RAG を活用した研究アイデア発想支援システム

今村 翔太 * 　　暦本 純一 * †

概要. 本研究では，研究アイデアの概要を入力することで関連研究データベースの中から近い研究をベクトル検索により抽出し，大規模言語モデルを用いて過去の研究動向や入力したアイデアとの関係を整理するシステムを提案する．本システムにより，類似点や相違点を容易に把握できるようになり，ユーザーは情報収集にかかる時間や労力を削減できる．近年，研究分野の進展が急速化し，特に人工知能など情報通信分野では過去の全研究を把握することが困難となってきている．本システムを活用することで，こうした急速に発展する分野でも，関連する研究を体系的に理解し，着実に新規研究に着手できるよう支援する．今後は，学会や大学の研究室など新しいアイデアの議論の場で本システムを活用してもらうことで，情報支援システムとしての有用性や効果について調査を行う予定である．

1 はじめに

研究者や技術者にとって，過去の研究や発明の動向を的確に把握することは，新規のアイデアを具体化し，独創的な知見を得たり，新しい発明を行うための重要なステップである．また，過去の知見を活かして研究や開発を効率的に行うこともできる．しかし，例えば人工知能や情報通信分野といった分野では，技術の進展速度が加速しており，膨大な研究成果が次々と生み出されている．そのため，分野全体の研究動向を理解したり，その中から自らのアイデアに近い研究を見つけることが困難となりつつある．

近年，埋め込みベクトル [4] や大規模言語モデル（LLM）[2]，それらを組み合わせた検索拡張生成（Retrieve-Augment-Generate; RAG）[3] といった技術が，情報の検索および生成の分野で注目を集めている．埋め込みベクトルは，テキストやその他のデータを高次元の空間にマッピングすることにより，データの類似性を数値的に表現する技術であり，意味的に類似した情報を効果的に抽出できる．この埋め込みベクトルを用いてベクトルの近傍検索を行うベクトル検索は，膨大な研究成果の中からユーザーの意図に近い情報を効率的に見つけ出すために有用である．また，RAG は，検索と大規模言語モデルによる生成を組み合わせた情報探索手法である．ベクトル検索等によりデータベースから関連情報を検索して大規模言語モデルに与えるコンテクストとして補完することで，大規模言語モデル単体で用いるよりも情報の精度と網羅性を向上させた関連する情報の生成や要約が可能である．

本稿では，関連研究の調査と内容の把握の情報処理負荷を低減するために，関連研究の調査と整理に特化した RAG のシステムを構築した．研究アイデアの概要を入力することで，関連する過去の研究をベクトル検索により抽出し，さらに大規模言語モデルを活用してそれらの研究動向や入力した研究アイデアとの関連性を整理するシステムを提案する．本システムは，類似点や相違点を把握しやすい形で提示することにより，情報収集の負担を低減し，研究者が新しいアイデアを検討し，新しい取り組み始めることを支援する．

本システムは学会や大学の研究室といった新たなアイデアの議論が活発に行われる環境において，その有用性を最大限に発揮すると期待される．今後は，学会や大学の研究室など新しいアイデアの議論の場で本システムを活用してもらうことで，情報支援システムとしての有用性や効果について調査を行なっていく．

2 実装

本稿で実装したシステムは研究概要を入力すると関連研究を調査し，調査結果の出力を行う（図 1）．入力は専用の Web サイトもしくは Slack Bot を招待した Slack チャンネルから行うことができる．本システムでは関連研究の調査はキーワード検索とベクトル検索を組み合わせて行う．

まずは対応する分野の主要な学会の予稿集から論文 PDF を取得し，それらを事前に OpenAI の埋め込みモデル [5] を仕様してベクトルデータベースを構築する．また，ユーザーが研究概要を入力する度に，入力された研究概要から OpenAI の大規模言語モデル GPT-4o [6] を用いて検索クエリを生成し，プレプリントの主要なデータベースである ArXiv [1] の API を使用して関連研究の検索を行う．複数の検索クエリの組み合わせを行い取得した概要をさらに埋め込み処理を行い，ベクトルデータベースに追加

Copyright is held by the author(s). This paper is non-refereed and non-archival. Hence it may later appear in any journals, conferences, symposia, etc.

* 東京大学
† Sony CSL

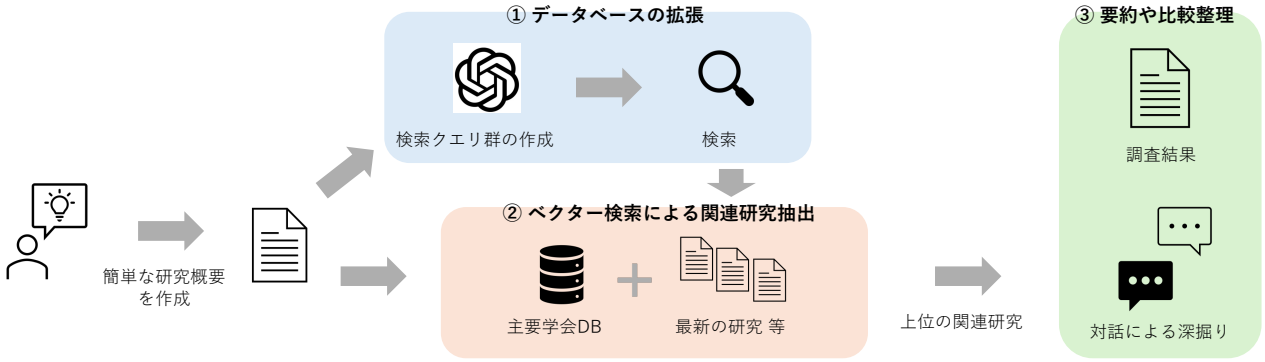

図 1. 関連研究の検索や比較整理を提示するまでの処理の流れ .1) ユーザーの入力した研究概要に近い関連研究をデータベースに追加する .2) 事前に構築した主要学会の関連研究 DB に最新を研究を追加し拡張されたデータベースにてベクトル検索を行い，関連研究を抽出する .3) 大規模言語モデルを用いて抽出された関連研究群をまとめ，ユーザーの入力した新しい研究アイデアの概要を比較整理する．さらにユーザーが追加で対話を行い深掘りを行うこともできる．

を行う．これにより，最新の研究や対象とする学会以外に関連した研究がある場合にも対象とすることができる検索データベースを効率良く構築する．この処理はリアルタイム性を確保するため，サーバーにおいて並列処理を行い数十秒程度で処理が完了するように実装している．また，キャッシングを行い，利用者の中で注目の集まっている分野については追加処理が早く完了するように実装した．

ベクトルデータベースの追加処理が完了したのち，入力された研究概要を埋め込んだベクトルを用いて，関連研究のベクトルデータベースから類似度の高いものを抽出する．抽出された上位の研究論文群を大規模言語モデルを用いてまとめると共に，入力された新しい研究アイデアと近い点や違う点を整理してユーザーに提示する．また，特に気になった関連研究や観点について関連研究のコンテクストを把握した大規模言語モデルと対話を行える．

本システムにより，ユーザーは思いついた研究アイデアの最新の研究動向について速やかに調査を行い，研究アイデアのブラッシュアップを効果的に行うことができる．

3 今後の展望

本システムは学会や大学の研究室といった新たなアイデアの議論が活発に行われる環境において，その有用性を最大限に発揮すると期待される．特に現在は Human Computer Interaction の研究分野における主要な学会の予稿集をもとに作成したデータベースを構築している．そこで著者の研究室内での検証と改善や，WISS Challenge での提供を行う．また，Web サービスとして一般公開を行い，広く実際に使用してもらうことを検討している．実際の研究アイデアの創発やブラッシュアップに活用してもらいフィードバックをいただくことで，情報支援システムとしての有用性や効果について調査を行なっていきたい．

謝辞

本研究は，JST ムーンショット型研究開発事業グラント番号 JPMJMS2012 の支援を受けたものです．

参考文献

[1] arXiv. arXiv. `https://arxiv.org`, 2024.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin eds., *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[5] OpenAI. Embeddings. `https://platform.openai.com/docs/guides/embeddings`, 2024.

[6] OpenAI. GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.