引用文脈の類似度に基づく局所的引用論文推薦

田中陸斗 [1] 杉山弘晃 [2] 平博順 [3] 有田朗人 [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, m1m21a02}@oit.ac.jp

概要

出版される論文の爆発的な増加により引用論文推薦の需要が高まっている．本論文では，対象論文の関連研究の章を執筆する際に，引用を付与すべき個所に対し適切な引用論文を推薦する局所的引用論文推薦のタスクを扱う．引用を付与すべき個所の周囲の文は引用文脈と呼ばれる．本研究では，ある特定の論文を引用する際に書かれる引用文脈同士は似通っていることが多いという仮定のもと，対象論文の引用文脈と既存論文の引用文脈の類似度を使って引用論文を推薦する引用文脈参照法を提案する．しかし，この手法には一度も引用されたことがない論文は推薦できないコールドスタート問題が存在する．本研究では，対象論文の引用文脈と既存論文のタイトル・要旨間の類似度を使った従来手法と組み合わせることによりコールドスタート問題に対処し，論文推薦の性能が向上することを示す．

1　はじめに

学術論文を執筆する際，論文中の主張を裏付けるために適切な引用を行うことは重要である．しかし，近年論文数が爆発的に増加し，研究者が関連する研究をすべて読み切ることは困難となっており，関連研究に関わる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．本稿では，その中の引用論文推薦タスクに着目する．これは与えられたテキストに対して適切な引用論文を推薦するタスクであり，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]．大域的引用論文推薦では，参考文献リストに入るべき論文を推薦するのに対し，局所的引用論文推薦では，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内に入る論文を推薦する．本研究では，関連研究の章を執筆する際の局所的引用論文推薦のタスクを扱う．また，用語の定義として，引用マーカーの前後 50 単語を引用文脈，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．

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，従来研究として，現在着目する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，本研究では，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使って引用論文を推薦する引用文脈参照法を提案する．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．しかし，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，本研究では，広く用いられてるタイトル・要旨参照法と組み合わせることで，一度も引用されていない論文も推薦できるように対処した．

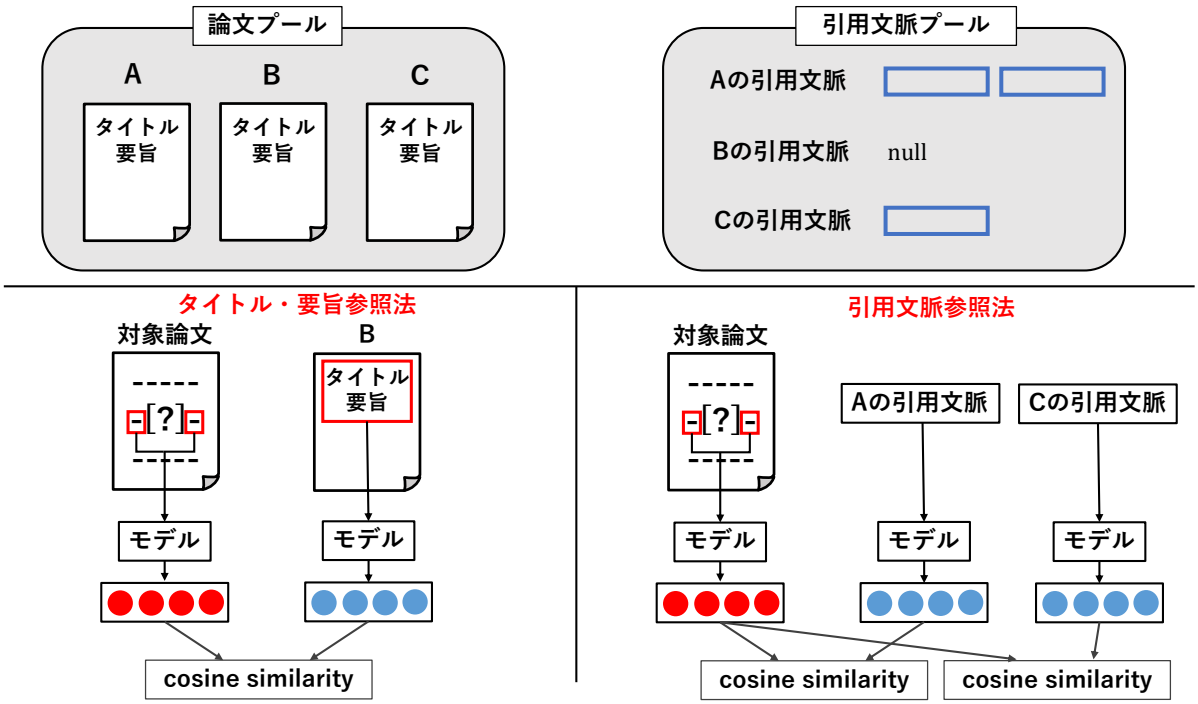

図 1　手法の概要

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は現在までに様々なアプローチがとられてきた．既存研究には，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 [3] 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[4] 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，Ebesu ら [5] は，引用文脈に応じた引用推薦のためのニューラル引用ネットワークを，引用文脈と著者ネットワーク，論文のタイトルを使用することで検討した．Zhang ら [6] は，草稿時の状況を想定しており，引用文脈と既に自分で引用している被引用論文の情報，引用文脈を書いている章に着目した推薦手法を提案している．Jeong ら [7] は，引用文脈の埋め込み表現を，BERT[8] と GCN[9] を用いて取得し，引用論文推薦に取り入れている．しかし，既存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．本稿では，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す．

3　提案手法

本稿では，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，新たに提案する，類似した引用文脈で引用された論文を取得する手法を引用文脈参照法と呼ぶ．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．

推論の概要図を図 1 に示す．この図は，論文プール内に A，B，C の 3 件の論文があり，B のみ一度も引用されていない状況である．よって，引用文脈プールに B の引用文脈は存在しない．引用文脈が存在する A および C は引用文脈参照法を用いてスコアを計算し，引用文脈が存在しない B はタイトル・要旨参照法を用いてスコアを計算している例を示している．

タイトル・要旨参照法　引用文脈と論文プール内の既存論文のタイトル・要旨の類似度を比較するために，文の埋め込み表現に特化した BERT である SBERT[10](以下 SBERT) を用いてモデルを作成す

る．アンカーの引用文脈と対応する論文のタイトル・要旨を正例とし，論文プール内の正例論文以外からランダムに選んだ論文のタイトル・要旨を負例として，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．また，プーリング層には，BERT の出力ベクトルの平均値を使用する．

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

ここで，C はアンカーである引用文脈の埋め込み，P^{+} は正例の埋め込み，P^{-} は，負例の埋め込みを示す．また，元論文 [10] に従い，距離にはユークリッドを使用し，マージン ϵ は 1 とした．

推論時は，対象論文の引用文脈と，論文プール内の論文のタイトル・要旨をそれぞれ上記のモデルに入力して得られたベクトル間のコサイン類似度を計算し，これをスコアとする．最後に，スコアが高い順に k 件の論文を取得する．

引用文脈参照法　引用文脈同士の類似度を比較するために，タイトル・要旨参照法と同様に SBERT を用いてモデルを作成する．この手法では，アンカー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，引用文脈プール内の正例の引用文脈以外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．

$$Loss = max\{(||C - C^{+}|| - ||C - C^{-}|| + \epsilon), 0\} \quad (2)$$

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，C^{-} は，負例の埋め込みを示す．タイトル・要旨参照法と同様に，距離にはユークリッドを使用し，マージン ϵ は 1 とした．

推論時は，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の類似度をモデルに入力し，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，引用文脈プール内の論文で，引用文脈が複数所持しているときは，最も大きい類似度をスコアとする．最後に，スコアが高い順に k 件の論文を取得する．

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．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，対象論文の引用論文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．未引用の論文には，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 k 件の論文を取得する．

4　実験

4.1　データセット

研究者の学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を使用する．このデータセットには，arXiv から取得した論文の関連研究の章が約 30000 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 13000 件を使用する．これを訓練データ，検証データ，テストデータにそれぞれ約 10400，1300，1300 件ずつに分割する．また，引用文脈の数は訓練データ，検証データ，テストデータそれぞれで約 70000，10000，7800 件である．テストデータの引用文脈のうち，コールドスタート問題が起こり得る数は 692 件であり，約 9%を占めている．

4.1.1　論文プール

本稿では，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．総数は約 38000 件である．

4.1.2　引用文脈プール

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．よって，総数は約 70000 件である．引用文脈プール内の論文は約 10000 件存在し，最大引用回数は 830 回，最小引用回数は 1 回，平均引用回数は 7.8 回である．

4.2　評価手法

提案した論文推薦システムが既存の論文の引用をどの程度予測できるかを測定する．本稿では，推薦された候補の上位 5 件と上位 10 件に対して Recall と MRR を計算して評価する．Recall は以下の式で表される．

$$Recall@k = \frac{|\alpha \cap p_k|}{|\alpha|} \quad (3)$$

ここで，k は考慮する上位ランキング件数，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

表 1 評価結果

モデル	手法	@5		@10	
		Recall	MRR	Recall	MRR
TF-IDF	タイトル・要旨参照法	0.102	0.071	0.136	0.077
	引用文脈参照法	0.324	0.220	0.415	0.232
	組み合わせた手法	0.328	0.223	0.421	0.236
SBERT	タイトル・要旨参照法	0.386	0.280	0.489	0.294
	引用文脈参照法	0.524	0.423	0.610	0.434
	組み合わせた手法	**0.539**	**0.434**	**0.628**	**0.456**

ある．また，MRR は以下の計算で表される．

$$MRR@k = \frac{1}{|\alpha|} \sum_{u \in \alpha} \frac{1}{rank_u} \tag{4}$$

ここで，u が正解論文の 1 つ，$rank_u$ が論文 u の候補論文のうち，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

対象論文の引用文脈と論文プール内の論文のタイトル・要旨の類似度並びに対象論文の引用文脈と引用文脈プール内の引用文脈との類似度を計算するためのベースラインとして TF-IDF を使用する．ベースラインと本手法の評価結果を表 1 に示す．まず，タイトル・要旨法も引用文脈参照法も，SBERT をモデルとして使った場合がベースラインを上回っていることが確認でき，SBERT による類似度の学習に効果があることが分かる．次に，引用文脈参照法はタイトル・要旨参照法よりも性能が良く，対象論文と既存論文の引用文脈同士の類似度を活用することに効果があることが分かる．

さらに，2 つの手法を組み合わせることで引用文脈参照法単体の場合よりも性能が向上していることが分かる．このことは，一度も引用されたことがない論文を推薦する場合に，タイトル・要旨参照法に組み合わせることによって，引用論文推薦の性能が向上する可能性があることを示している．

また，コールドスタート問題が起こる 692 件の引用文脈の評価結果を表 2 に示す．これは SBERT モデルのみでの評価となっている．引用文脈参照法での性能が 0 ではないのは，正解の論文が複数ある場合，すべての正解の論文が引用文脈をもたないとは限らないためである．この表から分かるように，コールドスタート問題が起こり得る場合において，タイトル・要旨参照法を組み合わせることで，引用文脈参照単独よりも性能が改善したことが分かる．

表 2 コールドスタート問題が起こる引用文脈の評価結果

手法	@5		@10	
	Recall	MRR	Recall	MRR
タイトル・要旨参照法	0.361	0.280	0.433	0.290
引用文脈参照法	0.017	0.027	0.024	0.030
組み合わせた手法	0.215	0.185	0.258	0.191

5 おわりに

本稿では，引用マーカーが与えられたときに，適切な被引用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．同じ被引用論文が引用されている引用文脈は類似する傾向があるという仮定に基づき，SBERT を用いて引用文脈同士の類似度を計算し，適切な論文を推薦するという引用文脈参照法を提案した．ただし，この手法には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ため，既存研究で広く用いられているタイトルと要旨を使用する手法と組み合わせることにより，引用文脈参照法単独よりも論文推薦の性能が向上することを示した．今後の課題としては，どのような意図で引用されているかを考慮した手法の検討や，引用回数に着目した手法の検討などが挙げられる．

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，電気通信大学 小山康平氏，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 南泰浩教授，工学院大学 大和淳司教授，名古屋大学 東中竜一郎教授，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 石井雅樹准教授，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] 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.

[4] 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.

[5] Travis Ebesu and Yi Fang. Neural citation network for context-aware citation recommendation. In **Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval**, pp. 1093–1096, 2017.

[6] Yang Zhang and Qiang Ma. Dual attention model for citation recommendation. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3179–3189, 2020.

[7] 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.

[8] 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.

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

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