データセットを対象としたマルチモーダルナレッジグラフの生成と活用
Generation and Utilization of Multimodal Knowledge Graphs for Datasets

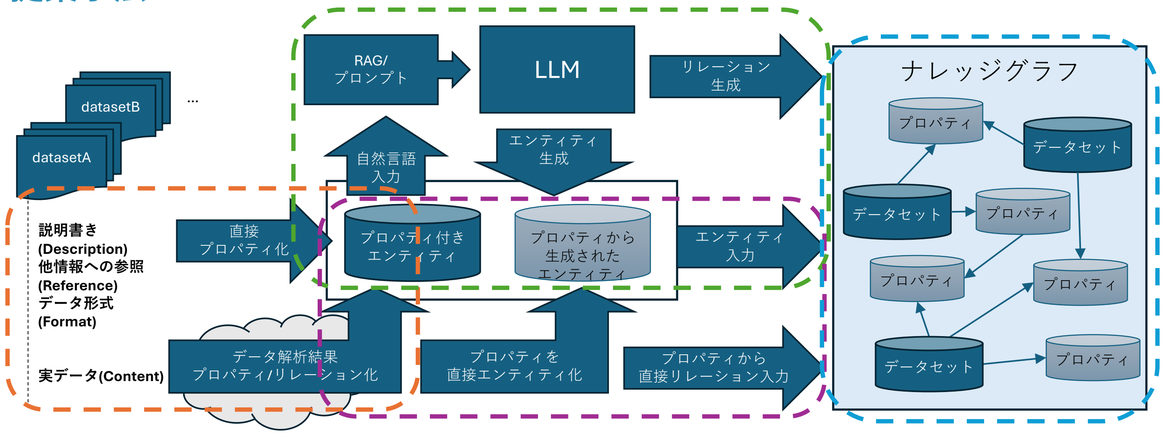
山崎貴史※1, 八重森洋毅※1
※1株式会社コンピュータマインド

Abstract

研究概要
- データを構造化するナレッジグラフ（KG）と自然言語を扱う大規模言語モデル（LLM）を組み合わせる研究が近年盛んに行われている
- 一般的なデータ分析・機械学習への活用を想定したオープンデータセットは多いが, 品質の評価には専門知識が必要で手続き的に選定することが難しい
- データセット選定の自動化のため, LLMを用いたKG生成方法を利用してデータセットをエンティティとするKGを生成する手法を提案する
- 実際のオープンデータセットに付属する説明文や属性データからKGを生成し, サブグラフを抽出してデータセット同士の関係を可視化した

Method

提案手法



① データセットから説明文, 参照, 形式等の情報をプロパティ付きエンティティとして抽出
データセット毎のデータの品質や偏りの情報を加える場合はデータセット毎に
自己回帰等の分析を行ってプロパティを作成することができる
② データセットから直接抽出したプロパティ, エンティティからプロンプトを作成し,
LLMに渡してエンティティ, リレーションを生成
③ データセットの分析結果からルールベースでプロパティ, エンティティを作成し,
②で生成したリレーションとの整合性を取ってKGの強化に用いることができる
④ 生成されたエンティティ, リレーションからKGを構成し, 可視化や検索に用いる
例）ユーザーが指定した属性を持つデータセットを検索する
例）ユーザーが用意したデータセットと類似するデータセットを検索する

Experiments

実験条件
- データセットとしてPhysioNet, NASA Prognostics Data Repository等のデータセット公開サイトから100組を使用
- プロパティとしてURL, 公開サイト, カテゴリ（医療系, 産業系等）, サブカテゴリ（心電図, 脳波, 機械, 電気等）, チャネル数, タスク, 目的を使用
- 下表descriptionが〇の項目はさらに公開サイトによる説明文をプロパティとして追加している
- 下表promptが〇の項目はKG生成のプロンプトをチューニングして実験を行った
- GB10はNvidia製Grace Blackwellアーキテクチャの CPU・GPUの統合チップ, 本研究ではNvidia DGX Sparkに搭載されたものを使用

実験ID	LLM	GPU	description	prompt	結果概要
1	qwen3:4b	RTX 3070ti	×	×	100組から一部（13組）のデータセットとそれらのタスク, カテゴリをエンティティとして検出した
2	qwen3:30b	GB10	×	×	全てのデータセットとプロパティ及びそれらの所属関係をエンティティ, リレーションとして検出した
3	deepseek-r1:32b	GB10	×	×	データセットと関係なく整数IDをエンティティとして検出し, 隣り合ったID間のリレーションを検出した
4	gemma3:27b	GB10	×	×	エンティティとして様々なデータセット, プロパティを検出したがリレーションは全く検出しなかった
5	qwen3:30b	GB10	〇	×	実験2と同じく全てのデータセット, プロパティを所属関係を検出した
6	qwen3:30b	GB10	〇	〇	実験2と同じく全てのデータセット, プロパティを所属関係を検出した

Results

実験6の結果
- 実験6: qwen3:30b, 説明文をプロパティに追加, プロンプトチューニング有り（表1）での結果
- 図1は検出されたKG全体で, 中央付近にデータセットノードが集まり周囲にプロパティノードが分散している
- 図2はPhysioNetをsourceとするデータセットとそのcategory, taskのプロパティを抜き出したサブグラフ
- 抜き出されたエンティティ, リレーション一覧は表2を参照

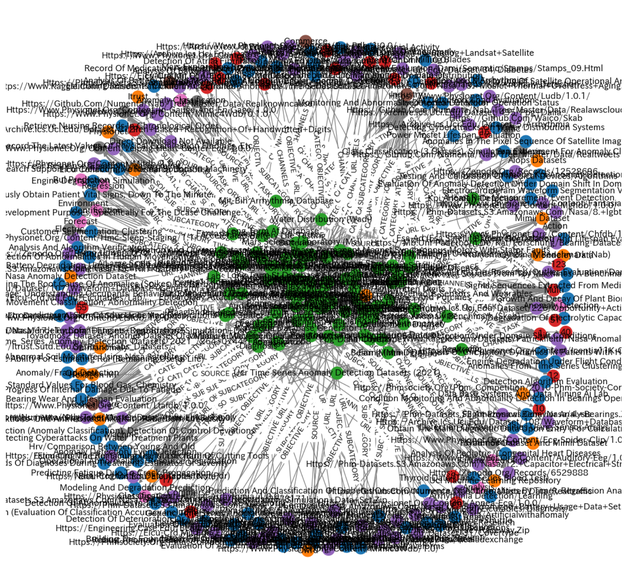


図1. 検出されたKG全体

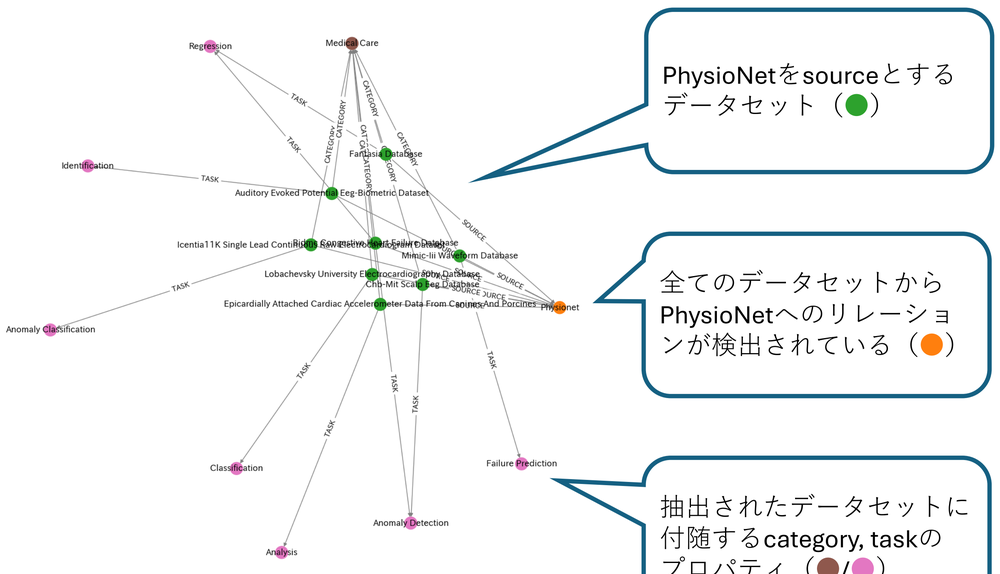


図2. PhysioNetをsourceとするデータセットのサブグラフ

```
"# Knowledge Graph Instructions for {}¥n"
"## 1. Overview¥n"
"You are a top-tier algorithm designed for extracting information in structured formats about various datasets to build a knowledge graph.¥n"
"Try to capture as much information from the text as possible without sacrificing accuracy.¥n"
"Do not add any information that is not explicitly mentioned in the text.¥n"
"The aim is to achieve simplicity and clarity in the knowledge graph, making it accessible for a vast audience.¥n"
"## 2. Nodes and Relationships¥n"
"### Nodes¥n"
"**Labels**: Nodes represent entities and concepts centered around dataset.¥n"
"Ensure you use basic or elementary types for node labels.¥n"
"For example, when you identify an entity representing a category, always label it as '**category**.'¥n"
"Avoid using more specific terms like 'academic' or 'industrial.'¥n"
"**IDs**: Never utilize integers as node IDs.¥n"
"Node IDs should be names or human-readable identifiers like dataset names found in the text.¥n"
"**Maintain Entity Consistency**: When extracting entities, it's vital to ensure consistency.¥n"
'If an entity, such as "Some University Dataset", is mentioned multiple times in the text but is referred to by different names or pronouns (e.g., "SUD", "it"), always use the most complete identifier for that entity throughout the knowledge graph.¥n'
"In this example, use "Some University Dataset" as the entity ID.¥n'
"Remember, the knowledge graph should be coherent and easily understandable, so maintaining consistency in entity references is crucial.¥n"
"### Relationships¥n"
"**Labels**: Relationships represent connections between entities or concepts.¥n"
"Ensure consistency and generality in relationship types when constructing knowledge graphs.¥n"
"Instead of using specific and momentary types such as 'COLLECTED_FOR_TASK', use more general and timeless relationship types like 'TASK'.¥n"
"Make sure to use general and timeless relationship types!¥n"
"## 3. Strict Compliance¥n"
"Adhere to the rules strictly. Non-compliance will result in termination."
```

表1. 使用したKG生成用プロンプト

	source	target	types
0	Bidmc Congestive Heart Failure Database	Physionet	SOURCE
1	Bidmc Congestive Heart Failure Database	Medical Care	CATEGORY
2	Bidmc Congestive Heart Failure Database	Regression	TASK
3	Bidmc Congestive Heart Failure Database	Anomaly Detection	TASK
4	Icentia11K Single Lead Continuous Raw Electrocardiogram Dataset	Physionet	SOURCE
5	Icentia11K Single Lead Continuous Raw Electrocardiogram Dataset	Medical Care	CATEGORY
6	Icentia11K Single Lead Continuous Raw Electrocardiogram Dataset	Anomaly Classification	TASK
7	Chb-Mit Scalp Eeg Database	Physionet	SOURCE
8	Chb-Mit Scalp Eeg Database	Medical Care	CATEGORY
9	Chb-Mit Scalp Eeg Database	Anomaly Detection	TASK
10	Auditory Evoked Potential Eeg-Biometric Dataset	Physionet	SOURCE
11	Auditory Evoked Potential Eeg-Biometric Dataset	Medical Care	CATEGORY
12	Auditory Evoked Potential Eeg-Biometric Dataset	Identification	TASK
13	Mimic-Iii Waveform Database	Physionet	SOURCE
14	Mimic-Iii Waveform Database	Medical Care	CATEGORY
15	Mimic-Iii Waveform Database	Failure Prediction	TASK
16	Fantasia Database	Physionet	SOURCE
17	Fantasia Database	Medical Care	CATEGORY
18	Fantasia Database	Regression	TASK
19	Lobachevsky University Electrocardiography Database	Physionet	SOURCE
20	Lobachevsky University Electrocardiography Database	Medical Care	CATEGORY
21	Lobachevsky University Electrocardiography Database	Classification	TASK
22	Epicardially Attached Cardiac Accelerometer Data From Canines And Porcines	Physionet	SOURCE
23	Epicardially Attached Cardiac Accelerometer Data From Canines And Porcines	Medical Care	CATEGORY
24	Epicardially Attached Cardiac Accelerometer Data From Canines And Porcines	Analysis	TASK

表2. 図2サブグラフのエンティティ・リレーション

Conclusion & Future Work

結論
- データセットからのKG生成とサブグラフ抽出を試し, データセット選定の自動化に繋がる手法が提案された
- 説明文からの単語レベルでのエンティティ生成やプロパティの間のリレーションの抽出のためには
さらに研究が必要である

今後の展望
- データセット毎に実データの分析を行い, その結果をプロパティとして追加する方法及びその自動化の研究
- KG生成及びデータセット検索に対する定量評価方法の検討・実験
- LLMを用いたプロパティ生成やRAGを用いた自然言語による検索等の発展研究

References

論文
[1] Darren E., et al., GraphRAG, arXiv preprint arXiv:2404.16130, 2024.
[2] Zirui G., et al., LightRAG, arXiv preprint arXiv:2410.05779, 2024.

データセットリポジトリ
[3] PhysioNet, https://physionet.org/.
[4] Prognostics Center of Excellence Data Set Repository – NASA, https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/.