# 学術論文の国際特許分類への自動分類：<br>第 7 回 NTCIR ワークショップ特許マイニングタスク成果報告

難波英嗣　藤井敦　岩山真　橋本泰一
広島市立大学　筑波大学　日立製作所　東京工業大学

## 1. はじめに

本稿では、国立情報学研究所(NII)が主催する評価ワークショップ「NTCIR」において、筆者らが行った特許と論文を対象とした情報処理のためのテストコレクションの構築研究について述べる。筆者らは本ワークショップにおいて「特許マイニングタスク」を企画し、国内外から参加グループを募り、2007 年から研究を開始した。

近年、大学研究者自身が関連論文だけでなく関連特許について情報を検索したり、特許を出願したりする機会が増えており、2008 年 6 月に政府の知的財産戦略本部が発表した「知的財産権推進計画 2008」においても、推進計画 2006、2007 に引き続き、大学研究における特許情報の重要性が謳われている。この計画で、大学研究者の利用を想定した特許・論文情報統合検索システムの整備が含まれていることからも、このような傾向が今後さらに強まっていくと思われる。

特許と論文を検索するのは、大学研究者に限った話ではない。例えば、特許庁の審査官は、出願された技術が特許権の取得に該当するかどうか判断するために、過去に同様の特許が出願されたり論文が発表されたりしていないか調査する。これは、一般に先行技術調査と呼ばれている。この他に、サーチャーと呼ばれる専門の担当者が審査官による審査を経た出願技術を再調査し、競合する他者の権利を無効化するために民間企業の社内で行われる無効資料調査でも、論文と特許が検索対象となる。

こうした状況を鑑み、特許と論文を対象にした検索や動向分析など、様々な目的に利用可能な言語処理技術の開発を最終目標とし、そのための第一歩として筆者らが位置づけているのが、本評価ワークショップの特許マイニングタスクである。

本タスクでは、日本語または英語論文抄録に、特許分類体系のひとつである「国際特許分類」(International Patent Classification: IPC)を自動的に付与することを目的とする。特許を分類するタスクは、これまでに NTCIR-5[Iwayama 2005]と 6[Iwayama 2007]における F ターム分類タスクが実施されてきたが、今回のタスクでは、分類対象となる文書が特許から論文に変わるため、特許と論文で使われる用語の違いについて新たに検討する必要がある。

特許では請求範囲をなるべく広く確保するため、一般性の高い特許用語を用いて記述する傾向がある。また、特許では学術用語よりも多様な表現が用いられることが多い。例えば、「機械翻訳」という学術用語に対する特許用語は「機械翻訳」の他にも「自動翻訳」「言語変換」などがある。このため、単純に表層的な単語の一致度を用いる従来の分類モデルでは、十分な分類精度が得られるとは限らない。本タスクでは、このような論文と特許の用語の使われ方の違いを吸収できる分類や検索のための基礎技術の確立を目指している。

## 2. 関連研究

ジャンル横断検索や文書分類に関しては、これまでにいくつかの先行研究がある。NTCIR-3 で実施された技術動向調査タスク[Iwayama 2002]では、与えられた新聞記事と関連する特許を検索する、という課題が設定された。このタスクにおいて、Itoh ら[Itoh 2002]は、"Term Distillation"という手法を提案している。例えば、「社長」という単語は新聞記事中では高頻度で出現するが、特許中では出現頻度が非常に低い。このため、「一般的な用語ほど重要ではない」という考えに基づいて単語の重要度を計算する tf*idf 等の手法を用いると、同じ単語でも新聞記事と特許では重要度が大きく異なる。そこで、Itoh らは、単語の新聞記事集合中での出現頻度と特許中での出現頻度の違いを考慮して単語の重み付けを行うことで、ジャンルを横断した文献の対応付けを行っている。

特許と論文を横断的に検索するための研究として Nanba ら[Nanba 2008:a]の研究が挙げられる。近年、特許中で関連論文を、また論文において関連特許を引用するケースが増えているが、このような文書間の引用関係をたどれば、論文や特許と関連する文書を集めることができる。そこで Nanba らは特許中で関連文献が引用される「従来の技術」という項目を解析して引用論文の書誌情報を抽出し、特許と論文間の引用関係を解析している。ただ、特許中の引用文献の中で論文が占める割合と、論文中の引用文献の中で特許が占める割合は数パーセント程度であるため、あるテーマに関する特許と論文を網羅的に収集するのに、引用関係をたどるだけでは十分とは言えない。

特許と論文を横断的に検索するための別のアプ

ローチとして、難波ら[難波 2009]は、論文用語を特許用語に自動変換する手法を提案している。例えば、論文用語「フロッピーディスク」を特許用語「磁気記録媒体」に自動変換する。難波らは、論文用語の特許用語への変換を実現するため、特許と論文間の引用関係に着目している。一般に、引用関係にある特許と論文は、同一トピック(分野)である可能性が高い。そこで、ある用語を表題に含んだ論文を収集し、それらと直接引用関係にある特許から、特許のトピックを示す用語を抽出すれば、入力された論文用語に関連する特許用語の変換が実現できる。

## 3. 特許マイニングタスク

### 3.1 タスクの概要

前述のとおり、特許マイニングタスクでは日本語または英語論文抄録に、特許分類体系のひとつである IPC のコードを自動的に付与する。IPC は、特許文献の技術内容によって上から順に「セクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の 5 階層から構成・分類されており、国際特許分類第 6 版ではサブグループのレベルで約 50,000[1]の IPC コードが存在する。本タスクでは、最下層の「サブグループ」レベルの IPC コードを論文抄録に付与することを目的とする。図 1 は日本語論文の例である。ここで、<TOPIC-ID>は論文の ID を、<TITLE>と<ABSTRACT>は論文表題と概要を、それぞれ示している。タスクの参加者は、図 1 のような入力が与えられると、対応する IPC コードを自動的に出力するシステムを構築することが求められる。

<TOPIC><TOPIC-ID>312</TOPIC-ID>
<TITLE>二値画像用高速符号化/復号 LSI</TITLE>
<ABSTRACT>二値画像データを高速で符号化,復号する LSI を開発した。参照ラインデータ上に「基準色変化点」を探すのと並行して、それを参照するランのイメージデータを生成する方式により、復号性能を向上させた。また、符号化時と復号時共に同じ方向にデータが流れるパイプライン構成とし、さらに主な回路は共通化する構成によって回路を簡略化した。</ABSTRACT>
</TOPIC>

図 1 システムの入力例

特許マイニングタスクでは、次のサブタスクが実施された。

- 日本語サブタスク(Japanese):日本語の論文を日本語で記載された特許データを用いて分類する。
- 英語サブタスク(English):英語の論文を英語で記載された特許データを用いて分類する。
- 言語横断サブタスク(J2E):日本語の論文を英語で記載された特許データを用いて分類する。

この他、英語の論文を日本語で記載された特許データを用いて分類するサブタスクも参加募集をしたが、参加者がいなかった。以上のサブタスクは図 2 にまとめられる。

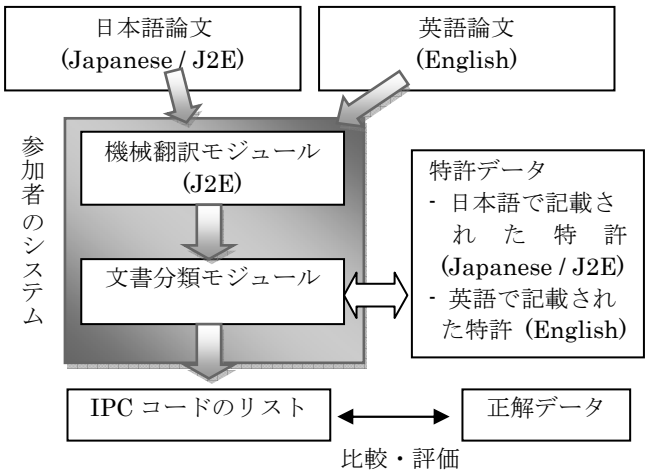


図 2 サブタスクのまとめ

### 3.2 特許マイニングタスクのポイント

特許マイニングタスクでは、システムを構築する上で、以下の 3 つのポイントを考慮する必要がある。

- **(ポイント 1) 膨大な IPC コード数および学習データ**

前述のとおり、IPC コード数が約 30,885 件と非常に多い。また、次節でも述べるが、参加グループに配布される訓練用データは、350 万～450 万件と膨大である。このため、機械学習を用いる場合には、非常に計算コストがかかるという問題がある。

- **(ポイント 2) IPC コード毎の文書数の違い**

1 つの IPC コードに属する文書数は、平均で約 80 件であるが、50 件未満のものが全体の 57%、10 件未満のものが全体の 25%と、IPC コード毎に、属する文書数が異なる。このため、文書分類手法として広く用いられている k-Nearest Neighbor(k-NN)法をそのまま適用した場合、入力論文には文書数の多い IPC コードほど付与される確率が高くなる、という問題がある。

[1] 特許マイニングタスクでは、これらのうち、学術分野とは関連性の低い分野を除外した 30,885 の IPC コードを対象とした。

- **（ポイント 3）特許と論文で使われる用語の違い**

2 節でも述べたとおり、特許と論文では使われる用語に違いがある。このため、単純に表層的な単語の一致度を用いる従来の分類モデルでは、十分な分類精度が得られるとは限らない。

### 3.3 文書データ

各サブタスクで使われた文書データを表 1 にまとめる。

表 1 文書セット

| データ名 | 年 | サイズ | 文書数 | 言語 |
|---|---|---|---|---|
| (1) 日本国公開特許公報 | 1993–2002 | 100 GB | 3.5M | 日 |
| (2) 米国特許 | 1993–2000 | 33 GB | 0.99M | 英 |
| (3) Patent Abstracts of Japan (PAJ) | 1993–2002 | 4.2 GB | 3.5M | 英 |
| (4) NTCIR-1、NTCIR-2 言語横断タスクテストコレクション(論文抄録データ) | 1988–1999 | 1.4 GB | 0.26M | 日/英 |

また、この他に、データ(1)～(3)の各文書に人手で IPC コードを付与したデータが訓練用データとして参加グループに配布された。

### 3.4 評価

- **正解データ**

976 論文に人手で IPC を付与したデータを用意し、このうち 97 件をドライランに、残りの 879 件をフォーマルランに用いた。なお、正解データの作成に関する詳細は、[Nanba 2008:b]を参照されたい。

- **評価尺度**

評価尺度は、MAP (Mean Average Precision)、再現率、精度を用いた。

### 3.5 参加グループ

日本語サブタスクには 24 システム、英語サブタスクには 20 システム、言語横断サブタスク(J2E)には 5 システムからの結果が、それぞれ提出された。また、参加グループ数は計 12 グループであり、その内訳は表 2 に示すとおりである。

表 2 参加グループ内訳

| | 日本 | アジア（日本以外） | 欧州 | 北米 |
|---|---|---|---|---|
| 大学 | 3 | 4 | 0 | 2 |
| 企業 | 2 | 0 | 1 | 0 |

## 4. 評価結果および考察

### 4.1 評価結果

日本語サブタスク、英語サブタスク、言語横断サブタスク(J2E)の MAP による評価結果を、表 3、4、5 にそれぞれ示す。

表 3 日本語サブタスクの各システムの MAP 値

| Run ID | MAP | Run ID | MAP |
|---|---|---|---|
| HTC13 | 44.02 | HTC04 | 41.65 |
| HTC11 | 43.71 | nttcs4 | 39.64 |
| HTC12 | 43.61 | *HCU1 | 39.13 |
| HTC07 | 43.60 | *HCU2 | 39.06 |
| HTC01 | 43.34 | HTC14 | 38.62 |
| HTC06 | 43.29 | nttcs3 | 35.72 |
| HTC05 | 43.26 | nttcs2 | 34.35 |
| HTC08 | 43.23 | nttcs1 | 33.03 |
| HTC10 | 43.18 | KECIR | 27.27 |
| HTC03 | 42.68 | *HCU3 | 14.12 |
| HTC02 | 42.36 | nut1-1 | 6.98 |
| HTC09 | 42.27 | nut2-1 | 4.06 |

(HCU1, HCU2, HCU3 はオーガナイザのシステム)

表 4 英語サブタスクの各システムの MAP 値

| Run ID | MAP | Run ID | MAP |
|---|---|---|---|
| NEUN1_S1 | 48.86 | rali2 | 14.37 |
| NEUN1_S2 | 47.21 | ICL07 | 14.36 |
| NEUN1_S3 | 44.53 | rali1 | 14.23 |
| xrce_e2j2e | 42.45 | ICL07_2 | 13.39 |
| xrce_en_lm | 42.09 | BRKLY-PM-EN-02 | 12.65 |
| xrce_en_filter | 41.83 | AINLP04 | 10.45 |
| xrce_en_pp | 41.49 | BRKLY-PM-EN-04 | 9.90 |
| nttcs2 | 34.79 | AINLP01 | 9.78 |
| nttcs1 | 33.74 | BRKLY-PM-EN-03 | 9.37 |
| KECIR | 29.03 | PI-5b | 3.79 |

表 5 言語横断サブタスクの各システムの MAP 値

| Run ID | MAP |
|---|---|
| xrce_j2e | 43.80 |
| AINLP05 | 10.70 |
| AINLP06 | 10.41 |
| AINLP02 | 9.41 |
| AINLP03 | 9.34 |

### 4.2 考察

- **「（ポイント 1）膨大な IPC コード数および学習データ」に関する考察**

"HTC13"[Mase 2008] 、 "NEUN1_S1"[Xiao 2008]、"xrce_j2e"[Clinchant 2008]は、それぞれ日本語、英語、言語横断サブタスクにおいて最も良い MAP 値を得ている。これらのシステムは、いずれも k-NN 法を用いている。また、これらのシステム以外にも、多くのシステムが k-NN 法を採用している。これは、IPC コード数や学習データが増えるほど計算コストも増加する機械学習手法と異なり、k-NN 法は、計算コストが IPC コード数や学習データの規模に依存しないためであると考えられる。

他方、機械学習を用いたシステムも少数ながら

ある。日本語サブタスクにおける"nttcs4"[Fujino 2008]はロジスティック回帰モデルを用い、MAP 値 39.14 を得ている。また、英語サブタスクにおける"nttcs2"[Fujino 2008]は、ロジスティック回帰モデルとナイーブベイズを組み合わせた手法により、MAP 値 34.79 を得ている。

英語サブタスクで 2 番目に良い MAP 値を得ている"NEUN1_S2"[Xiao 2008]も、一部、機械学習を取り入れている。"NEUN1_S2"は、k-NN 法を用いて、ある入力論文に対する IPC コードのリストを得た後、リランキング手法によりリスト内の IPC コードを並べ替えているが、このリランキングに機械学習のひとつである RankSVM [Herbrich 1999]を用いている。

- **「(ポイント 2) IPC コード毎の文書数の違い」に関する考察**

IPC コード毎の文書数の違いを考慮したシステムは、"NEUN1_S1"[Xiao 2008]と"NEUN1_S2"の 2 つである。これらのシステムは、いずれも、「文書数の多い IPC コードほど付与される確率が高くなる」という k-NN 法の問題を解消するため、「訓練データ中で文書数の多い IPC コードは、k-NN 法によるその IPC のスコアを小さくする」というパラメータを導入することにより、k-NN 法を改良している。

- **「(ポイント 3) 特許と論文で使われる用語の違い」に関する考察**

"HTC14"[Mase 2008]は、入力論文中の用語と関連する特許用語を収集するため、入力論文中の用語を用いて関連特許を検索し、検索結果として得られた特許中の用語を用いて再検索を行う、いわゆる疑似適合性フィードバックを用いている。再検索の後、k-NN 法により入力論文に対する IPC コードのリストを得ている。しかし、残念ながら疑似適合性フィードバックを用いない単純な k-NN 法の方が、高い MAP 値を得ている。

## 5. おわりに

本稿では、NTCIR-7 において、筆者らが企画した「特許マイニングタスク」について、タスクの概要、評価結果およびそこから得られた種々の知見について報告した。

## 参考文献


[Clinchant 2008] Clinchant, S. and Renders, J.-M. "XRCE's Participation to Patent Mining Task at NTCIR-7". Proceedings of the 7th NTCIR Workshop Meeting, pp.351-353 (2008)

[Fujino 2008] Fujino, A. and Isozaki, H. "Multi-label Classification using Logistic Regression Models for NTCIR-7 Patent Mining Task". Proceedings of the 7th NTCIR Workshop Meeting, pp.354-357 (2008)

[Herbrich 1999] Herbrich, R., Graepel, R., and Obermayer, K. "Support Vector Learning for Ordinal Regression". Proceedings of the 9th Ingernational Conference on Artificial Neural Networks, pp.97-102 (1999)

[Itoh 2002] Itoh, H., Mano, H., and Ogawa, Y. "Term Distillation for Cross-DB Retrieval". Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task (2002)

[Iwayama 2002] Iwayama, M., Fujii, A., Kando, and N., Takano, A. "Overview of Patent Retrieval Task at NTCIR-3'. Proceedings of Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task (2002)

[Iwayama 2005] Iwayama, M., Fujii, A., and Kando, N. "Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task". Proceedings of the 5th NTCIR Workshop Meeting (2005)

[Iwayama 2007] Iwayama, M., Fujii, A., and Kando, N. "Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task". Proceedings of the 6th NTCIR Workshop Meeting (2007)

[Mase 2008] Mase, H. and Iwayama, M. "NTCIR-7 Patent Mining Experiments at Hitachi". Proceedings of the 7th NTCIR Workshop Meeting, pp.365-368 (2008)

[Nanba 2008:a] Nanba, H., Anzen, N., and Okumura, M. "Automatic Extraction of Citation Information in Japanese Patent Applications". International Journal on Digital Libraries, Vol.9, No.2, pp.151-161 (2008)

[Nanba 2008:b] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. "Overview of the Patent Mining Task at the NTCIR-7 Workshop". Proceedings of the 7th NTCIR Workshop Meeting, pp.325-332 (2008)

[難波 2009] 難波英嗣，釜屋英昭，竹澤寿幸，奥村学，谷川英和，新森昭宏．"特許用語の論文用語への自動変換"．情報処理学会論文誌データベース(2009) (採録予定)

[Xiao 2008] Xiao, T., Cao, F., Li, T., Song, G., Zhou, K., Zhu J., and Wang H. "KNN and Re-ranking Models for English Patent Mining at NTCIR-7". Proceedings of the 7th NTCIR Workshop Meeting, pp.333-340 (2008)