

JST 論文データベースにおける共著者ネットワークから得られる論文引用要因の分析

○青木健 吉川厚 寺野隆雄 (東京工業大学)

Analyzing citation rules from researcher's co-authorship in the JST database

* K.Aoki, A.Yoshikawa and T. Terano (Tokyo Institute of Technology)

概要— 本報告では、論文引用の決定要因について共著者ネットワークを用いて分析する。我々のアプローチは、共著者ネットワーク上で研究者のコミュニティを構築し、そのコミュニティの中で起こる引用とそうでない引用を分類し、著者同士の関係が引用の決定に影響を与えるかを分析する。先行研究では、著者同士の関係は引用の決定に影響を与えないとされていたが、計算機利用技術分野で同様の結果が得られるのかを分析する。

キーワード: ネットワーク分析, コミュニティ, 共著関係

1 はじめに

研究者にとって研究資金は研究を行う上で欠かせないものであり、得られた研究資金を基に研究者は研究を行う。そのため、十分な研究資金を獲得できないことは研究者にとって大きな負担となることがある。研究者は研究資金によって新たな研究を行うことができ、更にその研究の成果が認められてまた資金が得られるという仕組みになっていることから、研究資金を獲得するためにはより多くの研究の成果を示す必要がある。そこで、研究の成果を示す指標について考えてみる。

現在、研究の成果を示す指標の一つとして論文の被引用数が用いられている。論文の被引用数とは、過去の論文がある論文から引用された時に数えられ、一般的にはその論文の影響度を測るものとされている。田村は、論文の被引用数と大学単位での研究資金獲得状況について分析を行い、それら二つに相関があるという結果を得た²⁾。つまり、研究者が研究の成果を示すことによって所属機関はより多くの研究資金を得ることができ、その資金は研究者に還元される。そのため、研究者が被引用数の多い論文を著すことは有益であり、今後の研究資金を獲得する上で必要なことであると考えられる。また、被引用数は雑誌の評価指標であるインパクトファクターや研究者の評価指標となるh-indexを求める数値として用いられている。インパクトファクターは、違う分野の論文を評価することは出来ないが同じ分野にある論文同士を評価する際の指標として用いる際には有効であり、h-indexは研究者の昇進の際に参考とされる数値である。このように論文の被引用数は多方面で大きな影響を与えている。

更に、論文の引用情報による評価は日本国内だけでなく世界でも扱われている。トムソン・ロイターは毎年、論文の引用情報から世界各国の大学や研究機関をランク付けしており、その結果はメディア等を通して報道されるため、このランキングは世界の学術的な動向を知る上で影響力の高いものとなっている³⁾。そのため、論文の被引用数を増やすためには何が重要であるかということに関する研究は海外でも多く行われている。そこで、我々は論文の被引用数を増やすためには引用の要因を知る必要があると考え、本稿では論文引用の決定要因について分析していく。

研究者が過去の論文を引用する理由はいくつか考えられる。例えば、自分の説を正当化する場合や他の研究者の意見に対して反論をする時、または過去の論文

を先行研究として紹介する時などの例が挙げられる。これらの引用の要因に対して、論文の特性を基に多変数の回帰分析から引用の要因を考えるという研究は数多く行われている⁴⁾。しかし、それらは著者同士の関係性が引用の要因と成り得るかについては検証していないものがほとんどである。研究者が過去の論文を引用する時の要因は論文の内容だけで決定され、研究者同士の関係は引用の決定要因に影響を与えないのだろうか。

そこで本研究では、研究者同士の関係が論文引用の決定に影響を与えるのかを計算機利用技術分野において明らかにすることを目的とする。また、Table. 1は計算機利用技術分野に含まれる分野を表したものである。

Table. 1: 計算機利用技術分野に含まれる分野

計算機利用技術分野		
計算機利用技術一般	数値計算	記号処理
図形・画像処理一般	医用画像処理	音声処理
自然語処理	パターン認識	人工知能
CAI	CAD,CAM	計算シミュレーション
医用情報処理	事務・経営情報処理	他の計算機利用技術
情報処理一般	分子・遺伝情報処理	その他の情報処理

2 関連研究

2.1 論文の引用要因に関する研究

研究論文の引用要因に関する研究はBaldi(1998)によって行われている⁵⁾。論文引用の決定要因は社会学者であるMertonが提唱したNormativeな考え方(以下、規範理論)とLatourが提唱したSocial constructivistな考え方(以下、社会構築主義理論)のどちらかに従うと考えられている。そこでBaldiは宇宙物理学分野における論文引用が、この二つの理論のどちらに従うかを分析した。今、論文引用の要因において規範理論に従うということは、研究のテーマや手法が類似している論文や被引用数が多い論文が引用され易いということの意味している。一方、社会構築主義理論に従うということは、著者同士が近い関係にあること、権威のある研究者や男性の研究者が引用され易いという特徴を持っている。この二つの理論に対して、Baldiは宇宙物理学分野の1965年から1980年までに発行された論文を100本無作為に抽出し、4950ケースの論文引用関係から得られるネットワークの分析と論文の内容や著者のランク等を数値化して回帰分析を行うことでその妥当性を

Table. 2: 変数の種類

変数
<従属変数>
論文の著者の人数
論文の大きさ・サイズ
論文を掲載した雑誌の権威
<規範理論に関わる変数>
引用された論文のテーマ
引用された論文の被引用数
研究の類似性
<社会構築主義理論に関わる変数>
著者が女性である割合
著者の権威
著者同士の結びつき

検証した。Table.2は分析において扱った変数の一部を抜粋したものである。分析の結果、宇宙物理学分野における引用の要因は規範理論を支持するものとなり、社会構築主義理論を否定する結果となった。唯一、社会構築主義理論で正しかったことは男性の研究者の方が引用され易いということのみであった。つまり、引用の要因は研究のテーマや手法によって決定されるため、著者同士の関係は引用の決定に影響を与えないという結果が得られた。しかし、異なる研究分野に対しても同様の結果が得られるかは分かっていないとされている。

そこで、我々の研究では対象分野を計算機利用技術分野に変更した時に研究者同士の関係が引用の決定要因となるか分析する。また、この研究において著者同士に関係があるということの定義は、同じ機関に所属している又は同じ大学院の博士号を取得したかによって定めるとされていた。しかし、この定義は著者同士の関係を表すには不適切であると考え、我々の分析では以下に述べる共著者ネットワークを用いて研究者同士の関係を表した。

2.2 研究者ネットワークに関する研究

研究者ネットワークに関する研究は数多く行われている⁶⁾⁻¹⁰⁾。原田(2015)は、2003～2013年の『計算機利用技術分野』における国内のJST学術文献データの著者IDと共著のデータを基に、共著者ネットワーク(Fig.1)を作成した¹¹⁾。共著者ネットワークとは、研究者をノードとして共著関係でリンクを張ることで作成された研究者ネットワークのことである。更に、2003～2013年を第1期(2003～2006)、第2期(2004～2007)、第3期(2005～2008)、第4期(2006～2009)、第5期(2007～2010)、第6期(2008～2011)、第7期(2009～2012)、第8期(2010～2013)の8つの期間に分け、各期間で共著者ネットワークにおける研究者の中心性指標による順位付けを行った。Fig.2の混合分布モデルは、第1期と第7期の順位変化によって研究者のクラスタリングをした様子を表している。縦軸、横軸共に研究者の

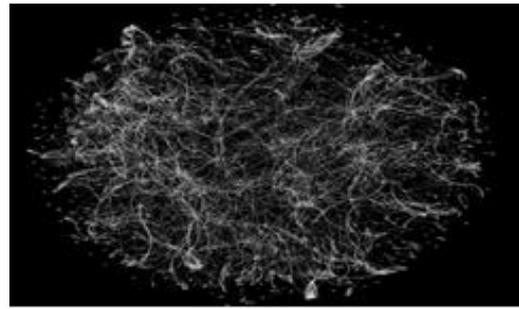


Fig. 1: 共著者ネットワーク

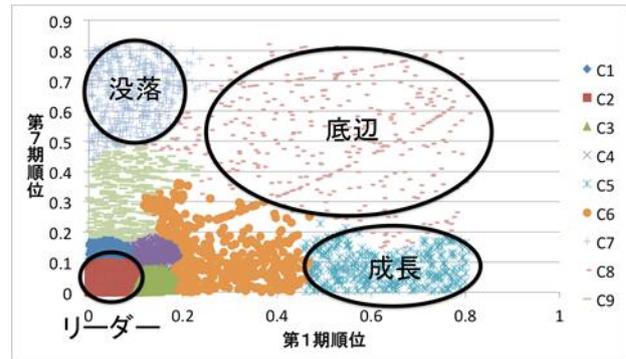


Fig. 2: 中心性順位の時系列変化による研究者のクラスタリング

Table. 3: 計算機利用技術分野の共著者ネットワーク推移

	2003-2006	2004-2007	2005-2008	2006-2009
ノード数	69629	77151	82724	87841
エッジ数	192203	224639	250099	271661
	2007-2010	2008-2011	2009-2012	2010-2013
ノード数	90257	90121	89536	82421
エッジ数	278845	273459	266320	221437

中心性順位を各期間のノード数で割った値となっているため、0に近づくほど高順位であることを表している。その中で、原田は4つのクラスタに注目し、それらのクラスタをそれぞれ『リーダー』『成長』『底辺』『没落』と名づけた。そして、それぞれのクラスタの所属、研究分野の変更情報を用いた分析の結果、『リーダー』群と『成長』群に分類される研究者は、その後の期間でも高順位であることが分かった。また、計算機利用技術分野の論文数、共著数の変化はTable.3に示した。後の分析では、この2003～2013年の『計算機利用技術分野』における国内のJST学術文献データを使用する。

3 分析手法

今回扱うデータは、2.2節で用いたものとする。更に、JST学術文献データを基に論文間の引用関係を整理する。まず、分析をするにあたって研究者のコミュニティを定義する必要がある。本研究では研究者のコミュニティを共著者ネットワークの二次までで繋がる研究者と定義する。この定義は、共著者ネットワーク上の二次までで結びつくことが出来る研究者は論文引用の

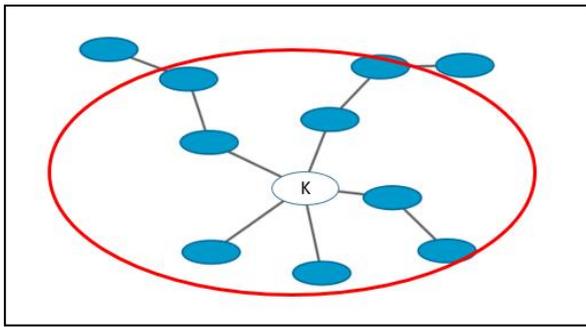


Fig. 3: 共著者ネットワーク上の研究者コミュニティ

決定に影響を与える近い研究者であると考えて定められたものである。この定義に関しては、Fig. 3を参照されたい。Fig. 3は共著者ネットワークにおける研究者Kの研究者コミュニティを表している。ここでは、二つのノードが研究者コミュニティに含まれない研究者であると考えられる。次に、分析の流れを説明する。まず、一人の研究者に対して研究者コミュニティを形成する。次に、その研究者に関する引用を調べる上で、コミュニティの中で起きた引用かそうでないか、そして自己引用かどうか調べる。これを引用関係に関わる全ての研究者に対して行うことで、論文引用を分類する。同じコミュニティに所属する研究者からの引用の割合が多いほど研究者同士の関係が引用の決定要因に影響を与えていると考えることが出来る。この分析によって、計算機利用技術分野における引用の決定要因が研究者同士の関係に影響されるかを測ることが出来ると考えている。

4 分析

4.1 予備実験と結果

本節では、4名の研究者に対して行った予備実験の結果について述べる。Table.4は、2003～2013年に発行された論文の中で1回以上引用された論文の数、発行された論文が引用された数の総計、研究者コミュニティに含まれない研究者から引用された数の総計、自分の論文を引用している回数をそれぞれ各研究者についてまとめている。また、Fig.4は各研究者の第1期、第4期、第7期における共著者ネットワーク上の中心性順位の時系列変化と各期間のノード数を示したグラフとなっている。本稿では、得られた実験結果に対してFig.4の中心性順位の時系列変化を用いて考察する。また、研究者Aと研究者Bは同じ『リーダー』群に所属する研究者で、研究者Cは『成長』群、研究者Dは『底

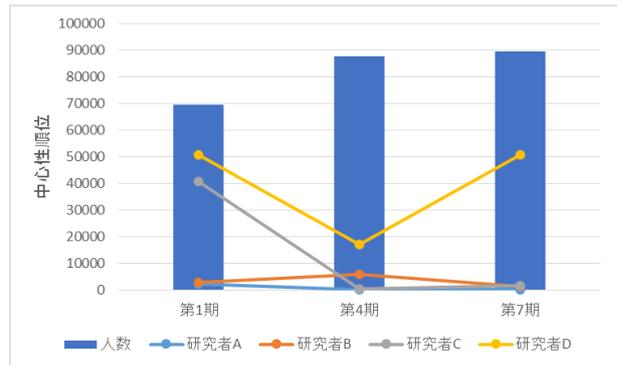


Fig.4: 中心性順位の時系列変化

辺』群に所属する研究者である。

予備実験の結果について考えてみる。全体の被引用数から自己引用を除けば、自分以外の研究者から引用された数が分かる。その中で、同じコミュニティに所属する研究者からの引用の割合を考える。例えば研究者Aであれば、自分以外の研究者から引用された数は51件あり、その内42件が同じコミュニティに所属していない研究者からの引用となっており、研究者同士の関係に影響されない引用が多いという宇宙物理学分野の引用決定と同様の結果が得られた。これは他のいずれの研究者でも同様の結果となった。特に、研究者Dは自己引用を除いた4件の引用全てが同じコミュニティに含まれない研究者からの引用であった。また、自己引用の数について研究者Cは被引用数に比べて少なく、研究者Dは非常に多くの自己引用があった。これらの得られた結果に対して、次節で考察を行う。

4.2 考察

予備実験の結果から、現状では計算機利用技術分野における引用の決定要因は研究者同士の関係に影響されないと予測を立てることが出来る。また、同じクラスターの研究者である研究者Aと研究者Bは引用される際の傾向や自己引用の傾向に類似性が見られた。また、研究者Cのように後に中心性順位が上がる『成長』群は積極的に他の研究者と関わっていくため自己引用の数が少なく、研究者Dのような『底辺』群は他の研究者との関わりが少ないため、自己引用の数が非常に多くなると予想される。そのため、自己引用の数は共著者ネットワークの中心性順位と相関を持つかもしれないと考えることが出来る。

Table.4:各研究者の引用情報

	論文数	被引用数	コミュニティに含まれない研究者からの引用	自己の引用
研究者A	18	64	42	13
研究者B	19	76	35	20
研究者C	7	28	18	3
研究者D	20	30	4	26

5 まとめと今後の展望

予備実験では、研究者4名に対して共著者ネットワークにおける研究者コミュニティを用いて、引用される時の要因が研究者の関係に影響されるか分析を行った。その結果、現状では研究者同士の関係は引用の決定要因に影響を与えないという結果になると予想されたが、研究者数を増やした時にどのような結果となるかは分析してみないと分からない。また、共著者ネットワークにおける研究者の中心性順位と比較したところ、今回の予備実験から二つの仮説が立てられた。一つ目は、同じクラスタの研究者は引用される際の傾向が似ているということ、二つ目は、自己引用の数と中心性順位には相関があるということだ。

今後は、これらの仮説にも注意しながら分析を行いたい。また、今回の分析では研究者コミュニティを共著者ネットワークの二次までで繋がる研究者と定義したが、この定義を変えたとき、例えば研究者コミュニティの範囲を二次から三次に変更した時に結果がどのように変わるか分析することを検討している。他にも、論文の内容や研究者同士の関係性以外の引用の決定要因が無いか検討する。

参考文献

- 1) https://www.jsps.go.jp/j-grantsinaid/27_kdata/data/1-1/1-1-1.pdf
- 2) 田村元紀：外部研究資金獲得に関する理工系国立大学のミクロ・ポテンシャル分析，第11回産学連携学会 37/44(2014)
- 3) <http://ip-science.thomsonreuters.jp/press/release/2015/esi2015/>
- 4) 小野寺, et al. : 論文の被引用数に影響する要因に関する統計学的研究, 第56回日本図書館情報学会研究大会発表要綱. 41/44(2008)
- 5) Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model, Stephane Baldi., *American Sociological Review* Vol. 63, No. 6, 829/846(1998)
- 6) 篠田孝祐：日本における人工知能研究の系譜，人工知能学会誌 26.6(2011)
- 7) 安田雪，松尾豊：人工知能学会における研究者ネットワークの分析，第19回人工知能学会全国大会(2005)
- 8) 安田雪，松尾豊，武田英明：人工知能学会におけるネットワークの構造と変化，人工知能学会全国大会論文集，1F2-1(2005)
- 9) 松尾豊, et al. 効果的な共同研究を支援するためのWebからの研究者ネットワーク抽出. 情報処理学会, 情報学シンポジウム (2005)
- 10) 西澤正己, et al. 科学研究費データベースによる採択と研究成果のパターン分析. 情報知識学会誌 1/6(2006)
- 11) 原田泰輔, 吉川厚, 寺野隆雄: ネットワーク分析を用いた計算機利用技術分野の変遷, 計測自動制御学会第8回社会システム部会研究会資料, 117/126(2015)