AIを用いた特許自動分類ツール（PatentNoiseFilter）のSDI調査への応用

Application to SDI of a atent automatic classification tool “PatentNoiseFilter” using AI.

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㈱［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。1997 年同社知的財産権部門に異動。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、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 年度から産業日本語研究会特許文書分科会委員。

1 はじめに

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 調査が多くの企業で実施されている。なお、特許情報とは、例えば、公開特許公報、登録特許公報である。

一方、我々は、AI 技術を用いた特許自動分類ツールである“PatentNoiseFilter®”（以下、適宜「PNF」と言う）を、既に開発している。

そこで、AI 技術を用いた PNF を、SDI 調査に効果的に利用するための機能について考察したので、報告する。

2 PatentNoiseFilter（PNF）の概要

PNF は、人工知能・自然言語処理技術を活用し、特許のリスト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。

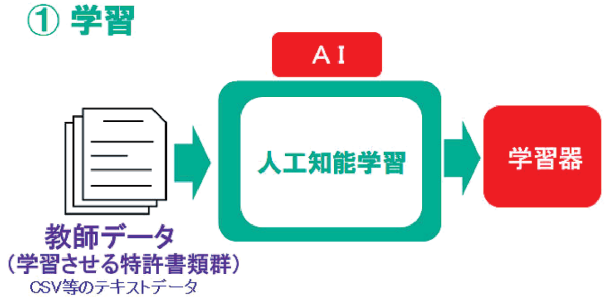

図 1　学習モジュールの概念図

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。

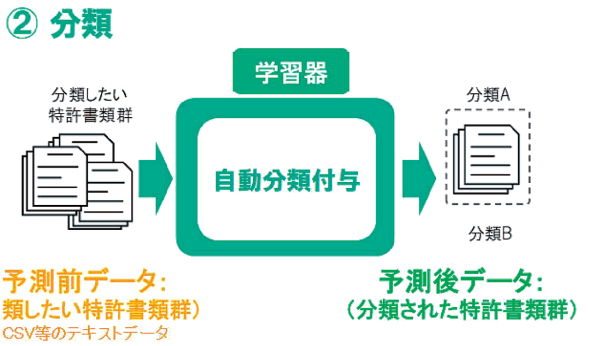

図 2　予測モジュールの概念図

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。

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 [1] 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[2)]。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[3)]。

（3）統計処理結果利用技術

PNF では、異なるアルゴリズムまたは異なるデータを用いた２以上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。

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）を用いれば分類の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 調査において、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、また 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）。

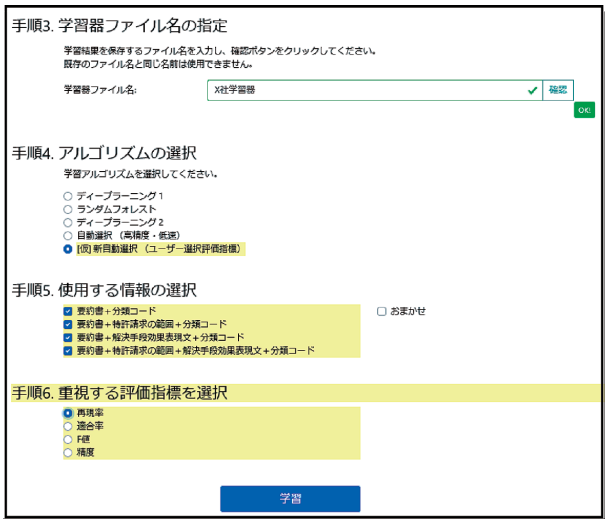

図３　PNF の学習処理の画面

つまり、PNF は、ユーザが最も重視する指標を指定でき、指定された指標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指標は、「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情報の組み合わせ、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、ユーザが指定した指標（「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、ユーザが必要な指標に対して、最も高い精度を有する学習器（以下、最適学

習器）を自動的に取得できる。

このような最適学習器を用いて、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。

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

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指標（「再現率」「適合率」「F 値」「正解率」）の値を算出し、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。

つまり、PNF では、使用するモジュールや使用情報等の多数の組み合わせを用いて学習器を構築してみて、評価した結果、最も精度の高いものを最適学習器として自動選択する。

4　SDI 調査における PNF の利用

4.1　最適学習器を用いた SDI 調査

SDI 調査におけるユーザ条件ごとに、ユーザによる過去の分類結果（教師データ）を PNF に与えて、学習処理を実行することにより、上述したように、ユーザ条件に対応する最適な学習器を得ることができる。なお、ユーザによる過去の分類結果は、例えば、関連特許（○）と非関連特許（×）とに分類されている。

そして、SDI 調査を支援する SDI システムにおいて、ユーザ条件に対応付いて、学習器が管理される。

次に、SDI システムにおいて、毎週発行される公開特許公報または登録公報に対して、ユーザ条件に基づいて、多数の公開特許公報等を自動抽出する。

次に、PNF の予測モジュールに、ユーザ条件と対になる学習器と自動抽出した多数の公開特許公報等とを与え、PNF の予測モジュールを実行する。すると、PNF の予測モジュールは、各公開特許公報等に対する分類結果（「○」または「×」）とスコアとを対応付けた CSV ファイルを出力する。

次に、SDI システムにおいて、分類結果およびスコアをキーとして降順にソートした CVS ファイルを構成し、蓄積する。

このようにすることに、毎週発行される公開特許公報等をチェックする必要がある研究者、技術者、または知財担当者は、AI が重要であると判断した特許情報から順番に内容を確認でき、また、スコアの高い（信頼性の高い）特許情報の確認が不要となり、SDI 調査の効率が大幅に向上する。

4.2　学習候補決定機能

過去に人手により分類された特許情報が検索された条件と同じ検索条件に合致する特許情報でも、時代の進展、技術の進歩により、過去に人手により分類された特許情報とは、かなり異なる技術的内容を有する情報となってくる場合も多い。しかし、このような状況で、同じ学習器を継続して使用していると、優秀な AI 技術によっても、分類精度が低下することは明らかである。

そこで、SDI 調査において、PNF を利用して自動分類し、かつスコアを付与するだけでは無く、新たに公開された特許情報の中から、人手により分類し、PNF に

学習させるべき特許（学習候補）を、ユーザに提示する。

また、PNF において、学習候補は、過去に人手により分類された各特許情報の平均ベクトルとの距離が閾値以上であるベクトルに対応する特許情報を学習候補として出力する。なお、ベクトルを取得する対象は、特許情報の全てではなく、PNF の最適学習器の構築に使用された情報だけであることが好適である。

4.3　学習器自動更新機能

PNF の大きな特徴として学習器（教師データ）を評価する機能が挙げられる。また、学習器を構築する際の教師データの数が多いほど、学習器の精度が向上することが多い。

そこで、さらに人手により分類された特許情報の数が閾値以上（例えば、10 以上）になった場合、または学習器を用いて自動分類した特許情報の中のスコアが閾値以上（例えば、0.95 以上）の特許情報の数が十分に多くなった場合に、過去に人手により分類された特許情報を合わせて、PNF を動作させ、最適学習器を取得する。

そして、その最適学習器の精度が、元の学習器の精度を上回っている場合、新たに作成した最適学習器を、元の学習器に上書きして、管理する。このような機能を、学習器自動更新機能という。なお、学習器自動更新機能は、さらに人手により分類された特許情報の数を考慮せずに、定期的に動作させても良い。

5　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（PNF）について紹介した。

また、PNF を用いた SDI 調査について説明した。特に、SDI 調査において有用な学習候補決定機能、学習器自動更新機能について説明した。

今後、PNF を使用した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）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）

2）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）

3）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）