グラフマイニング技術を用いたビッグデータ分析技術とその応用

研究室紹介

鬼塚　真*

Graph Mining Techniques for Big Data Analysis and Applications

Key Words：Big Data, graph mining, analysis, IoT

はじめに

Data is the new oil（データは新しい石油である）という言葉に代表されるように，ビッグデータを分析することで隠れた知識を発見し，社会的あるいは経済的なインパクトを生み出すことが期待されている．つまり，データは石油と同様に資源として活用することができるという考え方である．近年，ウェブやソーシャルメディア・センサ等の普及などに伴って，データの単位は従来の単純な表構造データから，実世界における多様な関係を表現可能なグラフ構造へと変化しつつあり，例えば人・モノ・場所からなる大規模なグラフ構造のデータを高速に分析することが重要な技術課題となっている．このような背景を捉え，我々は大量の計算機を効率的に利用する分散データ処理技術および高速グラフマイニングアルゴリズムの研究に取り組んでいる．更に，開発した技術を応用するため，認知症患者のケア情報の推薦に関する応用や，IoT（Internet of Things: モノのインターネット）を用いたスマートショッピング応用に取り組んでいる．本稿においては，これらの要素技術と応用事例について詳細に説明する．

1．高速グラフマイニング技術

近年，ウェブやソーシャルメディアなどの多様な情報の増加や，スマートフォンなどの端末の急速な普及に伴い，実世界における多様な関係，例えば人・モノ・場所の関係を表す大規模なグラフデータを高速に分析することが重要な技術課題となっている．例えばウェブグラフは全世界で100億ページ・1PBを超えると言われており，実際にcommon crawl corpus（50億ページ・541TB）はウェブページのアーカイブとしてamazon web serviceにおいて公開されている．世界最大のソーシャルグラフとしてはFacebookのソーシャルグラフが挙げられ，ユーザ数は15億人であると報告されている．このようなグラフデータを分析処理する代表的な方法として，クラスタ分析やPageRank計算による影響力分析が挙げられる．図1中央の政治家のソーシャル分析の例では，Wikipediaから得た政治家同士の繋がり（ソーシャルグラフ）をクラスタ分析することで政治家間の派閥などのコミュニティを抽出し，更にPageRank計算によって影響力を分析した結果を可視化した結果を表している．図における一つの円が一人の政治家を表しており，円の色はクラスタに相当し，円の大きさはPageRank値により政治家の影響力を表している．この図から得られる分析結果として，石原慎太郎氏の円が大きい，つまり影響度が大きいということと，色が他の政治家と異なるため特定の派閥には属していないということが挙げられる．一方，図1右は計算機科学の分野における過去50年間の技術の変遷を分析した結果である．書誌情報に関する論文データを年代毎に分割し，年代毎の論文集合をクラスタ分析することで技術領域を抽出し，隣接する年代間のクラスタの類似性を算出する際にクラスタに含まれる論文のPageRank値を用いることにより，精度の良い技術領域の時間変遷を導出している．

グラフデータのクラスタ解析やPageRank計算は計算量が膨大であるため，大規模なグラフデータに

* Makoto ONIZUKA
1968年8月生
東京工業大学　大学院情報理工学研究科
博士
現在、大阪大学　大学院情報科学研究科
マルチメディア工学専攻　教授
博士（工学）　データベースシステム
TEL：06-6879-7739
FAX：06-6879-7743
E-mail：oni@acm.org

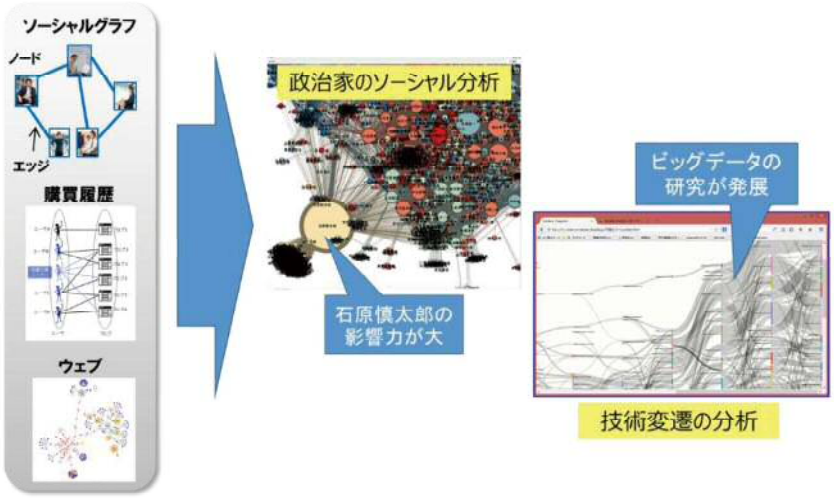

図1．グラフマイニングの分析例

対して高速にグラフマイニング処理を実行することは極めて重要な課題である．この課題に対して，我々は従来技術よりも2～3桁高速なクラスタ解析やPageRank計算のアルゴリズムを開発し，また大量な計算機を用いて高速にマイニング処理を実行する分散クエリ最適化の技術を開発してきた．その結果，データベース領域やデータマイニング領域における最難関の国際会議において多数の論文を発表してきており，特にデータベース領域で最難関のVLDBに2012から2015年まで4年間連続して論文が採択されている．

2．ビッグデータ分析の応用事例

グラフマイニングなどのビッグデータ分析の応用事例として，我々が現在取り組んでいる，1）大量に収集した認知症患者のケア情報の中から効果的なケア方法を発見して推薦する応用，および 2）センサやアクチュエータを活用することで，IoTを用いたスマートショッピングを実現する応用について紹介する．

2.1．認知症患者のケア情報の推薦

わが国では現在認知症患者が450万人以上おり，増加しつつある状況にある．このため，認知症患者に対するケアのノウハウを整備することは喫緊の課題である．しかし，現状ではケアに関するノウハウが整備されておらず，介護者は成功や失敗を繰り返しながら認知症患者のケアに取り組んでいる状況にある．このような背景を捉えて，我々は情報共有サイトを立ち上げ介護者からケア情報を収集し，得られたケア情報を分析することで有効なノウハウを自動抽出して介護者に提示する研究を進めている．具体的には，介護者からケア情報として，（直面した問題事象，実施した対処策，対処策の奏功）などの情報を収集し，これらのケア情報をクラスタ解析することでケアのノウハウを類型化する．更に，得られた類型（クラスタ）毎に奏功結果に基づいてケア情報のスコア付けを行うことで，代表的なケア情報を特定し介護者に提示する．H27年度は大阪や熊本において実証実験を行い，H28年度からは全国規模での実証実験を実施する予定である．

2.2．IoTを用いたスマートショッピング

加速度センサやBeaconなどの多様なセンサの進化に伴い，これらのセンサから取得される利用者の詳細な行動ログを収集して，実世界サービスと融合した情報推薦や行動ログからの人の行動に関する暗黙知の抽出，あるいはIoTと連携した新たなアートが創出される例が出現してきている．IoTを活用したスマートシティという観点においては，スペインのサンタンデル市のSmartSantanderプロジェクトの取り組みにおいて，市内に18000個のセンサを設置して多様なサービス（空き駐車場への車の誘導サービス，電子クーポンを用いた店舗推薦，バスの運行状況の把握）が実現されている．このように現在進化しつつあるセンサやアクチュエータを活用することで，我々は利用者の行動ログを収集し，どの利用者がどの店舗に訪れたかという情報を分析するこ

とで，利用者に適した店舗を推薦するスマートショッピングの研究を進めている．具体的には，Beacon を用いて利用者がどの店舗に何分間滞在したかの情報を収集し，利用者と店舗に関する情報を行列分解することで，利用者が訪問していない店舗の滞在時間を推定し，滞在時間の長いと推定される店舗を推薦する．更に，店舗の外的環境（照明，音楽，アロマ）が利用者の滞在時間に影響する要因を分析し，利用者の滞在時間が長くなるようアクチュエータを用いて店舗の外的環境を制御する．H27 年度はグランフロント大阪の研究展示スペースである The Lab. において実証実験を行い，H28 年度からはより広範囲なスペースで実証実験を実施する予定である．

今後の展望

ビッグデータを分析することで社会的あるいは経済的なインパクトを生み出すためには，多くの応用において時間をかけて分析技術を改善あるいは試行錯誤を繰り返しながらチューニングする必要がある．実際に IBM 社の質問応答システム Watson や Netflix 社のビデオ推薦の事例では，分析精度を向上するために数年が費やされている．今後の研究の方向性としては，人間が行う試行錯誤の工程を自動化する技術が重要であると考えられる．例えば，機械学習におけるハイパーパラメータの最適化や，多次元データ分析において特徴的な分析パターンを自動的に探索するなどの課題が挙げられる．一方，ビッグデータの分析システムに関しては，最新ハードウェアを活用した高速化と低コスト化の観点で技術が発展するとともに，グラフデータ構造や応用毎に専用化された要素技術が今後も発展するものと考えられる．