統計分野における研究者の氏名同定とその応用

高久雅生 † 相澤彰子 †† 大山敬三 †† 馬場康維 †††
† 情報・システム研究機構 新領域融合研究センター
†† 国立情報学研究所 コンテンツ科学研究系 / 総合研究大学院大学
††† 統計数理研究所 データ科学研究系
{masao, aizawa, oyama}@nii.ac.jp
baba@ism.ac.jp

概要

統計分野 3 学会の会員名簿延べ約 3000 件をもとに、統計分野の研究者データセットを作成した。データセットには科研費データベースにもとづく研究者番号を付与する同定作業を行った。会員名簿の氏名表記と科研費データベース中の研究者氏名とをマッチさせたところ 45–50% 程度の一致があり、これを同定対象候補として用いた。同定候補中 823 件までの人手判定結果によれば、2 割程度は複数学会に所属する研究者と判定された。重複を除いた同定対象候補のうち 9 割程度は、少なくともひとつは対応する研究者番号を持つものであると判定できた。複数の同定候補を持つ、なんらかの同姓同名の別人がいる者は全体の 15% 程度であった。また、同定結果の簡易な可視化として共同研究ネットワークによる表現を示し、これらのデータセットをもとにして統計分野領域の全体像を描くような知識ドメインが描ける可能性が示唆された。今後の課題としては、統計分野の特徴を示して分野の概観を可能にすること、および、同定作業そのものを自動化してより大規模な統計以外の領域の研究者同定にも適用可能な手法を開発することなどがある。

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「事項」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[*1]や、人物を対象とする Spock[*2]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における人物同定のための取組みについて述べる。

先に述べたように、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した、つまり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。

[*1] http://www.google.com/products
[*2] http://www.spock.com/

本研究では、日本の学術研究者の同定に科研費データベース[*3]を用い、科研費研究者番号をもとに人物同定を行った。実際に統計分野の研究者を対象に同定作業を行い、その事例を中心に報告するとともに、同定結果から得られたデータセットの応用可能性についても考察する。

2　統計分野研究者データセット

以下では、日本統計学会、日本行動計量学会、日本計算機統計学会の 3 学会の会員名簿に記載された人名集合を統計分野研究者データセット [2] 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　統計分野研究者データセット

学会	人数
日本統計学会 （2003 年）	1,545
日本行動計量学会 （2005 年）	1,070
日本計算機統計学会 （2002 年）	416

表 1 に統計分野研究者データセット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研究者情報サーバ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 1,399 件となった。それぞれ、日本統計学会 696 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。

以下では、この同定候補集合に対して科研費研究者番号を付与する作業を同定と呼ぶ。

[*3] 国立情報学研究所がサービス提供している。 http://seika.nii.ac.jp/

3　研究者同定

研究者同定では、研究者データセット中の氏名を、研究者情報サーバ上の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ため、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。

簡易な人手判定ツールを作成し、これを使って同定作業を行った（図 1)。

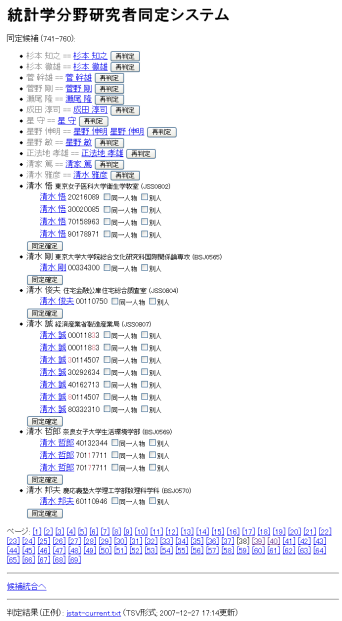

図 1　同定ツールのインタフェース

同定ツール上では、統計分野研究者データセットに含まれる研究者の氏名・所属が一ページあたり 20 名分のリスト形式で表示される。各研究者の情報の下に、科研費報告書等で氏名表記が一致した研

究者番号と研究者情報サーバ [3] の当該研究者情報へのリンクが示されている。氏名一致では複数の研究者番号と一致するものもあるため、複数を並べている表示する。各研究者番号との対応関係を人手判定したあとで記録していくため、各研究者番号のリンクの横に、その研究者で示される研究者と当該研究者が同一人物であるか否かを示すチェックボックスが用意してある。同一人物または別人のいずれかにチェックをいれて「判定確定」ボタンを押すと、判定結果が保存される仕組みとなっている。

データセットそのものが複数学会名簿をあわせたものであることから、複数学会に所属している研究者を含んでいる場合があるため、その重複判定を行う機能も持つ。

4　同定結果

本稿執筆時点で、統計分野研究者データセットのうち 823 件分（約 63%）の人手判定を完了しているため、以下ではこのサブセットの判定結果について述べる。

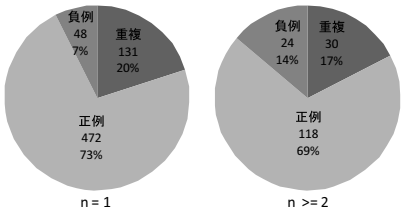

図 2　人手による判定結果 - 単独一致と複数一致候補の内訳

人手判定の結果、823 候補はそれぞれ、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。161 候補（19%）は重複、590 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、72 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 2)。図 2 は、もとも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 83.1% となっている。複数一致候補でやや下がっているものの、概ね同様の比率を示している。

また上記の複数研究者番号と一致した候補において、研究者情報サーバ上で同姓同名が見つかったものは、82 候補であった。また単独一致候補のうち負例候補についても同姓同名との扱いを取るとすると、823 候補中 130 候補（15.8%）で同姓同名となっていることとなる。

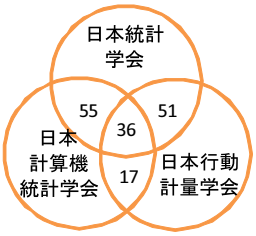

図 3　複数学会に所属している研究者の人数

複数学会に所属している研究者の人数を図 3 に示す。3 学会全てに所属している研究者は 36 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 55 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 51 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 17 名であった。

5　考察

5.1　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、かつ主題が統計分野であるとの条件があったため、研究者情報サーバ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、研究課題が統計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、研究者情報サーバ以外の一般の Web 空間での検索などをもとに、JST ReaD 研究者ディレクトリ[*4]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。

[*4] http://read.jst.go.jp/

5.2 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 [4]。今回のデータセットは統計分野に限ったものであり、統計分野の特徴を可視化できる可能性がある。

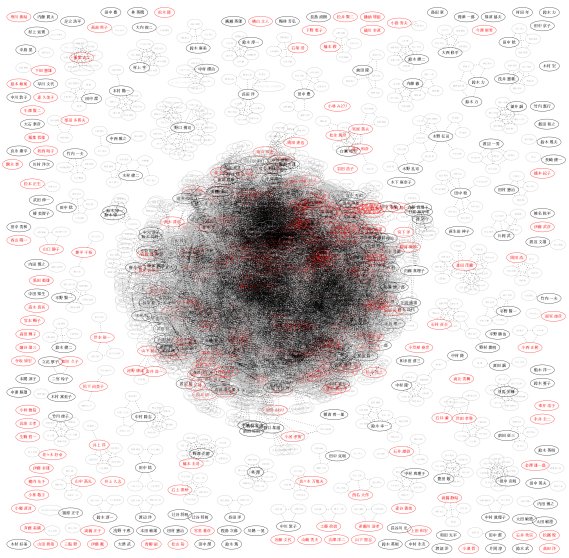

図 4　統計分野研究者の共同研究ネットワーク

図 4 は、本稿で対象とした統計分野研究者データベースを、研究者をノードとし、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いてみたものである[*5]。ここでノードの色や大きさは、その研究者の状態を表現している。赤い円で示されたノードは統計分野データセット内の同定済研究者を、黒い円で示されたノードは未同定の研究者を、グレーの小さな円で示されたノードは対象研究者集合から 1 ホップ先の研究者で候補集合外の研究者であることを示

[*5] 2007 年の報告 [4] と同様、対象期間は 1989 年度から 2003 年度まで約 15 年分の科研費実績報告書と科研費成果報告書からなる

している。これまでの分析から、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、当該研究者も統計分野に属する者であるという可能性が高いのではないかと思われる。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、統計分野のキープレイヤーなど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、統計分野を一目で概観し、その理解を促進できる [5]。

5.3 同定自動化に向けて

4 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。

6 おわりに

本稿では、統計分野 3 学会の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。

同定の自動化と可視化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、研究者を検索する際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。

参照文献

[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] 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.

[3] 高久雅生, 相澤彰子, 大山敬三. 科研費データベースにもとづく研究者情報ブラウジングツール. 「大規模データ・リンケージ、データマイニングと統計手法」研究会, pp. 89–96, 東京, 統計数理研究所, 2006.

[4] 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.

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

Identification and application of researcher's names in the fields of statistics

Masao Takaku[†] Akiko Aizawa[††] Keizo Oyama[††] Yasumasa Baba[†††]
[†]Research Organization of Information and Systems
[††]National Institute of Informatics / The Graduate University of Advanced Studies
[†††]The Institute of Statistical Mathematics

Abstract

The authors have constructed a dataset containing researchers in the fields of statistics, which is derived from members' lists of three Japanese academic societies. Researchers in these lists are identified with *Kakenhi* Database. About 45–50% of researchers' name in the lists match with the names in Kakenhi DB. As of 823 records of the dataset, about 20% are identified as a member of multiple societies. Except for duplicate records, over 90% of the dataset have one or more corresponding ID(s). About 15% of the dataset match to multiple researchers in Kakenhi DB. The authors also make a visualization of the identification result of the dataset, which shows joint research networks among researchers. Such kind of visualization based on the dataset might help a user to understand overview of the fields of statistics. Effective visualization methods and automated methodology for the identification of researchers remains as future works.