グラフマイニング技術を用いたビッグデータの応用と高速化技術の取り組み

特　集

大阪大学　大学院情報科学研究科　ビッグデータ工学講座
教授　鬼　塚　　真 氏

●はじめに

私は大阪大学でビッグデータ工学講座を担当しております。ビッグデータに関する研究は、社会への影響が大きいので、やりがいをもって日夜研究に励んでいます。本日は私が取り組んでいる研究内容とビッグデータの動向についても話していきたいと思います。まずは自己紹介ですが、私は今年6月までNTTの研究所で特別研究員として研究開発に取り組んでいて、7月から機会を頂いて大阪大学に移りました。これまでどんな研究をやってきたかというと、10年ほど前まではマルチメディアデータベースについて取り組んでいました。例えば画像検索、カラオケ検索、ホームラン検索などです。カラオケ検索では例えば私の上司がカラオケに行って歌を歌いたいが、リズムは分かるが題名が分からないから探せない。そのため当時、リズムを検索条件として入力して音楽を検索するシステムを開発しました。実際に実証実験をしてカラオケボックスに入れたことがあります。二つ目の例はホームラン検索で、2時間くらいある野球の番組の中からホームランのシーンを自動的に抽出しようというものです。野球は数十台のカメラで撮影されていて、ホームランの時はどのカメラの映像で、どうカメラアクションしたもの（ズームインなど）を組み合わせるかが決まっています。このようなホームランシーンのカメラ利用の特徴を利用して、ホームランのシーンを自動抽出するようなことをやっていました。その後にXMLの技術が普及したので、アラートシステム等に関するXMLデータを処理する技術に取り組みました。その後にウェブの検索システムをやりました。NTTグループでは、グーグルやヤフーと同様にウェブの検索サービスとして「goo」があります。検索サービスを実現する主要な技術として検索エンジンがあり、パソコンや携帯電話からのウェブ検索や画像検索を実現しています。その検索システムの中で、大量のウェブデータを格納してウェブ検索を実現するシステムを開発していました。このシステムは1000台規模のマシンを利用して、ウェブデータですから日本語のウェブデータですと100テラバイトくらいのデータを扱わなければなりません。このようなシステム開発の試験工程では、最初は10台、50台、100台と試験をしていくわけですが、マシン台数が増加していくと再現しにくいバグが出てきて、再試験するのに1週間程度を要するものもあり、大変だったのを思い出します。

現在どんなことをやっているかといいますと、グラフデータマイニングの高速化と分散処理の最適化に関する研究に取り組んでいます。ここでグラフというのは棒グラフや円グラフではなくて、例えばフェースブックにおける人と人の関係を表すようなグラフデータ構造、あるいはウェブページとウェブページの間の点と点を線で結んだようなグラフデータです。そうしたデータに対するクラスタリング処理や重要な情報を見つけるアルゴリズムの高速化について取り組んでいます。本日の内容ですが、まずビッグデータに関するトピックについて紹介し、その後に私が取り組んでいるグラフマイニングに関する応用事例と高速化の取り組みについて話したいと思います。

講師　鬼塚　　真 氏

●ビッグデータに関する話題

ビッグデータについて、国際会議などでよくネタになる話があります。ちょっと刺激的な言葉ですが、「Big data is like teenage sex」。これはビッグデータとティーンエイジャーのセックスをかけているわけで、それについてみんな話をしていて、他の人はそれがどういうものか知っていると言っていて、自分はよく知らないけど知っているよとうそぶいている状況を表しています。つまり、ビッグデータの実態は分からないのに、皆が知っているようだから知っていると言わざるを得ないという話です。もう1つは「Data is the new oil!」。新しい天然資源のように、データを活用することで価値を生み出せるということです。つまりデータは非常に重要だということです。ビッグデータを活用する上で重要なこととして、1つは良質なデータを保有することです。もう1つは集めたデータを使ってどのように分析するかです。例えば病院を例にとると、医療費の削減は至上命題です。病院のシステムには、電子カルテの情報もあれば、支払いに関する情報もあります。これらを統合して分析すると、例えば「軽いがんの医療費が高い」と言った意外な事実が分かることがあるそうです。つまり、今まで分析されていないバラバラにあるデータを統合して分析することによって新しい知識が見つかることにつながる。これは、そもそも医療データのような競争力があるようなデータがあって、かつどのように分析をするかによって、新しい発見を得ることができるという例です。また、企業に勤めていた時代には、お客様から「データはたくさんあるけど、どうにかこれを分析できないか」と言われることが時々ありました。しかし、データがあっても目的意識がないと良い分析はできません。本日の他の講師の講演の中でも出てくるかと思いますが、目的意識をもって仮説を立て実際に検証する、この仮説検証のサイクルを繰り返し回すことが必要になります。データがあるだけではだめで、データをどう使うか、どのあたりにキーポイントがあるかどうかを見すえたうえで仮説検証に回さないと分析は難しいし、良い結果が得られないということです。

●21世紀に注目されている職業「データサイエンティスト」

これもよく言われることですが、21世紀で最も注目されている職業がデータサイエンティストです。要するにビッグデータを分析することができる人であり、統計学、機会学習、データマイニングに詳しい、さらには分散処理に詳しいなど、多方面の能力を持っている人がデータサイエンティストだと考えています。これらの機会学習やデータマイニングの技術自身は10年前、20年前からありました。最近になってなぜこれらが脚光を浴びるようになったかというと、昔とくらべて膨大なデータを簡単に入手し活用できるようになったからです。実際、ハードディスクなどのストレージも安くなったので膨大なデータを低コストで格納できるようになってきています。もう1つ面白いのはムーアの法則に従ってCPUとメモリーの性能は年々向上していますが、データ量はムーアの法則より速く増加しています。つまり、ムーアの法則に従ってマシンが速くなったとしてもそれ以上のスピードでデータ量が増えてしまうので、このままではデータすべてを処理することができない問題があります。ビッグデータが注目されているもう1つの理由として、データ量を増やすことで分析精度が向上することが挙げられます。従来だと映像認識や音声認識の時に認識アルゴリズムを高度にチューニングすることで精度を上げていたのですが、単純にデータ量を10倍に増やすだけでアルゴリズムをチューニングなど工夫しなくても精度が上がることが分かってきた例があります。

●先端的な事例

皆さんがご存知のような典型的な例を挙げてみますと、コンピュータ将棋があります。コンピュータ将棋の特徴は何かというと、過去数年の対戦の棋譜データ、対戦データを入力にして、評価関数自体を

学習します。評価関数というのは将棋で序盤だったら駒得がいい、後半だったら王をとる方がいいというように、場面ごとに何を優先するべきかを判断する指標です。従来は評価関数を人間が作成していましたが、大量の過去データを利用して評価関数の重みをコンピュータに学習させることで、飛躍的にコンピュータ将棋が強くなったそうです。

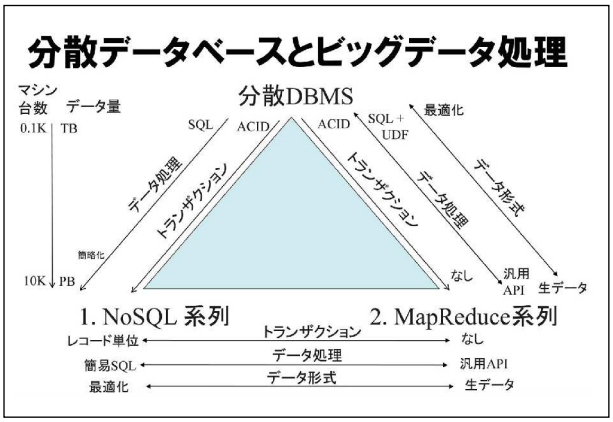
2つ目の例は、数年前に話題になったIBMの質問応答システム「ワトソン」です。日本なら昔の番組「クイズグランプリ」、今なら「アタック25」のようなクイズ番組において、コンピュータであるワトソンが人間と対戦して人間に勝利したという話です。このケースも番組の過去の正解集合を入力することと、もう1つは大量のコンテンツ。例えば歌詞とか聖書とか新聞とかウェブの情報を蓄積しておいて、やはり評価関数というものを過去データから学習することによって、人間に近い回答を決定するということをしています。余談ですが、最近では「ミニワトソン」という話があって、ウェブから誰でも使えるようになっています。

3つ目の例は、携帯電話の音声認識である「しゃべってコンシェル」です。音声認識に関しては、長年実用化が難しいという課題がありましたが、データが爆発的に増えたことによって、一気に認識精度が上がり携帯端末でも使えるようになったのは大きなブレークスルーだと思います。「しゃべってコンシェル」に関しても、膨大なデータを使って壁を越え、ようやく皆さんが使えるようなレベルになったということです。ここまでで、ビッグデータに関する大まかなトピックをおさらいしてみました。

皆様に近いところでですと、ビッグデータに関して比較的使いやすいのはアマゾンの環境のクラウドサービスがあります。クラウドサービスの中で主に2つのサービスがあり、1つはストレージ機能である大量なデータを格納する部分と、あとはストアしたデータを高速に分散処理してマイニングする部分です。それについて関係の説明をします。

●大規模データストア：NoSQL

1つは大規模ストアということで、NoSQLの話です。大量データを格納するということで分散データベースと何が違うかというと、分散データベースは100台規模くらいで動くものですが、NoSQLは



1000台や1万台で動くものであり、大規模計算機環境で使えるように機能が拡張されています。例えば数台のマシンが壊れても安定して動くような高可用性の機能が強化されています。その反面で従来のデータベースに比べると機能が簡略化されています。このように機能を簡略化することによって、マシンの拡張性や信頼性を上げています。どんなことを簡略化しているかというと、検索機能に関しては非常に簡略な機能しかないことと、あといわゆる簡略化したトランザクションしか保証しないこと。そうすることによって、必ずトランザクションが1マシンにしか行わないように簡略化し、高いスケーラビリティを達成しています。

●分散データ処理エンジン：MaqReduce

大量データを分析する代表的な技術にMaqReduceがあります。MaqReduceとは、分散処理のプログラミングモデルであり、かつシステムの両方をさしています。一般的には1万ノードマシンを用いてペタバイト級のデータ処理が可能です。従来の分散処理技術ですと、プログラムを記述する際には、どのマシンにどのタスクを割り当てる、あるいはどのマシンがダウンしたら、ダウンしたマシンに従ってどんなアクションをしたらよいかプログラマが実装しないといけません。対比的に、MaqReduceのフレームワークでは、これらの分散システム固有の複雑な処理はMaqReduceのフレームワークがシステム内に隠ぺいすることによって、プログラマはmap関数とreduce関数だけ記述するだけで、システムが分散処理をしてくれるし、負荷分散をしてくれます。また、マシンが落ちた時にも自動的に復旧してくれます。このようにMaqReduceを利用する

ことで、簡単に分散処理ができるようになっています。

MaqReduce自身はグーグルが開発したもので、もともとはウェブの検索エンジンのバックエンド用に開発されました。2008年時点で1日当り20PBのデータをグーグルでは処理しています。2004年に発表された技術で、すでに10年たっているため技術としてはそろそろ陳腐化していて、現在はMaqReduceの次の新たなエンジンが乱立する状況になっています。MaqReduceは一般的な分散処理エンジンなので、分析応用ごとに専用化された新たなエンジンが出現してきています。例えばメインメモリを活用して高速処理を実現する分散エンジンや、あるいはグラフデータ処理に特化したような高速エンジンなどが出てきています。

●グラフデータの有用性

ここからグラフデータに関するマイニングの話をします。現在の世の中にあるデータ構造は従来と比較して複雑になっています。例えばウェブのデータ構造、フェースブックにおけるソーシャルグラフ、友人関係のデータ構造、あるいは写真などという多様なデータ構造が増えているし、携帯端末などの急速な普及が背景にあります。そうすると従来の単純な表形式、いわばエクセルやデータベースのような表形式では、表現能力・処理能力に限界があります。そこで我々が注目しているのは、人、物、場所といった多様な情報のつながりを表現可能なグラフ構造です。例えば誰がどこで何を買ったといった情報をグラフ構造として表すことによって、そこから知識をマイニングしようと考えています。そのため大量のグラフデータを分析することが重要になってきています。

●高速グラフマイニング技術

大量なデータとしてどんなものがあるかというと、ウェブのグラフデータは全世界で100億ページ、1PB（ペタバイト）くらいあります。日本のウェブサイトでも100TB（テラバイト）くらいあります。フェースブックのソーシャルグラフは12億人規模です。アマゾンの例ですと、購買ログ、利用者ログということで、誰が何を買ったかという、利用者とアイテムのグラフをつくる大きなグラフデータです。このように巨大なグラフデータが出てきている状況にあります。

しかし、グラフマイニングに要する処理のコストは大きくて、階層型クラスタリングを考えてみると$O(N^2)$ということで非常にコストが高いという話があります。もう1つの代表的な分析方法としてページランクの計算があり、これは行列で一次元方程式を解く計算に相当しますが、これはランダムウォークの処理に相当します。ランダムウォークの処理、あるいは行列計算の繰り返し型で解くような処理は、グラフのサイズ×繰り返し数で処理コストが効いてくるので、やはり処理コストが高いという問題があります。だから高速なグラフマイニング技術を開発することが非常に重要だということになります。

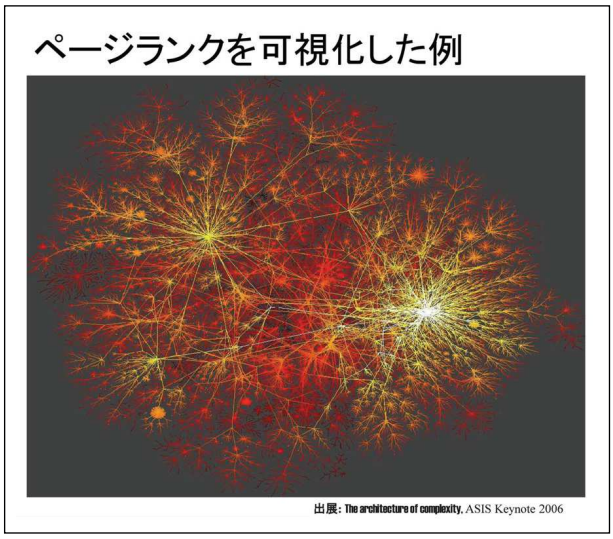
●我々の研究の取り組み

我々は3つの取り組みを行っています。1つはグラフクラスタリング処理の高速アルゴリズム、1つはページランクのトップK検索の高速化、もう1つが繰り返し型分析の分散高速化です。分散環境、複数のマシンで扱う時の高速化の最適化です。これらは全て世界のトップ会議で発表してきています。

●グラフマイニングの事例1

ここからはグラフマイニングの事例について見ていきたいと思います。1つはNHKの震災ビッグデータの例です。東北で震災があった後の企業間の取引を分析することで、震災の後にどの企業が早く復興しているかどうか、それはどういう企業がキーであるかを分析する例であり、「NHKスペシャル」で放送されています。NHKにはCGの部隊の方がいて、企業間の取引データを帝国データバンクのデータを使い、それをCGで可視化して、復興後に早期に立ち上がった企業を明らかにするということをやっています。これが可視化した図ですが、各点が企業に相当します。詳細にフォーカスして、どこの企業がキーの企業かを可視化することによって分析しているわけです。

2つ目はグーグルのランキング計算の典型的な例です。ウェブページの重要度を計算するのにページランクを使います。ページランクを計算して重要なウェブページにいくほど明るく可視化すると、この図のような感じになります。これは全世界のウェブ



ページを可視化した例で、点がウェブページを表し、線がウェブページ間のハイパーリンクを表したものです。明るい濃度ほどページランクのスコアが高いということで影響力があるページです。

3つ目はソーシャルネットワークの分析で、いろいろな例があるのですが本日のスライドでは、ダークネットワークの分析例を示しています。テロリストやバイヤーのネットワークを解析することで、中枢の人物を見つけ、とらえることで活動を阻止するというような例です。これはCACM（Communications of ACM）の記事に掲載されていた可視化の図ですが、「グローバル・サラフィ・ジハード・ネットワーク」といって、ピンクの色がビンラディン・グループで、黄色はアラブとか、青、緑などに色で分けてそれぞれのチームを可視化したものです。2002年バリ島爆弾テロ事件で関与しているのはこのグループだと分析して分かったことと、アメリカ同時多発テロはこのチームだと分析しているようです。うまくデータを分析して、可視化するといろんなものが見つかるという例であります。

●グラフマイニングの事例2

ここまで話したのは世の中の一般的な例でしたが、我々が取り組んでいる分析例は、1つは人間関係の分析ということで、有名人とか政治家のソーシャル関係の分析です。もう1つは論文の技術トレンド分析です。論文のメタデータを使って技術年表の自動生成をするということをやっています。この2つについてデモンストレーションをしたいと思います。

基本的には多様なグラフデータをグラフ分析すると、例えば政治家の勢力分析ができるとか、例えばこれは東京23区の道路網をメッシュ分割し、交通渋滞状況に応じて均等に負荷が同じように流れるように均等分割しています。道路のシミュレーションをして信号整備などをする際に、極力負荷を分散したいので道路網のメッシュ分割をします。もう1つはトレンド分析ということで、論文を入力にして技術トレンドを分析することを考えています。

●利用例1：Wikipediaの分析

これはその1例目です。ウィキペディアでは1ページ毎に1人の情報が記載されていて、それらがハイパーリンクでつながっているので、グラフをつくることができます。1回グラフをつくって、そこからクラスタ分析またはページランク分析をすることで、仲間を見つけたり、影響のある人を見つけることができます。この図が結果であり、データの規模としては3000人程度の情報となります。この図では円の大きさがページランクにおけるスコアを表していて、実際には影響力のある人というのが大きい円で表されています。また同じ色によって同じクラスタリング結果、いってみると同じコミュニティということを表しています。意味づけとしてはこの図のようになっています。この辺りが男性俳優、女性俳優、宝塚、ジャニーズ系、ハロプロ系、AKB系、こんな感じです。

少しフォーカスして男性俳優を見てみましょう。影響力が高い例として、勝新太郎さん、ここは三谷幸喜さんが出てきました。影響力ランキングからするとメジャーの方が上に多く出てくることが分かります。勝新太郎さんは映画・舞台系で影響力のある人で、三谷幸喜さんは脚本家なので、ドラマとか映画の両方に関係するところにレイアウトされています。次はジャニーズ系です。ジャニーズ系をクローズアップするとやはりスマップが多いことと、森光子さんはスマップと非常に仲がいいということで、ランキングが上の方に出てきているようです。

この図は政治家のソーシャルグラフを分析した例です。この箇所は安倍晋三首相と麻生太郎元首相を表しています。誰と誰が関係しているかどうかのデータをウィキペディアから抜いてきます。このようにリンク情報だけを入力して、gephiというツール

にリンク情報を入力すると、まずこのような特徴のない可視化結果が出てきます。各点が人を表していて、人と人が線によって接続されています。この入力データに対してgephiで分析することができます。gephiはグラフ分析に関する多様な機能を有していて、直径を計算したり、クラスタリングしたりといろんな機能があります。

●利用例2：論文データから技術年表生成

2つ目の例としては、自動的に技術年表を可視化するというものがあります。例えばIT分野の論文データを使います。DBLPというような論文のメタデータで、80年分くらいのデータがあります。そうしたDBLPの情報からこの図のようなグラフをつくります。中央は論文がずらっと並んでいて、この論文を書いた著者が3人いる。もう一方で論文のタイトルに入っている単語は何かを表します。この論文のタイトルには2つの単語が含まれている。このような3種類の情報から（3部）グラフをつくって、これに対してクラスタリングしてページランクによって重要度を計算すると、自動的に技術年表がつくれるというものです。

可視化結果はこの図になります。横軸は、1950年、60年、70年、80年、90年、2000年、2010年ということで、10年ごとに論文群を分割してからクラスタリングします。各年代毎に縦軸に列挙されているものがクラスタです。そして、隣接する年代間において類似度の高いクラスタ同士をリンクで接続します。そうしますと、このような技術年表の結果を得ることができ、例えばネットワークの研究というのを見ると、このような形で遷移しているのが分かります。

ここまで分析例について説明してきました。ここから我々が取り組んでいる技術について概要を説明します。

●クラスタ分析とは

クラスタ分析とは、関連性の高いデータの集まりを自動的に見つけるというものです。左側にある人と人の間の関係あるいはウェブページ、ある点と点の間を結ぶ線で情報を入力して、クラスタ解析をする。大雑把にいうとソーシャルグラフの場合ならクラスタ分析によって仲間を見つけてくれる。例えば、この図では野球派、サッカー派、バスケ派、テニス派というように仲間を見つけてくれる例を示しています。クラスタ分析の処理の直感的な意味としては、同じクラスタの中のエッジ数（線の数）が多ければ多いほど良いクラスタであり、異なるクラスタ間のエッジ数が少ないほど良いクラスタになります。このようなクラスタ分析の指標値を決めて、指標値ができるだけ大きくなるようにグラフをグループ分けする処理がクラスタ分析です。

●クラスタ分析高速化技術

我々はどのような高速化をやっているかというと、グラフに対しどういう順序でクラスタリングを計算したらよいかどうかを最適化し、次に2つの点を同じクラスタにすべきかどうかが判定された時点で、逐次的に2つの点を集約してクラスタリングをしていきます。このような2つの工夫をすることによって、従来技術に比べると15倍から50倍くらい速くなっています。横軸がデータの種類、縦軸が応答時間を表していて、従来手法に比べると我々の手法は全てかなり速いという結果を得ています。

● Personalized PageRank（PPR）とは

次にパーソナライズ・ページランクについてですが、ページランクというのは人の影響力を表す指標です。更にパーソナライズ・ページランクというのは、利用者ごとにページランクを計算する方法です。例えば、この図で説明しますと、この2人に影響を及ぼす人は誰かを計算するようなものです。パーソナライズ・ページランクのケースで、特に上位10件とか上位100件を特定するために実際にこれを入力にすると、この辺りの近くの人が見つかることに

なります。これを解くためにノード群を入力します。ページランクの計算は基本的には行列計算で解くことができるので、ここでは行列分解をして直接法で解くことをします。

● Top-k PPR 高速化技術

次に上位10件に入り得る候補データを選別して、最終的には上位10件を正確に特定するということをしています。データを選別する際には、各データのページランクスコアがとり得る上限値を見積もって、その上限値が上位10件に入らないことが判定できたら、そのデータの計算をスキップします。このようにすることで、上位10件に入り得る可能性があるデータだけ正確にページランクを計算することによって、従来よりも処理を高速化することを達成しています。

●繰り返し型処理の分散クエリ最適化技術

最後の技術は分散コンパイラに関する技術です。特に繰り返し型の処理、例えばk-meansクラスタリングのような繰り返し型処理、あるいはページランクの計算のような繰り返し型の処理を対象として、無駄な処理を自動的に見つけて排除するという技術です。繰り返し型の処理ですから、処理構造はこの図が表現するようなフローによって表現され、最初は初期化をして、指定された処理が繰り返されて、収束したら処理が終わるという形になります。この処理フローを入力として、2つの種類の無駄な処理を排除します。1つ目の工夫では、入力のプログラムの繰り返し処理の部分を分解して、ループに依存する部分と依存しない部分の2つに分けて、ループに依存しない部分に関してはループの外に追い出して1回計算して結果を蓄積し、その結果をループ内から再利用します。2つ目の工夫に移ります。

ページランクでもk-meansクラスタリングでも、繰り返しで処理をすると大半のデータについてはスコアあるいは重心点が収束していきます。2つ目の工夫では、この収束しているデータと収束していないデータを分離して、収束したデータの処理をスキップすることによって、処理量を減らしています。これら2つの工夫がどの程度効果があるかを示したのが次の図です。この図では時間経過に従って性能がどう変化するかを見ているのですが、従来手法に比べると1つ目の工夫と2つ目の工夫を行うと、処理量は非常に減っていることが分かります。

●まとめ：ビッグデータの潮流

最後にまとめますと、データ活用が大きなビジネスチャンスになると思います。とくに良質なデータ、他の人が持っていないような情報を持つことが重要だということと、もう1つは複数のデータを統合して分析することが重要になります。最近我々が研究として取り組んでいるのは、大規模データを分析・可視化するということです。従来のような単純な表構造データではなくて、人・物・場所といった多様な情報のつながりをグラフで表現して、そこからクラスタリング、あるいはページランクのような影響力のあるものを高速に実行するアルゴリズムやシステムの研究を行っています。クラスタリングに関しては1億ノード規模、日本の人口規模くらいのソーシャルクラスのクラスタリングは数分程度で実行することができます。私は今後、可視化の部分、分析の部分をつなげていくことに取り組んでいきたいと思っています。また、こうした場で産学連携の形でコラボレーションをしながら進めていきたいので、今後ともよろしくお願いしたいと思います。