1B2-1

大規模学術論文データの共著ネットワーク分析に基づく萌芽領域の中心研究者予測に関する研究

Predicting Citations using Citation Network and Text Analysis from the Academic Paper Database

森 純一郎*1 Junichiro Mori　原 忠義*1 Tadayoshi Hara　榊 剛史*1 Takashi Sakaki　梶川 裕矢*2 Yuya Kajikawa　坂田 一郎*1 Ichiro Sakata

*1東京大学大学院工学系研究科 The University of Tokyo　*2東京工業大学大学院イノベーションマネジメント研究科 Tokyo Institute of Technology

In this research, we aim to develop a method for predicting citations to detect emerging technology using academic papers. We assume the emerging research field grows off a highly and rapidly cited paper, which we call the "emerging paper". Our goal is to find such emerging paper in advance using a machine learning approach. We first extract a citation network of academic papers from a bibliographic database and then apply a clustering to the citation network to identify the research field as a cluster. Based on the citation network and its clusters, we design several features to predict citations. We conduct an experiment using the large amount of bibliographic data. Our preliminary result shows that our approach can predict the emerging paper in terms of increase of citations with F-value of 0.7-0.8.

1. はじめに

今日，科学技術イノベーションに関する情報は爆発的に増加している．例えば，太陽電池については，主要な国際論文誌に掲載される論文数は，90 年代には年間数百本に過ぎなかったが，今日では年間 4000 本に達している．こうした大量の情報は電子化され，世界のどこでも入手可能であり，イノベーションに関する経営戦略の立案や推進プロジェクトの評価等（技術経営）やイノベーションに関する政策形成に利用可能なものであると認識されている．しかしながら，実際には，情報量が多すぎ，知識の全体像や潮流，未来像が見えにくくなっている，自社又はイノベーション戦略を担う機関にとって有用と考えられる知識だけを抽出することが難しい，大量の情報に埋もれているため提携すべき相手又は潜在的な競合相手等も見出すことが難しくなっている，との意見が多く聞かれる．また，従来，技術の潮流の把握や予測等に用いられてきた専門家ワークショップ（代表的には，T-Plan 法）のような人的な活動を中心とした手法については，技術の変化の加速や専門家の知識の細分化により，限界に直面しているとの認識が強まってきている．こうした問題により，現状では，大量の有用な知識を科学技術イノベーションの効果的・効率的推進のために活かしきれていない状況にある．

特に，経営戦略の立案，技術経営，イノベーション政策の点から重要な点の一つは，現時点では未成熟で産業応用に制約が大きいが，関心を集め急速に立ち上がりつつある研究領域，萌芽領域，を早期に特定することである．萌芽領域は，技術シーズ発展のS字カーブ論でいう初期ステージにある技術群に当たり，こうした領域の中に，将来，経済・社会的に高い価値を生み出す技術群が含まれている．これまでは，萌芽領域の特定は学術俯瞰による成果と専門家の知見の融合により達成されてきた．しかしながら，専門家の知識の細分化が進み，全体像や補完的な技術や競合技術が見えにくくなっており，また情報量の増加から変化の激しい最先端を限られた数の専門家で常に追いかけるのは難しくなっていること現在，専門家の知見に頼るのみでは十分とは言い難い．

連絡先: 森純一郎，東京大学大学院工学系研究科，東京都文京区本郷 7-3-1，03-5841-1161，jmori@ipr-ctr.t.u-tokyo.ac.jp

従来研究においては，萌芽領域は中心となる萌芽的な論文から成長していると捉え，その中心的な萌芽論文を予測することによる萌芽領域の早期特定が行われている [Mori 14]．森らは，対象とする学術研究分野の大規模な論文群の引用ネットワークから抽出した様々な特徴量を用いて論文の引用数の増加を予測することで萌芽領域の中心論文を予測している．

萌芽領域の早期特定においては，技術シーズを含む論文に加えて，どの研究者あるいは研究グループが当該領域を牽引しているのかを特定することも重要となる．特に，萌芽領域の研究者を適切に把握することは，新規プロジェクトの立ち上げやチーム編成に貢献する．本研究では，萌芽領域における早期特定を目的とし，大規模な論文データの共著ネットワークを用いた萌芽領域の中心研究者の予測手法を提案する．これにより，科学技術イノベーションの効果的・効率的推進，すなわち，経営戦略の立案，プロジェクト評価等企業における技術経営の高度化や科学技術イノベーション政策の高度化を目指す．

2. 同姓同名と共著ネットワーク

共著ネットワークの基本となるものは論文の著者とそれらの共著関係であるが，大規模な論文データから共著関係を取得する際には著者の同姓同名が問題となる．図 1 は，コンピュータ科学分野を中心とした書誌情報のデータベースである DBLP *1の全論文データの共著者数分布を示している．多くは数名の共著者を持つ著者が大半であるが，数百を超える共著者数を持つ著者も存在する．これらの著者を見てみると，多くの同姓同名がある氏名であり，単純に書誌情報データベースから共著関係を抽出すると曖昧性を含む共著ネットワークとなってしまうことがわかる．Microsoft Academic Search *2や Google Scholar *3のように，著者が自己申告することで同姓同名を解決しようとするシステムも存在するが，やはりなお多くの同姓同名を含んでいる．著者の同定の問題に対して，著者にユニークな ID を付与する取り組みも進められている．トムソンロイターの学術文献データベースである Web of Science では一部に ResearcherID と呼ばれる，著者を一意に識別するための ID

∗1 http://dblp.uni-trier.de
∗2 http://academic.research.microsoft.com
∗3 http://scholar.google.com

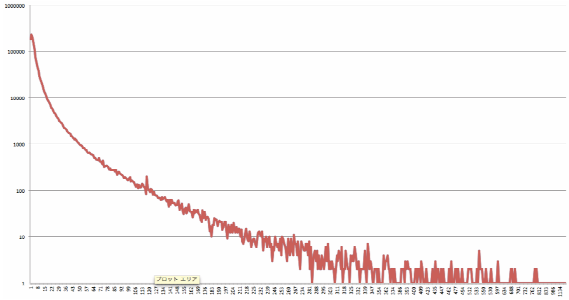

図 1: DBLP における共著者数の分布．X:共著者数，Y:延べ数

が一部付与されている．また，エルゼビアが提供するサービスの Scopus では著者の同姓同名を独自のアルゴリズムで処理しており，すべての著者にユニークな ID が付与されている．しかしながら，これらのサービスは有料であり大規模なデータ収集が困難という問題がある．

3. 共著ネットワークからの中心研究者の予測

3.1 研究者と共著データ

本研究では，共著ネットワークを用いて萌芽領域の中心研究者を予測するため，ACM Digital Library [*4]のデータを使用する．ACM Digital Library は，ACM に関連した論文のデータベースであり，コンピュータ科学分野の論文データを網羅している．著者の同姓同名については，著者の所属，発表論文の論文誌やキーワードなどを手がかりとして独自のアルゴリズムにより処理を行っており，著者ごとの同姓同名が分離された結果の共著者データが提供されている．また，同データベースには，著者ごとの発表論文数，被引用数，ダウンロード数などの計量書誌データが提供されている．本研究においては，これらを研究者のパフォーマンスと見なし，任意の領域において特に被引用数が高い研究者を領域の中心研究者とする．

3.2 共著ネットワークの構築と特徴量抽出

本研究では，共著ネットワークの構築に際して，まず任意の領域においてシードとなる研究者を決定し，その研究者から共著関係を逐次的にたどることにより共著ネットワークのノードとなる著者およびエッジとなる共著関係を収集する．その際，シードとなる研究者からたどる共著関係の深さを指定することで分析に応じた適切な大きさの共著ネットワークを取得する．

次に，共著ネットワークに対してクラスタリング [Good 10] や中心性の計算等を行い，中心研究者の予測に用いる以下の特徴量の抽出を行う．

- ノード特徴量
  次数，近接性，媒介性，クラスタリング係数，隣接ノードの平均次数，固有値，Pagerank，トライアド数
- クラスタ特徴量
  クラスターランク，クラスタサイズ，モジュラリティ

ノード特徴量は，共著ネットワークの個々のノードから抽出される特徴量で，ノードのネットワーク中心性などを含む．クラスタ特徴量は，共著ネットワークをクラスタリングした結果得られたクラスタから抽出される特徴量で，クラスタのサイズやクラスタのモジュラリティを含む．

∗4 http://dl.acm.org

提案手法では，共著ネットワークから抽出したネットワークの構造に基づく特徴量をもとにして，研究領域における研究者の被引用数を予測する学習モデルを構築する．

3.3 実験

共著ネットワークの構築あたって，まず「Machine Learning」の分野を対象に，「Vladimir Vapnik」をシードとなる研究者とした．同氏は Google Scholar Citations において「Machile Learning」を主要なタグとする研究者の中でも最も引用が多い研究者である．シードから共著関係を逐次的にたどりシードからの深さが 3 となる共著関係までの収集の対象とした．その結果，22,482 の研究者と 116,881 の共著関係からなる共著ネットワークを構築した．

共著ネットワークから特徴量を抽出した上で，被引用数が全体の上位 5%の研究者を中心研究者予測の正例データとし，正例データと同数の論文を負例データとしてランダムに抽出を 5 回行い，異なる 5 セットの学習データを作成した．学習は，ロジスティック回帰によって，被引用数が上位になるか否かの二値分類を行い，評価は交差検定によって precision, recall, F-value によって行った．

3.4 考察

交差検定の結果，精度はデータセットの平均 F-value が 0.8659 であり，一定程度の精度で共著ネットワークの構造から著者の被引用数の予測できることが示された．また，予測に有効な特徴量を見ると，共著ネットワークにおける著者の近接性や次数が有効であることが学習モデルから示された．これは共著ネットワークの中心にいる研究者が被引用数を多く獲得することを示唆していると考えている．

本研究で用いた ACM Digital Library のデータは同姓同名の判別精度は高いが，例えば異なる分野で活躍する同一研究者が分離されてしまう可能性がある．そのため，現在は著者にユニーク ID がふられた Scopus のデータを用いて提案手法の再実験を行っている．今後は，今回の実験を元に，H-Index などの複数の研究者パフォーマンス指標を用いて，萌芽領域の中心研究者の予測に有効な特徴量の精査とモデルの構築を行う．また，本実験では最新年の共著ネットワークおよび被引用数を用いているため，厳密な将来予測にはなっていない．今後は，共著ネットワークの時系列変化を考慮し，中心研究者について将来予測を行うモデルの構築を行う．

4. おわりに

本研究では，萌芽領域における早期特定を目的とし，大規模な論文データの共著ネットワークを用いた萌芽領域の中心研究者の予測手法を提案し，実データに基づき提案手法の評価を行った．大規模データに基づき萌芽領域を自動的に早期特定する提案手法は，企業の経営幹部や政府の政策担当者に対し，投資先候補と考えている技術分野の潮流を早期に把握するための技術経営基盤を提供する．こうした技術経営基盤は，企業の技術経営の高度化，政府間競争の中での政策形成の優位性の確保に対して，重要な貢献をしうるものと考えられる．

参考文献

[Mori 14] 森純一郎，榊剛史，梶川裕矢，坂田一郎 (2014). 萌芽研究領域特定のための大規模論文情報を用いた引用予測，人工知能学会全国大会.

[Good 10] Good, B.H., de Montjoye, Y.-A., and Clauset, A. (2010). The performance of modularity maximization in practical contexts. Phys. Rev. E 81, 046106.