1A3-4

萌芽領域特定のための大規模論文情報を用いた引用予測に関する研究

Predicting Citations to Detect Emerging Technologies using Academic Papers

森 純一郎[*1] 榊 剛史[*1] 梶川 裕矢[*2] 坂田 一郎[*1]
Junichiro Mori Takashi Sakaki Yuya Kajikawa 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いいがたい。

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

本研究では、萌芽領域の早期特定を目的とし、大規模な論文情報を用いた論文の予測手法を提案する。本研究では、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である萌芽領域を、領域の中心となる萌芽的な論文から成長している研究領域と捉え、その中心的な萌芽論文を予測することにより、萌芽領域の早期特定を行う。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. 関連研究

Boner らは、科学技術の科学のための分析ツールとして Sci^2 [Sci2 Team 09] の研究開発を行っている。同ツールは論文を含む科学技術政策に関する大規模な情報を分析可能であるが、主に可視化を目的としている。Chen らも、主に大規模な学術情報を分析しパターンやトレンドを可視化するツールである CiteSpace [Chen 06] の研究開発を行っている。学術情報の分析に特化したツールとして、Porter らは大規模な学術情報の統計処理を行うツールである Vantage Point の開発を行っている [Porter 04]。また、Boyack らは Elsevier と共同で研究機関の研究力評価に注力した SciVal の研究開発を行っている [Boyack 02]。これらのツールは現状の分析に特化した者であり、本研究が対象とする萌芽領域の早期特定のような将来の予測を汎用的に扱っていない。

3. 手法

3.1 引用ネットワークと研究領域抽出

本研究では、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である萌芽領域を、領域の中心となる萌芽的な論文から成長している研究領域と捉え、その中心的な萌芽論文を予測することにより、萌芽領域の早期特定を行う。萌芽論文の予測は以下のように行う。

まず、分析対象とする学術研究分野の論文群を取得し、それらの論文群内の論文間の引用関係に基づき引用ネットワー

クを構築する。次に、引用ネットワークに対してクラスタリング [Good 10] を行い、クラスタを抽出する。抽出されたクラスタは、当該研究分野の研究領域に対応している。複数のクラスタ（研究領域）の中で、どのクラスタが今後萌芽領域として急速に成長するかを特定することを、本研究では萌芽領域の特定タスクとして設定する。ここで、萌芽領域は、その領域の中心となる萌芽的な論文から成長している研究領域と捉え、本研究ではその中心的な萌芽論文を予測することにより、萌芽領域の早期特定を行う。萌芽論文については、当該論文が引用をどの程度得るか、その引用数の増加率を元に判別されるものとする。

3.2 萌芽論文予測のための特徴量設計

提案手法では、引用ネットワークから得られる以下の特徴量をもとにして、論文の引用数の増加率を予測する学習モデルを構築する。

- ローカル特徴量
  引用関係、ネットワーク中心性、テキスト等
- クラスタ特徴量
  クラスタサイズ、モジュラリティ等
- ネットワーク特徴量
  ネットワークサイズ、パス長、クラスタリング係数等

ローカル特徴量は、個々の論文から抽出される特徴量で、論文に対する引用関係の有無、論文のネットワーク中心性（次数、近接性、媒介性、固有値等）、論文のテキスト（タイトルや著者キーワードから生成される特徴語群）を含む。クラスタ特徴量は、論文が含まれるクラスタから抽出される特徴量で、クラスタに含まれる論文の数（クラスタサイズ）、クラスタ内の引用関係の数、クラスタのモジュラリティを含む。最後に、ネットワーク特徴量は引用ネットワーク全体の特徴量で、ネットワークの大きさ、パス長やクラスタリング係数を含む。

4. 実験

4.1 データ

本研究では、トムソンロイターが提供する学術文献データベースである Web of Science から論文データを取得する。Web of Science は引用文献検索機能を備えた学術文献データベースであり、大規模で質の高い学術情報を提供する代表的なデータベースの一つである。世界中の主要論文誌等 12,000 をカバーしており、また、書誌情報も高い品質で整備されている。

論文データの取得にあたっては、Web of Science が提供する Web API を用いて、同データベースにおいて”Artificial Intelligence” と分類がなされている学術論文とそれらの引用関係を取得した。取得した論文の総数は 179,290 であった。

取得した論文群から引用ネットワークを構築しクラスタリングを行って研究領域に対応するクラスタを抽出した上で、引用数の増加率が全体で上位 2%の論文を萌芽論文予測の正例データとして各クラスタから抽出した。一方、正例データと同数の論文を負例データとして抽出し、学習データを作成した。

学習データの論文から特徴量を抽出し、ロジスティック回帰によって引用数が急速に増加するか否かの二値分類器を学習した。

4.2 評価

学習は、1990 年から 2000 年（期間 1:1130 学習インスタンス）、2000 年から 2010 年（期間 2:3920 学習インスタンス）の

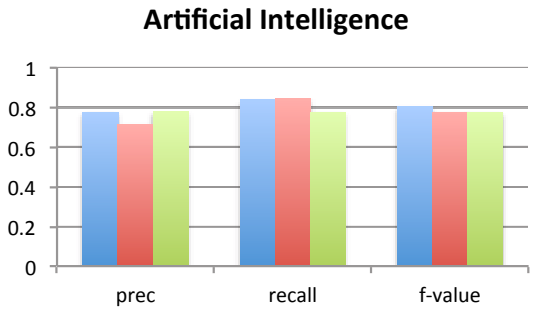

図 1: 萌芽論文予測の精度

2 つの期間と、それらを合わせた期間（期間 3:5050 学習インスタンス）の計 3 期間に対して行った。それぞれの期間について、学習データをもとに、基準年から過去 n 年のデータをもとにして m 年後において、対象論文の引用数の増加率が上位 2%になるか否かの学習を行った。評価は交差検定によって precision, recall, F-value によって行い、各期間内で n と m を変化させ、その精度の平均を評価した。

図 1 は、各期間における精度を示している。期間 1 では 0.802, 期間 2 では 0.772、期間 3 では 0.772 の F-value となっており、期間によらず一定程度の精度で論文の引用数の増加の有無を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、特に論文が属するクラスタの特徴量が有効であることが学習モデルから示された。今後は、今回の実験を元に、テキストの複数の特徴量も含めた萌芽論文の予測に有効な特徴量の精査とモデルの構築を行う。

5. おわりに

本研究では、萌芽領域の早期特定を目的とし、大規模な論文情報を用いた論文の予測手法を提案し、実際にデータの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

[Chen 06] Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

[Sci2 Team 09] Sci2 Team (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

[Porter 04] Porter, A.L., and Cunningham, S.W. (2004). Tech mining: exploiting new technologies for competitive advantage. Hoboken, NJ: JohnWiley and Sons, Inc.

[Boyack 02] Boyack, K.W., Wylie, B.N., and Davidson, G.S. (2002). Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology, 53(9), 764-774.

[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.