論文の内容の要旨

論文題目　論文の被引用数と引用持続性の間の相互関係、及びそれらと他の論文特性との関係に関する研究

氏　　名　小野寺　夏生

【目的】

論文の被引用数、あるいはそれに基づく指標は、その学術的影響度を測る尺度として最近注目されている。このことは統計学的に一定の妥当性を持つが、論文の被引用数をその評価に用いる場合には、その性質を十分に認識して注意深く取り扱うことが必要である。

その性質のひとつとして、論文の被引用数は、その質や内容に直接関係のない種々の要因（外在的要因と呼ぶ）の影響を受けるということがある。また、引用の持続性は論文により様々なので、被引用数を論文の評価に用いるとすれば、引用を計測する期間が評価に影響を及ぼす。本研究は、どのような「外在的」要因がどの程度論文の被引用数に影響するかについての知見を深めることにより、引用データの利用に資する情報を与えることを目的とし、以下の 3 つの研究目標を設定する。

(1) 引用影響要因の可操作化と正確な測定の検討

(2) 論文の被引用数に対する外在的要因の影響の検討

(3) 論文の引用持続性を示す指標の性質の検討

(1)に関して、論文の被引用数に影響を与える可能性のある外在的要因には様々なものがあり、かつ、それらの要因をどのような測度を用いて測定するか（要因の可操作化）にも選択の幅がある。そこで、この問題を扱った先行研究をレビューし、その結果を踏まえ、本研究の目的に照らして、分析対象とする外在的要因とそれらを表す測度を選択する。測度の正確な測定については、最も測定が難しい論文著者が過去に発表した論文数を取り上げる。著者の過去論文数を調べる一般的方法は、データベースを用いて著者名サーチを行うことであるが、その際、同姓同名の異なる著者の論文が混在することが大きな問題になる。本研究では、同定洩れを多少犠牲にしてもノイズを最小限に抑えることを目標に、真の著者による論文（真論文）と同名異人著者による論文（偽論文）を半自動的に識別する比較的簡便な方法を検討する。

(2)に関して、論文の被引用数へのいろいろな要因の影響について多くの研究がなされている。しかしながら、どの要因が被引用数に有意な影響を及ぼすか、その影響の強さはどの程度か、について明確な合意は得られていない。その理由の一つは、これまでの研究の多くが、ある単一の要因に着目している（あるいは複数の要因をそれぞれ独立に見ている）ために、要因間の相互作用が明らかでないことにある。もう一つの理由は、様々な潜在要因を総合的に考慮したいくつかの研究は、対象の論文集合が特定の分野や雑誌に限定されているため、結論の一般性に限界があることである。本研究は、複数の分野について、同

じ年に発表された同じタイプの論文（英語の原著論文）の被引用数に及ぼす種々の外在的要因の影響を体系的に分析することにより、影響を与える主要因について分野共通の傾向があるかどうかを探ることに特徴がある。

(3)に関して、論文の引用の老化や持続性についていろいろな観点から研究が行われているが、被引用数自体についての研究に比べると、系統的な知見の蓄積が不十分である。その本質的理由の一つは、論文の引用持続性を測る定量的な指標が確立されていないことにあると考えられる。Wang et al. (2015)による Citation Delay（記号として D を用いる）は、これまで提案されたものの中で最適の引用持続性測度と考えられるが、その性質についてはほとんど解明されていない。本研究では、D の分布の特徴、長期的被引用数との関係、(2)で用いた外在的要因への依存性を分析することにより、論文の被引用数に対する引用計測期間の影響を明らかにすることを目指す。(2)と同様、複数の分野において分析を行って、分野を超えた共通の傾向が見られるか否かを検討する。

【方法】

上述のように、本研究では複数の分野において分析を行い、分野共通の傾向が見られるかどうかを明らかにすることに主眼を置いているが、その分野には、Web of Science (WoS)の主題カテゴリーから次の 6 つを選択した : (a)物性物理学、(b)無機・核化学、(c)電気・電子工学、(d)生化学・分子生物学、(e)生理学、(f)消化器病学。そして各分野から、その主題カテゴリーのみに分類されている英語誌を 4 誌ずつ、発行国とインパクトファクターに大きな偏りがないように選択し、それらの雑誌からサンプルとする 1,395 論文（分野ごとに 230-240 件）を抽出した。サンプル論文はすべて、2000 年に発表された原著論文（WoS の記事タイプ”article”）である。

研究目標(1)における論文著者の過去論文数データの測定では、上記に示した 6 分野の全サンプル論文の著者が 1970-2000 年の期間に発表した論文を、著者名サーチ(姓＋イニシアル)により WoS を用いて検索した。検索された 60 万件以上の論文を真論文と偽論文とに識別するため、次の情報について、もとの論文との類似性あるいは関連性を数値化した : (a)共通の共著者の存在、(b)所属機関アドレス、(c)雑誌の引用関係、(d)タイトル語、(e)直接引用関係、(f)共引用関係、(g)発表年の間隔、(h)著者が特定の国の所属であること。まず、(a)、(b)、(c)を用いた一次フィルタリングにより「真らしい」論文に絞り込みを行い、通過した論文から 3,000 論文を抽出して、真偽の目視判定を行った。この判定結果を目的変数、上記(a)～(h)を説明変数とするロジスティック回帰モデルを考え、これによって得られた回帰式を判別関数として、二次のフィルタリングを行った。

研究目標(2)における論文の被引用数に及ぼす外在的要因の影響に関する研究では、上記の 6 つの分野それぞれにおいて、負の 2 項重回帰分析を行った。目的変数は、各サンプル論文の 6-7 年間(2000-2006)と 11-12 年間(2000-2011)の被引用数である。説明変数は、(1)で選択した外在的要因を表す次の測度である : 著者数 ; 著者所属機関数 ; 著者所属国数 ; 参考文献数 ; Price 指数（参考文献中最近 5 年のものの割合）; 図の数 ; 表の数 ; 数式の数 ;

規格化したページ数；第一著者の論文生産性；第一著者の論文の引用インパクト；第一著者の活動期間。このうち後 3 者の測度データは(1)の結果に基づく。この他、各雑誌をダミー変数とした。

研究目標(3)における引用持続性指標(*D*)の性質に関する研究でも、同様に 6 つの分野それぞれにおいて分析を行った。発表年である 2000 年から 2014 年までの毎年の被引用数のデータから各論文の *D* を計算し、その分布を正規分布と比較した。次に、*D* の変化に伴う長期被引用数(2000-2014)の変化を分析した。更に、*D* を目的変数とし、 (2)で用いた測度に加え長期被引用数と発表誌のインパクトファクターを説明変数とする重回帰分析により、どんな要因が *D* に影響を及ぼすか検討した。

【主な結果】

研究目標(1)について、検索で得られた約 62.9 万論文のうち約 9.0 万論文(14.3%)が最終的に真論文と判定された。二次フィルタリングに用いたロジスティック回帰モデルの再現率と適合率はともに 95%程度であることが実証された。直接引用関係または共引用関係があれば、確実に真論文であることが判った。同名の共著者の存在及びアドレスの類似度は重要な判別要素であり、タイトル語の類似性と雑誌間の引用関係強度も判別に有効であった。また、特定の国(日・中・韓・台)の著者に同名異人が極めて多いため、著者がそれらの国に所属するか否かを判別式に組み込むことが重要であった。

研究目標(2)に関する負の 2 項重回帰分析では、Price 指数が全ての分野で最も強力な被引用数への影響要因であり、次いで参考文献数が重要であることが見出された。著者数、第一著者の論文生産性と引用インパクトも弱い影響度が認められた。2 つの被引用数計測期間（6-7 年と 11-12 年）の間で、結果はほぼ同様であった。参考文献数は多くの先行研究で有意な関係が示されているが、Price 指数の予測変数としての重要性を全分野で示したことは、本研究の最も主要な発見の一つである。このように、分野を越えてある程度の共通性が見出されたが、有意となった変数はあまり多くない。その理由のひとつは、用いたサンプルが比較的小さい（各分野 230-240）ことである。しかし、逆にサンプルが大きすぎると、被引用数の予測にあまり重要でない変数まで影響を持つという結果が得られてしまうことがある。本研究で有意であった変数は、確実に被引用数に影響を持つ要因であると考えられる。

研究目標(3)について、まず *D* の分布はかなり正規分布に近いことが明らかになった。また、被引用数分布が分野により大きく異なるのに対し、比較的分野間格差が小さい。*D* と長期被引用数 *C* の間には正の相関があるがその関係は線形ではなく、ある *D* で *C* の平均値は最大値をとる。しかし、雑誌単位で見ると、引用インパクトの高い雑誌では *D* はむしろ低くなるという一見矛盾するが興味ある傾向が見出された。*D* を目的変数とする重回帰分析では、Price 指数が低く、図の数が少なく、表の数が多いほど *D*(引用持続性)が高くなる傾向があった。以上のことは概ね 6 つの分野に共通である。引用持続性指標の性質についてのこれらの発見は本研究の大きな特徴である。