最先端アルゴリズムを使って 小さなメモリでも迅速にビッグデータを分析

数理計画部　研究員　白川 達也さん
数理計画部　研究員　二反田 篤史さん

左から二反田篤史研究員，白川達也研究員

新しい分野について学びながら ビッグデータの分析ツールを開発

◆開発メンバについてお聞かせください．

二反田さん：主に私たちを含めて4名で開発しています．私と白川は数理計画部に属しており，アルゴリズムの研究開発という，計算のコアな部分を担当しています．残り2名はデータマイニング部に所属しており，ビッグデータの分析・集計・加工といった前処理的な操作や手法・ツールの整備を行っています．つまり，当社の中でも部をまたがって技術を集約する混成チームで開発しています．

◆Big Data Moduleとはどのようなツールですか．

二反田さん：大規模データに対して，ルールやパターンを見つけるための機械学習やデータマイニングを行うツールで，2012年8月にリリースしました．例えばコンビニ店舗の売上が変動する際に，気候や周辺人口，イベントの有無などが影響を及ぼしている可能性がありますが，その規則を何かしらの手法で発見するものです．こういったものはデータが増えると，より確かな分析を行うことが可能になります．

◆開発時のご苦労はありましたか．

白川さん：2年少し前から準備を始めたのですが，そのころは私も二反田も完全な門外漢で，勉強も兼ねて頻繁に打ち合わせを行うことが必要でした．ある程度知識が身についてからも，こういった機械学習やデータマイニングの分野では新しい技術が日々生まれては陳腐化していくので，毎日それに追いつくだけでも大変でしたね．最新の論文をずいぶんたくさん読みました．逆にいうと，毎日新しいことが学べてもっとも楽しかった時期でもあります．

二反田さん：私は，いかに計算速度を速くするか，という点が大変でした．データ分析においては，データに意味のない情報やノイズが含まれているために，分析がうまくいかない場合が多くあります．その問題を解決するため，分析の過程においては試行錯誤を繰り返すのですが，大規模データを対象としていると，試行錯誤のたびの計算時間の増加が顕著なため，高速化が大きな課題でした．

ミドルクラスのデータ分析に最適 分割すればどんなに大きなものも扱える

◆Big Data Moduleの特長を教えてください．

白川さん：世の中のビッグデータの分析は集計程度のレベルにとどまっている例が多く見受けられます．集計はもちろん重要なのですが，私たちのモジュールが目指すのはそれを超えた深い分析です．ビッグデータというとテラバイトやペタバイトオーダのデータというイメージがあるかもしれませんが，ある論文によると，Hadoopのような大規模分散環境で実行される分析タスクでも実際には10ギガぐらいのデータ量が平均的であるといわれています．分析が必要なデータに整形するとミドルデータといっていい程度のサイズになるんですね．しかし，ミドルデータ程度でも，多くの分析アルゴリズムは計算限界を超えてしまうので，適切なアルゴリズムを採用することが必要となってきます．

二反田さん：私たちのBig Data Moduleはそのミドルクラスである数ギガ～数百ギガを主なターゲットとしています．そして，大規模な分散環境を用意しなくても，１台のマシンで高速に分析を行うことができる点が大きな特長といえるでしょう．もちろん，その上のサイズとなると，１台のマシンでは取り扱えませんので，その基盤としてHadoopとの連携も視野に入れています．こちらは私たちとは別の2名が主に担当しています．

◆最先端のアルゴリズムを採用されたそうですね．

白川さん：オンラインマイニングアルゴリズムを採用しました．これはデータ１つを与えると微小な変化をモデルに加えるという操作を繰り返すものです．それによってどんなサイズのデータも処理が可能になります．

二反田さん：メモリにデータを載せて分析すると扱えるサイズが限られてしまいますが，オンラインマイニングアルゴリズムはデータを分割しながら少しずつメモリにロードして処理するので，全体のデータのサイズを気にしなくて済みます．メモリ領域自体を大きくするHadoopとはそこが異なる点ですね．

◆現在抱えている課題はありますか．

白川さん：ビッグデータのイメージが先行していて，ユーザがすぐにHadoopに飛びついてしまうことがもどかしいですね．Hadoopはオーバスペックになりがちで，10ギガ，100ギガ程度のデータをHadoopにかけると非効率です．そのような現実的なデータサイズであれば，当社のオンラインマイニングアルゴリズムのほうが効率的なので，これから普及させていきたいです．

教師なし学習を用いた全自動分析ツールの開発が夢

◆今後の夢や目標を教えてください．

白川さん：実現は難しい夢物語ではありますが，ボタンを1つ押すと欲しい結果が出るような全自動分析ツールをつくれたらいいですね．その第一歩としては，教師なし学習に基づく分析を強化していきたいです．機械学習には教師あり学習と教師なし学習という2つの手法があるのですが，教師ありのほうは，機械がルールを発見するために人間があらかじめ正解データを用意しないといけないものです．例えばスパムメールを検知する際には，「こういう文面はスパム」という知識を前もってインプットする必要があります．教師なしでは，入力データを受け取ったら人間が正解データを用意せずとも，アルゴリズムが自動的にルールを発見して分析結果を出してくれます．全自動分析を行う場合もそうですが，ビッグデータを扱うときにも，正解データを前もって用意することが難しい場面が多いのです．ですから，十分にビッグデータを活用しようとすると，教師なし学習が必要になると考えています．

二反田さん：私も教師なし学習に興味がありますが，直近の目標としてはBig Data Moduleのストリームデータ対応を考えています．センサデータ等のリアルタイムに送られ続けるデータを効率的に，分析をできればと思っています．

◆NTTデータグループとの連携に期待することは．

白川さん： NTTデータとの共同研究にも大いに期待しています．私たちはアルゴリズムを持っていますので，実際の業務システムとの連携など，NTTデータの持っている実データへの適用を多くこなしていきたいです．

二反田さん：NTTデータの所有する大規模な計算環境もぜひ使ってみたいですね．　　（インタビュー：村上百合）

NTTデータ数理システム ア・ラ・カルト

■和気あいあいと楽しむ社員旅行

2年に1回行われる社員旅行．1泊2日で，毎回50名ほどが参加するのだとか（**写真1**）． 2013年6月7～8日に行われた湯河原旅行のテーマは「近場でゴージャス」！　社員たちは宿の素晴らしさや食事の美味しさに大満足したそうです（**写真2**）．1日目は水族館かアーチェリー，2日目はお皿の絵付か蒲鉾づくりと，好きなアクティビティが選べるのも好評の秘訣．部署の垣根を越えた交流をたっぷりと楽しみました．

■他部署の活動が分かる成果発表会

毎年5月半ばに催される成果発表会では，各部署から1名ずつの代表者が出て，40分ほどの発表を行います．技術開発や研究に関する成果など，他部署でどんなことをしているのかを知ることができるチャンス．それに関する質疑応答も熱心に行われます．成果発表会がある日の夜は新人歓迎会で盛り上がり，公私両面でかかわりを深められる良い1日になるそうです．

■テニスや映画鑑賞などのサークル活動も

多趣味な社員の多いNTTデータ数理システム．テニスや映画鑑賞，スキー，茶道，華道といったさまざまなサークル活動が営まれています（**写真3**，**4**）．活動イベントが決まると回覧板やメールで告知され，参加者を募るというかたちで運営されているそうです．


写真1


写真2


写真3

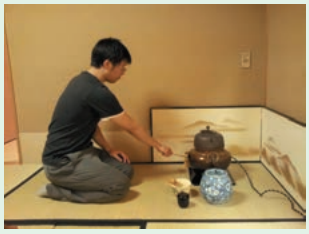
写真4