引用文脈の類似度に基づく局所的引用論文推薦の改良

田中陸斗 [1]　杉山弘晃 [2]　平博順 [3]　桒原龍生 [3]　堂坂浩二 [1]
[1] 秋田県立大学　[2]NTT コミュニケーション科学基礎研究所　[3] 大阪工業大学
{m24p010, dohsaka}@akita-pu.ac.jp　h.sugi@ieee.org
hirotoshi.taira@oit.ac.jp　m1m23a10@st.oit.ac.jp

概要

科学技術論文の数が急増する現代において，関連論文の効率的な推薦が重要になっている．本研究では，科学技術論文の関連研究の章を執筆する際，引用すべき個所が明示された場合，その周囲の文に基づいて適切な引用論文を推薦する局所的引用論文推薦タスクに焦点を当てる．我々は，これまで，対象論文の引用文脈と既存論文の引用文脈の類似度を活用して引用論文を推薦するという引用文脈参照法を提案してきた．この方法には，未引用の論文に対応できないという問題が存在する．本論文では，この問題に対処するために，引用文脈参照法と従来のタイトル・要旨参照の新たな組み合わせ方を提案する．大規模なデータセットを使った評価実験の結果，２つの手法の新たな組み合わせ方により論文推薦の性能が向上することが示された．

1　はじめに

科学技術論文執筆においては，適切な引用を行うことが重要だが，論文出版数の急増により関連研究をすべて把握することが難しくなっており，論文執筆支援の必要性が高まっている．

Narimatsu ら [1] は，研究者の論文執筆における関連研究の引用および生成に関わる統合的な執筆支援を目的として，関連研究に関わる様々な既存のタスクを統合した新たなデータセット構築方法および 5 つのタスクを定義した．本研究では，この中の引用論文推薦タスクに焦点を当てる．このタスクは，与えられたテキストに基づいて適切な引用論文を推薦するもので，大域的引用論文推薦と局所的引用論文推薦に分類される [2, 3]．本研究では，関連研究の章において，引用を付与すべき個所（引用マーカー）が与えられたときに，引用マーカーの周囲の引用文脈に基づいて，引用すべき論文を推薦する局所的引用論文推薦を扱う．また，用語の定義として，引用マーカーを含む文と前後の文の 3 文を引用文脈，引用を付与したい引用文脈をもつ論文を対象論文と呼ぶ．

局所的引用論文推薦タスクにおいて，推薦候補となる既存論文のタイトル・要旨を集めたものを論文プールと呼ぶ．2 節で説明するように，従来研究では，対象論文の引用文脈と論文プール内の既存論文のタイトル・要旨の間の類似度を計算し，対象論文の引用文脈と類似したタイトル・要旨をもつ論文を推薦するという手法が提案されてきた．この手法をタイトル・要旨参照法と呼ぶ．この手法には論文のタイトル・要旨は容易に収集できるという利点があるが，引用文脈とタイトル・要旨は別の意図で書かれた文章であるため，適切な引用論文を検索できない場合がありえる．

そこで，我々は，ある特定の論文の引用文脈同士は似通っていることが多いことを仮定し，対象論文の引用文脈と既存論文が引用された際の引用文脈の類似度を活用して推薦を行う引用文脈参照法を提案してきた [4]．この手法では，既存論文の引用文脈を引用文脈プールとして収集し，現在着目している対象論文の引用文脈と類似した引用文脈をもつ論文を引用文脈プールから探して推薦する．局所的引用論文推薦の従来研究では，タイトル・要旨を使った手法は提案されてきているが，知る限りにおいて，既存論文の引用文脈を収集し，それを活用することに着目した研究はない．ただし，引用文脈参照法には，過去に一度も引用されたことがない論文は推薦できないというコールドスタート問題が存在する．そこで，以前のアプローチ [4] では，一度も引用されたことがない論文でも推薦候補とするために，タイトル・要旨法と引用文脈参照を組み合わせた手法を提案した．本論文では，引用文脈参照法とタイトル・要旨参照法を組み合わせる従来の手法をさらに発展させた．具体的には，タイトル・要旨参照法で推薦論文候補を絞り込んだうえで，その候補に対し

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

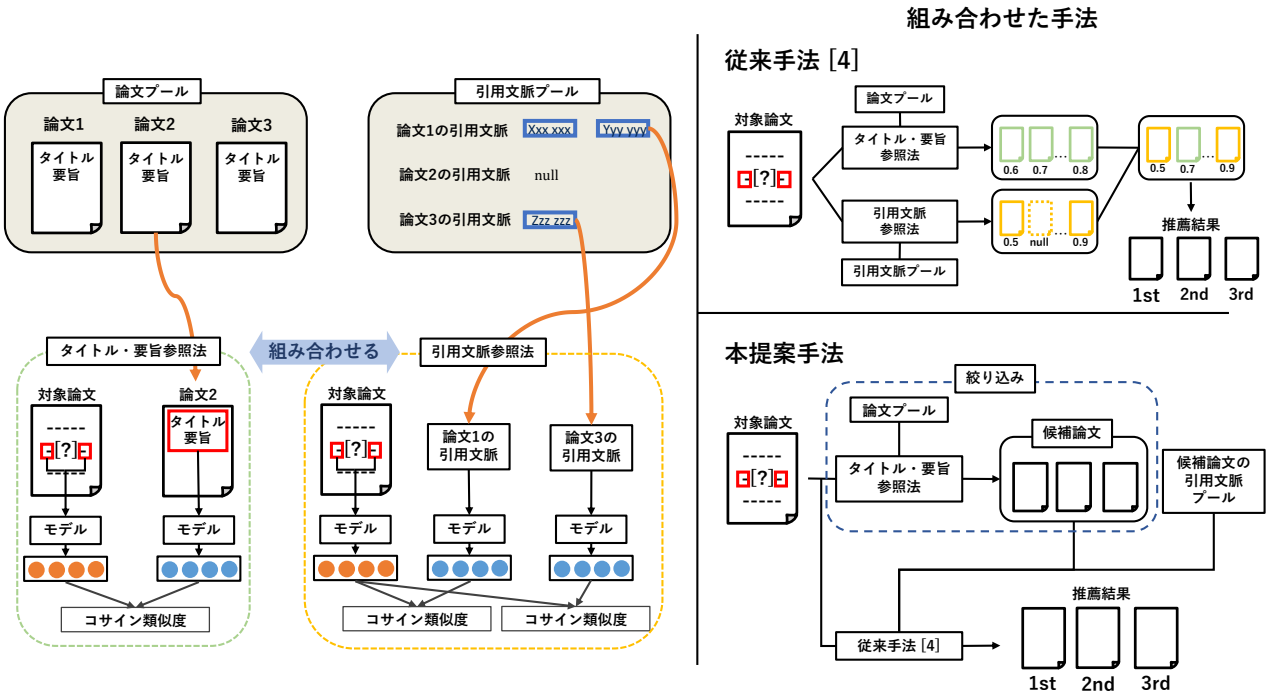

図 1　提案手法

て，タイトル・要旨参照法と引用文脈参照法を組み合わせて推薦論文を選ぶ．また，規模の異なる 2 種類のデータセットによる評価を行い，提案手法の有効性について考察した．

以下において，2 節で関連研究を述べ，3 節で提案手法を示す．4 節で，データセット並びに評価方法を説明し，評価結果について考察する．

2　関連研究

局所的引用論文推薦の従来研究には，引用文脈と論文間の関係に焦点を当てる研究が多くあり，Sugimoto ら [5] は，対象論文の引用文脈と推薦する候補の論文のタイトル・要旨の双方を独立に SciBERT[6] で埋め込み，候補の論文をコサイン類似度でランク付けするモデルを提案している．また，Zoran ら [7] は，論文のタイトルと要旨の他に，著者や引用数などの情報を組み込むことにより，引用推薦の性能が向上することを示した．Gu ら [8] は，階層型注意ネットワークを使用したテキスト埋め込みと SciBERT に基づく再ランキングを組み合わせた引用推薦システムを提案した．論文のタイトル・要旨以外に着目する研究として，Jeong ら [9] は，引用文脈の埋め込み表現を BERT[10] と GCN[11] を用いて取得し，引用論文推薦に取り入れている．

しかし，従来研究には著者が引用したい被引用論文が，他の既存論文ではどのような引用文脈で引用されているかに着目した研究は知る限りでは存在しない．我々の以前の研究 [4] では，対象論文の引用文脈と被引用論文の既存論文における引用文脈の比較を行うことが引用論文推薦に有効であることを示した．本研究では，以前の研究を改良した手法を提案する．

3　提案手法

本論文では，与えられた引用文脈に対して，類似したタイトル・要旨をもつ論文を取得する手法をタイトル・要旨参照法と呼ぶ．また，我々が新たに提案してきた，対象論文の引用文脈と既存論文が引用された際の引用文脈の類似度を使って引用論文を推薦する手法を引用文脈参照法と呼ぶ [4]．さらに，引用文脈参照法に存在する，1 度も引用されていない論文は推薦対象にならないというコールドスタート問題に対処するため，タイトル・要旨参照法と引用文脈参照法を組み合わせた手法を提案する．

タイトル・要旨参照法　この手法は，対象論文の引用文脈と論文のタイトル及び要旨間の類似度を計算し，対象論文の引用文脈と類似したタイトル・要旨をもつ論文を推薦することを目的とする．本研究では，従来手法である Sugimoto ら [5] の手法を使用する．この手法では，対象論文の引用文脈の埋め込

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　データセットの詳細

データセット	テストデータの引用文脈数	論文プール	引用文脈プール	コールドスタート問題が起こる割合
arXiv	7,869	38,620	77,411	9%
S2ORC	213,600	918,357	2,103,007	27%

みベクトルと，正解となる論文のタイトル・要旨の埋め込みベクトルの距離が近くなるようにそれぞれ SciBERT をファインチューニングしている．推薦時には，対象論文の引用文脈のベクトルと，各論文のベクトル間のコサイン類似度を計算し，高い順に推薦する．

引用文脈参照法　この手法は，対象論文の引用文脈と類似した引用文脈をもつ論文を引用文脈プールから探して推薦することを目的とする．引用文脈同士の類似度を比較するために，Sentence-BERT[12](以下 SBERT) を用いてモデルを作成する．対象論文の引用文脈で引用されている論文と同じ論文を引用している引用文脈を正例とし，引用文脈プール内からランダムに選んだ引用文脈を負例として，以下の式で表される損失関数で埋め込み表現を学習する．

$$Loss = max\{(||C - C^{+}|| - ||C - C^{-}|| + \epsilon), 0\} \quad (1)$$

ここで，C はアンカーである引用文脈の埋め込み，C^{+} は正例の埋め込みベクトル，C^{-} は負例の埋め込みベクトルを示す．距離にはユークリッド距離を使用し，マージン ϵ は 1 とした．

推薦時は，対象論文の引用文脈と，引用文脈プール内の論文の引用文脈をモデルに入力し，得られたベクトル間のコサイン類似度を計算する．その際，引用文脈プール内で複数の引用文脈を所持している論文は，最も大きい類似度をその論文のスコアとする．最後に，スコアが高い順に論文を取得する．

組み合わせた手法　引用文脈参照法において，1 度も引用されていない論文が推薦対象にならないという問題に対処するため，2 つの手法を組み合わせる．図 1 に提案手法を示す．従来の方法 [4] では，論文プール内のすべての論文にタイトル・要旨参照法，引用文脈参照法でスコアを付けた．このとき，1 度以上引用されている論文，つまり引用文脈を所持している論文には引用文脈参照法を使用し，対象論文の引用文脈と引用文脈プール内の論文の引用文脈とのコサイン類似度を計算する (図では論文 1 と論文 3)．未引用の論文 (図では論文 2) では，論文プール内の論文との類似度をタイトル・要旨参照法を使用しコサイン類似度を計算する．その後，得られたコサイン類似度をソートし，高い順に論文を取得していた．

本提案では，すべての論文を対象としていた以前の方法と異なり，まず，タイトル・要旨参照法で候補となる論文 (候補論文) の絞り込みを行った．今回，絞り込む論文数は 100 件とした．その後の処理は以前の研究と同様に，絞り込んだ候補論文のうち，1 度以上引用されている論文には引用文脈参照法を使用し，未引用の論文には，候補論文との類似度をタイトル・要旨参照法を使用しコサイン類似度を計算し，高い順に論文を取得する．

4　実験

4.1　データセット

研究者の科学技術論文の執筆支援を目的として，Narimatsu ら [1] によって作成されたデータセット (以下 arXiv) と，科学技術論文の大規模コーパスである S2ORC[13] を使用する．前者は，arXiv から取得した論文の関連研究の章が約 30,000 件と，関連研究の章で引用されている論文のタイトル，要旨が含まれている．すべての関連研究の章のうち，引用論文数が 1 件以上である約 13,000 件を使用する．後者は，コンピュータサイエンスの分野である約 214,000 件の関連研究の章を抽出し使用する．関連研究の章は訓練データ・検証データ・テストデータにそれぞれ 8:1:1 に分割する．データセットの詳細を表 1 に示す．

本研究では，論文のタイトル・要旨が得られた論文集合を論文プールと定義する．関連研究の章をもつ論文と，その章で引用されている論文が論文プールに含まれている．総数は arXiv では約 38,000 件，S2ORC では約 920,000 件である．

訓練データ内の関連研究の章で使われた引用文脈を集めたものを引用文脈プールと定義する．arXiv の引用文脈プール内の論文は約 10,000 件存在し，最大引用回数は 830 回，平均引用回数は 7.8 回である．また，S2ORC では，引用文脈プール内の論文は約 700,000 件存在し，最大引用回数は 1254 回，平均引用回数は 3.0 回である．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　評価結果

データセット	手法	@5		@10	
		Recall	MRR	Recall	MRR
arXiv	タイトル・要旨参照法	0.497	0.387	0.583	0.399
	引用文脈参照法	0.533	0.434	0.614	0.444
	組み合わせた手法 [4]	0.538	0.436	0.622	0.447
	組み合わせた手法 (本提案)	**0.603**	**0.493**	**0.684**	**0.504**
S2ORC	タイトル・要旨参照法	0.306	0.225	0.374	0.235
	引用文脈参照法	0.258	0.203	0.309	0.211
	組み合わせた手法 [4]	0.304	0.235	0.369	0.244
	組み合わせた手法 (本提案)	**0.327**	**0.252**	**0.390**	**0.261**

4.2　評価手法

本研究では，提案した論文推薦システムの性能を上位 5 件と上位 10 件の候補に対する Recall と MRR で評価する．Recall は以下の式で表される．

$$Recall@k = \frac{|\alpha \cap p_k|}{|\alpha|} \tag{2}$$

ここで，k は推薦する論文数，α は正解の被引用論文集合，p_k は上位 k 件の推薦リストである．また，MRR は以下の計算で表される．

$$MRR@k = \frac{1}{|\alpha|} \sum_{u \in \alpha} \frac{1}{rank_u} \tag{3}$$

ここで，u は正解論文の 1 つ，$rank_u$ が最初に正解論文が出現する順位を示している．

4.3　結果と考察

評価結果を表 2 に示す．arXiv での結果を見ると，引用文脈参照法は従来のタイトル・要旨参照法よりも性能が良いことから，対象論文と既存論文の引用文脈同士の類似性を活用することの有効性が示された．一方，S2ORC では，引用文脈参照法の性能がタイトル・要旨参照法よりも劣っている．この原因としては，引用文脈プールに対象論文と似ているが誤った文脈が多く存在するためで，例えば異なる分野の論文でも似た背景を持つ文脈が存在する可能性が考えられる．

さらに，両データセットにおいて，本提案の組み合わせた手法では，他の手法よりも性能が向上することが確認できる．タイトル・要旨参照法を使った絞り込みにより，より関連性の高い論文の引用文脈に焦点を当てることができ，正確な推薦結果を得やすくなったと考えられる．

また，S2ORC では，arXiv データセットと比較してコールドスタート問題の発生割合が約 3 倍にもなることが明らかになった．そのため，今後の研究では arXiv と同様の条件でデータを処理するか，あるいは別のデータセットを使用し，コールドスタート問題の発生率との関連性をさらに詳細に分析することが必要であると考える．

5　おわりに

本研究では，引用マーカーが与えられたときに，引用マーカー周囲の引用文脈に基づいて，適切な論文を推薦する局所的引用論文推薦のタスクに取り組んだ．特定の論文が引用されるときの引用文脈同士は似ているという仮定のもと，SBERT を使用して引用文脈間の類似度を計算し，適切な論文を推薦する引用文脈参照法を提案した．しかし，この手法には，過去に引用されたことがない論文を推薦することができないというコールドスタート問題がある．これを解決するために，広く用いられるタイトル・要旨参照法と引用文脈参照を組み合わせる必要があるが，本論文では、まずタイトル・要旨参照法を使って推薦候補論文を絞り込んだうえで，次にタイトル・要旨参照法と引用文脈参照法を組み合わせて引用論文を推薦するという方法を提案した．この新たな組み合わせる方法により，以前の研究と比較して推薦性能が向上することが確認され，この方法の有効性を確認した．

将来的な展望としては，著者情報やキーワードなど他の要素を利用することや，引用のされ方に焦点を当てることが考えられる．同じ論文を引用する際でも，データセットの説明や研究背景の提示など，引用の意図は多様である．この引用の意図を分析することは，論文推薦の性能を高める可能性があると考えられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究の遂行にあたり，ご助言・ご協力をいただきました，NTT コミュニケーション科学基礎研究所 成松宏美主任研究員，電気通信大学情報理工学研究科 小山康平氏，電気通信大学 南泰浩教授，工学院大学 大和淳司教授，国立研究開発法人科学技術振興機構 菊井玄一郎氏に感謝いたします．また，日頃より丁寧にご指導してくださる秋田県立大学 石井雅樹教授，伊東嗣功助教に感謝いたします．

参考文献

[1] Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. Task definition and integration for scientific-document writing support. In **Proceedings of the Second Workshop on Scholarly Document Processing**, pp. 18–26, 2021.

[2] Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. **International Journal on Digital Libraries**, Vol. 21, No. 4, pp. 375–405, 2020.

[3] Chaker Jebari, Enrique Herrera-Viedma, and Manuel Jesus Cobo. Context-aware citation recommendation of scientific papers: Comparative study, gaps and trends. **Scientometrics**, Vol. 128, No. 8, p. 4243–4268, jun 2023.

[4] 田中陸斗, 杉山弘晃, 平博順, 有田朗人, 堂坂浩二. 引用文脈の類似度に基づく局所的引用論文推薦. 言語処理学会第 29 回年次大会, 2023.

[5] Kaito Sugimoto and Akiko Aizawa. Context-aware Citation Recommendation Based on BERT-based Bi-Ranker. In **2nd Workshop on Natural Language Processing for Scientific Text at AKBC 2021**, 2021.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, 2019.

[7] Zoran Medić and Jan Snajder. Improved local citation recommendation based on context enhanced with global information. In Muthu Kumar Chandrasekaran, Anita de Waard, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert M. Patton, and Michal Shmueli-Scheuer, editors, **Proceedings of the First Workshop on Scholarly Document Processing**, pp. 97–103, Online, November 2020. Association for Computational Linguistics.

[8] Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, **Advances in Information Retrieval**, pp. 274–288, Cham, 2022. Springer International Publishing.

[9] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. A context-aware citation recommendation model with bert and graph convolutional networks. **Scientometrics**, Vol. 124, No. 3, pp. 1907–1922, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. **arXiv preprint arXiv:1609.02907**, 2016.

[12] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. **arXiv preprint arXiv:1908.10084**, 2019.

[13] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).