AIによる最適アルゴリズム探索技術を搭載した特許自動分類ツール(PatentNoiseFilter)

Patent automatic classification tool “PatentNoiseFilter “ equipped with AI-based optimal algorithm search technology

IRD 国際特許事務所 所長・弁理士/株式会社アイ・アール・ディー
谷川　英和

1986年神戸大学工学部システム工学科卒業。同年、松下電器産業(株)[現パナソニック]に入社し、中央研究所等において、データベース管理システム等の研究開発に従事。1999年弁理士試験合格。2002年1月、IRD国際特許事務所を開設。所長、弁理士。2003～2007年3月京都大学COE研究員、2007年4月～京都大学非常勤講師、2011年4月～大阪大学非常勤講師(現招聘教授)2019年4月～関西学院大学非常勤講師、2020年4月～兵庫県立大学非常勤講師。博士(情報学)。弁理士会、日本知財学会、情報処理学会各会員。2007年度から特許版産業日本語委員会委員。

長岡技術科学大学准教授
野中　尋史

2011年3月　豊橋技術科学大学大学院電子・情報工学専攻修了　博士(工学)。2011年4月　同大学産学官連携研究員。2011年7月　名古屋大学研究員。2012年4月　大分工業高等専門学校情報工学科助教。2014年4月　同校講師。2015年4月　長岡技術科学大学情報・経営システム工学専攻　講師。2018年12月　同大学准教授。現在に至る。この間、特許情報解析、画像情報解析などの研究に携わる。電子情報通信学会(信越支部委員、信越支部庶務幹事を歴任)、知財学会会員。

1 はじめに

我々は、2002年以降、発明の着想から権利化、権利行使に至る特許ライフサイクルにおける各作業について、工学的にアプローチを行う特許工学の研究を行ってきている[1)]。特許工学は、特許ライフサイクルにおける各種作業に対して、方法論を抽出し、ツール(以下、「特許工学ツール」という)と教育により、方法論の普及を図ることにより、各種作業の品質と効率の向上を目指すものである。

また、特許ライフサイクルの上流工程を構成する特許調査作業において、大量の特許データを対象とする大規模特許調査が各企業で頻繁に行われている。また、特許の検索条件を指定しておき、その検索条件に合致する特許情報を定期的にチェックし、必要なデータを収集するSDIも多くの企業で実施されている。

しかし、大規模特許調査やSDI等の多数の特許の調査において、全てを人手で行うことは非効率であることが企業の課題になっている。

一方、ハードウェアの処理能力の向上と機械学習の進歩とにより、使用可能になってきている人工知能(AI)技術を利用した特許調査支援が望まれている。

そこで、我々は、特に、大規模特許調査やSDIにおいて利用できる、AI技術を用いた特許自動分類ツール“PatentNoiseFilter®”(以下、適宜「PNF」と言う)を開発した。

2 PatentNoiseFilter(PNF)の概要

PNFは、人工知能・自然言語処理技術を活用し、特許のユーザによる分類結果である教師データをAIに学習させることで、特許データを自動的に分類できるツールである。

PNFは、ユーザが分類した教師データをAIに学習させ、学習器を取得する学習モジュール(図1参照)と、学習器を用いて特許データの分類予測を出力する予測モジュール(図2参照)とを有する。

3 PatentNoiseFilter(PNF)の特徴

PNFは、特許分類の精度を高めるために、以下に示す種々の技術を有している。

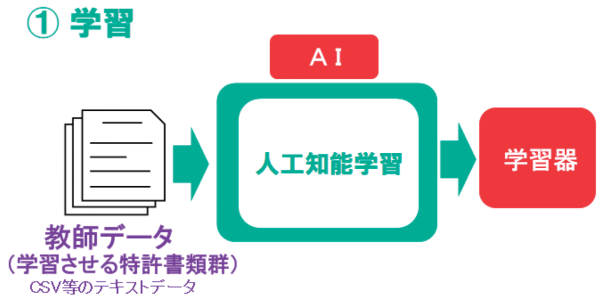

図１　学習モジュールの概念図

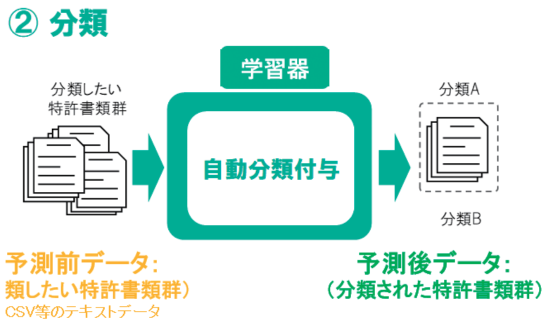

図２　予測モジュールの概念図

3.1　ユーザに安心を与えるための技術

PNF は、教師データを評価する機能を有する。この評価機能は、教師データに対してｋ - 分割交差検証[2] を行うことにより実現する。

PNF では、学習器を作成するために使用しようとする教師データを評価できるために、人手によりさらに特許の分類を行う必要があるか否かを判断することが可能になる。

3.2　精度向上のための各種の技術

（1）３つのモジュールの搭載

PNF は、２種類の深層学習のモジュールと１種類のランダムフォレストのモジュールとを保有し、これら３つのモジュールを適切に用いて、精度の高い分類を行えることを特徴とする。

（2）対象データ選択技術

PNF では、「要約書」「特許請求の範囲」「特許分類コード」「重要情報」等のうちの中からのデータの組み合わせを種々作成し、種々のデータの組み合わせを用いて、学習処理、予測処理を行える。

なお、重要情報とは、明細書から自動抽出した解決手段効果表現文、効果表現文、解決手段用語、効果用語である。

PNF では、手がかり句を用いて、明細書から解決手段効果表現文、効果表現文を抽出する[3]。また、PNF では、特許文書全体の構造情報とその意味関係をグラフで表現したグラフベースの教師なし重要技術語抽出手法を用いて、解決手段用語、効果用語を取得する[4]。

（3）統計処理結果利用技術

PNF では、２以上の PNF の予測結果に対して、「AND」「OR」「多数決」といった統計処理を行い、最終的な予測結果を得ることもできる。

（4）関連語辞書の使用

PNF では、約 30 年分の特許公開公報に対して、自然言語処理を行い、用語の上位概念語、下位概念語、同義語といった関連語の集合からなる関連語辞書を使用している。

つまり、機械学習のモジュールに与えるベクトルを構成する際に、関連語辞書を使用して、ベクトルの次元を調整して、精度向上に役立てている。

なお、関連語辞書を構築する構築機能では、まず、日本の特許公報（公開特許公報、登録特許公報）から手がかり句を用いて用語間の関係性を取得し、関連語辞書を構築する[5]。本機能では、例えば、「などの」と「等の」という表現の前後に記載されている用語は上位概念語、下位概念語の関係にあることに着目する。例えば、「HDD などの記録媒体」などである。この例では、「記録媒体」が「HDD」の上位概念語の関係にある。しかし、「などの」と「等の」だけでは、「パソコンなどのキーボード」のように上位概念語、下位概念語関係にない場合であっても登録を行ってしまう。そのため、「パソコンのキーボード」のように「などの」と「等の」を「の」に言い換えられる場合は、上位概念語、下位概念語の関係にないとした。また、「メモリや HDD などの記録媒体」のように共通の上位概念語、もしくは下位概念語を持つ用語が列記されている場合も上位概念語、下位概念語の関係にないとした。そして、共通の上位概念語と下位概念語を持つ用語は、同義語であるとした。

3.3　ユーザ指向の最適アルゴリズム探索技術

特許の分類において、どのような機械学習のアルゴリズム（例えば、深層学習、ランダムフォレスト）が有効であるか、また同じ機械学習のアルゴリズムでも、どの

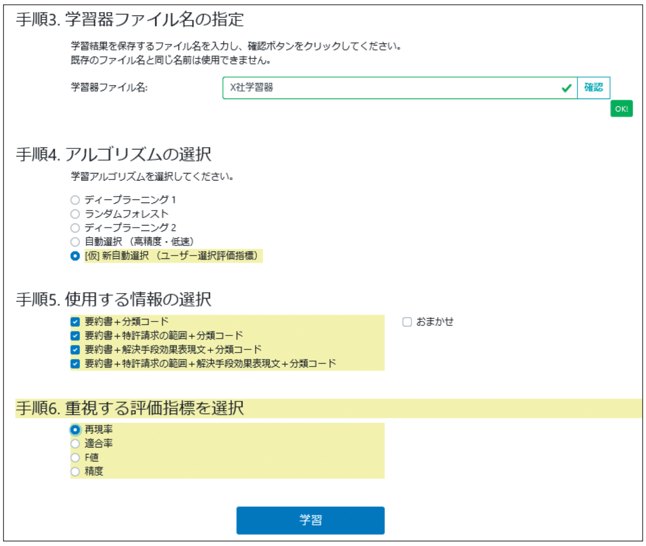

図３　PNF の学習処理の画面

モジュールが良いか、さらに教師データおよび予測データとして、どのような情報の組み合わせ（例えば、要約書、特許請求の範囲、明細書、分類コードなど）を用いれば精度が上がるかを判断することは極めて困難である。

つまり、分類対象の特許の技術分野や技術内容に応じて、適切な機械学習のアルゴリズム、適切な機械学習のモジュール、および適切な使用情報の組み合わせが異なってくる、と考えられる。

また、ユーザが特許自動分類ツールを使用する目的も種々あり得る。つまり、大量の特許情報から技術動向を大雑把に掴みたい場合には、効率的な調査を行うために適合率を上げたいと考え、SDI において、関連特許を決して漏らしたくない場合には、漏れを少なくするために再現率を上げたいと考え、Ｆ値や正解率を上げたいと考える場合もある。

そこで、PNF では、以下に説明するユーザ指向の最適アルゴリズム探索技術を採用する（図３参照）。

つまり、PNF は、ユーザが最も重視する精度を指定でき、指定された精度が最も高いアルゴリズム情報が選択され、このアルゴリズム情報に従った学習器を生成する。

ユーザが指定可能な精度は、「再現率」「適合率」「Ｆ値」「正解率」のうちのいずれかである。

また、アルゴリズム情報は、３種類のうちのモジュールのうちの使用するモジュール名、学習および予測で使用する情報の組み合わせに関する情報、統計処理の有無及び統計処理の内容を特定する情報である。

また、アルゴリズム探索とは、最も精度が高くなるアルゴリズム情報を決定することである。

つまり、PNF では、機械学習の３つのアルゴリズム、使用する情報の複数の組み合わせ、統計処理（「AND」「OR」「多数決」）の利用の有無といった３観点を組み合わせて多数のアルゴリズム情報の候補を自動作成し、各アルゴリズム情報に従った教師データを作成し、各教師データを、上述した評価技術により、ユーザが指定した精度（「再現率」「適合率」「Ｆ値」「正解率」のうちのいずれか）を評価することにより、ユーザが必要な精度に対して、最も高い精度を有する学習器を自動的に取得できる。

このような最適な学習器を用いて、予測処理を行うことにより、特許分類の予測精度が向上する。

3.4　PNF の出力例

以下の図４は、PNF の出力例である。

図４に示すように、PNF では、３つのアルゴリズムと種々の情報（要約書、特許請求の範囲、分類コード、重要情報等）と統計処理の有無との多数の組み合わせを構成し、各組み合わせごとにユーザが指定した精度（「再現率」「適合率」「Ｆ値」「正解率」）を算出し、提示できる。

また、図４に示すように、使用するモジュールにより各精度にかなりのばらつきがある。同じ特許のリストを対象として評価しているのであるが、正解率は 0.772 から 0.983 までのばらつきがあり、適合率は 0.982 から 0.998 までのばらつきがあり、再現率は 0.772 から１までのばらつきがあり、Ｆ値は 0.87 から 0.992 までのばらつきがあった。

アルゴリズム自動選択結果

以下の処理結果から、「3つのアルゴリズムのOR」（使用する情報は「要約書, 特許請求の範囲, 分類コード, 解決手段効果表現文」）が選択されました。

重視する評価指標は「再現率」です。

アルゴリズム	使用する情報	Average accuracy (精度)	Average precision (適合率)	Average recall (再現率)	Average F1 score (F値)
アルゴリズム2（ランダムフォレスト）	要約書, 分類コード	0.843	0.997	0.843	0.913
アルゴリズム1（ディープラーニング 1）	要約書, 分類コード	0.982	0.983	0.999	0.991
アルゴリズム3（ディープラーニング 2）	要約書, 分類コード	0.976	0.983	0.993	0.988
3つのアルゴリズムの多数決	要約書, 分類コード	0.979	0.984	0.995	0.989
3つのアルゴリズムのAND	要約書, 分類コード	0.841	0.998	0.84	0.912
3つのアルゴリズムのOR	要約書, 分類コード	0.981	0.982	0.999	0.991
アルゴリズム2（ランダムフォレスト）	要約書, 特許請求の範囲, 分類コード	0.799	0.996	0.799	0.885
アルゴリズム1（ディープラーニング 1）	要約書, 特許請求の範囲, 分類コード	0.977	0.983	0.994	0.988
アルゴリズム3（ディープラーニング 2）	要約書, 特許請求の範囲, 分類コード	0.98	0.984	0.996	0.99
3つのアルゴリズムの多数決	要約書, 特許請求の範囲, 分類コード	0.979	0.984	0.995	0.99
3つのアルゴリズムのAND	要約書, 特許請求の範囲, 分類コード	0.795	0.996	0.795	0.883
3つのアルゴリズムのOR	要約書, 特許請求の範囲, 分類コード	0.981	0.983	0.999	0.991

図４　PNF の出力例

つまり、図４によれば、一般的に、機械学習を用いた特許分類システムでは、分類した特許の分野や内容により、どのような情報を使用して、どのアルゴリズムを採用するのが最適であるかを、人が判断することは極めて難しいことが分かる。

4 まとめ

ユーザ指向の最適アルゴリズム探索技術を有する特許自動分類ツール（PatentNoiseFilter）について説明した。

このような技術を有する PNF は、ユーザが分類したい特許データに適合し、かつユーザが高めたい精度を高めることを示せた。

今後、特許分類の精度をさらに上げるために、機械学習の各モジュールの改善、第４のモジュールの導入等を行っていきたい。

参考文献

1）谷川英和他，特許工学入門，p1 ～ p7（2003），中央経済社

2）Kohavi, Ron（1995）. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2（12）: 1137–1143.（Morgan Kaufmann, San Mateo）

3）坂地泰紀他：Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法，電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp.742–755（2010）

4）邊土名朝飛他：特許構造を考慮したグラフベース教師なし重要技術語抽出，人工知能学会　全国大会（2020）

5）難波英嗣他，「特許、論文データベースを統合した検索環境および動向分析ツールの構築」，2006 年度 Japio「特許情報活用の時代の検索と機械翻訳技術」，pp116-119，2006.