２Ｈ１３

内容の類似性評価手法を利用した同一特許発明者の特定

○峯尾　翔太，中村　達生，片桐　広貴，大石　宏晶（VALUENEX 株式会社）
富澤　宏之，中山　保夫（文部科学省　科学技術・学術政策研究所）

1. はじめに

効率的な研究開発体制を把握するには発明者に着目した特許情報分析は有用な手段の一つとなる。しかし、信頼性の高い分析を行うには、同姓同名の別人の存在を考慮した同一発明者の特定が必要となる。同一発明者の特定に際しては、氏名や出願人名などの書誌情報の一致性が考慮されることがある。ところがこうした書誌情報の一致性だけでは、手がかりとなる情報が一切含まれない場合に対応できないという問題がある。そこで、書誌情報の一致性の他に、同一発明者による特許に対する内容の類似性が高い特許は同一発明者によるものである可能性を有すると判断するプロセスを加えた同一発明者の特定手法を開発した。本報ではその手法について紹介する。

2. 開発した同一発明者特定手法の全体像

開発した同一発明者特定手法は大きく分けて次の 5 つの工程から成り立っている。

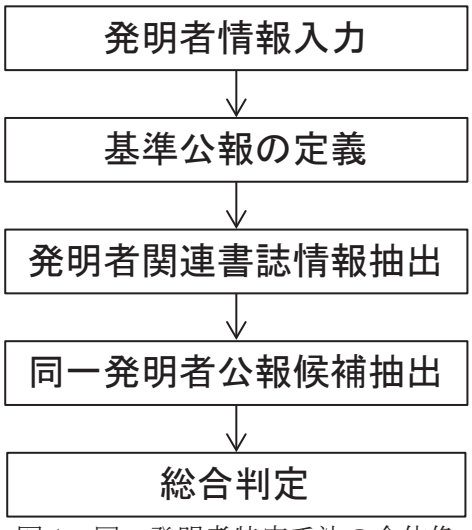

図 1. 同一発明者特定手法の全体像

(1)　発明者情報入力
関与した公開特許公報を特定したい発明者に関する情報を入力する。入力可能な項目としては次のものがある。

・発明者名
・所属機関名
・所属機関の所在地
・発明者住所
・発明者の関与が既知である公報番号

なお、入力後、所属機関名については、文部科学省　科学技術・学術政策研究所が提供する「NISTEP 企業名辞書」、及び国立情報学研究所が提供する「科学研究費助成事業データベース」を利用して、所属機関名の変遷及び所属機関の変遷を補完する。

(2)　基準公報の定義
同一発明者が関与した公報を特定するにあたって、その基準となる公報を定義する。具体的には、(1) において入力された発明者の関与が既知である公報か、発明者名と (1) において入力されたその他の情報が一致する公報のいずれかを発明者が関与したものであるとみなし、基準となる公報とする。

(3)　発明者関連書誌情報抽出
(2) において定義された基準公報から、(1) において入力されたもの以外の発明者関連書誌情報を抽出する。例えば、(1) において発明者住所として「東京都文京区○○」が入力された一方で、(2) において定義された基準公報の発明者住所が「東京都文京区××」であった場合、これを抽出する。なお、抽出された出願人名については、(1) と同様にして所属機関名の変遷及び所属機関の変遷を補完する。

(4)　同一発明者公報候補抽出
(3) において抽出された書誌情報に基づき、同一発明者が関与した公報の候補を抽出する。また、平行して (2) において定義された基準公報に対する内容の類似性に基づき、同一発明者が関与した公報の候

補を抽出する。これにより、発明者が関与したにも関わらず手がかりとなる書誌情報が含まれていない公報を抽出し得るようになる。なお、いずれの手法についても詳細は後述する。

(5) 総合判定
(4) において抽出された同一発明者公報候補について、書誌情報の一致性と内容の類似性を総合的に考慮して、抽出された公報が、同一発明者が関与したものであるか否かの確度を計算する。計算方法については後述する。計算後、(3) に戻り、(4) において新たな候補が抽出されなくなるまで再帰的にこの手順を繰り返す。

3. 書誌情報に基づく同一発明者公報候補抽出

次の 4 項目に基づき同一発明者が関与した公報候補の抽出を行う。

(1) 発明者住所
発明者名が一致する公報の内、2. (1) において入力された発明者住所または 2. (3) において抽出された発明者住所を含む公報を抽出する。

(2) 所属機関名
発明者名が一致する公報の内、2. (1) において入力された所属機関名または 2. (3) において抽出された出願人名を、出願人または発明者住所の一部に含む公報を抽出する。

(3) 共同発明者住所における所属機関名
発明者名が一致する公報の内、2. (1) において入力された所属機関名または 2. (3) において抽出された出願人名を共同発明者住所の一部に含む公報を抽出する。

(4) 発明者住所と所属所在地の近接性
発明者名が一致する公報の内、2. (1) において入力された所属機関の所在地または 2. (3) において抽出された所属機関の所在地と同一市区町村にある発明者住所である公報を抽出する。

4. 内容の類似性に基づく同一発明者公報候補抽出

2. (2) において定義した基準公報に対する内容の類似性が一定以上である公報を同一発明者公報候補として抽出する。

内容の類似性を評価する方法として、VALUENEX 株式会社が独自に開発した文書のクラスタ化手法を用いる。クラスタ化とは、文書を自然言語処理により特徴づけし、相互の類似度を評価する仕組みである。VALUENEX 株式会社の技術の特徴は情報処理が高速であり、解析に用いる語句の数が大量であること、および 2 次元化の際に類似性の高いものを精度よく表現していることにある。

クラスタ化の手順を次に示す。

(1) 形態素解析
形態素解析器を利用し、文書から文章を抜き出し、単語毎に分割する。

(2) 特徴語抽出
各文書内の単語について、tf-idf 法を利用し重み付けを行う。tf-idf 法においては、下記数式に従って、文書 i における単語 j のウエイト（重要性）を算出する。

$$W_{ij} = TF_{ij} \times IDF_j$$

TF : Term Frequency の略。文献 i における単語 j の出現頻度が高ければ高いほど、TF が大きくなる。
IDF : Inverse Document Frequency の略。Document Frequency の逆数。単語 j が多くの文献に出現すればするほど、DF が大きくなり、IDF は小さくなる。
当該文書に多く出現し、かつ他の文書にそれほど出現しない単語はウエイトが高くなる。

(3) 文書間の類似度算出
tf-idf 法により得られた各単語のベクトル に基づき文書ベクトルを生成し、得られた文書ベクトルの正規化を行った後、これらの内積（＝類似度 ）を算出する。

(4) 文書のクラスタ化
文書間の類似度を既定のしきい値に従いクラスタ化を行う。

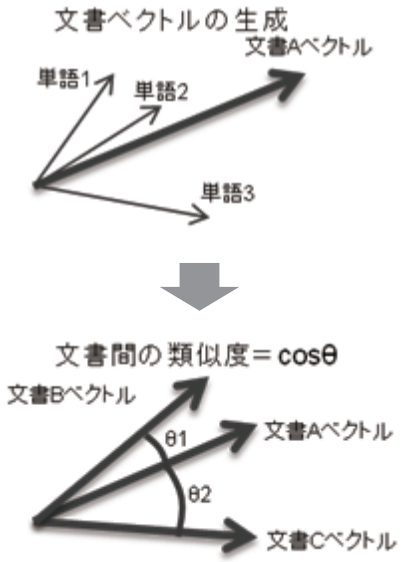

図 2. 文書間の類似度算出

5. 総合判定

同一発明者公報候補について、書誌情報の一致性と内容の類似性を総合的に考慮して、抽出された公報が、同一発明者が関与したものであるか否かの確度を計算する。具体的には、公報 i が、同一発明者が関与したものである確率 P_i を次のように表現する。

$$P_i=1-(1-P_j\times P_{ad})\times(1-P_j\times P_{aj})\times(1-P_j\times P_{ds})\times(1-P_j\times P_{ci})\times(1-P_j\times P_{nh})$$

ここで、P_j は抽出に用いた情報の取得元となる公報 j が対象発明者が関与したものである確率である。2.（1）において入力された情報を抽出に用いた場合には、これを 1 とする。

また、P_{ad}、P_{aj}、P_{ds}、P_{ci}、P_{nh} はいずれも定数であり、次の表における各項目と対応している。なお、「想定値」は、予め同一発明者が関与した公報が全て分かっている発明者について、当該項目に基づき抽出して測定した厳密な値ではない。

表 1　総合判定における判定項目

判定項目	同一発明者関与確率	想定値
発明者住所	P_{ad}	0.95
所属機関名	P_{aj}	0.8
内容の類似性	P_{ds}	0.7
共同発明者住所における所属機関名	P_{ci}	0.3
発明者住所と所属所在地の近接性	P_{nh}	0.3

6. 開発した手法の適用結果

開発した手法の適用結果として、国立大学に所属歴のある研究者を対象とした結果を示す。

表 2　開発した手法の適用結果例

対象発明者	母集団（件）	正解総数（件）	正解抽出数（件）	抽出総数（件）	再現率（%）	適合率（%）
後藤**	10	10	10	10	100.0	100.0
丸山**	25	20	16	16	80.0	100.0
柚木**	10	10	10	10	100.0	100.0
吉原**	18	18	11	11	61.1	100.0
樫木**	11	11	11	11	100.0	100.0
清水**	200	10	8	8	80.0	100.0
池田**	78	20	18	18	90.0	100.0
細見**	30	30	26	26	86.7	100.0
五十嵐**	34	34	26	26	76.5	100.0
永松**	13	13	13	13	100.0	100.0
計	429	176	149	149	84.7	100.0

国立大学に所属歴のある研究者のうち、1993 年から 2014 年 12 月 31 日までに公開された日本国公開特許公報、公表特許公報、再公表特許が 10 件以上である研究者 10 名を無作為に抽出して手法を適用した。なお、上記の表では個人の特定を防ぐため対象発明者名のフルネームでの記載と所属大学の記載を控えている。

正解総数は、対象発明者について特許内容や大学研究室のウェブサイト等の情報に基づき同一発明者のものであると考えられる公報をそれぞれカウントしたものである。また、抽出した結果のうち、総合判定時に出力される同一発明者が関与した確度が 0.6 以上である公報を抽出総数としてカウントした。抽出総数の内、正解に属するものを正解抽出数としてカウントし、再現率と適合率を計算した。

再現率については、10 例中 4 例が 100%であるが、中には 60%程度のものもみられる。これは、内容が一定以上異なる公報が多数ある場合に、内容の類似性に基づく同一発明者公報抽出手法によっても抽出し得なかったことが一因となっている。なお、2.（1）にて述べた発明者情報入力の段階において、内容が一定以上異なる公報群のいずれか 1 つでも入力されていれば、この再現率を向上させられる可能性はある。

一方、適合率は 10 例とも 100%であり、総合判定により出力される確度を利用して誤抽出を抑制できる可能性が示唆されている。

7. まとめと今後の課題

書誌情報の一致性の他に、同一発明者による特許に対する内容の類似性が高い特許は同一発明者によるものである可能性を有すると判断するプロセスを加えた同一発明者の特定手法を開発した。その結果、無作為に抽出した国立大学研究者 10 名について、再現率 84.7%、適合率 100%を記録した。再現率は 10 例中 4 例で 100%であったが、中には 60%程度のものもみられ、特許内容が大きく異なる場合の対応が課題の 1 つとして浮き彫りになった。また、総合判定時に出力される確度の精度向上や、利用する書誌情報の多様化

なども課題として挙げられる。

参考文献

[1]中村 達生,"JICST ファイル・特許 DB を用いた動向分析--データマイニング手法を用いた技術連関分 析",情報管理 Vol.46, No.2 (2003/5) pp. 97～106,科学技術振興事業団情報事業本部
[2]中村 達生,"データマイニング手法を用いたサイエンスと産業技術の連携分析",産業連関 Vol.12, No.2 (2004/6) pp. 50～61,環太平洋産業連関分析学会