据集时会面临困难，也容易出错。

在此背景下，一个关键的高层次问题是：如何利用从经过含噪光学字符识别处理的工程文档中提取的知识，即工程文档摘要及其关系，来促进工业数字化和生产系统的韧性？

知识提取预计需要高效完成，并尽可能接近实时。在这一过程中，同样重要的是利用文档之间的交叉引用数据来获取新信息。由于文档网络层次深、范围广，这项任务对人类而言本质上十分困难 [7]。

本文对此问题进行了探讨。本文提出的方法采用一种简单而有前景的机制，从各种工程数据手册中提取并组织数据，以便进行处理、知识提取和可视化。GraphLED 是一个负责数据处理、基于图的建模以及相关文档可视化的系统。基于图的方法有助于更好地理解关联信息，因为图结构提供了一种直接表示和建模底层数据属性的工具 [8]。
该系统假设能够以适中的正确率近乎实时地从扫描文档中提取知识，尤其是交叉引用数据。例如，光学字符识别处理产生的歧义数据会引入错误，因此应当妥善处理。

本文结构如下：第 2 节讨论相关工作。第 3 节探讨用户需求和 GraphLED 的设计选择。第 4 节介绍系统架构。最后，第 5 节和第 6 节展示验证结果并总结本文。

2 相关工作

工程数据手册是复杂的文档。用户对知识提取所得数据正确性的可接受程度仍是一项尚待解决的挑战，尤其是考虑到现有 OCR工具（如 Tesseract、Google OCR）的输出质量。问题不仅在于 OCR工具，扫描质量低劣的文档在业内仍十分常见。传统的结构化数据处理方法在这种情况下不够稳健，因为文档表格单元格中的噪声数据会导致相关知识有限 [4] [5]。
多种应用采用基于图的分析来提取知识并将结果可视化。例如， [9] 旨在提高公众对联合国成员投票模式的认识。该方法利用社会知识图谱和相似性度量，分析公开的成员投票信息。由于图适合以直观的方式进行可视化，这项工作从数据中探索了可视化辅助工具。然而，这种方法对数据噪声的鲁棒性较差，因为数据的采集、处理和管理阶段都要求事先准备好经过预处理、可直接使用的数据集。相互关联的建筑、工程与施工文档之间建立了各种联系（例如采购订单号、质量证书编号），可通过自动化软件工具 [10]追踪和分析这些联系。

基于图的分析探索了图数据结构上的算法 [11]。例如，基于图的文档表示系统（CGDR）可生成用于文档取证分析的分类报告 [12]。这项研究使用了总计1500份文件，对16类报告进行分类。该方法将信息组织为图节点，并采用层次文本分类来确定阅读顺序。然而，所提出的模型并非理想的解决方案，因为它仍依赖人类专家为每类报告识别有用的特征。
更加自动化且具备自适应能力的方法能够在许多方面促进该应用。此外，为使系统正常运行，数据应在高级处理阶段输入（例如经过增强和清理）。 [13] 提出了一种用于文档归纳式文本分类的基于图的方法。

该系统为输入的每份文档构建独立的图，并将其输入图神经网络（GNN），基于词语的局部结构学习细粒度的词表示。所提出的方法能够有效地为新文档中的未见词生成嵌入。然而，噪声数据限制了所提方法的稳健性。 [14] 引入知识图谱和本体，以应对建筑施工文档中复杂的决策网络。其他相关工作则探讨了工程图纸自动分析中的方法和算法，这些图纸通常用于石油和天然气工程行业。

[15] 中的作者承认，文档质量是问题的核心，因为除非采用一套稳健的预处理方法，否则质量低劣的图纸（例如，多次扫描过的纸质文档）几乎不可能实现数字化。

所提出的工作采用深度神经网络解决了这一问题，并取得了令人满意的改进。然而，该方法仍以模型为中心（即调优模型超参数），很少致力于通过预处理、清理、增强或重塑来提升原始数据质量。现有工作在很大程度上探索了基于 OCR（光学字符识别）的技术对文档进行数字化 [16] 和 [17]。以揭示知识提取中高度相关模式为目的，文档之间链接的益处在以往工作中尚未得到充分探索。

现有技术承认，要在扫描文档中实现近实时知识提取，尤其是交叉引用数据的提取，达到令人满意的数据正确性水平是一项挑战。数据噪声会沿着知识提取处理流水线传播，直至可视化和用户参与环节。

以往工作主要通过需要大量人工投入的校正 [4] [5] 或较新的以模型为中心的深度神经网络 [15], 来解决正确性问题；后一种方法的重点是微调模型超参数，而非提升输入数据质量。本文从以下方面推动了该领域的研究：

- 提出并验证了针对含噪光学字符识别扫描文档的数据清理和增强技术（即以数据为中心的方法），以克服所谓的垃圾进，垃圾出 错误传播问题。
- 提出了一种系统架构，旨在提供可大规模应用的基于图的工具，以近乎实时地提取和表示知识，帮助用户完成从数百到数千份文档中筛选信息这一艰巨任务。该系统验证重点关注性能和可扩展性问题。

3 系统设计选择

所提出的工作是一个更大型技术驱动项目的一部分，该项目旨在将按数据手册单位归类的工业文档（基于光学字符识别）数字化。这项工作实现了为石油和天然气行业量身打造数字孪生的愿景，使负责检查材料、设备、施工过程和质量控制相关文档的用户受益，这也是IT/OT融合趋势的一部分。2021年，项目与建筑、工程和施工领域的文档检查员举行了一系列会议和研讨会，由此整理出以下用户需求清单：

- 能够通过基于光学字符识别的系统上传大量扫描版数据手册，并从中提取关键信息。预期的正确性水平是知识提取数据中的歧义最小化。
- 能够通过直观易用的界面可视化、操作和分析数据手册。
- 以自动化方式执行常规检查任务（完整性和符合性）。
- 确认从数据手册中提取的特定信息是否符合现有技术规范和标准（例如 ASTM B16）。
- 了解产品和流程的制造地点及方式（可追溯性、异常检测）。

3.1 数据手册

感兴趣的数据空间包含会引用其他文档的技术文档。例如，设备采购订单文档可能与补充文档相关联，这些补充文档描述了所用材料的化学性质和物理性质。

另一个例子是用于记录海上石油钻井平台和平台施工过程的信息集合。在此情境下，数据手册是一组相关文档，涵盖特定产品或工艺周期的相关方面，包括材料生产、批量生产、产品符合性验证和采购订单描述。

例如，高压燃气管道所用螺栓的采购情况会记录在数据手册中。一系列文档详细说明了该螺栓部件的关键方面，包括制造工艺、材料性质、超声波检测，以及外观和尺寸特性。

3.2 设计选择

系统应检查数据手册的完整性及其是否符合相关标准，核查数据手册可追溯性信息，并在出现数据不一致或潜在欺诈时进行异常检测。非功能性要求包括低延迟运行、实时性能、数据一致性和完整性。

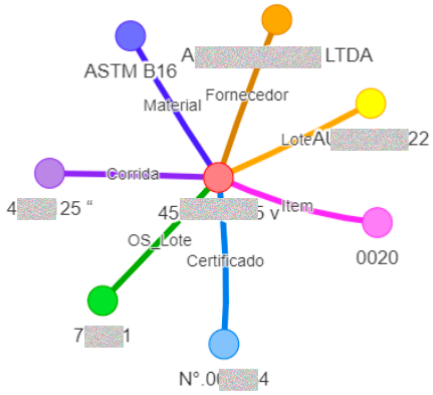

图 1：完整图：通过主题关联的文档（例如，批次号 OS_LOTE）

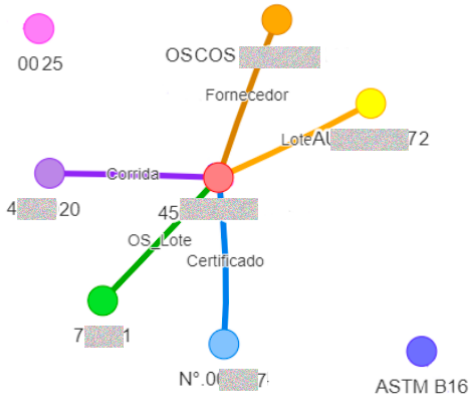

图 2：不完整图：存在缺失关联的数据手册。

3.2.1 完整性与符合性

检查完整性包括核实数据手册中是否随时可获取所有必需文件。此外，还要核查是否符合标准和指南，并检查多项合规规则，例如材料性能阈值、尺寸规格和目视检查。

3.2.2 可追溯性

此系统功能可确保产品或流程具有相关记录，以证明其从起始阶段开始的制造路径。可追溯性检查会搜索数据集中引用的所有数据手册文档，以分析关联数据与源数据的兼容性。此外，在可追溯性验证中，也应将不完整的数据手册或不符合规范的文档标记为故障。

3.2.3 数据一致性

由于文档需要经过扫描、光学字符识别（OCR）处理以提取相关文本，并以自动化方式进行结构化，因此数据不一致不可避免。例如，词语歧义可能源于语言变化或光学字符识别质量问题；当同一数据元素具有不同含义时，这会对所提出的系统构成挑战。在这种情况下，系统应不惜一切代价努力保持一致性和连贯性。

3.2.4 关联文档

基于图的结构和模型为指定技术文档中的数据元素及其关系提供了一种设计选择。此外，还可以探索多种图算法，以满足文档验证需求，包括

完整性、符合性和可追溯性。例如，中心性技术会遍历图，计算热门元素和瓶颈等属性。

基于图的数据表示在视觉上颇具吸引力，可为文档检查人员提供更直观的数据探索方式。图 1 和图 2 展示了一个数据手册样本的示例，该样本以图的形式呈现，取自一个真实世界的数据集（石油和天然气行业，标识符已被涂黑）。在此例中，图 $G(V, E)$ 由一组顶点 V（数据手册中的各份文档）和一组边 E 构成，这些边将顶点相互连接。图 1 展示了一个包含完整数据手册的图示例，因为存在一条路径，通过中心节点连接所有文档。相比之下，图 2 展示了一个不完整图，其中部分文档缺失，另有两份文档彼此孤立，因而损害了数据手册内可追溯性。需要注意的是，这些文档通过特定关系（例如属于同一质量保证实验室）与其他文档相关联。

4 GraphLED 系统

GraphLED系统架构（如图3所示）采用模块化设计，以确保从数据插入、知识提取到图可视化和处理等流水线功能在实现时具备互操作性。

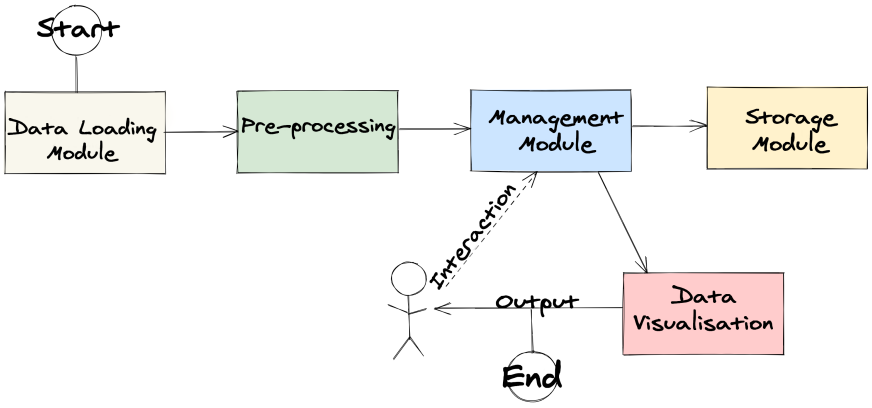

图3：系统架构模块。GraphLED采用鲜艳的色彩进行生动的可视化呈现（即以LED为喻）。

4.1 数据加载模块

光学字符识别工具（例如 Tesseract、Google图像阅读器）可在达到一定质量要求的前提下，将扫描文档转换为原始文本。为尽量减少与光学字符识别相关的错误，GraphLED采用了 [18], 中提出的步骤，对生成的原始文本数据进行预处理，包括噪声消除、二值化和倾斜校正。
输出文档会被送入表单理解流程，该流程使用预先定义的文档模型的基于表单的布局相似性方案对文档进行分类。该模块的输出是一个JSON文件，其中每份文档都表示为一组文本块，包含提取出的词语及其边界框，并指出是否存在指向另一份文档中某项数据的链接。

4.2 预处理

在此应用场景中，数据本身存在噪声。本模块负责数据清理和整理，通过转换输入数据集 来提升数据质量，使其成为适用于图生成的中间表示。预处理任务分为粗粒度（文档）和细粒度（文档词语）两个层级。挑战总结如下：（1）数据歧义：数据的表示方式不同，但语义上表示相同的信息；（2）无关数据：对基于图的分析内在价值较低的数据；（3）过度相关数据：在图生成阶段识别更有可能成为节点或边的数据。

对数据执行一系列操作，以降低噪声水平。例如，在使用消歧函数之前，必须先过滤掉停用词。消歧通过比较数据实例之间的相似度来实现，使用以下筛选器：

• Levenshtein 筛选器 [19] 依赖于比较长度相近（比例相同）的两个字符串之间的编辑距离，以尽量减少误报。此外，还会使用较低的接受阈值来检测非常相似的实例。
• 最长公共子序列（LCS）筛选器 [20] 通过比较任意两个字符串的最长公共子序列（保持完全相同的相对顺序）来检测相似度。
• 序列匹配器筛选器（LCS 的一种变体）使用不含干扰项的连续子序列作为相似度度量。

图 4 展示了歧义校正子模块的处理流程，其中各筛选器依次应用，以涵盖数据集中发现的主要歧义形式。另一个模块会将筛选步骤中所做的修改映射回原始文档中的相应实例，以便进行数据溯源。跟踪输入数据、转换后的数据及处理过程至关重要，这样才能保留数据及其来源的历史记录。

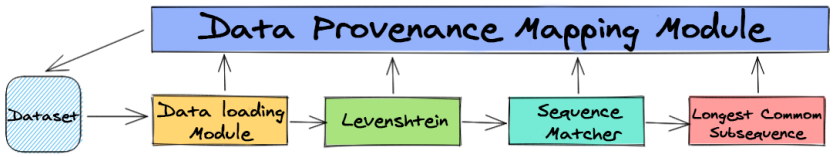

图 4：拟议的歧义模块架构。

面对海量数据，有效利用其中的有用信息可能令人望而却步。尽管如此，在通过筛选数据减少噪声后，基于图的算法便可尝试识别模式，并根据过往数据行为检测相关信息。相关数据节点的检测利用图中心性算法来识别趋势。这些算法可用于监测一组指标：(a) 节点的传递影响力（特征向量中心性）；(b) 节点对图中流动的影响力（介数中心性）；以及 (c) 节点的入向和出向关系数量（度中心性）。

用于 定义相关特征的标准，是上述指标所产生分数的总和 e.

4.3 管理模块

该模块负责系统集成及其正常运行。它通过 RESTful API（图 5）交换信息，即*graph-api*，该接口允许创建查询，并考虑通过 HTTP 请求提供的用户参数。

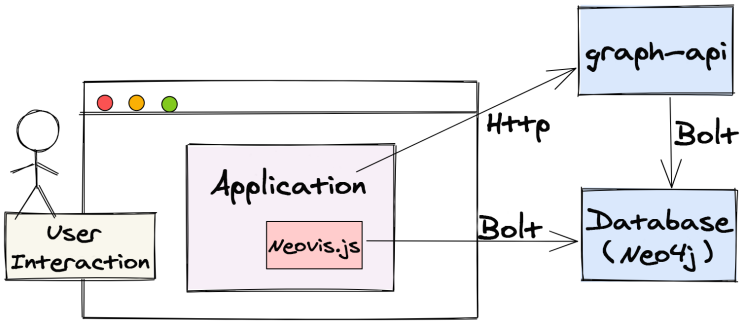

图 5：管理模块。数据通过 Bolt 协议与 Neo4j 图数据库交换

4.4 存储模块

基于图的存储和查询引擎是 GraphLED 系统中的核心功能。选择合适的图数据库系统时，采用了以下标准：

• 开发团队和社区所提供的文档与支持的成熟度和可用性 y.

- 易于与其他现有工具集成，并提供编程语言集成支持（例如 Python、Java 和 JavaScript）。
- 在中等数据量下表现令人满意。
- 易于使用、易于理解且便于维护的图查询语言。
- 无模式：在构建图时具备灵活性和敏捷性。

与传统数据库（类似 SQL）不同，面向图的数据库具有一些优势，但其中最主要的一项是对广度优先搜索进行优化；广度优先搜索是查找实体之间非显而易见关系的一种高度推荐的方法。我们考虑了几种选择，包括基于云的 Amazon Neptune、Azure CosmoDB 以及社区版 Neo4j。后者当时满足了项目要求。此外，即将推出的查询语言标准（Apache Gremlin 和 SparkQL）很快就能使 GraphLED 系统实现图数据库无关。

4.5 数据可视化

图天然适用于直观的数据可视化。本模块使用开源工具包，尤其是 Neovis 库，来呈现用于表示数据手册、相关文档及其关系的图节点和边。此外，GraphLED 可视化工具直接与 Bolt 协议的图数据库交互，以提升性能。不过，已有研究指出，随着图节点数量增加，图可视化可能成为性能瓶颈。

设计的用户界面通过（*i*）简单搜索来缓解潜在瓶颈，这类搜索旨在进行节点-边-节点遍历，用户可以指定节点的类别。

系统还提供（*ii*）以*Cypher* 语言编写的高级查询，向用户返回复杂且详尽的结果。计算得出的属性（*iii*），例如度中心性，也会以表格形式显示在图形用户界面底部。

4.6 用户界面

系统测试版（图 6）在美观性和用户体验方面均有所提升。系统添加了定制的图形用户界面元素，用户可以从多种工程文档类型中进行选择，并创建、可视化和删除数据手册/文档图。

侧边菜单中新增了一个探索表单字段，用户可以在其中编写复杂的图遍历查询，并选择节点和连接类型。此外，还添加了一个简单的搜索框，用于补充节点中心性信息，用户可以选择以下属性：中介中心性、节点度、接近中心性和特征向量中心性。 中心性可为用户识别结构异常提供线索，例如图不完整、关系缺失等。鉴于数据手册中的文档均以图的形式建模，中心性属性有助于回答有关完整性、符合性和可追溯性的问题。

5 系统验证结果

5.1 数据质量

此次验证旨在了解并进一步提升输入数据的质量。

所实现的数据流水线会接收扫描文档，并将其送入光学字符识别（OCR）阶段。随后，对 OCR 输出的文件进行分类并存储。验证数据集包含 81 份 PDF 格式的扫描数据手册文件，单份文档总计 *1.43GB* 。需要注意的是，数据手册是一组用途各异的工程类文档，涵盖采购订单规格和材料测试证书等内容。

基准测试采用了两种文档类别，分别标记为容易处理的文档 和困难文档。这些标签表示 Tesseract 光学字符识别系统在处理扫描文档时的难度，评估因素包括页面倾斜、潦草笔迹以及复杂的内部元素（例如线条较少的表格、页数较少的文档）。

为了分析数据噪声过滤和消歧技术（预处理模块）的性能，我们在一小组标注数据上开展了进一步的验证实验。此外，图中识别出的许多数据实体被用作比较指标，因为它能够反映数据集中重复节点和歧义节点的数量。数据由石油平台的零部件供应商构成（该数据集共有226行）。专家对数据集进行了分析，发现共有17家不同的供应商（共226家记录）。然而，生成的图将每家供应商表示为一个节点，其中包含128家不同的供应商。

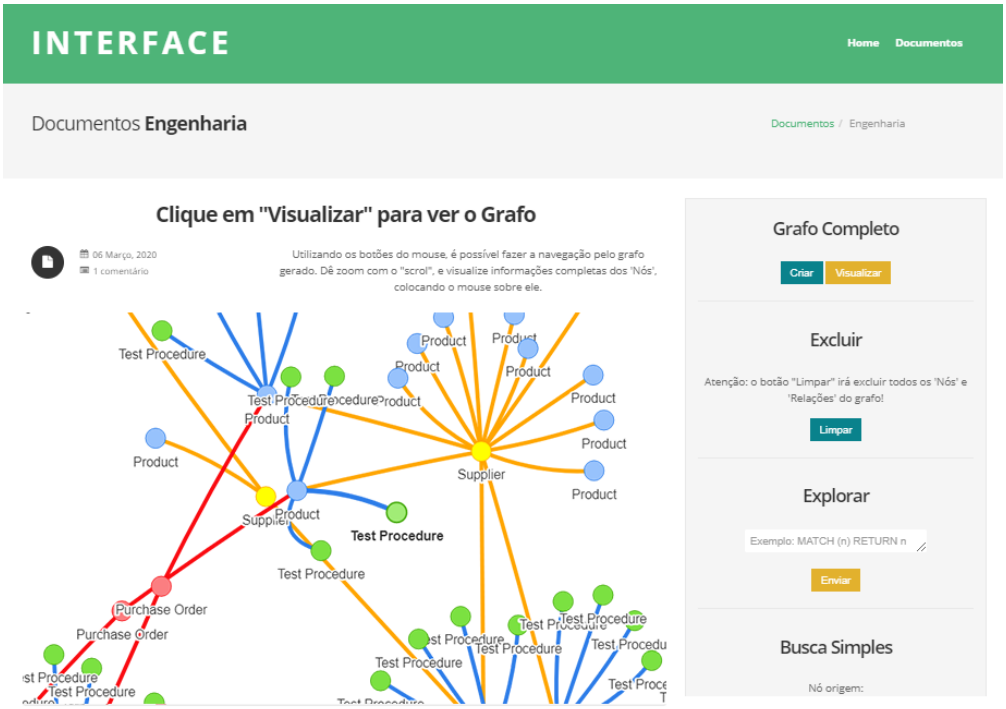

图6：测试版。

进一步分析发现，这种差异是由图中的歧义节点造成的，其原因是17家供应商存在128种语法变体。变体示例包括以下术语：“supplier-A”：“A-supplier”、“supplier- A.inc”、“group-A-supplier”。解决这一问题的关键思路是识别数据集中同一实体（供应商）的不同变体，并将其统一为一种标准书写形式。为完成这一任务，研究人员开发了歧义消除算法。该算法生成了18家不同供应商的列表，仅比专家列表多出一家，具体总结见图7。因此，该算法消除了图中的110个歧义节点——歧义节点总消除率达到99.09%，并使节点数量从128家供应商减少到18家，降幅达到85.93%。

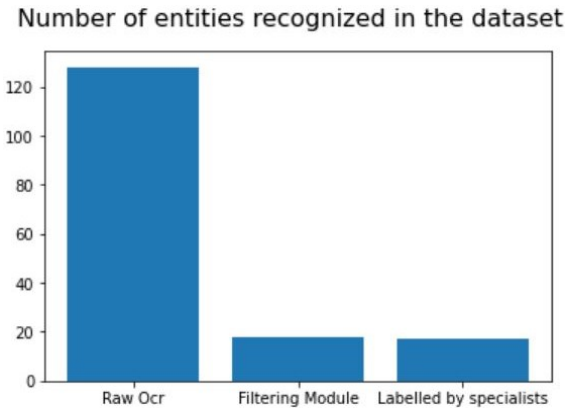

图7：　经过筛选后，光学字符识别系统识别出的实体与专家识别结果的比较　.

根据相似度（OCR文本与原始真实文本的比较），结果分为：准确率为*100%*（完全命中）、部分准确 （部分命中）以及不一致 （OCR文本与真实文本之间不存在任何关系）。如图8所示，在简单文档 的情境中，系统成功识别出文档内85.9%的数据元素并完全命中，12.72%部分命中，1.58%不一致（完全错误）。然而，对于困难文档，结果并不理想：完全命中的比例为25.67%，部分命中为24.32%，不一致的比例为50.01%。

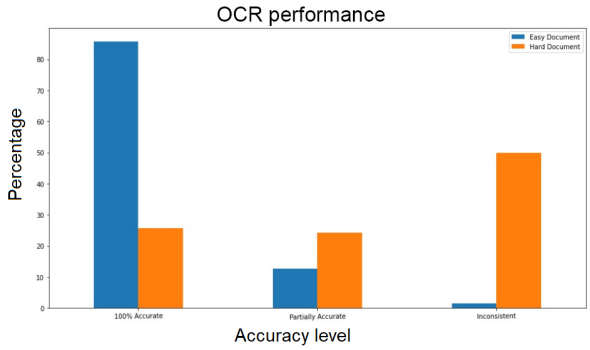

图 8：光学字符识别性能：简单文档与困难文档。

这些结果清楚地表明，噪声和歧义会在经过光学字符识别处理的文档中传播。这类错误无助于根据OCR数据构建图。

更复杂形式的歧义，例如实体标准化和首字母缩略词识别，均通过人工处理；GraphLED 系统的下一个版本将以自动化方式解决这些问题。

5.2 扩展 GraphLED

作为定量参考，我们在 Neo4j 数据库中分析了图节点、边和属性的读写速率。我们选择了 *graph-workload* [21]，这是一款用于运行负载测试的开源工具，能够以可配置且可扩展的方式生成随机数据负载和 Cypher 查询，以创建图节点、边和属性。

表 1 列出了图数据库的设置参数，其中堆初始大小设为 4GB DDR4 内存，堆最大大小限制为 8GB。

使用的实验参数为：$n = 1000$，其中“n”是数据库中的目标查询数量，而并发度 $= 10$ 是并行执行的查询数量。使用的数据是在运行时随机生成的，或以确定性方式（硬编码）生成的，与任何输入数据集都无关。所选的子图模式，例如中心辐射（图 1）和 N元树模式，也出现在行业运营文档的结构中。

表 1：图数据库设置配置。

设置	
堆初始大小	4 GB
堆最大大小	8 GB
页面缓存大小	4 GB
Neo4j 版本	3.5.14

可生成包含143,602个节点和87,427条边的完整图，近似执行时间为42286毫秒（表2）。图11展示了一个限制为10,000个节点的图样本的渲染结果。我们观察到以下图模式和操作：

- 可观察到中心辐射结构模式，其中一个中心节点连接一个“叶节点”，形成类似蒲公英的形状。此类节点以黄色表示（图9）。
- N元树模式，其中红色节点表示非平衡N元树结构，子图的层级较大是其主要特征，并以红色显示（图10）。

- 大量一对一单链结构用于填充图，其中节点在图10和图11中以蓝色、粉色和绿色表示）。此外，某些此类子图还可能具有其他属性：大量可索引属性（索引密集型）、包含大量字符节点的子图（胖节点追加）、节点通过随机链接连接的子图（随机链接）、同时包含随机节点和链接的子图（原始写入），以及通过 MERGE操作创建的子图。

• 其他类型的操作（聚合读取、长路径读取和随机访问读取）可作为直接对数据库执行的读取查询。这些操作仅用于模拟真实用户参与的生产环境，在该环境中，写入和读取操作可能同时发生。

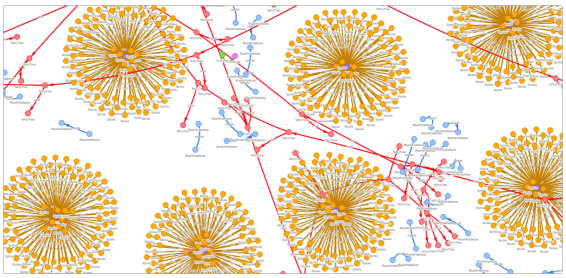

图9：中心辐射式子图模式。



图10：N元树子图模式。

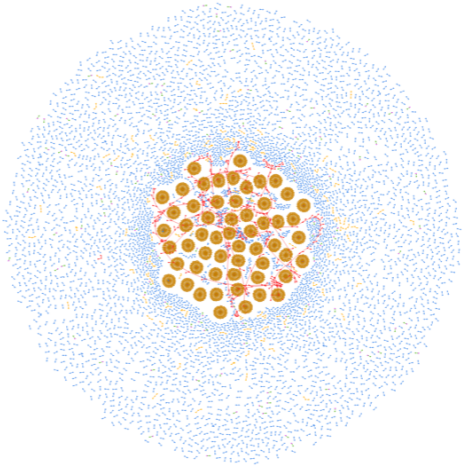

图11：包含一万个节点的子图样本。

6 结论

The GraphLED 系统满足石油与天然气检验在大数据 工程内容方面的需求 xts.

本研究提出了一种以数据为中心的机制，以提升从含噪声的OCR文档中提取的知识的正确性。所提出的数据清理和增强方法在消除数据歧义方面取得了令人满意的效果。

GraphLED 系统提供了一种基于图的工具，用于近实时的知识提取与表示 n.

表2：图数据库运行结果。

模式	RUNS	AVG	MIN	MAX
胖节点追加	1050	73.03	18	1990
N元树	1003	54.72	4	1993
合并写入	985	38.24	4	1991
随机链接	1010	31.90	5	249
中心辐射	481	60.74	20	719
原始写入	2933	41.51	3	1754
索引密集型	1002	41.10	4	1960
聚合读取	552	18.25	2	126
随机访问读取	952	37.44	7	114
长路径读取	62	24.66	2	127
总计（单位）	**10030**	-	-	-
平均值（毫秒）	-	**42.16**	**6.9**	**1102.3**

参考文献

[1] Martin Wollschlaeger、Thilo Sauter 和 Juergen Jasperneite。工业通信的未来：物联网和工业4.0时代的自动化网络。*IEEE*工业电子杂志，11(1)：17–27，2017年。

[2] Robert Amor 和 M Clift。作为建筑业并行工程赋能机制的文档。发表于第一届建筑并行工程国际会议 *CEC*，第 97 卷，1997 年。

[3] Pai Zheng、Chun-Hsien Chen 和 Suiyue Shang。面向智能产品服务系统的自动化工程变更管理——一种基于设计结构矩阵的学习方法。发表于《先进工程信息学》，39：203–213，2019 年。

[4] Kei Matsubayashi、Akihiro Yamashita、Hidetoshi Nonaka 和 Yohko Konno。基于从存储信息中提取特征词的文档摘要与演示系统研究。载于*2018*年人工智能技术与应用会议（*TAAI*），第 60–63 页。IEEE，2018 年。

[5] Qing Guan、Fuli Zhang 和 Enli Zhang。知识图谱技术在油气勘探与开发知识管理中的应用前景。载于*2019*年第二届人工智能与大数据国际会议（*ICAIBD*），第 161–166 页。IEEE，2019 年。

[6] Christian M Dahl 和 Emil Nørmark Sørensen。文档数字化与机器学习。2019 年。

[7] Eduard Hovy、Roberto Navigli 和 Simone Paolo Ponzetto。协作构建的半结构化内容与人工智能：迄今为止的发展历程。《人工智能》，194：2–27，2013 年。

[8] Elena Simperl、Oscar Corcho、Marko Grobelnik、Dumitru Roman、Ahmet Soylu、María Jesús Fernández Ruíz、Stefano Gatti、Chris Taggart、Urška Skok Klima、Annie Ferrari Uliana、Ian Makgill 和 Till Christopher Lech。面向公共采购的基于知识图谱的平台。载于 Emmanouel Garoufallou、Fabio Sartori、Rania Siatri 和 Marios Zervas（编），元数据与语义研究，第 317–323 页，查姆，2019 年。施普林格国际出版公司。

[9] Manjai Lee 和 Byung-Won On。公共数据的社交图可视化技术。《韩国人机交互学会期刊》，10(1)：5–17，2015 年。

[10] Masaki Eto。扩展共引搜索：在包含引文上下文信息的共引网络上进行基于图的文档检索。《信息处理与管理》，56(6)：102046，2019 年。

[11] Trevor Strohman、W Bruce Croft 和 David Jensen。《为学术论文推荐引文》。载于第*30*届*ACM SIGIR*信息检索研究与发展年度国际会议论文集，第705–706页，2007年。

[12] Ghulam Mujtaba、Liyana Shuib、Ram Gopal Raj、Retnagowri Rajandram、Khairunisa Shaikh 和 Mohammed Ali Al-Garadi。基于概念图的文档表示模型在法医尸检报告分类中的应用。《生物医学信息学杂志》，82:88–105，2018年。

[13] Yufeng Zhang、Xueli Yu、Zeyu Cui、Shu Wu、Zhongzhen Wen 和 Liang Wang。每篇文档都有自己的结构：通过图神经网络实现归纳式文本分类。*arXiv*预印本 *arXiv:2004.13826*，2020年。

[14] Mads Holten Rasmussen、Maxime Lefrançois、Pieter Pauwels、Christian AnkerHviid 和 JanKarlshøj。管理AEC知识图谱中的相互关联项目信息。《*Automation in Construction*》，108:102956，2019。

[15] Carlos Francisco Moreno-Garcia 和 Eyad Elyan。石油和天然气行业资产的数字化：挑战与机遇。载于*2019*年国际文档分析与识别研讨会（*ICDARW*），第7卷，第2–5页，2019。[16] Carlos Francisco Moreno-García、Eyad Elyan 和 Chrisina Jayne。复杂工程图纸数字化的新趋势。神经计算与应用，31(6):1695–1712，2019。[17] Carlos Francisco Moreno-García、Xavier Cortés 和 Francesc Serratosa。用于学习容错图匹配的图仓库。载于*IAPR*统计模式识别（*SPR*）与结构和句法模式识别（*SSPR*）联合国际研讨会，第519–529页。施普林格，2016。[18] Gabriel L Santos、Vanessa T Silva、Laura A Dalmolin、Ricardo N Rodrigues、Paulo LJ Drews 和 Nelson L Duarte Filho。一种面向印刷及结构化工程文档的表单理解方法。载于*2021*年第*34*届*SIBGRAPI*图形、模式与图像会议（*SIBGRAPI*），第330–337页。IEEE，2021。[19] Rishin Haldar 和 Debajyoti Mukhopadhyay。词典查找方法中的莱文斯坦距离技术：一种改进方法。*arXiv*预印本 *arXiv:1101.1232*，2011。[20] Jörg Tiedemann。加权字符串相似度量的自动构建。载于*1999*年*SIGDAT*自然语言处理经验方法与超大型语料库联合会议，1999。[21] David Allen。图工作负载。https://github.com/moxious/graph-workload，2019。