Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 309-315
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141030

Hans汉斯

企业审计知识图谱的构建

——以华神科技公司为例

李笑笑，安玉娥*，段祎然

上海立信会计金融学院，统计与数学学院，上海

收稿日期：2023年11月9日；录用日期：2023年11月29日；发布日期：2024年2月23日

摘　要

应对传统审计无法满足日益增长的全方位精确审计需求的挑战，智能审计相关技术应运而生，但仍处于弱智能化阶段，未能达到可用于审计知识推理的标准。本文基于BERT-BiLSTM-CRF实体识别模型，创新性地运用了改进的预训练模型BERT-WWM，提出了新的企业内控审计知识图谱的构建方法，获得最优序列标注，对审计报告、财务报表及内部审计文件等实体进行字符抽取，在知识图谱中引入深度学习、有效挖掘海量审计实体之间的复杂关系，并实现不同结构数据的融合。本文以华神科技公司为例，采用neo4j图数据库构建审计知识图谱，挖掘知识图谱推理价值、延拓审计知识图谱使用的广度和深度，为审计知识推理赋能，并为企业创造更高价值。

关键词

知识图谱，知识推理，智能审计，深度学习

The Construction of the Enterprise Audit Knowledge Graph

—Taking Huasen Technology Company as an Example

Xiaoxiao Li, Yu'e An*, Yiran Duan

School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai

Received: Nov. 9th, 2023; accepted: Nov. 29th, 2023; published: Feb. 23rd, 2024

Abstract

To address the challenges of increasingly comprehensive and precise auditing demands that traditional auditing methods cannot meet, intelligent auditing technologies have emerged. However,

*通讯作者。

文章引用: 李笑笑, 安玉娥, 段祎然. 企业审计知识图谱的构建[J]. 运筹与模糊学, 2024, 14(1): 309-315.
DOI: 10.12677/orf.2024.141030

these technologies are still in a stage of weak intelligence and have not yet reached the standards required for knowledge reasoning in auditing. In this paper, based on the BERT-BiLSTM-CRF entity recognition model, we innovatively utilize an improved pre-training model called BERT-WWM and propose a new method for constructing an enterprise internal control audit knowledge graph. This method achieves optimal sequence labeling by extracting characters from entities such as audit reports, financial statements, and internal audit documents. We introduce deep learning into the knowledge graph and effectively mine complex relationships among massive auditing entities, realizing the fusion of different structured data. Taking Huashen Technology Company as an example, we use the neo4j graph database to construct an audit knowledge graph, explore the inference value of the knowledge graph, expand the breadth and depth of its use in audit knowledge reasoning, empower audit knowledge reasoning, and create higher value for enterprises.

Keywords

Knowledge Graph, Knowledge Reasoning, Intelligent Auditing, Deep Learning

1. 引言

审计是企业发展的重要组成部分，对提高企业内部控制力度，降低违规风险，提升企业公众形象都有着重要作用。随着大数据技术在企业审计中的应用，智能审计在业务流程、内控机制、管理效能等多个方面都为传统审计带来了新思路，大幅减少了审计成本，提高审计效率和准确性[1]。但是，由于人工智能技术仍处于发展阶段，其技术的使用与场景应用并不完善，尤其体现在审计知识图谱关系单一、结构简单等方面，与知识推理所需要的多维度、深复杂度需求[2]相差甚远，导致多数审计知识图谱仅能用于向审计人员展示文件资料的数据关系，无法用于进一步自动化推理。因此，如何构建审计术语知识图谱、发掘审计信息潜在价值，从而提高智能审计技术壁垒成为当前亟待解决的问题。

目前相关研究显示，随着深度学习技术的不断发展，知识图谱等相关自然语言处理技术经历了快速的优化与迭代，其中 NER 提取效果的优劣直接影响后续审计语言处理质量效率。Humphreys 等人[3]提出了基于规则的 LaSIE-II 系统，但此类方法过多依赖语言学家指定的规则模板，程序繁琐。Zhang 和 Yang [4]提出了一种 LSTM 模型与字符间隔的变种(Lattice-LSTM)，利用外部词典中的词汇信息，建立了字符之间的关联联系，但其只能让词粒度信息影响当前位置前后的序列，并不能对当前位置的结果进行直接影响。高翔等[5]提出使用条件随机场(CRF)与长短时记忆神经网络(LSTM)相结合的 LSTM-CRF 模型，通过加入预先训练的字嵌入向量及不同词位标注集，对军事动向文本进行实体识别。2018 年 Google 团队结合不同语言模型的优点，提出 BERT 模型，充分描述字符、词语和语句之间的关系特征，表征不同语境中的相同词的语义，有效解决一词多义的问题[6]。姜同强等通过 BERT 层进行字向量预训练，根据上下文语义生成字向量，字向量序列输入 BiLSTM 层和 Attention 层提取语义特征，再通过 CRF 层预测并输出字的最优标签序列[7]。

以上研究成果为知识图谱构建中的复杂实体识别积累了经验，但实际应用审计领域的跨学科研究较少，为此，本文将挖掘审计知识图谱的潜在价值，向实现自动化推理分析推进。

2. 审计知识图谱构建方法

现有审计知识图谱多停留在理论与初步实践阶段，可用性不强且分析推理效果较差，而随着企业数

据不断增加，将知识推理应用于审计知识图谱迫在眉睫，构建灵活且表示结构复杂的知识图谱为多数企业所需要。

针对当前应用于审计行业人工智能技术不完善，实体识别存在偏移、关系挖掘不够精确，且现有知识图谱灵活性不足等问题，本文基于 BERT-BiLSTM-CRF 实体识别模型，创新性地将改进的预训练模型 BERT-WWM 与卷积神经网络相融合，有效的提高了知识结构灵活度与图谱表示效果。

由于审计报告等数据大多以文字等非结构化形式存在，首先需要对文章进行分词、去除停用词等清洗活动；清洗后的数据经由 BERT 改进模型通过命名实体识别、关系挖掘等一系列处理获取效能较好的三元组。最后，将单个三元组进行实体融合、关系合并等操作形成层次清晰的知识图谱并存入 neo4j 图数据库中，完成审计知识图谱的构建。

2.1. 审计数据的知识挖掘

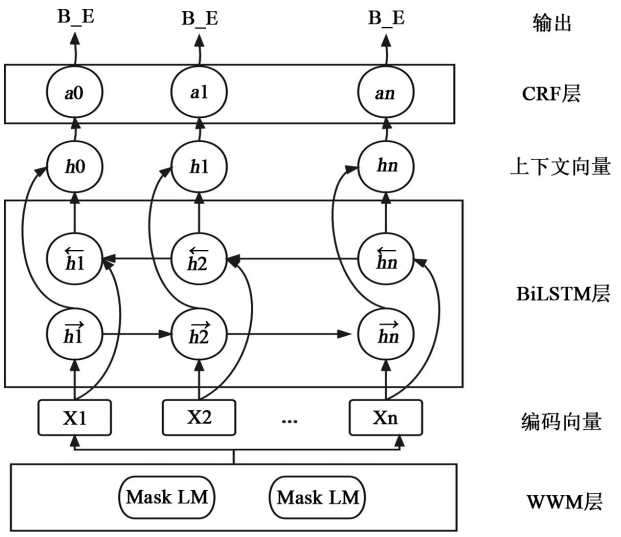
知识图谱是一种语义网络，近年来得到了广泛应用。它具有强大的表达能力，可以对审计数据中的实体、概念、属性以及它们之间的关系进行灵活建模。知识图谱的构建和使用更贴近人类的学习方式，旨在组织和结构化推理知识。构建知识图谱的基础是命名实体识别和关系提取。随着自然语言处理技术的快速发展，预训练语言模型(Pre-trained language model, PLM)对于语言信息的表示效果大幅提升。在下游任务中，PLM 表现出更好的性能，并且可以加快模型的收敛速度，从而提升整体模型性能。

2.1.1. 数据预处理模块构建

由于爬取的审计数据包括结构化数据和非结构化数据，非结构化数据无法直接运用到后续的研究，因此先对文本进行数据清洗，去除语气助词、标点、特殊字符等非必要信息，增强数据价值。同时本文采用哈工大 LTP 进行基本分词、词性标注等任务，在进行命名实体识别前完成一系列数据预处理工作。

2.1.2. 实体识别模块

BERT (Bidirectional Encoder Representations from Transformers)是一种预训练语言模型，它在训练过程中采用了遮蔽语言模型(Masked Language Model)。在预训练阶段，文本中大约 15%的词条会被遮蔽掉，其中约 80%的被遮蔽的词条会被替换为[MASK]标记，约 10%的词条会被随机替换，而另外 10%的词条则保持不变。除了 BERT 模型，研究人员还关注了 BiLSTM-CRF 模型，在实体识别研究中发挥了重要作用。然而，这种模型在某些方面存在问题，本文使用了一种改进的 BERT-WWM-BiLSTM-CRF 模型，如下(图 1)所示：



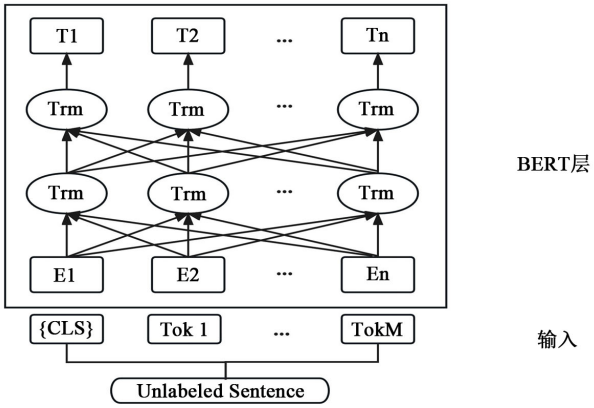


Figure 1. BERT-WWM-BiLSTM-CRF Model
图 1. 改进的 BERT 预训练模型

1) 词嵌入层

采用基于全词掩码(Whole Word Masking)技术的中文预训练模型 BERT-WWM 作为词嵌入层。而在 Google 发布的 BERT-BASE 中，由于是英文语料，通过 WordPiece 分词可以得到比词更小的粒度，应用到中文却无法获得同样效果，加之其对中文单字的掩盖导致模型对于词语信息学习的不完整性。为解决这一问题，本文在 BERT-WWM 中使用全词 Mask 的方法应用到中文分词中，将完整词未被 Mask 的部分也考虑在内，弥补了原有基于 WordPiece 分词方式的欠缺，保证词语完整性的同时又不影响其独立性。

本文首先对数据中的审计文本实体信息进行编码，能够将 BERT-WWM 输出的向量信息中的 head 实体和 tail 实体的向量进行平均，公式如下所示：

$$h_1 = W_1\left[\tanh\left(\frac{1}{j-i+1}\sum_{t=i}^{j} H_t\right)\right] + b_1 \tag{1}$$

$$h_2 = W_2\left[\tanh\left(\frac{1}{k-m+1}\sum_{t=m}^{k} H_t\right)\right] + b_2 \tag{2}$$

其中 W_1 和 W_2 为两个随机初始化向量，b_1 和 b_2 为随机初始化的偏置值。

2) 特征提取层

针对语义特征，本文采用的是双向长短期记忆神经网络(Bidirectional Long Short-Term Memory, BiLSTM)，由于 LSTM 是单向的，只能将信息按从前到后的顺序编码，以及审计内控数据文本是上下文相关的，使用 BiLSTM 同时获取从前到后以及从后到前的序列信息，进而将词嵌入矩阵送入 BiLSTM 网络提取更深层次的语义特征，得到前向 LSTM $\left\{\overrightarrow{h_1},\overrightarrow{h_2},\overrightarrow{h_3},\cdots,\overrightarrow{h_n}\right\}$ 和反向 LSTM $\left\{\overleftarrow{h_1},\overleftarrow{h_2},\overleftarrow{h_3},\cdots,\overleftarrow{h_n}\right\}$ 输出并根据对应位置拼接得到完整输出序列。

3) 特征分类层

虽然 BiLSTM 模型能够充分解析上下文本的信息，但是没有考虑标记标签间的依赖信息，所以本文采用条件随机场(Conditional random field, CRF)来使特征提取层的结果输出规范化。设 X 和 Y 为随机变量，$P(Y|X)$ 是条件概率分布，假若 Y 可以构成一个无向图 $G=(V|G)$ 表示马尔可夫随机场，则表示为 $P(Y_v|X,Y_v,w/=v)=P(Y_v|X,Y_w,w\sim v)$，式中：对任意的节点 v 成立，条件概率分布 $P(Y|X)$ 称为条件随机场；$w\sim v$；表示为与节点 v 有连接的所有 w；$w/=v$ 代表 v 以外的节点；Y_v,Y_w 为随机变量，在命名实体识别任务中，X 和 Y 都为线性条件随机场的随机变量序列，在给定 X 的条件下，Y 的概率分布 $P(Y|X)$ 为

$$P\left(Y_i \middle| X, Y_1, \cdots, Y_n\right) \approx P\left(Y_i \middle| X, Y_{i-1}, Y_{i+1}\right) \tag{3}$$

通过训练学习得到标签转移概率，然后为预测的标签设置约束条件以此来排除非法标签，来计算出最优的特征标签输出序列。

3. 知识图谱构建及分析——以华神科技公司为例

对于企业而言，健全和实施内部控制以及整体管理是完成企业审计系统建设尤为重要的一环。在方向选择和策略指导等多个方面，都企业有重要的意义。然而，在企业内部进行审计时，数据庞大且更新速度快，有价值的审计信息潜藏在海量行业数据之中，而数据之间的关系错综复杂，很难进行梳理和分析。

为了解决这个问题，知识图谱的推理应用变得越来越重要。知识图谱通过将文字性的非结构化数据转化为便于处理的结构化信息，方便对数据进行筛选、分析和推理，而构建关系更复杂、结构更灵活的知识图谱，是实现全自动化审计分析、推理并自动生成审计分析结果的必由之路。

对众多公司进行考察后，华神科技公司以其囊括丰富的审计案例，且各审计案例可研究性强等特征被选为本文案例进行分析。本文运用基于预训练模型的文本挖掘和分类技术以及深度学习算法来建立知识图谱，并详细说明了知识图谱构建的关系步骤。知识图谱的构建与优化，打破了知识推理的结构壁垒，为实现知识推理从而自动化分析审计报告创造了条件。

3.1. 数据采集

使用 Python 中的 Scrapy 框架，我们从新浪财经华神科技公司的网页上获取各种非结构化审计相关数据和结构化财务数据。首先，通过对数据的清洗，去除数据中的标点字符以及标识符等噪声，以减少对准确性的影响。在对数据进行预处理之后，通过哈工大 ltp 数据库对非结构化数据进行分词处理，并筛除修辞、停用词等无用词汇。

3.2. 实体获取

在获取经过初步清洗的数据后，采用改进的 BERT-WWM-BiLSTM-CRF 模型构建语义网络，用于从数据中抽取事件中重要的代表性词汇和词汇联系。同时，我们还使用了 BERT 模型来进行实体识别，对名称、组织机构、地名等进行筛选。然后，我们根据相似度计算的结果对实体进行融合，从原始数据中提取出有意义的实体，最终获得了完整的实体信息，为后续图谱构建与推理分析提供基础。

3.3. 关系挖掘

完成实体识别后，我们继续构建了一个新的 CNN 网络，用于实现关系抽取，即在两个实体之间建立关联。卷积神经网络(CNN)的结构通常由输入层、卷积层、池化层和全连接层组成。此外，为了增强 CNN 学习语义表示的能力，还会添加一些额外的特征，比如词法和语法信息。同时，位置向量也会被引入以考虑句子中词的语序，使得 CNN 能够更好地理解和表达句子的语义含义。使用基于 CNN 的关系挖掘方法，我们对输入的语句进行编码，并利用多层卷积层和池化层逐步提取特征，从而获得数据语句的高级语义表示。最后，我们使用一个基于全连接层的分类器对这些特征进行分类，以识别句子中存在的因果等多种关系。在这个过程中，由于数据在不同情况下的多样性和差异性，我们需要根据具体情况来选择模型、调整参数和设置，提高模型的性能和适应能力，从而更好地进行关系抽取任务。

3.4. 知识图谱构建

整合实体与关系组成的三元组，并利用深度神经网络挖掘数据对它们之间的特征进行表示，识别三元组之间的多层关系。最后链接 neo4j 图数据库，构建面向内控审计的知识图谱，并存入数据库中，形

成有组织结构的信息网络，为审计知识推理的实现提供更有力的支持。

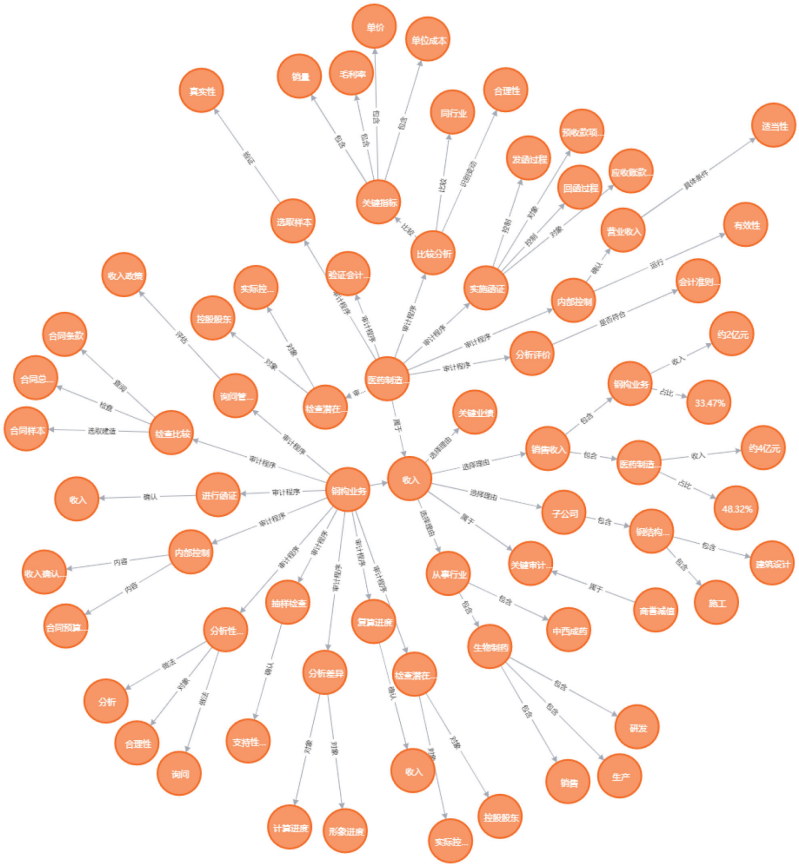


Figure 2. Huashen technology audit knowledge graph (part)
图 2. 华神科技审计知识图谱(部分)

本文图谱将篇幅较长、需大量时间阅读且难以处理的非结构化审计报告，通过三元组的形式，以事态逻辑为脉络进行连接，转化为便于处理、易于分析的、以逻辑与要素构成的结构化知识图谱数据形式，为全自动化审计推理奠定了结构基础。图谱的实体与关系并不被局限于某个特定的类别，而是根据输入数据灵活变换。以审计报告生成的部分图谱(图 2)为例：原始审计报告中，重要审计事件(即医药制造业、钢铁制造等)的事件原因、审计程序，以及公司采取措施的原因、内容和事件结果被转化为数据化形式存入知识图谱中(如：源文件中“收入方面的关键审计业务为医药制造业与钢构业务”，就分别以“关键审计业务－收入→医药制造业”和“关键审计业务－收入→钢构业务”的三元组形式被存入知识图谱)，知识推理时即可通过对各个三元组的分析推理，得到原审计报告的分析结果。

3.5. 知识推理

审计知识图谱构建完成后，后续尝试通过 RSKCCA 算法特征化两组变量的非线性关系，从而统计分析事理逻辑关系，从构建的知识图谱中进行分析与推理从而获取内控审计有效性评价。接着，通过基于行为模拟的方法(actor-mimic)，我们可以利用多关系深度迁移来预训练深度策略网络。这个方法通过指导监督信号，使得单个策略网络能够学习适应于不同任务的策略，并将其学到的知识迁移到相似的新任务中。这样可以实现对数据的即时性分析，并结合外部组织对企业相关措施的评价，生成其综合评价报告。

通过这种方法，快速生成对原非结构化审计数据的分析结果，直接性的省去了审计人员对大量繁琐的文字性数据的阅读与人工低效率分析，大大提高了企业审计的效率。

4. 总结

针对目前审计领域存在实体关系交叉关联、多源异构数据聚合能力差、利用率低、知识共享困难等问题，本文基于自然语言处理等相关深度学习技术，提出了 Bert-WWM-BiLSTM-CRF 预训练模型构建的审计知识图谱方法，优化了知识图谱结构简单、灵活度低等问题，解决了现有知识图谱结构复杂度与数据覆盖量无法达到知识推理要求的问题，为知识推理的实现创造了条件，也为构架企业立体化数智审计系统为智能审计的搭建提供新空间。

基金项目

校“互联网+”种子项目：基于图谱推理的立体化数智赋能审计系统。

参考文献

[1] 樊世昊. 基于知识图谱的审计方法研究[D]: [硕士学位论文]. 南京: 南京审计大学, 2018.

[2] 刘琦. 基于 Neo4j 的学科知识可视化检索系统的实现[D]: [硕士学位论文]. 郑州: 河南大学, 2019.

[3] Humphreys, K., Gaizauskas, R., Azzam, S., *et al.* (1998) University of Sheffield: Description of the LaSIE-II System as Used for MUC-7. *Proceedings of the* 7*th Message Understanding Conference*, Fairfax, 29 April-1 May 1998, 1-20.

[4] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Melbourne, 15-20 July 2018, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[5] 高翔, 张金登, 许潇, 等. 基于 LSTM-CRF 的军事动向文本实体识别方法[J]. 指挥信息系统与技术, 2020, 11(6): 91-95. https://doi.org/10.15908/j.cnki.cist.2020.06.017

[6] Devlin, J., Chang, M.W., Lee, K., *et al.* (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[7] 姜同强, 王岚熙. 基于双向编码器表示模型和注意力机制的食品安全命名实体识别[J]. 科学技术与工程, 2021, 21(3): 1103-1108.