会議報告

The 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining（KDD 2017）

開催地：World Trade and Convention Centre & Scotiabank Centre（ハリファックス，カナダ）
開催日程：2017 年 8 月 13 日（日）～ 17 日（木）
http://www.kdd.org/KDD2017/

1. KDD 2017

KDD はデータマイニングに関する国際会議で，この分野では最難関会議と位置付けられている．ICML や NIPS などの機械学習の国際会議ではデータマイニングに必要なアルゴリズムや手法が中心である．それに加え，データマイニング分野の会議では具体的な実問題も対象とし，その定式化やモデル化についての発表もなされる．ほとんどは北米で開催され，その他の地域で開催されたのは 3 回のみである．筆者は 10 回目の参加で，2009 年以降は続けて参加している．

開催地のハリファックスは，カナダの東海岸にあるノバスコシア州の州都である．カナダ東海岸の中心地である港町で，ロブスターが名物のようだ．カナダでの開催は，1995 年の第 1 回モントリオール，2002 年の第 8 回エドモントンに続き 3 回目となる．会場は町の中心部にあり，ホールはホッケー場にも使われているように見えた．今年も IT 系企業を中心としたスポンサーが多数あり，提供金額は約 54.4 万 USD であった．中国のライドシェアリング DiDi は今年もダイヤモンドスポンサーで，日本からのスポンサーはなく，ここ数年いた BOSCH がいなくなっていた．非 IT 系では American Express が加わっている．

去年から 5 日間の開催となり，初日がチュートリアル，2 日目がワークショップ，そして本会議が 3 日間である．2014 年ニューヨーク 2 134 人や 2016 年サンフランシスコの 2 792 人には及ばないが，米国外での開催では 2015 年のシドニーの 1 182 人を上回る 1 675 人の新記録とのことだった．筆者の見た印象では，日本からの参加者は 50 人を超えていたようだ．

去年，大幅に変更された企画を今年も踏襲していた．トラック名は去年から Research と Applied Data Science という名称になり，通常のチュートリアルに加え，ハンズオンチュートリアルも行われた．SIGKDD にはインド，中国，豪 + NZ の国別チャプターと，オースチンやシアトルの地方チャプターがあるが，シンガポールが新たに加わったようだ．ネットワーキングのため，著名研究者と若手研究者が対話できる Networking with Experts Forum という企画もあった．

2. 招 待 講 演

3 件の基調講演があった．最初の Bin Yu の講演では，計算機科学，数学·統計学，ドメイン知識を融合したデータ科学についてであった．機械学習を用いた結果を，科学として成立させるために結果の安定性の重要性を強調していた．安定性は，科学としての再現性や解釈可能性が担保のために必要なもので，データの小さな摂動に対して結果が大きくは変わらない安定性が求められる．続いて，脳科学分野で，スパース学習を用いた研究事例の紹介があった．一つ目は，映画を見たときの fMRI で計測した脳信号との対応を分析したもので，もう一つは，深層学習で抽出した画像特徴量から，脳の高次視覚野の信号を予測できるかといった問題であった．

二つ目は，機械学習の予測結果で公平性を担保する問題について，差分プライバシなどの業績で著名な Cynthia Dwork が講演した．再犯率を予測する COMPASS というソフトウェアでは，アフリカ系の人に対して厳しい予測をするという，データジャーナリズム NPO ProPublica の指摘をあげていた．形式的な公平性の規準として，人種のグループ間で有利な判定を受ける割合が等しくグループとして公平性を担保すること．さらに，類似した人は同様の扱いを受けるという fairness through awareness の概念を紹介し，これらの公平性を担保したうえで予測精度などの効用を最大化する手法を紹介していた．今後の課題としてはトロッコ問題のような道徳的ジレンマについて触れ，人間の平均的な行動に従うといった解決があるのではと指摘していた．最後の Renée J. Miller はデータベース分野の研究者である．複数の関係データベースの情報を，仮想的なスキーマに従ったビューを通じて統合して利用者に見せるデータ統合についてであった．join 演算をするときのキー不一致などの問題と，それらに対する解法の紹介があった．

去年から始まった，Applied Data Science というデータ分析の実務家の招待講演が企画され，11 件の講演があった．米スーパー大手 TARGET では，Ph. D 50 人，エンジニア 150 人，チーム全体では 900 人といった大規模なデータ分析チームを編成しているといった事例が紹介されていた．非 IT 系企業でも，米国ではデータ分析が重要視されるようになっていることを改めて実感した．

3. チュートリアル・ワークショップ

去年と同様に，通常形式のチュートリアル 22 件に加え，本会議と並列して 8 件のハンズオンチュートリアル

があった．筆者は，特徴選択に関するチュートリアルを聴講した．ランダムに割り当てられていない特徴はノイズではないとの仮定に基づき，データの類似性構造を保存している特徴を選択するというアプローチなど，いろいろと新しい試みがあることがわかった．特徴選択について各種のアプローチを俯瞰するとともに，各種の事前知識を反映したlasso系手法の紹介があった．

去年と同様に，今年もワークショップに1日が割り当てられ，20件のワークショップが開催された．筆者は現在取り組んでいる，Dworkの招待講演にもあった公平性に関するワークショップに出席した．第1回のときは部屋に行ったとき誰もいなかったほど注目されていなかったが，第4回となる今年は多数の投稿があり，立ち見もでるほどの盛況であった．2016年のEU General Data Protection Regulationによるアルゴリズム決定への説明の要求や，オープンネスへの世界的な注目などの理由があったのかもしれない．招待講演は二つあり，一つ目はMargaret Mitchellによるものであった．バナナを見たとき，ほとんどのバナナは黄色いので「黄色いバナナ」とはいわず，青いバナナのときだけ「青いバナナ」という傾向がある．このように人間の行動には多くの認知バイアスが存在するが，これをモデル化する方法．二つ目は，マルチタスク学習などで先駆的な研究のあるRich Caruanaによるものであった．医療分野では解釈可能性が重要視されるので，一般化加法的モデルで二次の交差項までを考える拡張をしたGA2Mを用いた分析であった．医学データ分析で見られる，介入要因以外の要因の影響を受ける例をあげ，これらを見つけ出すために解釈可能性は重要であると指摘していた．一般発表では，各グループが好ましい判定を受ける確率が下がらないという新しいタイプの公平性を提案したZafarらの発表が興味深かった．

4. 一般発表・受賞

KDDには，手法・理論・モデルなどの提案や改良を対象とした研究（Resarch）と，手法などを実問題に適用した事例を対象とした応用データ科学（Applied Data Science）の二つのトラックがある．去年のトラック再編に伴い，研究トラックに対して応用系の投稿割合が増えたが，この傾向は今年にも引き継がれ，研究トラックは747件に対し，応用データ科学トラックでは390という比率であった．研究トラックでの採録数は，口頭64（8.6%）ポスター50（8.8%），応用データ科学トラックでは口頭（9.2%）ポスター（12.8%）であった．採択率の推移は，研究トラックは19.5%→ 18.1%→ 17.4%と年々厳しくなっている一方で，応用データ科学トラックでは36%→ 19.9%→ 22.1%とばらつきがある．分野別の傾向では，例年ソーシャルネットワーク関連が多いのだが，今年は時系列が多かったようである．深層学習への注目は機械学習系のICMLやNIPSほどではなく，1セッションあるのみであった．

受賞についてまとめておく．研究トラックのベストペーパは“Accelerating Innovation through Analogy Mining”で，聴講していないので内容はわからない．Runner-upは時系列データのクラスタリングに関するものであった．応用データ科学のベストペーパは，アンドロイド系スマートフォンのマルウェア検出に関するもの，Runner-upは深層学習を用いた気候変動予測についてのものである．

引用数の多かった10年前の論文に与えられるTest-of-Time賞はThorsten Joachimsの線形SVMを線形時間で解く論文，会議運営への貢献により与えられるサービス賞はQiang Yang，いままでの業績に対して与えられるInnovation賞はPrefixSpanのJian Peiであった．データ分析コンペティションの先駆けであるKDD Cupは，中国の交通量予測の問題で，チームBlackswanが二つの部門の優勝を占めたほか，上位は中国勢が独占した．

個人的に関心のあった一般発表をいくつかあげておく．

- The Selective Labels Problem：Evaluating Algorithmic Predictions in the Presence of Unobservables：選ばなかった選択肢の結果は観測できない保釈の判定だと，保釈数が多い裁判官と少ない裁判官がいるので，その差の部分にいるデータを使って，観測できない事例の代わりとする．
- A Data Science Approach to Understanding Residential Water Contamination in Flint：Flint市の水道の鉛汚染問題の分析．多くの要因が関連する複雑な問題であったため，予測器の水質予測は安定せず困難だったようだ．
- Towards an Optimal Subspace for K-Means：低次元空間でのクラスタを見つけるため，特徴を使った空間でクラスタの良さを評価し，選択しなかった特徴空間はノイズとして分布するようにした．

5. お わ り に

発表論文は無償で公開されており，短い紹介ビデオとともにSIGKDDのサイトに掲載されている．招待講演やパネルのビデオはYouTubeで「KDD 2017 video」のチャネルを検索すると閲覧できる．会議関連のTwitterのtweetは`https://togetter.com/li/1137556`にまとめておいたので参考にされたい．2018年は，イギリスの首都ロンドンにて8月19～23日に開催される．北米以外の開催は，パリ，北京，シドニーに続き4回目となる．2019年は，アメリカ，アラスカ州のアンカレッジで8月3～7日に開催とのアナウンスもあった．データ分析技術は，あらゆる分野の基盤技術としてその需要は高まり続けているので，今後も本会議は発展していくであろう．

〔神嶌 敏弘（産業技術総合研究所）〕

The 11th ACM Conference on Recommender Systems（RecSys 2017）

開催地：Villa Erba（コモ，イタリア）
開催日程：2017 年 8 月 27 日（日）～ 31 日（木）
http://recsys.acm.org/recsys17/

1. RecSys 2017

ACM Recommender Systems（略称 RecSys）は，推薦システムを専門とする国際会議である．推薦システムとは，どのものや情報に価値があるかを特定するのを助ける道具である．この会議では，HCI（ヒューマンコンピュータインタラクション），情報検索，および機械学習などの関連研究を，推薦システムへの応用という観点からまとめている．Chris Anderson の“We are leaving the age of information and entering the age of recommendation.”の言葉に示されるように，推薦システム研究は，既存の分野から独立してきたとの考えに基づき，推薦システム研究が盛んなミネアポリスにて 2007 年に第 1 回を開催した．これ以降，2012 年の香港を除いて，北米と欧州交互に開催されており，第 11 回となる今回は欧州で 5 回目開催である．筆者は 8 回目の参加で，2010 年から続けて参加している．開催地のコモ市はミラノの北方にあり，アルプスの水をたたえたコモ湖のほとりにある．電圧のボルトのもとになった Volta ゆかりの地であるらしい．

例年どおり，5 日の開催期間のうち，初日と最後がワークショップに割り当てられ，中 3 日間が本会議であった．オープニングで発表された参加登録者数は 43 か国から 627 人とのことで，新記録であった去年の参加者数 560 から 1 割ほど増えた．国別やインダストリ・アカデミアの比率などの情報の発表は今年はなかった．日本からの参加者は，筆者の印象では去年より多く 20 名前後であろうか．昨年は日本のアカデミアからの参加は 5 人ほどに増えたのだが，今年は 3 名ほどに減少してしまった．

スポンサーは 11 社で，去年からはやや減った．ネット関係の企業が並ぶが，Google や MSR などの大きな研究所が見られなかった．去年から引き続き，日本からはシルバーエッグが加わっていた．

2. 招 待 講 演

招待講演は 4 件あった．最初の講演は行動経済学の George Loewenstein であった．情報の獲得・共有と，これらの行動を望む・望まないの二つの軸で四つの場合に分けて行動経済学的視点から情報について論じた．情報の獲得を望む場合を好奇心（curiosity）と呼び，これを定量化する心理実験を紹介していた．画像の解像度が徐々に上がるビデオと，徐々に下がるビデオでは，徐々に詳細が明らかになる前者のほうが視聴時間が長いというものであった．情報の回避（information avoidance）は情報の獲得を望まない場合である．株価が下がっているときに，株価を見たくなくなる ostrich 効果などの紹介があった．情報の共有を望まない場合をプライバシと呼び，自身が認識している情報公開のリスクと，実際のリスクには大きな乖離があるという問題を紹介していた．最後は，情報の共有を望む desire to reveal という場合で，自分が死んだときの情報を誰や，どこで伝えたいかといった話題について調査していた．

2 件目の George Karypis は推薦システムの研究者であり，大学教育の質の向上についての研究を紹介した．現状の知識である単位を取得できる可能性を予測することで，他の関連クラスを取得するように進めたりといったことを目指しているとのことであった．3 件目のイスラエル企業 Outbrain の Ronny Lempel はニュースサイトでの広告配信について紹介した．ニュースサイトの内容と利用者の個人嗜好を示す中間表現と，目的の広告を表す中間表現の一致を調べる情報検索の技術を用いた方法を採用している．利用者の関心とマーケッタのキャンペーンは一般には一致しないが，キャンペーンに関心があるであろう利用者を見つけるのに利用者プロファイルの一致性を利用する．最後は，Jason Weston は対話インタフェースを用いた個人化について紹介した．基本的には end-to-end のデータを収集して回帰結合型のニューラルネットワークで学習させるもので，実際に物事を知っているわけではないが，対話は実行できる．同じ枠組みで通常の対話に加え，質問応答なども可能にしている．また，ボットからの問いかけにより能動的に学習情報を収集したりする．

3. インダストリアルセッション・チュートリアル・ワークショップ

この RecSys に特徴的なものとして，査読で採録する研究発表とは別に，企業の研究者や技術者を招待しシステムの運用に関する講演をするインダストリアルセッションがある．今年は三つのセッションが一般講演とパラレルに行われた．ファッション系の推薦である Dressipi 社と Farfetch 社を紹介しよう．畳込みネットワークにより画像特徴が容易に抽出可能になったことで，服飾の画像の類似性などの判断ができるようになったため，ファッション系にも推薦技術がここ数年適用されるようになってきたように思う．商品のサイクルが早いうえに，顧客の趣味もよく代わると，予測の難しい対象である．Dressipi 社はスタイリストなどによるドメイン知識を重視していた．また，なぜこれが似合いそう

なのかという理由付けの重要性を述べていた．一方のFarfetch社は，画像特徴などとクラウドソーシング的に集めたラベルを使い，服の組合せの良し悪しを識別するなど，ドメイン知識より技術的解決を目指しているようだった．

チュートリアルは4件で，今年は一般セッションのあとに開催された．ここでは聴講した推薦におけるプライバシについて紹介する．最初は，データの悪用について，敵対者が利用者にとってどのような不利益をもたらすかという分類の紹介であった．その後，プライバシの保護技術として，匿名化，無関係なデータを入力したりするobfuscation，秘密関数計算などの暗号化技術，そして乱数による摂動を用いる差分プライバシの紹介があった．後半はヒューマンインタフェース的観点からのプライバシについてであった．サーバに個人情報を送信しないクライアント処理などをしても，そうした安全性は認識されずプライバシ的な信頼を得るのは難しく，他の人の行動に従うといった実際には有効とはいえない方法で判断してしまうとのことであった．説明が複雑になりがちな透明性の確保や，利用法が難しくなりがちな制御権の付与もあまり有効ではなかった．この対策も個人の知識などに合わせるしかないのではということで，その方策の分類体系などの紹介があった．

初日には7件，最終日には6件，合計13件のワークショップが開催された．筆者はResponsible Recommendationというワークショップで発表の機会を得た．このワークショップは，推薦システムの運用や研究における社会的問題について，問題の指摘や，その技術的解決について論じるものである．公平性には決定過程の公平性・透明性と，結果の公平性などがある．決定仮定を透明にするための，推薦の説明や，不公平と考えられる情報の排除などが論じられた．ジョブマッチングでは求職者とともに雇用者側の視点があり，ソーシャルカーシェアリングでは利用者だけでなく，運ぶ側も公平にマッチングされなければならない．また，性別や人種間で結果の公平性を担保するという考えも必要である．その他，コンペティションRecSys Challengeは，XING社のジョブマッチングを対象としたもので，日本からは3位に佐藤らが入賞していた．

4. 一般発表・受賞

発表はロング，ショート，ポスターがあり，去年試みられたPast, Present and Future（PPF）というポジションペーパの部門は廃止された．この会議では，観測されたデータ上の予測誤差が下がればよいというのではなく，いかに利用者の情報要求を考慮しているかという点に着目している．ロングとショートは論文のページ数の差で口頭発表の機会が与えられ，ポスターは会議録に採録されないもので，意見交換などのための発表である．総投稿数は247件，ロングペーパは125件のうち26件が採録，ショートペーパは122件の投稿で20件（16.4%）が採録であった．採択率は，ロングは昨年から18.2% → 20.8%，ショートは20% → 16.4%とロングは平年並み，ショートは年々厳しくなっているようだ．日本からはショートペーパで2件の口頭発表があった．この会議は日本からの発表は弱く，前回の第10回までは口頭発表は1本だけしか採録されていなかった．

受賞についてまとめておく．学生ベストペーパ“Translation-based Recommendation”：回帰結合ニューラルネットワークを用いた系列予測系での推薦で，語の内部表現ベクトルの系列学習の手法を推薦に適用したもの．ベストペーパ“Modeling the Assimilation-Contrast Effects in Online Product Rating Systems：Debiasing and Recommendations”：利用者の評価はその過去の評価値に影響されてバイアスが生じるが，この効果を明示的に組み込んだものである．

個人的に関心のあった一般発表をいくつかあげておく．

- Educational Question Routing in Online Student Communities：MOOCでの学生間のQAフォーラムを活発化させるための推薦で，edX上でユーザ実験を行った．
- The Magic Barrier Revisited：Accessing Natural Limitations of Recommender Assessment：人間が評価するときに生じてしまう自然なばらつきをMagic Barrierといい，これについて調査．
- User Preferences for Hybrid Explanations：協調・内容ベースのハイブリッド型推薦での推薦理由の説明を生成する．
- An Elementary View on Factorization Machines：factorization machineで高次の交互作用を考慮する．
- Understanding How People Use Natural Language to Ask for Recommendations：利用者が自身の嗜好を音声とテキストで伝えてもらい，その傾向を分析．

5. おわりに

基調講演やチュートリアルなど一部の資料はホームページにて公開されている．会議関連のTwitterのtweetは`https://togetter.com/li/1142825`にまとめておいたのでご参考にされたい．2018年は，カナダのバンクーバーで例年よりやや遅く10月3～7日の開催である．世界的には注目される会議にもかかわらず日本からの発表や参加が，特にアカデミアから少なく寂しい現状だが，来年は発表や参加が増えればと思う．

〔神嶌 敏弘（産業技術総合研究所）〕