学術情報推薦におけるグラフ構造の有効性検証

長尾浩良 [1]　桂井麻里衣 [2]
[1] 同志社大学大学院理工学研究科　[2] 同志社大学理工学部
{nagao21,katsurai}@mm.doshisha.ac.jp

概要

研究者への情報推薦は論文調査や関連研究者の探索に有用である．個々の研究者の業績一覧が入手可能な場合，そのコンテンツ情報を各人の専門興味のモデル化に用いることができる．一方，共著者情報が適切に整備されている場合，研究者や論文をノードとしたグラフ構造に基づくモデル化手法を採用できる．近年はこのようなグラフベース手法が注目を集めているが，コンテンツ情報のみでのモデル化手法と同一条件下で比較した例は未だ報告されていない．本論文では，学術情報推薦をグラフ内のリンク予測問題とみなし，これら二種類の手法を比較する．実験方法を情報推薦の観点から再設計した結果，コンテンツベース手法がグラフベース手法より高い性能を示した．また，リンク予測の観点での推薦性能と推薦結果の意外性はトレードオフの関係にあることが示唆された．

1　はじめに

インターネット上では膨大な量の学術情報が蓄積されており，個々の研究興味に基づいて論文や関連研究者を容易に取得可能なシステムが必要とされている．初期の研究では，ユーザがキーワードを入力し，所望の情報を自身で検索することを前提としていた [1, 2]．近年は，論文テキストなどのコンテンツ情報から特徴量を抽出し，論文間の推薦 [3] や論文に対する査読者の割り当て [4] などに活用した例がある．また学術ドメインに限らず，ユーザとアイテムの関連性をモデリングし推薦に活用する研究はこれまで盛んに行われている [5, 6]．以降，このような手法を**コンテンツベース手法**と呼ぶ．

各論文の著者，所属機関，発表場所などの情報が整備された学術データベースを利用できる場合，それぞれの実体をノード，それらの関係をエッジで表した異種混合グラフが構築できる．グラフニューラルネットワーク（GNN）[7, 8] を用いると，コンテンツ情報をノードの初期特徴量とみなしたうえで，実体間の関係を考慮して各ノードをモデリングできると考えられる．こうした**グラフベース手法**を情報推薦に導入する動きが近年活発化しており，一般には異なるノード間のエッジの有無を予測する問題（以降，リンク予測）として定式化される [9]．しかし，従来研究は主にグラフベース手法に限定してリンク予測性能を検証しており，コンテンツベース手法と比較した例は報告されていない．加えて，評価用データはグラフ要素全体からのランダムサンプリングに基づいて作成されるため，性能評価指標に対する各ノードの寄与率がばらつく可能性がある．手法の実応用を考えると，情報検索・推薦の文脈でよく用いられてきた指標である，個々のノードのリンク予測性能の平均を算出する方が望ましい．

そこで本研究では，(1) 研究者ノード・論文ノード間のエッジ有無と，(2) 研究者ノード間のエッジ有無という二種類のリンク予測を対象に，学術情報推薦への応用に即した条件下でグラフベース手法とコンテンツベース手法の性能を比較する．なお，(1) の予測は論文推薦，査読者推薦，著者同定などに有用であり，(2) の予測は共同研究者候補の探索などが応用先に挙げられる．実験では，リンク予測性能に加え，推薦結果の意外性・多様性を定量評価する．

2　データセット

リンク予測のためのデータセットとして，SemOpenAlex-SemanticWeb（SOA-SW）[10] を用いる．SOA-SW は大規模な学術知識グラフである SemOpenAlex [11] のサブセットであり，次の二種類のノードとその関係データが含まれる：(i) セマンティックウェブ分野の論文を少なくとも 1 件持ち，合計 3〜200 件の論文を持つ著者．(ii) これらの著者が持つ論文のうち，アブストラクトがあり，発表年が 2005 年以降で，被引用数が 10 件以上の論文．表 1，2 にノード，エッジについての統計量をそれぞれ示す．

Work ノードにはテキスト特徴量として，論文の

表 1 SOA-SW のノードについての統計量

ノードタイプ	ノード数
Work	95,575
Author	19,970
Concept	38,050
Source	10,739
Institution	5,846
Publisher	786

表 2 SOA-SW のエッジについての統計量

エッジタイプ	エッジ数
Work-Concept	1,320,949
Work-Source	247,667
Work-Work	115,271
Author-Work	112,565
Author-Author	38,632
Author-Institution	19,281
Source-Publisher	1,781

タイトルとアブストラクトを連結したテキストの SciBERT [12] による埋め込み[1)]が与えられている．それ以外のノードには日付などの数値や著者の貢献度などのカテゴリ値が付与されている．エッジに関しては，特定のエッジタイプのみ数値やカテゴリ値が付与されている．本研究では，特徴量の利用可否が各手法の性能に影響を与えることを避けるために，全ての手法で Work ノードのテキスト特徴量のみを用いることとする．

3 手法

3.1 グラフベース手法

グラフのノード表現を獲得するために，本研究では Heterogeneous Graph Attention Network（HAN）[7] および Heterogeneous Graph Transformer（HGT）[8] を採用する．いずれも，Attention に基づく GNN である Graph Attention Network [13] の，異種混合グラフ向けの拡張である．HAN はノードタイプ間の経路（メタパス [14]）をユーザが事前に定義する必要があり，HGT はその必要がない．メタパスの詳細は付録 A で述べる．

研究者・論文間の推薦，および研究者間の推薦を，それぞれ SOA-SW 上の Author-Work エッジ，Author-Author エッジに対するリンク予測タスクとして定義する．実験では，Work ノードはデータセット内のテキスト特徴量を用いることとし，それ以外のノードについては，ノード ID に基づく one-hot ベクトルで初期化する．ノード対 (u, v) のエッジに関するスコア関数 $\mathrm{score}(u, v)$ として，パラメータ不要で計算が容易な内積を採用する．

$$\mathrm{score}(u, v) = h_u^T h_v \tag{1}$$

ここで，h_u，h_v はそれぞれ，GNN エンコーダから得られたノード表現である．学習時には，グラフ上に存在するエッジを正例，存在しないエッジを負例とみなし，損失関数として二値交差エントロピー損失を用いる．具体的には，正例と負例をまとめたエッジ集合とそのスコア $\mathrm{score}(u, v)$ に対しシグモイド関数を適用し，エッジへの予測確率を得る．そして予測確率と正解ラベルから損失関数を最適化する．

3.2 コンテンツベース手法

コンテンツ情報のみに基づく研究者・論文の特徴表現手法として，BERT4Rec [6]（以降，単に BERT と呼ぶ）と，その BERT 部分を LSTM に変更したモデルの二種類を比較する．これらは本来ノード（情報推薦の文脈ではアイテム）の ID を入力として受け取るが，各ノードに何らかの特徴量を用意できる場合，ID から特徴量への変換層を省略する形でそれらを初期特徴量として用いることが可能である．

本研究では，Author ノードの表現を，その 1-hop 近傍にあたる Work ノード集合から算出することを考える[2)]．具体的には，Work ノード集合をノード ID 順に並べ，系列としてモデルに入力する．BERT については，出力された各 Work ノードの表現を平均プーリングによって一つの Author 表現へと集約する．LSTM については，系列に対する次ノード予測の結果として一つの Author 表現を得る．Work ノードに関しては，対応する特徴量を，Author ノードの表現と同じ次元となるよう線形変換する．それ以外のノードについては存在しないものとみなす．以上はグラフベース手法と同様，リンク予測タスクとして定式化し，スコア関数には内積を用いる．

[1)] 平均プーリングが使用されている．また，最初の 128 次元のみが提供されている．

[2)] ここでは便宜上グラフ構造の部分的な使用として説明しているが，実際には各研究者の論文リストを入手すればよいだけであり，グラフ構造が十分整備されていない学術データベースにおいても適用しうる手法である．

3.3 ベースライン手法

単純な方法として，Author ノードの表現を，その 1-hop 近傍にあたる Work ノードの特徴量を平均し一つの表現に集約することを考える．Work ノードの表現は，データセットの特徴量をそのまま用いる．グラフベース手法と同様，リンク予測タスクとして定式化し，スコア関数には内積を用いる．

4 実験

4.1 評価における負例サンプリング戦略

一般にリンク予測の評価では，実験対象とするエッジタイプ（Author-Work または Author-Author）について，データセット内に存在する全てのエッジを正例とみなす．負例については，エッジの存在しないノード対 (u, v) をそのまま用いると数が多すぎることから，ランダムサンプリングすることがほとんどである．しかし，この単純な方法では，各ノードが関係する負例の数にばらつきが生じてしまい，ノードによっては負例が一つも存在しないこともありえる．この状況で評価すると，各ノードが評価指標に与える影響が一様とならず，特定のノードでの性能が支配的になる可能性がある．そこで本実験では，個々のノードからみた負例エッジの数が全てのテストノードで同一となるように強制したうえで，テストノードをクエリとみなしてそれに関するリンク予測性能を評価し，最後に全てのテストノードで平均を算出する．これは Mean Average Precision（MAP）に相当し，従来の方法に比べて実際の推薦に即した評価となる．以降，4.2 節では従来研究と同様の評価方法を，4.3 節では上で述べたような情報推薦的な評価方法をそれぞれ採用する．

4.2 従来の評価方法でのリンク予測性能

表 2 の Author-Work エッジ，Author-Author エッジをそれぞれ 8 : 1 : 1 の割合で訓練用，検証用，テスト用へと分割した．そして，従来の評価実験と同様，訓練用，検証用のそれぞれで正例と負例の割合が 1 : 1 となるよう負例をサンプリングした．その他の実験設定は付録 B を参照されたい．評価指標として ROC-AUC を算出した．

Author-Work，Author-Author エッジそれぞれに対するリンク予測性能を表 3 に示す．ここで，上位二位となる数値を太字で示す．コンテンツベース手法とグラフベース手法を比較すると，予測対象とするエッジタイプによらず一貫してコンテンツベース手法の方が性能が高い．続いて，ベースライン手法とグラフベース手法を比較すると，両エッジタイプともに HGT がベースライン手法を凌駕する．しかし，HAN については，Author-Author エッジではベースライン手法および HGT を上回る一方，Author-Work エッジにおいてはベースライン手法よりも著しく性能が低下した．HAN のような手法は利用するメタパスが性能に影響するため，ドメイン知識に基づき慎重にメタパスを設計する必要があるといえる．

表 3 Author-Work エッジおよび Author-Author エッジに対する従来の評価方法でのリンク予測性能

	Author-Work	Author-Author
Model	ROC-AUC	ROC-AUC
Baseline	0.901	0.847
LSTM	**0.976**	**0.972**
BERT	**0.991**	**0.967**
HGT	0.934	0.912
HAN	0.700	0.938

4.3 情報推薦としてのリンク予測性能

4.1 節で述べた負例サンプリング戦略を導入し，クエリノードごとに正例と負例の割合が 1: 100 となるよう負例をサンプリングした．情報推薦的な評価尺度として Hit Rate@1，Hit Rate@10，MAP@10 をそれぞれ算出した．その他の設定は前節の実験と同様である．Author-Work エッジ，Author-Author エッジに対する推薦の性能をそれぞれ表 4，表 5 に示す．表中の HR は Hit Rate を指す．コンテンツベース手法とグラフベース手法を比較すると，前節の実験と同様，どちらのエッジタイプにおいてもコンテンツベース手法の方が高い性能を示した．一方，前節の実験とは異なり，グラフベース手法はベースライン手法に劣る結果となった．つまり，情報推薦に関する評価指標ではグラフ構造の有用性が示せておらず，これは従来研究の評価実験と実応用の間のギャップを示唆している．

4.4 推薦結果の意外性・多様性の評価

最後に，推薦結果の意外性・多様性という観点に基づいて各手法の性能を評価した．意外性の指標として，新たに 1-Hop Ratio および Mean Hop を定義す

表 4 情報推薦の観点による Author-Work エッジのリンク予測性能

Model	HR@1	HR@10	MAP10
Baseline	0.481	0.848	0.581
LSTM	**0.552**	**0.926**	**0.659**
BERT	**0.647**	**0.966**	**0.739**
HGT	0.236	0.807	0.397
HAN	0.145	0.448	0.241

表 5 情報推薦の観点による Author-Author エッジのリンク予測性能

Model	HR@1	HR@10	MAP10
Baseline	0.186	0.642	0.306
LSTM	**0.461**	**0.903**	**0.588**
BERT	**0.544**	**0.919**	**0.651**
HGT	0.067	0.388	0.141
HAN	0.074	0.518	0.181

る．1-Hop Ratio は，クエリノードとのスコアが高いものとして推薦されたノードのうち，そのクエリノードからグラフ上で 1-hop（距離 1）であるノードの割合を表す．この値が高いほど，グラフ上で隣接したノードばかりが推薦されることを意味し，結果の意外性が低いと考えられる．反対に，1-Hop Ratio が低いほど意外性が高いことを表す．Mean Hop は，推薦されたノードとクエリノードの間の平均距離である．Mean Hop の値が大きいほど，クエリの隣接ノードよりも遠くのノードを推薦できていることを表し，結果の意外性が高いといえる．また，推薦結果の多様性の指標として，推薦されたノード間の平均距離である Intra-List Distance（ILD）[15] を採用する．ILD が高い値であるほど推薦されたノード同士がグラフ上で局所に集中しないということを意味し，推薦内容の多様性が高いといえる．

Author-Work および Author-Author エッジに対する推薦の意外性・多様性の評価結果をそれぞれ表 6，表 7 に示す．表中の 1-HR，MH はそれぞれ 1-Hop Ratio，Mean Hop を表す．まず，いずれのエッジタイプにおいてもグラフベース手法はコンテンツベース手法より高い意外性を示した．ベースライン手法と比較すると，推薦結果上位 1 件ではグラフベース手法の方が意外性が高い．Author-Author エッジの推薦結果上位 10 位の MH ではベースライン手法が最高スコアを示したが，それ以外の意外性の指標はグラフベース手法の方が良い性能を示した．ただし，

表 6 Author-Work エッジに対する推薦の意外性・多様性

Model	1-HR		MH		ILD
	@1	@10	@1	@10	@10
Baseline	0.483	0.105	2.035	2.826	2.298
LSTM	0.554	0.119	1.894	2.794	2.302
BERT	0.651	0.128	1.696	2.784	2.349
HGT	**0.236**	**0.098**	**2.579**	**2.885**	**2.373**
HAN	**0.146**	**0.060**	**2.822**	**3.003**	**2.418**

表 7 Author-Author エッジに対する推薦の意外性・多様性

Model	1-HR		MH		ILD
	@1	@10	@1	@10	@10
Baseline	0.191	0.080	3.383	**3.722**	**3.993**
LSTM	0.477	0.133	2.459	3.506	3.832
BERT	0.559	0.136	2.250	3.515	**3.873**
HGT	**0.078**	**0.071**	**3.641**	**3.712**	3.768
HAN	**0.077**	**0.076**	**3.655**	3.690	3.839

上位 10 位での意外性に関する値の差は，上位 1 位の指標での差と比べて小さい．

まとめると，リンク予測に基づく推薦性能（4.3 節）ではコンテンツベース手法，ベースライン手法，グラフベース手法という順に性能が高かったのに対し，推薦の意外性においてはその順序が逆転していた．このことから，推薦性能と推薦結果の意外性はトレードオフの関係にあることが示唆される．

一方，推薦結果の多様性に着目すると，Author-Work エッジについてはグラフベース手法が最も高いが，Author-Author エッジについてはベースラインが最も高く，次いで BERT となった．エッジタイプの性質によって指標の挙動が異なる可能性を確認できたことから，従来のリンク予測評価実験の枠組みに多角的な評価を導入することが必要といえる．

5 おわりに

学術情報推薦をリンク予測問題と考え，コンテンツベース手法とグラフベース手法の性能を比較した．情報推薦としての評価方法を採用した結果，コンテンツベース手法の方が高い性能を示した．従来の実験方法ではグラフベース手法の有用性主張に不十分であった可能性と，推薦評価指標と結果の意外性がトレードオフの関係にあることが示唆された．今後は各手法の性質について，グラフ構造の複雑さや，ノード特徴量の点から検証する予定である．

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP20H04484）の助成を受けたものです.

参考文献

[1] Suzanne Fricke. Semantic scholar. **Journal of the Medical Library Association: JMLA**, Vol. 106, No. 1, p. 145, 2018.

[2] Jian Wu, Kunho Kim, and C. Lee Giles. Citeseerx: 20 years of service to scholarly big data. In **Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse**, AIDR '19, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics.

[4] Ivan Stelmakh, John Wieting, Graham Neubig, and Nihar B. Shah. A gold standard dataset for the reviewer assignment problem, 2023.

[5] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In Yoshua Bengio and Yann LeCun, editors, **4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings**, 2016.

[6] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In **Proceedings of the 28th ACM International Conference on Information and Knowledge Management**, CIKM '19, p. 1441–1450, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In **The World Wide Web Conference**, WWW '19, p. 2022–2032, New York, NY, USA, 2019. Association for Computing Machinery.

[8] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In **Proceedings of The Web Conference 2020**, WWW '20, p. 2704–2710, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. **IEEE Trans. on Knowl. and Data Eng.**, Vol. 34, No. 10, p. 4854–4873, October 2022.

[10] Michael Färber, David Lamprecht, and Yuni Susanti. Autordf2gml: Facilitating rdf integration in graph machine learning. In Gianluca Demartini, Katja Hose, Maribel Acosta, Matteo Palmonari, Gong Cheng, Hala Skaf-Molli, Nicolas Ferranti, Daniel Hernández, and Aidan Hogan, editors, **The Semantic Web – ISWC 2024**, pp. 115–133, Cham, 2025. Springer Nature Switzerland.

[11] Michael Färber, David Lamprecht, Johan Krause, Linn Aung, and Peter Haase. Semopenalex: The scientific landscape in 26 billion rdf triples. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, **The Semantic Web – ISWC 2023**, pp. 94–112, Cham, 2023. Springer Nature Switzerland.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In **International Conference on Learning Representations**, 2018.

[14] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. Pathsim: meta path-based top-k similarity search in heterogeneous information networks. **Proc. VLDB Endow.**, Vol. 4, No. 11, p. 992–1003, August 2011.

[15] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In **Proceedings of the 14th International Conference on World Wide Web**, WWW '05, p. 22–32, New York, NY, USA, 2005. Association for Computing Machinery.

表 8　ハイパーパラメータ

バッチサイズ	2048
最大エポック数	300
オプティマイザ	Adam
学習率	1e-3
エンコーダレイヤ数	2
Attention ヘッド数	8
隠れ層の次元	64
最終的なノード表現の次元	64

A　メタパスの説明

例えば，Author を著者ノード，Work を論文ノードと定義し，Author と Work の間に，「論文を書く」（あるいは「論文が書かれる」）という関係を定義する．このとき，AWA というメタパスは，同じ論文を書いた，つまり共著者であるという関係を表す．このように，ドメイン知識に基づいてメタパスを定義することによって，特定のノード間でメッセージの伝達を行うことができ，タスクにとって有用な処理が可能となる．

B　実験設定

B.1　ハイパーパラメータ

実験に使用したハイパーパラメータを表 8 に示す．訓練時の最大エポック数は 300 とし，30 エポックの間損失に改善が見られない場合訓練を中断する．また，検証データでの評価値が最も高かったエポックのモデルパラメータを使用してテストを行う．

B.2　HAN のメタパス

表 9 に，HAN においてメッセージ伝達の際の近傍を制御するために用いるメタパスを示す．ここで，A は Author，W は Work，C は Concept，S は Source，P は Publisher，I は Institution をそれぞれ表す．

表 9　HAN のメタパス

タスク	メタパス
Author2Work	AWAW
	AWCW
	AWSW
	AWSPSW
	AIAW
Author2Author	AWA
	AWCWA
	AWSWA
	AWSPSWA
	AIA