知財データの活用・分析を通じた各国施策への貢献

Contribution to government policy makings through utilization and analysis of IP data

経済協力開発機構　科学技術イノベーション局知財アナリスト
名和　大輔

平成 20 年特許庁入庁。生命工学、医療、食品等の特許審査に従事の他、調整課、経済産業省生物化学産業課等を経て、令和 2 年 2 月から現職。

1 はじめに

経済協力開発機構（OECD）は、国際的な政策課題について、調査研究の結果に基づき、経験・ベストプラクティスを共有して意見交換を行う各種会合を通じ、各国の政策立案に貢献している。

筆者が所属する STI/PIE（Directorate for Science, Technology and Innovation / Productivity, Innovation and Entrepreneurship Division）は、生産性、イノベーション、人材育成、雇用、国際貿易等の多岐にわたるテーマについて、各種ミクロデータに基づき経済分析することを主な役割としている。同データには、OECD 加盟国から提供される、匿名化された企業データ、産業連関データ等が含まれ、OECD 独自での国際比較分析を可能とするデータベースを構築している。STI/PIE において筆者は、知財データを専門的に扱うチーム（以下、「IP チーム」という。）に属し、知財関連データベースの構築、同データベースと企業情報データベースとの連結、これらを用いた各種分析等に関わっている。

本稿では、所属部署、特に IP チームにおいて知財データがどのように扱われてきたか、基本的な取組を整理するとともに（2）、そこから派生した新しい取組と言える部分を紹介する（3）。

なお、本稿中の見解等は、筆者の個人的なものであり、組織の見解等を表すものではない点ご了承いただきたい。

2 知財データを用いた基本的な取組

知財データを用いた一般的な分析として、OECD を含む国際機関・政府機関や民間調査会社において、各産業における知財活動の統計的な把握、技術トレンドの把握、それらの国際比較、各種経済指標との相関分析等が行われてきた。OECD での知財関連分析の特徴又は強みとしては、特許では欧州特許庁（EPO）から世界の特許出願情報を収集した PATSTAT データの提供を受けるとともに、商標・意匠でも欧州連合知的財産庁（EUIPO）、日本国特許庁（JPO）及び米国特許商標庁（USPTO）からの提供データを独自に管理して分析に用いている点、並びに知財データベースを OECD で構築されているその他のデータベースと連結して総合的な経済分析が可能である点が挙げられる。

以下に、OECD での基本的な知財関連分析について、筆者が関わる業務内容を絡めて簡単に紹介する。これらは、同じく特許庁から赴任した前々任者によって当時の詳細がまとめられており[1]、提供データの形式や調査に必要とするデータの変化に伴って、現在までに細かく改良が加えられている。

2.1　知財データベースの構築

前述のとおり、特許に関しては EPO が管理する PATSTAT データを元に、商標・意匠に関しては EUIPO、JPO 及び USPTO からの提供データを元に、データベース言語の SQL で操作される、OECD 独自の知財データベースを構築している。各庁から提供される

データは、その形式や構造に違いがあり、それらを統一的にデータロードするため、プログラミング言語のPython や R を用いて前処理を行っている。筆者は特に商標・意匠のデータベース構築・管理を担当している。

2.2　企業情報データベースとの連結

知財データを経済分析で有効に活用するため、企業情報と結びつける必要があるところ、所属チームでは、財務、産業分野、資本関係等の情報が格納された企業情報データベースを用いている。両者に共通する企業 ID は存在せず、さらに知財データにおける企業名は表記揺れがあることから、基本的には文字列の類似度から企業名を照合している。この際、類似度の計算によく使われるレーベンシュタイン距離、ジャロ・ウィンクラー距離等を採用しているが、類似度のみで判断できない場合には、他の情報も含めて人手で判断しており、他のデータ処理と同様、こうしたプロセスが最も時間を割かれる部分といえる。

さらに、上記過程で作られる特許出願人のグルーピング情報は、OECD HAN（Hamonised Applicant Names）database という名称で管理され、PATSTAT にも取り込まれている。

2.3　各種指標を用いた分析

知財データを用いて、国・地域別、産業別、技術分野別等の観点で比較分析を行う場合、出願数や登録数を種々の切り口で算出することが基本である一方で、特許の質を測る指標として、被引用文献数、ファミリー出願数、特許分類から算出される被引用文献の技術範囲、引用されるまでの期間等が OECD での分析において活用されてきた[2、3]。

近年の取組として OECD は、欧州委員会の共同研究センター（JRC）と協力しながら、研究開発費の大きい企業に対象を絞った知財関連の調査研究の結果を、2015 年から隔年で刊行物としてまとめている[4、5、6]。本調査研究は、企業の知財活動（特許、商標及び意匠）を国・地域、出願知財庁、産業、技術、時系列等の切り口で分析するもので、2019 年の刊行物では、AI 関連知財に一つの焦点を当てた上で、企業における AI 技術開発について、時系列推移、国・地域別又は産業別での相違等を明らかにしており、本年は、「持続可能な開発目標」への貢献が各国で求められていることを念頭に、グリーン関連知財に焦点を当てた刊行物を公表する予定である。

3　知財データを用いた新しい取組

データ処理において比較的機械的に数値化できる指標は、再現性が高く、それらが示す意味も明快である場合が多いため、今後も広く中心的に用いられると予想される一方で、知財情報であれば、図表を含む記述内容までデータ分析の対象になり得るところ、より高度な分析手法を用いるニーズは高い。この点については、従前から多くの研究がなされてきたが、自然言語処理を含む、ビッグデータに対する分析手法の開発が大きく加速する中で、知財情報に対してもこれら手法を適用する研究・調査例は、新たな有用な洞察を与え得る意味で注目に値する。

図 1　研究開発費の大きい企業の知財活動に関する刊行物

以下では、知財データを用いた分析手法について、2.で述べたような基本的な取組に対して新たな試みといえる部分を、筆者のOECDでの取組や他組織での調査研究も含めながら紹介する。

3.1　キーワードの時系列分析

特許文献や学術文献から新興技術の動向を調査する際に、あらかじめ体系立てられた分類を用いて出願数や論文数の推移を調査することは有効な手法であるものの、分類が追いついていない技術は分析が難しいことから、目的に応じて収集したコーパスに対する時系列分析により、出現頻度が上昇するキーワードを抽出する手法が一つの選択肢として用いられる。

IPチームはこれまでに、技術開発の将来予測、新興技術と同時に新興フェーズを経て勢いが弱まった技術も含めた検出、及び学術文献と特許文献に記述される新興技術の時間差の解析が、研究開発における予算・資源配分にあたっての重要な指標となるとして、時系列分析に用いられる一つのアルゴリズムを用い、文書中の出現頻度が時間軸で大きく変化するキーワードを抽出する手法を提案している[7]。

また、英国国家統計局内に2017年に設立されたデータサイエンスキャンパスは、特許文献を含むコーパスから、時系列分析によく用いられる状態空間モデル等によって、頻度の変化に関するスコアとともにキーワード又はキーフレーズを抽出する手法を提案し、これをオープンソースツールとして開発プラットフォームのGitHub上で公開している[8]。

3.2　クラスタリング分析

企業レベルでの経済活動を分析する際、各企業に紐づく複数の変数から、目的変数と説明変数を設定し、重回帰分析を行うことは、相関関係に加えて因果関係を見出すための標準的な手法であり、知財データを含むデータセットについても同手法を用い、企業の生産活動と知財活動の関係性等が研究されてきた。一方で、機械学習の種々の手法の発展に伴い、これらを経済分析でも取り入れる研究・調査が多く見られ、ここでは、知財データに対してクラスタリング（機械学習の中で教師なし学習に大別される、データ間の類似度によって対象をグループ分けする手法）を適用した例を挙げる。

2.3に述べたとおりOECDと共同で調査研究を行うJRCは、各企業の国際特許分類別の特許出願数を変数として、企業をクラスタリングによってグループ分けする手法を発表している[9]。同手法では、類似する対象を順番にまとめて樹形図のように表現する階層的クラスタリングや、予め定めた数のクラスターに対照群を分割する非階層的クラスタリングを用いており、特許出願に付与された国際特許分類の傾向から企業を分割することで、異なる産業でも技術的に類似する企業群を抽出している。

EUIPOは、OECDと同様に知財情報と企業情報を紐付けた上で、計量経済分析を行い、少なくとも一つの知財権を有する企業は、そうでない企業と比べて、従業員一人当たりの利益や賃金が高いこと等を示すレポートを公表している[10]。この中で、非階層的クラスタリングによって、特許権、商標権及び意匠権の保有傾向が似ている企業群をグループ化し、群間で利益や従業員数等の企業指数の違いを分析している。

3.3　トピックモデル分析

トピックモデルはテキストマイニングの一つの手法であり、特定の文書群に対して複数のトピックが各文書に一定の確率で潜在的に存在することを仮定した上で、設定したトピック数に応じて、各トピックにおける単語・語句の出現確率、及び各文書におけるトピックの存在確率を計算するものである。同手法は、ウェブ情報から、財務情報、学術文献、特許文献まで、幅広い文書群に対して適用されており、特許文献に関しては、特許分類には表れない観点のトピックを単語・語句とともに抽出すること等が期待される。

IPチームはマックスプランク研究所との共同調査において、AI技術の発展を学術文献、特許文献及びGitHub情報から分析する中で、3.1で述べたキーワードの時系列分析を行うとともに、ソフトウェアに付属されるreadmeファイルにトピックモデル分析を適用し、AI技術の多様な開発分野を導いた[11]。

さらに、筆者が直接関与する取組では、AI関連企業のホームページに表現されている内容にトピックモデル分析を行い、その活動を知財データに絡めて分析している。ウェブスクレイピング等の技術を扱い、ウェブ情報をデータマイニングするサービスを提供する企業と提携

し、AI 関連企業がホームページ上で公開する情報を収集し、2. で紹介した知財データベース及び企業情報データベースと結合した上で、AI 関連企業の動向を総合的に分析している。調査研究中であるため結果のイメージのみを示すと、トピックモデルによって図２のように、トピックが単語の組合せ（縦軸）、及びそれらの存在確率（横軸）とともに抽出される。ここでは、他にもトピックがある中で二つのみ示したが、単語・語句の組合せから、いかなるトピックであるかは人が判断することとなり、トピック１はヘルスケア関連、トピック２はロボット関連であると推測できる。また、ホームページの各記述に対してトピックの存在確率が計算されるため、各企業を最も確度の高いトピックによって図３のように分類することも可能となる。各ドットが一企業を表し、t-SNE という手法で次元を圧縮しており、縦軸横軸に特に意味はなく、ドット間の距離がトピックという観点からの企業間の距離といえる。こうした手法により、例えば知財関係の変数と各トピックとの関連を視覚的に捉えることも可能となる。

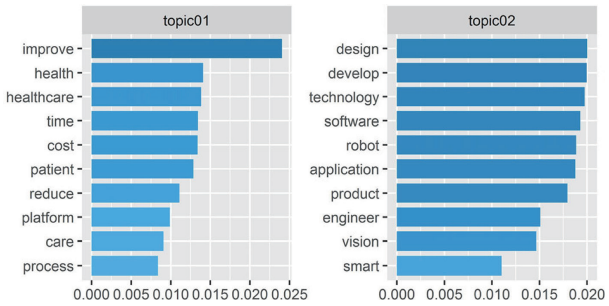

図２　トピックモデルにより抽出された単語群（一例）

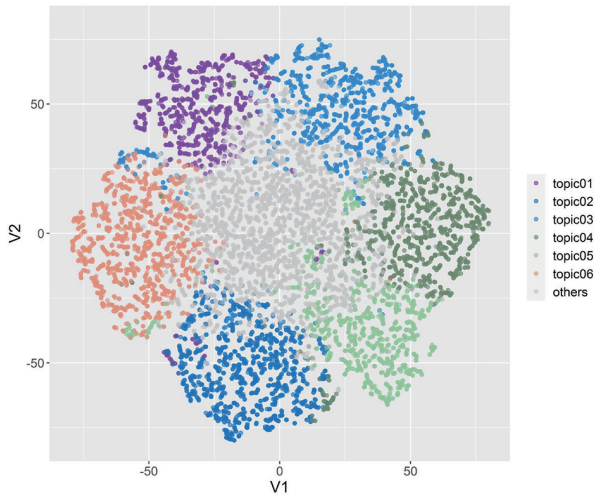

図３　トピックモデルに基づく企業マッピング（一例）

3.4　ネットワーク分析

ネットワーク分析は、データ要素の関係性を表現する手法の一つであり、テキスト、出願国、出願人国籍、発明者国籍、分類、引用・非引用文献等の多彩な知財情報に対しても、同分析を適用した多くの研究例が報告されている[12、13、14]。共起関係からキーワード間の関係を視覚化することに加えて、例えば、引用・被引用文献情報から企業間、出願人間、発明者間等の関係を視覚化している。同分析は、中心性、密度等の指標によって各要素を評価できることにも特徴がある。

IP チームにおいても、AI 関連商標の調査では、対象とする商品・役務の記述に対してネットワーク分析を適用し、AI 関連キーワードと共起する単語や語句の関係性を視覚化した[15]。さらに、OECD の別チームでは、AI 関連の求人情報に対して同手法を適用し、企業が求める多様なスキルの関係性を調査する取組も進められている。

3.5　アソシエーション分析

アソシエーション分析は、マーケティングでよく利用されるデータ分析手法であり、顧客の購買情報から、ある商品が購入される際に一緒に含まれる確率の高い商品を特定する等、商品間の関連性を抽出する目的に用いられている。同手法を特許情報に適用する例として、関連性の高い技術を抽出するために、各出願に付与される国際特許分類の組合せから相関の強いものを同定する調査

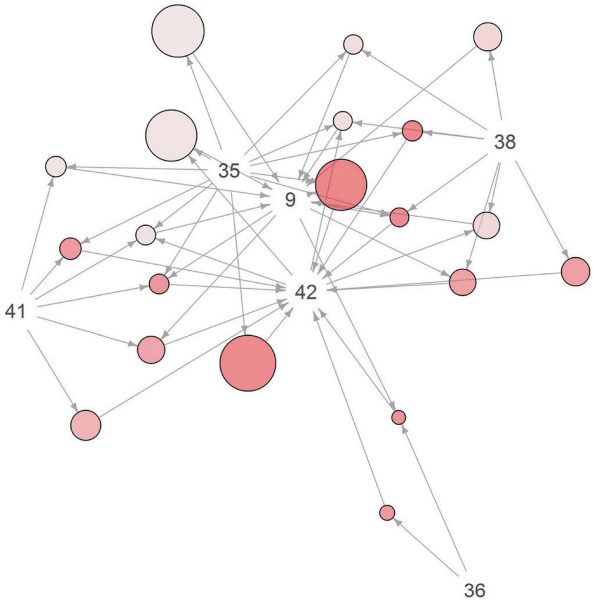

図４　商標分類に対するアソシエーション分析（一例）

研究が発表されている[16]。

筆者は特許分類だけではなく、商標分類に対しても同手法を適用する分析に取り組んでいる。図４は、ある条件で絞った商標出願群に対して、アソシエーション分析を行い、結果を可視化した一例である。数字が商標における国際分類、円が数字間の関係、円の大きさが支持度と呼ばれる指標、円の色の濃さがリフトと呼ばれる指標をそれぞれ表している。例えば、第９類と第42類が他の分類との関係で類似していること、第35類と第42類の組合せが他の組合せに比較して相関が強いこと等がわかる。

2. で述べたように、OECDでは特許に加えて商標・意匠のデータベースを構築している他、これらを企業情報とも結びつけているため、特許分類と商標分類との関連性を抽出し、相関の高い技術と商品・サービスの組合せを同定して、技術開発が市場につながる流れを追うことも検討できる。

4 おわりに

本稿では、知財データ分析に関するIPチームや同チームに属する筆者の取組について、関連する他の調査研究を含めて、基本的な取組から、機械学習等を取り入れた新しい取組までを、主に分析手法の観点からまとめた。（特に後者に関しては、実験的な取組といえる部分もあるため、個人的な見解に基づきまとめられたものであること、改めてご了承いただきたい。）

「エビデンスに基づく政策形成」という考え方が広がり、各国政府や国際機関にとってデータ分析の重要性がより意識されるようになっているところ、知財情報は、特に企業における技術開発や対象市場に関する膨大な情報を含み、政策検討への貢献が期待されている。一方で、高度な分析手法を用いても、目的に沿った手法の選択や結果の解釈の難しさから、各分野の専門家が有する既存の知見を超えない分析に留まってしまうことも多い。また、AI・機械学習を用いた分析手法の発展が著しいといえども、例えば自然言語処理に関しては、文献を人手で読み込むことで得られる理解のレベルには至っていない。これらの点から、具体的なアクションにつなげる分析結果・考察を提供するためには、データ分析の現状の技術レベルを把握した上で、適切な分析手法の選択とその解釈が必要であり、これらスキルを有する人材によって知財データをより有効に活用することが、同データを扱うOECDを含む機関・組織に今後ますます求められる。

参考文献

1 大光太朗、「OECD経済統計課での業務について」、特技懇289号、2018

2 Squicciarni, M., et al., “Measuring Patent Quality: Indicators of Technological and Economic Value”, OECD Science, Technology and Industry Working Papers, 2013, No. 2013/03

3 長部喜幸、治部眞里、「イノベーションシステムの可視化に向けた分析」、特技懇278号、2015

4 Dernis, H., et al., “World Corporate Top R&D Investors: Innovation and IP bundles”, Publications Office of the European Union, 2015, JRC94932

5 Daiko, T., et al., “World Top R&D Investors: Industrial Property Strategies in the Digital Economy”, Publications Office of the European Union, 2017, JRC107015

6 Dernis, H., et al., “World Corporate Top R&D investors: Shaping the Future of Technologies and of AI”, Publications Office of the European Union, 2019, JRC117068

7 Dernis, H., et al., “Detecting the emergence of technologies and the evolution and co-development trajectories in science (DETECTS): ‘burst’ analysis-based approach”, J. Technol. Transf., 2016, 41: 930-960

8 Anthopoulos, T., et al., “An open source tool for discovering emerging technology in large text datasets”, 2020, https://datasciencecampus.ons.gov.uk/projects/pygrams-an-open-source-tool-for-discovering-emerging-terminology-in-large-text-datasets [Accessed: 25-Aug-2021]

9 Gkotsis, P., et al., “A technology-based

classification of firms: Can we learn something looking beyond industry classifications?", Entropy, 2018, 20:887
10 Office for Harmonization in the Internal Market, "Intellectual property rights and firm performance in Europe; an economic analysis", Firm-level analysis report, 2015, https://euipo.europa.eu/ohimportal/documents/11370/80606/Intellectual + property + rights + and + firm + performance + in + Europe [Accessed: 25-Aug-2021]
11 Baruffaldi, S., et al., "Identifying and measuring developments in artificial intelligence: Making the impossible possible", OECD Science, Technology and Industry Working Papers, 2020, No. 2020/05
12 Zhou, X., et al., "A hybrid approach to detecting technological recombination", Scientometrics, 2019, 121: 699-737 based on text mining and patent network analysis
13 Chakraborty, M., et al., "Patent citation network analysis: A perspective from descriptive statistics and ERGMs", PLoS ONE, 2020, 15 (12) : e0241797
14 Gaviria, M. and Kilic, B., "A network analysis of COVID-19 mRNA vaccine patents", Nat. Biotechnol., 2021, 39: 546–548
15 Nakazato, S. and Squicciarini, M., "Artificial intelligence companies, goods and services: A trademark-based analysis", OECD Science, Technology and Industry Working Papers, 2021, No. 2021/06
16 Ampornphan, P. and Tongngam, S., "Exploring Technology Influencers from Patent Data Using Association Rule Mining and Social Network Analysis", Information, 2020, 11 (6) : 333