

「情報リンケージ」が生み出す新たな価値

同一の人やものに関する情報を集めて一元管理する「情報リンケージ」を利用して、情報に新たな価値を見出そうとする研究者たち。「情報リンケージ」というテーマで集まった彼らは、どのように価値ある成果を生み出そうとしているのだろうか。



Akiko Aizawa
相澤彰子
コンテンツ科学研究系教授

一般には「名寄せ」という名で知られる「情報リンケージ」(*1)。「目立ちませんが、じつは非常に重要な操作です」と語るのは、国立情報学研究所新領域融合プロジェクト(*2)で「大規模リンケージ情報の収集・分析手法」グループの代表を務める相澤彰子教授。もともと「情報リンケージ」を用いることになったきっかけは、「論文のデータベースや図書のオンラインカタログなどの重複登録をなくすこと(クリーニング)でした」という。

たとえば論文の著者に同じ名前があった場合、それが重複なのか同姓同名の別人なのかは、データベースを利用する側にとって重大な問題である。情報リンケージによってデータベースの質を高めることは、データを使用するときに大きなメリットとなる(次ページ図)。

コンピューターの普及とネットワークの巨大化により情報の量は増え続け、しかも手に負えないほど乱雑にばらまかれている。情報を必要に応じて整理整頓する情報リンケージが必要とされる場面は、それこそ無数にある。

情報リンケージとデータベース

一口に情報リンケージといってもさまざまなパターンが考えられる。相澤教授は「難易度から問題設定は4段階に分けられます」という。

まずは先の例で示した、1つのデータベース内での操作である。データベースは、ある目的のために項目や形式を整えたデータの集まりである。だから、各データを相互に照らし合わせて同一のデータかどうかを特定する(情報の同定)といった操作も比較的容易である。次は、2つあるいはそれ以上のデータベース間での統合である。図らずも情報リンケージ(名

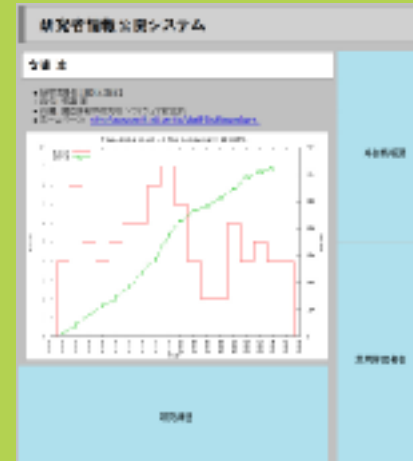
寄せ)が有名になった年金納付記録の照合は、各所がもつデータベースの間で個人情報の同定を行う。そして、同一人物と特定できたものを統合する。そのようにしてデータベース全体を統合していく。

ここまでは形式が整ったデータベースでの操作である。しかしデータベース以外にも、ウェブ上などにはたくさんの有用な情報がある。まったく整理されていないこれらの情報をいかに取り入れていくかが、残る課題である。つまり、データベース内とデータベース外との情報リンケージが第3段階、完全にデータベース外での操作が第4段階となる。

具体的なサービスに

この情報リンケージだけでも十分価値があるようにみえるが、プロジェクトの目的は、情報リンケージによって情報を集め、そのうえで新たに生み出される“何か”を有効利用することである。この“何か”を具体的な形にすることを主目的に共同研究を行っているのが、学術コンテンツサービス研究開発センター(センター長・武田英明教授)の蔵川 圭特任准教授である。

「学術をテーマにしたコンテンツサービスの開発を目的に、まず国内の研究者の名寄せを行っています。これは、たとえば共同研究者を探すときや、あるいは研究者の評価などに役立てられるでしょう」と蔵川准教授。文部科学省にある科研費の研究者番号データベースや、各大学の研究者総覧から名寄せを行っており、これは第2段階を具体的に形にしたものといえる。
高久雅生・新領域融合研究センター研究員は、



さまざまなソースからの情報を集めて作成した「研究者情報公開システム」の例。一覧表、グラフ、図などで表示される。左のグラフは参加プロジェクト数(赤の棒グラフ)と成果リスト登録件数(緑の折れ線グラフ)。右の地図は共同研究者(上位100名まで)の分布。安達 淳教授は、NII学術基盤推進部の部長で、新領域融合プロジェクトのメンバーである。

Google Maps

同じく科研費のデータベースを中心に、人どうしのつながりに焦点を合わせて再編成する研究を行っている。共同研究者や、研究代表者と分担者の関係などに着目することで、「たとえば、だれをボスにすると研究費が獲得しやすいか、なんてこともわかるかもしれません」と高久研究員。情報リンケージの仕方によって、得られる情報はいろいろ違ってくるようだ。

相澤研究グループでは、海外からのインターン学生も活躍している。Dang Bac Vanさんはベトナム国家大学ホーチミン市校の修士課程の学生で、テキストデータからの情報抽出をテーマに研究している。

これは応用範囲が広い魅力的な研究テーマである。たとえば論文に書かれた手法、ツールなどの名称や関係を抽出することで、そこから他分野への広がり期待できる。現在は、いかに人手によるコストをかけずにテキストデータを抽出するかというテーマに力を注いでいる。

「情報リンケージ」に集まった仲間

相澤教授を中心とする研究グループの面々は、みなはじめから情報リンケージ研究を行っていたわけではなく、共通点も少ない。相澤教授は通信から始まり、機械学習、最適化、テキスト・自然言語処理など。蔵川准教授は、機械における設計学、ソフトウェアの設計など。相澤教授との共通点は同じデータに触れていることだけ。そんな研究者たちが集まっている。

また、日ごろの研究風景も、同じ部屋で毎日顔を突き合わせている理系の一般的な研究

室のイメージとは違い、部屋は全員別々で必要なときに集まり議論する。Vanさんは学生であるため定期的に相澤教授とのディスカッションを行い、「教授からはアイデアをたくさんもらい、充実した研究ができています」というが、基本的には必要なときにのみである。しかし集まったときには、「みな異なるバックグラウンドをもっているのだから、議論するたびに新たな刺激を受け、とても楽しく研究をしています」(相澤教授)と団結力は強い。まさに「情報リンケージ(名寄せ)」に“名寄せ”られた研究グループなのだ。

より使いやすく、より価値のあるものを

「現在、第2段階までは問題なく情報リンケージができるシステムを作り出しています。ただ、人手による検証が必要ですので、どの程度の品質でどの程度のコストをかけて行うかという問題になってきます」と相澤教授。研究費が特定の個人に集中する問題や年金など、データベースを共有することで解決する問題は多い。あとは運用方針次第なのだろう。

今後の方針について、「情報リンケージによって、いままでわからなかったことが断片的に捉えられるところまでできています。次は、その断片的な情報の価値を判断して、本当に役に立つ情報を社会に提供できるようにしていく予定です」と相澤教授。一方で教授は、新しいサービスのアイデアも練り始めている。

「情報リンケージ」に集まった研究者たちがどんな新しい価値を生み出していくのか、今後が楽しみである。

(取材・構成 吉戸智明)



Masao Takaku
高久雅生
新領域融合研究センター研究員



Dang Bac Van
国際インターンシッププログラムによるインターン学生