論　文

機械学習による特許の質の定量評価と統計分析

比戸　将平[†a)]　今道　貴司[†]　鈴木　祥子[†]　高橋　力矢[†]　金平　裕介[††]
葉田 琳樹[††]　田島　玲[†]　上野 剛史[††]　渡部 俊也[†††]

Quantitative Evaluation and Statistical Analysis of Patent Quality using Machine Learning

Shohei HIDO[†a)], Takashi IMAMICHI[†], Shoko SUZUKI[†], Rikiya TAKAHASHI[†], Yusuke KANEHIRA[††], Rinju YOHDA[††], Takeshi UENO[††], Akira TAJIMA[†], and Toshiya WATANABE[†††]

あらまし　特許の質に対する関心が高まる中で，実際の特許データを用いた定量的な質の分析が注目されている．本論文では，成立した特許の安定性を表すバリディティモデルに加えて，特許出願の成立可能性に対応する質の基準であるパテンタビリティモデルを導入する．出願明細書からパテンタビリティの値を予測するための特徴量として，一般的な文書特徴量に加え，単語の新しさや文章の複雑さも考案した．パテンタビリティを予測する機械学習モデルとしてロジスティック回帰モデルを構築し，その性能を評価する．さらに結果を表示し，ユーザーに分析機能を提供するための可視化ツールについて紹介する．

キーワード　特許の質，パテント・クォリティ・インデックス，データ分析，機械学習，可視化ツール

1. まえがき

技術開発を原点とする企業の競争的優位性の源泉として，そして企業の経済活動を支える国のインフラとして，知的財産制度の重要性は年々高まるばかりである．製薬や半導体など競争の激しい高付加価値産業を中心に，企業が自社の重要な技術要素の法的な保護をおこなうことで，他社からの侵害行為を排除し，技術を活用するための知的財産戦略は，事業戦略の重要な要素である．優れた知的財産戦略を実行するためには，より安定で，質の高い知的財産を生み出し，管理することが重要である．さらに最近では，オープンイノベーション戦略の進展に伴って，知的財産を自社で利用するだけでなく，他社にライセンスしたり，プールして活用したり，譲渡を行って活用を図るなど，多様な利用方法が開発されている．このことは知的財産が様々な権利主体に流通され，様々な目的で活用されることを前提とするため，その質の向上が，知的財産を利用する産業界全体にとって，今まで以上に重要になりつつある．

本論文では主要な知的財産として対象を特許に限定する．知的財産の重要化に伴って特許の出願件数も国際的に増加傾向にある．2005 年から 2006 年にかけては全世界で出願件数が 4.8%増加し，特に経済発展著しい中国においては 32.1%の伸びを示している[1]．このような状況が審査期間の長期化など，特許制度における新たな問題を引き起こしている．実際，日本において特許出願に対し審査請求が行われてから審査官が着手するまでの待ち期間は 2005 年度において平均 25.7 ヶ月にもなっていた[2]．2008 年のリーマンショック以降，一時的に出願件数は減少するなど年によりある程度の変動はあるが，過去に比べれば相変わらず出願の絶対数は多い．そのため特許庁では特許出願から審査を短期間で終了させるために審査官の増員など様々な対策を講じており[3]，これらの情勢は欧米においても同様である[4]．一方，特許審査の迅速化は，特許の質を低下させることなくなされなくてはならない．仮に，質の低い特許が登録されてしまったとすると，それ以降の同じ分野での事業展開が制約を受け，技術の進歩の妨げともなりうる．例えば，技術の開示が乏しいにもかかわらず不当に広い権利範囲の特許が登録されると，本来認められるべき第三者の正当な事業活動が制限ないし萎縮する事態が考えられる．そのような制約を受ける事業者側がその特許処分に承服できない場合，無効審判を請求することもできるが，その結果が出るまでにはさらに長い年月がかかるため，目に見えないものも含めて関係者の被る時間的経済的コストは膨大なものとなる．つまり特許審査は，スピードと質の両立という極めて難しいトレードオフについて，限られたリソースの中での対処が求められていると言える[5].

そのような状況の中で，客観的な視点から見て質の高い特許とはどのようなものか，特許の質とは何かについての関心が高まっている[6]．無論，どのような特許が質の高い特許なのか定義は異なり得るが，たとえば，米国特許商標庁による資料においては，出願に対して徹底的な審査がなされたことが出願記録からも明らかで，保護の範囲が適切・明確である，という項目が質の高い特許（quality patent）の要件としてあげられている[7]．特許庁のように審査を担う政府機関は，真に権利を付与するべき発明に対してのみ適切かつ迅速に特許を成立させる特許審査を可能にする特許制度と審査体制を構築することが望まれる．一方，出願者側にも最終的に成立する特許の質を高めることについて果たすべき役割があると考えられる．このように，関係者全員の意識が質の向上に向けられて初めて，質の高い特許を中心とした健全な特許制度が実現されると考えられる．

特許の質への関心を高めるためには，質の高い特許のあり方に関して具体的な形の合意形成が必要であり，知的財産関連学界において近年盛んに研究が行われている．最も中心的なアプローチとしては，特許成立後に特許性が争われた事例の調査がある．何を理由に有効・無効の判断が下され，その原因は権利化の過程においてどのようにして見過ごされたのかを複数の事例において分析することで，特許の質に影響を与える傾向を定性的に把握し，考察を加えることができる．しかしながら，調査対象となったいくつかの事例がそのまま特許の質について全体的な傾向を示しているとは限らない．このような事例調査は出願経過記録の調査など情報の分析を手作業に頼らざるを得ないため，多くの事例を調べ上げることは難しく，肝心の特許の質に関する知見を見逃してしまう可能性がある．そこで，別のアプローチとして，コンピュータを用いた特許情報データ処理による特許の質に対する定量的な分析がある．特許に関する情報を出願明細書や出願経過の情報などを含めて電子データとして収集し，分析の結果得られる質の評価を何らかの数値に帰着させ，全体的な傾向を定量的に把握することが目的である．以前は計算負荷の観点から実現困難だった膨大な特許関連情報の自動処理も，現在では市

[†] IBM 東京基礎研究所，神奈川県
IBM Research – Tokyo, 1623–14, Shimotsuruma, Yamato-shi, Kanagawa, 242–8502 Japan
[††] 日本 IBM 知的財産，神奈川県
Intellectual Property Law Department, IBM Japan, Ltd, 1623–14, Shimotsuruma, Yamato-shi, Kanagawa, 242–8502 Japan
[†††] 東京大学先端科学技術研究センター，東京都
RCAST, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904 Japan
a) E-mail: hido@jp.ibm.com

販のパーソナルコンピュータと同等のシステムでも取り扱えるほどに価格対性能比が向上してきた．本論文においても，人手では扱えない数十万件規模の出願明細書情報を扱うために，コンピュータによる定量分析アプローチを採用する．

以下に本論文の概要を説明する．まず 2 節で特許出願の質の基準として，特許出願が最終的に特許される確率を意味するパテンタビリティを導入する．3 節で関連研究について述べたのち，4 節において実際の特許出願 10 年分のうち，特許処分がなされたものと拒絶処分が確定したものを区別する予測モデルを構築してパテンタビリティモデルとし，実験では特許になるか否かを明細書のみからどの程度正しく予測できるかを示す．最後に，5 節でパテンタビリティ予測を行えるツールを紹介し，考察と PQI プロジェクトの紹介を 6 節 7 節で行う．

2. 特許の質と特許制度

本節では，本論文で着目する特許の質の定義について述べる．

近年，特許ないし特許出願の質への関心が世界的に高まっており，様々な活動や研究が行われている．例えば，知的財産高等裁判所（知財高裁）において権利無効と判断されたケースについて，その理由を分類した調査研究などがある[8]．しかしながら，個々のケースを調べ上げるアプローチでは，定性的な分析を詳細に行える一方で，出願人，代理人，特許庁，裁判所から関連情報を全て集めるための手間や労力が非常に大きいという課題がある．また，扱うことのできる案件数も相当に限定的となりうる．そこで本論文では，コンピュータを用いた定量的分析に着目する．分析の対象とする電子データは，特許電子図書館（IPDL）において公開されている特許公報など，全ての特許出願に関する情報のうちの一部である．また，その他の審査状況などについても特許関連情報サービスとして提供されているものがある．これらの電子データを収集，加工，分析することによって，人手では難しい網羅的な分析を定量的に行うことが可能になる．

特許に関する指標としては，価値の側面と質の側面がある．特定の市場から他社を排除できる特許，多くの実施料収入に繋がる特許などの経済的価値の高いものが価値の高い特許であり，特許権者にとって重要な側面である．一方，登録された特許の安定性の高い特許が質の高い特許であり，特許権者および第三者双方にとって重要な側面である．特許権者は，研究開発から生まれた成果を特許として保護しその一方で事業化にも取り組む中で多大な投資を行うが，一旦成立した特許が実際に権利行使する段階になって無効となる場合にはそれで長年にわたって行ってきた投資が無に帰することにもなりかねない．また，質の低い特許が多く成立する場合には，第三者もその権利が行使されることに対処することが必要となり本来費やす必要のなかった労力をかけることにもなりうる．さらには，特許のライセンス，流通の活動が健全な発展を遂げるためにも，特許の質が高いことは不可欠である．また，技術的内容が詳細に記述され，特許請求の範囲が明確で，先行技術の調査と比較も十分に行われ，新規性，進歩性，明細書の記載要件などの特許性を具備する質の高い特許出願が，かかる質の高い特許を成立させるために重要であると考えられる．本論文では特許の経済的価値ではなく，権利の安定性などの法的な側面における質に絞って議論する．以下に，バリディティとパテンタビリティという 2 種類の質の基準を定義する．

2.1 バリディティ（Validity）

バリディティは成立した特許の安定性を表す．具体的には，特許庁もしくは裁判所において第三者より特許に対する無効の主張を受けた場合に，権利が有効であると判断される確率を，バリディティのスコア（評価値）とする．バリディティについては，これまでも定量的評価の研究が行われてきた[9, 10, 15]．主に，実際の無効審判ケースを集め，それぞれに関する何らかの統計的な特徴量を数値として抽出するアプローチがある．特徴量の数値と，無効審判が請求された特許が有効と判断されるか否かの関連性を検定することにより，統計的見地から特許の安定性とそれに寄与する要因を結びつけられる．あるいは，自然言語処理と呼ばれる文書解析技術を利用して明細書文章の特徴からバリディティのスコアを計算するバリディティスコア予測モデルの研究も行われている[11]．計算されたスコアによって権利の有効性をある程度の精度で言い当てることができれば，そのモデルは特許の安定性に関する何らかの傾向を捉えていると結論付けられる．

しかしながら，バリディティに関しては調査対象となる事例数が少ないという課題がある．その理由は，実際に無効審判や訴訟において特許の有効性が争われるケースは日本ではまれであり，ほとんどの成立した特許に関しては安定性に関わる情報を得ることが困難だからである．実際，上記の研究においてもケース数は数百件の規模に止まっており，成立済み特許のほんの一部に過ぎない．この理由から，全ての特許について有効な質に関する知見をバリディティ評価のみに頼るアプローチには限界がある．そこで，我々は次にパテンタビリティと呼ぶ別の基準を導入し，追加の知見を得ることを試みる．

2.2 パテンタビリティ（Patentability）

本論文ではパテンタビリティと呼ぶ質の基準を新たに導入する．これは特許出願に対して最終的に特許処分がされ特許が成立する可能性（特許成立可能性）に対応している．より正確に言えば，全ての特許出願のうち，審査請求された特許出願が最終的に特許登録される確率に相当する．つまり，スコアが高く質が良いとされる特許出願は，新規性，進歩性，明細書の記載要件など特許要件を満たしているものである．一方，特許請求された発明自体がこれらの条件を十分に満たしていないか，もしくは出願明細書の記述が不十分または不適切であった場合，パテンタビリティスコアの値は低くなるとする．これらの分析では審査請求された特許出願ほぼ全てを対象にできるため，バリディティと比べて事例数は桁違いに多く，数十万件から数百万件を用意できる．コンピュータによる定量評価のために用いる電子データとしては，公開公報に含まれる出願明細書の文字情報を基本とする．出願明細書は，特許庁の審査において最も重要な文書であるため，特許成立可能性を評価するためにも欠かせない情報といえる．また，パテンタビリティ評価の正解データとなる審査状況，つまり審査請求の有無や，現在の査定状況（特許査定か拒絶確定か係属中か）等の出願経過情報もデータの取得は可能である．出願書類からパテンタビリティスコアを計算できるとすれば，実際には出願前の発明に関する作成中の出願明細書についても，パテンタビリティ評価が可能である．つまり，パテンタビリティ予測モデルは既存の特許出願における質の分析だけでなく，出願前の作成中の明細書の質向上にも役立つ可能性がある．一方，パテンタビリティ評価においても審査過程の詳細な情報は有用であると考えられる．例えば，拒絶理由通知書や拒絶査定には審査官の意見が書かれているため，その情報を使えば拒絶査定後に拒絶査定不服審判を請求した場合の特許成立可能性をより正確に予測できると考えられる．しかしながら，詳細な審査過程情報の全件入手は困難であるため，今回のパテンタビリティ評価では明細書のみを対象とする．

海外では小規模な調査研究がある[14]ものの，ある期間の日本国内における特許出願全てを対象に，統計的予測を用いる非常に大規模な特許成立可能性の定量評価は，これまであまり行われてこなかった．これは，事例数に比例してデータ量が増大するため，計算処理に必要とする時間やデータを格納するストレージの大きさの問題等が影響していると考えられる．しかしながら，コンピュータの演算処理プロセッサの性能向上やストレージの容量単価の低下により，市販のパーソナルコンピュータと同様のシステムおいても，これらの大規模なデータ処理を現実的な時間で行うことはすでに可能となっている．

3. 関連研究

特許出願を行う企業，審査する政府機関，及び知的財産関連学会の全てにおいて，また日本，米国，欧州など国の東西を問わず，知的財産の価値や質に対する注目が高まるにつれて，多くの特許制度に関する現状調査や改善方針の発表，提言，及び研究報告が成されている[1, 2, 3].

日本の特許庁においては 2007 年に品質監理室という知的財産の早期権利化と共に特許審査自体の質の維持向上を図る新しい組織を立ち上げている[5]．そこでは第三者による審査内容のサンプルチェックや審判に発展した事例の統計分析，審査済みの一部の案件に対するユーザーの満足度調査などによって現状把握を行い，その結果を審査部にフィードバックして改善に役立つ情報を提供するシステムの導入が試みられている．米国特許商標局（USPTO）でも特許審査の迅速化と品質向上を大きな目標として掲げており，新人審査官に対する新たな研修プログラムを実施するなどしている[4].

一方，産業界からも特許の質に関する提案が発表されている．例えば IBM は，2006 年に特許の質向上を目指すイニシアティブを創設し，USPTO との協力で第三者コミュニティがオンラインで特許審査に貢献できるワークショップ（Open Patent Review，現在は Peer-to-Patent と呼ばれている）や，オープンソースソフトウェアを先行技術として審査で利用することを支援する仕組みを構築する活動，そして，特許の統一的な品質インデックスを作成する取り組みを開始している[12].

アカデミアにおける関連の研究報告としては，特許の価値と企業の競争力の関係について渡部が論じている[13]．ほぼ同じような潜在的な経済的価値を持つ特許を所有していても，その活用戦略が企業によって大きく異なれば，結果としてもたらされる金銭的利益が劇的に変わりうることなど，知的財産活用と企業組織，および事業戦略の間に相互依存が生じていることが示されている．一方，特許の価値と質との関係や違いに関する報告もなされている[6]．代表的な質の低い特許としては，効果の十分な実証が行われないまま，過度に範囲が広く曖昧なクレームのまま成立してしまっているものなどが挙げられる．原因としては，審査において考慮すべき先行技術文献の数が増加していることや，無効かも知れないと自覚しながらも可能な限り広いクレームを入れる出願人や代理人の傾向などがある．実際に世界中で成立している特許の中にも有効性が疑われる特許，無効と判断される特許は存在する．そこで，何らかの方法で特許の安定性をその特許の質として評価し，質の低い特許の成立ができる限りなされない仕組みを目指すことが提案されている．元の文献[6]では専門家による詳細な評価ではなく，データ利用によって特許出願内容の充実度を推定するアプローチが提案されている．実際に知財高裁における有効性判断をベースに事例データを分析したところ，出願明細書の中で効果を示す表現の数や先行技術文献の引用数などが質に与える影響が報告されている[9, 10]．ここで計算されているのは特許の安定性，つまりバリディティ評価に相当する．このようにデータから計算した特許の質に関する情報の活用方法としては，企業側で質の低い特許出願を事前に修正する目的や，特許庁において審査前の段階で明らかに質の悪い特許出願を見分けることなどが挙げられている[6]．同様のアプローチで出願明細書を用いたバリディティの定量評価において．新たな文書特徴量の導入などが研究され，その有用性が国際会議で発表されている[11].

表 2　明細書等軽量に基づく特徴量の例
Table 2　Examples of Specification-based Features

発明名称の文字数
明細書全体の文字数
図表の枚数
請求項の数
請求項全体の文字数
発明の詳細な説明の文字数
国内優先権が主張されているか否か
パリ条約等による優先権主張の有無
定義された IPC 分類項目の種類数
発明者の人数

海外においても，特許成立可能性や質に関する先行研究が存在し，例えば国際的な研究活動や共同出願人による出願の特許成立可能性の調査[14]や，特許引用数が特許の質に与える影響の分析[15]が挙げられる．本論文では，より多くの特徴量と明細書データ集合を用いることで，多様な観点での定量的な知見を提供することを目的としている．

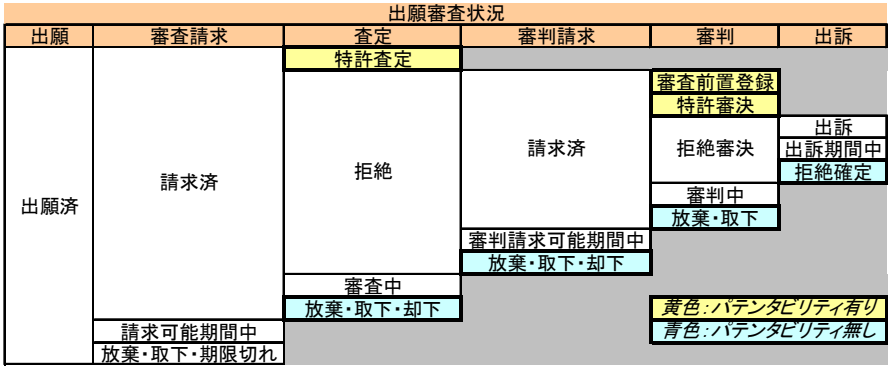
4. パテンタビリティ予測

4.1 明細書データ

パテンタビリティ予測モデルの構築のために，正解ラベル付きデータ集合を公開公報から作成した．構築に用いたのは 1989 年から 1998 年にかけて出願された 10 年分の特許出願である．

まず各データ（出願）が特許登録されるか拒絶されるかというパテンタビリティの有無を表す正解ラベルを取得した．以降，特許成立した出願はパテンタビリティが有ってスコアが 1.0，成立しなかった出願はパテンタビリティ無しでスコアが 0.0 とする．注意すべき点は，正解ラベルを定義できないケースが多く含まれることである．例えば，現在も審査係属中で最終的な判断が下されていないケースに関しては除外する必要がある．また，未だ審査請求が行われていない出願に関しては，特許権利化に対する特許庁の審査も行われておらず，未審査請求のまま確定する可能性もあるため，正解ラベル付きデータ集合には含めない．実際，半数近くの出願は審査請求されないままである．

表 1　出願審査状況に応じたパテンタビリティラベルの定義
Table 1　Definition of Patentability Labels based on Examination Status



したがって，全出願のうち審査もしくは審判が終了して判断が確定しているもののみを抽出し，正解ラベルを与える必要がある．ただし，一般にそのような審査状況の完了状態を直接的に示す情報は記録されていない．そこで，我々は取得した中間記録などを基に，表 1 に示した審査状況分類表を用いて出願の正解ラベルを定義した．パテンタビリティ有りとなっているのは特許査定（審査前置含む）または特許審決となった出願である．一方，パテンタビリティ無しとなるのは拒絶理由通知に対する応答，拒絶査定に対しての拒絶査定不服審判請求，拒絶審決に対する審決取消訴訟提起が無いまま期限が過ぎて拒絶処分が確定したケースと，放棄や取り下げが行われたケースである．データ取得作業の制約から，今回は対象出願期間を日単位でランダムに間引くことで，平均して約 1 割の出願に対する正解ラベルを得た．最終的に我々が得た正解ラベル付き出願は約 30 万件であった．この数は一般の機械学習の問題と比べて大きいが，学習アルゴリズムの適用にかかる計算負荷は許容範囲で，かつ学習結果の安定性という意味でも十分な量と言える．ここで 1999 年以降の出願に関する特許情報データを採用しなかった理由は，出願から審査請求を経て特許査定もしくは拒絶査定が確定するまでには長い審査期間が必要であることから，未だ審査係属中である可能性が高い最近の出願をデータに含めても正解ラベルを定義できる可能性が低いためである．

一方，パテンタビリティ予測の対象とするラベル無し出願データ集合に含めたのは 1993 年から 2007 年にかけて公開された特許出願であり，合計で約 530 万件の出願が存在する．パテンタビリティ予測モデルは，次に述べるように明細書から定義される特徴量を元に，これらパテンタビリティ有り無しのラベルを予測するものである．

4.2 特徴量

予測の根拠となる説明変数の計算方法について，ここでは明細書統計量，構文複雑性，単語年齢，TF-IDF の 4 種類の説明変数群を説明する．

明細書統計量は，明細書の構造や文章から導かれる統計量を出願の特徴とする説明変数である．今回は永田ら[9, 10]によって提案された説明変数群を用いる．表 2 に代表的な明細書統計量の内容をまとめた．その中には請求項の数や，一番目の請求項に含まれる文字数，図面の枚数などがある．これらの明細書統計量は，成立済み特許のバリディティ予測において説明変数として有効であることが先行研究で示されており[9, 10, 11]，パテンタビリティ予測においても効果があると考えられる．

次に，構文複雑性について説明する．これは明細書の個々の文章の意味的な構造の複雑性を扱う統計量である．1 つの日本語文章は，形態素解析によって単語に分割し，係り受け構造解析によってそれら単語間の意味的な係り受け構造を推定することができる．ここでは，両方の機能を持つテキストマイニングツール，IBM Content Analytics®を用いて明細書の主要な部分の文章の構造を解析した．その上でいくつかの統計量を設計し，実際に値を計算して予測モデルの構築に使用する．統計量を決める上で採用した仮説は，熟練した弁理士が代理人として作成した明細書は，特に請求項などにおいて日本語としての明快さよりも論理的厳密性が優先されるため，一文が長く複雑な構文となる傾向があり，その結果構文の複雑性とその明細書のパテンタビリティには正の相関がある，というものである．具体的には，係り受けの最大深さ，1 文中に含まれる文節の最大数，1 文節に含まれる単語の最大数等を，構文複雑性を表す統計量とした．例として，図 1 に係り受け深さ計算の例を示した．明快な日本語であれば係り受け深さはそれほど大きくならない一方，厳密さを要求される請求項の文章では係り受け深さが大きくなる場面が現れることが分かる．よって，これらの構文複雑性を反映した統計量は，パテンタビリティ予測に寄与するものと考えられる．

続いて単語の時間的な新しさを表す単語年齢を導入する．研究開発の世界では技術の進歩に伴い，新しい素材や手法を表す新語が日々生まれ，それぞれに有用性が比較検討され，効果の大きいものは業界に定着していく．その過程では，技術的新規性を必要とする特許出願においてもそれらの新語が現れ始めるタイミングがあり，一般により新しい単語が高いパテンタビリティに結びつくものと考えられる．そこで，各明細書で使われている単語全てに関して，他の出願を含め初めて明細書に現れてからの期間を月単位で計算し，対象の明細書の全単語の年齢をヒストグラムとして求めた．ただし，一般的な常用語に関してもデータ集合に含まれる最初の明細書を基点として年齢が計算されることに注意する必要がある．求められたヒストグラムにおいて，6 ヶ月期間で平滑化した値を 1 ヶ月刻みでずらしながら計算し，最終的な説明変数とした．

最後に，自然言語処理において標準的に用いられるデータ表現形式である TF-IDF を導入する．ある単語がその明細書に出現する回数を，その単語が出現する全明細書の数で割った値が 1 つの TF-IDF 値となる．単純な単語の出現数や頻度に比べて，それほど多くの明細書では使われない語の出現を明細書の特徴として重要視することに相当する．ほとんどの単語に対する TF-IDF 値はゼロになることに注意する．また，単語毎に変数が一つ増えるので，高々1,000 種類の単語の TF-IDF だけでも 1,000 変数となるため，計算負荷を考慮して単語数を調整する必要がある．

正解ラベル付きデータ集合において，上記のような特徴量を明細書毎に計算し，その値の列を説明変数ベクトルとして，モデル構築の入力とする．全ての説明変数は公開公報から計算可能であり，パテンタビリティの予測対象である正解ラベル無しの出願明細書に対して同じ方法で説明変数ベクトルを求め，モデルを適用することでパテンタビリティスコアが得られる．

ただし明細書データは「【請求項１】」等のタグで構造化された文章ながら，記入忘れや形式の誤りによって特徴量の計算ができない場合がある．例えば，発明の名称が誤って空欄となっている出願においては，名称の文字数などの一部の明細書特徴量を正しく計算できない．また，計算できたとしても何らかのエラーの影響により，間違った値である可能性もある．仮に発明者数を数えて 0 人だとしても，正しい値とは言えない．そのため，得られた説明変数の信頼性については別途検証する必要がある．今回の計算においては，全体の 1%を超えるような頻度の高い特徴量計算エラーは見受けられなかったため，そのまま除外して対応した．

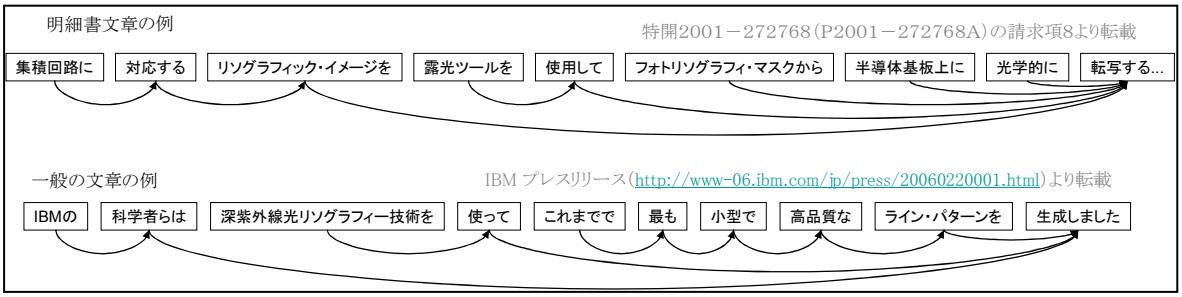


図 1　明細書と一般の文章における係り受け構造とその複雑さの例
Figure 1　Example of dependency structure for specification and general documents

4.3 学習モデル

各明細書に対し，パテンタビリティの正解ラベル y と説明変数ベクトル x の組{y, x}を用意し，それらの集合から予測モデルを学習（構築）する．モデルの適用時には訓練データに現れなかった新たな明細書の説明変数ベクトル x'に対し，ラベル y'を予測する．今回は確率的な判別モデルを導入し，x'がパテンタビリティ有り（y=+1）である可能性を確率の値で評価し，その値が 0.5 を上回ればパテンタビリティ有りと判定する．学習モデルとして，確率的な判別モデルの中でも広く用いられており，安定性と性能が認められているロジスティック回帰モデルを使用する．ロジスティック回帰モデルにおいて，パテンタビリティ有りである確率 p の計算式は次の式 1 ように与えられる．

$$p = \frac{1}{1+\exp(-(w_0 + w_1 x_1 + w_2 x_2 + \cdots))}$$

式 1　ロジスティック回帰モデルによる出力確率
Equation 1　Probability Output of Logistic Regression Model

モデル学習時には，訓練データセットとして与えられたデータのラベル予測誤差が可能な限り小さくなるように，上式の重みベクトル $w=\{w_0, w_1, w_2, \ldots\}$を調節する．さらに，モデルが訓練データ集合に過度に適合して一般性を失う過学習による性能低下を防ぐために，正則化と呼ばれる手法を導入することで，モデルの学習結果を安定化させる．ロジスティック回帰の特長としては，線形モデルであるため，学習結果の解釈が容易な点がある．具体的には，学習結果である重みベクトル w の値の符号と大きさを調べることにより，各説明変数がパテンタビリティ確率の増減どちらに，どの程度寄与しているのかを検証できる．その情報により，個別の明細書に対する予測を与えるだけではなく，全体としての特許審査の傾向分析も可能になる．

4.4 実験結果

パテンタビリティ予測モデルの精度評価実験結果を示す．用いる説明変数群を変更したいくつかの予測モデルとして，明細書統計量のみのもの，それに構文複雑性を加えたもの，単語年齢を加えたもの，TF-IDF を加えたものの 4 種類を用意する．実験は 10 段階交差検定（クロスバリデーション）によって行う．まず正解ラベル付きデータセットをランダムな 10 つのサブセットに分割する．そのうち 1 つを評価セットとして取り置き，残りの 9 セットを訓練データセットとしてモデルを学習させる．そのあと，取りおいた最初の 1 セットにモデルを適用し，予測結果から評価値を計算する．これを

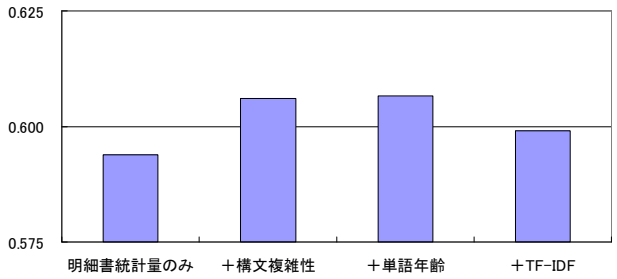


図 2　パテンタビリティ予測モデルの AUC 性能
Figure 2　AUC Performance of Patentability Models

全 10 セットに対して行った結果得られた 10 つの評価値の平均を，最終的な評価値として，4 つのモデルを比較する．モデルの評価値としては，Area Under The ROC Curve（AUC）の値を用いる．この値は，ある 2 つのグループに分かれたデータ集合の個体それぞれに対して何らかのスコアが与えられているとき，そのスコアがグループの判別にどの程度有効かを表す指標であり，この場合のスコアとはモデル出力のパテンタビリティ確率である．つまり，パテンタビリティスコアによってパテンタビリティの有無が正しく見分けられているか否かを評価している値と言える．スコアによって完全にグループが見分けられる場合，AUC の値は 1.0 となる．一方，全くグループの判別に役立たないランダムなスコアリングの場合，AUC は 0.5 である．各モデルの平均 AUC の値を図 2 に示す．明細書統計量のみを使用したモデルでは平均 AUC は約 0.585 である．一方，特徴量として構文複雑性または単語年齢を加えたモデルでは，AUC は 0.605 を超えている．これは，構文複雑性や単語年齢が明細書統計量にはない特徴を捉えており，パテンタビリティ予測に有効であることを示している．一方，TF-IDF の追加による効果はやや小さく，AUC の値は 0.6 を超えていない．

AUC の値が 0.6 前後というのは，予測精度という意味では決して高いとは言えないが，その値が 0.5 よりも有意に大きければ，特許成立確率について何らかの傾向を捉えていると考えられるため，パテンタビリティスコアとして意味を持つと言える．AUC 精度自体の向上については 5 節で今後の課題として述べる．

5. PQI ツール

5.1 概要とシステム構成

このツールは，知的財産の専門家の利用を想定して開発している．目的は，全ての公開された明細書についてパテンタビリティ・バリディティのスコアを計算，比較できるツールを用意

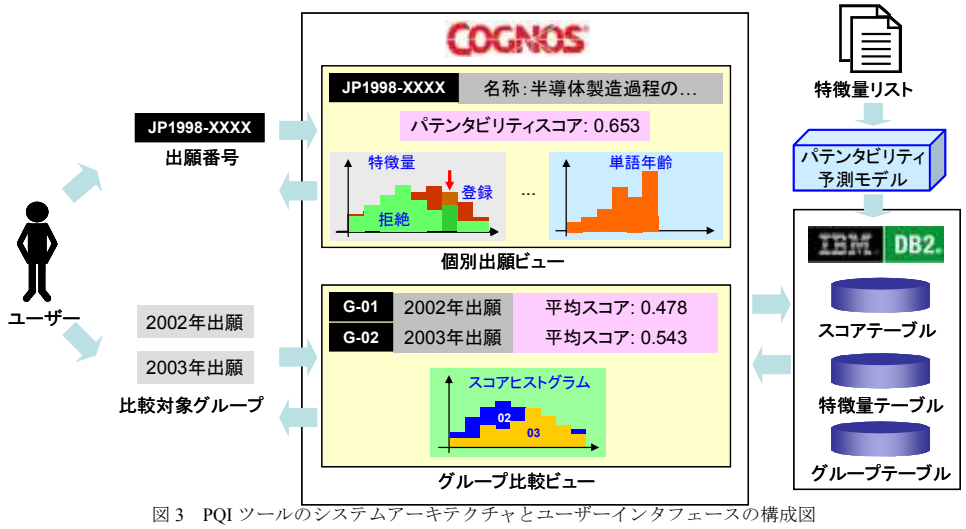


図 3　PQI ツールのシステムアーキテクチャとユーザーインタフェースの構成図
Figure 3　Diagram of System Architecture & User Interface of PQI Tool

することで，知財の専門家へ特許の質に関する知見を提供することである．使い方の例としては，様々な特徴量に関して 4.3 のモデル構築に用いたパテンタビリティ有りデータとパテンタビリティ無しデータではどのような傾向があって，着目している出願の明細書はそれらと比較してどのような位置づけにあり，結果としてパテンタビリティスコアが予測値になったのかを確認することで，その明細書の強みや弱点などを知ることができる．この機能は個別出願ビューとして実装されている．他には，複数の出願を特定の属性を持つグループに分割し，グループ間でパテンタビリティスコアの分布を比較することで，グループの傾向の違いを調べられる．こちらはグループ比較ビューとして実装している．現在は主にこの 2 種類の分析を対象にツールを開発している．スコア計算の対象としたデータは 1993 年から 2007 年までに公開されたほぼ全ての明細書約 530 万件である．それぞれの明細書に対して 4.2 で示した特徴量を計算し，4.3 で述べたように学習したモデルを適用することで，パテンタビリティスコアが求められる．図 3 に PQI ツールのシステム構成とユーザーインタフェースの概要をしめす．サーバー上では必要な情報が全てデータベースに保存されており，ユーザーの入出力は Web ブラウザを経由して行われる．個別出願ビューの利用については，まずユーザーは出願番号を PQI ツールに入力する．すると，該当する明細書に関する特徴量の値やスコアがデータベースから読み込まれ，情報が Web ブラウザ画面を通じてユーザーに提示される．グループ比較ビューの場合は，あらかじめ指定された明細書グループの中から，比較検討したい複数のグループを選択し，実行することで比較結果が得られる．ここで，システム中のデータベースとしては IBM DB2®を，Web 上のユーザーインタフェースとしては IBM Cognos®を使用している．

5.2 個別出願ビュー

まず個別出願ビューについて説明する．図 4 がその実行画面の例である．上部の表にユーザーが指定した出願番号，名称，出願日が表示されている．ここでは匿名性を保つため，出願年以外をダミー情報で上書きしている．その横に審査状況として「前置登録」と示されている．つまりこれは登録特許となった出願に対する出力結果である．その横に 2 種類のモデル（M01 と M05）に対するパテンタビリティスコアの値が表示されている．それぞれ，明細書統計量と構文複雑性を用いたモデル，及びそれを IPC 別に構築したモデルに対応する．ここでは M01 においてスコアが約 0.48 なのに対し，M05 では約 0.87 と，非常にパテンタビリティ評価が高いことがわかる．

左下のグラフは，この明細書の単語年齢ヒストグラムである．横軸が単語年齢で右にいくほど古い単語を，縦軸がその頻度を示している．一般用語も全て今回のデータ期間の最初に始めて出現されたとみなされるため，最も古い単語の頻度が非常に大きくなっていることに注意する．右下の図は，特徴量のヒストグラムを表す．実際には全ての特徴量にヒストグラムがあるが，ここでは構文複雑性特徴量のうち，係り受け深さの最大値が表示されている．黄色の薄いバーはパテンタビリティ有りの出願におけるその特徴量の頻度を示している．一方，青色の濃いバーがパテンタビリティなしの出願における頻度に対応している．これら 2 種類のバー群を比較することで，その特徴量がパテンタビリティの有無判定にどの程度寄与しているのかが視覚的にわかる．バーの分布が大きく異なっていれば，パテンタビリティスコアへの寄与度が高いと言える．一方，バーの分布がほぼ等しければ，パテンタビリティの有無にはあまり関係ないと言える．バーの上にある赤色の下三角（▼）が，対象の出願における特徴量の値を示している．その地点のバーの高さを比較すると，パテンタビリティ有りのバーの方が高くなっており，この特徴量はこの出願のパテンタビリティスコアを増大させる方向に働いていると考えられる．このように，各特徴量が対象の出願のパテンタビリティスコアの増減にどのように影響しているかを把握することができる．

一方，図 5 に示したのは拒絶査定が確定した出願に対する個別出願ビューの例である．審査状況の欄が「拒絶査定（確定）」となっていることに注目されたい．実際，両モデルにおけるスコア値は図 4 に比べ大幅に低くなっている．また，右下図における特徴量の値を比べても最大の係り受け深さの値が小さく，低いパテンタビリティスコアに繋がっていることがわかる．

5.3 グループ比較ビュー

次にグループ比較ビューを説明する．図 6 がその実行画面例である．ここでもある匿名の出願人企業（メーカーA）が IPC 分類 G06F において出願したうち，2002 年の出願明細書群と 2003 年の出願明細書群を比較する 2 つのグループとして採用した．上の表にはそれぞれのグループ 2002，2003 の平均パテンタビリティスコアを表示している．この結果から，グループ 2003 の方が高い平均スコアを持ち，グループ 2002 に比べてパテンタビリティの高い出願を多く含むことがわかる．それらスコアの分布における違いが，下のヒストグラムに示されている．横軸がスコアであり左のバーほど低いスコア，右のバーほど低いスコアの出願を含んでいる．バーの下の数字は，そのバーに含まれる出願の最大スコアを表す．また，縦軸はそれらス

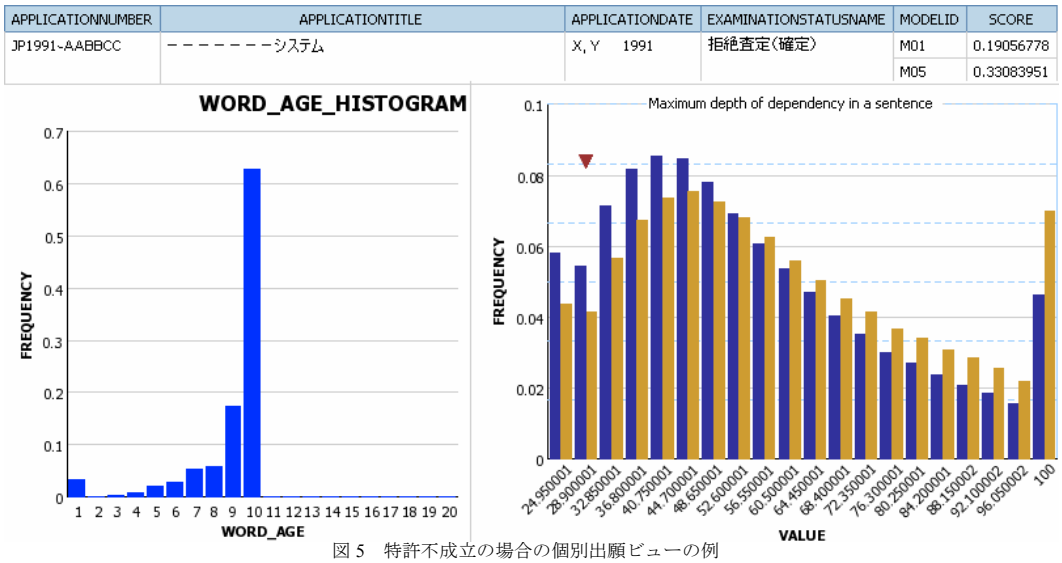


図 5 特許不成立の場合の個別出願ビューの例

Figure 5 Example of Feature Analysis View for Rejected Applications

MODELID	GROUPNAME	average_score
M01	メーカーA (2002,G06F)	0.27719194
M01	メーカーA (2003,G06F)	0.32400821

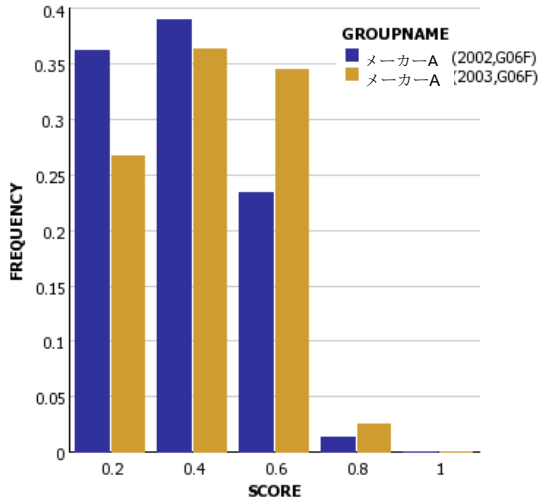


図 6　年度による出願グループ比較ビューの例
Figure 6　Example of Group Comparison View based on Application Year

コアに属する出願の頻度である．このグラフを見ると，低いスコアのバーにおいてはグループ 2002 の頻度が，高いスコアのバーではグループ 2003 の頻度が高い傾向があることがわかる．

このように，PQI ツールにおいて多様な視点で定義された出願グループを比較することによって，様々な知見を得ることができる．グループ比較ビューにおいて，現在はあらかじめ DB 上で定義されたグループの比較しかできないが，将来的にはユーザーが自ら調べたいグループ情報をツールに組み込める機能を追加する予定である．

6. 考察

この節では現状の結果に対する考察と，課題点に対する今後の改善方針について述べる．

まずパテンタビリティ予測モデルについて述べる．4.4 で示したように，パテンタビリティ予測モデルの精度は AUC 基準において 0.6 前後であり，決して高精度とは言えない．そもそも明細書の情報のみから特許成立の可否を推測することが難しいことは明らかなので，AUC が 1.0 近くまで上がることは現実的にありえないが，考えられる改良の余地について以下で紹介する．第一に，今回の実験では別々に用いた構文複雑性や単語年齢などの特徴量を組み合わせることにより，お互いに予測が苦手な部分を補うことで，精度が向上すると思われる．また他のアプローチとしては，対象とする出願の限定がある．例えば対象とする明細書をクラス分けすることで，5.2 節で用いた M05 モデルのように，各クラスに特化した精度の良いモデルを構築する方法が考えられる．IPC 分類は明細書においてすでに定義されているので，IPC 大項目 A から H でクラス分けすることで，8 つの個別モデルを生成できる．使われる単語情報やその時間的変化などは IPC 毎に大きく異なり，それらに関する単語年齢や TF-IDF の特徴量がより有効に働くため，クラス別のパテンタビリティ予測モデルの精度は向上すると思われる．ただし，個々のクラスにおいて十分な量の訓練データ集合が与えられることが条件となる．さらに，明細書以外の情報から追加の特徴量を抽出し，予測に利用することも有効である．例として，審査状況に関するデータが与えられたとして，審査において最初に拒絶査定を受けた出願が審判請求を経て審査前置登録または特許審決を受けて特許として成立するか否かという予測は，元の問題よりも限定された条件下でのパテンタビリティ予測と見なせる．審判請求後の予測であれば，全ての対象出願は一度拒絶査定を受けているため，審査請求から査定を受けるまでの期間の長さは把握できる．またもし拒絶理由通知書を参照することができるとすれば，その内容から審査官の拒絶判断の理由や意図を今回用いたものとは別の特徴量として取り出し，予測に用いることができる．このように新たな特徴量を増やし，またそれらを組み合わせることで，より高い精度のパテンタビリティ予測結果を用いた特許の質の分析が可能になる．

究極的にはパテンタビリティとバリディティは同一の質の評価基準を指すべきである．それはつまり，安定的な特許になり得る出願（バリディティが高いもの）のみが特許として成立する，言い換えれば，一旦成立した特許（パテンタビリティが高いもの）は安定的であるべき，ということを意味する．そういう観点では，パテンタビリティ予測とバリディティ予測の比較も今後の課題となる．例えば，両モデルの特徴量別の重みや出力結果を比べることで，パテンタビリティで重視されるがバリディティではそれほど有効ではない特徴量の発見などが可能となる．その結果をより詳しく分析すれば，特許庁の審査官・審判官，地裁・知財高裁の裁判官の評価基準の差異などが浮かび上がることが期待される．ただし，バリディティとパテンタビリティについて取得済みの特徴量集合の差異がここでの課題となる．これまでバリディティ予測で用いられている事例には手作業で集められた詳細なデータが多く用意されており，予測モデルでもそれらの特徴量が一部用いられている．一方，パテンタビリティ予測においては明細書情報のみしか用いられていないため，特徴量集合の違いからこれらの予測モデルを単純な方法で定量的に比較することは難しい．一つの方針としては，バリディティ予測において特徴量集合を削減し，パテンタビリティ予測と同様に明細書情報のみに限定した場合の結果を検証することがある．バリディティ予測の精度は低下すると考えられるが，パテンタビリティ予測と同じ条件下で平等に比較することが可能になる．また逆に，明細書以外の審査状況に関するデータを全特許出願に関して集め，パテンタビリティ予測モデル用の特徴量集合をバリディティの方に近づける方法も考えられる．さらに，パテンタビリティ予測においては，審査において発見される先行技術により特許性が否定される場合や，出願人毎の特許取得手続遂行における行動パターンの差異をいかに予測において考慮するのか，という問題も今後の検討項目となりうる．

上記のように，パテンタビリティ予測の精度向上やバリディティ予測との比較を考えると，審査状況データは非常に大きな意味を持つ．しかしながら，それらの情報は一般に大量入手が難しい．個別の出願に関して手作業で調べることは可能でも，公開公報のように全出願について審査過程の電子的情報を入手できるサービスは，我々の知る限り存在しない．これらのデータへのアクセスが今後容易になることを期待するものである

一方，すでに入手済みの明細書データの質も，分析における障壁の 1 つとなっている．最近の明細書は全て電子化され，文字列データまたは XML 形式データとして入手できるものの，そこには一定の割合で形式の不備や誤表記が含まれている．例えば発明の名称が空のものや，出願人の項目が抜けているなど，必須の情報が欠けているものがある．これらは出願として成り立たないエラーであり，後日訂正されていると思われるが，公開公報ではそのまま残っている場合がある．大量文書のデータ処理中にこのようなエラーを生じると，その度に元の文章を参照して原因を探り出し，エラーを回避する処理を施して再度実行する必要があり，非常に多くの手間がかかる．その他にも，特許制度や出願方式に変更があったため，ある年を境に明細書の形式が変わっている問題などは，データ分析の研究者がそれらの変更に関して知識を持たないため，どのように対処するか知的財産の専門家のアドバイスを受けながらの対応策の検討が必要である．また，現在は特徴量がほぼ文字情報に限定されている一方で，図表情報が未活用であるという問題もある．明細書に添付されている図に関してはその枚数が明細書特徴量に入っているのみで，図表の内容は一切考慮されていない．また，化学物質を扱う出

願の明細書において化学式が画像として埋め込まれている場合がある．それら化学式画像は文字情報として抽出困難なため，現在の特徴量計算の中では無視されている．このように，明細書を人間が読む場合は問題なくとも，コンピュータで処理する場合に初めて現れる明細書形式の様々な課題がある．今後，特許情報の電子データ利用に対する要求はさらに高まっていくと考えられるため，それに伴ってコンピュータ処理の容易な特許情報公開フォーマットの導入が期待される．

7. PQI プロジェクト

本論文で紹介したパテンタビリティ予測やツール実装を含む活動は，パテント・クォリティ・インデックス(PQI)プロジェクトの日本における活動の一環として進められている[16]．PQI プロジェクトは，特許の質に関する調査研究と指標の策定を目的として開始された．特許実務の豊富な経験を持った知的財産の専門家が中心メンバーであり，現行の特許制度に関する知見を提供している．一方，データを用いた定量評価のために，IBM 研究所の研究者も参加しており，特許情報データにおける特徴量抽出からモデル構築，予測結果出力，ツール実装を担当している．さらに日本の PQI プロジェクトにおいては，文書分析の専門家であるテキストマイニングチームのアドバイスを仰ぎつつ，出願明細書の分析を行っている．国毎に特許法，出願明細書の形式，言語などが異なるため，現在は日米で別々のチームがそれぞれの国の特許についてバリディティやパテンタビリティに関する研究を進めている．アカデミアからは，米国のコロンビア大学ロースクールの Ronald J. Mann 教授，日本からは東京大学の渡部俊也教授と連携し，外部の視点からの意見を頂き，また一部データの提供を受けている．

PQI はこのように多彩なバックグラウンドを持つエキスパートのコラボレーションによって初めて成り立っているプロジェクトである．知的財産の専門家とデータ分析の研究者がチームを組んでプロジェクトを進めることの利点は，お互いに足りないスキルやノウハウを補えることがある．何故なら一般的に考えて，知的財産の専門家が，大量のデータをコンピュータで処理して定量評価を行うことは困難である．同様に，データ解析の研究者は十分な量の特許情報データを与えられたとしても内容に関する専門知識が無いため通り一辺倒な定量評価を行うことになるが，その結果から有用な知見を得られる可能性は低い．そのため，日本の PQI プロジェクトにおいてはデータの意味や分析アプローチについて定期的に確認し議論する場を設けることで，価値ある成果を生み出す努力を続けている．

8. むすび

本論文では特許出願の質をターゲットに，実際の特許情報データを用いた定量的な評価の実現についての取り組みを紹介した．特に，特許出願の特許成立可能性を表すパテンタビリティという質の評価基準を新たに導入し，出願明細書からパテンタビリティのスコアを予測する意義とその方法について述べた．予測に用いる各出願の特徴量としては，従来の明細書統計量や自然言語処理で用いられる一般的な単語特徴量のほかに，明細書に現れる単語の新しさを表す単語年齢ヒストグラムや，文章の複雑さを表す構文複雑性などを考案した．パテンタビリティ予測モデルとしてはロジスティック回帰を採用し，実際の出願データを用いて予測精度を検証した．それら特徴量の値やパテンタビリティ予測の結果を，一般ユーザーが任意の出願について分析できるよう，Web ブラウザ上で操作可能なグラフィカルなツールを開発した．これによって知的財産の専門家が興味ある特許出願についてパテンタビリティスコアと特徴量の関係を調べることや，共通する属性を備える特許出願のグループの比較を行うことが可能となった．

今回は主にパテンタビリティ予測を対象としたが，より良い特許制度の実現を考える上では成立特許の安定性，つまりバリディティが重要であることに変わりは無い．特許の法的安定性という特許の質の観点では，パテンタビリティは特許が成立するまでに着目するものであるが，そこで得られた知見がバリディティの考察において応用しうることも予想される．また今後の検証によって，パテンタビリティの高い出願が特許となった場合には，それが法的にも安定である傾向が示される可能性があり，そうであればパテンタビリティは特許の質にも密接に関連する一指標となりうると考える．

謝辞

本研究の実験において，IBM 東京基礎研究所テキストマイニングチームから日本語文書解析ツールの提供と，貴重な助言を受けた．武田浩一，村上明子，西山莉紗，海野裕也各氏へ感謝の意を表する．また，東京大学の鹿島久嗣准教授，永田健太郎氏，学習院大学の久保山哲二准教授からは実験データと処理プログラムの提供を受けた．合わせて謝意を表したい．

文　献

[1] WIPO，“World Patent Report: A Statistical Review，” http://www.wipo.int/ipstats/en/statistics/patents/wipo_pub_931.html，2008.

[2] 特許庁，“特許審査迅速化の中・長期目標を達成するための平成１８年度実施計画，” http://www.jpo.go.jp/torikumi/zinsoku/h18zinsoku_plan.htm，2006.

[3] 特許審査迅速化・効率化推進本部，“イノベーション促進のための特許審査改革加速プラン 2007（ＡＭＡＲＩプラン 2007）” http://www.jpo.go.jp/torikumi/hiroba/pdf/sinsa_kaosku/siryou_2_1.pdf，January，2007.

[4] USPTO，“Strategic Goal 1: Optimize Patent Quality and Timeliness，” http://www.uspto.gov/web/offices/com/annual/2009/mda_02_02.html，January 2009.

[5] 服部智，“「特許の品質監理」について，”特技懇，no.246，pp.141-147，August 2007.

[6] 渡部俊也，“特許の質の評価　誰のために，何に使うのか，”情報管理，vol.52，no.5，pp.304-307，2009.

[7] USPTO，“Request for Comments on Enhancement in the Quality of Patents，” US Federal Register，vol.74，no.235，http://www.uspto.gov/web/offices/com/sol/og/2010/week01/TOC.htm#ref12，December 2009.

[8] 寒河江孝允，“新制度化における権利無効の抗弁の判決事例一覧表，”パテント，vol.59，no.12，pp.71-72，2006.

[9] 永田健太郎，渡部俊也，“日本特許の質に関する実証分析，”日本知財学会第 6 回年次学術研究発表会，pp.326-329，2008.

[10] K. Nagata，M. Shima，N. Ono，T. Kuboyama，and T. Watanabe，“Empirical Analysis of Japan Patent Quality，” Proc of the 18th International Association of Management of Technology (IAMOT)，April 2009.

[11] H. Kashima，S. Hido，Y. Tsuboi，A. Tajima，T. Ueno，N. Shibata，I. Sakata，and T. Watanabe，“Predictive Modeling of Patent Quality by Using Text Mining，” Proc of the 19th International Association of Management of Technology (IAMOT)，March 2009.

[12] 日本 IBM，プレスリリース，“特許の品質向上を目指すイニシアティブを創設，”http://www-06.ibm.com/jp/press/20060111002.html，January 2006.

[13] 渡部俊也，“特許の価値と企業競争力　イノベーション戦略との関係，”情報管理，vol.52，no.9，pp.562-565，2009.

[14] D. Guellec，B. van Pottelsberghe，“Applications，Grants and the Value of Patent，”Economics Letters，vol.66，pp.109-114，June 2007.

[15] B.N. Sampat，“Determinants of Patent Quality: An Empirical Analysis，” http://siepr.stanford.edu/programs/SST_Seminars/patentquality_new.pdf_1.pdf，2005.

[16] 日本 IBM，プレスリリース，“IBM の研究員が「特許の質」を測るプロジェクトに参画，” http://www-06.ibm.com/jp/press/2009/01/1401.html，January 2009.