特許分類の自動推定の取り組み

Automated classification of patent documents

一般財団法人工業所有権協力センター　研究所総括研究員　近藤　裕之

平成 27 年 10 月より現職

Kondoh-hiroyuki@ipcc.or.jp　　TEL 03-6665-7870

1 はじめに

今年の６月に人工知能が書いた特許を特許庁が受理というニュースに衝撃が走った。つい１年ほど前までは、特許の出願は機械には書けないとも言われていただけに、IT 分野の技術進歩の速さに改めて驚かされる。

政府も本年４月に「人工知能技術戦略会議」を発足し、研究の加速と実用化･産業化の推進を強力に進めており、さらに特許庁でも、特許の事務処理に人工知能が活用できないかの検証事業が開始されている。

一般財団法人工業所有権協力センター（IPCC: Industrial Property Cooperation Center、以下「財団」という。）では、特許文献の分類付与を事業の大きな柱としているが、研究所を中心として、最新の技術動向を注視しながら、人工知能をはじめとした IT を活用した業務の効率化及び高精度化について常に検討を行っている。

本稿では、過去に財団が公表・寄稿させていただいた記事のうち分類推定についてのものをまとめ、財団がどのようなシステムを実用化し、どのような研究を行ってきたのかについて俯瞰する。

2 分類推定についてのシステム開発と研究

特許は、年間 30 万件以上出願されるが、原則この全てについて分類が付されて公開される。出願された書類に分類を付与するためには、各技術分野の専門家が詳細な分類を付与するが、その前処理として、どの技術分野の担当者に配布するのかテーマコードレベルで粗ぶるいする必要がある。この粗ぶるいの作業は、出願全件に対して行わなければならないことから、人手で行うと非常に大きな負担となるため、システム化できないかと検討が開始された。

このようなニーズから、分類推定については、財団では 10 年以上前から研究開発し実用化をしている。

日本で主に用いられている分類は、①Ｆタームのテーマコード（約２千２百）、②IPC（約７万）、③FI（約 19 万）、④Ｆターム（約 40 万）であるが、分類を推定する技術レベルは分類の数が増えると急速に難易度が高くなる。

したがって、最終的には FI の分類が推定できることを目指としても、まずは粗ぶるいできればよいということで、Ｆタームのテーマコードレベルでの分類推定が出来ないか検討を行うことから始め、研究の結果、業務に適用できる精度であることが確認されたため、平成 11 年に実用化した。それらは、Japio YEAR BOOK 2007 や、IPCC 創立 20 周年記念誌、同 30 周年記念誌、で紹介している。

そして、次にＦタームレベルでの分類が推定できないかの研究を行った。これについては、Japio YEAR BOOK 2012 － 2015 で紹介している。こちらについては、人手による付与の精度に達していないため、現在のところ実用化のめどはたっておらず、さらなる精度向上がないか検討が必要な段階である。

3 分類推定の研究・開発

3.1 テーマコードレベルでの分類推定

財団では、F タームのテーマコードレベルでの分類推定ができないか研究し、実用化のめどが立ったことから、「大分けシステム」と称して開発し、実用化している。

大分けシステムは、明細書テキスト中の技術用語を基にテーマコードを自動的に推定するものである。

その概要を図 1 に示す。

その処理概要は下記のとおりである。

(1) まず、本願の分類を推定するためのデータベース作成を行う。具体的には、過去の公開公報（例えば、過去 10 年分）を解析してどのテーマコードにどの文献が分類されるのかという統計情報を大分けデータベースに蓄積する。

(2) 次に、本願を分類するため、本願テキストから単語抽出を行う。

(3) 大分けデータベース内の単語と比較して、IPC クラス別に類似度を計算してあらぶるい処理（最も近い IPC クラスの算出）を行う。例えば、A44（類似度 0.32）、A45（類似度 0.13）、A46（類似度 0.24）、A47（類似度 0.96）と算出された場合、A47 を本願が属する IPC クラスと判定する。

(4) 特定された IPC クラスに属する各公報との類似度計算を行う。例えば、A 公報（A47J27/60[テーマコード：4B055]、類似度 0.855）、B 公報（A47J37/04[テーマコード：4B040]、類似度 0.652）、C 公報（A47J27/21[テーマコード：4B055]、類似度 0.532）、D 公報（A47J37/10[テーマコード：4B059]、類似度 0.301）・・・というように各公報との類似度計算を行う。

(5) 各テーマコード毎に公報と比較した類似度を積算してヒストグラムを作成する。

(6) 類似度の積算値の大きい順にテーマコードを並べ、当該テーマコードを担当するグループを突き合わせて、担当グループを算出する。

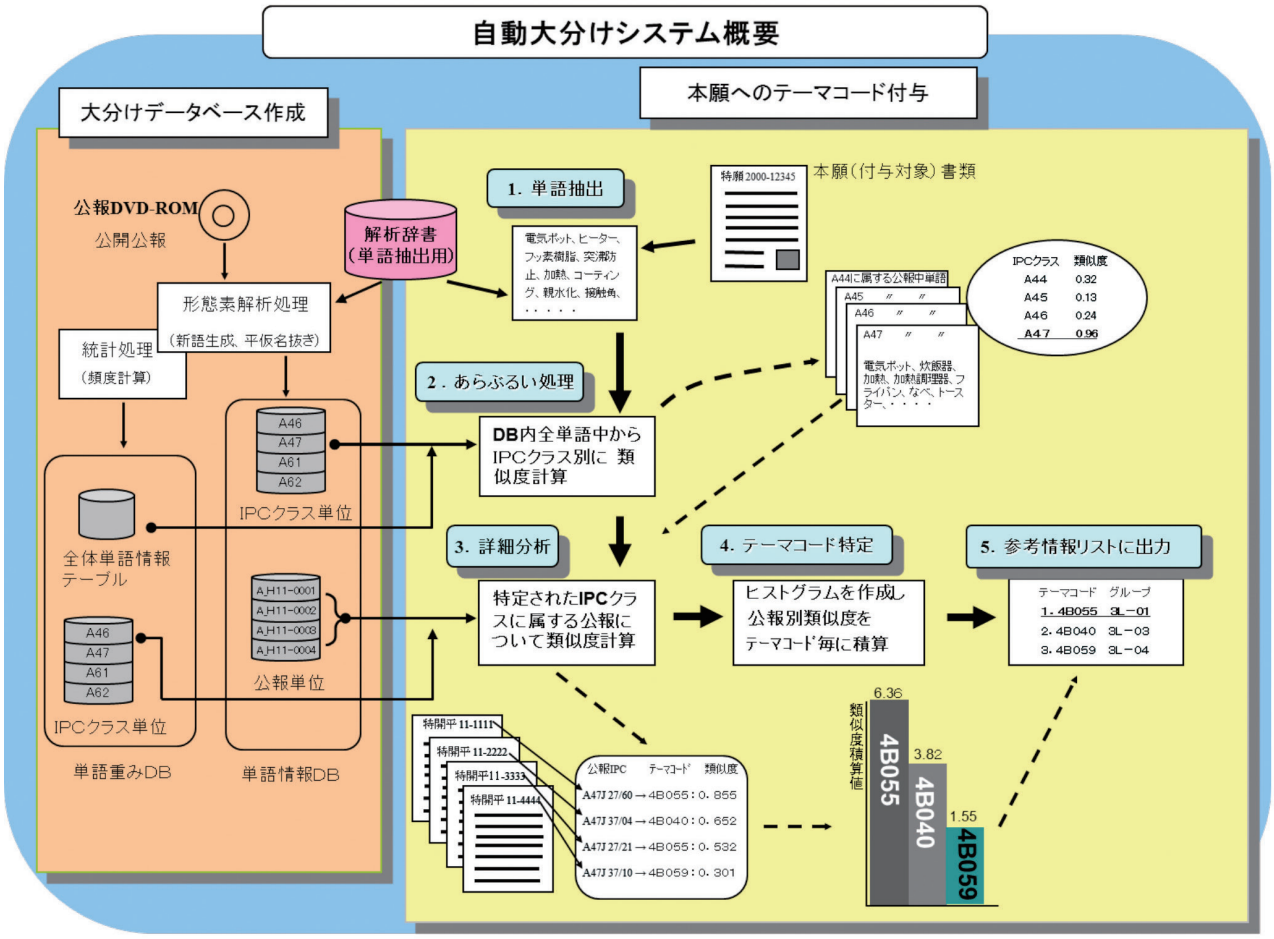

図 1

このシステムは、平成 10 年に研究開発に着手し、平成 11 年 4 月に一部の文献への付与（年間 10 万件を対象）で運用を開始し、平成 13 年度から全特許出願に拡大している。当初のテーマコードの推定精度（推定したテーマコードが主分類である精度）は約 40% であったが、その後、推定精度向上のために、大きく以下の 4 つの段階を経て、平成 27 年度にはテーマ特定精度 70% を達成している。

第 1 段階（平成 14 ～ 17 年度）

大グループ（全 39 グループ）レベルの精度向上に重点的に取り組み、国際出願・分割出願・出願人 IPC の分類情報の利用、発明の名称の利用等を行うように改造した。（テーマコードの推定精度約 40%）

第 2 段階（平成 18 ～ 20 年度）

テーマコードレベルの推定精度向上に取り組み、要約書の利用、類義語の検索機能、辞書機能の拡充等を行った（同精度約 60%）。

第 3 段階（平成 21 ～ 24 年度）

小グループ（全 78 グループ）及びテーマコードレベルの推定精度向上に取り組み、新たなテーマ特定方法の導入、機械学習モデルの導入等を行った（同精度約 70%）。

第 4 段階（平成 25 ～ 27 年度）

副分類のテーマコードレベルの推定精度向上に取り組んだ。副分類は、誤った分類が記載されていると、分類付与者が分類が誤っていることを確認・判断するために文献を読んでしまうという業務の無駄が発生する。この無駄を低減させるための取り組みである（同精度約 70%）。

上記したように、テーマコードレベルの推定精度向上に重点をおいて改造を行った時期に精度が向上し、主分類のテーマコード推定精度以外の機能改善に取り組んだ時期は精度が変わっていない。しかしながら、その時期においても、精度が低下しないよう考慮しつつ主分類のテーマコード推定精度以外の機能改善に取り組んでおり、直接的には精度向上対応を行っていなくとも、精度の低下防止には腐心している。

3.2　F タームレベルでの分類推定

財団では、大分けシステムにおいて機械学習モデルを導入しているが、テーマコードより詳細な分類である F タームのレベルでも分類を推定できないか研究を行った。

特許文献に（人手によって）付与されている分類情報を正解データとして機械に自動学習させて分類付与ルールを作成、そして、その分類付与ルールに基づいて、新たな特許文献に対して機械が分類を推定するという点については、大分けシステムとコンセプトは同じであるが、学習するデータとして、所定の分類が付与されているもの「正例」と、所定の分類が付与されていない「負例」の両方を利用する点が大きく異なる。

機械学習などの自然言語処理には SVM（Support Vector Machine）という 2 つのクラス（ここでは「正例」と「負例」）のパターンを識別するのに適したパターン認識モデルを用い、機械学習には財団が有する情報資産である付与根拠データを用いた。

F タームの推定結果は、光学系、機械系の技術分野では F 値[1] で 5 割、化学系、電気系で 3 割～ 4 割であり、テーマによるばらつきが大きいことが判明した。このため、精度自体が低く、ばらつきの大きさの解消が課題として浮彫になった。

このため、付与根拠データだけでなく、公報全文を用いて機械学習させる手法を試みたところ、光学系、機械系の技術分野で 5 割、化学系で 6 割、電気系で 4 割に向上した。5 割を下回るものがあり、精度向上策の検討が必要であった。

次に、TF・IDF（素性出現頻度）法を活用し、「1」又は「0」ではなく素性を重要度（実数値）で表現して正例と負例の境界をハッキリとできる計算手法を用いることを試みたところ、全分野で精度向上が見られたが、極めて少ない上昇にとどまった。

1　・Precision（精度）･･･ 付与すると推定したもののうち、正解分類に存在していた割合。割合が高いと、ノイズが少ないと評価できる。
・Recall（再現率）･･･ 正解分類に存在したもののうち、付与すると推定できた分類の割合。割合が高いと、漏れが少ないと評価できる。
・F 値（Precision と Recall の調和平均）･･･ 下記式で示される、Precision と Recall との総合評価値。

$$F\text{値} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

そして、Fタームの観点はドット「・」で示される階層構造で規定されていることから、この点を考慮してFターム推定を行うこととし、機械系と化学系で評価したところ、それぞれ6割弱と5割強ということで若干の精度向上が見られたところである。

4 まとめ

分類推定について、テーマコードレベルのものを実用化させ、さらに詳細なFタームレベルでのものを研究してきたが、様々な精度向上策を施しても精度が向上しにくくなっていることから、一度立ち止まって、言語の解析手法、機械学習の方法を再検討することが必要な状況になっているようにも思われる。

参考文献

[1] 情報システム部，IPCCシステムの変遷，IPCC創立20周年記念誌－IPCC20年のあゆみ－，p.106-107

[2] 調整部，一元付与業務の経緯－自動大分けシステム・サポートチーム・Σシステム－，IPCC創立30周年記念誌―IPCCのあゆみ―，p.179

[3] 情報システム部，IPCCシステムの変遷，IPCC創立30周年記念誌―IPCCのあゆみ―，p.206

[4] 笹野秀生，特許分類の自動推定に向けた取り組み－機械学習による自動分類技術の特許文献への適用－，Japio YEAR BOOK 2012，p.208-211

[5] 小林英司，特許分類の自動推定に向けた取り組み－機械学習による自動分類技術の実用化に向けて－，Japio YEAR BOOK 2013，p.234-237

[6] 小林英司，特許分類の自動推定に向けた取り組み－特許分類の階層構造を利用した自動推定－，Japio YEAR BOOK 2014，p.200-203

[7] 小林英司，特許分類の自動推定に向けた取り組み－機械学習による自動分類推定の課題と今後の展開－，Japio YEAR BOOK 2015，p.272-275