Focus 01

巨大ネットワークの解析で情報学を牽引

数年前よりネットワーク上で増え続ける「ビッグデータ」が脚光を浴びている。膨大なデータを解析することで、社会の課題を解決するような価値ある情報が取り出せると期待されているのだ。この巨大なネットワークを点と辺からなる「グラフ」で表わし、数学を使って解析を試みようとするのがERATO「河原林巨大グラフプロジェクト」の目的である。研究総括を務める国立情報学研究所の河原林健一教授がその概要と意義、成果について語る。

河原林 健一
（かわらばやし けんいち）

国立情報学研究所　情報学プリンシプル研究系教授
（ビッグデータ数理国際研究センター　センター長）

2001年慶応義塾大学大学院理工学研究科後期博士課程修了 博士（理学）。01年バンダービルト大学客員研究員、02年プリンストン大学博士研究員、03年東北大学大学院情報学研究科助手。06年国立情報学研究所情報学プリンシプル研究系助教授、09年同研究所情報学プリンシプル研究系教授。12年よりJST ERATO 河原林巨大グラフプロジェクト研究総括。

増え続ける大量の情報を素早く解析するために

コンピューターの技術の進展で情報量が爆発的に増え続けている。その背景には、世界中の10億以上もの人々がインターネットやスマートフォンを日常的に使い、フェイスブックやツイッターなどのソーシャルネットワーキングサービス（SNS）に文章や写真を投稿したり、ウェブサイトを通じて買い物やゲームをしたりするようになったことによる。インターネットのようなネットワークはもはや日常生活に欠くことができないものであり、人々の活動にともない情報量が増え続けているのだ。

しかしここへきて、莫大に増え続ける情報量に対して、コンピューターの性能が追いつかなくなり、処理に時間がかかるようになってきている。

ビッグデータの解析に挑んでいるのが、2012年10月にスタートしたERATO「河原林巨大グラフプロジェクト」である。ビッグデータの中でも、特に10の10乗、すなわち100億以上の頂点を持つネットワークを「巨大グラフ」として表現し、情報学や数学の最先端の成果を駆使して解析の高速化をめざしている。ただし、「グラフ」とは、私たちがよく思い浮かべる棒グラフや折れ線グラフのことではない。グラフ理論分野の牽引者としてプロジェクトの研究総括を務める国立情報学研究所の河原林さんはこう説明する。

「ここで言うグラフとは、点と辺で表現できるネットワークのこと。離散数学と呼ばれる、計算機科学と密接に結びついた数学の分野で扱う対象です。例えば、鉄道で言えば、それぞれの駅は点、駅同士を結ぶ線路は辺としてグラフに置き換えることができます。フェイスブックの場合も、登録している人を点、友達同士を辺で結ぶことでグラフとして表すことができます」。

このプロジェクトでは、グラフの中でも特に巨大なもの、つまり情報量が膨大なものについて、その関係性を数学的に読み解いたり、点同士の最短距離を探ったり、グラフの変化を予測したりすることをテーマとしている。

「しかも、できるだけ素早く、厳密な答えでなかったとしても、正解に近い答えを見つけることで、新たな価値を生み出そうとしているのです」。

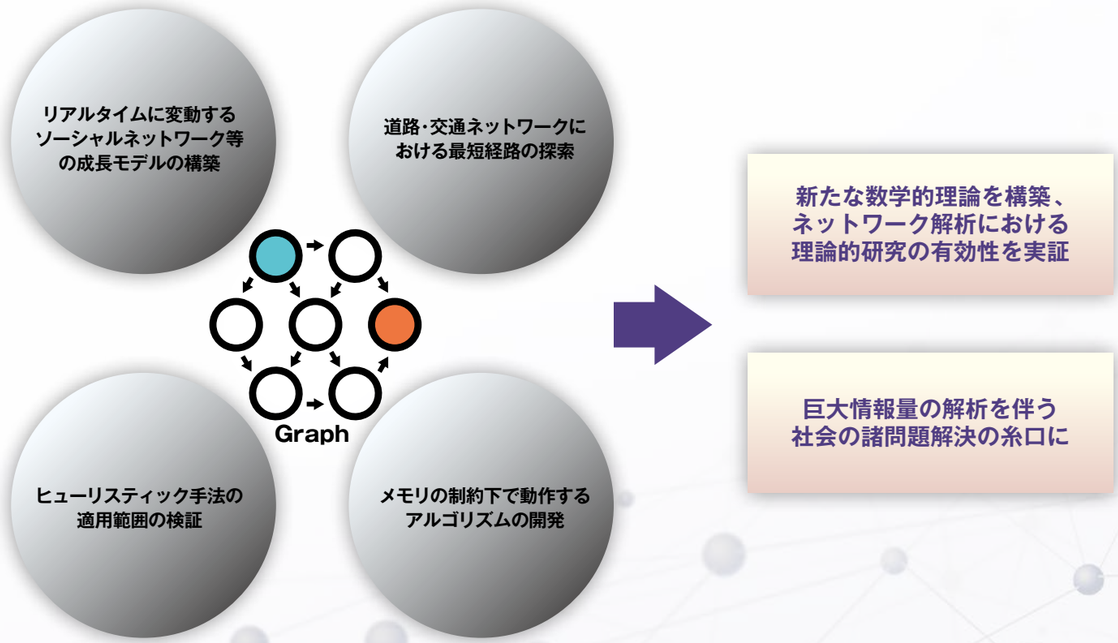
「アルゴリズム」の工夫で計算をより速く

この巨大グラフの解析の肝となるのが、「アルゴリズム」だ。アルゴリズムとは、コンピューターが計算問題を解くための手順、すなわち算法のこと。例えば、3の16乗を計算する場合、3×3の答えに3を掛け、さらにその答えに3を掛けてという具合に順に計算していくと、計算回数は全部で15回になる。ところが、3の2乗に3の2乗を掛けるというやり方をとれば、計算回数は4回で済む。このように、アルゴリズムの工夫によって、計算時間を大幅に短縮できる。

「実世界では、交通網にしろ脳の神経回路にしろ、一気にネットワークが広がることはほとんどありませんが、インターネットの世界ではたったワンクリックでつながるのです。時々刻々とネットワークが変化し、膨張していきます。SNS上では、一般の人でもせいぜい数人を介せばオバマ大統領にまで行き着くといわれるように、物理的な距離は関係ありません。そうしたネットワークの性質を踏まえた上で、高速に最適な答えを導き出すには新しいアルゴリズムの開発が不可欠であり、理論計算機科学や離散数学の分野の最先端の知見が必要なのです」。

プロジェクトで使う道具は最先端の数学だが、その目的は、最終的に社会に役立てることにある。さらに河原林さんは、情報学の基盤を担うことで先端の科学技術領域を牽引していきたいと言う。

「いま、人工知能が注目されていますが、われわれのプロジェクトはこの人工知能の進展を支える基盤研究と言い換えることもできます。コンピューターにより、膨大な情報から欲しい情報を取り出したり、最適な答えを導き出したりするのは、まさに人工知能の1つの姿ですからね」。



「理論計算機科学」、「離散数学」を基礎としたグラフ解析

4つのグループで多角的にネットワーク解析を行う

プロジェクトは以下の4グループからなる。
①「グラフマイニング＆WEB＆AI」、
②「複雑ネットワーク＆地図グラフ」、
③「グラフ・ネットワークにおける理論と最適化」、
④「ネットワーク・アルゴリズム」

①の「マイニング」とは、もともと鉱山の採掘を意味する言葉で、巨大グラフから有用な情報を探すことを目的としている。例えば、SNSのネットワークから特定の事柄に関心を持つ人々の集団を探し、ウェブ上で影響力を持つ人の存在を見つけるなど、時々刻々と変化し膨張していくネットワークの中から価値ある情報を探したり、ネットワークの変化を予測したりするのだ。

「ただし、それを自動的に、ネットワークの接続関係だけから推測しようとしています。いちいち中身を見ることなく、変化し続けるグラフから統計的な手法などを使って有用な情報だけを素早く取り出そうというのです。そうやって情報の変化や伝わり方が見えるようになると、誰に伝えれば情報がいち早く伝わるかなど、平常時・災害時を問わず、最適な情報伝達などに役立てることができます」。

アルゴリズムで20,000倍の高速処理化を実現

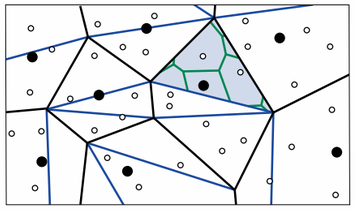
②では、SNSのネットワークやウェブグラフなどに代表される「複雑ネットワーク」と、交通網など実世界の「地図グラフ」を対象とする。例えば、現在のカーナビは10の5乗から6乗もの巨大ネットワークを持つが、こうした巨大グラフの最短経路を探るために、理論からの予測と実験の両面から研究を行う。

「最短経路の探索やさまざまなつながりの見える化は、最適で効率的な物流や交通網の開拓はもちろんのこと、物質同士の組合せを探って創薬の開発に役立てるなど、社会の幅広い場面での応用につながります」。

続いて③では、組合せ最適化※1やグラフ理論などの最先端の手法を、通信ネットワークなどの分野に応用することを目的としている。

「この研究はもともとオペレーションズ・リサーチ（OR）と呼ばれる学問から発展した分野です。第二次世界大戦と冷戦中の米ソで発展した学問で、戦略の立案や効率的な物資の輸送などに役立てられてきました。現在では、エネルギーの需給バランスの算出や日程計画などに広く使われています。米国では現在も莫大な研究予算がついているほどです。そうした中で、われわれは最適化に役立つ高速かつ理論的に保証されたアルゴリズムを開発することで、人工知能を支える機械学習※2の発展にも役立てたいと考えています」。

最後の④は、ネットワークの構成やネットワーク上のさまざまな問題の最適化に関連したアルゴリズムに関する研究を行う。特に幾何学的な問題を対象とするところに特徴がある。「サッカーボールが正五角形と正六角形を組合わせることで球体に似た形をつくっているように、世の中の複雑な形もいくつかの定型の部分を重ね合わせたと理解することで、かなり近い答えを導き出すことができます。そうすることで、脳の構造のような複雑な3次元の形も解析できるのです」。



計算時間と使用メモリのトレードオフを考慮した、ボロノイ図計算のアルゴリズム

いずれのグループも、4～5人の研究員と十数名のリサーチアシスタントで構成されている。4つのテーマは相互に関連するものであり、年1回、研究成果を発表する「ERATO感謝祭」のほか、グループ合同でワークショップを開催するなどして、互いに交流しながら研究を進めている。

※1　組合せ最適化
有限個の解の中からもっとも良い解を見つけるための方法論。有名な問題に、「巡回セールスマン問題」がある。

※2　機械学習
人間の学習能力と同様の機能をコンピューターで実現しようとする技術・手法のこと。大量に集めたデータから学習することにより、そこに潜む規則やルール、パターンなどを見つけ出し、予測などに活用する。第3次人工知能ブームを支える深層学習は機械学習の一種。

広告費の最適化やネットから意味ある話題を抽出

プロジェクトが採択されてから4年目を迎え、目覚ましい成果も出ている。その1つが、③の最適化グループによる、広告の予算配分に関する研究である（p.8参照）。この研究では、企業がマーケティングを行う際に、どのメディアにどのように効率的に広告を予算配分すればいいかを、「劣モジュラ性」と呼ばれる数学の概念を使って素早く計算できるようにした。

「劣モジュラ性を持つ関数というのは、最初は効果があってもだんだん効果が減ってくるような性質を表現できる関数です。例えば、マラソンを始めたばかりの人は、最初のうちはぐんぐんタイムを上げることができますが、ある程度、タイムが上がってくるとそれ以上伸ばすのは次第に難しくなります。経済対策も同じで、最初は効果があってもだんだんと薄れていきます。マラソンのタイムや経済対策、広告の予算配分などを個別に対処するのではなく、劣モジュラ性という共通の性質を用いて汎用的に定式化することで、最適な答えを素早く計算できるようにしたというわけです」。

もう1つ、①のグループ（p.9参照）では、ツイッター上で流行している話題を自動的に、しかも高速に取り出す手法を開発した。現在もツイッター上では話題のキーワードが表示される機能はあるが、これは単語やハッシュタグ（記号付きの文字列）に限ったもので、話題そのものを取り出しているわけではない。

「現状のシステムはたくさんつぶやかれた言葉を単純に数えて、その数が多いものの中から重要だと思われる言葉を人が選んで表示しています。そうではなく、単語同士の関係性を探り、自動的にトピックを取り出そうというのがわれわれの狙いです。実験では、毎分6万件程度投稿される日本全国のつぶやきから、そのときの重要な話題が拾えることを確認しました」。

ここでのポイントは、雑多な情報から意味のある話題だけを取り出すことにある。ビッグデータ解析の障壁は、まさに雑多なものが多数混じっているデータの中から、いかにして有用なものだけを取り出すかにある。

「もっとも、コンピューターには重要かどうかまでは判断できません。しかし、膨大な情報の中から価値ある情報を絞り込むのには大いに役立つはずです。健康診断などでも、要注意の人を自動的に絞り込めれば、相当に効率化できるでしょう。病気の診断は最終的に医師である人間の仕事ですが、判断の支援に役立つと思っています」。

その他にも、複雑ネットワークの性質を用いることで、最短経路の計算を数百倍に高速化することにも成功。計算機科学分野でトップクラスの国際会議を中心に、すでに80本以上の論文が採択されており、さらなる成果が期待される。

世界に通用する若手研究者の育成をめざして

河原林さんがプロジェクト開始当初から目論んでいた、若手研究者の育成でも、このプロジェクトは大きな役割を果たしてきた。

「グーグルにしろアマゾンにしろ、いまや米国のIT企業が数学者を高い年俸で引き抜いて囲い込んでいるように、IT社会の発展には数学の基礎研究に通じ、プログラミングができる人材が不可欠です。しかもIT関連の研究者の平均年齢は若く、欧米で活躍している人の多くは20～30代です。一方、日本ではIT業界を支える人はもっと年配で、特に応用数学の分野では後れを取っているといわざるを得ません」。

そもそも、日本の高校生や大学生が数学オリンピックやプログラミングコンテストなど国際的なコンテストでトップクラスの成績をおさめているにもかかわらず、修士課程を終える頃には世界のトップクラスに大きく差をつけられている状況に、強烈な危機感を抱いてきたと河原林さんは言う。

「残念なことに、数学オリンピックの参加者の多くが将来は医者になってしまうのです。数学や情報学などの研究者が魅力的に見えないからなのでしょう。この状況を変えたいと思い、このプロジェクトを始める時から20代の若手研究者を積極的に採用し、高校生などへのアウトリーチ活動も積極的に行ってきました。そもそもERATOの役割の1つは、世界に通用するエリート研究者を育てることにありますから」。

その言葉の通り、プロジェクトを構成するメンバーはほとんどが35歳以下と若い。すでにプロジェクトでの成果を携えて、国立情報学研究所や東京大学、京都大学などで助教などの教員に採用された研究者も十数名にのぼっている。

「研究者の育成には、ただ教えるよりも一緒に研究を進める中で身につけていくほうが効果的です。だからこそ、自分が30代のうちに彼らと一緒に働くことができて良かったと思っています」と振り返る。

トップクラスの研究者がマネジメントに携わる意義

ところで、河原林さん自身、国際的に活躍するトップクラスの研究者として20代から多くの論文を発表し、情報学や数学の分野で大きな成果をあげてきた。このプロジェクトに対してどのような思いを抱いて臨んでいるのだろうか。

「実は私が研究総括に選ばれたときに思ったのは、『自分で自分の実験をしてみたい』ということでした。トップクラスの研究者が自身も研究に携わりながらマネジメントをすることで、世界的に活躍できる人材を育成できるかどうか、さらには産業界でも活躍できるようなリーダーを輩出できるかどうか、試してみたかったのです。まだ志半ばですが、人材が育っていくのを目の当たりにして、それなりに意義のある活動ができていると思います」と自負する。

もう1つ、プロジェクトを進める中で大きな原動力となっているのが、先述の日本が置かれた状況に対する強烈な危機感だ。

「今世紀に入り、日本は情報分野でずっと負け続けています。かつて世界を席巻してきた日本の名だたる大企業のエンジニアが、欧米のIT企業に圧倒されていく姿を目の当たりにして、なんとかしなければという強い思いがあります。日本の技術力はいまだに非常に高い水準にもかかわらず、なぜそのような差が生じてしまうのでしょうか。それは基礎研究、特に数学の基礎研究の差に他なりません。その部分の底上げに貢献したいですね」。

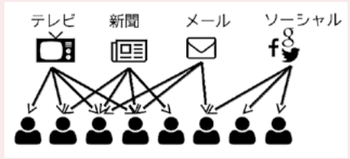
40代となった現在、対外的な仕事も増え、研究に没頭できる時間は減ったが、その分、質を追求しているという。その言葉の端々に、日本の情報学をこのプロジェクトで牽引する強い決意が感じられる。今後のさらなる成果を大いに期待したい。

今年3月の成果報告会で。
若手研究者が着実に育っている。

グラフ・ネットワークにおける理論と最適化グループ

広告の持つ「劣モジュラ性」を発見し、効果的な広告予算を割り出す

相馬 輔 東京大学 大学院情報理工学系研究科 助教（左）と
垣村 尚徳 東京大学 大学院総合文化研究科 講師



最適予算配分問題
広告費をメディアに分配してなるべく多くの人に影響を与えたい。

今回使った手法では、最大15%既存手法より多くの人に影響を与えられる。

グループの研究課題の1つである組合せ最適化は、与えられた条件を満たす限られた組合せの中から、最も良い組合せを見つけるための方法論だ。組合せ最適化問題の例としてよく知られているのが巡回セールスマン問題。n個の地点を1回ずつ通って元の地点に戻る巡回路の中から、総移動距離が最も短いものを見つけ出す。この問題は、原理的にはすべてのルートを列挙して距離を比べれば解くことができるが、nの数が大きくなると、膨大な量の計算が必要になってしまう。組合せ最適化の研究では、このような問題に対し効率よく最適な解を求めるためのアルゴリズムをつくることをめざしている。

限られた広告予算を、どのメディアにどう配分すると最も効果が上がるかという広告予算配分問題も、組合せ最適化問題の1つである。

「新聞、雑誌、テレビ、ラジオ、ウェブなどさまざまな広告媒体がある中で、できるだけ重複をなくし、多くの人の目に触れるように広告を打つというのは複雑な問題です。われわれは、組合せ最適化でよく知られる劣モジュラ性という概念に着目することで、この難問をうまく数理的にモデル化し、計算できることを示しました」。研究協力者の相馬さんは、今回の成果を説明する。

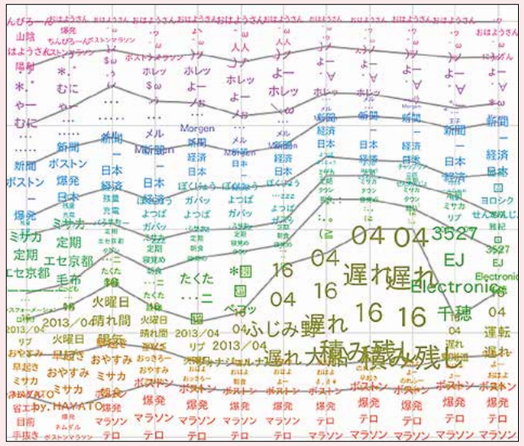
グループリーダーを務める垣村さんは、広告の効果にも劣モジュラ性が見られるという。

「広告というのは、最初の数回は大きな反応があるけれど、すでに多くの広告を出している時は、新しく出す広告の効果は薄れていきます。そういう性質が劣モジュラ性であり、われわれの成果は、広告の持つ劣モジュラ性を発見したことによって、比較的単純なアルゴリズムで、広告予算配分のような複雑な問題を解けることを数学的に証明したという点にあります」。

劣モジュラ性は、組合せ最適化の分野では古くから知られた概念だが、近年は機械学習の分野でも注目されている。グループでは今後も、劣モジュラ性を含めた組合せ最適化のさまざまな手法や、グラフ理論の先進的な手法の応用に取り組んでいく。

グラフマイニング＆WEB＆AIグループ

ツイッター上で流行している話題を、自動的に高速抽出する手法を開発



トピック抽出の可視化例。文字が大きいほど頻出度が高い。

林 浩平 産業総合研究所 研究員

グループでは、人やモノ、場所などの「つながり」から、新たな知識や有用な情報を発見するグラフマイニングと、その技術のウェブやAI分野への応用を研究テーマとしている。グラフマイニングの土台は、大量のデータの中から意味ある情報を取り出すデータマイニングと、データを繰り返し解析することで有用なパターンやルールなどを見つけ出す機械学習などの技術だ。

それらの研究成果の1つに、ツイッター上で流行している話題を自動かつ高速に抽出する手法を開発した。手軽なメッセージ発信ツールとして、全世界で3億人以上、国内では約3500万人が利用しているツイッターで、毎分毎秒、膨大な数のつぶやきが飛び交っている。

「その中には価値ある情報も多いはずで、それをできるだけ早く、人手をかけずに見つけたいとの思いから始めた研究です」。サブグループリーダーである林さんは、開発の経緯をそう話す。

開発したのは、ツイッター上で多くつぶやかれているキーワードを、単体ではなく3つ程度の単語のグループとしてリアルタイムに抽出する手法だ。1つのつぶやきの中に含まれる複数の単語は、互いに関連性があると考えられる。それらの中で同時に出現する頻度が高い単語を、グループとして抽出する。

「キーワード単体なら簡単に抽出できますが、それでは情報量が少ないですよね。複数の単語のグループとして取り出すことで、文脈が見え、周辺情報も含めたホットなトピックを知ることができます。スパム広告などは、統計的手法を用いて自動的に判定し、排除する仕組みとなっていますから、意味のない投稿に影響されず実態をつかむことができます」。

SNSのリアルタイム分析は、災害時の被害状況の把握や救援要請などにも役立つと期待されている。データの山の中から有用な情報を自動的に見つけ出す技術は人の命を救うことにも直結し、社会に欠かせないものとなるだろう。