統計分析用特許データベースの進展

IIP パテントデータベース

東京大学工学系研究科教授　元橋　一之

PROFILE

東京大学工学修士、コーネル大学 MBA、慶応大学博士（商学）。経済産業省、OECD エコノミスト、一橋大学助教授などを経て 2006 年から現職。主な著書に『IT イノベーションの実証分析』（東洋経済新報社）、『日本経済競争力の構想』（日本経済新聞社）など。

✉　TEL

1 はじめに

特許データベースは企業などにおける先行技術調査などで利用されることが多い。また、最近では特定技術分野におけるパテントマップの作成やそれらの情報を利用した技術経営戦略の立案などに用いられてきている。このような特許データに対するニーズを反映して、一部の大企業や特許を専門とする調査会社では、独自のデータベースを構築して、社内での活用や外部ユーザーに対する提供を行っている。

特許データベースに対するニーズはこのような企業ユーザーのみならず、大学などにおける社会科学研究者の間においても高まっている。特許データを用いることによって、産学連携などのオープンイノベーションに関する定量分析や産業クラスターや国際的な技術スピルオーバーなどの幅広い研究を行うことが可能となる。また、特許データと企業財務諸表を接続したデータベースを用いることによって、企業の無形資産に関する定量分析など幅広い分析を行うことが可能になる。企業の研究開発やイノベーション活動は企業にとって重要な秘密情報であることが多いが、詳細な技術情報が公開されているという点で、研究者にとって重要は情報源といえる。

「IIP パテントデータベース」は、こうした要望に応えるべく、後藤晃氏（東京大学名誉教授）と筆者が中心となって構築した、我が国で最初の公開型の本格的な実証分析用特許データベースである（当データベースは、財団法人知的財産研究所の HP 上で公開されている。構築プロセスの詳細は、Goto and Motohashi(2007) “Construction of a Japanese Patent Database and a first look at Japanese patenting activities”, Research Policy, Volume 36, Issue 9, November 2007, Pages 1431-1442 を参照されたい）。

2 データベースの概要とアップデート

IIP パテントデータベースは、概ね月２回のペース公表されている「整理標準化データ」をベースに作成されている。「整理標準化データ」は SMGL や XML などのタグ付きテキストファイルとして特許情報が収録されたものである。ここでは、これらのテキストファイルをデータの統計的処理を容易にするために SQL データベースに変換し、更に研究者においてもっともニーズの高いと思われるものを CSV 形式のテキストファイルとして公開している。現時点では、1964 年 1 月以降の出願から 2009 年 10 月時点で公開されたもの（整理標準化データの 2009 年度第 15 回公表分）までを取り込んだものとなっている。

IIP パテントファイルとして CSV 形式で公開しているデータには、特許出願データ（出願番号、出願日、審査請求日、技術分野、請求項数等）、特許登録データ（登録番号、権利消滅日等）、出願人データ（出願人名、個

法官コード、国・県コード等）、権利者データ（権利者名等）、引用情報データ（引用・被引用特許番号等）、発明者データ（発明者名称、住所）が含まれている。データベースの構成とテーブル毎のデータ数については、図１のとおりである。例えば出願特許数でいうと11,254,825 件の特許データが収録されており、そのうち 3,507,336 件の特許が登録されている。それぞれに出願人、権利人に関するテーブルが接続しており、また引用データは審査官引用（審査請求があった特許に対して、審査官が拒絶理由を付す際に引用された過去文献）に関するデータが収録されている。

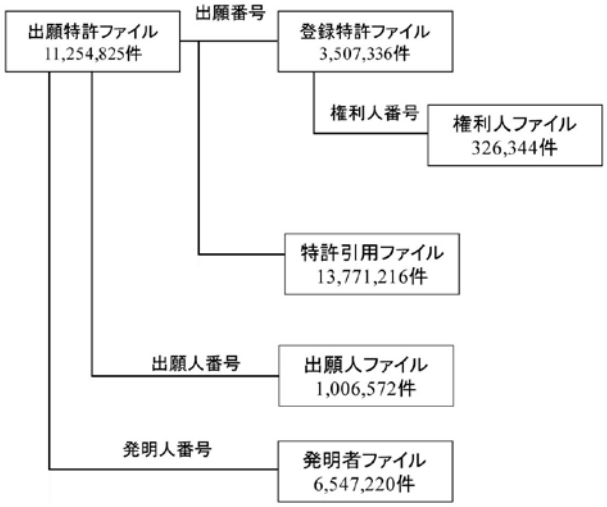

図１：IIP パテントデータベースの構成

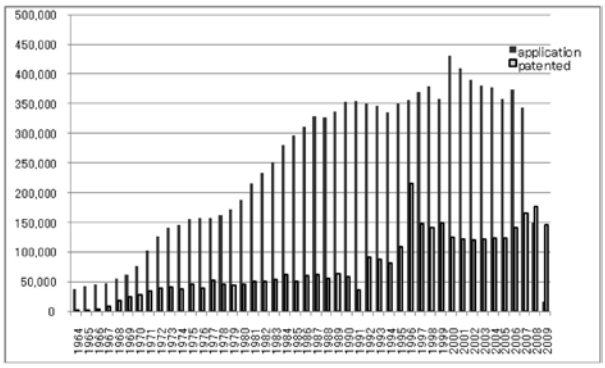

図２：出願特許推移と登録特許数の推移

このデータを用いることによって、出願年や技術分野別の特許数などの特許データに関する記述データを容易に作成することができる。図２は同データによる出願年別の特許出願数と登録公開年別の登録特許数の推移をみたものである。なお、出願特許数が 2008 年から大きく減少しているのは、本データベースが 2009 年 10 月までに公開された特許をベースにしていることによる。

3 IIP パテントデータベースの改良

IIP パテントデータベースは「整理標準化データベース」における情報を忠実に取り出して、データベース化したものであるが、このデータを用いて分析を行う上ためには、オリジナルのデータにおいていくつかの問題がある。そのうち最も重要なのが、出願人、権利人、発明者などの情報の標記の揺れの問題である。例えば 1960 年代などの古い時代のデータはこれらの名称がカタカナ表記されているのに対して、最近では漢字標記になっていているのでオリジナルのテキスト情報のみからは名寄せはできない。また、企業の名称変更や表記方法の変更によっても、本来であれば同じ企業であってもデータベース上では違うものとして認識されてしまう。そこで我々は主に出願人情報について名寄せ作業に取り組んでいる。図３は出願人名称の名寄せフローをしめしたものである。

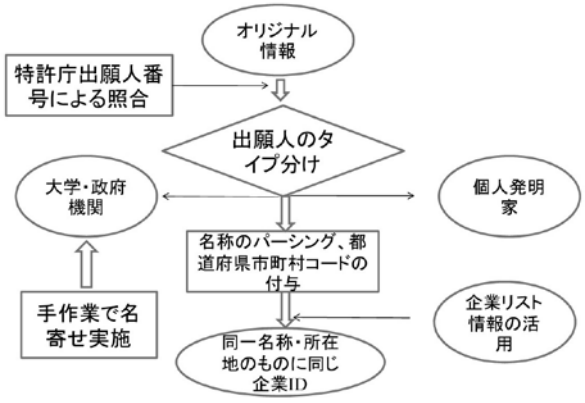

図３：出願人名称の名寄せ作業フロー

ここでの作業は、まず特許庁の出願人コードを活用することから始める。ただし、この出願人コードは現在の９ケタコードに至るまで、コードの変換が何回か行われていることから、これをまず補正する。なお、特許庁の出願人コードは、False Negative（本来同一の出願人

に対して違うコードが振られる）という問題があるが、False Positive（違い出願人に対して同じコードが振られる）という問題はない。

この状態から出願人の名称情報を用いて出願人のタイプとして、(1) 個人、(2) 企業、(3) 非営利機関（官庁、公的研究機関など）(4) 大学に分類を行う。このうち、(3) と (4) については件数が少ないためマニュアルで名寄せ作業を行っている。

更に企業出願人を取り出して企業名称の標準化を行い、そのうえで住所情報から同一市町村の存在に存在する同一名称の企業を同一企業として新たな ID 番号を付与している。なお、この方法によると企業名称の標準化が不十分な場合や名称変更が行われた場合などにおいて False Negative の可能性がある。また、同一名称で違う企業が同一所在地に存在する場合は False Positive の可能性もある。

これらの問題を解決するためには、所在地情報を含んだ正確な企業名称に関する情報が必要である。イノベーションデータベース整備にあたって企業活動基本調査との接続を行っているが、現時点ではこの情報を特許データの企業名寄せには用いていない。また、日本におけるすべての事業所・企業をカバーする事業所企業統計の名簿情報を用いればより広範囲の出願人名寄せに関する False Positive の問題を解決することができる。更に、ここでの作業は主に日本に所在する出願人に対して行われたものであることに留意することが必要である。欧米の企業などの外国における出願人についても今後の作業として残っているところである。なお、欧米の企業の名寄せについては OECD や NBER グループなどによって作業が進んでいるところで、これらのグループとの連携によってある程度の対応が可能となる。

4 イノベーションデータベース基盤の整備

東京大学においては、経済産業省の委託研究（産業技術調査事業）などを受けながら、IIP パテントデータベースと統計調査などの他のデータベースとの接続を行うデータベース基盤整備事業に取り組んでいる。ここでの中核的なデータは、IIP パテントデータベースの他、科学技術研究調査（総務省）と企業活動基本調査（経済産業省）の企業レベル個票データである。

まず、科学技術研究調査は企業の他、大学や公的研究機関における研究開発活動を総合的に調査しているものであるが、ここでは資本金 1 億円以上の企業に対して行われている企業等 A の調査項目について、1984 年からのパネルデータを作成している。科学技術研究調査においては、科学コードという番号で企業データの整理が行われているが、コードの付け替えが行われている年があり、パネルデータの作成にあたっては新旧コードの対応関係を把握することが適当である。この作業によって、5414 企業のアンバラストパネルデータ（最多年の 2002 年の企業数が 3650）を作成した。

一方、企業活動基本調査は 1991 年に開始された比較的新しい統計調査であるが、こちらは永久企業番号という期間を通じて統一的なコード体系が整備されており、パネルデータの作成は容易である。資本金 3000 万円以上でかつ従業員数 50 人以上の製造業または卸小売業（2001 年から一部のサービス産業に対して業種が拡大）に属するすべての企業に対する調査であり、毎年約 2.5 万社のサンプル数となっている。

これらの統計調査の企業パネルデータに IIP パテントデータの出願人（日本に所在する企業のみ）約 60 万社を接続させたものがイノベーションデータベース基盤の中核的な構成要素となっている。

また、これらまでのデータ接続作業としては、これらの中核データの整備に加えて、「知的財産活動調査」（特許庁）や新規に行ったライセンスに関するアンケート調

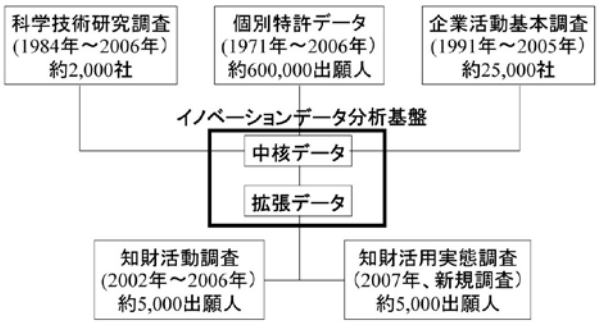

図３：イノベーションデータ分析基盤のフレームワーク

査（知財活用実態調査）を接続し、データ拡充を行っている。知財活動調査は特許庁において行われている承認統計であり、毎年一定数の特許、実用新案、商標及び意匠の出願を行っている出願人（企業の他、大学や個人発明家含む）に対してして毎年行われている。内容は、各種知財の出願状況やライセンシングなど知財利用状況、知的財産活動費や知財侵害の実態に関するものである。このデータは特許庁における出願人番号をベースとして管理されているため、中核データの個別特許データ（同じ出願人番号の情報が存在する）との接続が可能である。

知的財産活動調査は出願数やライセンスに用いられる特許数など定量的な情報として有益なものが存在するが、特許データをイノベーション分析に用いるためには、特許性向（すべての発明が特許化されるわけではないので、発明案件のうちどの程度の割合のものを特許出願するか）や特許利用の実態に関する詳細な情報（例えば特許保有の目的として防御的な目的なのかライセンシングアウトを目的としているのか）が必要となる。これらの情報は知的財産活動調査から得ることができないので、新規にアンケート調査（ライセンス活動実態調査）を行いデータベースの拡充を行っている。

5 課題と今後の取り組み

IIP パテントデータベースは実証分析用のデータベースとして極めて有用であるが、利用に際しては次のような問題を含んでいることも理解しておかなければならない。

まず、出願人名に関して、同一企業にもかかわらず、出願人名や住所の表記に揺れがある場合や異なる事業所から出願されている場合などに、異なる出願人番号が付けられていることがある。前述したように IIP パテントデータの出願人名称の名寄せについてはかなりの作業を行ってきているが、まだ改良すべき点は残っている。

欧米の特許データベースも同様な問題を抱えており、各国において効率的な名寄せ手法に関する研究が進められているその成果の一部は、OECD や EPO が毎年主催している国際学会等において報告されており、標記の揺れを勘案したテキストマッチング（approximate matching）の手法についても開発が進んでいる。

IIP パテントデータの出願人について特に大きな問題として残っているのは、日本以外の出願人をどうするかである。とくに外国出願人はカタカナ表記となっていることから、標記の揺れが大きい。この点については、上記の欧米の特許データベースを開発しているグループとの連携を進めているところである。上記のグループにおいては、EPO や USPTO に対する出願特許において、欧米企業を出願人とするものの作業が進んでいる。日本に出願してくる欧米の企業はそれぞれの国・地域における出願特許を優先権として国際出願することが通常だと考えられるので、国際的な特許ファミリーの情報を用いることによって、欧米グループの欧米企業に関する名寄せ結果を我々のデータベースに移送することができる。逆に、欧米グループは欧米特許における日本企業出願人の名寄せに苦労しており、当方の情報を移送することによって、両者においてメリットが大きい。現在、主に欧州チームとの連携によって、上記のプロジェクトを進めているところである。

また、特許データと企業データの接続については、これまでは企業活動基本調査の対象企業に限られていた。従って、今後は日本の存在するすべての事業所・企業を包含する事業所・企業統計との接続について行っていくことを予定している。これは OECD において進められている特許データとビジネスレジスターの接続とデータ分析プロジェクトにも対応するものであり、我が国企業の特許活動の状況を欧米における状況と比較できるという点でも意義が高い。

更に、OECD においては、商標に関するデータベースの構築など、特許以外の知的財産権データに対する取組も始まっている。国際的に特許を中心とした研究者向けのイノベーションデータベースの整備が進む中、我々としても積極的に新しいプロジェクトに取り組み、その成果を公表していきたいと考えている。