KDD2017会議報告

横山 大作

KDD 2017 (23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining) は，知識発見，データマイニングなどに関するトップレベルの国際会議である．近年のトレンドであるビッグデータ，機械学習に関わるトピックも多く含まれ，注目が高まるとともに年々その規模を拡大している．ここでは，2017 年 8 月に開催された会議について紹介する．

1 会議概要

KDD はデータからの知識獲得・利用などのトピックが対象であり，主としてアルゴリズムに関する研究が扱われるが，後述するように実装やアプリケーション，データサイエンスなどの応用分野に関する発表も多い．1 年に 1 回，北米を中心に開催され，本年は 8 月 13 日から 17 日の日程で，カナダの東岸，ハリファックスにて行われた．1 日目はチュートリアル，2 日目はワークショップが行われ，本会議は 3 日目から 5 日目というスケジュールであった．

ビッグデータ，機械学習などのキーワードに呼応して，この分野に対する注目度は近年ますます高まっている．会議最終日の発表によると今年の参加者は 1675 人であり，サンフランシスコで行われた 2016 年の 2792 人 (過去最多記録) には及ばないが，米国以外で実施された中では最大の規模となった．論文投稿数は総計 1143 本 (過去最多) であり，採録率はオーラル発表 8.8%，ポスター発表 10.2%であった．

KDD2017 Conference Report.

Daisaku Yokoyama, 東京大学生産技術研究所, Institute of Industrial Science, The University of Tokyo.

コンピュータソフトウェア, Vol.35, No.1 (2018), pp.86–89.
2017 年 9 月 6 日受付.

多数のスポンサーがあり，本年は合計 54.4 万米ドルの提供が集まった (過去 2 番め，米国外では最多)．昨年に引き続き，中国のシェアライド企業である DiDi がダイヤモンドスポンサーであった．プラチナスポンサーはアマゾン，マイクロソフト，アリババ，Facebook の 4 社で，以下，グーグル，SAS や Oracle などの IT 系を中心に，SIEMENS などの工業系，アメリカン・エクスプレスなどの金融系も含む幅広い企業がスポンサーに並んだ．

この会議の目的として強調されていたのは，“Research and Application”, “Science and Industry” の間を橋渡しすることであった．この研究分野は実世界で発生する様々なデータを取り扱うことが目的であり，アカデミックとデータを持つ企業，自治体等との距離が非常に近い．最先端技術を追求したい研究者と，それをすぐさま実問題に応用したい利用者とが集まり，情報共有とネットワーキングをする場としてこの会議がある，との目的意識である．このために会議構成にいくつかの工夫がなされていた．

- 論文募集は Research Track と Applied Data Science Track への 2 つの枠組みに分けられ，それぞれの基準で採否が決定される．前者はアカデミアに馴染み深いいわゆる研究論文の募集であり，新規性などの基準で判定されるのに対し，後者は企業，自治体などにおける実問題への応用に関する研究が求められ，「実世界の難しい問題に適用した事例」「単に性能が上がった等ではない，実データ利用によって得られた新しい発見」「実問題への応用可能性を明確にした手法やシステム」というカテゴリの研究が募集された．また，論文募集では双方とも「研究の再現性」を

表 1 トラックごとの投稿・採録状況 (カッコ内は採録率)

論文トラック	投稿数	採録数: oral	採録数: poster
Research	747	64 (8.6%)	66 (8.8%)
Applied Data Science	390	36 (9.2%)	50 (12.8%)

明確にするように，との要求があり，近年の研究論文における問題意識が感じられる．それぞれのトラックでの採録率等を表 1 に示す．Applied Data Science Track の方が若干採録率が高いが，いずれもポスターを含めて 20%そこそこの難関会議である．

- Applied Data Science Track では，スーパーマーケット経営の Target がデータ分析チームの紹介をするなど，社会での応用事例を中心に 10 件の招待講演が企画された．
- ネットワーキングのために，研究やキャリアについてエキスパートと相談できるラウンドテーブルセッションや，中国，インドなどの地域ごとの参加者交流用セッションが設けられた．
- 学生に加えて，スタートアップ企業のための旅費補助グラントが設けられた．
- Job matching の場が用意され，あらかじめ希望などを登録した参加者と企業との間でマッチングが行われた．今年は 627 人がマッチングされたとのことであった．また，会議参加者用に準備されたモバイルアプリでは，プログラムや会場マップなどに加えて交流用のメッセージングサービスが提供されていたが，全参加者に向けたチャンネルでは企業からの求人が大量に投稿され，この分野の盛り上がりが感じられた．
- 大規模データ解析ツール，フレームワークなどに関するハンズオンチュートリアルが多数用意された．

また，この会議では “Data science” がキーワードとして多くの場所で用いられていた．オープニングにおいて Google trend (Google でユーザが検索のために入力している言葉の頻度の移り変わり) が紹介されていたが，“Data mining” は 2012 年頃からほぼ横ばい，ここ 1 年程度はやや減少しているのに対し，2015 年頃から “Machine learning”， “Data science”， “Deep learning” が次々と “Data mining” の検索数を抜いていっており，このコミュニティを取り巻く環境が変化していることが示唆されていた．

2 招待講演

本会議では以下の 3 件の招待講演が行われた．

Bin Yu (UC Berkeley) は “Three Principles of Data Science: Predictability, Stability, and Computability” として，データサイエンスを科学たらしめるための要件について話し，特にデータとモデル双方において stability が必要であると述べた．stability とは，入力を変動させた時の出力への影響の度合いであると定義し，機械学習に用いるデータが変化した場合，あるいは適用するモデルを他のものに変えた場合でも得られる結果に現れる影響が少なくなるような手法を目指すことが，科学であるために必要であると強調していた．

Cynthia Dwork (Harvard & MSR) は “What’s Fair?” として，公平性を担保した機械学習に関する講演を行った．異なるグループ間での判定割合が等しくなることを求める Group fairness，似た個人どうしは等しく扱われることを求める Individual fairness の概念を定義し，これらを担保するような判定手法が求められていることを紹介していた．

このような公平性は，この会議でもワークショップが併設されるなど注目度が上がっているトピックである．機械学習による判断過程は人間にとってブラックボックスとなりつつあり，誰もが納得できる判断基準となっているのか，不公正な判断が紛れ込んでいないか，を確認したいという要求が強まりつつある．例えば EU は，「AI による判断が行われる場合，判定理由の説明を受ける権利がある」とする EU General Data Protection Regulation を制定し，2018 年 5 月

から発効する予定である †1．招待講演はこのような背景を反映したものと言える．

Renée J. Miller (Toronto) は “The Future of Data Integration” の講演を行った．複数のデータソースから情報を抽出して，共通 ID 等を通じて結合し，結果を統一的なスキーマで取得するための手法について，関係代数的な整理を行ったこれまでの研究について紹介した．近年提供が増えているオープンデータの利用について，データの基数 (cardinality) がデータソースごとに非常に異なった歪んだ分布をしており，極めて大量の種類のデータからごく数種類のデータで構成されたデータベースまでを取り扱わなければならないことが今後の課題である，としていた．

3 論文発表セッション

採録論文に関して，会場でいくつかの項目に関する word cloud が示されており，著者が設定した論文のトピックでは “temporal and time”, “series data”, “graph algorithms” などが多数を占めていた．また，論文タイトルでも networks, time, series などの単語が見て取れ，時系列データやグラフデータの解析が現時点で課題となっていることがわかる．アブストラクトになると social, deep などの単語の出現頻度が上昇しており，深層学習の流行がやはり感じられたが，全体としては深層学習一辺倒ではなく，システムの部品として当たり前に組み合わせて利用するような研究が多いように思われた．

研究内容としては大規模グラフや時系列データに対するクラスタリングやモデル化，教師あり学習などのアルゴリズムに関するものが多いが，スポーツや映画製作，都市設計などの新しい領域への適用事例や，大規模データ処理基盤に関する発表も見られた．以下，私が興味を持った発表をいくつか紹介する．

Platforms and Infrastructure というセッションでは，Google, アリババ等が社内で利用しているデータ解析のためのプラットフォームに関する事例紹介があった．[Jun Zhou et al., KunPeng: Parameter Server based Distributed Learning Systems and Its Applications in Alibaba and Ant] はアリババの事例である．アリババが提供する電子決済プラットフォーム Alipay では，2016/11/11 (中国での「独身の日」，EC が最も盛んになる特異日) に 10 億トランザクションの決済が行われ，マスターカード等クレジットカード会社を超えた規模に成長している．そのような大規模データのためのフレームワークシステムの説明は迫力を感じさせ，多くの参加者を集めていた．Google が発表した [Denis Baylor et al., TFX: A TensorFlow-Based Production-Scale Machine Learning Platform] は，深層学習用フレームワークである Tensor Flow の拡張のような印象のタイトルではあるが，機械学習のアルゴリズムを適用する環境のみならず，収集データ，実験用中間データのマネジメント，可視化等，大規模データを使って実験を行いアプリケーションを構築するまでの面倒を見るためのフレームワークの紹介であった．日々の実験で作られる様々な中間データについて，GC を行って管理しているとの言及もあり，データのライフサイクル全体を扱うプログラミング環境のような側面が感じられた．

Applied Data Science Track の最優秀論文である [Yanfang Ye et al., HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network] は，マーケットにあるアンドロイドアプリの静的解析を行い，コードブロック内で利用している API の組み合わせをグラフとして特徴量化し，マルウェアと善良なアプリの 2 つのクラスに分類する学習を行った研究である．ソフトウェアセキュリティの研究においては，動的な振る舞い解析や精密な静的解析などの技術はもちろんより進展していると想像でき，その意味での先進性は私には判断できなかったが，現存している大量アプリケーションを解析対象にし，データ解析からのアプローチで迫っているところが評価されたのかな，という印象を受けた．

Social good を追求する研究も毎年いくつか見受けられ，印象深く感じられた．[Rebecca S. Portnoff et al., Backpage and Bitcoin: Uncovering Human Traffickers] は，Bitcoin のトランザクションログを解析して同一人物のものを見つけ出そうという研究で

†1 http://www.privacy-regulation.eu/en/13.htm

あった．性的搾取などの犯罪行為に関わるような広告は多く，どの広告がどのようなグループにつながっているのか，その構造を調べることは難しい．この研究は，その資金の流れを調べ，あまたある Bitcoin の wallet(匿名取引のための「財布」) がどのようなグループ構造を持っているのかをヒントに，広告の構造に迫ろうとするものであった．[Jacob Abernethy et al., A Data Science Approach to Understanding Residential Water Contamination in Flint] は，昨年問題になった，米国 Flint 市の水道水に鉛が溶け出し中毒を引き起こした問題について，住民が水道水の危険度をチェックできるようなサイトの構築過程とアルゴリズムを紹介していた．住居につながるような末端の細い水道管については，その位置や規格等が記録されておらず危険度の推定が困難であったが，市庁舎で打ち捨てられていた紙の図面が発見され，システムに入力し利用可能となった過程などが紹介され，聞いていてワクワクするような内容であった．

このような Social good への取り組みは，研究者が行政などの機関と深く連携し，長時間にわたって取り組むことで初めて，価値ある，迫力のある結果が得られるものである．研究の姿勢について考えさせられるものがあった．

4 コンテスト等

多くの会議と同様，この会議でも中国の存在感は強いものがあった．トップスポンサーである DiDi，アリババなどの中国企業の存在感や，中国人発表者，参加者の割合の多さも近年のトレンドという感がある．驚いたのは，KDD Cup という会議併設データ解析コンテストでの上位入賞者が全て中国から出ていたことである．このコンテストは KDD では恒例となっており，今年のコンテストは有料道路の料金収受に関わるデータが与えられ，交通量や平均通過時間を予測するという 2 つのタスクについて競われた．データセットが中国の道路で取得されているため，何らかの地理的，文化的な知見の有無が成績に影響した可能性はあるが，2 部門のトップ 3 が全て中国で占められるというのは，情報教育に対する注力が感じられる結果であった．

また，KDD で採録された論文はそれぞれにウェブページが設けられ，会議開催前に無料で論文が公開される．このページは論文著者にコンテンツ編集権限が与えられ，ソースコードや実験データなどの公開の他，読者とのコメント機能を利用した議論も可能である．さらに，著者には論文内容紹介のための 2〜3 分間のビデオの作成が求められ，前述のウェブページ及び KDD2017 Youtube channel で公開される．この視聴回数上位のビデオについて内容，創造性，品質を基準に審査が行われ，最優秀作品には 1000 ドルの副賞が出るコンテストが開催された．最優秀作品 [Daniel Hill et al., An Efficient Bandit Algorithm for Realtime Multivariate Optimization] が会場で上映されたが，たいへん素晴らしいものであった．アイテムの組み合わせを考慮したバンディットアルゴリズムに関する研究であるが，理論研究をこんなに面白く説明できるのか，と感心する必見作品である．

5 終わりに

採録論文等は KDD 2017 のウェブページ，Accepted Papers リストから無料で取得可能である．次回 KDD2018 は 2018/8/19〜23 にロンドンで開催される予定である．また，再来年の KDD2019 はアンカレッジで開催されるとのアナウンスがあった．

KDD は本学会のコミュニティには必ずしも馴染み深いわけではない会議であろう．しかしながら，アルゴリズムとアプリケーションの幸福な関係を知ることは，今後の計算機科学の方向性を探る上で有用であると確信している．本誌の読者にとってこの報告が役立てば幸いである．

横 山 大 作

2006 年東京大学より博士号取得．博士 (科学)．2002 年より同大学新領域創成科学研究科助手などを経て，現在同大学生産技術研究所特任助教．並列・分散プログラミング環境，ゲームプログラミング，データマイニングに関する研究に従事．ACM，IEEE CS，情報処理学会各会員．