会議報告

25th International World Wide Web Conference（WWW 2016）

開催地：The Palais des congress, Montreal, Canada
開催日程：2016 年 4 月 11 日（月）～ 15 日（金）
http://www2016.ca

1. 会 議 概 要

WWW（International World Wide Web Conference）は，Web 分野のトップカンファレンスで，近年は毎年，開催地をおおよそヨーロッパ→北米→それ以外の地区という順で変えて開催されている．25 回目にあたる WWW 2016 は，カナダのモントリオールで開催された．参加者は 55 か国から約 1 000 名であった．Technical Peper の発表は 31 セッションに分かれ，115 件であった．なお投稿は 700 件強あった（採択率約 15%）．また Poster とデモ展示が合計 100 件程度であった．日本からの参加者は 10 数名であった．

WWW 会議のトピックは多岐にわたる．セッション名と発表件数は表 1 に示す．これらが 3 日にわたって発表された．最初の 2 日は Tutorial と Workshop が開催され，21 の Workshop が開催された．

表 1　WWW 2016 でのセッション名と発表数

Social Networks and Graph Analysis	14
Security and Privacy	13
Content Analysis	13
Behaviour Analysis and Personalization	14
Economics and Markets	9
Web Search System and Applications	10
Crowdsourcing Systems and Social Media	10
Web Science	7
Semantics and Big Data	9
Web Mining	10
Mobility	6

このように多岐にわたるので，今回の会議報告では 5 名の報告者がそれぞれの興味で報告したものをまとめている．

2. Keynote と Semantic Web 関連（報告者：武田）

今回の Keynote には久しぶりに Tim Berners-Lee が登場した（WWW 2012 以来）．Tim Berners-Lee は今回は Web security, Web of things, redecentralization of the Web の三つをテーマに話をした．今回は WWW 会議 25 周年，セマンティック Web 15 周年ということで，やや懐古的に過去を振り返りつつ，これらのトピックの話を行った．現在 Google に所属する Peter Norvig は，“The Semantic Web and the Semantics of the Web：Where Does Meaning Come From?”と題して語った．前半はセマンティック Web がなぜ普及しなかったかを，人間は怠け者であるなど，人間の生来の特性（Cory Doctrow の七つの超え難き障害）で説明した．それに対して schema.org はそういった特性を回避するような仕組みを用意することで成功したということを述べた．さらに自動的にセマンティクスを獲得するには深層学習が有効であると述べたものの，詳細には触れなかった．

セマンティック Web 関連の Workshop では Learning & Education with the Web of Data（LIFE 2016），Usage Analysis and the Web of Data（USEWOD），Linked Data on the Web（LDOW 2016）などがあり，盛況であった．LDOW は 2008 年以降毎年 WWW 国際会議で開催され 9 回目になるもので，Linked Data に関する重要な会議となっている．この Workshop は Paper 発表中心で，今回は 24 件中 14 件が採択され発表した．Workshop の最後では Linked Data の今後を三人の Organizer が一つずつあげ議論を行った．Blockchain との関係がトピックに出たのは意外であった．

3. Security & Privacy（報告者：鳥海）

Security & Privacy のセッションは，3 日間にわたって 4 セッション 13 件の発表があった．主にセキュリティに注目した研究が多く，プライバシーの研究はあまり盛んではないようである．

さて，Security & Privacy のセッションで興味深かった二つの研究について紹介しよう．一つは，Understanding the detection of view fraud in Video Content Portals である．この研究では，さまざまなビデオコンテンツポータルにおいて，どの程度ボット検知能力があるのかを調べている．具体的には，(1) クラウドソーシングを使って自分達の登録したビデオを複数人に見てもらう．(2) ボットを使って同じく登録したビデオを見る．以上から，各ビデオコンテンツポータルがどの程度ボットを検出できるか，また，どの程度人間をボットと誤認するかを分析している．対象としているポータルは YouTube, Dailymotion, Myvideo.de, UOL, Vimeo の五つである．結果，YouTube のみがボットを正しく判別でき，それ以外のサイトではボットであろうと人間であろうと検知せずに再生回数を増加させていることが示された．さらに，この論文では広告付き動画についても同様にボットを検知しているかどうかを調べている．動画再生回数に

応じて広告主からコンテンツ提供者にお金が支払われることから，広告付き動画はよりボットを厳しく判別しているだろうと筆者らは予想していた．しかしながら，YouTubeではボットを正しく判別できるにもかかわらず，広告の再生回数に関してはその検知能力が著しく悪化していることが示された．

もう一つは“Tell Me About Yourself：The Malicious CAPTCHA Attack”である．この研究ではiframeをうまく配置することで，ブラウザに表示されるメールアドレスなどの情報を，CAPTCHAに見せかけ，利用者自らに入力させることで情報を抜き取ろうというクラッキング技術の研究であった．CAPTCHAによる承認が一般的になりつつある今，このようなクラッキング技術を実際に使って情報を抜き取ることは容易に可能であろう．クラッキング技術そのものが論文として採択されるというのは筆者が普段顔を出しているセッションでは考えられないことで，研究分野の広さを実感させられた．

4. Wikipedia Workshop & Twitter Analysis（報告者：鈴木）

主にWikipediaやWikidataに関するワークショップであるWikipedia Workshopが行われた．5件の招待講演と9件の研究発表であり，Wikipediaを運営しているWikimedia財団とスタンフォード大学ソーシャルネットワーク分析プロジェクトが中心となっている．

このワークショップで発表されている研究の中心はWikipediaであるが，現在では特に新しいプロジェクトであるWikidataが着目されている．また，もう一つのキーワードとしてリアルタイム解析があり，現在行われているWikipediaの編集がリアルタイムで可視化されるシステムListen to Wikipedia：`http://listen.hatnote.com`, Wikipedia Live Monitor：`http://wikipedia-live-monitor.herokuapp.com`など，WebSocketを用いたさまざまなアプリケーションが紹介された．Wikipediaの創始者の一人であるJimmy Walesがサプライズゲストで登場し，Wikipediaは現在編集者が減少していること，著者への報酬などに関する議論が行われた．

本会議で興味深かった論文であるTweet Properly：Analyzing Deleted Tweets to Understand and Identify Regrettable Onesを紹介する．Twitterにおいて，投稿した後にその投稿を行ったことを後悔することは少なくない．例えば，人の悪口であったり酔った勢いにおける発言などである．ところが，それらはRetweetされてしまい拡散され，消去することができなくなる．そこで，そのようなTweetを投稿する前に事前に自動でチェックする機構が提案されている．そこで，削除されたTweetを解析することによって，どのようなTweetが消去されやすいかを分析している．実験の結果，性的な言葉や暴力的な言葉に関係する単語数は150個程度であるのに対して，人を罵るために使われる単語数は800個近くもあるという点は興味深い．さらに，分類手法にあまり関係なく70％程度で消去すべきTweetを発見できている点は興味深い．

5. Content Analysis（報告者：諏訪）

現実社会での利用可能性を示唆する研究も多く見受けられた．例えば，Content Analysisのセッションで発表されたSantosh, K. C.とArjun Mukherjeeの“On the Temporal Dynamics of Opinion Spamming：Case Studies on Yelp”は，その一つである．タイトルからもわかるとおり，この研究ではYelpのデータ（レストラン）を対象として，スパムの発見を試みている．注目すべきは，まずスパムのタイプとして，初期，中期，後期に分けて分析している点である．さらに彼らは，causal time-series analysisを用いて，スパムをbuffered spammingとreduced spammingに分類している．結果として，これらの特性を考慮に入れることで，効率的なスパム発見ができることを示している．この研究は，収集したデータの基礎的分析を行うことで事象の特性を把握し，その特性に基づく特徴量を用いることで予測精度（今回はスパム発見精度）を向上させており，実運用に向けた基礎的分析の重要性が示唆されていると考える．

6. Social Media Analysis（報告者：豊田）

ソーシャルメディアの分析に関する研究はいまだ人気が高く多くの論文が発表されている．ソーシャルメディアにおける特定の現象を分析しているものから，ユーザのメンタルヘルスに関するものまで幅広い発表があったが，その中からいくつか興味深い論文を紹介したい．

Hong-Han Shuaiらの Mining Online Social Data for Detecting Social Network Mental Disordersは，ソーシャルネットワークサービスに過剰に熱中してしまうある種の精神障害を，そのユーザのソーシャルメディア上での複数種類の振舞いから検知する手法を提案している．3 000人を超えるユーザのログに基づいて精神科医がラベル付けを行うという大規模な実験を行っており，Multi-source semi-supervised learningの枠組みを用いて，複数種類のユーザ行動と心理学の知識とを統合した学習を行うことで高精度での分類を実現している．ソーシャルメディア分析と医療分野との共同研究の試みとして興味深い研究である．

Justin Chengの Do Cascades Recur? は，Lada Adamic, Jon Klienberg, Jure Leskovecら，著名な研究者との共著で，ソーシャルメディア上で時間をおいて繰り返し同じコンテンツが流行する現象を分析している．Facebookでは，五つに二つの画像，三つに一つのビデオが，繰り返し流行することが観測されており，著者らはこの現象はソーシャルネットワーク上で異なるコミュニティに時間差を置いて情報が到達することにより起き

ると仮定して，さまざまな分析を行っている．非常に大きなカスケードを起こした reshare はネットワーク全体にいき渡っているため繰返しは起こらず，中程度のカスケードが繰り返し起きやすいといった興味深い分析が行われている．さらに，繰返しが起こるかどうかを機械学習を用いて予測可能かどうかも調べており，繰返しの発生は AUC 0.89 で予測でき，2 回目のカスケードが 1 回目よりも大きくなるかどうかは 0.78 の AUC で予測できるが，いつ繰返しが起きるかは AUC 0.58 程度でしか予測できないことが示されている．カスケードの繰返しとネットワークコミュニティの関係に着目した点で興味深い研究である．

Travis Martin らの Exploring Limits to Prediction in Complex Social Systems は Duncan Watts との共著で，ソーシャルネットワーク上での情報カスケードのサイズ予測がなぜ難しいのかを分析した．情報カスケードが何人のユーザまで拡散するかを予測する研究はこれまでに多く行われてきたものの，回帰などでそのサイズを予測することは難しいことが知られている．そのため近年の研究では，サイズが一定のしきい値を超えるかどうかを予測するなど，どの程度であれば予測可能であるか模索が行われていた．本論文では，サイズを予測する問題に立ち返り，カスケードのシミュレーションにより予測モデルを立てることが難しいことを示している．

7. Social Network（報告者：森）

ソーシャルネットワークに関連する研究の中から特に，ベストペーパならびにその候補となった二つの研究を紹介する．

まず，ベストペーパに選ばれた「Social Networks Under Stress」は，外部のイベントがネットワーク構造やその内部のコミュニケーションにどう影響するか，という問題意識を着想としている．例えば，テロ，流行病，あるいは競争環境の突然変化のような外部の「刺激」に対して，関連する人々や組織のネットワークがどのように変化するか，という実応用においても重要な状況を想定している．同研究では特にデータセットとして，ヘッジファンド会社の数千万件のインスタントメッセージのやり取りから社内外の人々のソーシャルネットワークを構築し，株価の変動という外因がそのネットワークにどのような影響を及ぼしたかを定量的に考察している．社会学的知見に基づけば，非常時には情報共有やリスク管理を促すために，ネットワークの弱い紐帯や外部とのつながりが活発化することが期待できるが，同研究が示した結果はその逆であった．つまり，株価の変動時には，ネットワークは高いクラスタリング係数を示し，強い紐帯，そして外部よりも社内でのつながりがより強化された．著者らは，このようなネットワークを“turtle up”であると表現している．外部のイベントに対するネットワークのダイナミクスについて，実データに基づいて得られたこれらの知見は非常に興味深い．

次に，ベストペーパの候補にもなった研究「Visualizing Large-scale and High-dimensional Data」は，大規模な高次元データ，数百次元の数百万のデータポイント，を二次元に可視化する手法である LargeVis を提案している．同研究では，従来同様の可視化に用いられている t-SNE 法の問題点として，その前処理となる K 近傍グラフを構築する計算コストが高いこと，データが膨大であるとグラフの可視化処理が非効率であること，そしてパラメータ調整の困難性などを指摘し，それらを解決するための新たな手法を提案している．具体的には，元の高次元データから K 近傍グラフを高い精度で近似するための効率的な手法，ならびにその K 近傍グラフに基づいてデータを二次元に可視化するための確率的なモデルを導入している．提案手法を，テキスト，イメージ，ネットワークなどさまざまなデータセットに適用し，従来手法である t-SNE 法よりも高い精度でより早く可視化が可能であることを示している．特に，提案手法を用いて，学術文献のメタデータである DBLP データを二次元に可視化することでコンピュータ科学分野の主要な国際会議マップを生成しており，その内容はとても興味深い．提案手法のグラフ可視化のモデルに用いている目的関数は非対称の確率的急勾配法を用いてデータ数に対して線形で最適化できることを示しており，数百万のデータポイントであっても，単一の計算機だけで数時間で処理可能であると述べている．

なお，同研究の著者は，昨年の WWW 国際会議においてネットワーク分散表現の新たな手法である LINE を提案しており，同研究はその着想を大規模データの可視化へ拡張したものであり，ネットワーク分散表現の有効な応用例を示している．

〔武田　英明（国立情報学研究所），
鈴木　　優（奈良先端科学技術大学院大学），
諏訪　博彦（奈良先端科学技術大学院大学），
豊田　正史（東京大学），
鳥海　不二夫（東京大学），
森　純一郎（東京大学）〕