ナレッジグラフ RAG における語彙的マッチングを用いた検索精度の改善

高見昂季 [1]　滝口哲也 [2]　有木康雄 [2]　岩﨑雄介 [3]
小島幹 [3]　山本昌輝 [3]
[1] 神戸大学工学部情報知能工学科　[2] 神戸大学大学院システム情報学研究科
[3] 株式会社デンソーテン
2285043t@stu.kobe-u.ac.jp　{takigu, ariki}@kobe-u.ac.jp
{yuusuke.iwasaki.j7p, motoki.kojima.j3h, masaki.yamamoto.j3h}@jpgr.denso.com

概要

従来のナレッジグラフ RAG は埋め込みベクトルによる検索が主流であり，語彙的マッチングの欠如による検索漏れが懸念される．本稿では，ナレッジグラフ内の構造化データに対して BM25 によるパス重み付けを行い，グラフ検索の精度を改善する手法を提案する．さらに，グラフ検索と BM25 検索を統合するハイブリッド検索を行うことで，直接的な語彙マッチングによる更なる精度改善を実現する．評価実験の結果，提案手法は従来手法と比較して高い検索精度を達成し，その有効性が確認された．

1　はじめに

近年，大規模言語モデル（LLM）に外部知識を与える手段として Retrieval-Augmented Generation（RAG）が広く用いられている．とりわけ，知識をエンティティと関係からなるグラフとして表現できるナレッジグラフ（KG）は，エンティティ間の関係を明示的に表現できるため，クエリに関連する根拠（文書・事実）を探索・整理しやすく，KG に基づく RAG（KG-RAG）が注目されている．

一方で，KG-RAG の検索は，起点ノードの同定，重要パスの探索，チャンク選定といった各段階で埋め込みベクトルに基づく類似度計算に依存する設計が一般的である．しかし，ベクトル検索に偏った検索では，固有名詞や専門用語などに対する語彙的マッチングの観点が不足し [1]，必要なノードやパスが候補から漏れる可能性がある．特に，厳密な一致が求められる製品名や人名などの固有名詞を含むクエリにおいては，意味的な類似性だけでは捉えきれず，検索精度の低下が顕著となり，結果として十分な回答生成が困難になるという報告もある [1, 2, 3]．

そこで本研究では，KG に保存された構造化データあるいはチャンクに対し，BM25 を用いた語彙的マッチングによるスコアリングを導入し，グラフ検索の精度を改善する手法を提案する．具体的には，クエリとパスの語彙的整合性に基づいてパスへ重み付けを行い，探索・ランキングに反映させることで，ベクトル表現のみでは拾いにくい関連候補を上位に押し上げる．さらに，提案するグラフ検索と BM25 検索の結果を統合するハイブリッド検索を行うことで，より強化された検索を実現する．

実験の結果，提案する BM25 によるパス重み付けは従来の KG-RAG に対して検索精度を向上させ，ハイブリッド検索により更なる改善が得られることを確認した．

2　関連研究

2.1　NaiveRAG

RAG は，外部コーパスから取得した文脈を入力条件として生成を行う枠組みであり，知識集約的タスクなどにおいて性能向上や根拠提示に寄与しうる．Lewis らは，密検索器により関連文書を取得し，その文書に条件づけて系列生成を行う RAG モデルを提案し，多様な知識集約的 NLP タスクで有効性を示した [4]．本稿では，ベースラインとして「クエリに対し文書チャンクを検索し，上位 k 件を LLM への入力として生成する」単純な RAG を想定し，とりわけ検索器が埋め込みベクトルによる密検索のみに基づく場合を NaiveRAG と呼ぶ．NaiveRAG は実装が容易で汎用的である一方，固有名詞や専門用語などの語彙的手がかりが十分に効かず，関連情報の

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

検索漏れが生じうる．

2.2 GraphRAG

Edge らは，大規模なテキストコーパス全体の理解を必要とするタスクにおいて，従来のベクトルベースの RAG の限界を指摘し，グラフ構造を活用した GraphRAG [5] を提案した．彼らのアプローチは，ソースドキュメントからエンティティと関係性を抽出してナレッジグラフを構築し，Leiden アルゴリズムを用いて階層的なコミュニティ構造を作成する点に特徴がある．

GraphRAG は，検出された各コミュニティに対して LLM で要約を生成し，これをインデックスとして保持する．ユーザーからの質問に対しては，これらのコミュニティ要約を Map-Reduce 形式で統合して回答を生成することで，従来の RAG よりも網羅性（Comprehensiveness）と多様性（Diversity）に優れた結果を示している．

2.3 LightRAG と MiniRAG

また，検索の効率化や軽量化を目指した手法も提案されている．LightRAG [6] は，グラフ構造を使用した低レベル・高レベルの二段階検索により，検索の精度と効率の両立を狙う．さらに，データの増分更新に対応しており，新しいデータが追加されるたびにグラフ全体を再構築する必要がないため，運用コストを大幅に削減できる利点がある．

MiniRAG [7] は，小規模言語モデル（SLM）やエッジデバイスでの利用を想定した，さらに軽量なフレームワークである．SLM を用いても LLM ベースの手法と同等の精度を維持しつつ，ストレージ使用量を LightRAG 等の既存手法と比較して約 25%に抑えることに成功している．

3 提案手法

本節では，KG-RAG における検索段階に対し，語彙的マッチングを導入して検索漏れを抑制する手法を提案する．本研究の検索処理は MiniRAG のプログラム実装をベースに設計し，その上で語彙的マッチングに基づく拡張を加える．

3.1 パス重み付けの現状

汎用的な埋め込みモデル（Embedding Model）は，単語の意味的な類似性を捉えることには長けているが，学習データに含まれないドメイン固有の専門用語や，厳密な一致（Exact Match）が求められる製品名・人名などの固有名詞に対しては，適切なベクトル表現を生成できない場合が多い．特に，構造化データにはこうした固有名詞が多く含まれるにもかかわらず，ベクトル化によってその特異性が希釈されてしまうと，ユーザーが本当に知りたい特定のエンティティへと続くパスが，ノイズとして切り捨てられてしまう可能性がある．

3.2 BM25 の概要

本研究では，語彙ベース検索の代表的手法である BM25 スコアを用いて，クエリ q と文書 d（本稿ではエッジテキストやチャンク）との語彙的一致度を数値化する．BM25 は TF-IDF を改良したスコアリング手法の一種であり，語の出現回数（TF）に対して飽和を入れつつ，文書長に基づく正規化を行うことで，文書長の影響を制御する [8]．

BM25 の標準形は，クエリ q に含まれる語 t についての和として次式で与えられる：

$$\mathrm{BM25}(q,d) = \sum_{t \in q} \mathrm{IDF}(t) \cdot \frac{\mathrm{tf}(t,d)}{K + \mathrm{tf}(t,d)} \tag{1}$$

$$K = k_1 \left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right) \tag{2}$$

ここで，$\mathrm{tf}(t,d)$ は文書 d における語 t の出現回数，$|d|$ は文書長，avgdl はコーパス全体の平均文書長である．$k_1 > 0$ および $0 \le b \le 1$ はハイパーパラメータであり，それぞれ TF の飽和度と文書長正規化の強さを制御する．また，$\mathrm{IDF}(t)$ は語 t の逆文書頻度であり，

$$\mathrm{IDF}(t) = \log \frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5} \tag{3}$$

のように定義される（N は文書数，$\mathrm{df}(t)$ は語 t を含む文書数）．本稿では，クエリおよび検索対象テキストを単語列へ分割し，その上で BM25 を計算する．

3.3 BM25 によるパス重み付け

前節で述べた BM25 を，MiniRAG のグラフ探索で得られる候補パスの順位付けに組み込み，語彙的マッチングの手がかりを追加する．狙いは「クエリと語彙的に強く対応するエッジを重要パスとして抽出し，そのエッジを含むパスに重みを付けることで最終チャンク選定に活用する」ことである．なお，実際の重み付け手法は従来手法である MiniRAG と同様の手法を用いる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

まず，従来手法（ベクトル検索）の検索設定について述べる．前提として，グラフ構築時に各エッジは属性として「キーワード」「両端エンティティ」「説明文（関係性の説明）」を持つ．検索時は以下の 2 者間で類似度計算を行う．

- 検索クエリ：入力テキストの埋め込みベクトル
- 検索対象：エッジ情報（キーワード+両端エンティティ+説明文）の埋め込みベクトル

上記検索により，類似度上位のエッジを重要候補として扱う．

一方，本研究で導入する手法は，埋め込みを介さず BM25 を用いることで，クエリと文書の語彙一致に基づいてスコアリングを行う．具体的な検索設定を以下に 2 通り示す．

- **手法 1**:
  - 検索クエリ：入力テキスト
  - 検索対象：エッジ情報（キーワード+両端エンティティ+説明文）
- **手法 2**:
  - 検索クエリ：入力テキスト
  - 検索対象：エッジ両端ノード（エンティティ）が保持するテキストチャンク

本稿では，従来手法とこれら 2 つの設定を比較評価するとともに，エッジ情報とコンテンツ（チャンク）のいずれが語彙的マッチングの検索対象として適しているかを明らかにする．

3.4 BM25 のハイブリッド統合

本手法では，3.3 節で述べたグラフ検索（ベクトル検索または BM25 重み付けを適用したグラフ探索）に加え，グラフ構造を用いない独立した BM25 検索を併用し，両者の結果を統合するハイブリッド検索を行う．

具体的には，以下の 2 つの検索結果を統合する．

1. **グラフ検索**: 3.3 節の手法に基づき，グラフ構造を探索して得られたテキストチャンク集合．
2. **BM25 検索**: グラフ探索とは独立して，クエリと文書（チャンク）との語彙的類似度のみに基づいて抽出されたテキストチャンク集合．

これら 2 つのテキストチャンク集合に対し，順位に基づく統合手法である Reciprocal Rank Fusion（RRF）[9] により統合後の順位を計算する．一般的な RRF は複数の検索結果の総和として定義されるが，本手法では 2 つの検索結果を統合するため，次式を用いる．

$$\mathrm{RRF}(q, c) = \frac{1}{k + \mathrm{rank}_{\mathrm{graph}}(c)} + \frac{1}{k + \mathrm{rank}_{\mathrm{bm25}}(c)} \quad (4)$$

ここで c は，検索結果のチャンクであり，$\mathrm{rank}_{\mathrm{graph}}(c)$ および $\mathrm{rank}_{\mathrm{bm25}}(c)$ は，それぞれグラフ検索・BM25 検索での最終検索結果における c の順位（1 が最上位）を表す．k は RRF の定数（$k > 0$）であり，上位への寄与の強さを調整する．なお，ある検索器で c が候補に含まれない場合は，当該項の寄与を 0 とみなす．RRF スコア上位 K 個のチャンクを選択し，最終的な検索結果とする．

また，本研究では RRF による統合に加え，リランクによる統合も検討する．具体的には，グラフ検索と BM25 検索で得られたチャンクの和集合に対し，Qwen3-Reranker-4B を用いてクエリとの関連度を再計算し，そのスコアに基づいて最終的な順位を決定する．RRF が順位情報のみを用いるのに対し，リランカーはチャンクの意味内容を直接考慮してスコアリングを行うため，より高精度な統合が期待される．

4 実験

本節では，提案手法（BM25 によるパス重み付け・ハイブリッド統合）が，KG-RAG における検索漏れを抑制し，検索精度を改善できるかを検証する．評価は「クエリへの回答に必要な情報を上位で検索できるか」に焦点を当て，SLM，LLM による最終回答生成は本稿の範囲外とする．

4.1 データセット

本実験では，MiniRAG [7] で構築された，ローカル RAG シナリオ向けのデータセットである LiHua-World を使用する．本データセットは，仮想ユーザー LiHua の 1 年分（1 月から 12 月）のチャット記録で構成されているが，本実験ではそのうち半年分（1 月から 6 月）を使用した．データには社会的交流，フィットネストレーニング，娯楽活動，生活事務など，日常生活の多岐にわたるトピックが含まれている．

具体的には，時系列順に整理されたチャットメッセージ（タイムスタンプ，送信者，メッセージ内容）と，評価用の QAE（Question-Answer-Evidence）ペアからなる．実験では，全てのチャット記録をナレッジベースとして構築し，QAE データセットの Q に

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

対し，正解となる E を検索できたかどうかに基づいて検索精度の評価を行う．

4.2 比較対象

提案手法の有効性を検証するため，以下の手法と比較を行う．

- **NaiveRAG**: クエリと各チャンクの埋め込みベクトル間の類似度計算により上位 k 件を取得する．グラフを使用しない最も単純な RAG 手法．
- **MiniRAG**: 3.3 節で述べた従来手法．重要パスの探索はベクトル検索により行われる．
- **BM25 によるパス重み付け（提案）**: 3.3 節で述べた提案手法．重要パスの探索は BM25 を用いた語彙的マッチングにより行われる．BM25 の検索対象として，以下の 2 種類を比較する．
  - **エッジ情報**: エッジ情報（キーワード+両端エンティティ+説明文）を BM25 の検索対象とする．
  - **チャンク**: エッジ両端のエンティティが保持するテキストチャンクを BM25 の検索対象とする．
- **グラフと BM25 のハイブリッド（提案）**: グラフ検索と BM25 検索を独立に実行し，RRF またはリランクにより候補を結合する．

4.3 評価指標

各クエリに対し，上位 K 個の検索結果（テキストチャンク）を取得したときに正解根拠をどれだけ含むかを評価する．評価には以下の指標を用いる．

- **Recall@K**: 正解根拠が上位 K 個の検索結果に含まれる割合．
- **nDCG@K**: 検索結果の順位を考慮した評価指標．正解根拠がより上位に出現するほど高いスコアとなるよう，順位に基づく減衰を加えて算出する．

4.4 結果

表 1 に結果を示す．提案手法である BM25 によるパス重み付けは，ベースラインである NaiveRAG および MiniRAG と比較して，Recall@10 および nDCG@10 の双方で精度が向上した．特に，エッジ情報を対象とした BM25 パス重み付けにハイブリッド検索（Rerank）を組み合わせた手法が最も高い精度を示した．また，BM25 の検索対象としてエッジ情報を用いた場合とチャンクを用いた場合を比較すると，エッジ情報を用いた場合の方が全体的に高い性能が得られた．

表 1 提案手法の検索精度比較

手法	Recall@10	nDCG@10
NaiveRAG	0.636	0.532
MiniRAG	0.675	0.479
手法 1：エッジ情報	0.755	0.575
+ハイブリッド (RRF)	0.808	0.695
+ハイブリッド (Rerank)	**0.861**	**0.814**
手法 2：チャンク	0.748	0.651
+ハイブリッド (RRF)	0.795	0.713
+ハイブリッド (Rerank)	0.834	0.798

4.5 考察

実験結果より，従来のベクトル検索に依存した手法（NaiveRAG, MiniRAG）と比較して，BM25 による語彙的マッチングを導入した提案手法が優れた検索精度を示した．これは，ベクトル検索に偏った検索では，固有名詞や専門用語などに対する語彙的マッチングの観点が不足しがちであるのに対し，BM25 がその欠点を補完できたためと考えられる．特に，ユーザーのクエリに含まれる具体的なキーワードが，ナレッジグラフ内のエンティティやエッジの属性と直接マッチすることで，関連性の高いパスを効果的に探索できたと言える．

また，BM25 の検索対象として「チャンク」よりも「エッジ情報」を用いた場合の方が良好な結果となった．この要因として，エッジ情報はグラフ構造に由来する構造化データであり，ハイブリッド検索においてチャンクを対象とする検索とは異なる視点を提供できたことが挙げられる．すなわち，パス重み付けにチャンクを用いた場合（手法 2），ハイブリッド検索側の BM25 検索と検索対象が重複するため，統合による相補的な効果が得られにくかったと考えられる．

5 おわりに

本稿では，ベクトル検索に偏りがちな KG-RAG の検索に対し，語彙的マッチングを導入して検索精度を向上させる手法を提案した．これにより，従来のベクトル検索では捉えにくい語彙的な特徴を考慮した検索が可能となった．今後の展望として，(1) 提案手法を他の KG-RAG フレームワークに適用し，汎用性を検証する，(2) 検索精度向上が最終的な回答生成に与える影響を評価する，などが挙げられる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pages 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 6769–6781. Association for Computational Linguistics, 2020.

[3] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. **arXiv preprint arXiv:2211.14876**, 2022.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. **CoRR**, abs/2005.11401, 2020. arXiv:2005.11401 (Accepted at NeurIPS 2020).

[5] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. **CoRR**, abs/2404.16130, 2024. arXiv:2404.16130.

[6] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. **CoRR**, abs/2410.05779, 2024. arXiv:2410.05779.

[7] Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. MiniRAG: Towards extremely simple retrieval-augmented generation. **CoRR**, abs/2501.06713, 2025. arXiv:2501.06713.

[8] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In **Overview of the Third Text REtrieval Conference (TREC-3)**, pages 109–126. Gaithersburg, MD: NIST, January 1995.

[9] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, **Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009**, pages 758–759. ACM, 2009.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　実験設定の詳細

本実験で使用した詳細なパラメータおよび環境を以下に示す.

- **Embedding Model**: sentence-transformers/all-MiniLM-L6-v2
- **LLM (Entity Extraction)**: Qwen3-1.7B
- **BM25**: $k_1 = 1.5, b = 0.75$
- **RRF**: $k = 60$

B　ケーススタディ

本節では，ベースライン手法（MiniRAG）では検索に失敗し，提案手法（BM25 パス重み付け）によって正解根拠を検索可能となった具体的な事例を紹介する.

ケース 1：''Freelancer Group Meeting''

Question: When is the Freelancer Group Meeting scheduled for according to the conversation...
Evidence: LiHua: ... attend the Freelancer Group Meeting this Wednesday at 3 pm?
分析: "Freelancer Group Meeting" という具体的なイベント名が含まれているが，ベクトル検索では一般的な会議の文脈に埋もれてしまい，正確な日時を含むチャンクを上位に取得できなかったと考えられる．一方，提案手法では "Freelancer", "Group", "Meeting" といった語彙の重なりを捉えることで，当該イベントに言及しているエッジおよびチャンクを正しく特定できた.

ケース 2：''Overwatch 3''

Question: What time does Li Hua watch the movie "Overwatch 3"?
Evidence: ... "Overwatch 3" ...
分析: 架空あるいは特定の作品名である "Overwatch 3" は，汎用的な埋め込みモデルではその固有性を十分に表現できない場合がある．提案手法では，クエリ内の "Overwatch 3" とナレッジグラフ内の同単語が直接マッチすることで，関連する視聴時間に関する記述をピンポイントで検索することに成功した.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).