研究者同定とその応用
— 統計分野と材料科学分野を例として —

Identification of Researchers and Its application
— In the Case of Statistics and Materials Science field —

高久 雅生 *1　相澤 彰子 *2　馬場 康維 *3　蔵川 圭 *2　谷藤 幹子 *1
Masao Takaku*1　Akiko Aizawa*2　Yasumasa Baba*3　Kei Kurakawa*2　Mikiko Tanifuji*1

*1　物質・材料研究機構
National Institute for Materials Science
*2　国立情報学研究所
National Institute of Informatics
*3　情報・システム研究機構 新領域融合研究センター / 統計数理研究所
Research Organization of Information and Systems / The Institute of Statistical Mathematics

Abstract: As an application for data mining from a large scale database manually built, we have constructed and analyzed academic researcher networks. We have used Kakenhi database as a master database for this purpose. And then we have picked two dataset as a first target: One is researchers from members of three academic societies in Statistics related field; and the other is from members of National Institute for Materials Science. An analysis and comparison on these types of researcher networks shows their different characteristics.

1　はじめに

通常のサーチエンジンのように検索対象をサイトやページに限定せず、Web 上の大量の情報を整理してよりアクセスしやすくなるように、実世界に存在する「実体」そのものを取り扱うエンティティ検索が提案されている [1]。たとえば、このような検索を行うサーチエンジンとして、商品やその価格を対象とする Google Product Search[2] や、人物を対象とする Spock[3] などがサービスをはじめている。エンティティ検索を行うためには、単にページやサイトを特定するだけでなく、エンティティの特定や同定をも行う必要がある。エンティティとしては様々な事項を対象としうるが、その対象領域毎に領域固有の知識を獲得したり、エンティティの同定手法を確立したりする必要がある。著者らはこの領域として「人物」を対象とし、とりわけ学術分野で活動する研究者を対象としたエンティティ検索のための手法の研究開発に取り組んでいる。以下では、学術分野の研究者検索のための人物同定の取組みについて述べる。

人物エンティティ検索のためには、まずその参照点となる人物情報の同定が必要となる。これには、あらかじめ作成した人名 ID 付きリストをもとに同定を行う手法と、クラスタリング等により半自動で同定を行う手法とに分けられる。学術研究者の場合、あらかじめ人手により整備された人物データベース（後述）があるため、これを用いる前者の手法を採用する。なお、ここで重要な点は、ある特定の人物の氏名はその人物を特定するに足りないものであるという点にある。つまり、同姓同名の存在があるために、人名だけでは一意に人を特定することが難しい点にある。そこで、人手でこれらの同姓同名の関連を解決するため、氏名以外をキーとして一意に特定できるように構築したデータベースを用い、このデータベースを参照点として用いることとする。

本研究では、日本の学術研究者の同定に国立情報学研究所がサービス提供している科研費データベース [4] を用い、科研費研究者番号をもとに人物同定を行った。対象として、1）統計分野の研究者、2）物質・材料研究機構所属の研究者の 2 つのデータセットを対象に同定作業を行い、その事例を中心に報告する。同定にともなって得られた結果を報告する。また、科研費データベースに研究者間の共同研究の関係性が記録されている性質を利用して、同定結果をもとに科研費データベース上での各研究者コミュニティの可視化例を示す。

2　科研費データベース

本研究では、科研費データベースの全レコードに含まれる研究者番号をもとに研究者同定を行った。科研費データベースは、採択課題情報、研究実績報告書、研究成果報告書の 3 点のデータからなるが、採択課題情報からは研究者番号を得られなかったため、各年度終了時に提出される研究実績報告書と、研究課題終了後に提出される研究成果報告書とに記載されている情報を元にした。なお、科研費データベースは報告書提出後にデータ更新が随時行われるが、本研究では、2007 年 12 月時点の科研費データベースを取得し、これを元に同定および分析を行った。

本稿で対象として扱うデータは科研費データベースから取り出した情報の一部であり、含まれる研究者番号の異なり数は 154,400 名分であった。これは、採択課題の研究代表者だけでなく、その研究分担者として研究者番号が付与されている者を含む人数である。なお、これは主に 1985 年から 2006 年までの総数 312,460 件の科研費研究課題の情報に基づくものであり、この間に科研費補助金の採択を一度も受けなかった研究者は含まれていない。

科研費データベースの特徴の一つは、科研費研究者番号と呼ばれる個々の研究者に一意な番号が付与されている点にある。研究者番号は一研究者につき 8 桁の番号が付与されることとなっており、氏名の変更、所属の変更等を含む個人同定のための情報が変わったとしても、原則として同一の番号を使用する運用がされているため、この研究者番号を用いることによって同一人物としての判定が容易におこなえる。ただし、科研費データベースでは所属機関を異動した際に誤って研究者番号を再取得してしまった場合や、報告書に含まれる誤記によって一人の研究者が複数の番号を持っている場合があるため、実際の同定作業においてはこれをも考慮して対応を行った。

3　研究者同定

研究者同定の作業は、研究者データセット中の氏名を、科研費データベースに記録されている氏名表記に対し検索し、一致のあったペアを出力し、それらの間の関係を人手で判定することにより行った。なお、これにより氏名表記が一致しない研究者ペアは同定対象人物がいなかったものとして扱った。

同定作業は、おおまかにいえば識別 ID（ここでは科研費研究者番号を指す）が付与されていない氏名と所属のみが記載された複数のリストに対して、ID を付与する作業であった。この際、複数の名簿により構成されている場合は、複数名簿間での同一人物の同定を行う必要がある。同定時には、研究者番号を割り振る同定作業と重複候補の統合作業とを並行的に行った。

4　統計分野データセット

統計分野研究者としては、日本統計学会（JSS）、日本行動計量学会（BSJ）、日本計算機統計学会（JSCS）の 3 学会の会員名簿に記載された人名集合を統計分野データセット [5] と呼び、このデータセット中の人名の同定処理について述べる。なお、会員名簿はそれぞれ、日本統計学会会員名簿（2003 年版）、日本行動計量学会会員名簿（2005 年版）、日本計算機統計学会会員名簿（2002 年版）による。

表 1: 統計分野データセット

学会	人数	科研費データベースと一致	
		氏名表記	同定研究者
JSS	1,545	697 (45.1%)	619 (40.1%)
BSJ	1,070	519 (48.5%)	468 (43.7%)
JSCS	416	184 (44.2%)	168 (40.4%)
（計）	3,031	1,400 (46%)	1,008

表 1 に各学会名簿に記載された研究者数を示す。

上記データセット中の氏名表記を科研費データベースの登録者とマッチングし、一致する氏名表記を同定候補として抽出した。これは延べ 1,400 件となった。それぞれ、日本統計学会 697 件 (45%)、日本行動計量学会 519 件 (49%)、日本計算機統計学会 184 件 (44%) であり、おおよそ各学会会員の 45–50%は、表記上で科研費データベースの研究者番号に対する候補となった。さらに、このうちから所属・氏名表記の両者が完全に同一の者は重複所属の研究者として除去し、延べ 1,307 名からなる統計分野研究者の人名リスト（同定候補集合）が得られた。

4.1　同定結果

次に、同定候補を人手で判定する。同定結果は、複数学会に所属しているもの（重複）、同一人物が見つかったもの（正例）、同一人物が見つからなかったもの（負例）の 3 種に大別できる。統計分野データセットの 1,307 候補に対して、266 候補（19%）は重複、1008 候補（72%）は対応する研究者番号が 1 つ以上見つかった候補で、126 候補（9%）は対応する研究者番号が見つからなかった候補であった (図 1)。図 1 は、もとも

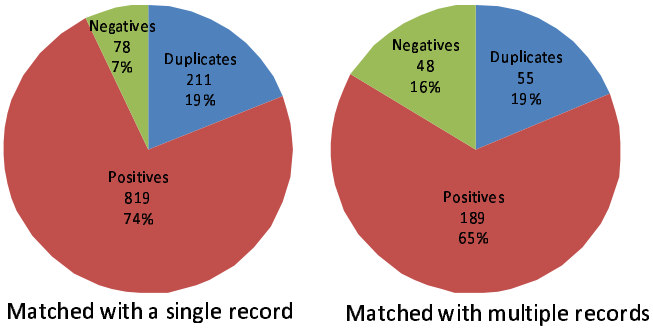

図 1: 単独一致候補（左）と複数一致候補（右）に対する人手判定結果の内訳

との氏名表記において複数の研究者番号とマッチしたものとそうでないものの内訳を示している。単独一致と複数一致の双方で、重複候補の割合はそれほど変わらないが、正例を持つ候補の割合は重複分を除くと、単独一致の候補で 90.8%、複数一致の候補で 79.7%となっている。複数一致候補でやや下がっており、重複案件において一致する候補氏名がありふれたものであるケースもあり、それが負例候補を増やすことにつながっている可能性が示唆される。

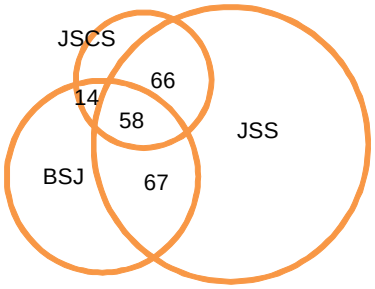

図 2: 複数学会に所属している研究者の人数（JSS: 日本統計学会、BSJ: 日本行動計量学会、JSCS: 日本計算機統計学会）

複数学会に所属している研究者の人数を図 2 に示す。3 学会全てに所属している研究者は 58 名、日本統計学会と日本計算機統計学会の両者に所属している研究者は 66 名、日本行動計量学会と日本統計学会に所属している研究者が 67 名、日本行動計量学会と日本計算機統計学会に所属している研究者が 58 名であった。

5　物質・材料研究機構所属研究者（NIMS）データセット

材料科学分野の研究者として、筆頭著者の所属する独立行政法人物質・材料研究機構（以下、NIMS）に所属する研究者の集合を取り上げる。NIMS 所属研究者の名簿としては、正確を期すために所内人事データベースにおいて運用されているものを使用し、2009 年 4 月 28 日時点に抽出したデータを対象とした。なお、元とした人事データベースには 2,127 名の情報が含まれているが、これには事務職員等の研究職以外の者も含まれるため、前処理として、所内の職制コードを参考に、事務職、事務業務員、研究業務員、研修生、派遣職員を除外したリストを作成し、1,229 名の人名リストを得た。これを NIMS データセットにおける同定候補集合とする。

表 2: NIMS データセット

所属	人数	科研費データベースとの一致	
		氏名表記	同定研究者
NIMS	1,229	348 (28.3%)	256 (20.8%)

この 1,229 名分の同定候補集合に対し、前節で述べた統計分野データセットと同様の同定手順で同定を行う。まず、科研費データベースに対して氏名照合したところ、氏名が一致するレコードは 348 名分あった。さらに 348 名分の同定候補について人手で同一人物判定を行ったところ、256 名において科研費データベース中の研究者番号を持つレコードと同一人物だと判定できた（表 2）。

なお、NIMS データセットが統計分野データセットに比べて科研費データベースとの照合率が低いのは、NIMS の法人化以前の前身機関である、金属材料技術研究所と無機材質研究所において科研費補助金への申請が行なわれていなかったために、旧国立研究所時代からの所属研究者の科研費データベースへの収録は 2001 年の法人化以降のこととなっているためである。

6　考察

6.1　同定の判断基準

同定をおこなう際にはなんらかの情報を基準にする必要がある。今回の同定作業においては、同定対象データセットに氏名と所属の情報が含まれており、かつ主題が当該分野であるとの条件があったため、科研費データベース上の科研費課題報告書に含まれる「氏名」「研究分野」「研究課題タイトル」「所属」などの情報に加え、複数の研究課題の記録が残っている者に関しては時系列的にどのような研究を行ってきたか、所属がど

のように移動したかの情報も得られるため、これらの情報を手掛かりに同定判定を行った。たとえば、氏名が同一であり、データセット中の所属と研究課題に記載された所属も一致し、かつ、研究課題が当該分野であると判断できた場合には、容易に同定を行うことができる。一方で、上記 3 つの情報（氏名・所属・研究分野）が合致しなかったり、採択課題が 1 つしか無いような場合には、同定に関する判断をつけるのは難しくなる。そこで、科研費データベース以外の一般の Web 空間での検索などをもとに、JST ReaD 研究者ディレクトリ [6] や、判定対象人物が所属する大学などが提供している研究者紹介ページ、当該人物が執筆した著書のプロフィール、本人が自身のウェブサイトで提供している経歴情報など、様々な情報をもとに判断を加えることになる。たとえば、所属が移動していると思われる場合は、氏名とともに移動前・移動後の所属先の組織名を Web 検索し、移動したことを確認した。またたとえば、同姓同名の別人であると思われる場合には、JST ReaD などのデータベース上で別々のエントリとなっているか確認したり、所属履歴上の年月などをもとに別の組織に所属していることを確認したりした。同定作業において、当該氏名表記での同定対象候補が 1 つしかない場合は比較的、判定が楽であったが、候補が複数ある場合、とりわけその所属先が一致しない場合には、確認作業を行う必要がある。これには時間がかかるだけでなく、その根拠を確定させること自体が困難であった。今回の同定作業では主に Web 上の情報源を用いたが、Web 上の情報だけでは人物情報の明確な判定には限界があり、判定根拠が得られない場合にどのような対応を取るかを考える必要がある。この問題は同定結果をどのように利用するかを考えた上で、その用途に合わせて考慮しなければならない問題であろう。また、判定時間としては、容易に判別可能なケースでは 1 人あたり 0.5〜1 分程度だったが、候補が多かったり所属を確認しなければならないケースでは 10〜30 分程度の時間を要することもあった。今回の作業においては時間計測を行わなかったため正確な所要時間は不明だが、大規模な同定作業を行う場合にはあらかじめ作業に要する時間を概算し、見積る必要が出てくるものと思われる。前述の判定基準や使える情報源等と合わせ、時間見積りについても考慮を要する。

6.2 同姓同名研究者

同定作業において判定が困難なものは同姓同名の研究者がある場合のことが多い。とりわけ、類似の研究分野に同姓同名者がいた場合、その判別を行うためには同姓同名の別人であることを確認する必要が出てくるため、その人物同定にはコストがかかる。

たとえば、統計分野データセットの場合、同定作業の過程において複数研究者番号と一致した同定候補に対して、科研費データベース上で同姓同名となる複数の研究者番号が存在したものは、105 候補であった。また単独一致候補のうち負例候補となったものに対しても、科研費データベース中には統計分野データセット中の研究者とは別人の同姓同名者がいたこととなると考えられ、このケースは 126 候補で同姓同名者の存在を確認した。これら 2 つのケースを合わせると、1134 候補中 231 件（20.4%）となり、比較的多くの場合に同姓同名者との照合を行う必要があることが判明した。

これらの同姓同名者の照合のコストについては、所属機関名での一致を見るなどのヒューリスティクスを用いて、労力軽減を図ることも有用であると思われる [7]。

6.3 同定結果の可視化

筆者はこれまで科研費データベースに含まれる研究者代表者・分担者の情報をもとにした研究者の共同研究ネットワークの可視化、分析を進めてきた [8]。今回のデータセットは統計分野や特定の機関所属者に限定したものであり、当該分野の研究者コミュニティの特徴をとらえられる可能性がある。

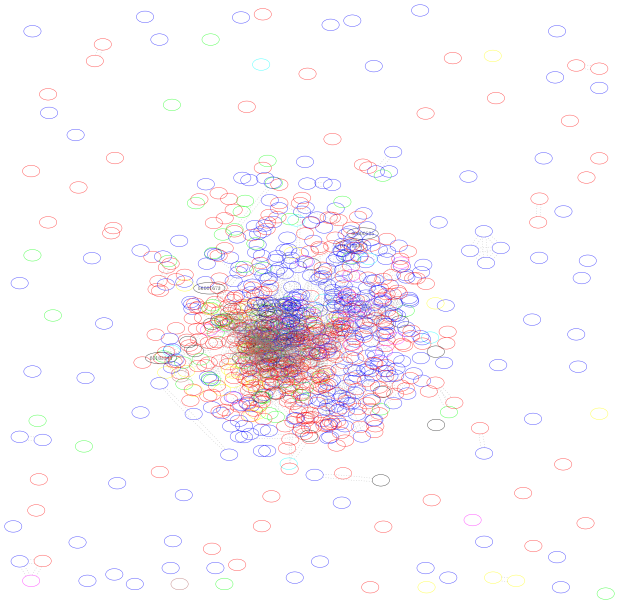

図 3: 統計分野データセットにおける研究者共同研究ネットワーク

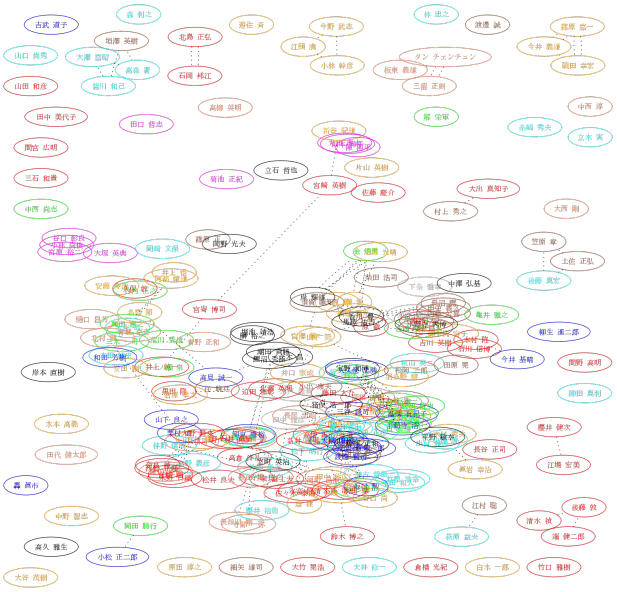

図 4: NIMS データセットにおける研究者共同研究ネットワーク

研究者間の関係は、研究者をノードとして過去に研究者間で共同研究を行ったことがある研究者同士をエッジで結ぶようなネットワークとして描くことができる。グラフレイアウトツール GraphViz[9] を用いて描いた可視化図を以下に示す。図 3 は、統計分野データセットにおける共同研究ネットワークであり、データセット内の研究者から 1 ホップ先の共同研究者との関連までを考慮して共同研究回数の多い者同士が近傍になるよう配置している。図 4 は NIMS データセットにおける共同研究ネットワークであり、データセット内の研究者から 2 ホップ先の共同研究者まで考慮して、同様に描画したものである。

また、各共同研究ネットワークにおけるノードの色はその研究者の所属を表現している。統計分野データセット（図 3）では、日本統計学会は赤色、日本行動計量学会所属者は青色、日本計算機統計学会は緑色として描画している。NIMS データセット（図 4）におけるノードの色は研究所内の研究領域を表現し、ナノテクノロジー基盤領域は赤色、ナノスケール物質領域は緑色、情報通信材料研究領域は青色、生体材料研究領域は赤紫色、環境・エネルギー材料領域は空色、材料信頼性領域は茶色、萌芽ラボはオレンジ色、共用基盤部門は灰色、MANA/ICYS はサンゴ色、その他は黒色として描画している。

著者らのこれまでの分析 [8] から、一般に科研費における共同研究ネットワークは密に広い範囲で連結していることが分かっているが、統計分野のような特定の分野に限って見た場合でも密なネットワークが観察できた。

このようなネットワーク可視化には、2 つの利用可能性が考えられる。

1 つは同定作業そのものへの貢献である。あとで述べるように、同定作業そのものは難しい作業の確認を除けば、半自動化できると考えられるが、この場合、研究者間のつながりとしてのネットワークの特徴を、同定を自動化するための特徴の一つとすることができるのではないか。特に密になったネットワーク内に配される研究者と、それ以外とでは研究者の位置付けがやや異なると考えられ、中心に配されるような研究者は統計分野の多くの研究者とつながっており、当該研究者も当該の分野や機関に属する者である、という仮説が考えられる。いったん同定候補を得た後に、そのネットワーク中心性指標や接続性指標など当該研究者のネットワーク特徴量を計算して、同定作業に役立てることが可能ではないか。この仮説については現在確認できていないため、この点については今後の検討していく予定である。

2 番目の利用可能性は、分野構成の理解という観点である。こういった特定分野上のエンティティを可視化することにより、当該の分野や機関のキープレイヤーを可視化することができるものと考えられる。これにたとえば、所属組織、学会、投稿論文誌といった様々な視点からの情報を加えれば、当該の分野や機関を一目で概観し、その理解を促進できる可能性が考えられる [10]。

6.4 同定自動化に向けて

4.1 章で述べたように、同定候補対象のうち大半は正例となる候補を持っており、重複分を除けば、その割合は 9 割程度になる。これらの結果から、同定作業のうち一部については自動化することができるとおもわれる。特に、所属組織名と氏名が合致し、かつ、他に同定対象候補が存在しないケースでは、そのまま正例としてしまったとしても、判定誤りはほとんど含まれない。また、複数学会に所属している重複分の判定についても、大半は同一人物のものであり、同姓同名の別人であったケースは、判定済のサブセット中では 1 件のみであった。同定作業においては、単に人名リストと研究者番号を結び付けるだけではなく、別人である旨も判定しているので、これらの負例情報も用いれば、一定の教師付き学習の枠組みで自動化することも

考えられる。これらの同定自動化についても今後、検討を進めていく予定である。

7 おわりに

本稿では、統計分野データセットおよびNIMSデータセットを対象に研究者番号を付与するような同定作業と、そこから得られた知見を報告した。

同定の自動化と得られた結果の応用が今後の課題として残されたが、ほかにもこれらの同定結果をもとに、Webページや Webサイトの関連を同定していったり、研究者検索に同定結果を用いるなどの応用も考えられ、今後取り組んでいく予定である。

参考文献

[1] 原田昌紀; 佐藤進也; 風間一洋: 「Web 上のキーパーソンの発見と関係の可視化」, 情報処理学会研究会報告 DBS-130/FI-71, pp. 17–24, 2003.

[2] “Google Product Search”. http://www.google.com/products (2009-05-14 accessed).

[3] “Spock”. http://www.spock.com/ (2009-05-14 accessed).

[4] 国立情報学研究所: 「KAKEN: 科学研究費補助金データベース」. http://kaken.nii.ac.jp (2009-05-14 accessed).

[5] 馬場康維; 坂口尚文: 「マッチングによる共通メンバー数の推定」, 日本計算機統計学会 第 21 回シンポジウム, 2007.

[6] 科学技術振興機構: 「ReaD 研究開発支援総合ディレクトリ」. http://read.jst.go.jp (2009-05-14 accessed).

[7] 蔵川圭; 武田英明; 高久雅生; 相澤彰子: 「研究者リゾルバー α のコンセプト」, ディジタル図書館ワークショップ, No. 36, pp. 15–21, 2009.

[8] 高久雅生; 相澤彰子; 大山敬三: 「研究者情報サーバの構築: ネットワーク構造可視化と解析の試み」, 「シンボリック・データ解析と周辺技法」研究会, pp. 35–41, 東京, 統計数理研究所, 2007.

[9] “GraphViz”. http://www.graphviz.org/ (2009-05-14 accessed).

[10] Hook, Peter A.; Börner, Katy: “Educational knowledge domain visualizations: tools to navigate, understand, and internalize the structure of scholarly knowledge and expertise”. “*New Directions in Cognitive Information Retrieval*”. pp. 187–208. Springer, 2005.

連絡先

高久雅生（物質・材料研究機構）

E-mail: TAKAKU.Masao@nims.go.jp