２Ｃ０６

# アウトカムに基づくイノベーション測定手法の開発
## ―特許および製品データベースを用いた類似度測定―

○原泰史（政策研究大学院大学）[1]
小柴等（科学技術・学術政策研究所）
池内健太（経済産業研究所）

## １．はじめに

イノベーションの生産性評価に係る従来の研究の多くは、特許あるいは論文の書誌情報が用いられてきたが、それらはあくまで研究開発プロセスにおけるアウトプットのひとつであり、イノベーションそのものを指し示す指標ではなかった。本研究では、特許と製品データの接合を目的に、製品データベースの代替としてプレスリリースデータベースを用いることで、企業や大学・研究機関による知の中間生産物である特許が、実際のイノベーションにどの程度結びついたのか識別するための新たな手法を提案する。具体的には、製品の説明文および特許明細の概要に対して、最新の深層学習に基づく手法を適用してベクトル（数値）化した上で、両文書間の意味的な距離について定義することにより、新製品・新サービスといったイノベーションとその技術的な基礎となる発明（特許）との対応関係を推測する手法である。さらに本研究では、提案した手法を実際の日本の特許データおよび新製品データベースとしてのプレスリリースデータに適用する実験を行う。

## ２．先行研究

テキスト情報に基づいて類似性を測定する。Arts et al. (2017) は特許の記事のテキストマッチングを行うことで特許同士の類似度を測定する方法を提案している。特許と製品の関係性を測定する取り組みとしてはまず、医薬品分野においては多くの研究の蓄積がある（Azoulay et al. 2007）。また近年、Ecole polytechnique fédérale de Lausanne（EPFL）の Gaétan de Rassenfoss 氏と David Portabella 氏の「IPRoduct: Linking Patents to Products」プロジェクトでは、米国特許法で 2011 年に施行された Leahy-Smith America Invents Act により認められた、「仮想特許マーキング（Virtual Patent Marking: VPM）」を活用した製品と特許の対応データベースの構築を進めている。VPM とは、実際の製品にはインターネットのウェブサイトへのアドレス情報のみを記載し、ウェブサイト上でその製品に関わる特許番号を記載することによって製品の特許表示を可能とするものである[2]。これにより、特許権が切れた場合に製品のパッケージを変更する必要性がなくなったり、パッケージのデザイン性を損ねることなく特許表示を行うことができたりするなど、特許表示を実施する費用が低下した[3]。インターネット上の情報をクローリングすることで、製品と特許の関係性を把握できるデータベースを構築している。

しかしながら、製品と特許の関係性をプレスリリースの情報を用いて、テキストマッチで行う既存研究は我々の知る限り行われていない。また、Arts et al. (2017) などの既存研究で使われているジャカード係数などを用いた単語の出現頻度に基づく単純な方法ではなく、深層学習に基づく分散表現の獲得手法を用いることにより、テキスト間のより概念的な類似性を考慮している点も本研究の特徴である。

---

[1] yasushi.hara1982@gmail.com, ya-hara@grips.ac.jp

[2]大阪・北区　森本聡特許事務所のブログ「特許の仮想表示（virtual patent marking ）について」（2013 年 4 月 30 日、http://morimotopat.blogspot.jp/2013/04/virtual-patent-marking.html）

[3] The United States Patent and Trademark Office, “Report on Virtual Marking,” (September 2014, https://www.uspto.gov/sites/default/files/aia_implementation/VMreport.pdf)

## ３．分析に用いたデータ

特許データは、出願特許数のトップ 20 の出願人の 1990 年以降の出願特許のうちアブストラクトがある特許と全プレスリリースである[4]。なお、全特許は量も多く，個人の出願人も含まれている．プレスリリース発行のない出願人の特許を扱うことは本研究の目的に合致しない．また，仮に個人出願などを除外しても量が多く，計算コストが膨大になる．そこで今回は出願人（Applicant）テーブルをベースに特許出願数の上位 20 位までの出願人を取得し，これらの出願人の特許を対象とした．また，1990 年以降に出願された特許を対象としている．さらに，これらの特許のうち特許明細の概要データのないものは除外することとした．上記の条件にマッチする出願人ごとの特許件数を表 1 に示す．

**表 1：累計特許出願数トップ 30 の出願人**

| 順位 | 出願人名 | 特許出願数（累計） |
|---|---|---|
| 1 | 松下電器産業株式会社 | 210,143 |
| 2 | キヤノン株式会社 | 202,434 |
| 3 | 株式会社東芝 | 193,733 |
| 4 | 株式会社日立製作所 | 156,407 |
| 5 | ソニー株式会社 | 138,311 |
| 6 | 三菱電機株式会社 | 132,528 |
| 7 | 株式会社リコー | 123,815 |
| 8 | トヨタ自動車株式会社 | 122,721 |
| 9 | 日本電気株式会社 | 122,687 |
| 10 | セイコーエプソン株式会社 | 108,394 |
| 11 | 富士通株式会社 | 101,985 |
| 12 | シャープ株式会社 | 94,838 |
| 13 | 三洋電機株式会社 | 77,248 |
| 14 | 株式会社デンソー | 63,128 |
| 15 | 本田技研工業株式会社 | 61,970 |
| 16 | 富士写真フイルム株式会社 | 58,647 |
| 17 | 日本電信電話株式会社 | 55,945 |
| 18 | パナソニック株式会社 | 55,591 |
| 19 | 三菱重工業株式会社 | 55,356 |
| 20 | 日産自動車株式会社 | 53,449 |

注）共同出願の特許については，各出願人の特許にそれぞれ計上されるため，重複カウントが存在する．

プレスリリースについては、2003 年から 2014 年までに「日経新聞電子版：プレスリリース」のウエブサイト[5]に掲載されたすべてのプレスリリースを対象とする。

## ４．イノベーションに直接結び付いた発明を識別するための新たな手法の提案

**特許・リリース単位の分散表現の算出方法**

あらかじめ比較的多数の文書から名詞単位の分散表現を算出した上で，文書の分散表現を，これら「名詞の分散表現を線形に加算し，正規化したもの」として定義した．算出手順の概要は以下の通りである．

まず、ベースとなる分散表現辞書の準備を行う。辞書のベースとして，日本語版 Wikipedia，KAKEN に登録されて

[4] 特許データについては科学技術振興機構（JST）の「J-Global データベース」より提供を受け、プレスリリースについては日本経済新聞社から購入した。

[5] 「日経新聞電子版：プレスリリース」ウエブサイトの URL：https://www.nikkei.com/pressrelease/

いる研究課題，CiNii に登録されている研究概要などの文書データを収集した．これらの文書データに形態素解析器（mecab，辞書には mecab-ipadic-NEologd を適用）を適用することで，各文書データの名詞句（キーワード）を抽出する．キーワードが抽出された文書データに深層学習に基づく分散表現の獲得手法を適用した。具体的には、本研究では Facebook 社の「fastText」（Bojanowski et al. 2016）を適用し、各キーワードの分散表現（300 次元）を算出した[6, 7]．

次に、上記で得られたキーワード単位の分散表現を用い、特許・リリース単位の分散表現の算出を行う。まず、上記ベースとなる分散表現辞書の作成に用いたのと同様の形態素解析器「mecab」により、文書（特許・リリース）ごとにキーワードを抽出する。次に、キーワードごとに上記で算出した各キーワードの分散表現を割り付けて，文書単位で合算し、正規化したベクトルを文書の分散表現とする．

**特許とプレスリリースの類似度の計算方法**

本研究では類似度の代わりに，分散表現（ベクトル）の距離を用い，距離が近いほど似ているものとして扱う．通常，文書間の類似度算出は，「cos（コサイン）類似度」や「Jaccard 係数」など，キーワードの重複をベースに行われる．これらの手法は有用で，多くの分野で適用されている．ただし，計算機上では「りんご」と「リンゴ」，「林檎」，「アップル」，「Apple」，「APPLE」などは表記が異なるため，すべて全く異なるものとして扱われる．そのため，意味的には類似するキーワードが用いられていても，表記が一致しない場合，類似しないとして扱われる．

他方、分散表現の利点として、分散表現を用いると，「りんご」と「リンゴ」やその他の表記，品種なども含めて，比較的近い位置にマッピングされることが期待でき，この分散表現について cos 類似度を算出すると，単純なキーワードベースの cos 類似度に比べて，意味的な類似度がより反映されやすいと期待できる．

ところで，本論文であつかう特許の数は 200 万件以上，プレスリリースも約 23 万件である．単純に対象特許と，プレスリリースの類似度を計算してしまうと，特許件数とプレスリリース数の積の回数の類似度計算を要することになり，計算コストが膨大となって現実的でない．そこで，今回は Yahoo! Japan 社の開発した高次元ベクトル近傍探索（NGT: Neighborhood Graph and Tree）を用い，空間上で近くに位置するベクトルを近似的かつ高速に検索することで，全組み合わせを探索することなく，高速に類似度が高いと考えられる文書を検索している[8]．具体的には，プレスリリースの分散表現を NGT に登録し，適時，任意の特許の分散表現を NGT に与えて，空間上の距離が近い（類似度が高いと期待される）プレスリリース（厳密にはプレスリリースの ID）を算出している[9]．

また本論文で着目する特許とプレスリリースの関係性を考慮すると，2 つの関連性が高い（分散表現の距離が近い・類似度が高い）としても，プレスリリースは特許出願以降のものである必要がある．一方，近傍ベクトルの取得に際して，事前にプレスリリースの年の絞り込みは困難である．そこで，今回は NGT により近傍 30 件のプレスリリースを取得し，その後で特許公開日とプレスリリース発行日の比較を行っている．近傍 30 件であるため，特許ごとで取得してくるリリースの分散表現の取り得る範囲，ベースとなるプレスリリースの件数が異なることになる[10]．

また，簡易的に出願人とプレス対象のマッチングも行った．リリースの先頭に会社名略称があるので，「松下電器産業株式会社」は「松下」など，出願人名と会社名略称との 1 対 1 の対応表を作ってマッチさせた．なお，現状の方法の場合，特許上の「松下電器産業株式会社」とプレスリリース上の「松下ソリューションズ」は同一の会社としてみなされ

6 https://github.com/facebookresearch/fastText

7 キーワードに対する分散表現の獲得手法としては「doc2vec」手法がよく知られているが、今回用いた「fastText」は分散表現を獲得するのに深層学習を用いているという点など基本的な考え方は「doc2vec」と同様である。しかし、「doc2vec」などと比較し CPU でも高速な処理ができるなどの利点があるため、本研究では「fastText」を用いている。

8 https://github.com/yahoojapan/NGT/releases

9 なお、類似度では 0 が独立，1 が同一で，値が大きいほど似ているが，今回はベクトルの距離をそのまま用いているため，0 に近いほど似ていることになる点に，注意が必要である．

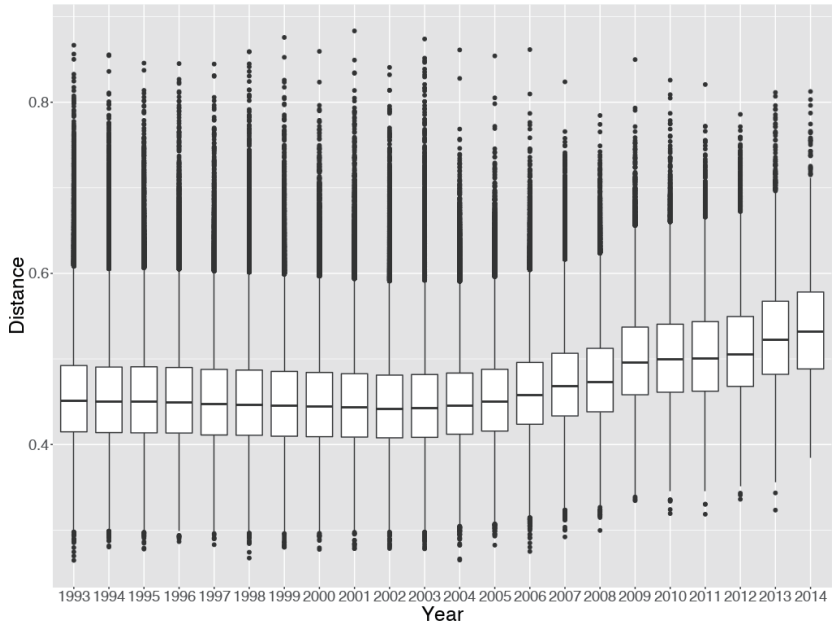
10 例えば，特許 A の近傍 30 件は距離 0.1 以内に入っている一方，特許 B の近傍 30 件は距離 100 以内にひろく分布していることがある．また，30 件のうち，特許公開後に発行されたものが 0 件の場合から 30 件すべての場合までの複数のケースがあり得る．

てしまうことに注意が必要であり、厳密にはリリースの記事全文を用いたマッチングを行う必要がある.

## ５．予備的な分析結果

図 1 では、各特許にマッチしたプレスリリースとの最小距離の分布を特許の出願年別に確認した。1993 年から 2004 年までは、最小距離の中央値は 0.45 であり、全特許のうち 10%はプレスリリースとの最小距離が 0.3 以下、25%の特許は最小距離が 0.4 以下、半分はプレスリリースとの最小距離が 0.45 以下となっている。一方、2004 年以降は時間がたつにつれて最小距離が全体的に上昇している。これは、本研究で用いたプレスリリースデータが 2003 年発行以降に限定されるため、2004 年以降は直近年になるにつれてマッチング対象のプレスリリースの数が減少することを示している。

**図 1：出願年別の特許のプレスリリースとの最小距離の分布**




## ６．まとめと今後の課題

本研究では、イノベーションに直接結び付いた発明の識別を目的に、特許のタイトル及び概要とプレスリリースのタイトル（ヘッドライン）及び本文のテキスト情報を用いて、プレスリリースが意味するイノベーション（新製品・新サービス）の技術的な基礎となった発明（特許）を推測する手法を提案し、実際に日本の 1990 年以降に出願された約 200 万件の特許のタイトル及び概要と 2003 年から 2014 年までに発行された約 23 万件のプレスリリースのデータに適用する実験を行った。

本研究の主な課題は、特許とイノベーションの間の対応関係に関する精度の検証用の正解データを整備した上で、本研究の実験結果を評価することである。また、全てのイノベーションが今回対象としたプレスリリースとして公表されるとは限らないため、プレスリリースデータの特性を踏まえて対象領域を限定する必要がある可能性がある。加え、文書に対する分散表現の獲得手法は本研究の予備的分析で用いた方法の他にも代替手法が提案されている（例えば、doc2vec など）。それらの代替的な手法を用いた場合の結果の頑健性についても検証が必要であろう。

**参考文献**

Arts, S., Cassiman, B., & Gomez, J. C. (2017). Text matching to measure patent similarity. Strategic Management Journal, in press.

Azoulay, P., Ding, W., & Stuart, T. (2007). The determinants of faculty patenting behavior: Demographics or opportunities? Journal of Economic Behavior & Organization, 63(4), pp. 599-623.

Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas (2016) “Enriching Word Vectors with Subword Information,” arXiv preprint arXiv:1607.04606.