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。

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、東京大学を中心とする研究者グループにおいて構築された、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）。なお、データのアップデートや更新については、筆者が委員長を務める IIP パテントデータベース運営委員会において行われている。毎年のデータベースアップデートを行うと同時に、同委員会においてはデータの問題点の指摘や解決案の検討、新たなデータ整備の方向性について定期的な検討を行っている。以下、同データベースのアップデートと改訂の状況について述べる。

2 IIPパテントデータベースの概要

IIP パテントデータベースは、年間 25 回（2 週間に 1 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。

また、2011 年 3 月時点まで公開されたもの（整理標準化データの 2010 年第 23 回公表分）については β バージョンとして一部のユーザーに公開し、必要な修正を加えたものについて 2011 年末をめどに公開する予定である。

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。現在公開さ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。

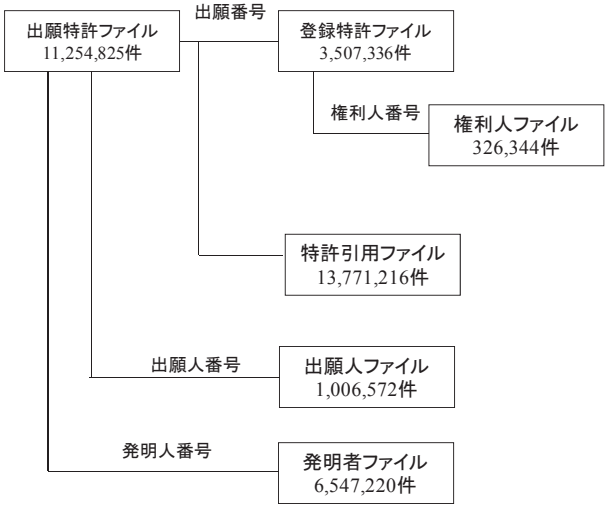

図 1 ：IIP パテントデータベースの構成

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。

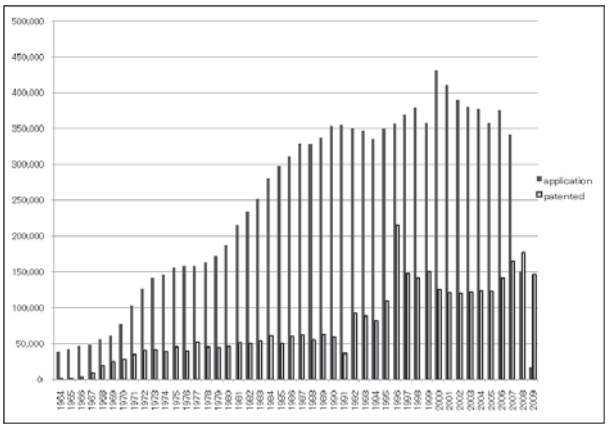

図 2 ：出願特許推移と登録特許数の推移

3 出願人名寄せの状況

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。

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、

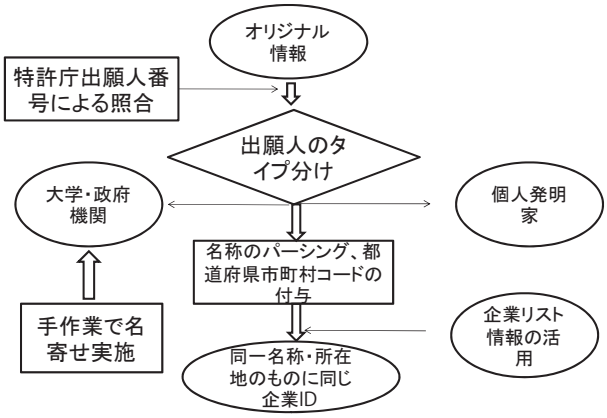

図３：出願人名称の名寄せ作業フロー

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。そのためには、財務データベースなどの企業データベースと特許データの出願人情報を接続し、より正確な出願人名寄せを行うことが有益である。現行においては、上場企業における財務データベースや政府統計の個票データ（企業活動基本調査や事業所・企業統計など）と特許データを接続した事例がみられるが、その結果がまとまった形式で提供されていない。今後の検討事項として重要な点である。

なお、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きた。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。今年度から当該作業を、文部科学省科学技術政策研究所（NISTEP）の「政策のための科学」データベース基盤整備事業に引き継ぐこととしており、そこでの作業内容を IIP パテントデータベースの出願人名寄せに利用することを検討している。

更に、ここでの作業（企業データベースと特許データの接続）は日本企業の出願人に対するものであることに留意する必要がある。日本特許庁には、米国や欧州などの海外の企業も多数の特許を出願しており、これらの企業の名寄せをどうするかという問題がある。整理標準化データにおいては、それらの企業名称はカタカナで記述されていることから、標記ゆれの問題が大きく、名寄せを行うことの必要性は高い。

そこで OECD パテントデータタスクフォースなどと連携しながら、欧米の出願人に関する名寄せ作業についても取り組んでいるところである。具体的には、国際出願（パテントファミリー）の情報を用いて、同じ発明が日本特許庁と例えば欧州特許庁の両方に出願されているものを特定する。欧州特許庁のデータにおける出願人の名寄せについては、OECD の作業によって行われているので、同じ発明は同じ出願人によって出願されているという前提のもの、欧州特許の名寄せ情報を日本特許に移送するという手順で行う。

この方法で欧州企業出願人について予備的な作業を行ったところ、IIP パテントデータベースにおいて 7898 件の名称が 4980 件まで名寄せされた。現在、この方法を本格的に活用して、米国特許と欧州特許の情報を IIP パテントデータベースに取り入れることを検討している。（詳細については、Grid Thoma, Kazuyuki Motohashi, and Jun Suzuki, “Consolidating firm portfolios of patents across different offices. A comparison of sectoral distribution of patenting activities in Europe and Japan”, IAM Discussion

Paper Series #019, 2010/11 を参照のこと）

4 その他の課題と今後の方向性

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。

企業の名寄せ以外の問題として大きいのは、特許の分割や延長があった時の特許出願日の取り扱いの問題である。また、PCT 経由で日本に出願された特許については、優先権特許の情報を取り入れることも必要になろう。この点については、整理標準化データにおける上記のような例外的な特許の取り扱いを整理するとともに、必要な情報を抽出すべく IIP パテントデータベース運営委員会でも検討を行っているところである。

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。