論文の被引用数予測に基づく共同研究者推薦についての一考察

武田　悠佑† 　岩田　具治†† 　澤田　宏††

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
†† 日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2-4
E-mail: †takeda.yusuke.tr0@is.naist.jp, ††{iwata.tomoharu,sawada.hiroshi}@lab.ntt.co.jp

あらまし　適切な研究者同士が協働して取り組んだ共同研究は良い成果を生む傾向にあることが知られている．一方で，研究者が自身に適合する共同研究者を見つけることは容易ではないため，研究者に対して適切な共同研究者を推薦することは有用である．本論文では，より良い成果を生んだ研究は結果としてより高い被引用数を得ると仮定した上で，効果的な共同研究者をより高い被引用数を得るような共同研究を実現する研究者と定義し，効果的な共同研究者の推薦手法を提案する．共同研究者を推薦するための手法はこれまでにも数多く提案されている．しかし，先行研究の多くでは共同研究者推薦を共著ネットワーク上のリンク予測問題として定式化しているのみで，共同研究による成果やその結果としての論文の被引用数については着目されてこなかった．提案法では，推薦を受ける研究者と共同研究者の候補者による共著論文の被引用数を予測し，予測結果に基づいて候補者を順序付けすることで効果的な共同研究者の推薦を実現する．実験の結果，論文の被引用数予測について可変長の集合データを入力とできる確率的ニューラル回帰モデルがより良い性能を示すことが確認された．

キーワード　共同研究者推薦, 被引用数予測, ニューラル回帰モデル

1. はじめに

適切な研究者同士が協働して取り組んだ共同研究は良い成果を生む傾向にあることが知られている [1,2]．一方で，研究者が自身に適合する共同研究者を見つけることは容易ではない．そのため研究者に対して効果的な共同研究者を推薦することは有用であると考えられ，これまでに数多くの共同研究者推薦の手法が提案されてきた．しかし，先行研究の多くにおいては，共同研究者推薦は共著ネットワーク上のリンク予測問題として定式化され，共同研究による成果やその結果としての論文の被引用数については着目されてこなかった．一方で，共同研究者を推薦する目的が良い成果を生む研究の促進であるならば，共同研究者推薦をリンク予測問題として扱う，つまり研究者が将来的にどの研究者と共同研究するかを予測するだけでは不十分だと考えられる．

本研究では，より良い成果を生む共同研究の促進を目的として，共同研究による成果の予測に基づく共同研究者推薦手法を提案する．ただし，研究の成果を定量的に表現することは難しいと考えられるため，本論文では，より良い成果を生んだ研究は結果としてより高い被引用数を得ると仮定した上で，予測する研究の成果をその研究についての論文の被引用数で代替する．提案法では，推薦を受ける研究者と共同研究者の候補者による共著論文の被引用数を予測し，予測結果に基づいて候補者を順序付けすることでより良い成果を生む研究の実現に効果的な共同研究者の推薦を実現する．提案法の概要を図 1 に示す．

論文の被引用数を予測するモデルとして，本論文では，Deep Averaging Network（DAN）[3] を基礎とするニューラル回帰モデルを提案する．提案モデルでは，研究者は潜在ベクトルで表現され，論文の被引用数は著者集合内の研究者の潜在ベクトルの平均をニューラルネットワークで変換した値として表現される．潜在ベクトルを著者集合について平均して用いることからこのモデルでは可変長の入力を扱えることが特徴である．また提案法は，潜在ベクトルの生成を確率的にして，事前分布を設定することで，研究者ごとの論文の被引用数の分散を考慮できる．

以下の本論文の構成は次のとおりである．2 章では共同研究者推薦の関連研究について述べ，本研究との差異を示す．3 章では，提案法における共同研究者推薦の定式化と提案法の詳細について説明する．4 章では，実世界データとしてコンピュータサイエンス分野における論文を用いた実験について述べ，提案法が論文の被引用数予測と共同研究者の順序付けにおいて比較手法と同等以上の性能を示すことを確認する．5 章では，結論と今後の展望について述べる．

2. 関連研究

共同研究者推薦の手法についてはこれまでに様々な研究がなされてきた．ネットワークベースの手法としては，共著関係から研究者についてのネットワークを構築し，Random Walk with Restart [4] を用いてノードとなる研究者同士の類似度を算出することで共同研究者を探索する手法が提案されている [5,6]．共著ネットワーク上にリンクを持たないような研究者に対する共同研究者推薦では，研究者の属性情報を用いたコンテンツベースの推薦手法が提案されている [7]．また，ネットワークベースの手法とコンテンツベースの手法を組み合わせた共同研究者推

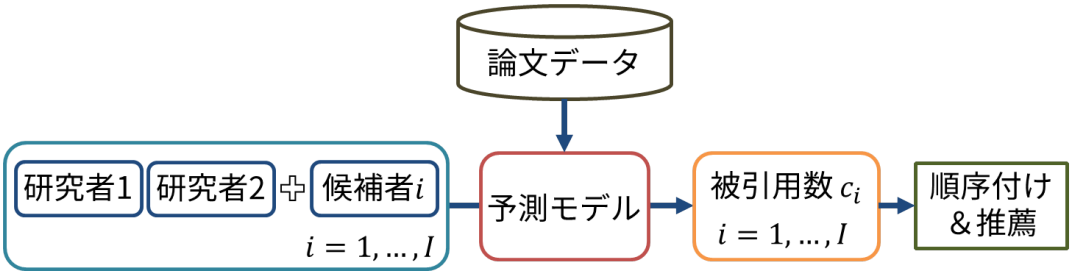

図 1: 提案法の概要

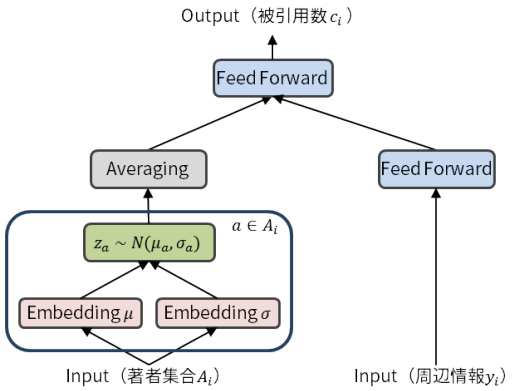

図 2: 提案モデル

薦手法も複数提案されている [8–10]．一方で，Araki らは，異分野の共同研究者推薦について，共著ネットワーク上におけるリンク予測問題として定式化するのではなく，研究内容の類似性に基づいて共同研究者を推薦する手法を提案している [11]．しかし，いずれの研究においても，共同研究による成果や共著論文の被引用数は着目されておらず，評価実験においてはリンク予測のタスクを用いて手法の評価を試みている．これらの研究に対して本研究は，共同研究の成果と論文の被引用数の相関を仮定した上で，論文の被引用数予測に基づく共同研究者推薦に取り組み，リンク予測ではなく論文の被引用数予測を手法の評価タスクに用いる点で異なる．

3. 提 案 手 法

本研究では，より良い成果を生んだ研究についての論文は結果としてより高い被引用数を得ると仮定し，より高い被引用数を得る論文に結びつく共同研究を実現する研究者を効果的な共同研究者として定義する．提案法では，まず推薦の前段階として，著者集合を入力として論文の被引用数を出力する予測モデルを構築し，過去に発表された論文についてのデータを用いて予測モデルを学習する．推薦の際には，推薦を受ける研究者と共同研究者の各候補者が共著で論文を執筆した場合の被引用数を予測モデルで予測し，各候補者を被引用数の予測値によって順序付けして推薦する．すなわち，提案法において共同研究者推薦は著者集合からの論文の被引用数を予測する回帰問題として定式化される．なお，従来の共同研究者推薦が目的としていた将来的な共同研究者の予測は，提案法では考慮しないものとする．

3.1 予測モデル

本節では，提案法において著者集合からの共著論文の被引用数予測において用いる予測モデルを説明する．提案モデルは，文書分類において用いられる Deep Averaging Network（DAN）[3] を基礎としたニューラルネットワークを用いる回帰モデルである．DAN は，入力されたデータの各要素に潜在ベクトルを割り当てた上でその平均ベクトルを算出し，ニューラルネットワークによって平均ベクトルを固定長の出力ベクトルへと変換するモデルであり，可変長の入力を扱えることが特徴である．提案モデルは，DAN を基礎として用いることで，論文ごとに大きさが異なる著者集合からの被引用数への変換を表現できる．これにより，従来の共同研究者推薦手法の多くが 1 人の研究者に対して 1 人の共同研究者を推薦する 1 対 1 の推薦であったのに対して，提案モデルでは推薦する側の人数と推薦される側の人数を 1 に限定しない多対多の推薦を扱える．

提案モデルのベースとなるモデルは以下の式で表現される．

$$\hat{c_i} = f\left(\frac{1}{|A_i|}\sum_{a \in A_i} z_a\right). \tag{1}$$

ここで，A_i は入力となる著者集合，z_a は研究者 a の潜在ベクトル，$\hat{c_i}$ は著者集合 A_i による共著論文 i の被引用数の予測値，関数 f はニューラルネットワークによる変換を表すものとする．このモデルでは，各研究者は固有の潜在ベクトルを持ち，著者集合はそこに含まれる研究者の潜在ベクトルを平均した代表ベクトルで表現され，論文の被引用数は著者集合の代表ベクトルを変換したものとして表現される．このモデルの学習は，被引用数が既知の論文 i について以下の損失関数 L を最小化することで行える．

$$L = \sum_{i \in I}(c_i - \hat{c_i})^2. \tag{2}$$

ここで，c_i は論文 i の真の被引用数であり，I は論文の集合である．

このモデルをベースモデルとして提案モデルでは以下の 2 つの拡張を加える．

（1） 論文の周辺情報を追加利用

（2） 研究者の潜在ベクトル z を確率的な生成

1 つ目の拡張は，発表媒体や発表日時など論文 i についての周辺情報を y_i として，次の式のように表される．

$$\hat{c_i} = f\left(\text{concat}\left(\frac{1}{|A_i|}\sum_{a\in A_i} z_a, g(y_i)\right)\right). \quad (3)$$

ここで，関数 g は f とは異なるニューラルネットワークによる変換であり，$concat$ は 2 つのベクトルを結合する操作を表す．この拡張では，論文 i の周辺情報をニューラルネットワークによって変換したベクトルを著者集合の代表ベクトルと結合することで，ベースとなる予測モデルよりも多くの情報を用いて論文の被引用数予測を行える．

2 つ目の拡張は，事前分布を仮定した上で，研究者の潜在ベクトルを確率的に生成することで，研究者ごとの論文の被引用数の分散を考慮したものであり，次の式で表現される．

$$z_a \sim \mathcal{N}(\mu_a, \text{diag}(\sigma_a)). \quad (4)$$

ここで，μ_a は研究者 a の平均値パラメータベクトル，σ_a は研究者 a の分散パラメータベクトルである．式 4 は，研究者ごとの潜在ベクトル z が $\mu, diag(\sigma)$ のパラメータを持つ正規分布からサンプリングされることを表している．また z の生成分布に $N(0, I)$ という事前分布を設定する．提案モデルではこれらの拡張により，頑健な被引用数予測モデルの構築を試みる．

なお，これらの拡張により損失関数 L は，式（2）の損失関数に事前分布と事後分布の Kullback-Leibler divergence の項を加えた次の式となる．

$$L = \sum_{i\in I}\left((c_i - \hat{c_i})^2 + \sum_{a\in A_i}\text{KL}(\mathcal{N}(\mu_a, \text{diag}(\sigma_a)|\mathcal{N}(0, I))\right). \quad (5)$$

提案モデル全体の構造を図 2 に示す．ここで，過去の論文データによって学習されるパラメータは，研究者ごとに割り当てる平均値パラメータベクトル μ と分散パラメータベクトル σ，そして f, g の 2 つのニューラルネットワークのパラメータとなる．

4. 実　　験

4.1 データセット

実験では，Arnetminer [12] で公開されている，DBLP に登録された論文についてのデータセットを用いて，提案モデルによる論文の被引用数予測の性能を検証する．本論文では，1996 年から 2007 年までの間に発表された論文のうち，表 1 に示す，データマイニングや機械学習に関連する計 14 の国際会議で発表された論文を用いる．訓練データと検証データには，1996 年から 2005 年までに発表された論文のデータを 90 : 10 の割合で分割して用いる．また，評価データには 2006，2007 年に発表された論文のデータを用いる．なお訓練データと検証データには合計で 10,320 本，評価データには 4,564 本の論文情報が含まれている．

4.2 実 験 設 定

実験では，訓練データと検証データの組を無作為抽出によって 10 組作成し，各データ組を用いてモデルを学習した場合の評価データに対する予測性能を評価する．なお，データ中の論文の被引用数 c_i については，1 を加えて対数変換した上で，学習と評価を行なう．評価では，各データ組について，異なるハイパーパラメータを持つ複数のモデルで学習を行った後，検証データについて最も高い性能を示したモデルを利用する．

表 1: 実験に用いる論文の発表会議

”AAAI”，”NIPS”，”ICML”，”KDD”，”IJCAI”，”ICDM”，”ACL”，”CIKM”, ”ECML”，”PAKDD”，”VLDB”，”PKDD”，”WWW”，”EMNLP”

表 2: 被引用数の予測誤差

手法	RMSE
線形回帰	1.352
MLP	1.622
LSTM	1.447
ベースモデル	1.458
ベースモデル＋拡張 2	1.372
線形回帰＋拡張 1	1.321
MLP ＋拡張 1	1.582
LSTM ＋拡張 1	1.310
ベースモデル＋拡張 1	1.376
ベースモデル＋拡張 1 ＋拡張 2	**1.308**

比較手法には次に示す 5 つの手法と，それぞれに拡張 1 を適用した手法の計 10 手法を用いる．

線形回帰： 著者集合を Bag-of-Words のベクトルとして表現したものを入力として被引用数を線形回帰で予測するモデル．

MLP： 著者集合を Bag-of-Words のベクトルとして表現したものを入力として被引用数を隠れ層 2 層の多層パーセプトロンで予測するモデル．

LSTM： 研究者の潜在ベクトルを平均する代わりに，可変長の入力を扱えるリカレントニューラルネットワークの一種である双方向 LSTM を用いて著者集合の代表ベクトルを計算するモデル．このモデルでは論文における著者の順序が考慮される．

ベースモデル： 式 1 で表現されるモデル．

ベースモデル＋拡張 2： 提案モデルに拡張 2 を適用したモデル．

なお，拡張 1 における論文の周辺情報としては，論文が発表された会議を one-hot ベクトルで表したものと，論文が発表された年をデータセットの構築年から引いた値を結合したベクトルを利用する．また，ベースモデルと拡張 1 におけるニューラルネットワークとしては，隠れ層 2 層の多層パーセプトロンを用いる．評価では，これらの手法について回帰問題と順序付けの 2 タスクについて比較する．なお，実験では既存の論文を評価に用いるため評価時においても著者順の情報を利用するが，本来はまだ存在しない論文についての被引用数を予測するため推薦時において著者順の情報は利用できない．このため，LSTM のモデルによる実験結果は参考値となる．

表 3: 順序付けの正答率

手法	順序正答率
線形回帰	0.515
MLP	0.512
LSTM	0.492
ベースモデル	0.512
ベースモデル＋拡張 2	0.513
線形回帰＋拡張 1	0.560
MLP ＋拡張 1	0.546
LSTM ＋拡張 1	0.577
ベースモデル＋拡張 1	0.577
ベースモデル＋拡張 1 ＋拡張 2	**0.588**

4.3 実験結果

回帰問題のタスクでは，著者集合から論文の被引用数を予測する性能について評価する．評価指標には，回帰問題の評価として予測値と正解値の平均平方二乗誤差（RMSE）を用いる．10 組のデータについての実験結果の平均を表 2 に示す．表中で太字の数字は比較手法間で最も性能が高いことを表す．ベースモデルに 2 つの拡張を適用した提案モデルが最も高い性能を示し，提案モデルが著者集合からの論文の被引用数予測において有用なモデルであることが示唆される．

順序付けのタスクでは，共同研究者の候補者を効果的な共同研究者として順序付けする性能を評価する．評価では，著者が 1 人違う論文のペアを評価データから収集し，実際の被引用数の大小関係と各論文に対する被引用数の予測値の大小関係が一致したペアの割合を順序正答率として算出する．10 組のデータについての実験結果の平均を表 3 に示す．順序付けのタスクにおいても提案モデルが最も高い性能を示すことが確認された．この結果から，提案モデルはより良い成果を生む共同研究者の推薦に有用であると考えられる．

拡張 1 については，いずれのモデルでも適用することで性能が向上することが確認されたことから，論文の被引用数予測において論文の周辺情報を用いることは有用であると考えられる．拡張 2 についても，ベースモデルに適用することで両タスクにおいて性能の向上が確認された．また，LSTM を用いたモデルよりも拡張 2 を適用した提案モデルが高い性能を示した．この結果から，潜在ベクトルの確率的な生成は，本来は利用できない著者順の情報の利用と同等以上に，論文の被引用数予測において有用だと示唆される．

5. おわりに

本論文では，より良い成果を生むことが期待される効果的な共同研究者を推薦する手法を提案した．提案法では，より良い成果を達成した論文はより高い被引用数を得るという仮定のもと，共同研究者推薦を，著者集合から共著論文の被引用数を予測する回帰問題として定式化した．実験の結果，提案モデルがベースラインと同等以上の性能を示すことを確認した．また，論文の周辺情報を用いること，研究者の潜在ベクトルの確率的な生成が予測性能の向上に寄与することが示唆された．今後の課題としては，提案手法が高い性能を示した理由の解析や，他のデータセットの使用等による実験の拡充による仮定の追証があると考えている．また，従来の共同研究者推薦手法との比較や個々の研究者についての属性情報を活用したモデルへの拡張を展望している．

文　献

[1] S. Lee and B. Bozeman, “The impact of research collaboration on scientific productivity,” Social studies of science, vol.35, no.5, pp.673–702, 2005.

[2] R.B. Duque, M. Ynalvez, R. Sooryamoorthy, P. Mbatia, D.-B.S. Dzorgbo, and W. Shrum, “Collaboration paradox: Scientific productivity, the internet, and problems of research in developing areas,” Social studies of science, vol.35, no.5, pp.755–785, 2005.

[3] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, “Deep unordered composition rivals syntactic methods for text classification,” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol.1, pp.1681–1691, 2015.

[4] H. Tong and C. Faloutsos, “Center-piece subgraphs: problem definition and fast solutions,” Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data miningACM, pp.404–413 2006.

[5] J. Li, F. Xia, W. Wang, Z. Chen, N.Y. Asabere, and H. Jiang, “Acrec: a co-authorship based random walk model for academic collaboration recommendation,” Proceedings of the 23rd International Conference on World Wide WebACM, pp.1209–1214 2014.

[6] Y. Guo and X. Chen, “Cross-domain scientific collaborations prediction with citation information,” 2014 IEEE 38th International Computer Software and Applications Conference Workshops (COMPSACW)IEEE, pp.229–233 2014.

[7] T. Huynh, A. Takasu, T. Masada, and K. Hoang, “Collaborator recommendation for isolated researchers,” Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference onIEEE, pp.639–644 2014.

[8] H.-H. Chen, L. Gou, X. Zhang, and C.L. Giles, “Collabseer: a search engine for collaboration discovery,” Proceedings of the 11th annual international ACM/IEEE joint conference on Digital librariesACM, pp.231–240 2011.

[9] J. Tang, S. Wu, J. Sun, and H. Su, “Cross-domain collaboration recommendation,” Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data miningACM, pp.1285–1293 2012.

[10] X. Kong, H. Jiang, W. Wang, T.M. Bekele, Z. Xu, and M. Wang, “Exploring dynamic research interest and academic influence for scientific collaborator recommendation,” Scientometrics, vol.113, no.1, pp.369–385, 2017.

[11] M. Araki, M. Katsurai, I. Ohmukai, and H. Takeda, “Interdisciplinary collaborator recommendation based on research content similarity,” IEICE TRANSACTIONS on Information and Systems, vol.100, no.4, pp.785–792, 2017.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, “Arnetminer: extraction and mining of academic social networks,” Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data miningACM, pp.990–998 2008.