様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ５ 月 １３ 日現在

機関番号：１４５０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ０３６５９

研究課題名（和文）データサイエンスを利用した特許データの経済分析

研究課題名（英文）Patent Data and Economic Analysis based on Data Science

研究代表者
田中　克幸（Tanaka, Katsuyuki）

神戸大学・経済経営研究所・特命講師

研究者番号：８０４４８１６７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、１）１億件に及ぶPatentデータと5000万件に及ぶ大規模な特許申請者情報をaggregateすることで【類似技術企業検索システム】を構築し、SystematicかつScalableな技術分析を行う新たなフレームワークを構築した。また、２）類似技術企業検索システムで最も技術的に類似する企業を検索することで、どのような企業と技術的に競合をするか分析を行う【Systematicな技術競合分析手法】を確立した。さらに、３）技術競合度合いによってランキングされた検索結果を用いて、企業間の競争度合いを表す指標を作成することで【新たな技術競争分析方法】を確立した。

研究成果の学術的意義や社会的意義
経済分析の分野ではあまり取り入れられていない、Data Science手法の１つである検索技術を、新たな応用経済学の技術・経済分析手法として紹介することで、学術的な貢献を行うことができた。
本研究で構築した類似技術企業検索システムやSystematicかつScalableな技術分析手法は、R&D戦略、M&A戦略、イノヴェーション伝播などのさまざまな分析に応用可能で、企業のみならず国家単位での技術分析への応用が期待できる。技術は経済と密接につながっているため、今後の技術と経済の関係を読み解くうえで包括的かつ客観的な分析手法として貢献できることが期待できる。

研究成果の概要（英文）：In this research, 1) We developed a novel technological analysis framework by constructing the technological competitor retrieval system by aggregating over 100 million patents and about 50 million corresponding patenters, which enable to perform various technological analysis systematically, effectively, and efficiently, 2) Based on the technological competitor retrieval system, we developed a novel and the most comprehensive technological competitor analysis by searching for the most likely technological competitors, 3) we developed a novel approach of measuring technological competition based on the ranking results of competitors searched by technological competitor retrieval system.

研究分野：Data Scienceと経済学の融合

キーワード：技術分析　情報検索　Data Science

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

テクノロジーの発展は、企業の活動を活性化するだけでなく、経済活動の活発化にも影響を及ぼす重要な要素の１つとなっている。近年、それらのテクノロジーを文書化した特許データが注目されて、企業の技術成長やパフォーマンスとの関連性を計測・予測を行ううえで、特許データを用いる重要性が認識され始めている[1,2]。特許に関する文献の引用・参照文献数、特許が依拠する技術分野の広さなどの分析をもとに、特許の技術・経済的価値を測定し、それらを指標として様々な経済分析に用いられている [3]。

ところが従来の研究では、あらかじめ選択された特定の企業や技術分野に関する特許データをもとした限定的な技術分析が多く、Systematic かつ Scalable な技術分析は活発に行われていない状況にある。この原因は、大量で複雑な構造をもつ特許データの扱いにくさと大量なデータの分析を行う技術的な難度が大きく影響していると考えられる。肥大化する特許データを利用した､包括的かつ大規模な経済･技術分析を簡単に行フレームワークと方法論の開拓が必要である。

そこで、検索技術や機械学習といったデータサーエンス由来の技術と特許データに伴う経済・技術分析の融合がこれらの問題を解決し、より実世界を投影した経済・技術分析が可能になるのではないかという着想にいたった。

２．研究の目的

本研究では、大量かつ複雑な構造を持つ特許データと情報検索技術と融合することによって特許データを基にした技術検索システムの構築を行い、情報検索技術に起因した距離の計測、共起関係、相関関係や機械学習、ネットワーク理論などの技術を利用し、従来とは異なるデータサイエンス的なアプローチでさまざまな経済・技術分析を Systematic かつ Scalable に行うことが可能となるフレームワークを構築することを目的としている。

検索技術とは、ユーザが求める情報をインターネット上にある膨大な情報（ドキュメント）の中から正確かつ効率的に発見する技術である。検索技術は Google などで利用されているとおり、大規模のデータを扱うこと関して成熟した技術を有する分野である。インターネット上における爆発的な情報量の増加にともない、検索の分野では BigData が話題になる以前より、膨大な情報の扱いに関して独自の技術発展を遂げてきた。

このようなデータサイエンス的なアプローチを技術分析に用いることで、大量の特許データに基づくさまざまな経済･技術分析を行う事が可能となり、新たな手法によって従来とは異なる分析や新たな知見を得られることが期待される。

３．研究の方法

研究方法は 2 つのフェーズでデータサイエンスを応用した Systematic かつ Scalable な技術分析方法を確立していく。

１）Systematic&Scalable な技術分析を可能にするフレームワークの構築

大規模かつ複雑な構造を有する特許データを用いた分析を、Systematic かつ Scalable に行えるように特許データを整備し、MySQL と検索技術を利用した技術分析用の検索システムの構築を行う。本研究では、技術分析には頻繁に利用されている大規模な特許データである PATSTAT を利用し、以下の手順で特許データの整備を行う。

1. PATSTAT は膨大かつ複雑な構造をもったデータベースである。PATSTAT に収納されているさまざまな情報を取り出しやすいように、特許情報を表す各種 csv データをもとに MySQL で PATSTAT DB を作成する。
2. 作成された PATSTAT DB を利用し、申請された特許と特許申請者を紐づける。

3. 大規模な特許データを用いた Scalable な技術分析を行うため、申請された特許と特許申請者が紐づけされたデータをもとに、類似する技術をもつ企業が簡単かつ Systematic に検索できる技術類似検索システムの構築を行う。検索システムには Java をベースとした Lucene を利用する。

２）Data Science を応用した技術分析の確立

構築した技術類似検索システムと情報検索に起因する技術を利用して、データサイエンス的なアプローチによる以下に示す技術分析を提案することで、新たな方法論の開拓を行う。

1. 言語モデルを利用した企業間の技術距離の新たな測定方法。
2. 技術類似検索を行った結果を用いた、Systematic かつ Scalable な企業技術競合分析方法。
3. 技術類似検索の検索結果の検索ランキングを用いた、企業間の技術競合度合いの新たな指標と分析方法。

４．研究成果

約 1 億件に及ぶ特許データと約 5000 万件におよぶ特許申請者を有する膨大かつ複雑な構造をもった PATSTAT を整備し、情報検索やそれに起因する技術を用いることで、１）Systematic かつ Scalable な技術分析を行うことを可能とするフレームワークを構築するとともに、２）Data Science を応用した新たな技術分析方法の確立を行った。Figure1 は概要を示している。

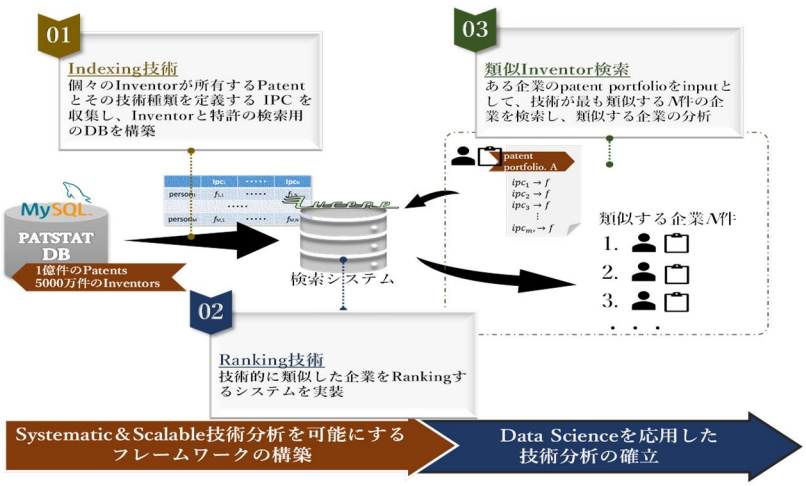

FIGURE 1　データサイエンスを応用した技術分析フレームワーク概要

１）Systematic&Scalable な技術分析を可能にするフレームワークの構築

主に、PATSTAT のデータ整備を行い、技術類似検索システムの構築を行った。具体的には以下のとおりである。

- PATSTAT によって提供されている csv raw データを、より特許情報を取り扱いやすいようにするため、MySQL で DB 化を行い PATSTAT DB を構築した。
- 作成された PATSTAT DB のうち、申請された特許と特許申請者情報を保持する tls201_appln、tls206_person、tls209_appln_ipc などを利用して、企業がある技術分野に申請している特許数を集計し、それぞれの特許申請者がすべての特許と特許分野の紐づけができるよう Person DB を作成した。
- 類似する技術をもつ企業を簡単かつ Systematic に検索することで、特許データを用いた大規模な技術分析が行えるように、Person DB を基に Java をベースとした Lucene によって技術類似検索システムの構築を行った。具体的には PersonDB に収録された約 2500 万件特許申請者（名寄せ後）それぞれに対して：
  1. 申請した全ての特許とそれに紐づく技術分野を抽出。
  2. 検索システムで用いられる、1 つのドキュメントを単語の羅列とその頻度で表す Bag of

words を参考に、特許申請者が所有する全ての特許を約 7 万種類におよぶ技術分野に細分化された International Patent Classification (IPC) の頻度、Bag of IPCs で表現する。これを patent portfolio と呼び、*n* 種類の IPC は$P = (p_1, \cdots, p_n)$と表わせる（p_nは *n* 番目の IPC の頻度を表す)。特許は特許申請者（とくに企業）のこれまでの技術開発の取り組みを表す情報で、Patent portfolio はこれまでの技術開発の歴史、すなわち技術の軌跡を表す重要な表現方法である。

3. 情報検索技術の１つである inverted indexing 方法を用いて、patent portfolio とそれを所有する特許申請者を index 化する。

この結果、1 億件に及ぶ特許データを集積し、2500 万件に及ぶ特許申請者を対象とした技術類似検索システムの構築ができた。検索技術を用いることで、このような大規模データにおいても Systematic かつ Scalable な技術類似検索システムが可能となり、本格的な技術分析を行うフレームワークを構築できた。

２）Data Science を応用した技術分析の確立

構築した技術類似検索システムと情報検索に起因する技術を利用して、データサイエンスなアプローチによる技術分析を行う方法論の開拓を行った。以下に具体的な内容を示す。

- 新たな技術類似距離の計測方法

  特許データの利用方法として、各企業が保有する特許情報を利用した企業間の技術の違い（企業間技術距離）の計測があり、正確な距離測定は，テクノロジーやイノベーション分析における重要な要因の１つである。

  従来の特許データを用いた企業間技術距離(*d*)の測定方法は，企業 *i* からみた *j* の距離と *j* からみた *i* の距離が同じ対称的距離測定である(*d(i,j)=d(j,i)*)。ところが現実には，企業それぞれの見方によって独自の距離感覚があると考えられるので対称的ではなく非対称的距離測定(*d(i,j)≠d(j,i)*)のほうが正確な距離を表していると考えられる。対称・非対称の違いは例えば、2 人のフィジカル距離と愛情・友情距離の違いのようなもので、前者は双方同じだが後者はそれぞれの見方によって異なる。

  本研究では、検索技術の１つである言語モデルを応用し、１）企業 *j* が持つ patent portfolio に対して、２）*j* の技術的特徴を確率分布として定義し、３）この確率分布から *i* の patent portfolio を生成するプロセスにより企業 *i* から *j* の距離を表現し *d(i,j)≠d(j,i)*となるような非対称的な距離測定法を提案した[4]。企業 *j* から *i* の距離を測定するとき、patent portfolio の生成が逆のプロセスとなり対称となる確率分布が異なるため、距離が非対称的となる。

- Systematic かつ Scalable な技術競合（類似）分析手法

  技術的に類似あるいは競合する企業を特定することは、R&D などの企業における技術管理だけでなく、技術や市場スペースにおける企業のポジションを分析し、M&A やパフォーマンスなどさまざまな戦略を練るために重要な役割を担う。

  従来の技術分析では通常、特定の企業を業種などであらかじめ選定し、技術分析を行うのだが、この方法は意図的に競合対象を特定した case-study にすぎない。近年では特に企業の技術やビジネスの多様化がすすんでいるため、分析が不正確・不十分であることが否めない。

  本研究では、構築された技術類似検索システムを用いて Systematic に競合相手の検索を行うことが可能である。具体的には、ある企業の patent portfolio を類似検索システムに入力することで、2500 万件の特許申請者のうち最も技術的に類似する企業、すなわち技術的に競合する相手を検索することができる。これにより多様化した企業の技術力を考慮した真の競合相手を分析することができる。

- **企業間の技術競合度合いの新たな測定方法**
  技術類似検索システムによって検索された競合相手は、競合の度合いによってランキングされている。これは、検索技術の１つであるランキング手法によってランキングを決定しているためである。相手企業を高くランクする・されるほど競合度が高いとみなせるので、競合度合いを知るうえで有用な指標となる。新たな技術競合度合いを測定する方法として、ランキングが高いほど競合が強くなるようにランキングの逆数を競合度合いとみなし、あるターゲット企業間、例えば後述する IT Giants 間、の競合度合いを測定する方法を提案した。
- **IT Giants の技術競合分析**
  本研究では、IT Giants と呼ばれている、Google、Apple、Facebook、Amazon、Microsoft （以後 GAFAM と呼ぶ）の技術分析を上記した方法論を用いて行った[5]**。今や GAFAM は IT だけではなく世界の経済を大きく動かす社会的に重要な企業となっているため、これらの企業がどのような企業と技術的に競合しているのかどのような技術が活発なのか、GAFAM 間ではどのような技術競合状態にあるのか分析することは、経済活動の展望を予測するうえで重要な役割をはたす。この結果、GAFAM 間では当初予想していたような強い技術競合は存在せず、競合を避けるよう技術分野でそれぞれが発展している傾向がみられた。**

３）研究成果の意義と今後の展開

本研究では、約 1 億件に及ぶ特許データと約 5000 万件におよぶ特許申請者を有する膨大かつ複雑な構造をもった PATSTAT を整備し、さまざまな技術分析を Systematic かつ Scalable な行うことを可能とするフレームワークを構築することができた。さらに、情報検索やそれに起因する技術を用いたデータサイエンスを応用することで、従来ではできなかった真の技術競合相手を探しだすことで技術競合分析を行うことが可能となった。

本研究で構築された技術類似度検索システムは汎用性が高いだけでなく、国内外全ての特許データと特許申請者はカバーしているため、さまざまな技術分析分野の研究に応用できる。特に、技術競合分析は，R&D マネッジメント、製品開発、マーケット戦略、M&A など企業におけるさまざまな戦術・戦略に利用可能な重要な役割を果たす。今後、以下のトピックを軸として研究を進めていく予定である．

１）GAFAM だけでなく、半導体産業、製薬産業、自動車産業は経済を動かす重要な役割を果たしているため、これらの技術分析を行う

２）Deep learning の機械学習やネットワーク理論を、特許データや citation 分析に応用したイノベーションやテクノロジーの伝播分析

３）産業単位・国家単位のマクロ的な技術分析

[1] Longhui Zhang, Lei Li, and Tao Li, “Patent Mining: A Survey”, *SIGKDD Explor. Newsl.* 16, 2, pp.1-19, (2015).

[2] Adam B. Jaffe, “Characterizing the ”technological position” of firms, with application to quantifying technological opportunity and research spillovers.”, *Research Policy* 18, 2, pp. 87-97, (1989).

[3] Nicholas Bloom, Mark Schankerman, and John Van Reenen, “Identifying Technology Spillovers and Product Market Rivalry”, *Econometrica* 81, 4, pp. 1347-1393, (2013).

[4] Katsuyuki Tanaka, Takuji Kinkyo, and Shigeyuki Hamori, “Asymmetric technological distance measure based on language model”, Applied Economics Letters, (2019).

[5] Katsuyuki Tanaka and Takashi Kamihigashi, “Measuring Technological Competition among Big Five Using Patent Data: A Systematic and Scalable Approach Based on Information Retrieval Technology”, RIEB Discussion Paper Series DP2021-06, Kobe University, (2021).

5．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

1．著者名 Tanaka Katsuyuki、Kinkyo Takuji、Hamori Shigeyuki	4．巻 -
2．論文標題 Asymmetric technological distance measure based on language model	5．発行年 2019年
3．雑誌名 Applied Economics Letters	6．最初と最後の頁 1～4
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/13504851.2019.1584364	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Katsuyuki Tanaka and Takashi Kamihigashi	4．巻 DP2021-06
2．論文標題 Measuring Technological Competition among Big Five Using Patent Data: A Systematic and Scalable Approach Based on Information Retrieval Technology	5．発行年 2021年
3．雑誌名 RIEB Discussion Paper Series	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関