特集《人工知能》

特許品質評価及び特許からの情報抽出における自然言語処理のアプローチ


日本アイ･ビー･エム株式会社　研究員・理学博士　**鈴木　祥子**
日本アイ・ビー・エム株式会社　主席研究員・工学博士　**那須川　哲哉**

要　約

大量の特許文書を解析するために，人工知能の一分野である自然言語処理のアプローチを用いた特許分析の手法が近年多く提案されている。特許文書は多くの専門的な情報を含むため，特許分析の目的も手法も多様である。本稿では，数ある特許分析のうち，特許文書の品質評価を自然言語処理と機械学習のアプローチを用いて行う手法を解説する。また，重要性の高い特許の抽出手法について述べる。更に，特許分析を行う上で重要な各種情報を特許文書から抽出する手法をいくつか紹介する。

目次

1．特許分析における自然言語処理

大量の特許文書から各種情報を得るために，人間が使う言葉をコンピュータで扱う自然言語処理のアプローチを用いた分析手法が近年注目されている[(1)]。類似特許検索やパテントマップ作成，特許文書の品質評価などのスコアリング，特許分類といった分析を行うために，従来の人手による作業や簡単なルールの適用に代わる自然言語処理のアプローチが提案されてきた。

一方でSNSや新聞記事といったテキストデータと比較して，特許文書は独自の構造を持ち，また内容が高度に専門的であるため，既存の自然言語処理のアプローチを単純に特許文書に適用してもうまくいくとは限らない。例えば類似特許検索では，通常の類似文書検索手法をそのまま適用するよりも特許文書の構造を利用する手法がよい精度を出すと報告されている[(2)(3)]。また，特許文書は多くの情報を含むため，一概に特許分析といっても目的や手法は多様である。このため，分析目的に応じて特許文書の特性を反映した手法を選択する必要がある。

本稿では，特許の品質評価というタスクに対して，分析の目的を正確に定義した上で，定量的な評価指標を導出する。また，特許の重要性という指標に関連する研究として，重要性の高い特許の自動抽出を試みた結果を紹介する。更に，各種特許分析に利用可能な情報抽出の例を紹介する。

2．特許の品質評価

特許文書をなんらかの指針でスコア付けするというニーズは従来から存在している。例えば自社の保有特許の中から商業的価値の高い特許のみを取り出してポートフォリオを形成する，競合他社が開発した重要技術を検知する，などのニーズは広く存在する。これらのニーズに応じて，人手によるノウハウに基づいたスコアリング手法も多く提案されている（例えば[(4)(5)]）。これらは人間の直感に合う形でデザインされており，広く利用されている。例えば被引用数，国際出願有無などの特徴は，外部からの着目度や出願人自身の評価を反映していると考えられるため，各種スコアリング手法に組み込まれている。ただし，多くの手法ではテキストからの情報を利用していないため，特許の内容に踏み込んだスコアリングをしているとは

言いがたい。

本節では，テキストを自然言語処理により解析し，その上で定量的・客観的な特許の指標を機械学習により構築する手法を紹介する。一般に定量的・客観的な指標を得るには，具体的にどの特許を選択すべきかという正解データを多数集める必要がある。上記で挙げた商業的価値のような指標は，この特許はこの程度の金銭的価値を持つという正解データを得るのが困難であるため，ここでは議論の対象としない。一方で，品質指標には次小節で定義するような正解データを得ることが可能である。

（1） 特許の品質とは

特許の品質評価は，様々な局面で重要となる。明細書作成の評価時や保有特許のポートフォリオ評価時のほか，各国の特許庁において審査の質を上げるためにも定量的で客観的な品質スコアは重要な指針となる。特許の品質の定義として，validity という概念が提案されている[6]。これは，成立した特許が無効審判や裁判で無効とならない安定性を表すスコアである。特に米国では特許に対する裁判が多いため，特許が無効となる・ならないという validity の正解ラベルが得られやすい。一方日本では無効審判や知財高裁裁判のケースが少なく，同様の正解ラベルを得ることは難しい。そのため，validity に代わる概念として patentability という概念が導入された[7]。patentability とは，特許庁で審査された特許が成立するかどうかという成立可能性の度合いを表す。特許庁の審査結果は多数あるため，これを正解ラベルとして利用することができる。より正確には，図 1 のような分類で正解ラベルを抽出する。

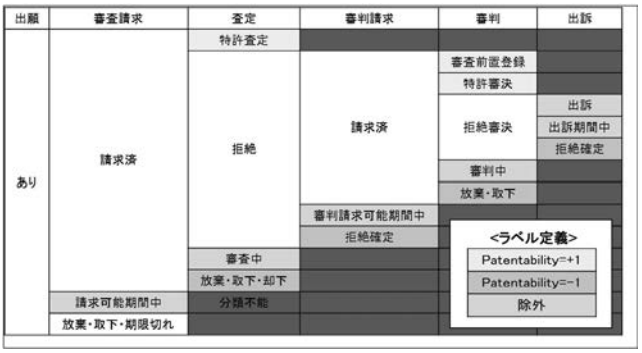

図 1：patentability 正解ラベルの定義

（2） 特許の成立可能性の予測モデル

機械学習のモデルを作成するために，各特許文書から文書を特徴づける以下の（ｉ）から（iv）の各種特徴量を抽出する。

（ｉ） 明細書の統計量

請求項の数や，明細書の文字数，図表の数，優先権主張の有無，付与された IPC の種類数など，テキストの中身に関わらない統計量を特徴量とする。すでに validity のモデルで効果があることが確かめられている。

（ii） 単語年齢の導入

成立特許には新規性・進歩性がある，という特徴を反映するために，単語年齢という特徴量を定義する。単語年齢は，当該明細書の各単語に対して，コーパス（調査対象としている文献集合）中に初めて出現してからの年月を計算したものである。当該特許の全単語の単語年齢をヒストグラム化し，特徴量とする。

（iii） 文書の複雑さ指標

1 文中の文節数や係り受けの平均深さなどを特徴量とする。

（iv） TF-IDF

TF-IDF とは各単語に対して定義される指標であり，通常の自然言語処理でよく用いられる。これは文書中の出現頻度（Term Frequency）である TF と，当該単語を含む文書数の逆数（Inverse Document Frequency）である IDF の積を取ったものである。そのため，単語が当該文書中に多く出ているほど，またコーパス中の出現頻度が低いほど大きな値をとる。すなわち，珍しい単語が特定の文書内に頻出していると TF-IDF が大きくなり，TF-IDF が大きな単語は，その文書を特徴づける語と解釈できる。

これらの特徴量を各文書について数千次元抽出した上で，正解ラベルを再現するような予測モデルを学習する。ある特許文書を入力とし，学習された予測モデルが返すスコアを，その特許の品質と解釈することができる。ランダムに選択した数十万件から学習した結果では，スコアがある程度特許の品質の傾向を表していることが分かっている。

3．重要性の高い特許の候補抽出

2 節では特許の品質のスコアリングのために，正解データを利用し機械学習による予測モデルを構築する手法を解説した。一方で，ビジネス的には競合他社や業界中で重要性の高い特許をいち早く検知したい，というニーズが根強い。2 節で定義した特許の品質スコアは特許の成立可否の可能性を表すものであるから，

技術の重要性とは必ずしも一致しない。既存の特許スコアには，特許の被引用数，不服審判，無効審判請求など，外部からの着目度を組み込んだものが存在するが，これらの指標は特許公開後比較的時間が経たないと見積もることができないため，いち早く検知するという目的にそぐわない。また，多数の正解データ自体を得ることが非常に難しいため，2節のような機械学習によるモデル構築のアプローチも困難である。

既存の特許スコアには，国際出願の有無や早期審査の有無，優先権主張の有無など出願人が力を入れている度合いの指標を特徴量として組み込んでいるものがあり，これらの指標はある程度重要性と関連があると考えられる。しかし重要特許と見なされる特許でもこれらの指標に合致しないことも多く，補完して重要特許を抽出するための手法が求められる。本小節では，これまでの国際出願や早期審査などの指標と異なる観点で重要性の高い特許を抽出する方法を紹介する。

著者らは産業業界における業界内外で重要性が高いと評価されている特許の出願方法を調べた。ここで，重要特許を客観的に判断するために，全国発明表彰[8]で表彰された特許を調査対象とした。調査により，重要特許は，近い出願日で複数の類似特許の固まりとなって出願される可能性が高い，という傾向があることが分かった。この傾向の一つの形態は分割出願であるが，分割出願ではないケースも多々みられる。これらの傾向が見られる理由として，重要発明であると出願人側が認識している場合に，広く権利を取っておく，後のライセンス料を増やす，などのことが考えられる。

図2は，IBM Watson Explorer Advanced Edition Analytical Components(注)[1]で特許のテキスト中の語を品詞別に抽出し，各語が時系列的に過去からどの程度急増しているかの度合いをトレンド分析で見た結果の例である。急増度合いを数値で計算し，急増度合いの高いものから順に並べてある。この例ではある分野において，ある時期に急増した語が抽出されているが，これらの語（この例の場合“気体”，“高強度”，“循環”，“希釈”，“下限”など）が実際に同一の類似特許群に含まれる語であることが確かめられる。更にこの類似特許群は全国発明表彰で受賞した特許に関連するものであることも確かめられる。

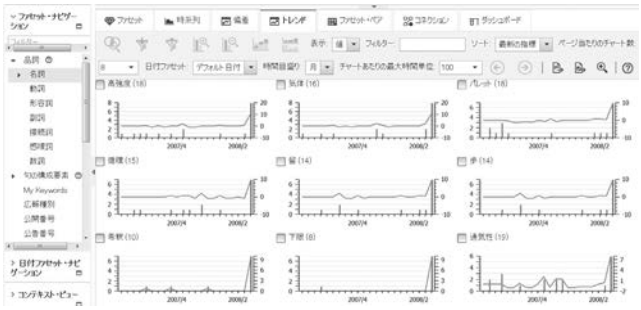

図2：名詞のトレンド分析

重要発明のコアとなる部分は人に紐づいていることが多いため，発明人の粒度での分析も効果的である。発明者のトレンド分析を行うことで各発明者の出願が急増しているものを深堀し，時間的に局在している特許群の内容が類似していることをテキスト解析で判断することで，出願人が力を入れている特許を抽出することが可能になる。

このようなトレンド分析を利用した重要特許の抽出手法は，これまでの国際出願，優先権主張，早期審査の有無といった指標と相補的に利用することが可能である。

4．特許からの各種情報抽出

2節および3節では各特許文書の品質や重要性といった，各特許文書の評価指標をみてきた。その一方で，各文書にどのような内容が記述されているのかという，より詳細な情報抽出を行うことは，幅広い特許分析に役立つ。例えば2節で紹介した各種特徴量は，特許文書からの情報抽出の具体例と言える。本節では，よりテキストの内容に踏み込んだ情報抽出の手法を紹介する。これらはパテントマップの作成や類似特許検索にも役立つと考えられる。

（1） 特長表現の抽出

特許公報には，なんらかの技術的課題を克服し，技術的優位点を持つ発明が記載されていることが多い。このような技術的優位点の特徴の記述表現を，本稿では「特長表現」と呼ぶ。これらの特長表現を特許文書から自動抽出することは，検索やパテントマップの作成に効果的である。例えば自動車に関する発明の特長表現として“振動を抑制する”，“燃費が向上する”のような表現が挙げられる。これらの特長表現は，大きく以下の2クラスに分類され，手がかり表現を用いて精度良く抽出できることが分かっている[9]。

・Improve クラス：好ましい点を更に伸ばすことで特長とする
・Reduce クラス：望ましくない点を抑えることで特長とする
利用する手がかり表現と特長表現の例を表１に示す。

手がかり語	抽出される表現例
Improve クラス	
～[助詞]+向上する	ユーザの使い勝手を向上する
～[助詞]+高める	光の利用効率を高める
～[助詞]+優れる	冷熱サイクル性に優れる
～[動詞]+できる	円滑な空気の流れを確保できる
～[*]+実現する	回路の安定動作を実現する
Reduce クラス	
～[助詞]+防止する	画像の劣化を防止する
～[助詞]+抑制する	変動による影響を抑制する
～[助詞]+低減する	消費電力を低減する
～必要+[助詞]+ない	手作業で試行錯誤的に作成する

表１：特長表現抽出のための手がかり表現と特長表現の例

（２） 特許請求項の構造解析を利用したキーワード抽出

特許文書の中でも，請求項は権利の範囲を記述した最も重要な箇所である。しかし，請求項の記述は通常のテキストと大きく異なる記述方式を用いているため，専門家以外には可読性が非常に低いという問題がある。このため，可読性向上や請求項記述の支援を目的として，請求項独自の構造を解析するという試みがされている[(11)]。また通常の自然言語処理のアプローチにおいても，請求項の構造を利用した方が性能が良いことが主に類似特許検索のタスクで報告されている[(2)(3)]。

これまで行われてきた請求項の構造解析は，主に発明の対象の特定，Jepson 形式などの形式の判別，請求項内の構成要素の分解，従属請求項の依存関係抽出，などが主体であった。またこのような請求項の構造解析を利用して，発明の対象のキーワード抽出や，類似特許検索の枠組みでのキーワード重み付けなどが試みられてきた。

更に著者らは，特許請求項の構造解析を用いて，請求項から発明の新規性や進歩性に関わるキーワード（ここでは新規性に関わるキーワードと呼ぶことにする）を抽出する手法を開発した[(10)]。このような新規性に関わるキーワードは特許に含まれる最も重要な情報の１つであるにも関わらず，これまであまり抽出のアプローチは提案されていなかった。その理由として，請求項の可読性の低さ以外にも，新規性に関わる箇所の抽出が人手でも困難であることが挙げられる。このため多量の正解データを用意することが難しく，客観的な評価が出来なかった。

著者らは，まず独立請求項の構成要素への分解，分解された各構成要素間の依存関係および従属請求項との依存関係抽出を行った。その上で，構造上の特徴を元に，仮説に基づき各キーワードへの重み付けを行った。図３に構造解析の例を示した。赤字のキーワードは，次に説明するフレームワークで得られた正解データである。

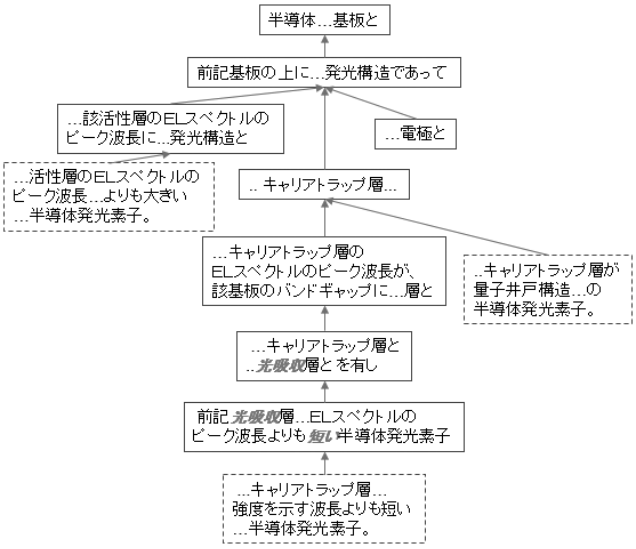

図３：請求項構造解析の例

また，評価用の正解データを大量に収集するためのフレームワークを提案した。このフレームワークでは，審査請求後，一旦新規性・進歩性がないという拒絶理由のみで拒絶され，補正後に最終的に登録されたという経過情報を持つ特許に対し，公開公報と登録特許を比較する。これらの公開公報と登録特許との第一請求項の違いの多くは，明細書から語句を追加することや，元の第一請求項を限定した形で書かれていた従属請求項を第一請求項にすることで成り立っている。そこでこれらの特許の公開公報と登録特許で現れるキーワード集合を比較し，登録特許でのみ出現したキーワード集合があるとすると，これは登録特許の新規性･進歩性に関するキーワード集合，あるいはその一部と解釈できる。つまり，公開公報のままでは新規性･進歩性がないと判断されていたのに対し，これらのキーワードが付加されたことで新規性･進歩性が生じた，と考えることが可能である。このようにして，登録特許に対して，新規性に関するキーワードの近似

的な正解データを大量に収集することが出来る。図4に提案するフレームワーク概要を示す。

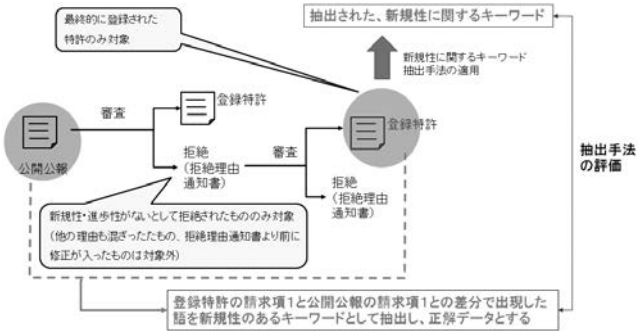

図4：新規性に関するキーワード抽出の評価フレームワーク

新規性に関するキーワード抽出手法を正解データで評価すると，既存の自然言語処理のキーワード抽出手法と比較して大幅に性能が向上していることが分かった。

新規性に関わるキーワードの抽出は，請求項の可読性を高めるだけでなく，類似特許検索や，特許の評価にも役立つと考えられる。

5．まとめ

自然言語処理のアプローチを利用した特許分析の試みは，権利上のデータ利用の手軽さやニーズの高さから，以前から行われている。しかし，特許文書の記述の特殊さ及び各産業分野の専門性の高さが障壁となり，広く使われるには未だ至っていない。

本稿でみてきたように，特許文書にはリッチな情報が存在し，分析目的も多様である。大量の文書から必要な情報を取得するためには，自然言語処理や機械学習のアプローチが有用であるが，目的に応じて適した分析手法も異なってくる。また，客観的な手法の評価，および機械学習での学習のために，大量の正解データを必要とする。このため多様な分析手法を目的に応じて開発することは一朝一夕で出来るとはいえない。しかし，確実に分析手法は進歩しており，新たに得られる情報を組み合わせていくことで，少しずつコンピュータが行えることは広がっている。今後はより人工知能を使った特許分析が盛んになると考える。

※免責

本稿に記載されている内容は，全て，著者個々人の見解に基づいており，著者の所属組織の見解を示すものではありません。

（参考文献）

(1) A. Abbas, L. Zhang and U. S. Khan, “A literature review on the state-of-the-art in patent analysis,” *World Patent Information*, 2014.

(2) T. Takaki, A. Fujii and T. Ishikawa, “Associative Document Retrieval by Query Subtopic Analysis and Its Application to Invalidity Patent Search,” in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, 2004.

(3) F.. Lin and F. Huang, “The Study of Patent Prior Art Retrieval Using Claim Structure and Link Analysis,” in *Pacific Asia Conference on Information Systems*, 2010.

(4) 佐藤 祐介，岩山 真，“特許固有の引用情報を考慮した特許文献の重要度算出方式の検討，” *情報管理*, 2008.

(5) 株式会社パテント・リザルト，“パテントスコアとは，” [Online]. Available: http://www.patentresult.co.jp/about-patentscore.html.

(6) K. Nagata, M. Shima, N. Ono, T. Kuboyama and T. Watanabe, “Empirical Analysis of Japan Patent Quality,” in *Proc of the 18th Internationl Association of Management of Technology (IAMOT)*, 2008.

(7) S. Hido, S. Suzuki, R. Nishiyama, T. Imamichi, R. Takahashi, T. Nasukawa, T. Idé, Y. Kanehira, R. Yohda, T. Ueno, A. Tajima and T. Watanabe, “Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining,” *Journal of Information Processing*, 2012.

(8) “全国発明表彰，” [Online]. Available: http://koueki.jiii.or.jp/hyosho/zenkoku/zenkoku.html.

(9) 西山 莉紗，竹内 広宜，渡辺 日出雄 and 那須川 哲哉，“新技術が持つ特長に注目した技術調査支援ツール，” *人工知能学会論文誌*, 2009.

(10) S. Suzuki and T. Hiromichi, “Extraction of Keywords of Novelties From Patent Claims,” in *International Conference on Computational Linguistics*, 2016.

(11) A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. 2003. "Patent Claim Processing for Readability: Structure Analysis and Term Explanation". Proceedings of the ACL Workshop on Patent Corpus Processing, 2003.

（注）

(1) IBM ® Watson Explorer Advanced Edition Analytical Components V11.0 は International Business Machines Corporation の米国およびその他の国における商標。

（原稿受領 2016. 10. 17）