URL https://journal.ntt.co.jp/article/29841
DOI https://doi.org/10.60249/24101301

NTTコミュニケーション科学基礎研究所
特別研究員

藤原靖宏 Yasuhiro Fujiwara

迫り来る大規模データ時代に必要な「高速かつ正確なデータ分析基盤」

IOWN（Innovative Optical and Wireless Network）構想の構成要素であるオールフォトニクス・ネットワークでは多くの大規模データを集める未来が描かれています．低遅延のオールフォトニクス・ネットワーク上では計算リソースやデータが手元になくとも高速にアクセスできればユーザが必要な知識や情報を取得できるようになります．データ分析を高速に実現するため，近似解ではなく正確なデータの分析を取り組む藤原特別研究員から，研究の軸として変化しない部分と，柔軟に外部環境を研究して取り組む部分についてお話を伺いました．

◆PROFILE：2001 年早稲田大学理工学部電気電子情報工学科卒業．2003 年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年，日本電信電話株式会社入社．2011 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．博士（情報理工学）．データ工学，データマイニング，人工知能などの研究開発に従事．平成 29 年度科学技術分野の文部科学大臣表彰若手科学者賞，電気通信普及財団第 27 回テレコムシステム技術賞，情報処理学会 2015 年度長尾真記念特別賞など受賞．

データベース×機械学習：高速かつ正確な大規模データ分析基盤のための研究

■今行われている研究を始めるまでの経緯を教えてください．

私は現在「大規模データに対する高速データ分析基盤」という，データベースを軸とした研究を進めています．

自身の研究についてさかのぼると20年ほど前にデータベースの研究チームに入り，多くのデータの中からユーザの好みに基づいた情報の提示や，ユーザの指定した検索条件に近しいデータを探すなど，データ内容を分かりやすくかつ要約してユーザに提示するという研究を行っていました．

私が研究を始める前の1990年代のデータベース分野では格納されている静的なデータが研究の対象でしたが，私が研究を始めた2000年代になるとセンサなどのデバイスが発達し，データストリーム（例：インターネットから時系列で送信されてくるデータ）などの動的に流れてくるデータに対する研究が盛り上がっていました．

私も時代の流れに沿ってデータストリームを研究していましたが，2010年ごろからはソーシャルネットワークやWebなど，当時流行りのサービスやアプリケーションと親和性が高い，グラフでデータを表現するグラフデータベースに対する研究がその分野で人気の研究課題となりました．

それならば，と私もグラフデータの研究を進めていましたが，データベースを利用する立場であるアプリケーションからの観点で技術課題や研究論文などを調べると，レコメンデーションなどの多くのタスクに機械学習が応用できる可能性が示されていました．そのときに「どうやらこれは機械学習の時代がきているのでは」と感じ，現在の機械学習を意識した大規模分析基盤の研究にシフトしてきたという状況です．

またデータベース分野の研究も近年の機械学習の影響で変わってきたように感じます．データベースは，アプリケーションとOSの間にあるミドルウェアで，データを蓄積・加工するような補佐的な位置付けにあります．20年前のデータベース研究が対象としていたデータ分析は，似たデータを探す類似検索など初歩的なものでしたが，10～15年くらい前にビッグデータという概念が出てきたころから，大量のデータに対して高度な分析をしてユーザが想像もできないデータに潜む情報を提示する，いわゆる「機械学習に基づいた分析技術」が注目されている状況にあると思っています．

ただ私としてはOSとアプリケーションの間というデータベースが担うポジション自体は変わっていないと考えています．そのため，集まった大量のデータを分析しその結果をユーザに提供するというデータベースのミドルウェアとしてのポジションのまま，データベースと機械学習をつなげて高付加価値を生み出す大規模データの情報処理基盤の研究を進めています．

■大規模データ時代にはなぜ分析基盤研究が必要なのでしょうか．

大規模データを取り扱うことはデータベース研究の醍醐味ともいえます．データベースの研究者の発想はシンプルで，「膨大な

量のデータを高速に処理する」ことにフォーカスしています．極端にいえばそこにだけ興味がありスモールデータには興味がありません．

これは今に始まったことではありません．1970年代からあるもっとも古いデータベースの国際会議の１つにVLDBというものがありますが，これはVery Large Data Basesの頭文字を取ったものになります．このように昔からデータベースの研究には「多くのデータを高速に処理する」という世界観があり，その世界観を現代風にアレンジしているのが私の研究だと思います．

低遅延のネットワークにより膨大なデータを集めることが可能になる大規模データ時代には，集めたデータを活用するために必要となる正確かつ高速なデータの分析技術への関心がこれまで以上に高まると考えています．その集めたデータの分析手法として，機械学習が今よりさらに重要になってくるでしょう．

機械学習を研究している方は国内外に非常に多くいらっしゃいますが，その研究のほとんどは「精度を上げること」に興味が向いている傾向があり，処理速度の高速化をメインで研究されている方は，実をいうとさほど多くありません．なぜなら高速化を追求するためには機械学習だけでなく，データベース分野などにおける手法も知らないといけないため手を出しにくいところがあるからです．

私はその「高速化をメイン」とした分析基盤を研究テーマとしています．大規模データ時代は人間が知覚不可能なデータを，人間が知覚可能な情報に変換できるように処理することが重要になってきます．そのためには精度を高める研究はもちろん，機械学習が大規模なデータを正確かつ高速で分析処理ができる基盤を構築する研究が必要だと考えています．

■この研究はNTTの掲げるIOWN構想の１つであるオールフォトニクス・ネットワークにも関連しているのでしょうか．

IOWN（Innovative Optical and Wireless Network）構想ではオールフォトニクス・ネットワーク（端末からネットワークまですべてにフォトニクス（光）ベースの技術を導入し，エンド・ツー・エンドでの光波長パスを提供することで実現する，低消費電力，高速大容量，低遅延伝送のネットワーク）により大規模なデータを集める未来が描かれていますが，これはデータベース研究における「膨大な量のデータを高速に処理する」という世界観と非常に親和性が高いと考えています．そのため私のデータ分析基盤の研究がIOWN構想に貢献できると考えています．

具体的に言えばオールフォトニクス・ネットワークにより低遅延が当たり前になった場合，手元に計算機リソースやデータがある必然性はなくなり，低遅延のネットワークの先のサーバにある計算機リソースやデータにアクセスすればさまざまな処理を行うことが可能になります．またIOWN構想におけるオールフォトニクス・ネットワークが立ち上がれば大規模データの収集も可能になります．私の研究は大規模データの高速データ分析基盤であり，ネットワークの先につながったサーバにある「莫大なデータの中に人が思ってもいなかった知識や情報」を獲得することにつながります．超高速低遅延のIOWNオールフォトニクス・ネットワーク，そしてそこに集まるデータから価値を生み出すデータ分析の基盤として，「ネットワークにおける高付加価値」をめざした研究を進めています．

■「大規模データに対する高速データ分析基盤」はどのような特徴があるのでしょうか．他の処理技術との違いやポイントを教えてください．

よくあるデータベース研究との大きな違いは，近似解（高速性を優先させた，誤差を許容した解）ではなく「高速かつ正確をめざしていること」です．

初めてデータベース分野の研究に取り組んだ当時は，データストリームに対する研究が流行っていたとお伝えしました．そのときのデータベース分野の共通認識として「データは流れていくから処理が追いつかない．そのため近似解で良い」というものがありました．20代そこそこの私はデータベースのあるべき姿としてその考えに違和感を抱き，なんとかしたいと考えていました（**図１**）．

私は一時的に研究所から出てNTT西日本に２年ほど所属していたのですが，当時はNGN（Next Generation Network）の開発が盛り上がっている一方で，NGNの開発を終えた次はどのようなサービスを開発するのかが課題でした．

フレッツ・グループアクセスなどのネットワークを活用した新

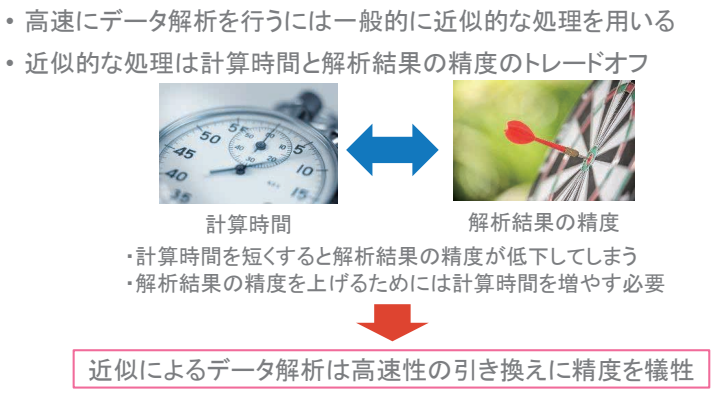


図１　近似によるデータ解析

- 不要な計算を枝刈りすることでデータ解析の高速性と厳密性を両立
- 計算の枝刈りを用い膨大なデータを高速に解析する機械学習基盤を実現

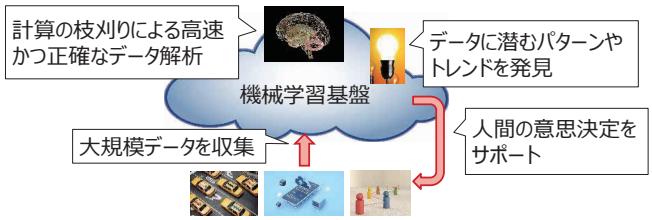


図２　高速かつ正確なデータ解析技術

しい利用用途などを創出するのは１つの方法でしたが，私としては当時自分が担当していた「既存サービスを新しい技術で裏支えして既存事業に貢献する」というサービスに意義を感じていました．具体的にそのサービスは，NTTが提供していたラジオ中継網がメーカ機器の老朽化により維持困難になったため，お客さまから見てインタフェースが変わらないような機器を新規に開発することで延命させていくというものでした．

そのときに既存のサービスをそのままの動作条件で，バックエンドを新しいものに置き換えて提供するというのは，企業のサービスの１つになるのではと考えました．ユーザに提供する機能は変わらないけれども，より早く，より確実に，という見えない部分の価値提供も１つの大きなイノベーションではないかと考えをめぐらせたともいえます．

もし代替システムで置き換えるとき，ユーザとしては処理結果が近似ではなく正確なほうが良いし，さらに必要な計算機リソースを減らしつつ速度を上げ，イニシャル（初期導入）コストやランニングコストを下げられるのならより喜んでいただけるはずですよね．

このような思いもあり，データベース分野において主流ではあるが私としては必ずしも同意しにくかった「高速な近似解で良い」という手法ではなく，枝刈り技術を取り入れ「高速かつ正確」な手法にこだわりをもって研究しています．

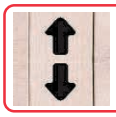
この高速かつ正確なデータ分析技術がIOWNの世界観の中でも他の機能の裏側を支え，処理速度やコストを少しでも下げることで，IOWN自体のアドバンテージになると思っています（**図２**）．

■高速化のポイントとなる枝刈り技術について教えてください．

データベースの世界では枝刈りという概念がありますが，それが研究のコンセプトでもあり，手法の考え方の中心の１つでもあります．枝刈りの手法はいくつかありますが，代表的なものですと上限値・下限値による計算の省略，解になり得ない計算の打ち切り，楽観的処理による高速計算の３つがあります（**図３**）．

枝刈り技術の特徴は分かりやすくいうと基本的には「サボること」しか考えていないことです．どうやったら「計算をサボれるか」，「何がいらないのか」ということを考えて高速化しています．

例えば，辞書の中から「鉛筆」という言葉を探すときに，最初から１ページずつめくる人はあまりいないでしょう．まずは「え」



上限値・下限値による計算の省略
厳密にスコアを計算した結果，値がゼロになるような処理をスコアの上限値と下限値を用いて高速に特定

解になり得ない計算の打ち切り
検索処理の過程で解になり得ないパターンを保持しておきそのパターンが再び検索処理に現れたときに処理を打ち切る

楽観的処理による高速計算
制約条件を外したうえで高速に解を求めた後に，得られた解が制約条件を満たすかを確認する

図３　代表的な計算の枝刈り手法

のページを見つけて，そこから「ん」がありそうなページをペラペラめくる，このようなcoarse-to-fineのように荒い探索から徐々に細かい探索を繰り返すのがデータベースとしてよくある手法です．また皆さんが日常生活で行っている物探しとも同じです．鍵をなくしたからといって家中ひっくり返して探すのではなく，最後にどこにあったかを思い出して，当たりをつけて探しますよね．枝刈りの考え方として明らかにないところは省き，当たりがつけられそうなところを探すというコンセプトで無駄を省きます．このような考え方をコンピュータで処理する数式に落とし込むような研究をしています．

変化し続ける技術に対応する挑戦

■この研究を遂行するうえでの課題と，心掛けていることなどがあれば教えてください．

私が他の研究者の方と特に違うなと思うのは，良くも悪くも特定の分野に特化しないことでしょうか．多くの研究者は特定の分野を専門的に研究されていますが，私は「枝刈りをして計算を速くしよう」という考え方を基盤として研究しているので，必ずしも特定の分野に特化していません．そのため，時代で求められているデータ処理や将来的にポピュラーになりそうな技術トレンドをとらえて，その中で自分が貢献できる分野がないかキャッチアップし続けることが常に課題です．

以前指導していた研究者に，「藤原さんは自動車を片輪走行さ

せながらパーツを変えるというパフォーマンスに似ている」と言われたことがあります．研究者は皆同様で，常に走り続けながらパーツを変える作業をしていかなければなりません．私の場合はデータベース分野における枝刈りという研究領域を半歩踏み締め，もう半歩は機械学習などに代表される新しい研究領域に踏み出しながら研究を進めています．

現在，研究的にホットな深層学習（ディープラーニング：コンピュータが自動で大量のデータを解析し，データの特徴を抽出する技術）はもしかしたら５～10年後には別の技術に置き換えられているかもしれません．その場合はそのときに求められている技術をキャッチアップして勉強しないといけないと思っています．常に勉強，そして研究です．

■この研究の展望をお聞かせください．

最近さまざまな分野の方と話をしていて頻繁に出てくるのが，やはり深層学習についての話題です．深層学習が出てきた10年ほど前は，まず深層学習をやるかやらないかという選択肢がありました．当時，深層学習はやらないという研究者の方も結構いましたが，時代は変わり，そういう方はだいぶ減ってきました．今の若い研究者の多くは深層学習を研究していますよね．ただ深層学習はとにかくゴリゴリ計算を頑張るという処理であり，「計算をサボれるところ」があまりなく，私自身は必ずしも深層学習に積極的ではありませんでした．

しかし，深層学習の登場から10年以上経った今，技術の分化が進み多くの周辺技術が出てきました．大規模言語モデル（LLM：Large Language Models）でも，LLMだけではなくて外部データを使って知識を生成する，そういう時代になってきて少しずつ計算をサボる隙というのが出てきたように思います．深層学習において外部と連携するデータが多くなればなるほど，自分の枝刈り技術が活用できるのではと可能性を感じています．

また「条件にマッチしたものを探す」というのはおそらく人間がデータを計算機で扱ううえで本質的な処理であるため，今後出てくるであろうあらゆる研究でも重要になると思っています．将来的にも「高速に探索をする」ことにフォーカスして研究するのだと思います．

■最後に，研究者・学生・ビジネスパートナーの方々へメッセージをお願いします．

一般論でいうと仕事では，できること（can），すべきこと（should），したいこと（want）の３つのバランスが取れていると理想的だと思っています．

学生のうちは「したいこと」が大きなウェイトを占めており，自分に何ができるかは分からない方が多いと思います．私の場合は「したいこと」を仕事にするために事業会社に行きましたが，そこでの仕事は必ずしも自分が「できること」ではないと分かりました．同じように若いうちは自分の「できること」が分からず，往々にしてそれが自分の「したいこと」とズレていて，「したいこと」に対する気持ちに裏切られることが多いかなと思っています．そのため特に若いうちは食わず嫌いをすることなくまずはさまざまな仕事に取り組むことで，自分が「できること」を探すのが良いかなと思います．

また情報系の研究者としては，今は世の中でコンピュータやインターネット技術の活用が進んでいるので，自分が「したいこと」でかつ「できること」が，世の中に必要とされるような「すべきこと」になっている機会に恵まれていると思っています．さまざまな研究の選択肢のある中で自分の研究者としてのフィールドを探せば良いのだと思います．

NTTは情報系の広い分野で数多くの研究をしているので，自分の専門外の研究をされている方が多くいらっしゃいます．何せ2000人以上も研究者がいるわけですから．NTTの研究者は人間性が素晴らしい方が多いので，仲間集めや情報収集もしやすい環境です．さらにNTTには多くの事業会社があるため，研究者以外にエンジニアやデータアナリストなどになる道もあります．

この業界を志す方はぜひ選択肢が多いことを利点として幅広い分野に興味を持ち，長い目で前向きに取り組んでほしいと思います．NTTで現在活躍されている研究者の方は必ずしも学生のころから研究者として多くの実績を出していたというわけでもなく，研究を粘り強く継続することで花開いた方も数多くいらっしゃいます．私も研究に本格的に取り組み始めたのは事業会社から研究所に戻った30代からでした．粘り強く継続していれば，見てくれる人は見てくれます．

私は若いころ，初めて論文が通ったとき，１カ月くらいそれはもう天にも昇る気持ちでふわふわしていました．学生や若き研究者の方々も，メモに書いていたアイデアが論文になりそれを眺めたときに，嬉しいなと感じられると思います．論文が通る前の段階だとしても，自身の頭の中の考えが具現化されたときに，格別の達成感を感じられる仕事だといえるでしょう．

研究者で成功している方はそれなりに皆さん軸があります．たとえもし世の中がある一定の見方をしていても別の視点から見るとこうだよね，と言えるような方はこの業界に向いているのではないでしょうか．今の時代は理系の研究者が技術により新しい世界をつくる時代になってきました．自身の軸がある若い研究者の方が増えてくれることを願っています．

（今回はリモートにてインタビューを実施しました）