

学術論文における引用文脈を用いた文献の推薦

生駒流季

名古屋大学大学院情報学研究科
ikoma.tomoki@h.mbox.nagoya-u.ac.jp

松原茂樹

名古屋大学情報連携推進本部
matubara@nagoya-u.jp

1 はじめに

学術論文では文献が引用されるが、引用される文献には、

- 文献の著者が自身の研究において参考にした概念やアイデアを提案した文献
- 研究の過程において利用したデータやツールなどの研究資源に関する文献
- 自身の研究に関連のある先行研究（関連研究）に関する文献

などがある。参考文献の探索は、発表される文献の増加とともに難しくなっており、その支援に向け、文献推薦システムが提案されている [1, 2, 3, 4, 5, 6, 7, 8].

これまでに提案された文献推薦システムの多くは、入力として利用者が執筆する論文のタイトルや要旨を受け取り、その論文で引用するのに適した文献を出力として提示する [9].

本稿では、論文における引用文脈を利用する文献推薦手法を提案する。引用文脈とは、論文において被引用論文が引用された箇所のテキストを指す。引用文脈には、引用した論文の著者がその文献を引用した目的や、自身の研究との関連性についての記述が含まれる。引用文脈を併用することにより、論文の著者による記述とは異なる観点や、他に引用した文献との関係を考慮した推薦文献の選定が可能となる。医学系論文を対象とした実験により、引用文脈を利用することによる推薦性能の向上を確認した。

本稿の2章では、本研究において想定する文献推薦の場面設定を説明する。3章では、引用文脈を用いることの利点について記す。4章では、提案する文献推薦システムの構成について述べる。5章では実験の結果を報告する。

2 問題の設定

文献推薦の研究は、推薦システムの利用者が執筆する論文のタイトルや要旨などの情報に対して、その論文において引用するのに適した文献を推薦する研究 [1, 2, 3, 4] と、論文の本体の一部を入力して、その箇所で引用するのに適した文献を推薦する研究 [5, 6, 7, 8] に大別される。前者は論文のタイトルと要旨が完成して本文を執筆中に利用する状況を想定し、後者は本文を執筆した利用者が、その部分の論旨の補強などを目的として文献を探索する状況を想定する。

本研究は前者を前提とする。すなわち、論文のタイトルと要旨を入力とし、その論文において引用するのに適した文献を出力する。

3 引用文脈とその利用

本章では、引用文脈について説明し、本研究における利用方法を記述する。

3.1 引用文脈

引用した文献における被引用文献の引用文脈とは、引用した文献の本文中における、被引用文献が引用された箇所の記述のことである。本研究では、被引用文献への引用を示す標識（引用タグ）を含む段落のテキストを引用文脈と定義する。

図1に、引用文脈の一例として、引用した文献 [10] における被引用文献 [11] の引用文脈の一部を示す。この図において、下線を引いた箇所が被引用文献に対応する引用タグである。また、図2に、被引用文献の要旨の一部を示す。

引用文脈には、引用した文献の著者がその文献を引用した目的や、自身の研究との関連性など引用した文献の著者にとって重要な情報が記述される。引用文脈中の記述は、被引用文献の内容のうち、引用した文献において重要な点をまとめたものとして捉えることができる。図1に例示した引用文脈では、

devastating impact on human health and healthcare systems. Take the example of the 1918 influenza pandemic, famously known as the “Spanish flu.” It is estimated that this outbreak killed between 17 and 50 million people worldwide.[3,4] Although the case mortality rate with this virus was estimated to be only 3%–5%, the virus was highly infectious and infected more than a third of the

図1 引用した文献 [10] における文献 [11] の引用文脈 (一部)

Abstract

Go to: ☑

Mortality estimates of the 1918 influenza pandemic vary considerably, and recent estimates have suggested that there were 50 million to 100 million deaths worldwide. We investigated the global mortality burden using an indirect estimation approach and 2 publicly available data sets: the Human Mortality Database (13 countries) and data extracted from the records of the *Statistical Abstract for British India*. The all-cause Human Mortality Database

図2 被引用文献 [11] の要旨 (一部)

COVID-19 と過去に発生した感染症の状況の比較を目的としてスペイン風邪の記録についての文献を引用した旨が記述されており、引用目的が示されている。

一方、要旨にはその文献の研究の概略や、提案した手法や概念、実施した実験の結果など、文献の著者がその文献において最も重要と考える情報が記述される。図2に例示した文献の要旨は、1918年に発生したスペイン風邪の当時の拡散状況などについての記録を再考した旨が、研究の概略として記されている。

3.2 引用文脈の利用

本研究では、引用文脈を入力論文の要旨と組み合わせて利用する。先行研究 [1, 2, 3, 4] の文献推薦システムは、入力論文の要旨と推薦候補文献の要旨を比較するが、推薦候補文献の要旨のみではその文献がどのような目的で、どのような点に着目して引用されたかを考慮した推薦文献の選択ができない。

一方、入力論文の要旨と推薦候補文献の引用文脈を比較することで、要旨には含まれない内容の記述を考慮に加えた推薦候補の選択が可能となる。図1および図2に示した例では、引用文脈を利用することにより、被引用文献の記述内容が他の感染症の記録との比較に利用できるという点を考慮した上で、推薦文献として選択することができる。

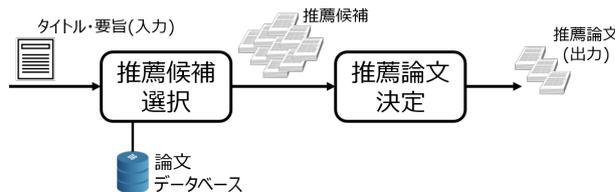


図3 提案システムの処理の流れ

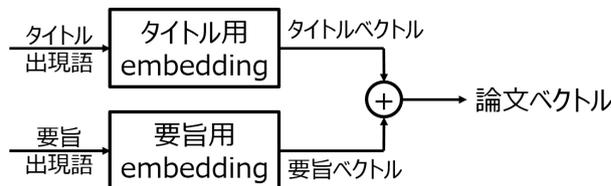


図4 推薦候補選択による文献のベクトル化

4 提案手法

本研究で提案する推薦システムは、推薦文献の候補を絞り込む段階と、推薦候補文献の中から推薦する文献を決定する段階を通して文献を推薦する。図3に処理の流れを示す。

4.1 推薦候補の選択

推薦候補選択の段階では、入力論文や文献データベース中の各文献を、タイトルや要旨の出現語の bag-of-words をもとにベクトル化する。図4にその概略を示す。推薦候補選択モデルはタイトルの出現語を処理するための embeddings と要旨の出現語を処理するための embeddings を学習し、タイトルと要旨のそれぞれについて、出現語の embeddings の重み付き和を算出してタイトルベクトルと要旨ベクトルを生成し、それらの重み付き和によってその文献をベクトル化する。

推薦候補選択のモデルの学習では、引用関係にある2文献のベクトルの cos 値が高く、引用関係にない2文献の cos 値は低くなるように各 embeddings と各語の重み、タイトルベクトル及び要旨ベクトルの重みを学習する。推薦候補選択モデルは、タイトルと要旨をもとに入力論文をベクトル化し、データベース中の各文献と入力論文のベクトルの cos 値を算出して、cos 値の高い上位一定数の文献と、それらによって引用されている文献からなる集合を推薦候補文献として出力する。

4.2 推薦文献の決定

推薦文献決定の段階では、各推薦候補文献が入力論文での引用にどの程度適しているかを示す推薦度

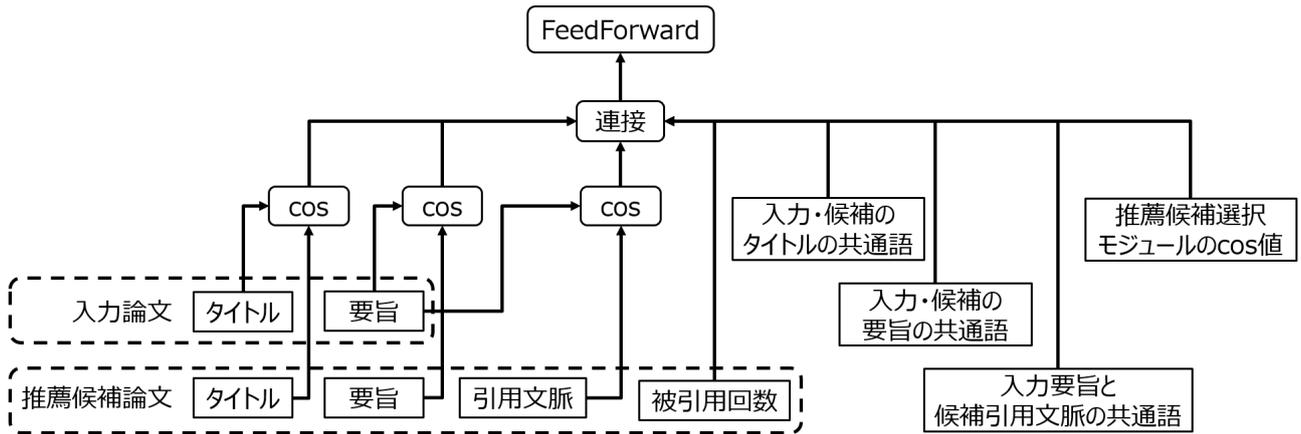


図5 推薦文献決定に用いるモデルの構成

を FeedForward により算出し、推薦度が一定の閾値を上回る文献を推薦文献として出力する。図5にモデルの構成を示す。本図において、入力論文は推薦システムに入力された文献のタイトルや要旨を指し、これらは推薦候補選択における入力と同じである。また、推薦候補文献は推薦候補選択で選出された各文献を指す。

文献 [1] で提案された手法は、以下の各項目を素性値として推薦度を算出し、推薦文献を決定している。

- 入力論文と推薦候補文献のタイトルベクトルの cos 値
- 入力論文と推薦候補文献の要旨ベクトルの cos 値
- 推薦候補選択モデルで計算された、入力論文・推薦候補文献間の cos 値
- 推薦候補文献の被引用回数 (FeedForward へは被引用回数の対数を入力)
- 入力論文と推薦候補文献のタイトルに共通して出現する語の重み値の和
- 入力論文と推薦候補文献の要旨に共通して出現する語の重み値の和

本研究の提案手法では、文献 [1] で採用されていた素性項目に加え、以下の各項目を素性として利用する。

- 入力論文の要旨ベクトルと、推薦候補文献の引用文脈ベクトルの cos 値 (引用文脈ベクトルは、推薦候補文献の引用文脈を要旨と同様の方法でベクトル化したものとする)
- 入力論文の要旨と、推薦候補文献の引用文脈に共通して出現する語の重み値の和

推薦文献決定モデルにおいて用いるタイトル用および要旨用の embedding は、推薦候補選択モデルで用いたものとは共通しない。学習では、タイトル用および要旨用の embeddings と各語の重みに加え、FeedForward 内で用いる各素性値の重みを学習する。

5 実験

本章では、文献推薦における引用文脈を利用することの効果を検証することを目的とした実験について記述する。

5.1 データセット

実験には医学系のオープンアクセス文献を集めた PMC Open Access Subset [12] を用いた。約 32 万本の文献を無作為に抽出し、文献の発行年で学習・開発・テストデータに分割した。本実験では、2018 年までに発行された文献 (約 25 万本) を学習データ、2019 年に発行された文献 (約 4.5 万本) を開発データ、2020 年に発行された文献 (約 2.5 万本) をテストデータとして用いた。

5.2 実験の概要

学習データを用いて、推薦候補選択モデルと、推薦文献決定モデルを、それぞれ 10 エポックで学習した。選択する推薦候補文献の数を 100 とし、テストデータ中の各文献を入力して出力された推薦文献と入力論文において引用されていた文献 (参考文献リスト) を比較して、以下の各指標で性能を評価した。

- 適合率: 推薦文献決定モデルによって推薦された文献のうち、参考文献リストにあるものの割合

表1 実験結果

	適合率	再現率	F 値	MRR
ベースライン	.0414	.1340	.0633	.1358
提案手法	.1130	.1472	.1279	.2749

表2 閾値の設定による性能の比較

閾値	適合率	再現率	F 値
.99	.0374	.0244	.0295
.95	.1212	.1515	.1347
.90	.1051	.2447	.1470
.85	.0753	.2976	.1202
.80	.0529	.3310	.0913
.75	.0383	.3524	.0691
.70	.0284	.3677	.0527
.60	.0173	.3838	.0330
.50	.0119	.3931	.0232

- 再現率: 参考文献リストにある文献のうち、推薦文献決定モデルによって推薦されたものの割合
- F 値: 適合率と再現率の調和平均

文献 [1] で提案された、引用文脈を用いない文献推薦システムの推薦性能をベースラインとし、各評価値を比較した。推薦文献決定における閾値は、ベースラインにおいて最も高い性能が記録された 0.95 と設定した。

5.3 実験結果

表 1 に実験結果を示す。本研究で提案したシステムはいずれの指標でもベースラインを上回っており、引用文脈を用いることの有用性が示された。

5.4 考察

5.4.1 推薦文献決定における閾値の設定

推薦文献決定で各文献を推薦するか否かを定める閾値と推薦性能の関係を考察するために、閾値の設定を様々に変えて、開発データにおける評価値の変化を観察した。表 2 にその結果を示す。閾値が下がるにつれて推薦される文献数は多くなるため、適合率は下がり再現率は上がっている。F 値では、閾値が 0.9 のときに最も高い評価値が記録された。

5.4.2 推薦候補選択における引用文脈の利用

本研究の提案手法では、引用文脈は推薦文献決定でのみ利用するものとしている。その一方で、推薦文献決定だけでなく推薦候補選択でも引用文脈を利

表3 引用文脈の利用方法による性能の比較

引用文脈の利用	適合率	再現率	F 値
なし	.0426	.1344	.0647
推薦候補選択のみ	.1096	.1291	.1186
推薦文献決定のみ	.1212	.1515	.1347
両方	.1046	.1191	.1114

用する手法も考えられる。そこで、推薦入力論文の要旨と各候補文献の引用文脈を推薦文献決定時と同様の方法でベクトル化し、両者間の cos 値を推薦候補選択の素性に追加し、この素性の有無による性能の変化を観察した。

本考察では推薦の閾値を 0.95 に設定し、引用文脈を使用しない、推薦候補選択でのみ用いる、推薦文献決定でのみ用いる（提案手法）、両方で用いるの 4 つの条件で、開発データにおける評価値を比較した。その結果を表 3 に示す。最も高い評価値が記録された条件は、推薦候補選択では引用文脈を用いず、推薦文献決定でのみ利用する条件であった。また、推薦候補選択のみで引用文脈を用いる条件と、両段階で用いる条件で記録された閾値はほぼ同じであった。この結果は、引用文脈を二重に考慮することの効果は薄いことを示唆する。

6 まとめ

本稿では、各候補文献を引用した他文献における引用文脈を利用して推薦文献を選定する文献推薦システムを提案した。また、医学系論文を対象とした実験により、引用文脈の利用が推薦性能の向上に貢献することを示した。

今後の課題として、引用文脈の取得のより適切な方法について検討することを挙げる。本研究では、引用文脈は一律に引用タグを含む段落と定義した。一方、文献 [13] では、引用を含む段落において、被引用文献の内容に言及した範囲を抜き出す手法が提案されている。このような手法を引用文脈の抽出に適用することで、引用文脈としてより適切な部分のみを用いた文献推薦が可能となると期待される。

参考文献

- [1] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 238–251. Association for Computational Linguistics, 2018.

- [2]Dai, Tao, Cai, Xiaoyan, Pan, Shirui, and Yuan, Sheng. Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. *J Ambient Intell Human Comput*, Vol. 9, pp. 957–975, 2018.
- [3]Mu, Dejun, Guo, Lantian, Cai, Xiaoyan, and Hao, Fei. Query-focused personalized citation recommendation with mutually reinforced ranking. *IEEE Access*, Vol. 6, pp. 3107–3119, 2018.
- [4]Guo, Lantian, Cai, Xiaoyan, Hao, Fei, Mu, Dejun, Fang, Changjian, and Yang, Libin. Exploiting fine-grained co-authorship for personalized citation recommendation. *IEEE Access*, Vol. 5, pp. 12714–12725, 2017.
- [5]Yang, Libin, Zheng, Yu, Cai, Xiaoyan, Dai, Hang, Mu, Dejun, Guo, Lantian, and Dai, Tao. A LSTM based model for personalized context-aware citation recommendation. *IEEE Access*, Vol. 6, pp. 59618–59627, 2018.
- [6]Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. hyperdoc2vec: Distributed representations of hypertext documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2384–2394, 2018.
- [7]Travis Ebesu and Yi Fang. Neural citation network for context-aware citation recommendation. Association for Computing Machinery, 2017.
- [8]Daniel Duma and Ewan Klein. Citation resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 358–363, 2014.
- [9]M.Färber, A. Jatowt. Citation recommendation: approaches and datasets. *Int J Digit Libr*, Vol. 21, pp. 375–405, 2020.
- [10]Khan, G., Sheek-Hussein, M., Al Suwaidi, A. R., Idris, K., and Abu-Zidan, F. M. (2020). Novel coronavirus pandemic: A global health threat. *Turkish journal of emergency medicine*, 20(2), pp. 55–62.
- [11]Peter Spreeuwenberg, Madelon Kroneman, John Paget. Re-assessing the global mortality burden of the 1918 influenza pandemic. *American Journal of Epidemiology*, Volume 187, Issue 12, 2018, pp. 2561–2567.
- [12]PMC Open Access Subset. <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>.
- [13]Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6181–6190. Association for Computational Linguistics, 2020.