



Building an ontology for ESG investment using securities reports and integrated reports and extracting information using generative AI

Ryosuke Ohori

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

July 3, 2025

有価証券報告書と統合報告書を活用した ESG 投資のための オントロジー構築と生成 AI による情報抽出

English Title Ontology construction and information extraction using generative AI
for ESG investment using securities reports and integrated reports.

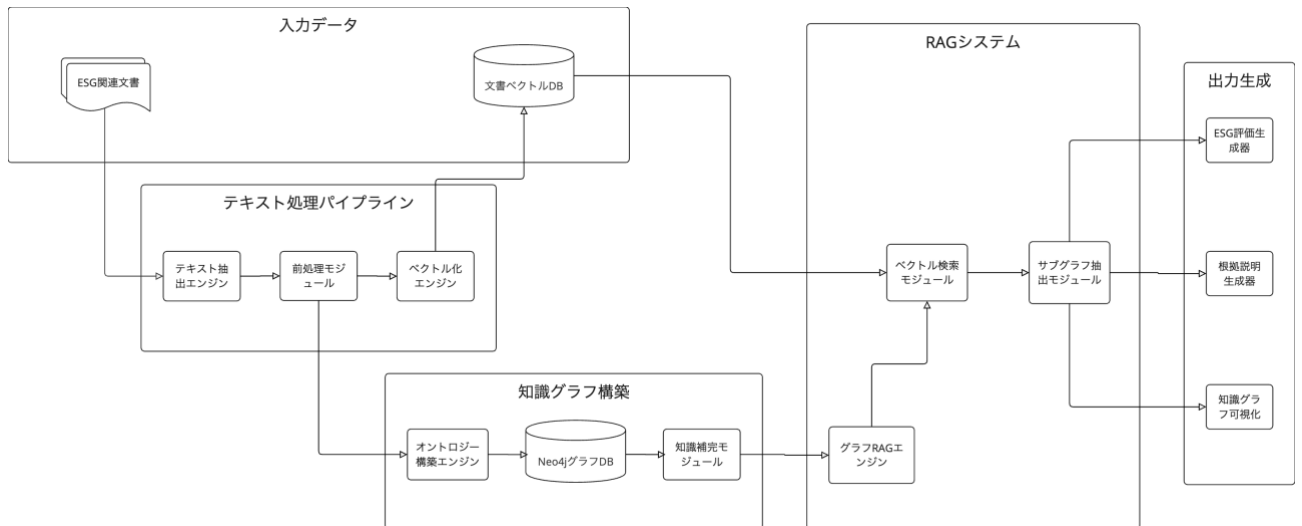
大堀遼介¹

Ryosuke Ohori¹

¹株式会社 ulusage

¹ulusage Inc.

Abstract: 環境・社会・ガバナンス（ESG）分野では、企業のサステナビリティに関する膨大な情報が報告書やニュース記事などで提供されている。しかし、この情報は多種多様かつ分散的であり、意思決定に活用するには統合的な知識基盤が必要である。本研究では、ESG ドメインに特化したナレッジグラフを構築し、大規模言語モデル（LLM）とグラフニューラルネットワーク（GNN）による知識補完を行った上で、構築した知識グラフを活用した GraphRAG（グラフベース RAG）QA システムを提案する。提案手法では、テキストから ESG 領域のオントロジーと関係を抽出してナレッジグラフを生成し、Neo4j グラフデータベースに保存する。さらに、LLM と GNN を組み合わせて未発見の関係を推論・補完し、知識の網羅性を高める。最後に、このナレッジグラフを QA 利用する GraphRAG アーキテクチャを設計することで、質問に対して信頼性が高く説明可能な回答を生成する。



1. はじめに：サステナビリティと ESG に対する関心が近年高まっており、企業は環境・社会・ガバナンスに関する取り組みを積極的に開示し始めている。ESG 情報は投資判断や企業価値評価に組み込まれ、長期的な成長とリスク管理に不可欠な要素となっている。しかし、ESG データは年間報告書、ニュース記事、評価レポートなど様々な情報源に散在しており、

その形式も定性的な記述から定量的な指標まで多岐にわたる。ナレッジグラフは多様なデータをエンティティ（実体）間の関係として統合し、知識の体系化と推論を可能にする

1. ESG ドメイン知識グラフの構築：文書から ESG に関するエンティティ・関係を抽出し、既存の基準（例：GRI スタンダード）

を参考にオントロジーを設計してナレッジグラフを構築する手法を提案する。グラフは Neo4j 上に実装し、Cypher クエリによりエンティティとリレーションを格納・管理する。

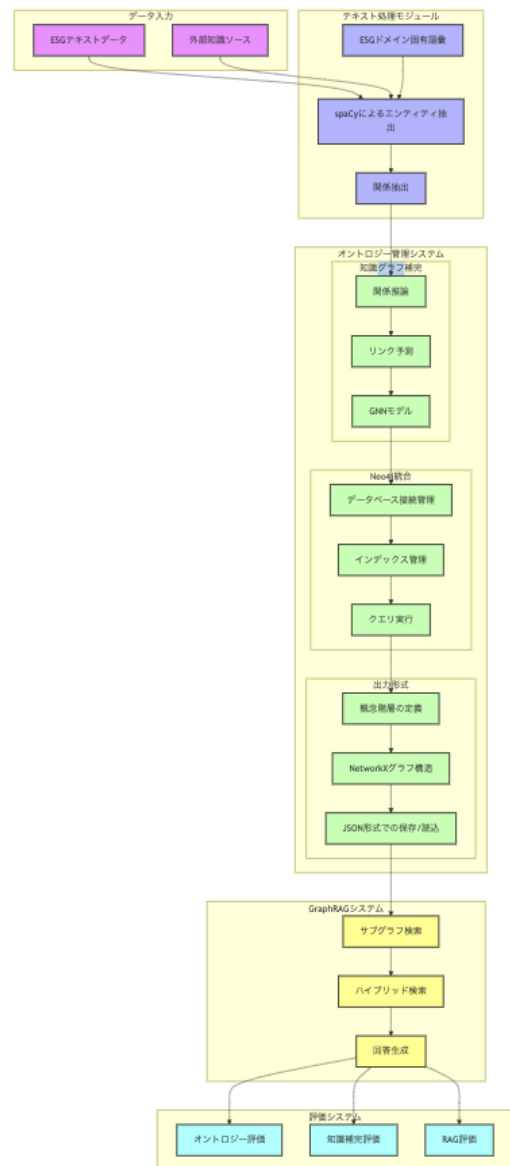
2. **知識補完のための LLM・GNN 統合:** 構築したナレッジグラフの欠損を補完するため、グラフ構造に基づく関係推論手法（グラフ埋め込みモデルや GNN）と、大規模言語モデルを用いた知識推論を組み合わせ、新たなエッジ（関係）やノードを発見するアプローチを開発する。LLM を用いてグラフを自然言語表現に変換し多段のリンク予測を行う手法など、最新の知識補完技術を ESG グラフに応用する。
3. **GraphRAG による QA システム:** ナレッジグラフを質問応答に統合する GraphRAG アーキテクチャを設計する。ユーザ質問からグラフ内の関連ノードを検索し、その情報をコンテキストとして LLM に与えて回答を生成するパイプラインを構築した。知識グラフに基づく検索により回答の根拠を提示し、LLM 単体の場合に比べ正確かつ説明可能な回答を得られるようにした。
4. **包括的評価:** ESG データセットを用いて提案手法の効果を評価した。知識グラフ構築の精度を人手アノテーションにより検証し、LLM・GNN によるリンク予測の性能をヒット率や F1 スコアで定量評価する。また、QA システムについて、ベースライン（テキスト検索+LLM）と比較して正答率の向上と誤答（幻覚）の減少、および計算資源コストを評価する。

2. 関連研究: ナレッジグラフと知識補完

ナレッジグラフ構築に関する研究は近年多くの分野で盛んである。一般ドメインでは Wikipedia 等を基にした汎用的なナレッジグラフ（例：DBpedia, Wikidata）が構築されているが、ESG のような専門領域ではドメイン知識に即したカスタムなオントロジー構築が必要となる。企業のサステナビリティ報告に含まれる指標・項目は GRI や SDGs などの階層的枠組みで整理でき、これを基に ESG 概念のタクソノミー（分類体系）を構築する取り組みが報告されている。典型的なアプローチとしては知識グラフ埋め込みモデルが挙げられる。また近年はグラフニューラ

ルネットワーク（GNN）を用いて構造的特徴を学習しリンク予測を行う手法も注目されている。

提案手法: 本章では、ESG 領域におけるナレッジグラフ構築と GraphRAG 質問応答システムの詳細な手法について述べる。まず、テキストデータから ESG ドメイン知識を抽出しオントロジーに基づくナレッジグラフを構築するステップを説明する。次に、構築したグラフを Neo4j データベースに保存する方法について、スキーマ設計とクエリ例を含めて示す。続いて、LLM と GNN を用いて知識グラフの補完（リンク予測・新規関係発見）を行う手法を述べる。最後に、このナレッジグラフを統合した GraphRAG 質問応答システムのアーキテクチャと動作を説明する。



3. 実験：テキストからのドメイン知識抽出 (オントロジー構築と関係抽出)

1) **データ収集と前処理**：最初に、ESG 領域の知識源となるテキストデータを収集する。本研究では主に (a) 企業のサステナビリティ報告書・統合報告書、(b) ESG 関連のニュース記事・プレスリリース、(c) サードパーティによる ESG 評価レポートを対象データとした。これらの文書は PDF や HTML、テキストなどの形式で得られるため、事前にテキスト抽出・クリーニングを行う。

2) **ドメインオントロジーと概念階層の設計**：ESG 領域の包括的な知識表現を目指し、まずドメイン専門家の知見や既存の標準を参考にオントロジー（概念階層）を構築する。具体的には、GRI（Global Reporting Initiative）スタンダードの指標分類や、SDGs（持続可能な開発目標）、TCFD（気候関連財務情報開示タスクフォース）ガイダンスなどを参照し、ESG に関わる主要なトピックを上位から下位へ階層化した。例えば、最上位には環境・社会・ガバナンスの3つの基本カテゴリを置き、その下位に *Energy*（エネルギー）、*Emissions*（排出）、*Labor Practices*（労働慣行）、*Ethics*（倫理）等の中位概念を配置する。さらに中位概念を具体化する下位概念として、*renewable energy*（再生可能エネルギー）、*greenhouse gas emission*（温室効果ガス排出）、*child labor*（児童労働）、*corporate bribery*（企業贈収賄）といった具体的トピックを列挙した。それぞれの関係性として、上位概念から下位概念へは「包含する (Contains)」「一種である (IsA)」等の階層関係を定義した。

3) **エンティティ抽出とリンク付け**：オントロジーが定まったら、収集したテキストからエンティティ（主語・目的語候補）とそれらの関係を抽出する。まずは固有表現抽出 (Named Entity Recognition) により企業名、組織名、人名、地名などを検出する。企業名については収集対象企業のリストをもとにパターンマッチングも行った。次に、ESG 特有の用語（例：「カーボンニュートラル」「労働安全」）については、オントロジーの中位・下位概念リストを辞書としてキーワードマッチングを実施した。さらに、テキスト中の重要フレーズを漏れなく抽出するため、RAKE (Rapid Automatic Keyword Extraction) アルゴリズムを適用し、名詞句や重要語の抽出を行った。

抽出したエンティティをオントロジー内のノードとリンク付けする作業には、語彙の多様性や同義表現の問題がある。そこで各エンティティ表現について、Sentence-BERT を用いて文脈埋め込みを計算し、オントロジー内既存エンティティの埋め込みとコサイン類似度を算出した。

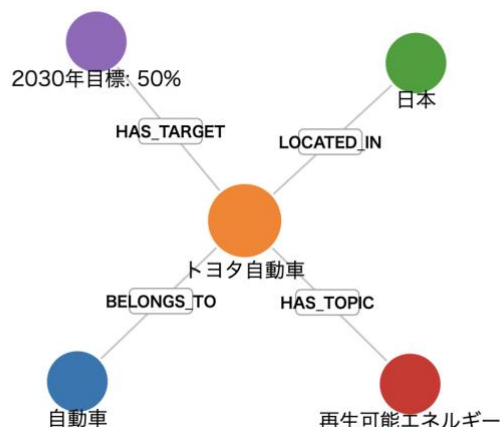
4) **関係抽出**：エンティティ同士の関係については、文章中の述語や構文パターンから抽出した。まず、IE（情報抽出）の手法としてオープン IE ツール（例えば Stanford OpenIE）を適用し、テキストから \$(主体, 関係詞, 客体)\$ の三項組を抽出した。ここで主体・客体に当たるテキストが上記エンティティ抽出で得たものと一致（または類似）する場合に、その2つのノード間に関係エッジを付与した。

5) **知識グラフデータの生成**：上記のエンティティと関係抽出の結果を組み合わせ、ナレッジグラフに取り込むデータを生成した。

- **企業ノード**：各企業を表すノード（例：*Company: トヨタ自動車, Company: Tesla Inc.*）。属性として社名、所在地、業種などを保持。
- **ESG 概念ノード**：オントロジーで定義した ESG の概念（基本・中位・下位の各カテゴリ）。属性として名称（日本語・英語）、レベル区分、対応する指標 ID (GRI コードなど) を持つ。
- **記事ノード**：情報源となった文書（報告書）を表すノード。属性にタイトル、発行日、開示内容等。記事テキスト本文そのものは長大なので格納せず、別途全文検索インデックスで管理。
- **関係エッジ**：上記ノード間の種々の関係を表すエッジ。主なものは以下の通り：
 - 企業ノード → 概念ノード：企業が特定の ESG トピックに関連する場合のエッジ。例えば企業がそのトピックに関する目標設定や実績報告を行っている場合にリンクする。関係ラベルは行動に応じて *hasGoal*, *hasIssue*, *violates*, *implements* 等を付与。
 - 記事ノード → 企業ノード：記事がどの企業に関するものか。例えばニュース記事から企業名を検出できれば、その企業ノードとの間

に *mentions* (言及) エッジを張る。

- 記事ノード → 概念ノード: 記事中で言及されている ESG 概念との間に *mentions* または *discusses* エッジを張る。テキスト中のキーワード抽出結果に基づき、記事がどのトピックに言及しているかを表現。前述の RAKE→埋め込みマッチで記事内フレーズを概念ノードに対応付けた結果をそのままエッジにしたものである。
- 概念ノード間: オントロジーの階層構造に基づき、上位・下位概念を *Contains* や *IsA* 関係で接続。また、同レベルで関連の深い概念同士 (例: *CO2 emission* と *methane emission*) は *RelatedTo* 的なエッジで結ぶ場合もある。これらは事前定義または知識補完フェーズで追加。



LLM・GNN による知識補完 (関係推論、リンク予測、新しい関係発見)

初期ナレッジグラフには、元データ中に明示的に記述された関係しか含まれないため、暗黙的な知識や欠落したつながりを補完する必要がある。

Neo4j への保存 (Cypher クエリ)

1) スキーマ定義とデータインポート: Neo4j ではスキーマは厳密に強制されないが、一貫性のためノードラベルとリレーションタイプを明確に定義した。例えば企業ノードには :Company ラベル、ESG 概念ノードには :ESGTopic ラベル、記事ノードには :Article ラベルを付与する。リレーションタイプは上記の関係種別 (例: MENTIONS, HAS_GOAL, CONTAINS 等) を大文字表記で登録した。

```
```cypher
MERGE (c:Company {name: "トヨタ自動車"})
ON CREATE SET c.industry = "Automobile",
c.country = "Japan";
ESG 概念ノードの作成 (例: 再生可能エネルギー)
MERGE (t:ESGTopic {name_en: "Renewable Energy",
level: "Intermediate"})
ON CREATE SET t.name_ja = "再生可能エネルギー",
t.category = "Environment";
企業が概念に目標を設定した関係の作成
MATCH (c:Company {name: "トヨタ自動車"},
(t:ESGTopic {name_en: "Renewable Energy"}))
MERGE (c)-[:HAS_GOAL {year: 2030, target:
"50%"}]->(t);
```
```

1) グラフ埋め込みとリンク予測: まず既存グラフに対して、従来手法である知識グラフ埋め込みモデルを適用し、リンク予測を行った。Neo4j 上のグラフデータを抽出し、PyTorch-Geometric を用いて R-GCN (Relational Graph Convolutional Network) の学習を実施した。R-GCN では各エンティティを低次元ベクトルにエンコードし、隣接ノードからメッセージパッシングで表現を更新する。(リンク予測タスクによる自己教師学習)。

2) LLM によるテキスト知識からの補完: 次に、大規模言語モデルを用いて、グラフ外の知識やテキスト文脈から新たな事実を発見するステップを行った。これはある意味で、構築済みグラフを再度自然言語プロンプトとして LLM に問い合わせ、新規知識を引き出す作業である。

実験と評価: 本章では、提案手法の有効性を評価するために行った一連の実験結果について述べる。評価は (A) ナレッジグラフ構築の精度、(B) 知識補完 (LLM・GNN) の性能、(C) GraphRAG QA システムの回答品質と効率、の 3 つの観点から行った。

A. ナレッジグラフ構築の精度評価

まず、テキストから抽出したナレッジグラフの**構築精度**を評価した。構築精度とは、抽出されたエンティティや関係がどれだけ正確であるか（真の知識を反映しているか）を測るものである。評価には人手によるアノテーションデータおよび既存の知識ソースを用いた。

評価データ：10社の大手企業（トヨタ、三菱UFJ、等）について、それぞれESG報告書から抽出されたエンティティ・関係のセットを評価対象とした。また、それら企業に関する主要なESGニュース記事20件ずつを別途収集し、そこからも抽出結果を得た。専門家（サステナビリティ分野のコンサルタント2名）に協力を仰ぎ、各抽出項目について正解か否かをラベル付けしてもらった。正解の定義は、「企業XがトピックYに関して実際にその関係を持つ」と信じられる場合を真とした。

4. 評価指標：Precision（適合率）、Recall（再現率）、F1スコアを算出した。ここでは1企業あたりの抽出関係セットを1単位とし、その中で真陽性・偽陽性・偽陰性をカウントして平均値を出した。Precisionは抽出関係のうち専門家が真と認めた割合、Recallは専門家がリストアップした全真関係（抽出漏れも含む）に対する抽出ヒット率である。（※表1）

表 1

| Equation | Formula |
|-----------|---|
| Precision | $TP / (TP + FP)$ |
| Recall | $TP / (TP + FN)$ |
| F1 | $2 * (Precision * Recall) / (Precision + Recall) = 2TP / (2TP + FP + FN)$ |

結果：抽出ナレッジグラフ全体評価は(表2)に示す。抽出された関係の約8割は正しい知識であり、また既知の関係のおよそ2/3は抽出できていたことになる。Precisionに関してはESG概念のリンク付けエラーが主な偽陽性要因だった。RAKE+Embeddingによるマッチングは概ね良好だったが、意味の異なる語が近いベクトル位置にある場合の誤対応が散見された。Recallに関しては間接表現の見落としが原因の偽陰性が多かった。また、OpenIEでは検出できないような長距離依存の関係（文章をまたぐ因果関係など）は抽出されていなかった。これらのRecall不足部分は、知識補完フェーズである程度カバーされることが期待される。

表 2

| 項目 | 値 |
|-----------|------|
| Precision | 0.81 |
| Recall | 0.68 |
| F1 | 0.74 |

B. LLM・GNNによる知識補完の性能評価

次に、知識グラフ補完の性能を評価した。

リンク予測モデルの評価：R-GCNモデルを用いて未知の関係を予測するタスクでの性能を測定した。グラフのエッジをランダムに20%除去して学習に使い、除去したエッジをテストセットとしている。評価指標はHit@KおよびAUC、そして前述のPrecision/Recall/F1である。R-GCNモデルはエンティティ埋め込み次元128、層数2で学習し、エッジの存在確率をスコアリングするよう調整した。また比較として、TransEとComplExの埋め込みモデルも同じデータで学習した。

結果：R-GCNモデルのリンク予測精度は次の表3に示す。全体として、構造情報を直接活用できるR-GCNがわずかに優位だった。特に、TransEでは二次以上の隣接関係を読み取るのが苦手なためか、離れたノード間のリンクはRecallが低かった。

表 3

| モデル | Hit@10 | AUC | Precision | Recall | F1 |
|---------|--------|------|-----------|--------|------|
| R-GCN | 0.92 | 0.88 | 0.73 | 0.69 | 0.71 |
| TransE | - | - | 0.67 | 0.60 | 0.63 |
| ComplEx | - | - | - | - | 0.70 |

LLM補完の評価：LLMを用いた知識補完については定量評価が難しいが、いくつかのケーススタディで有効性を確認した。LLMのみが発見した関係は10件程度あり、専門家評価ではその約70%が妥当な関係であった。

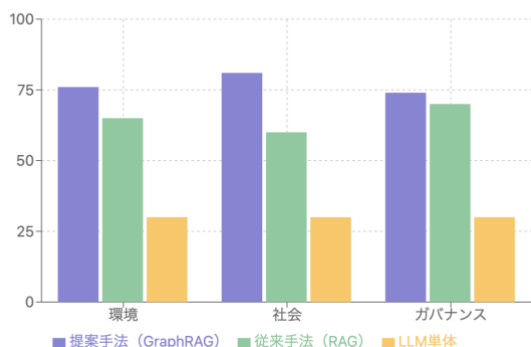
C. GraphRAGによるQAシステムの改善効果

最後に、GraphRAGを組み込んだ質問応答システムの性能を評価する。評価項目は回答の正答率、回答の

信頼性（根拠提示や一貫性）、および計算コスト（推論時間やリソース使用量）である。

評価データ： ESG 分野の多様な質問 144 件を用意した。これは先行研究で専門家が作成した環境カテゴリ 59 問、社会 41 問、ガバナンス 44 問の質問集合に準拠している。

カテゴリ別QAシステム正答率の比較



5. むすびに： 本研究では、ESG（環境・社会・ガバナンス）領域における情報統合と高度な質問応答のために、ナレッジグラフ構築と GraphRAG を組み合わせたシステムを提案した。まず、テキストデータから専門知識に基づくオントロジーと関係抽出により ESG ナレッジグラフを構築し、Neo4j に格納する手法を確立した。次に、構築したグラフに対し GNN によるリンク予測と LLM による知識推論を適用し、欠落する知識を補完してグラフを強化した。最後に、この知識グラフを LLM と統合する GraphRAG 質問応答システムを開発し、ユーザの質問に対してグラフから関連事実を検索・提示することで、正確かつ説明可能な回答生成を実現した。

今後の課題としては、時間情報の組み込み、グラフとモデルの自動更新、さらなる精度向上のためのモデル改良などが残されている。

参考文献：

[1] Jana Driller, Simon T. Trang: Unlocking sustainable reporting: Leveraging knowledge graphs for ESG metrics extraction, INFORMATIK 2024, pp. 1877-1884, (2024).
[2] Semantic Web Company: PoolParty Meets ChatGPT: ESG Knowledge Graph Application, Semantic-Web.com Blog, (Feb. 2023).

[3] Tanay K. Gupta et al.: Knowledge Graph aided LLM based ESG Question-Answering from News, Proc. of 2nd Workshop KG4S (ESWC), (2024).
[4] Stack Overflow Blog: Retrieval augmented generation: Keeping LLMs relevant and current, (Oct. 2023).
[5] Dong Shu et al.: Knowledge Graph Large Language Model (KG-LLM) for Link Prediction, arXiv:2403.07311, (2024).
[6] M. Zhang, Y. Chen: Link Prediction Based on Graph Neural Networks, NeurIPS, (2018).
[7] A. Bordes et al.: Translating Embeddings for Modeling Multi-relational Data, NeurIPS, (2013).
[8] Darren Edge et al.: From Local to Global: A Graph RAG Approach to Query-Focused Summarization, arXiv:2404.16130, (2024).