２Ｉ１５

科学研究費助成事業データベースと科学論文書誌データベースの高精度データ連接

○富澤宏之，伊神正貫，阪彩香（文科省・ＮＩＳＴＥＰ）

1. はじめに

政府の研究開発ファンディング・システムは、公共財的な科学技術知識の生産において、中核的役割を果たしている。これらは、政府資金を配分するシステムであることから、その効果を国民に示すことが求められている。しかし、これまで、ミクロ・レベル（プロジェクト・レベル）でのインプットーアウトプットデータの関係性についての網羅的なデータが存在しないことから、政府の研究開発ファンディング・システムとアウトプットの関係性については、分析が困難であった。

科学研究費助成事業(科研費)は、我が国における最大の研究ファンドであり、その予算額は、政府の競争的資金の予算額全体の約半分を占めている。また、科学研究費助成事業は、研究課題の情報とそこから生み出された成果についての情報が時系列で収集され、かつデータベース化(科学研究費助成事業データベース: https://kaken.nii.ac.jp/)されている唯一の制度である。

科学研究費助成事業データベース(以下、KAKEN と呼ぶ)は、Web 上で一般に公開されており、科学研究費助成事業の成果を把握する上で、貴重なデータベースといえる。しかしながら、科学研究費補助事業の成果についての統計的な分析を行うには以下のような困難があった。

A) 論文等の成果情報に重複があり、KAKEN に収録されている成果数が、そのまま科学研究費助成事業の成果数にはならない。通常、成果情報には表記の揺れや情報の欠落が存在するため、重複を排除することは容易ではない。
B) 多様な成果が報告されているため、統一した基準のもとで、科学研究費助成事業の成果数やその時系列変化を把握することが難しい。

これらの困難を克服するために、科学技術・学術政策研究所において、KAKEN に収録されている成果情報とトムソン・ロイター社の Web of Science (WoS)を著者情報、論文タイトル、書誌情報等の類似性に基づいて高精度データ連接するプログラムを開発した。本報告では、用いたデータ、データ連接の方法論、データ連接の精度について説明したのち、日本の WoS 論文に注目し、その中で科学研究費助成事業が関与している論文の割合について分析した結果について紹介する。

2. データ連接の対象データ

データ連接の対象として、KAKEN とトムソン・ロイター社の科学論文データベース Web of Science を用いた。それぞれの概要は以下のとおりである。

① 科学研究費助成事業データベース（KAKEN）

国立情報学研究所よりデータベースの貸与を受けた。使用したデータは、KAKEN_XML(2012 年 3 月 16 日更新)である。

KAKEN_XML は、採択課題（研究課題番号、研究課題名、研究機関、研究分野名、種目名、配分額など）と報告書（実績報告、研究成果報告書概要、研究成果報告書など）のデータから構成されている。今回用いた 2012 年 3 月 16 日更新版には、1965-2011 年度の採択課題の情報、1985-2009 年度の報告書の情報と、2010 年度の報告書情報の一部が収録されている。

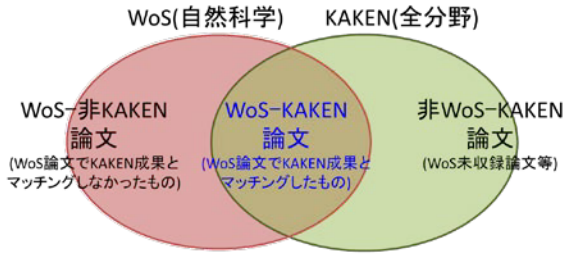
② Web of Science データベース(WoS)

使用したデータベースは Web of Science(2011 年 12 月末バージョン)である。論文の収録期間は 1981-2011 年（データベース年）となっている。

WoS のうち、自然科学系の雑誌を収録対象としている SCIE(Science Citation Index Expanded)を分析対象とした。文献種類のうち Article, Article & Proceedings, Review, Note, Letter について KAKEN とのマッチングを行った。

WoS 論文と KAKEN 成果の包含関係について、図1に示す。ここで、WoS は WoS 論文の集合であり、KAKEN は KAKEN 成果の集合である。WoS 論文は自然科学を対象とする一方で、KAKEN は全ての分野を対象としている点に注意が必要である。

図 1 WoS 論文と KAKEN 成果の包含関係



3. データ連接の方法論

データ連接を個別論文の書誌データのレベルで行うためにはデータのマッチングが必要である。その方法としては、書誌データの一致性をコンピュータのプログラムによって判定する方法が広く用いられるが、マッチング対象の書誌データの状態によって具体的な方法は多少異なってくる。

まず、論文を同定するために最小限必要な書誌データを用いる方法が考えられる。例えば、(1)掲載誌名、(2)出版年、(3)巻号、(4)ページ、の 4 項目をマッチ・キーとして用いることにより、論文を同定できる [1]。この4項目は、理論的には論文を同定するための情報として充分であるが、4項目とも完全に一致していないと同一の論文とは判定できないため、データの不備が多い場合には有効でない。しかし、KAKEN成果データには、出版年、巻号、ページの情報のいずれかに不備があるデータが相当数ある [2]。また掲載誌名については、略記がなされている場合が多いが、略記方法が統制されておらず、一つの掲載誌名について、様々な表記が存在する。

以上のような状況であるため、本研究では、論文の同一性判定を人間が目視で行う方法を参考にし、書誌データの不備の多い場合に適したマッチング方法として、論文タイトルと著者名を含めた書誌情報全体を活用するアプローチを採用した。

特に、論文タイトルと著者名を重要な判定基準として活用したが、これらのテキストデータはスペルミス等を多く含むため、テキストの完全一致性でなく、テキスト類似性を一致度の指標とした。具体的には、単語の一致率を基本的な指標とした。

しかし、用いられている単語はよく似ていても、語の順番まで考慮すると違いが大きい場合もあるため、テキストのトリグラム一致率を補助的な指標として用いた。これは、テキスト中の連続する3つの単語を構成単位(これをトリグラムと呼ぶ)として扱い、両テキストにおいてどの程度、トリグラムが一致するかを測定する方法である。更に、多少のスペルミスがあっても全体として類似している場合を見落とさないようにするため、単語ではなく文字(アルファベットと数字)を要素としたトリグラム(連続する3つの文字が構成単位)の一致率も併用した。

テキスト類似性は、論文タイトルだけでなく、掲載誌名についても適用した。掲載誌名については、略語が用いられる場合が多いため、英語においてよく用いられる略語の辞書を参照して、類似度を測定した。

以上のように、本研究では、これらの複数の項目(前述の4項目に論文タイトル、著者名を加えた6項目)を同一性判定の基準とした。ここで問題となるのは、これらの項目のうち何項目が一致し、また、テキスト類似度がどの程度であれば同一論文と判定できるかである。

最終的には、6項目の様々な組み合わせについて調べ、何項目がどの程度一致していれば同一論文と判定できるかを、経験的・実証的に決定した。

以上の方法によるマッチングのコンピュータ・プログラムを用いて、表1に示すような結果が得られた。元の KAKEN 成果データ約 175 万件のうち、53%に当たる約 93 万件が WoS 論文として同定された。KAKEN 成果では、同一の論文の書誌情報が重複して記載されることがあるが、重複を除いて数えると、KAKEN 成果全体のうち 38%が WoS 論文(WoS-KAKEN 論文)であった。

表1　KAKEN 成果データの WoS とのマッチング結果

	重複排除前		重複排除後	
KAKEN 成果全体	1,749,135	(100.0%)	790,838	(100.0%)
WoS 論文	929,049	(53.1%)	303,426	(38.4%)
非 WoS 論文	820,086	(46.9%)	487,412	(61.6%)
うち英語論文	374,095	(21.4%)	210,818	(26.7%)
うち日本語論文	445,743	(25.5%)	276,354	(34.9%)

注：一定精度のコンピュータ・プログラムによる集計値であるため不定性がある。

4. マッチングの精度

第3節に示したマッチング結果の精度を評価するために、無作為抽出を行い、それらについて、目視によるチェックを行った。無作為抽出は、分野による偏りを避けるために、KAKEN データベースより8つの研究領域のそれぞれについて 140 件(合計で 1,120 件)の書誌データを抽出した。

1 実際に、欧州委員会が実施した特許による科学論文の引用データと科学論文データベースとのマッチングでは、この方法が用いられている(富澤, 2010)

2 例えば、WoS 論文と確認された KAKEN 成果の 33%は、論文の開始ページが正しく記載されていなかった。

この 1,120 件のデータのうち、WoS とのマッチングの対象となる英語タイトルの論文は 833 件であり、そのうちの 622 件がコンピュータ・プログラムにより対応する WoS 論文が同定されていた(表2)。そのうち目視により正しい結果であることが確認できたのは 619 件(判定結果の 99.5%)であり、3 件(判定結果の 0.5%)については“誤判定”(正しくない WoS 論文が同定)であった。

表2 無作為抽出と目視によるマッチング結果の評価

	件数	割合
日本語タイトル論文	287	-
英語タイトル論文	833	-
WoS 論文と判定	622	100.0%
うち正しい結果	619	99.5%
うち誤判定	3	0.5%
見落とした WoS 論文	23	(3.7%)
うちデータ不備の無いもの	5	(0.8%)
総数	1,120	-

マッチングのエラーにはもう一つのタイプがある。すなわち、実際にはWoS論文であるにもかかわらずマッチングされなかった“見落とし”であり、それは 23 件であった。これは、目視によりWoS論文と確認した全642件(“正解データ”)の 3.6%に相当する。

この 23 件について詳しく見ると、そのうち 15 件は、KAKENデータベースにおいて、「印刷中」や「投稿中」などと付記された論文であり、掲載誌における巻号やページ番号、出版年などの書誌情報が記載されていなかった [3]。また 3 件については、「印刷中」等ではないものの、掲載誌における巻号やページ番号、出版年などの書誌情報が不十分であった。これら 18 件については、データ側の不備と考えるべきであり、目視での判定においてもWoS論文と判断すべきかどうか疑問もある。その意味で、“見落とし”に該当するのは、23 件ではなく5 件(判定結果に対する割合は 0.8%)と考えるべきかもしれない。

これらのマッチング結果は、データ分析の精度にどのような影響を及ぼすのだろうか。これは、データ分析の目的によっても多少、異なってくる。

まず、「WoSに収録された日本の論文全体のなかに、科学研究費助成事業の成果論文はどの程度、含まれているのか?」という量的な状況を明らかにしたい場合、このマッチング結果データは、WoS論文を 3.7%程度(あるいは不備データを除くと0.8%)見落としているため、このデータを用いた場合、科学研究費助成事業の成果論文の数を、その程度、過小評価していることになる。その一方で、誤判定の部分については過大評価になるが、それは0.5%程度であるので、全体としては3.2%程度(あるいは不備データを除いた場合は 0.3%)の過小評価となっている [4]。ここで知りたい量的状況については、「科学研究費助成事業の成果論文は少なくともx%である」といった下限値を明らかにできれば良いことが多いので、このように過小評価となっていることは、実用上、大きな障害にならないであろう。

一方、「WoS に収録された科学研究費助成事業の成果論文の性質を明らかにする」ことが目的である場合、このデータは誤判定の部分を 0.5%含むのみであり、したがって概ね 99.5%の精度のデータと扱うことができるであろう。

5. 指標としてのデータ連接結果

以下では、データ連接結果から得られた事実として、日本の WoS 論文のうち科学研究費助成事業(科研費)が関与している論文の割合やその時系列変化を紹介する。

日本の論文数におけるKAKEN成果の関与度をみるため、日本のWoS論文をKAKEN成果とそれ以外に区別して示した(図2)。ここで、日本のWoS論文のうち、KAKEN成果とマッチングした論文を「WoS-KAKEN論文」、日本論文のうちKAKEN成果とマッチングしなかった論文を「WoS-非KAKEN論文」と表記した [5]。

(A)は日本の論文数の内訳を積み上げグラフで示している。日本の論文数は、1996 年以降緩やかな上昇を見せたが、2000 年代に入り横ばいとなっている。

(B)は日本の論文数を 100%とした積み上げグラフで示した結果である。1996 年以降、日本の論文のなかでWoS-KAKEN 論文が占める割合が年々増加しており、2008 年には日本の WoS 論文の 47.3%が WoS-KAKEN 論文となっている。

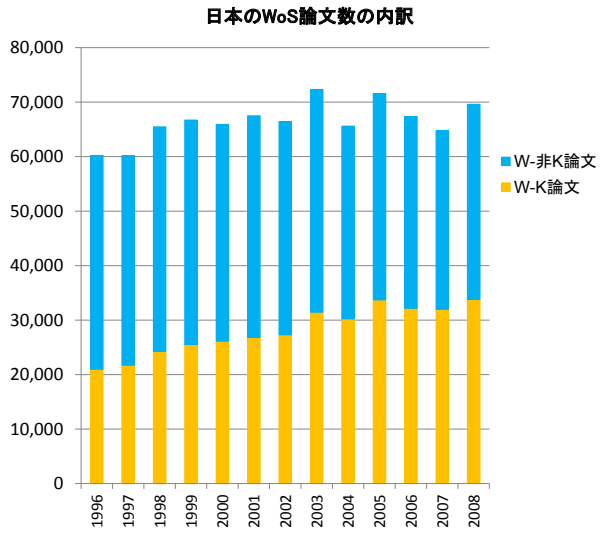
3 なお、「印刷中」などと付記された論文でも WoS とのマッチングが正しくなされたものも相当数あるが、それらは、(1)掲載誌名が適切に書かれている、(2)出版年が正しく書かれている、という2つの条件を満たすものに限られていた。

4 WoS 論文の“見落とし”と“誤判定”はトレードオフの関係にある。すなわち、見落としは、単純にマッチングにおける一致度の閾値を下げることにより、減少させることができるものの、同時に誤判定の件数が増加することになる。“見落とし”よりも“誤判定”の方が深刻な問題であることからも、“誤判定”が“見落とし”を下回っていることが望ましいであろう。

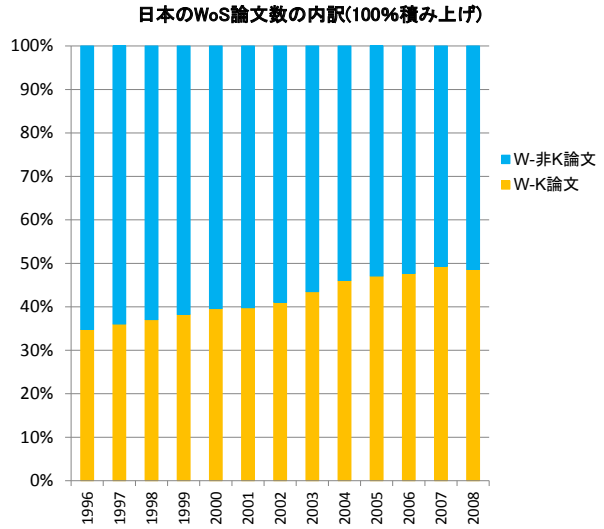
5 なお、WoS-KAKEN 論文は、科研費を獲得した研究者が科研費が関与した論文として報告したものであるが、科研費が研究成果にどの程度寄与しているかは不明である。実際、論文や研究プロジェクトを対象とした質問票調査からは、研究者が研究を実施する際に様々な財源を活用していることが明らかになっている(富澤ら, 2006; 長岡ら, 2010)。したがって、ここで示すのは、日本の論文生産における科研費の関与度合いを見るためのデータであって、科学研究費助成事業のインプットとアウトプットの関係性の分析を行う為に充分なデータではない。

図 2 日本の論文における KAKEN 成果の関与度

(A)



(B)



6. まとめ

科学研究費助成事業データベース(KAKEN)は、科学研究費助成事業の成果を把握する上で、貴重なデータベースといえる。しかしながら、論文等の成果情報の重複や、統一した基準のもとで成果数やその時系列変化を把握することが難しいといった理由から統計的な分析を行うことは困難であった。

これらの困難を克服するために、科学技術・学術政策研究所において、KAKEN に収録されている成果情報とトムソン・ロイター社の WoS を著者情報、論文タイトル、書誌情報等の類似性からデータ連接するプログラムを開発した。このプログラムを用いると「誤判定(偽陽性率)」は 0.5%、「見落とし(偽陰性率)」は高めに評価した場合でも 3.7%と、高精度で KAKEN と WoS のデータ連接が可能であることが確認された。

このプログラムを用いることで、KAKEN に成果情報が重複して収録されている場合でも、それに対応する WoS 収録論文(WoS 論文)がユニークに決定されるので、成果情報の重複が排除される。このことは、計量不可能であったデータが計量的研究になったことを意味する。また、WoS は一定の基準を満たした論文雑誌が収録対象となっているので、統一した基準のもとで、科学研究費補助事業の成果数とその時系列変化を把握することが可能となる。重複を排除したのちに、日本の論文数に占める WoS-KAKEN 論文の割合をみると、近年では約半分の論文に科研費が関与していることが分かった。

本研究において開発したデータ連接手法は、不定形の文献書誌情報を定量的に分析するために極めて効果的であり、様々な応用が考えられる。例えば、KAKEN 成果データには、WoS 論文だけでなく、日本語論文や各種の文献が収録されているが、それらのデータに対して、本研究で開発したデータ連接プログラムを重複排除プログラムとして適用することにより、論文・文献の実数の集計が可能になる。また、特許ドキュメントに引用された科学論文の情報は、科学研究と特許発明の関係を示す貴重なデータ(いわゆるサイエンスリンケージ)と考えられるが、本研究で開発したデータ連接プログラムを適用することにより、そのようなデータの精度の高い分析が可能になる(富澤, 2010)。

このように、関連データの相互連接は、極めて有用な分析を可能にするものであり、今後、一層の精度向上と適用可能性拡大を探求する意義があると考えられる。

参考文献

富澤宏之, 科学論文を引用することは特許の影響力を増大させるか, 研究・技術計画学会第 25 回年次学術大会, 講演要旨集, pp. 499-501, 2010 年 10 月.

富澤宏之, 林隆之, 山下泰弘, 近藤, 正幸, 優れた成果をあげた研究活動の特性:トップリサーチャーから見た科学技術政策の効果と研究開発水準に関する調査報告書, 科学技術・学術政策研究所(調査資料-191), 2006 年 3 月

長岡貞男, 伊神正貫, 江藤学, 伊地知寛博, 科学における知識生産プロセスの研究 -日本の研究者を対象とした大規模調査からの基礎的発見事実 -, 科学技術・学術政策研究所(調査資料-191), 2010 年 11 月