2J3-4

大規模論文データからの異種ネットワーク組み合わせによる萌芽論文の推定

Predicting Citations using Heterogeneous Citation Networks

森 純一郎[*1] 原 忠義[*1] 梶川 裕矢[*2] 坂田 一郎[*1]
Junichiro Mori Tadayoshi Hara 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

大規模な学術情報の増加に伴い，従来，科学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いる．

萌芽領域の早期予測は，科学技術のフォーサイトやホライゾンスキャニングの点からも特に重要であり，データマイニングや機械学習を用いた自動化アプローチを含めて近年さまざまな研究が行われている．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．これらの手法は対象分野ごとに予測モデルを学習するものであり，当該分野がすでに確立されており十分な学習データがあることを前提としている．

一方，近年の学際的な研究や異分野融合に基づく新興学術分野においては，対象分野を具体的に定義するのが困難であり，従来手法による分野に応じた予測モデルの学習が難しい．本研究では，任意の学術分野における萌芽論文の早期特定を目的とし，引用ネットワーク特徴量抽象化による萌芽論文予測の分野横断適用手法の提案と評価を行う．

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

2. 引用ネットワーク特徴量抽象化による萌芽論文予測の分野横断適用

萌芽論文の予測に関して，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，出版直後の論文の被引用数増加を予測する手法が提案されている [Mori 14]．それらの引用ネットワーク特徴量は，次数，近接性，媒介性，クラスタリング係数，固有値，Pagerank，Hub，Authority などの中心性指標やネットワークのクラスタリングに基づくクラスタ情報，クラスタランク，クラスタサイズ，モジュラリティなど，が用いられ，これらの組み合わせが論文の被引用数増加を予測するのに有効であることが示されている．

従来手法においては，予測対象の分野ごとに，専門家の知見に基づく検索クエリーを設定し，それらの検索クエリーを元に対象分野の書誌データを取得し，予測モデルの学習を行っていた．これらのアプローチは，すでに確立された学術分野においては有効であるが，学際的な融合分野や新興分野においては書誌データを取得するための検索クエリーの設定が困難であり，すでに学習済みの他分野の予測モデルを適用することが必要になる．予測モデルを分野横断で適用可能であることの仮説は，引用数の増加を引用ネットワークの成長と捉えると，引用ネットワークの特徴量により抽象化されている従来の予測モデルは，分野に固有の特徴に依らず任意の分野において，その引用ネットワーク成長予測，つまり個々の論文の引用数の増加予測に寄与可能である，というものである．

以下では，引用ネットワーク特徴量抽象化による萌芽論文予測の分野横断適用の可能性について，実際の学術論文の書誌データを用いた実験を通して，予備的な検証を行う．

3. 実験

7 つの異なる研究分野，複雑ネットワーク，ソーラーフォト，ガリウム・ナイトライド，ナノカーボン，ソーシャルシステム，リアルタイムロボット，サイバーフィジカル，の書誌データを

予測モデル学習分野・年		予測対象分野・年																																		
		複雑ネットワーク					ソーラーフォト					ガリウム・ナイトライド					ナノカーボン					ソーシャルシステム					リアルタイムロボット					サイバーフィジカル				
		'01	'03	'05	'07	'09	'01	'03	'05	'07	'09	'01	'03	'05	'07	'09	'01	'03	'05	'07	'09	'01	'03	'05	'07	'09	'01	'03	'05	'07	'09	'01	'03	'05	'07	'09
複雑ネットワーク	'01	–	84.4	86.6	83.6	79.8	72.0	75.4	75.7	75.2	73.8	70.6	70.8	70.9	71.7	72.2	80.4	79.8	76.0	74.7	73.0	82.5	85.4	82.7	83.0	82.5	79.6	79.1	76.0	78.0	75.2	73.4	75.0	78.3	81.4	80.2
	'03	77.1	–	86.8	84.1	80.1	76.4	77.0	76.1	75.0	73.5	70.8	71.0	71.1	71.4	72.1	80.4	79.1	75.3	74.6	73.9	82.5	85.4	82.5	83.0	82.4	79.4	79.0	76.2	78.8	76.4	73.3	74.5	77.8	79.6	75.7
	'05	82.1	85.0	–	86.2	83.9	74.1	73.6	73.9	75.0	75.0	70.6	70.6	70.8	71.6	73.1	83.6	81.7	78.0	76.1	74.7	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.8	76.6	73.4	75.1	78.4	81.4	80.6
	'07	79.7	79.1	83.2	–	85.7	73.7	73.3	72.7	72.2	72.3	70.6	70.6	70.7	70.9	71.6	79.8	81.4	80.1	79.4	76.8	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.8	76.6	73.4	75.1	78.4	81.4	80.9
	'09	79.6	80.2	81.1	88.4	–	74.5	73.5	72.5	71.7	71.2	70.5	70.6	70.7	70.8	71.0	76.2	79.0	80.9	82.2	80.2	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.8	76.6	73.4	75.1	78.4	81.4	80.6
ソーラーフォト	'01	25.3	35.3	44.8	49.7	55.6	–	91.0	88.2	78.9	71.9	72.2	73.4	71.4	73.4	73.5	76.1	70.1	65.3	62.3	60.7	82.5	85.4	82.5	83.1	84.2	79.6	79.1	76.2	78.7	79.5	70.7	59.8	42.2	32.2	17.2
	'03	20.0	47.5	64.1	68.4	71.1	80.6	–	89.5	82.5	76.8	71.7	71.7	70.3	73.3	74.0	81.4	77.1	73.5	72.2	73.1	75.9	80.2	73.8	75.6	75.7	65.0	60.3	53.7	55.1	54.1	36.0	25.2	14.0	10.1	6.5
	'05	7.2	18.2	27.6	51.0	67.5	72.0	86.4	–	88.8	81.7	73.1	72.4	72.5	74.8	74.4	86.7	77.4	73.0	70.8	72.1	82.5	85.4	82.5	83.1	84.8	79.6	79.1	76.3	79.6	77.8	72.0	59.3	34.9	24.1	8.3
	'07	6.8	28.6	33.5	52.5	70.4	60.0	78.8	92.6	–	86.8	73.3	71.4	68.1	72.9	73.4	86.8	84.4	79.3	75.3	73.9	82.5	85.3	82.6	84.1	84.1	79.6	78.9	73.2	70.4	67.7	62.9	48.8	26.9	16.3	5.7
	'09	8.0	26.9	37.1	54.1	68.2	57.3	71.1	84.4	85.7	–	76.9	73.2	68.4	72.3	70.0	81.8	87.4	84.1	80.5	77.7	82.5	84.5	82.9	86.0	86.5	79.6	78.5	70.9	68.4	62.9	57.5	40.6	21.3	15.4	5.5
ガリウム・ナイトライド	'01	38.3	49.7	55.3	56.4	60.4	44.8	43.3	58.4	63.4	66.0	–	73.4	66.1	67.5	62.4	66.5	64.9	62.4	62.1	60.2	82.4	81.5	82.8	78.9	83.0	79.8	78.9	75.4	73.1	74.9	71.7	67.2	51.9	43.8	29.1
	'03	40.5	53.0	60.5	60.0	58.6	62.7	66.6	75.1	77.0	79.7	78.5	–	61.8	52.2	51.9	62.3	61.3	61.5	57.5	61.9	82.5	84.3	82.6	83.1	82.9	79.6	79.1	76.0	76.6	74.0	72.5	72.1	66.0	61.1	45.1
	'05	49.3	60.1	72.9	71.9	73.3	58.4	60.4	67.3	77.7	82.6	76.8	77.9	–	78.1	73.7	74.3	74.8	75.1	74.3	76.1	82.5	83.5	82.8	82.8	83.1	79.6	78.9	75.8	75.1	73.6	70.3	69.1	63.4	58.5	45.2
	'07	44.9	51.0	67.3	73.4	79.3	60.1	65.5	71.2	78.7	84.4	76.2	77.2	76.6	–	77.1	86.4	85.2	81.5	78.0	76.0	82.6	83.4	82.9	82.9	82.6	79.7	79.1	75.9	76.2	73.0	73.2	74.4	73.3	70.5	54.1
	'09	74.9	71.2	57.5	41.9	36.9	70.3	75.1	80.2	82.5	84.5	72.9	74.2	75.8	79.2	–	82.1	64.5	51.4	41.8	33.3	82.5	85.1	82.6	83.1	82.6	79.6	79.1	76.2	78.4	75.3	73.2	74.4	76.2	80.0	76.9
ナノカーボン	'01	76.8	65.4	47.2	36.2	27.3	60.2	64.5	68.7	75.5	83.7	71.3	72.0	73.2	75.8	77.6	–	90.0	82.3	76.2	78.5	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.8	75.9	73.4	75.1	78.3	81.4	79.8
	'03	82.9	79.1	68.8	53.4	28.3	71.2	72.7	73.2	74.8	78.4	70.6	70.7	70.8	71.6	73.4	88.2	–	82.8	63.8	45.2	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.2	78.8	76.4	73.5	75.3	78.4	81.4	80.9
	'05	80.7	84.5	87.9	84.2	68.6	74.5	73.5	73.0	73.8	74.9	70.6	70.6	70.8	71.6	72.9	84.6	88.4	–	82.2	55.5	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.8	76.6	73.4	75.1	78.4	81.4	80.6
	'07	82.1	78.4	73.3	69.6	63.3	67.4	71.9	74.3	75.6	77.1	70.8	71.1	71.8	72.7	73.8	87.1	88.6	87.4	–	87.5	82.5	85.4	82.5	83.0	82.4	79.6	79.1	76.3	78.7	76.5	73.4	75.2	78.4	81.4	80.9
	'09	56.9	54.0	43.2	35.7	34.5	56.1	60.2	62.4	67.3	73.0	74.3	73.7	74.6	77.2	76.4	78.5	70.7	67.1	82.2	–	82.4	82.1	82.8	80.3	82.7	79.6	78.7	75.6	75.8	72.2	73.2	74.7	78.0	79.6	74.9
ソーシャルシステム	'01	79.6	80.9	80.2	78.8	76.1	75.1	73.8	72.8	71.9	71.3	70.4	70.6	70.8	70.9	71.1	76.0	75.3	72.9	71.8	71.0	–	84.0	82.5	83.5	82.9	76.5	78.9	75.6	78.7	77.0	69.3	71.0	74.5	79.2	79.3
	'03	77.9	79.4	79.1	77.4	75.0	75.4	73.8	72.8	71.7	70.9	70.0	70.2	70.2	70.7	70.7	75.7	74.8	72.7	71.5	70.6	76.0	–	80.5	81.5	81.4	74.6	76.5	71.1	75.6	76.2	50.3	58.6	64.6	72.6	74.1
	'05	80.4	82.1	81.4	79.6	76.7	77.6	75.7	74.1	72.7	71.8	70.6	70.9	70.9	71.4	71.4	77.3	76.2	73.6	72.3	71.4	76.7	79.9	–	84.7	84.6	62.9	76.3	70.4	75.8	78.0	43.1	52.2	59.6	72.1	75.6
	'07	79.2	81.8	81.3	79.5	76.7	79.6	77.1	74.4	72.6	71.3	70.9	71.3	71.0	71.6	71.7	77.3	76.1	73.6	72.2	71.4	67.9	68.7	79.3	–	84.5	59.7	63.8	64.9	69.4	76.4	33.0	42.1	47.2	63.3	69.8
	'09	76.3	82.2	83.5	80.4	77.9	81.8	79.5	77.4	74.6	72.9	71.3	71.2	70.6	71.9	71.9	79.3	78.3	74.6	73.1	72.2	79.5	76.9	80.3	82.7	–	73.5	72.3	64.2	65.0	72.0	34.2	30.9	30.8	43.8	54.9
リアルタイムロボット	'01	72.7	70.3	66.9	66.8	66.9	73.1	68.4	66.3	65.8	62.5	70.0	70.0	70.0	70.8	70.7	70.4	68.8	67.6	67.6	66.9	76.3	77.0	81.8	82.2	82.2	–	77.8	72.4	74.1	75.7	56.4	60.8	66.6	72.0	73.2
	'03	75.7	76.5	73.9	73.0	72.0	76.7	74.9	73.8	72.7	69.8	70.7	70.9	70.9	71.4	71.5	75.9	73.0	70.9	70.2	70.4	71.2	79.9	84.0	84.5	83.7	70.2	–	72.0	77.1	78.1	46.9	53.8	64.3	70.9	74.2
	'05	73.8	78.2	79.2	77.8	75.5	79.4	78.0	76.5	74.5	72.5	70.3	70.2	69.0	69.1	69.2	78.0	75.9	72.6	70.2	70.3	81.7	84.8	81.4	84.1	83.4	79.7	79.0	–	77.1	76.2	56.5	53.0	51.5	59.0	64.9
	'07	64.5	67.9	66.6	66.1	67.1	80.5	78.5	76.4	73.5	71.2	71.6	71.6	70.7	71.8	71.9	77.4	73.8	68.9	68.7	70.2	82.5	85.4	82.1	83.0	83.8	79.6	78.9	74.5	–	79.5	49.4	44.9	41.8	48.6	54.1
	'09	63.0	65.2	62.4	61.6	64.5	79.5	75.5	69.3	65.3	60.9	70.9	71.0	70.2	71.7	71.2	68.7	68.5	66.9	66.4	67.4	81.3	83.9	81.5	83.6	84.7	74.5	73.1	70.4	73.8	–	46.1	38.0	32.7	38.3	44.7
サイバーフィジカル	'01	79.8	81.0	80.3	79.0	76.2	74.6	73.6	72.6	71.8	71.2	70.6	70.6	70.7	70.9	71.0	75.8	75.2	72.6	71.8	71.0	80.7	78.3	81.2	82.5	82.5	79.9	79.0	75.9	76.1	75.3	–	75.0	77.9	81.1	80.2
	'03	79.9	80.0	79.6	79.1	76.3	72.2	70.4	69.3	69.6	69.5	68.2	69.7	70.4	70.8	71.0	76.1	75.6	72.9	72.1	71.1	73.3	69.1	70.5	74.8	77.8	79.6	79.0	74.7	70.2	68.6	73.3	–	77.5	80.0	79.2
	'05	79.3	79.3	76.0	73.9	70.8	74.2	72.5	68.9	66.4	64.7	69.1	67.6	67.2	67.1	66.0	71.0	68.5	64.7	63.8	62.1	79.3	74.7	77.6	81.1	82.3	80.0	79.1	75.3	73.2	72.5	73.2	75.0	–	79.7	78.8
	'07	77.2	77.9	77.9	74.1	70.0	66.0	70.5	73.0	72.4	71.2	70.8	71.0	69.5	70.4	70.3	77.4	74.8	70.1	66.4	63.6	80.1	82.4	79.5	80.0	81.0	79.7	79.0	75.7	75.9	73.1	73.4	75.0	78.2	–	76.9
	'09	80.5	78.4	76.5	73.8	71.2	64.0	70.5	72.7	71.0	69.4	71.6	70.6	69.0	68.6	67.8	74.8	71.9	68.3	66.5	65.3	81.1	81.3	80.9	79.9	81.9	79.7	79.0	75.7	74.6	72.1	73.4	75.3	78.3	81.2	–

図 1: 論文の被引用数増加の予測モデルの学習対象分野および年と予測対象分野および年の組み合わせによる予測精度の変化

Web of Sciece Web Services を用いて取得した．書誌データの取得にあたっては，専門家の知見を元に設定した検索クエリーを用い，各分野ごとに数千から数万論文の規模の書誌データから引用ネットワークを構築した．

学習の対象年 (2001 年，2003 年，2005 年，2007 年，2009 年）に出版された論文を対象に，各分野ごとに対象年の引用ネットワークからネットワーク特徴量 [Mori 14] を抽出した上で，　出版後 3 年の被引用数が分野の上位 5%になる論文を正例データとし，出版後 3 年の被引用数が下位の論文から正例データと同数を負例データとしてランダムに抽出を行い，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，各分野において学習の対象年ごとに予測モデルを構築した．

任意の分野の学習データから構築した予測モデルを用いて，他分野の任意の対象年に出版された論文の 3 年後の被引用数の増加可否の予測を行うことで，予測モデルの分野横断適用可能性の検証を行った．予測の評価には F 値を用いた．

4.　考察

図 1 は，引用ネットワークのネットワーク特徴量に基づく論文の被引用数の増加予測モデルの分野横断適用における精度を示している．表中の各行は，予測モデルの学習に用いた対象分野と年度に対応しており，各列はそれらの予測モデルを適用する対象分野と年度に対応する．表中の値は，任意の予測モデルを適用した予測精度を F 値で示したのものであり，そのセルの色が濃いほど予測精度が高いことを表している．例えば，複雑ネットワーク分野の 2003 年に出版された論文を用いて学習した 3 年後の被引用数の増加可否を予測モデルを同分野において F 値 0.8 以上の精度を達成している一方で，他に分野に適用した際も F 値 0.7 以上を達成していることがわかる．

特定の分野間においては，論文の被引用数の増加を予測する際に，引用ネットワークの特徴量を用いた予測モデルは，分野を横断した適用の可能性が，結果から示されている．これは，中心性やクラスターのようなネットワーク特徴量を用いることにより，引用ネットワークの構造から分野を抽象化できているためと考えられる．今回得られた基礎的な知見を元に，今後は，任意の学術分野をより抽象化し予測モデルの分野横断適用に寄与する特徴量の設計を行うとともに，分野横断適用の可否を評価するために，分野の内容，成熟度，成長度などを考慮した分野間の類似度や分野の適切な組み合わせによる予測モデル学習の検討を行う．

5.　おわりに

本研究では，大規模な学術論文データを対象にした萌芽論文の早期特定を目的とし，被引用数の増加予測モデルの分野横断適用手法の提案と評価を行った．今後は，提案手法を用いて，学際的な研究や異分野融合に基づく新興学術分野における萌芽論文の推定を行うことで，最終的には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への貢献を目指す．

謝辞

本研究は，JST，CREST の支援を受けたものである．

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.