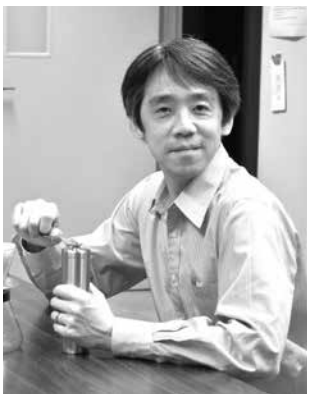
ビッグデータの先を見る

数理・計算科学専攻
脇田 建 研究室



脇田 建 准教授　1965年東京都生まれ。東京大学大学院理学系研究科情報科学専攻にて博士（理学）を取得。東京工業大学理学部情報科学科助手を経て、同大学大学院情報理工学研究科准教授。

近年、通信技術の発達によりビッグデータと呼ばれるものが登場してきた。ビッグデータは多方面から注目されており、研究方面では情報分野に限らず他の分野でも利用されていたり、企業ではビジネスなどに活用されたりしている。本稿では、脇田研究室で行われている研究の中から、ビッグデータに関する研究を取り上げて紹介していく。

世界にあふれるビッグデータ

現代に生きる私たちの身の回りにはたくさんの情報があふれている。それらを手に入れる手段はここ数十年で急速に発達してきた。例えばパソコンやスマートフォンを用いてウェブサイトを閲覧したり、メールやSNSで他者と交流したりすることで、さまざまな情報を得ることができる。情報通信技術の発達により私たちの身の回りにある情報は膨大な量になった。これらはビッグデータと呼ばれている。

ビッグデータとは、その名の通り大容量のデジタルデータのことである。ビッグデータにはいくつかの特徴があり、当然データの容量の大きさもそのひとつに挙げられ、また世界中のデータは日々増え続けている。ほかにも、データの形式が多いという特徴がある。例えば、数値データを整理する場合でも、文書ファイルに羅列されているものと表計算ソフトで表にまとめられているものでは形式が異なっている。このようなものだけでなく、音声や動画などのさまざまな種類があり、またそれぞれの形式も異なっている。

現在では、さまざまな企業がビジネスの一環としてビッグデータを活用している。例えば検索エンジンでは、検索ワードに関連した広告を表示している。ほかにも、検索ワードから景気の動向をつかむという取り組みを行なっている企業もある。またインターネットショッピングサイトには、過去にユーザと同じ商品を購入した人がほかにどのような商品を購入したかを表示し、ユーザの興味を惹くような工夫をしているところもある。このように、ビッグデータを活用することで利益につなげることができるので、現代社会においてビッグデータに関する研究は注目されている。

では、大学ではどのような研究が行われているのか、脇田研究室で現在行われているビッグデータに関する二つの研究がどのようなものかを見てみよう。

大規模社会ネットワークの可視化

まずは、ビッグデータをどのように扱うかについての研究を紹介する。先生の研究のひとつに大規模社会ネットワークの可視化がある。社会ネットワークとは、何らかの関係性をもつ人間や物事をつなげた社会的な構造である。社会ネットワークを視覚化する例として、人や物事を頂点で表し、2つの頂点の間に何らかの関係があれば辺で結ぶ方法がある（**図1**）。このようにして可視化されたものはグラフと呼ばれる。大規模社会ネットワークとは、社会ネットワークにおいて人や物事の数が膨大になったものである。例えば人の交友関係のグラフにおいては、対象となる人数が数千万人、数億人単位の社会ネットワークであり、これらをデータに落とし込むことでビッグデータとなるのだ。先生はこの大規模社会ネットワークを可視化、つまりグラフに表すことができるのではないかと考えた。

頂点の数が少ない場合は、視覚化することは容易である。しかし大規模なデータになると、可視化を行う際に三つの問題が生じる。それらは、計算量の問題、ピクセル数の問題、毛玉問題の三つである。

まず一つ目として計算量の問題だ。社会ネットワークを可視化するためには、頂点や辺を空間に配置するための計算を行わなければならない。このような計算では、頂点や辺の数に対して計算量が2乗や3乗、あるいはそれ以上の大きさになることが多い。つまり、データ量が10倍になると計算量が100倍、1000倍、あるいはそれ以上になってしまうことがある。計算量が増えると、それだけ計算時間がかかってしまうので、効率が悪くなってしまう。

二つ目はピクセル数の問題である。例えば、200万ピクセルまで表示できるディスプレイに対して、2000万個もの頂点を表示させようとしても、頂点の数がディスプレイのピクセル数よりも多いので表示することができない。そのため、点同士のつながりがわからなくなってしまうのだ。

三つ目は毛玉問題と呼ばれる問題だ。人間社会において、5人の知人を経由すると世界中の誰とでもつながっている、という話が存在する。これに基づいて、3次元空間に原点の1人からつながった1000万人分のグラフを映し出そうとする。例えば、知人同士のつながりを長さが1の辺で表し、頂点同士をつなげると、半径が6の球の内側に1000万人分の頂点が収まってしまう。その内部を見ようとすると、頂点や辺が密集していて見づらくなってしまう（**図2**）。このように、データ量が大きくなるとさまざまな問題が生じる。

そこで、先生はクラスタリングという手法を取った。クラスタリングとは、ひとつの大きな集合を何らかの共通点を基準にしてグループに分けることである。このグループのことをクラスタと呼ぶ。先生はクラスタリングを用いてグラフを分割することにより、大規模社会ネットワークを可視化しようと考えた。

先生は、クラスタリングを導入したSocial Cosmo Browserというソフトを開発した。宇宙をイメージしてこの名前を付けたと先生は語る。宇宙には銀河団があり、その中に銀河、またその中に多数の

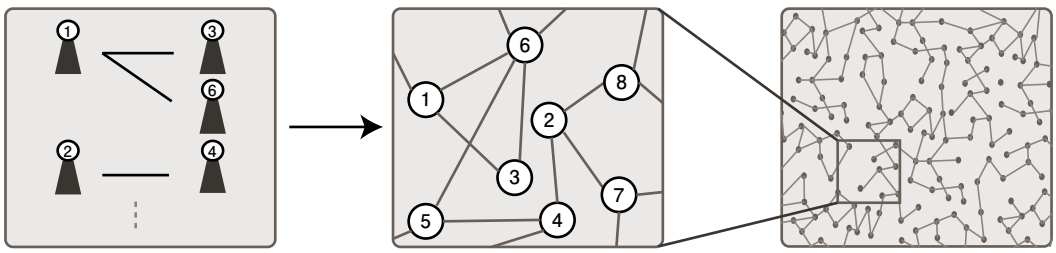


図1　社会ネットワークの可視化

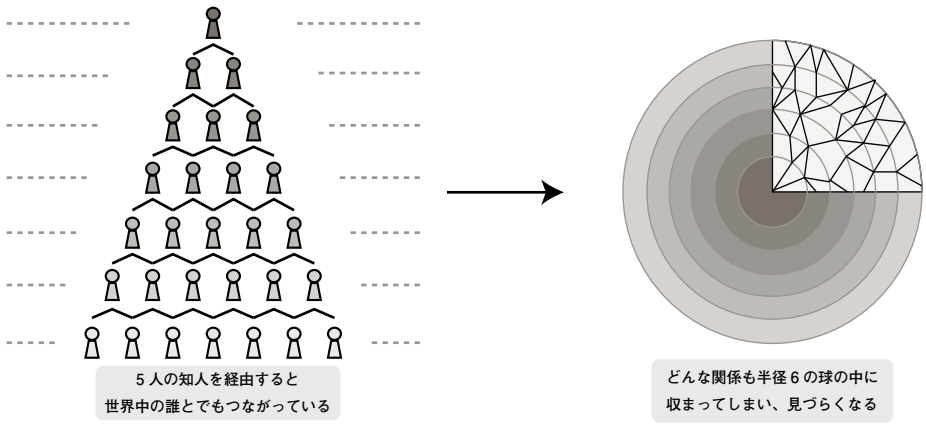


図2　毛玉問題

大規模社会ネットワークを単純に可視化させると見づらくなってしまう。

星というような構成になっている。クラスタリングを使えば、銀河団をクラスタが入った大きな箱、銀河をクラスタ、多数の星をクラスタの内部の頂点、というように、大規模社会ネットワークを対応させることができる。視点が銀河団に位置していれば、多くの銀河、つまりクラスタが画面に表示されて、ある銀河の中へと入るとその銀河内の星を詳しく見ることができる。これを段階的詳細化という（**図3**）。

このSocial Cosmo Browserを使うことで、ビッグデータの可視化における三つの問題が解決できる。社会ネットワークを画面上に高速に配置するアルゴリズムは以前から存在していて、データ数が1万個以下なら高速に配置することは可能であった。そこで、先生は段階的詳細化を使ってこのアルゴリズムを応用した。表示したい視点でのデータ数を1万個以下に抑えることができれば、画面上に高速に表示することが可能である。例えば500万個のデータを1000個のクラスタに分割したとすると、各クラスタ内のデータ数はおよそ5000個になる。このように段階的詳細化を用いれば、一度に表示するデータを5000個程度に収めることが可能となるのだ。

Social Cosmo Browserを使うことで、現段階では87万人に対しての可視化には成功している。しかし、Facebookなどの億単位のデータ量をもつ超大規模社会ネットワークを可視化することにはまだ至っていない。そのような超巨大社会ネットワークを可視化することが先生の目標である。そのために、クラスタ内やクラスタ間での関係性の解析や省メモリ化によって問題点を解決しようとしている。

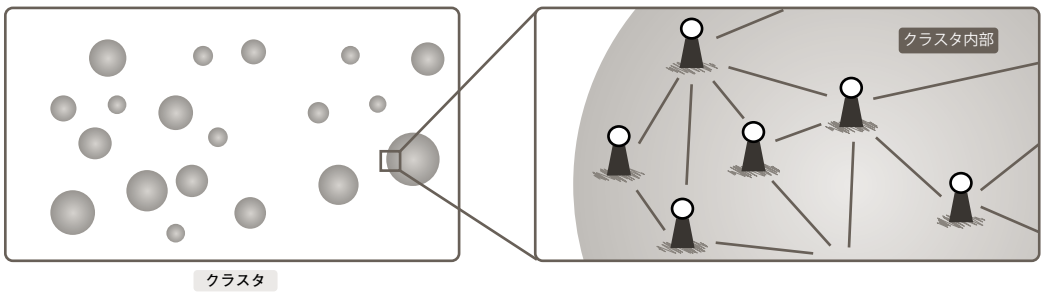


図3　段階的詳細化

クラスタの内部には何らかの関係を共有している人たちの社会ネットワークがある。

通話情報を用いた社会の分析

次に、携帯電話の通話情報というビッグデータを用いて社会を分析するという研究を覗いてみよう。先生が研究に使用したデータは、コートジボワールという国の携帯電話の通話情報である。

コートジボワールは以前フランスの植民地であったが、1960年に独立を果たした多民族国家であり、大小含めて60ほどの部族がある。10年ほど前に国内で民族問題から紛争が起きたため、政府が正確な人口を把握できていないなどの問題がある。

研究に使われた携帯電話の利用情報は、Orangeという携帯会社が研究者向けに提供していたデータである。先生は提供されていた複数のデータのうち二つを利用した。一つ目は携帯の基地局間での1時間ごとの通話回数、通話時間はどれくらいなのかというデータ、二つ目は2週間ごとに無作為に選ばれた携帯のユーザ5万人が、1時間ごとにどの基地局の周りで通話を行なったかというデータである。先生はこれらのデータを利用して、大きく分けて二つの研究を行なった。

まず一つ目の研究は、主要都市の位置および規模の推定である。主要都市の位置については、始めにある基地局を中心として円を描く。その円の内側にほかの基地局がある場合、その基地局を中心にして新たな円を描く。この操作を繰り返してできた範囲から都市の位置を定めた（**図4**）。日本を見てみると東京などの大都市には基地局が多いので、同様の操作をすると円でつなげられる範囲が広くなることがわかる。一方、山岳部などでは基地局が少ないので、大都市とは逆に狭くなるのがわかるだろう。

次に、基地局間の通話回数を用いて都市の規模を推定した。先生は人口が多ければ都市の規模も大きいと仮定して、各都市の人口を求めることにした。そのために2つの基地局i,j間において、基地局間の人々の仲の良さを数値化した$w_{i,j}$を、基地局周りの人口をそれぞれN_iとN_j、基地局間の通話回数を$t_{i,j}$とし、これら3つの数を用いて

$$w_{i,j}=\frac{t_{i,j}}{N_i\times N_j}$$

と定義した。わかっていたのは基地局間の通話回数のみであるから、このままでは都市の人口を求めることができない。しかし、人口の比が求まれば都市の規模の推定は可能であるため、先生は人口の比を求めようとした。

先生は、まず先ほどの2つの基地局i,j間に対応する方程式を、1つの基地局i内での式

$$w_{i,i}=\frac{t_{i,i}}{N_i^2}$$

にした。$w_{i,i}$は基地局i内での人々の仲の良さを数値化したもの、$t_{i,i}$は基地局i内での通話回数である。こうすることで、基地局j周りでの人口を考える必要がなくなるため、変数が1つ減ることになる。ここで、同じ基地局周りに住んでいる人の仲の良さはどこも同じであるとして$w_{i,i}$を一定値とした。また通話回数はデータから読み取ることができる。

すると、N_iは$t_{i,i}$にのみ依存する値となるので、通話回数から基地局周りの人口の比を求めることができる。各都市にどの基地局が位置するかはわかっているので、あてはまる基地局の人口の比がわかれば都市全体の人口の比もわかり、都市の規模の比較も可能となる。結果として、求められた各都市の人口の比を現時点でわかっている実際のデータと照らし合わせると、少し結果がずれる都市はあったものの、人口の比はほぼ同じになった。

二つ目の研究として、先生は人口動態の推定を行なった。人口動態とは、人がどのように移動しているのかということである。例えば日中は仕事をするので人々はオフィス街に集中し、逆に夜には人々は住宅地に集まるといったように、一日の中で人々は移動している。先生はどこの基地局の周りで人々が通話をしているかというデータを用いて、人々がどのように移動しているかを調べた。ここで先生は、人口動態を調べるにあたって、コートジボワール全体を見るのではなく各都市内の移動だけに注目した。その理由は、コートジボワールは経済発展が遅れていて交通が発達していない地域があり、都市間の移動が少ないからである。一方都市内であれば、朝に仕事へ向かい、日中は仕事をして、夜は帰宅するといったような動きが見られる。以上を踏まえて、各都市で人口動態の推定を行なった。

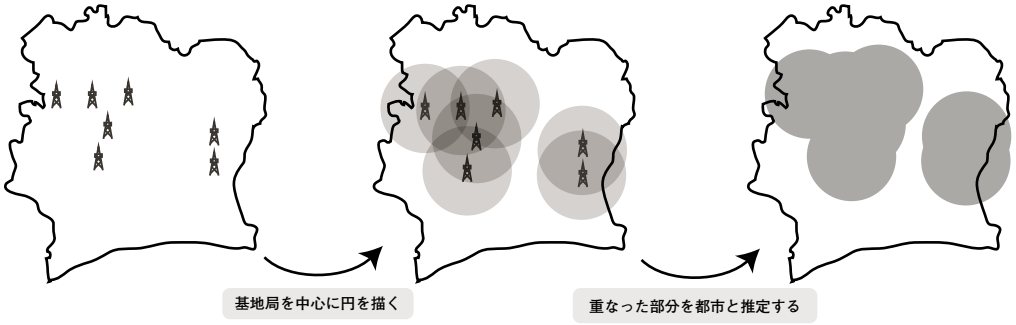


図4　都市の位置の推定
近くにある基地局同士から都市の位置を推定する。

ここでひとつ問題が生じる。それは時間による通話回数の変動である。日中は仕事などで通話を頻繁に行うが、夜中はほとんどの人が寝ているので通話頻度が落ちる。そこで、先生は正規化という方法を取った。正規化とは、データを利用しやすい形に変形させることである。今回のデータでは昼と夜で携帯電話利用者数が一定ではないので、昼と夜の利用者が一定になるように揃えた。例えば、ある基地局の周りで夜は昼の10%の人数しか通話していなければ、グラフに落とし込む際に夜間の通話量のデータを10倍にして表示させる。このように正規化を行いグラフの形を比較することで都市を職場圏、居住圏、そしてその中間の三つに分けることができた。

これからの展望として、基地局間の通話の回数と長さのデータを組み合わせることで人口動態の推定をもっと詳細化できるのではないかと先生は語る。その応用として、宗教に関する分析をしようとしている。コートジボワールはイスラム教、キリスト教をはじめさまざまな宗教が入り混じった多宗教国家である。宗教によって礼拝の時間が違うので、先生は礼拝行動と携帯電話の使用を関連付けるための研究を行なっている。その第一歩として、通話情報から基地局の周りにある宗教施設を見つけるということに挑んでいる。

脇田先生の研究理念

これまでビッグデータに関連した研究を紹介してきた。先生は単にビッグデータをグラフなどに表すことだけではなく、研究の背景にあるものも大切であると考えている。例えば先に挙げた研究においては、コートジボワールという国がどのような国かわかっていなければ分析もしづらく、情報分野以外のことも必要となってくる。

また、脇田研究室ではビッグデータ以外を扱う研究も行われている。先生は、過去に色覚異常者をソフトウェアによってサポートするという研究を行なっていた。ほかには、現在行われているものとしてJavaScriptのマクロ拡張という、C言語などにあるマクロという機能をJavaScriptにも搭載する研究もある。

先生は、自身の専門分野に加えて、常に何か自分の分野とは違うものを吸収しようと思っている。例えば色覚異常の研究に関して言えば、物理学や生理学が絡んでくるので情報分野の知識だけで問題に取り組むことは難しい。専門分野だけにとどまらない研究をやるからには先生は、専門分野以外の勉強もしたうえで研究に取り組むそうだ。自分の専門分野だけでなくほかの分野も取り入れた先生の研究が、私たちの社会をより便利にしてくれることを期待したい。

執筆者より

取材では紙面の都合上紹介できなかったものも含め、大変興味深い話をして頂きました。お忙しい中、快く取材を引き受けていただいた脇田先生に心より御礼申し上げます。

（川邉 航市）