# 中国特許文献に付与されている国際特許分類情報の精度に関する調査
# 報告書 概要版

平成25年2月

特　許　庁

目次

1.1. 調査の目的 ..... 1
1.2. 調査の考え方 ..... 1
1.3. 調査の概要 ..... 2
1.3.1. 調査方法の概要 ..... 2
1.3.2. 機械的判定手法 ..... 5
1.4. 調査結果概要 ..... 7

## 1.1. 調査の目的

企業の事業展開や特許出願等のグローバル化が加速する現代において、知的財産の分野で課題となっているのが、知財大国化する中国の特許文献への対応である。急速な経済成長に伴い中国特許文献は急増しており、同時に、知的財産関連訴訟件数も増大している。特許出願人が出願前において、あるいは特許審査官が審査時において、適切な先行技術文献調査を行うためにも、また、日系企業が中国で事業展開するに当たり、中国国内における特許権を侵害してしまうリスクを避けるためにも、中国特許文献へのアクセス性を向上させることの必要性は年々増加している。

中国特許文献へのアクセス性を向上させて先行技術調査を効率化させる有効な手段としては、言語に依存しない分類情報を活用することが挙げられる。しかし、中国文献に付与されている国際特許分類（IPC）情報を特許情報として利用することを想定した場合、中国における特許文献へのIPC の分類付与と、日本における特許文献への FI の分類付与との間で、運用に差があるため、以下の問題点が認識されている。

1. 中国で所定の IPC が付与された案件の中に、当該 IPC に基づいた所定のテーマに対応する FI の付与対象でない案件（テーマ外案件）が含まれる。
2. 日本における分類付与の運用では所定のテーマの FI が付与されるべきにもかかわらず、中国では所定のテーマに対応する IPC が付与されない案件（付与漏れ案件）が生じる。

これらの問題はサーチ漏れやノイズの発生につながる大きな懸念事項となるため、本調査では、これらの問題を解決するための機械的手法を検討するとともに、その実効性の検証を行う。

## 1.2. 調査の考え方

中国で所定の IPC が付与された案件の中から、当該 IPC に基づいた所定のテーマに対応する FI の付与対象でない案件（テーマ外案件）を機械的に抽出することができれば、また、中国では所定のテーマに対応する IPC が付与されない案件の中から、日本における分類付与の運用では所定のテーマの FI が付与されるべき案件（付与漏れ）案件を機械的に抽出することができれば、上記問題を解決することができる。

そこで、本調査では、テーマ外案件及び付与漏れ案件を機械的に抽出するための方法について検討し、以下の仮説を立てた。

1. 所定のテーマの FI が付与される特許文献には、そのテーマ特有のキーワードが多く含まれている。

2. よって、ある特許文献が所定のテーマのテーマ内であるかテーマ外であるか(即ち、所定のテーマの FI が付与されるか否か)は、その特許文献において、所定のテーマ特有のキーワードが多く含まれるか否かによって判断することができる。
3. 具体的な手法としては、(1)全てのテーマについてテーマ特有のキーワードからなるキーワードテーブルを作成し、(2)特許文献に対して、全てのテーマについてテーマ特有のキーワードが出現する頻度を計算し、(3)全てのテーマを頻度によって順位付けする。
4. その特許文献が所定のテーマのテーマ内であり、そのテーマ特有のキーワードを多く含む場合は、そのテーマの順位は上位となると考えられる。
5. したがって、順位付けの結果、所定のテーマの順位が上位である場合は、その特許文献はテーマ外案件でない可能性が高く、そうでない場合は、その特許文献はテーマ外案件である可能性が高い。
6. 同様に、順位付けの結果、所定のテーマの順位が上位である場合は、その特許文献は付与漏れ案件である可能性が高く、そうでない場合は、その特許文献は付与漏れ案件でない可能性が高い。

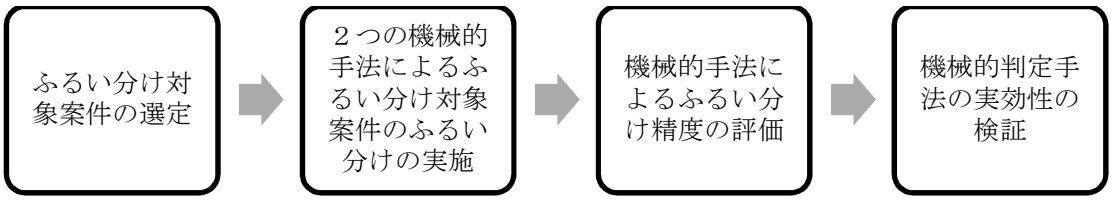
このように、本調査では、テーマごとにキーワードテーブルを作成し、そのテーマでよく用いられるキーワードを用意した上で、以上の仮説を用いてふるい分けを行い、テーマ外案件と付与漏れ案件の特定を行う。

なお、文献のボリューム等に関わらず、各文献を同一基準でふるい分けするために、テーマ内／外の判定基準は、全テーマにおける順位を採用した。

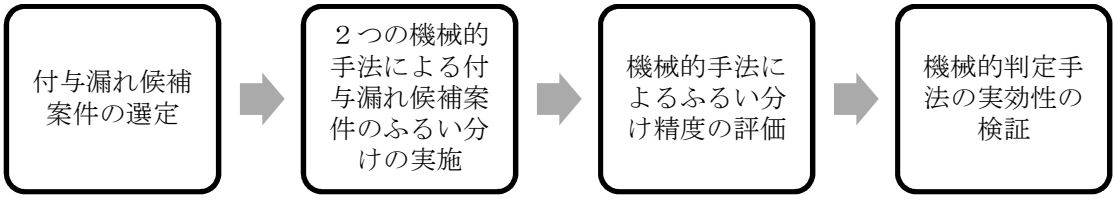
## 1.3. 調査の概要

### 1.3.1. 調査方法の概要

本調査では、テーマ外案件に関する問題を解決する方法として、以下の方法で調査を実施した。



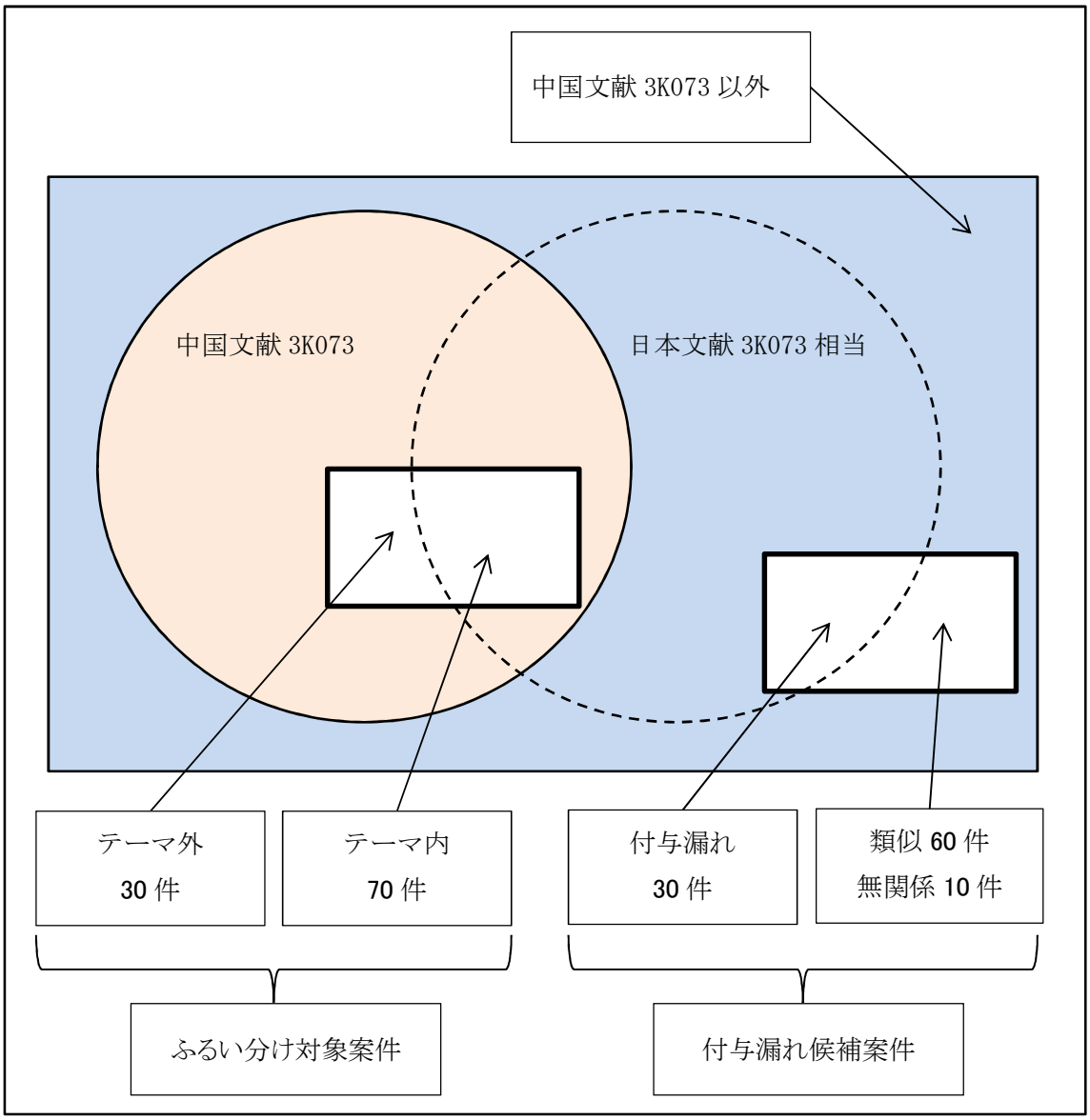

また、付与漏れ案件に関する問題を解決する方法として、以下の方法で調査を実施した。



それぞれのステップの調査内容については、以下のとおりである。

| 調査単位 | 調査内容 |
|---|---|
| ふるい分け対象案件の選定 | 中国特許文献に付与されているIPCから、対応するFIのテーマを当該文献に自動的に付与することで、検証テーマが自動付与された特許文献を選定する。これらの特許文献を、ふるい分け対象案件という。ふるい分け対象案件には、日本の分類付与でも検証テーマが付与される文献（テーマ内案件）と、日本の分類付与では検証テーマが付与されない文献（テーマ外案件）とが含まれる。ふるい分け対象案件は、JPファミリーのあるものとする。 |
| 2つの機械的手法によるふるい分け対象案件のふるい分けの実施 | ふるい分け対象案件が、テーマ外案件である可能性が高い案件か否かを機械的に判定する手法を2つ検討し、それらを用いて、ふるい分け対象案件のふるい分けを行う。 |
| 付与漏れ候補案件の選定 | 中国特許文献に付与されているIPCから、対応するFIのテーマを当該文献に自動的に付与することで、検証テーマが自動付与されなかった特許文献を選定する。これらの特許文献を、付与漏れ候補案件という。付与漏れ候補案件には、日本の分類付与では検証テーマが付与される文献（付与漏れ案件）と、日本の分類付与でも検証テーマが付与されない文献とが含まれ、後者はさらに、中国特許文献のIPCまたはJPファミリーのFIが検証テーマと関連するもの（類似案件）と、特に関係しないもの（無関係案件）とを含む。付与漏れ候補案件は、JPファミリーのあるものとする。 |
| 2つの機械的手法による付与漏れ候補案件のふるい分けの実施 | 付与漏れ候補案件が、付与漏れ案件である可能性が高い案件か否かを機械的に判定する手法を2つ検討し、それらを用いて、付与漏れ候補案件のふるい分けを行う。 |
| 機械的手法によるふるい分け精度の評価 | ふるい分け対象案件のふるい分け及び付与漏れ候補案件のふるい分けの精度を評価する。 |
| 機械的判定手法の実効性の検証 | それぞれの機械的判定手法の実効性の検証を行う。 |

以下の図はテーマコード 3K073 を例とした場合の本調査の概念図である。




今回の調査では検証テーマとして以下のテーマを用いた。

| テーマコード | 説明 | FI カバー範囲 |
|---|---|---|
| 3K073 | 光源の回路一般 | H05B37/00-39/10 |
| 4F100 | 積層体(2) | B29D9/00;B32B1/00-35/00 |
| 4J002 | 高分子組成物 | C08K3/00-13/08;C08L1/00-101/14 |
| 5C025 | TV送受信機回路 | H04N5/38-5/46 |
| 5F142 | LED素子のパッケージ | H01L33/00@H;33/00@L;33/00,400-33/00,450 |

### 1.3.2. 機械的判定手法

本調査では、F タームリストのキーワードを用いてテーマ外案件及び付与漏れ案件の判定を行う方法と、日本特許から学習したキーワードを用いてテーマ外案件及び付与漏れ案件の判定を行う方法の 2 種類の機械的手法を用いて判定を行った。

キーワードは、テーマコードと紐づけてキーワードテーブルとし、キーワードを検索するとそのキーワードを含むテーマコードの一覧が得られるようにしておいた。例えば、3K073 に紐づけられたキーワードとしては「光源」「回路」「照明」「応答性」「ちらつき」などがあり、4F100 に紐づけられたキーワードには「アクリル酸」「エポキシ樹脂」「セラミック」「外層」「樹脂」などがある。一方「ちらつき」は 3K073 だけでなく、2F013 や 2F041 や 3K005 や 5C080 にも紐づけられており、「ちらつき」で検索するとこれらのテーマコードの一覧を返す。

それぞれの手法のふるい分けツールを作成して判定を行ったが、判定手法の具体的な流れは以下のとおりである。

まず、以下の手順で、中国特許文献から日本語の名詞を抽出する。

(1) 判定対象の中国特許文献（ふるい分け対象案件または付与漏れ候補案件）を中日機械翻訳エンジンで日本語に機械翻訳する。
(2) 機械翻訳結果の日本語を、日本語形態素解析器で形態素解析する。
(3) 形態素解析結果の日本語の単語から名詞を抽出し、名詞群を得る。

次に、抽出した名詞と 2 つのキーワードテーブルのいずれかを用いて、以下の手順でスコア計算を行う。

(4) スコア計算
  (ア) 各テーマのキーワードテーブルの各キーワードが、上記名詞群の中に何回出現するかをカウントし、それを各テーマの各キーワードの出現頻度とする。
  (イ) テーマごとに、各キーワードの出現頻度の合計値を算出する。
  (ウ) 各テーマの出現頻度の合計値を、全テーマの出現頻度の合計値を足し合わせたもので割り、それを各テーマの確率とする。
  (エ) 全てのテーマの確率が算出できたら、確率の高い順に全テーマを順位付けする。

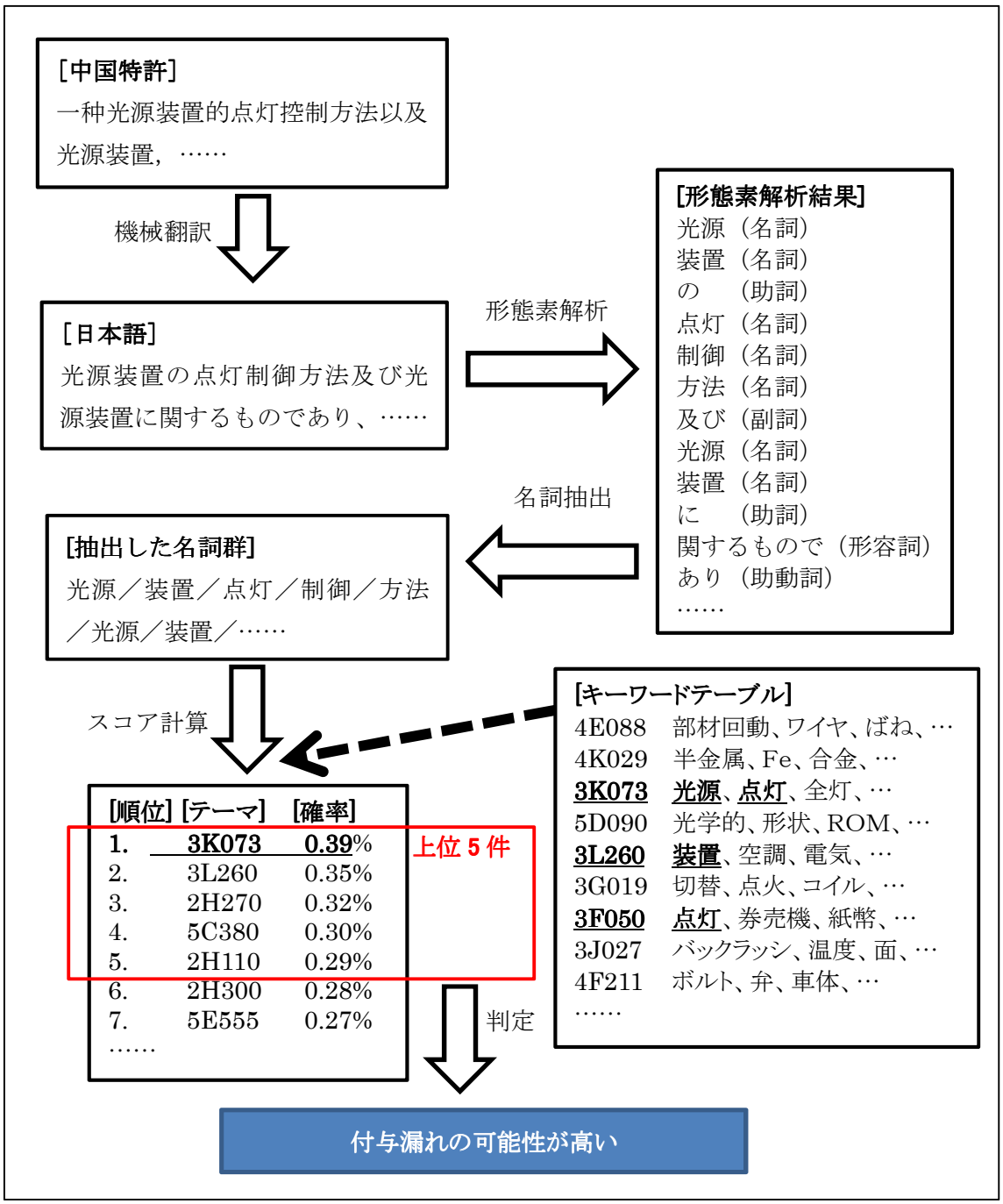
最後に以下の手順で判定を行う。

(5) 判定閾値となる所定の順位までのテーマ群に、検証テーマが含まれるか否かを判定する。
(6) ・ふるい分け対象案件のふるい分けでは、前ステップで検証テーマが含まれる場合は「テーマ外である可能性が高いとは言えない案件」であると判定し、含まれない場合は「テーマ外であ

る可能性が高い案件」であると判定する。

・付与漏れ候補案件のふるい分けでは、前ステップで検証テーマが含まれる場合は「付与漏れの可能性が高い案件」であると判定し、含まれない場合は「付与漏れの可能性が高いとはいえない案件」であると判定する。

テーマコード 3K073 の付与漏れ候補案件を、判定閾値上位 5 位で判定する場合を例に以上の手順を図示すると、以下のようになる。

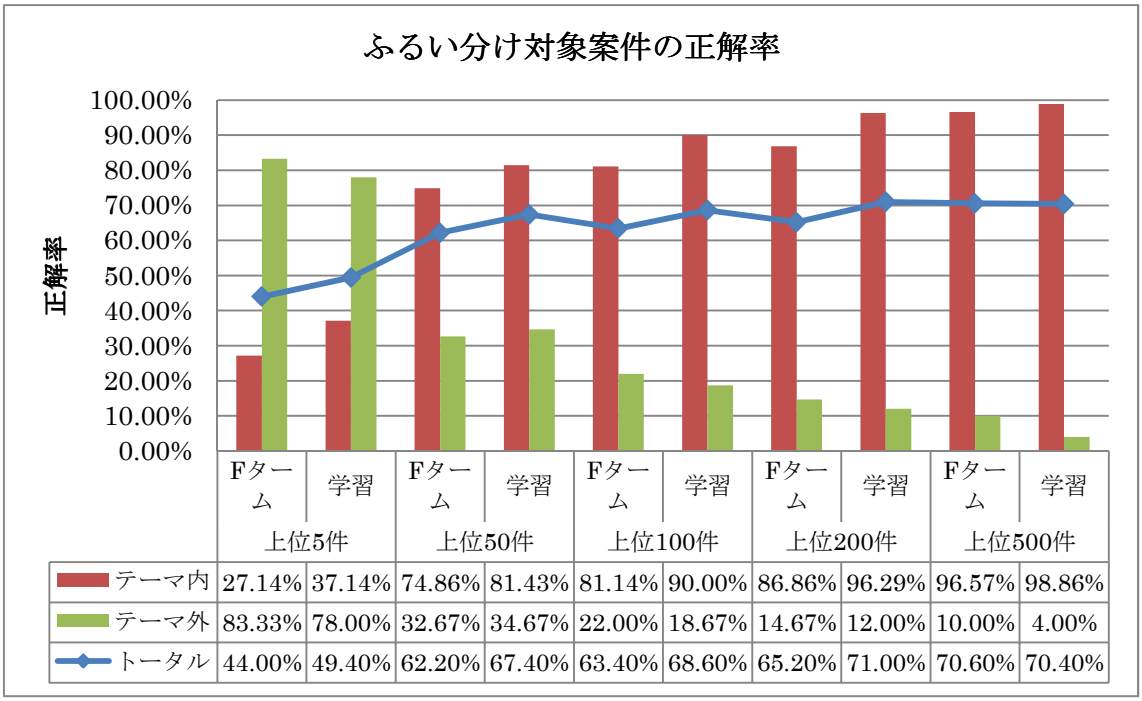


## 1.4. 調査結果概要

本調査では、機械的手法による2種類のふるい分け手法で、判定閾値（上位5件、上位50件、上位100件、上位200件、上位500件の5種類）ごとにふるい分けを行った後、ふるい分け精度の測定を行った。ふるい分け精度の測定にあたっては、JP ファミリーの FI と突き合わせることで、判定結果が正解であるか不正解であるかを特定した。即ち、あるテーマのふるい分けで中国特許文献が「テーマ外である可能性が高いとは言えない」あるいは「付与漏れである可能性が高い」と判定された場合は、JP ファミリーの FI をテーマコードに変換したものの中に当該テーマが含まれていれば正解、含まれていなければ不正解とした。一方、中国特許文献が「テーマ外である可能性が高い」あるいは「付与漏れである可能性が高いとは言えない」と判定された場合は、JP ファミリーの FI をテーマコードに変換したものの中に当該テーマが含まれていなければ正解、含まれていれば不正解とした。判定結果の正解／不正解から正解率を算出することで、ふるい分け精度の測定を行った。

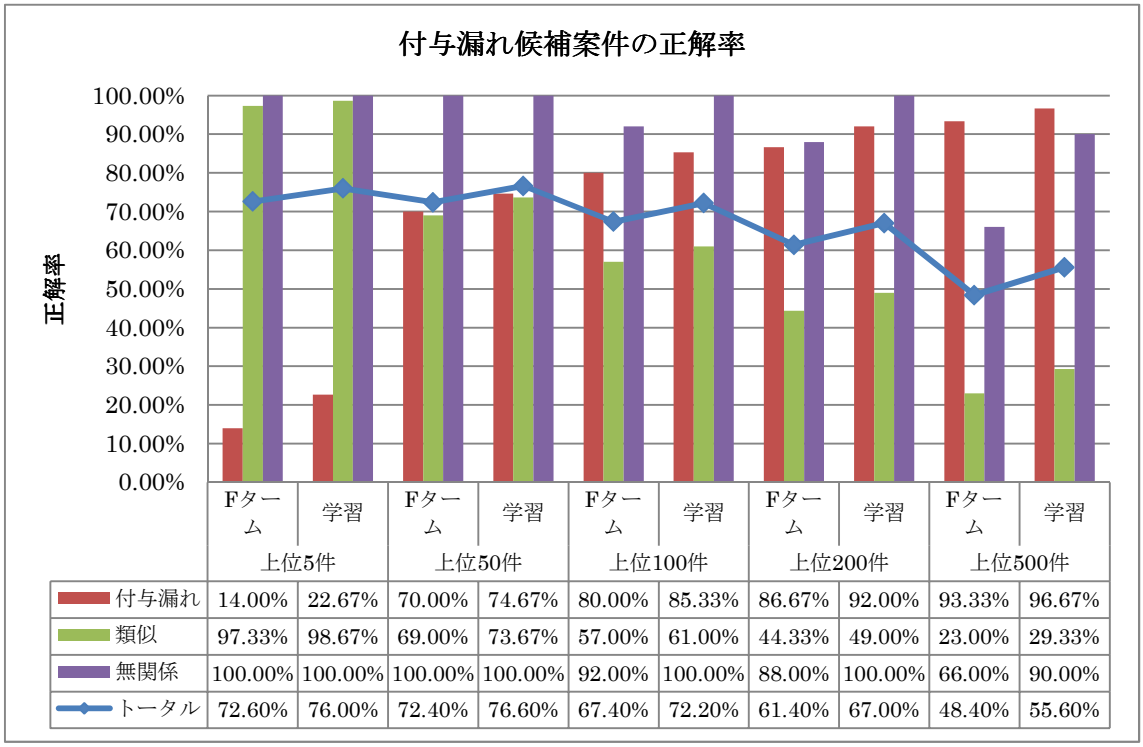
・精度の評価について

調査結果を以下に示す。まずは、ふるい分け対象案件のふるい分け結果である。




| | 上位5件 | | 上位50件 | | 上位100件 | | 上位200件 | | 上位500件 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 |
| テーマ内 | 27.14% | 37.14% | 74.86% | 81.43% | 81.14% | 90.00% | 86.86% | 96.29% | 96.57% | 98.86% |
| テーマ外 | 83.33% | 78.00% | 32.67% | 34.67% | 22.00% | 18.67% | 14.67% | 12.00% | 10.00% | 4.00% |
| トータル | 44.00% | 49.40% | 62.20% | 67.40% | 63.40% | 68.60% | 65.20% | 71.00% | 70.60% | 70.40% |

次に、付与漏れ候補案件のふるい分け結果である。




| | 上位5件 | | 上位50件 | | 上位100件 | | 上位200件 | | 上位500件 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 | Fターム | 学習 |
| 付与漏れ | 14.00% | 22.67% | 70.00% | 74.67% | 80.00% | 85.33% | 86.67% | 92.00% | 93.33% | 96.67% |
| 類似 | 97.33% | 98.67% | 69.00% | 73.67% | 57.00% | 61.00% | 44.33% | 49.00% | 23.00% | 29.33% |
| 無関係 | 100.00% | 100.00% | 100.00% | 100.00% | 92.00% | 100.00% | 88.00% | 100.00% | 66.00% | 90.00% |
| トータル | 72.60% | 76.00% | 72.40% | 76.60% | 67.40% | 72.20% | 61.40% | 67.00% | 48.40% | 55.60% |

本調査では 2 種類のふるい分けツールを用いて調査を行ったが、図の「F ターム」及び「学習」が、この 2 種類のツールを用いたふるい分け結果に該当する。また、上位何件までに検証テーマが含まれていれば「テーマ内」または「付与漏れ」である可能性が高いとみなすかという判定閾値を、図の「上位 5 件」「上位 50 件」「上位 100 件」「上位 200 件」「上位 500 件」で示している。「テーマ内」「テーマ外」「付与漏れ」「類似」「無関係」は、ふるい分け対象案件または付与漏れ候補案件が、それぞれ、1.3.1 で述べた「テーマ内案件」「テーマ外案件」「付与漏れ案件」「類似案件」「無関係案件」であることに対応する。例えば「テーマ内」の正解率は、「テーマ内案件」のうち正しく「テーマ外である可能性が高いとは言えない（即ち、テーマ内である）」と判定された割合を示し、同様に、「テーマ外」の正解率は、「テーマ外案件」のうち正しく「テーマ外である可能性が高い」と判定された割合を示す。「付与漏れ」の正解率は、「付与漏れ案件」のうち正しく「付与漏れの可能性が高い案件」であると判定された割合を示し、「類似」及び「無関係」の正解率は、「類似案件」及び「無関係案件」のうち正しく「付与漏れの可能性が高いとはいえない案件」であると判定された割合を示す。

ふるい分けの実施時には、「テーマ外」の文献を的確にテーマ外であると判定しつつ、「テーマ内」の文献を誤ってテーマ外と判定しないようにする必要がある。従って、「テーマ内」の文献の正解率を 100%に近づけつつ、「テーマ外」の文献の正解率もいかに高くできるかが肝要である。この両者の正解率が高くないと判定精度が高いとは言えない。

付与漏れの判定の実施時には、「付与漏れ」の文献を的確に付与漏れであると判定しつつ、「類似」や「無関係」の文献を誤って付与漏れと判定しないようにする必要がある。従って、「類似」と「無

関係」の文献の正解率を 100%に近づけつつ、「付与漏れ」の文献の正解率もいかに高くできるかが肝要である。この両者の正解率が高くないと判定精度が高いとは言えない。

・精度向上策について

本調査では、さらに、(1)語と句、(2)ゴミ除去、(3)ストップワード、(4)重複削除の 4 つの精度向上策を考え、その効果を調べた。

(1) 「語と句」は、ふるい分けの際のキーワードとして語と句のどちらを用いた方が精度が高いかを検討した。

(2) 「ゴミ除去」は、F タームリストからキーワードテーブルを作成する際にゴミ除去を行うと精度が上がるのではないかということを検討した。

(3) 「ストップワード」では、指示詞等の特定の語を検索の際に無視するようにすると精度が上がるのではないかということを検討した。

(4) 「重複削除」は、複数のテーマに出現するキーワードは判定の役に立たないのではないかと考え、指定数以上のテーマに出現するキーワードを除去してキーワードテーブルを作成すると精度が上がるのではないかということを検討した。

その結果、(1)「句」より「語」を用いた方が正解率が高く、(3)「ストップワード」は使用した方が正解率が高かった。(2)「ゴミ除去」は正解率にはほとんど影響がなく、(4)「重複削除」は機械的な重複削除だけでは正解率が落ちることが分かったため、最終的な手法としては採用しなかった。

上記の正解率は、「語」と「ストップワード」を使用し、「ゴミ除去」と「重複削除」は行わなかった場合の結果である。

・実効性の検証について

図を見ると、全体的にFタームを使用したツールより学習を使用したツールの方が正解率が高いことが分かる。ふるい分け対象案件では、「テーマ内」の正解率が 90%以上となるのは、学習の「上位 100 件」、学習の「上位 200 件」、F タームと学習の「上位 500 件」であった。付与漏れ候補案件では「無関係」はいずれも高い正解率であったが、判定閾値の件数が多くなるほど「類似」の正解率が下がった。

本調査の手法の是非については、おおむね有効であると考えられる。ふるい分け対象案件のふるい分けでは、例えば学習で判定閾値を上位 200 件とすると、3.71%は間違えて「テーマ内」をテーマ外と判定してしまうものの、12.00%は「テーマ外」を正しく判定できた。また付与漏れ候補案件のふるい分けでは、学習で判定閾値を上位 5 件とすると 1.33%は「類似」を間違えて「付与漏れ」と判定してしまうものの、「無関係」を間違えて判定することはなく、「付与漏れ」の 22.67%を正しく付与漏れであると判定できた。したがって、用途により判定閾値をうまく調整すると、本調査の手法によりふるい分けを行うことは有効であると考えられる。

ただし、いずれの判定閾値を用いればよいかは一意には決められないので、例えば、『「ふるい分け対象案件」では「テーマ内」をテーマ外と判定して除外してしまうことは極力避けたいので、テ

ーマ内を除外してしまう比率を 10%以下に抑えた上でテーマ外を極力除外したい』という場合は、「学習」の「上位 100 件」を用いるといった判断を行う必要がある。一方の「付与漏れ候補案件」については、例えば、『「類似」を間違えて付与漏れとみなしてしまうのはしょうがないとしても、「無関係」を付与漏れとみなしてノイズを増やすのは極力避けたいので、「無関係」は正解率 100%を採用したい』という場合は、「学習」の「上位 5 件」～「上位 200 件」の判定閾値で運用するといった判断を行うことになる。

・今後の課題について

本調査のふるい分けの手法自体は有効であると考えられるものの、今回の調査での判定精度については高精度であるとは言い難く、まだ改善の余地がある。

今回の調査では、2 種類のキーワードテーブルはともに機械的に作成したものであり、人手による選定などは行っていない。しかし、F タームキーワードテーブルを用いた場合の結果と学習キーワードテーブルを用いた場合の結果に差が出ていることから分かるように、キーワードテーブルの中身によって判定精度が変わってくる。したがって、キーワードテーブルの中身の改善を行うことで判定精度を改善させる余地がある。

今後の課題としては、学習データの増強や、人手によるキーワードの選定、シソーラスを活用した同義語・表記ゆれ対応、ストップワードの追加、機械翻訳精度の向上、スコア計算方法の改良、付与漏れ候補案件の判定閾値の改良などがあり、これらにより判定精度の向上が見込まれる。また、コスト面では、学習データの増強は既存の特許文献を増やして学習させるだけなので安価であり、一方、人手によるキーワードの選定は人力の作業なので高価である。シソーラスやストップワードは、既存のものがあれば安価に導入できる上に継続的なコストがかかるものではない。