様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ３１ 日現在

機関番号：１０１０１
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３４９１

研究課題名（和文）巨大知識グラフに対するクエリ検索の近似的な高速化

研究課題名（英文）Acceleration of Query Search on Large Knowledge Graphs

研究代表者
林　克彦（Hayashi, Katsuhiko）

北海道大学・情報科学研究院・准教授

研究者番号：５０７２５７９４

交付決定額（研究期間全体）：（直接経費）　　7,200,000円

研究成果の概要（和文）：電子情報へのアクセスは現代社会において不可欠である。ここでは知識として集積されたデータへのアクセスを高速かつ柔軟に行うための方法論を確立するための研究に取り組んだ。初年度では、知識データベースから登録事例を高速に検索するため、事例をバイナリベクトル空間で表現し、ビット演算に基づくハミング距離によって高速検索する手法を確立した。一方、2年目以降は、ChatGPTの登場により、分野の状況が一変したため、データベースや検索そのものを見直すこととなったが、画像や音楽などのマルチメディアデータを含めた知識源の検索をGPTの根幹となる大規模言語モデルによってどのように扱うかを検討した。

研究成果の学術的意義や社会的意義
知識データベースは法令、娯楽、化学など世の中の多様な知識を集積しており、その情報へのアクセスは現代社会において必要不可欠な要素となっている。また、ChatGPTのようなツールの基盤となっている大規模言語モデルは情報へのアクセスのあり方を大きく変えるほどの影響力を持っている。そのような言語知識源を活用して、画像、マンガ、音楽などのさまざまな知識源への効果的で新しいアクセス方法を実現しようという試みは社会的意義も大きいと考えている。学術的な意義としては、このような研究の過程において、構築したデータを公開情報として残しており、本研究プロジェクトの後にも引き継いで利用することができることにある。

研究成果の概要（英文）：Access to internet information is essential in modern society. This research focuses on establishing methodologies for fast and flexible access to accumulated data as knowledge. In the first year, a method was established for rapid retrieval of registered examples from a knowledge database by representing fact examples in a binary vector space and using bit operations based on Hamming distance for fast information retrieval. However, from the second year onwards, with the advent of ChatGPT, the landscape of the field changed dramatically, prompting a reconsideration of databases and search methods. We explored how to handle searches for knowledge sources, including multimedia data like images and music, using large-scale language models that underpin GPT.

研究分野：自然言語処理

キーワード：知識グラフ　表現学習　マルチメディアデータ　情報検索　情報推薦

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

(1) 実世界の事実関係を集めた知識グラフは情報化社会における知識インフラとして企業、大学問わず大きな関心を集めている。知識グラフにおける事実は「経営者(David, PestBye)」のようなTriple形式で表され、経営者は述語、DavidやPestByeはエンティティと呼ばれる。Wikipediaや新聞などのテキストデータから自動的に知識獲得を行う技術が発達し、YAGOやGoogle's Knowledge Graphのような数億から数百億規模の事実が登録された巨大知識グラフも開発されている。このような知識グラフは情報検索（推薦・セマンティックWeb）、自然言語処理（質問応答・対話）、社会科学（WWW・ソーシャルネットワークの解析）、データマイニング（一般規則の抽出）、機械学習（ニューラルネットワークの解釈性）など多岐に渡る分野で応用が検討され、情報化社会において必要不可欠となる次世代人工知能（AI）技術の中核を担うことが期待されている。一方、知識獲得の自動化に伴う弊害もあり、巨大知識グラフには膨大な事実の欠損や誤り（ここでは「知識グラフの不完全性」と呼ぶ）が存在することも報告されている。知識グラフの応用を進める上では、この不完全性の問題とどのように向き合うかが重要な課題の1つとなる。

(2) 知識データベースに格納するデータが多様化している。テキスト、音楽、画像など多様な形式のデータを知識化して整備することも重要な研究課題となっている。

２．研究の目的

(1) 知識グラフの不完全性に対処する有効な手段の1つとして、事実が真となる確からしさを確率として扱う方法が考えられる。例えば、Google Knowledge Vaultでは、Webコンテンツや既存の知識グラフの情報を統合する確率モデルを考えることで、事実が真となる確率を与えている。さらに、近年盛んに研究されている知識グラフ埋め込み技術でも、知識グラフにおける欠損の補完問題を確率的なテンソル分解でモデル化し、事実に確率を与えている。これらは関係データベースにおける各タプルに確率を付与することに相当し（図1(a)）、タプル独立型の確率関係データベースとして定義することができる。そして、情報検索や質問応答などの応用も論理式に基づくクエリ検索として見通しよく定式化される。しかし一方で、知識グラフに確率を導入すると、真となる確率が一定値以上の事実を全て考慮する必要があり、データベースへ登録される事実は爆発的に増加するため、クエリ検索やメモリ効率性は著しく損なわれる。

(2) さまざまなマルチメディアデータを知識データベース化することが重要であり、本研究では、マンガデータを対象として考える。マンガは日本から世界へ発信される強力なコンテンツであり、データベースを整備することで、新たな情報検索の応用

を探ることができる。マンガに付与する情報としては、コマやキャラクタ間の関係性などを考える。これらは物語構造を理解するために必要な情報となる。

3．研究の方法

(a)

経営者 (x, y)		確率
David	PestBye	0.6
Elga	KwikEMart	0.9
Fred	Vulgari	0.8

(b)

$X(x)$	確率
経営者	0.3

$Y(x)$	確率
David	0.4
Elga	0.9
Fred	0.7
PestBye	0.5
KwikEMart	0.8
Vulgari	0.6

図1．(a) 確率関係データベースの例、 (b) (a)の二項関係テーブルを擬似的な単項関係テーブルX，Yに分解した例．

(1) 本研究では、巨大知識グラフに対する確率の導入方法を捉えなおし、メモリの効率性に優れ、高度なクエリ検索を可能にする確率関係データベースの構成方法について考える。 本構想の根幹となるアイデアは、図1(a)のような多項関係を図1(b)のように単項関係へと近似的に分解する学習モデルを新たに提案することにある。このとき、N項関係のタプルt= r(e1，e2，...，eN)に付与された確率は$P(t) \approx P(X(r)) \cdot P(Y(e1)) \cdot P(Y(e2)) \cdot \ldots \cdot P(Y(eN))$ というように新たに導入された各単項関係X、Yにおけるタプルの確率の積で近似することが単純には考えられる。例では、単項関係のタプルに確率を与えたが、これは確率を要素としたベクトルに拡張することもでき、この学習モデルは特殊なテンソル分解ともみなせる。このような分解を行うことで、確率関係データベースは述語とエンティティの単項関係テーブルのみで構成され（図1(b)）、確率化に伴う事実の増加を防ぐことができる。また、後述するように、単項化はクエリ検索の効率性の面でも大きな利点がある。

図2．マンガへの読み順や物語構造のアノテーションツール．

(2) Manga109と呼ばれるマンガデータが公開されているが、これには読み順に関する情報が付与されていない。マンガを知識として記述して、データベース化するにはその物語構造の理解は不可欠であると考えられる。そこで、マンガの読み順情報や物語構造をアノテーションするために必要となるツール開発の開発を行う。このツールはブラウザアプリとして作成する（図2）。将来的には、コマやキャラクタレベルの関係性を記述し、これらをデータベース化する。

４．研究成果

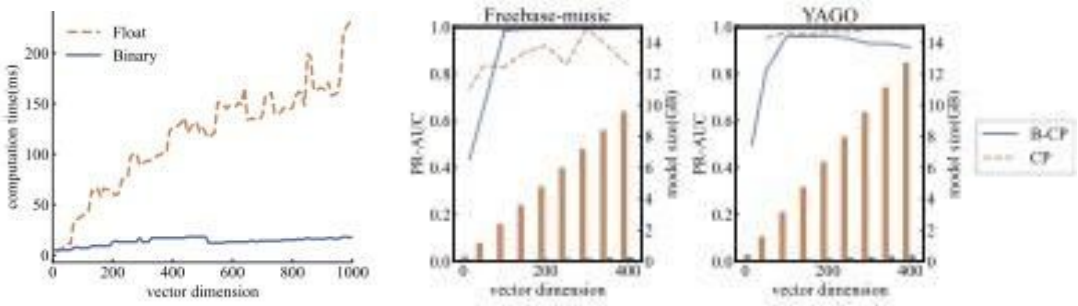

図3．左図は提案手法の検索速度、右図はメモリ消費量を表す．

(1) 知識グラフを潜在的な特徴空間に埋め込み、知識を高速に検索するための手法を開発した。この研究では、0と1の値のみを取る特殊な確率値を埋め込みベクトルの要素として考えるモデルとなっており、研究目的として考えていた内容を達成する成果となっている。また、0-1の値はビット演算を用いて高速に処理できるため、検索速度と消費メモリ量を大幅に向上させることに成功した（図3）。この成果を論文化し、IEEE TKDEに採録された。

Loss	Objective Distribution	$\Psi(z)$ or $\Psi(\mathbf{z})$	Remarks
NS w/ Uni	$p_d(y\|x)$	$\Psi(z) = z\log(z) - (1+z)\log(1+z)$	
NS w/ Freq	$T_{x,y}^{-1} p_d(y\|x)$	$\Psi(z) = z\log(z) - (1+z)\log(1+z)$	$T_{x,y} = p_n(y\|x) \sum_{y_i \in Y} \frac{p_d(y_i\|x)}{p_n(y_i\|x)}$
SANS	$(1-\lambda)p_d(y\|x) + \lambda\omega\{1, \|Y\|\}$	$\Psi(z) = z\log(z) - (1+z)\log(1+z)$	Approximately derived. λ increases from zero in training.
SCE	$p_d(y\|x)$	$\Psi(\mathbf{z}) = \sum_{i=1}^{len(\mathbf{z})} z_i \log z_i$	
SCE w/ BC	$T_{x,y}^{-1} p_d(y\|x)$	$\Psi(\mathbf{z}) = \sum_{i=1}^{len(\mathbf{z})} z_i \log z_i$	$T_{x,y} = p_n(y\|x) \sum_{y_i \in Y} \frac{p_d(y_i\|x)}{p_n(y_i\|x)}$
SCE w/ LS	$(1-\lambda)p_d(y\|x) + \lambda\omega\{1, \|Y\|\}$	$\Psi(\mathbf{z}) = \sum_{i=1}^{len(\mathbf{z})} z_i \log z_i$	λ is fixed.

図4．自己敵対的な負例サンプリング法に基づく損失（SANS）の一般的な性質．

(2) (1)で開発した埋め込みモデルを効率的に学習する方法についても研究を行った。Softmax Cross Entropy損失に対して、自己敵対的な負例サンプリング法を考えることで、パラメータの発散を抑えることに成功している。また、この損失の一般的な性質を明らかにし、他の損失関数との違いを整理している（図4）。これは国際会議AC

L21や国際会議ICML22に採択された。また、言語処理学会第28回年次大会で優秀賞を受賞した。

(3) 21年度から進めていたマルチメディアデータベースの整備として、漫画の読み順データセット、漫画キャラクタのヴィジュアルエンティティリンキングデータセットなどの構築を進めた。この成果の一部は2023年3月開催の情報処理学会・第67回EC研究会で発表しており、学生優秀賞、研究奨励賞を受賞した。作成したデータセットの一部はgithub https://github.com/mu-perori/Manga109_AnnotationApp で公開済みとなっている。また、この成果の一部は、ITMedia NEWSに取り上げられ、約12のウェブ ニュースサイトで報道された。

(4) 22年度から行っていた知識を用いた画像生成（あるいは逆の画像から知識データベースの生成）に 関わるデータセットをWikipediaから構築した。この成果は2023年7月開催に開催された国際会議ACL23に採択された。また、知識と画像の併用に関する研究とし て、画像の批評文生成、および、芸術画像の説明文生成に関する課題に取り組んだ。どちらもWikipediaから知識と画像に関わる情報を抽出することで実施した。 それぞれ2024年3月に開催された言語処理学会年次大会で発表を行い、芸術画像の説明文生成に関する研究は国際会議ACL24に採択された。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Hayashi Katsuhiko、Kishimoto Koki、Shimbo Masashi	４．巻 35
２．論文標題 Binarized Embeddings for Fast, Space-Efficient Knowledge Graph Completion	５．発行年 2023年
３．雑誌名 IEEE Transactions on Knowledge and Data Engineering	６．最初と最後の頁 141 - 153
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TKDE.2021.3075070	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計13件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Katsuhiko Hayashi
２．発表標題 Rethinking correlation-based item-item similarity for recommender systems
３．学会等名 Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval（国際学会）
４．発表年 2022年

１．発表者名 Hidetaka Kamigaito、Katsuhiko Hayashi
２．発表標題 Comprehensive analysis of negative sampling in knowledge graph representation learning
３．学会等名 Proceedings of the 39th International Conference on Machine Learning（国際学会）
４．発表年 2022年

１．発表者名 竹内皓紀、林克彦
２．発表標題 Wikipedia編集者情報を用いた協調フィルタリングによるエンティティ類似度推定
３．学会等名 第18回Webインテリジェンスとインタラクション研究会
４．発表年 2022年

１．発表者名
竹内皓紀、林克彦

２．発表標題
Wikipedia協調フィルタリング法

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
飯塚柚稀、林克彦、永野清仁、宮尾祐介

２．発表標題
服飾の色情報に基づいた ポエティックな商品名の作成支援システム

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
上垣外英剛、林克彦、渡辺太郎

２．発表標題
視覚と言語の融合モデルにおける知識の振る舞いを調査するための表と画像の生成タスクの提案及びその調査結果

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
Xincan Feng、上垣外英剛、林克彦、渡辺太郎

２．発表標題
知識グラフ補完のためのモデル予測に基づくサブサンプリング

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
吉永瑛哉、林克彦、鷲尾 光樹、上垣外 英剛、新保 仁

２．発表標題
マンガのコマ割りのみから作品分類は可能か?

３．学会等名
情報処理学会第67回エンターテイメントコンピューティング研究会

４．発表年
2023年

１．発表者名
上原瑞歩、倉石怜実、林克彦、鷲尾光樹、上垣外英剛

２．発表標題
マンガの読み順評価データセットの開発

３．学会等名
情報処理学会第67回エンターテイメントコンピューティング研究会

４．発表年
2023年

１．発表者名
Hidetaka Kamigaito, Katsuhiko Hayashi

２．発表標題
Unified Interpretation of Softmax Cross-Entropy and Negative Sampling: With Case Study for Knowledge Graph Embedding

３．学会等名
In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (国際学会)

４．発表年
2021年

１．発表者名
上原瑞歩，鷲尾光樹，林克彦，上垣外英剛，木曽鉄男，小田悠介

２．発表標題
漫画の読み順データセット公開に向けた調査

３．学会等名
第6回コミック工学研究会

４．発表年
2021年

１．発表者名 上垣外英剛，林克彦
２．発表標題 知識グラフ埋め込みにおける負例サンプリング損失の分析
３．学会等名 言語処理学会第28回年次大会
４．発表年 2022年

１．発表者名 Hidetaka Kamigaito, Katsuhiko Hayashi, Taro Watanabe
２．発表標題 Table and Image Generation for Investigating Knowledge of Entities in Pre-trained Vision and Language Models
３．学会等名 In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics（国際学会）
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	能地　宏 (Noji Hiroshi) (00782541)	国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員 (82626)	削除：2021年10月14日

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関