NumColBERT: マルチベクトル検索パイプラインを維持した数値条件対応検索モデルの提案

藤巻晴葵 [1]　加藤誠 [2,3]
[1] 筑波大学大学院　[2] 筑波大学 図書館情報メディア系　[3] 国立情報学研究所
fujimaki.haruki.tkb_eg@u.tsukuba.ac.jp　mpkato@acm.org

概要

数値条件を含むクエリに対する密検索において，既存モデルは性能が低く，専用モジュールを用いる従来手法は既存検索パイプラインとの統合が困難である．本研究では，既存の検索パイプラインに影響を与えずに数値条件検索の性能を向上させるNumColBERT を提案し，数値トークン検出・数値情報予測・対照学習・埋め込み設計を導入した．実験の結果，提案手法は通常の微調整手法と比較して数値条件検索性能を改善し，専用モジュールを用いる従来手法と同等以上の性能を達成した．本研究の成果は，既存の検索パイプラインを維持したまま数値条件検索を改善できる可能性を示すものであり，既存システムに適用しやすく，保守性の高い検索システムの提供が期待できる．

1　はじめに

近年，密検索モデルが商品検索や金融文書の検索，医療文書の検索などの幅広い分野で利用されている [1, 2, 3]．密検索モデルはクエリや文書をBERT [4] などの言語モデルを利用し，密ベクトルに変換，そのベクトル同士でのコサイン類似度などの値を元に検索するモデルであり，BM25 といった従来の語彙一致の検索手法に比べ，クエリや文書に含まれる文脈を考慮した検索が可能である．一方，言語モデルは数値の理解が不十分であることが以前から指摘されている [5, 6]．このことから，言語モデルを基盤とする密検索モデルも数値を含んだクエリの検索に対して十分な性能を出すことができないことが考えられる．実際に，Fujimaki ら [7] は，数値条件付きクエリに対する密検索モデルの性能を分析した結果，数値条件を含んだクエリに対して，密検索モデルは期待される検索性能と実際の性能とで大きな乖離があることを明らかにした．特に数値の比較能力が密検索モデルにおいて課題があることが示されている．

この課題は，現実の検索ユースケースにおいて，正しい情報の提供が行えない懸念として考えられる．具体的な情報要求の例として，ニュース記事群から「死者数が 1,000 人以上と報じられている災害」を検索する，あるいはレビューと商品説明文から「カメラでの撮影下で 12 時間以上バッテリーが持続するスマートフォン」を検索する，といったシナリオが挙げられる．これらは，自然文中に非構造的に散在する数値・単位・比較条件を正しく解釈し，それらを満たす文書を優先的に提示する必要がある．このような要求は，構造化データベースに対するSQL では完全に解決することは困難であり，非構造化テキストを扱う密検索モデルそのものが数値条件に対応できる能力を獲得することが不可欠となる．

数値条件付き検索の精度改善に向けた既存のアプローチとして，Almasian ら [8] や Agrawal ら [9] の研究が挙げられる．これらは，自然文からの数量・単位抽出モジュールや，数値比較のための専用ニューラルネットワーク，あるいは互換性のある単位系ごとのスコアリング機構を導入することで，高い検索性能を実現している．しかし，これらの手法は検索モデルの外部に専用の処理機構を必要とするか，推論時のスコアリング計算において，密検索での類似度計算とは別の数値専用の処理を加えるなどの検索処理自体を変更する必要があり，既存の検索パイプラインへの統合やレイテンシや運用コストの観点で課題が残る．

そこで本研究では，上記のような専用モジュールの導入や推論ロジックの変更を行わず， 既存のColBERT [10, 11] の検索パイプラインを維持するという制約の下で，数値条件付き検索の性能をどこまで改善できるかを明らかにすることを目的とする．本研究において，検索パイプラインを維持すると

は，文書のインデクシング手法や推論時の類似度計算（MaxSim）といった密検索システムの根幹には変更を加えず，数値条件の処理をすべてトークン埋め込み表現の生成プロセス内で完結させることを意味する．具体的には，学習時においてのみ数値情報の理解を促す補助タスクと埋め込み設計の改良を導入し，推論時は標準的な ColBERT と同様にトークン埋め込み表現を利用した検索手法の構成で動作させるモデルを提案する．

実験の結果，金融および医療の 2 つのベンチマークにおいて，提案手法は標準的な ColBERT の微調整手法と比べて優位な性能改善を達成し，数値処理に特化した専用モジュールを持つ先行研究と比較しても同等以上の性能を達成することを示した．本論文では，検索パイプラインを変えないという強い制約下においても，学習方法や埋め込み表現の工夫のみで数値条件クエリに対する能力を向上させられることを示し，この方向性の発展を，将来的に低コストかつ高精度な数値条件付き検索を実現する有望なアプローチとして提案する．

本論文の貢献は以下の通りである．

- 既存の ColBERT 推論システムと完全な互換性を持ちつつ，数値対応能力を強化したモデル NumColBERT を提案した．
- 金融および医療ドメインの数値条件検索タスクにおいて，専用モジュールを必要とする手法と同等以上の精度を，より低い運用コストで実現できることを実証した．
- 既存パイプラインの維持という制約下における数値処理の可能性を明らかにし，実用的な検索システム設計に資する新たな知見を与えた．

2 提案手法：NumColBERT

本研究では，既存の ColBERT の検索パイプラインを変更することなく，数値条件付きクエリに対する検索性能を向上させる手法として NumColBERT を提案する．NumColBERT は，ColBERT の基本設計であるトークンごとの埋め込み生成と MaxSim 演算によるスコアリング機構を継承し，文書のエンコードとインデックス化，クエリのエンコードと検索という一連のパイプラインも ColBERT と同一である．本手法の独自性は，モデルの学習過程において，数値情報の理解を促進するための複数の補助タスクと，数値に特化した埋め込み設計を導入する点にある．

NumColBERT では以下の 4 つの主要な構成要素を導入する．(i) 数値トークン検出，(ii) 数値ゲート機構，(iii) 数値情報・条件予測，(iv) 数値対照学習．これらの構成要素は，ColBERT の検索パイプラインとの互換性を保ちながら数値理解を強化するよう設計されている．具体的には，推論時に利用されるのは数値トークン検出と数値ゲート機構のみであり，残りの要素は学習時の補助タスクとして機能する．これにより，既存のベクトル検索基盤を変更することなく，文書側のインデックスも従来と同じ方法で生成できる一方で，学習時には数値情報・条件予測や数値対照学習といった補助タスクを通じて，エンコーダが数値の構造と意味を深く理解した埋め込みを生成することが可能となる．以下では，各構成要素の詳細を説明する．

数値トークン検出　入力テキスト中の数値表現を特定するため，各トークンが数値の一部であるかを予測する二値分類ヘッドを設け，対応する損失を $\mathcal{L}_{\text{detect}}$ とする．このヘッドは推論時にも利用され，数値トークンと検出された部分に対して後述の数値ゲート機構の処理が適用される．

数値ゲート機構　数値トークンの重要度に応じてスコアへの寄与を調整するため，数値ゲート機構を導入する．この機構は，クエリ側の数値トークン検出で検出された各数値トークンの埋め込みに対して，ゲート機構によって予測されたゲート値をノルムとなるよう調整し，検索損失 $\mathcal{L}_{\text{retrieval}}$ における検索スコアの算出時に対応する数値トークン埋め込みのスコア重みとして機能する．これにより，クエリのトピックとの関連性が高い数値表現の影響を強め，そうでないものは影響を弱めることで，より適切な数値条件検索を実現する．この機構は独立した損失項を持たず，検索損失の計算過程に組み込まれる形で学習され，推論時にも利用される．

数値情報・条件予測　数値トークンの埋め込みがその値の大きさや単位，さらにはクエリにおける比較条件といった意味情報を保持するよう，複数の補助タスク $\mathcal{L}_{\text{info}}$ を導入する．具体的には，数値トークンの BERT の隠れ層の埋め込みから，その値の仮数 (mantissa) と指数 (exponent) を予測する回帰タスク，および単位 (unit) を予測する多クラス分類タスクを解かせる．さらに，クエリに対してのみ，含まれる数値条件の種類（例：より大きい “>”，より小さい “<”，等しい “=”）を予測する多クラス分類タスクを追加する．これらの補助タスクにより，エン

コーダは数値の持つ多面的な情報やクエリの意図をトークン埋め込み空間内で体系的に表現することを学習する．

数値対照学習　クエリが要求する数値条件と文書中の数値が意味的に合致するかを埋め込み空間上の距離として直接学習させるため，数値トークン間の対照学習 $\mathcal{L}_{\text{contrast}}$ を導入する．クエリ中の数値トークン埋め込みをアンカーとし，そのクエリが指定する条件（例：「1000 以上」）と単位（例：「人」）を満たす文書中の数値トークン埋め込みを正例，満たさないものを負例とするペアを構築する．本研究では，一つのクエリに対して複数の正例が存在しうる状況を考慮し，複数の正例を許容する InfoNCE 損失を用いて学習を行う．

NumColBERT の学習では，ColBERT 本来の検索損失に加え，上述した数値に関する複数の補助タスクの損失項を組み合わせた合成損失関数を最適化する．全体の損失関数 $\mathcal{L}_{\text{total}}$ は次式で定義される．

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{retrieval}} + \lambda_1 \mathcal{L}_{\text{detect}} + \lambda_2 \mathcal{L}_{\text{info}} + \lambda_3 \mathcal{L}_{\text{contrast}} \quad (1)$$

ここで $\mathcal{L}_{\text{retrieval}}$ は ColBERT の検索損失であり，数値ゲート機構が検索スコアの算出時に組み込まれている．λ は各損失の重み係数である．

以上の構成により，NumColBERT はエンコーダがテキストの文脈的意味だけでなく，文中に埋め込まれた数値情報の構造と意味を深く理解したトークン埋め込みを生成することが期待される．重要な点として，これらの拡張のうち推論時に利用されるのは数値トークン検出ヘッドと数値ゲートヘッドのみであり，文書側は従来と同じ方法でインデクシングを行い，検索パイプライン自体は ColBERT と完全に同一のまま維持される．これにより，既存のベクトル検索基盤を変更することなく，数値条件への感度を高めることが可能となる．

3　実験

本章では，提案手法 NumColBERT の有効性を，数値条件付きクエリを含む検索ベンチマーク上で検証する．特に，ColBERT の検索パイプラインを変更することなく，学習時の補助タスクおよび数値ゲート機構の導入のみで，数値条件検索がどの程度改善できるかに焦点を当てる．

データセット　実験には，Numbers Matter が公開した FinQuant および MedQuant データセットを用いる [8]．FinQuant は金融ドメイン，MedQuant は医療ドメインの非構造テキストからなる文書集合とクエリ集合から構成され，クエリには数値・単位・比較条件（例：以上，以下，等号）が自然文として含まれる．評価はテストクエリに対する検索性能指標により行い，既存研究に倣って nDCG@10，MRR@10，P@10，Recall@100 を性能指標として評価する．

比較手法　ベースラインとして，以下の手法を比較対象とする．(i) 同一データセットでファインチューニングした ColBERT（$\text{ColBERT}_{\text{ft}}$），(ii) 数量抽出パイプラインを導入した QColBERT [8]，(iii) 数量比較専用ニューラルネットワークを持つ DeepQuant [9]，(iv) 参考として，大規模言語モデルを用いた検索手法として GPT-4o-mini によるリランキング手法を比較する．これは Agrawal ら [9] で検証された実験での結果である．提案手法 NumColBERT は，第 2 章で述べた数値トークン検出，数値ゲート機構，数値情報・条件予測，数値対照学習を学習時に導入し，推論時は ColBERT と同一の埋め込みによる検索手順で検索を行う．

主実験結果　表 1 に FinQuant および MedQuant テストセットにおける各モデルの検索性能を示す．FinQuant において，提案手法 NumColBERT は nDCG@10 で 0.59 を達成し，ベースラインである $\text{ColBERT}_{\text{ft}}$（0.50）を上回り，DeepQuant と同等の結果を示した．また，MedQuant においては，NumColBERT は nDCG@10 で 0.50 を記録し，DeepQuant（0.44）や $\text{ColBERT}_{\text{ft}}$（0.44）を上回り，すべての指標において最高値を達成した．注目すべき点として，専用の数値比較モジュールを持つ DeepQuant と同等以上の性能を達成しており，検索パイプラインを変更しない制約下においても，学習時の補助タスクと数値ゲート機構の導入により，専用モジュールに匹敵する性能を実現できることが示された．

また，大規模言語モデルである GPT-4o-mini を用いた手法では，どの密検索モデルよりも低い性能を示している．これは，数値条件付きのクエリに対し，密検索モデルが推論コストと検索性能の両面において，有効性の高いアプローチであることを示唆している．

これらの結果は，本研究の主要な貢献である既存の検索パイプラインを維持したまま数値条件検索を改善できるという主張を強く支持するものである．特に，QColBERT や DeepQuant が推論時に専用の数値抽出・比較モジュールを必要とするのに対し，提

表 1 FinQuant および MedQuant テストセットにおける各モデルの検索性能．太字は最高値，下線は 2 番目の値を示す．

Model	FinQuant				MedQuant			
	nDCG@10	MRR@10	P@10	R@100	nDCG@10	MRR@10	P@10	R@100
$\text{ColBERT}_{\text{ft}}$	0.50	0.62	0.26	0.81	<u>0.44</u>	0.52	<u>0.22</u>	<u>0.80</u>
QColBERT	<u>0.56</u>	0.69	0.30	<u>0.87</u>	0.37	0.51	0.18	0.73
DeepQuant	**0.59**	**0.73**	**0.32**	**0.88**	<u>0.44</u>	<u>0.59</u>	0.21	<u>0.80</u>
GPT-4o-mini	0.36	0.52	0.17	0.69	0.26	0.36	0.13	0.62
NumColBERT	**0.59**	<u>0.71</u>	<u>0.31</u>	<u>0.87</u>	**0.50**	**0.61**	**0.25**	**0.84**

表 2 NumColBERT のアブレーション実験

Method	nDCG@10	Recall@100
NumColBERT	0.592	0.874
– w/o info/cond	0.581	0.857
– w/o contrast	0.518	0.802
– w/o gate	0.545	0.830

表 3 一般的な検索と数値条件検索での検索性能

Method	MRR@10	
	MSMARCO	FinQuant
ColBERT	0.376	0.373
DeepQuant (zero-shot)	0.210	0.734
DeepQuant (joint training)	0.332	0.731
NumColBERT (zero-shot)	0.268	0.710
NumColBERT (joint training)	0.356	0.708

案手法は標準的な ColBERT の MaxSim 演算のみで同等の性能を達成しており，実装の簡潔性と運用コストの観点で大きな利点を持つ．

アブレーション分析　提案手法を構成する各要素の寄与を明らかにするため，表 2 にアブレーション実験の結果を示す．完全な構成では，数値情報・条件予測，数値対照学習，数値ゲート機構のすべてを含み，nDCG@10 で 0.592，Recall@100 で 0.874 を達成した．数値対照学習を除去した場合（w/o contrast），nDCG@10 は 0.518 まで低下した．この結果は，数値トークン間の対照学習が数値条件検索の性能向上に重要な役割を果たしていることを示している．さらに，数値ゲート機構を除去した場合（w/o gate），nDCG@10 は 0.545，Recall@100 は 0.830 となり，両指標で顕著な性能低下が観察された．これは，数値トークンの重要度を動的に調整する機構が，クエリの意図に応じた適切なスコアリングに不可欠であることを示唆している．

3.1　一般検索性能

数値条件検索能力の獲得が，一般的な文書検索性能に影響を与えるか，および複数データセットを用いた同時学習（Joint Training）によってその抑制が可能かを検証する．一般的な検索タスクは MSMARCO [12] Passage Reanking タスク，数値条件検索タスクは FinQuant を用いる．

比較対象として，MSMARCO のみで学習した ColBERT，および FinQuant のみで学習を行い MSMARCO に対して Zero-shot で適用する設定（Zero-shot），そして MSMARCO と FinQuant の両方を用いて学習した設定（Joint-training）を比較する．

なお，DeepQuant についての結果は Agrawal ら [9] の報告値を引用する．Joint-training を適用したモデルでは，FinQuant における高い検索性能を維持しつつ，MSMARCO における性能低下を抑制することに成功している．加えて，NumColBERT は DeepQuant と比較して，MSMARCO における検索性能をより高い水準で維持している．以上の結果は，提案手法 NumColBERT が学習データの構成を適切に設計することで，汎用的な検索能力と高度な数値条件処理能力を両立可能であることを示唆している．

4　まとめ

本研究では，ColBERT の検索パイプラインを維持しつつ，数値条件を含む検索性能を向上させる手法である NumColBERT を提案した．提案手法は，数値埋め込みゲート機構および学習時の補助タスクを導入することにより，通常の ColBERT と同様のマルチベクトル MaxSim 検索の枠組みの中で検索処理を完結させる．これにより，専用モジュールを必要とする従来手法において課題となっていた，運用コストの増大や既存システムへの統合の困難さを解消した．評価実験の結果，金融および医療ドメインにおいて提案手法は，既存手法と比較して同等以上の検索性能を達成することを確認した．本研究の成果は，既存の ColBERT ベースのシステムとの互換性を保ちながら，高精度な数値条件検索を実現するものである．今後の展望としては，他の検索モデルへの適用や，日付などの数値以外への構造的情報への拡張が挙げられる．

謝辞

本研究は JSPS 科研費 JP23K28090 と JP24K03048 の助成を受けたものです.

参考文献

[1] Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. **arXiv preprint arXiv:2206.06588**, 2022.

[2] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. In **EMNLP**, pp. 3697–3711, 2021.

[3] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **Applied Sciences**, Vol. 11, No. 14, p. 6421, 2021.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In **EMNLP-IJCNLP**, pp. 5307–5315, 2019.

[6] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In **NAACL-HLT**, pp. 2368–2378, 2019.

[7] Haruki Fujimaki and Makoto P. Kato. Investigating the performance of dense retrievers for queries with numerical conditions. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, **Advances in Information Retrieval**, pp. 210–218, Cham, 2025. Springer Nature Switzerland.

[8] Satya Almasian, Milena Bruseva, and Michael Gertz. Numbers matter! bringing quantity-awareness to retrieval systems. **arXiv preprint arXiv:2407.10283**, 2024.

[9] Prayas Agrawal, Nandeesh Kumar K M, Muthusamy Chelliah, Surender Kumar, and Soumen Chakrabarti. Dense retrieval with quantity comparison intent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 23825–23839, Vienna, Austria, July 2025. Association for Computational Linguistics.

[10] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '20, p. 39–48, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics.

[12] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne, editors, **Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016**, Vol. 1773 of **CEUR Workshop Proceedings**. CEUR-WS.org, 2016.

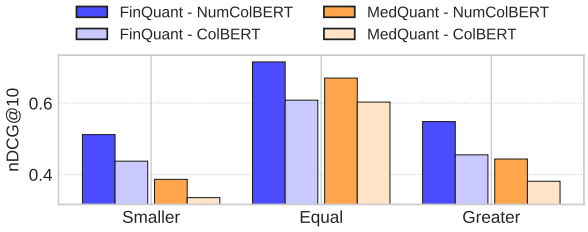

図 1　FinQuant における各モデルの数値条件別の検索性能

表 4　FinQuant における数値トークン対照学習の学習パターンの性能比較.

Method	nDCG@10	Recall@100
Joint	0.581	0.866
Separate	0.590	0.866
Numeric only	0.580	0.864
Unit only	0.592	0.874

A　追加実験

A.1　数値条件別性能

数値を条件とした検索において，その条件の種類（等号，大小関係など）によってモデルの性能がどのように変化するかを詳細に分析する．図 1 に，FinQuant データセットにおける数値条件（Equal, Greater, Smaller）ごとの nDCG@10 スコアを示す．比較対象は，ベースラインである $\text{ColBERT}_{\text{ft}}$ と提案手法 NumColBERT である．

図より，両モデルともに Equal（=），Greater（>），Smaller（<）の順で性能が高い傾向にあることが確認できる．等号条件は特定の値を対象とするため比較的容易である一方，大小比較（Greater，Smaller）は数値の範囲を考慮する必要があるため，検索難易度が高いと考えられる．特に Smaller 条件での性能が最も低い点は，既存の数値埋め込みにおける大小関係の学習の難しさを示唆している．

提案手法 NumColBERT は，Equal，Greater，Smaller のすべての条件において $\text{ColBERT}_{\text{ft}}$ を上回る性能を達成している．これは，提案手法で導入した数値トークンに対する対照学習や数値条件予測という補助タスクが，特定の条件に偏ることなく，汎用的に数値理解能力を向上させていることを示している．

A.2　数値対照学習のパターン分析

数値対照学習における学習パターンの違いが性能に与える影響を調査するため，表 4 に 4 つの学習パターンの比較結果を示す．Joint は数値と単位を統合した条件でペアを構築，Separate は数値条件と単位条件を別々に扱う，Numeric only は数値条件のみを考慮，Unit only は単位条件のみを考慮する設定である．

結果として，Unit only（単位条件のみ）が最も高い性能を示し，nDCG@10 で 0.592，Recall@100 で 0.874 を達成した．これに対し，Separate（数値と単位を分離）も nDCG@10 で 0.590 と高い性能を示した．一方，Joint（統合）および Numeric only（数値のみ）はやや性能が低く，それぞれ nDCG@10 で 0.581，0.580 となった．この結果は，単位情報が数値条件検索において極めて重要であり，数値の大小関係よりも単位の一致が検索性能に大きく寄与することを示唆している．