

報道関係者各位

国立大学法人 筑波大学
文部科学省研究振興局参事官(情報担当)

文部科学省事業「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」 前半期の主要成果

研究成果のポイント

多様な異種ビッグデータを高度に連携活用するための以下の新技術を開発しました。

1. Twitter、センサーデータ等の多様なストリームデータを処理するための高性能・高機能基盤システム
2. ソーシャルメディア利用者の位置情報、年齢等の属性を世界最高水準の精度で推定する手法
3. オンラインニュースと Twitter を高精度・高性能でリアルタイムマッチングするシステム

研究の背景

情報技術の発展及び情報化の進展に伴い、実社会で生成・取得されるデータは爆発的に増加しており、これらビッグデータの利活用が強く求められています。一方、実社会から収集されるビッグデータは、その種類、量、質において極めて多様であり、リアルタイムデータを含む異種データを連携利用するためのデータ統合・解析技術の研究開発が必須です。

平成 26 年度から 4 年間の計画で始まった文部科学省事業「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」では、筑波大学(代表校)、東北大学、東京大学、慶応義塾大学の 4 大学が連携して、共通的なデータ統合・解析技術の研究開発と、藤沢市における実証実験を行っています(図1)。

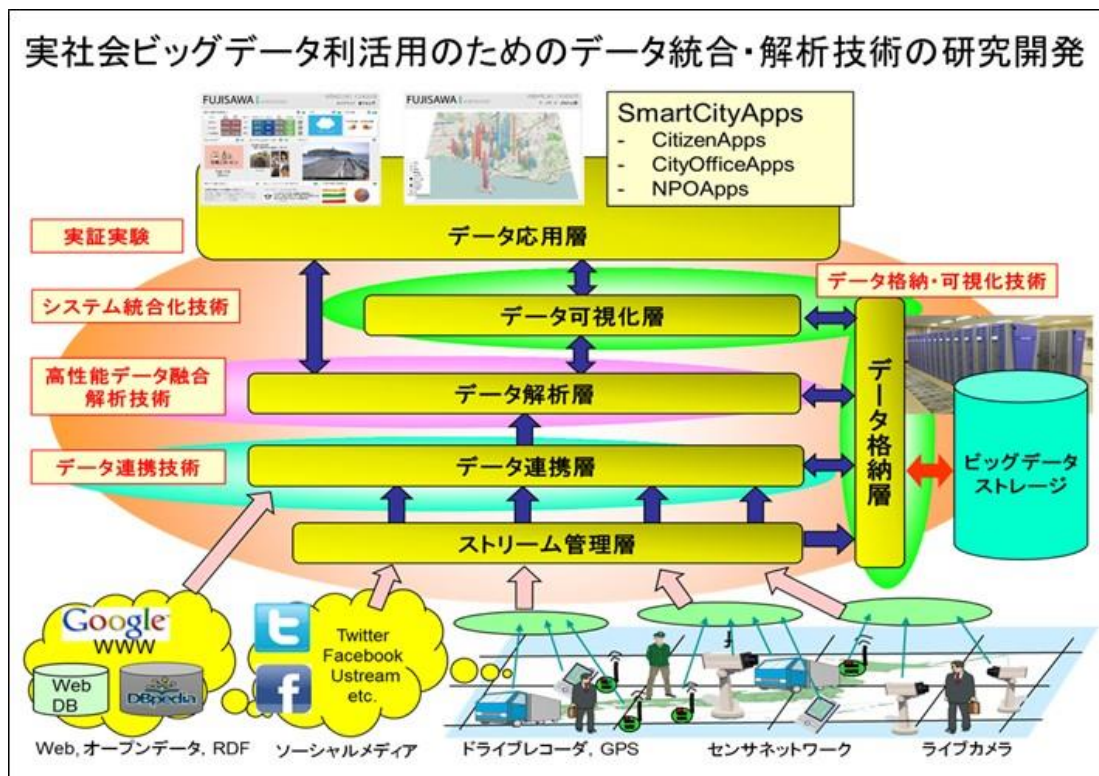


図1 文部科学省事業「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」

研究内容と成果

本事業において、筑波大学は、実社会における多様な異種ビッグデータの高度な連携技術の研究開発を担当しています。これまでの2年間における研究開発成果として、以下のような新技術の開発に成功しました。

1. Twitter、センサーデータ等の多様なストリームデータを処理するための高性能基盤システムの開発 [1, 2]

今日、実世界から連続的にリアルタイム情報を発信する Twitter、センサーデータ等の多様なストリームが存在し、これらの高度な利活用は実社会ビッグデータを扱う上で極めて重要になっています。本事業では、多様なストリームデータを統合的に処理し、様々な情報利用を実現するための基盤システム「JsSpinner」を開発しました。本システムを用いることにより、ストリームを対象とした多様なアプリケーション開発のコストを大幅に低減することが期待できます。JsSpinner の特長は、1)異なる構造を有するデータを柔軟に表現するためのデータ形式としてグローバルに普及が進んでいる JSON 形式のストリームデータを直接扱えること、2)特定のイベントが生じた場合にだけデータを詳細にモニタリングする際(イベント駆動型問合せ)の処理性能を従来よりも大幅に向上(実験環境において従来比最大 300%の処理効率。以下、いずれも実験環境における比較)したこと、の2点です(図2)。

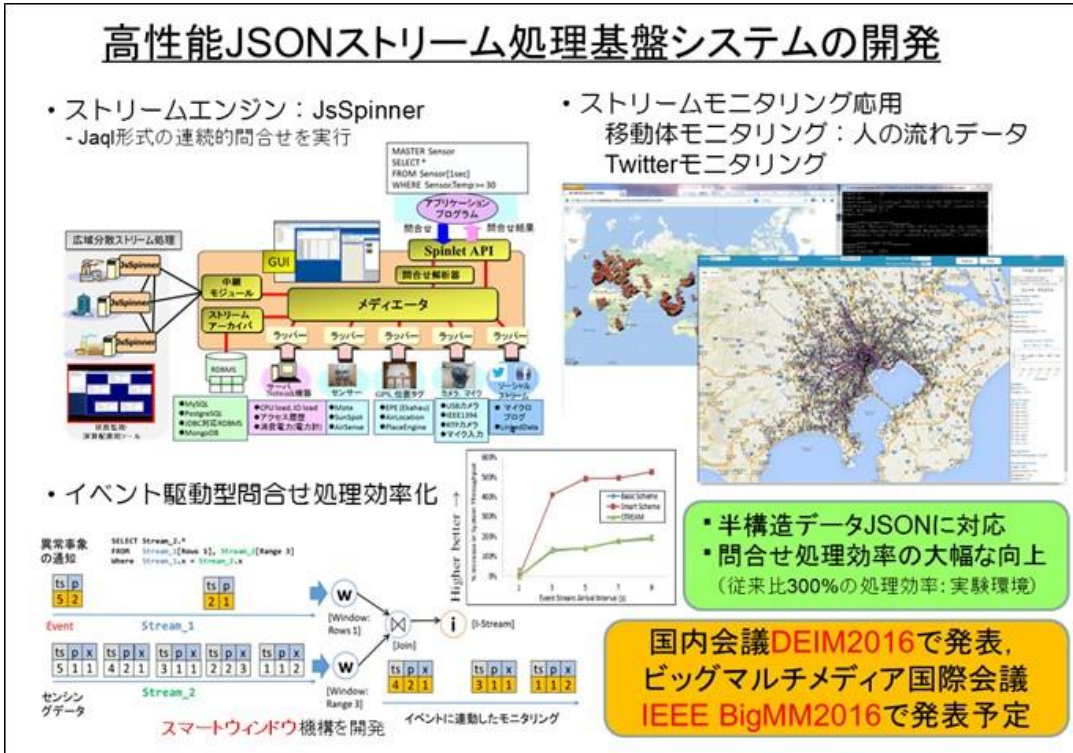


図2 多様なストリームデータを処理するための高性能基盤システムの開発

2. ソーシャルメディア利用者の位置情報、年齢等の属性を世界最高水準の精度で推定する手法を開発 [3, 4]

実社会ビッグデータ利活用において、多くの有益な情報が埋もれているソーシャルメディアの活用の重要性が増加しています。その際、利用者の位置情報、年齢等の属性は記載された内容を意味解釈する上で重要な補足情報となりますが、多くの場合、明示的には開示されていません。本事業では、プライバシー保護に十分留意しつつ、これらの欠損した属性を高精度で推定する新たな2つの手法を研究開発しました。一つ目は、Twitter 利用者の位置を発言内容を手がかりとして推薦する新たな手法で、地名やランドマーク等に加えて、一定の時間帯のみ地域局所性が高くなる単語に着目することで、世界最高水準の推定精度(従来比推定誤差最大 32%削減)を実現しました(図3)。二つ目は、ソーシャルメディアにおける利用者同士の関係と既知の利用者属性を基に未知の利用者属性を推定する新たな手法で、隣接関係にある利用者同士が同じ属性を持つ場合だけでなく、異なる属性をもつような場合でも高精度で属性推定(従来比推定精度最大 42%向上)ができる画期的手法です(図4)。

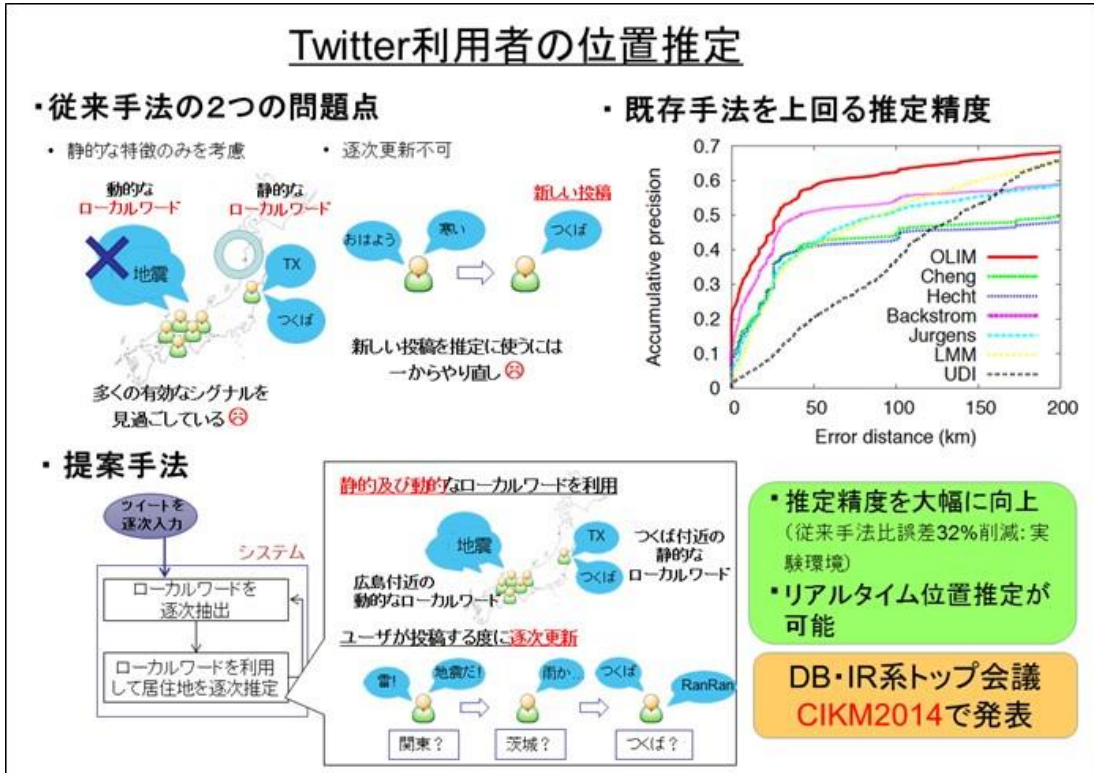


図3 Twitter 利用者の位置推定

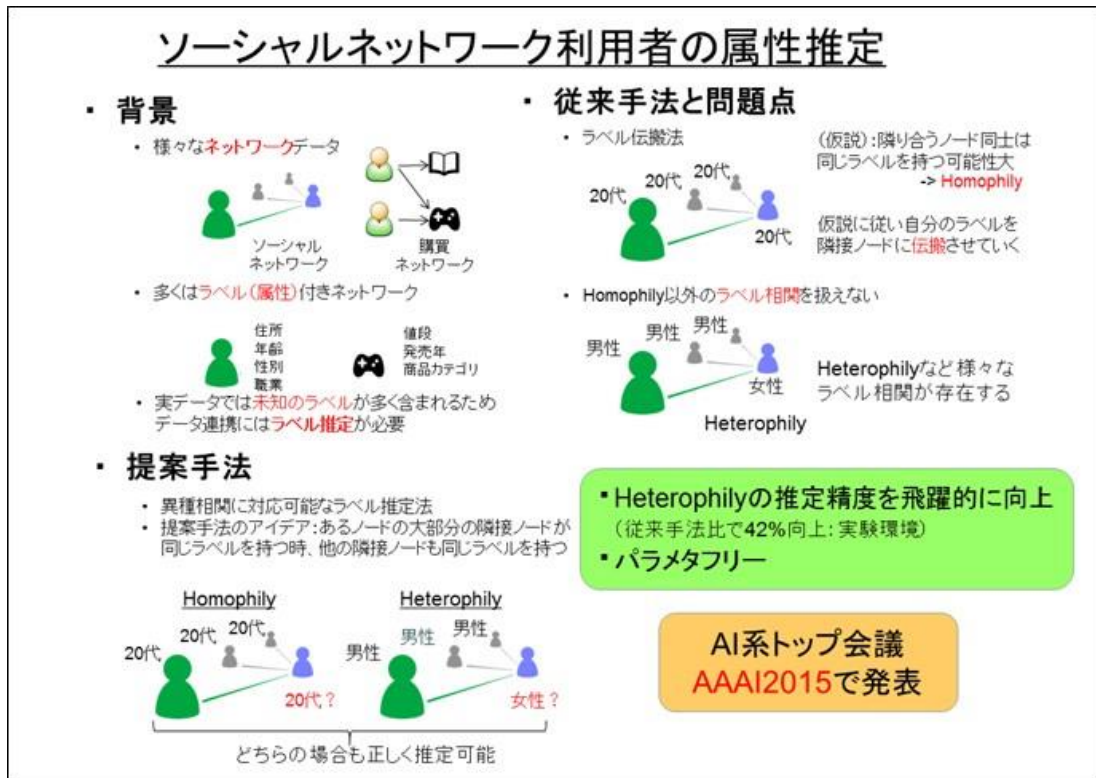


図4 ソーシャルネットワーク利用者の属性推定

3. オンラインニュースとTwitterを高精度・高性能でリアルタイムマッチングするシステムを開発 [5, 6]

複数の実社会ビッグストリームデータ連携の具体的事例として、オンラインニュースとTwitterを高精度・高性能にリアルタイムマッチングするシステムの開発に成功しました。本システムでは、各ニュース記事とそれに関連するツイートをリアルタイムで紐付けすることが可能です。これにより、各ニュース記事に対する社会の関心の高さや様々な

意見を容易に集約することができます。技術的には、ニュース記事とツイート内容の関連度合いをどのように判定するかという課題と、大量のオンラインニュース記事とツイートをいかに効率的に照合するかという課題の2つをクリアする必要があります。本研究で開発した新たな手法により、既存の手法よりも精度(従来比約 6%向上)、性能(ニュースに関して従来比最大 89%処理時間削減、ツイートに関して従来比最大 33%処理時間削減)の両面で大幅に改善できることを示しました(図5、図6)。

ニュースとツイートの高精度・高性能リアルタイム照合

○ 背景と目的

ニュースに言及するツイートが多く存在

これらを高速かつ高精度にマッチング

これによりニュースの分析にソーシャルメディアを活用可能になる

○ 提案手法

ニュースごとに適切な類似度閾値を推定 → 高精度

適切な索引構造により類似度計算を枝刈り → 高速

○ チャレンジとアイデア

- ニュースごとに人気度が違う → ニュースごとに適切な数のツイートをマッチング
- 膨大な数のニュースとツイート → 索引構造による効率化

○ マッチング精度を向上
(従来手法比約6%向上: 以下いずれも実験環境)

- 新規ニュース到着時の索引更新時間を大幅削減 (従来手法比最大89%削減)
- 新規ツイート到着時の処理時間を削減 (従来手法比最大33%削減)

協調システム系主要会議CooplS2015, 国内会議DEIM2016で発表

図5 オンラインニュースと Twitter を高精度・高性能でリアルタイムマッチング

The interface displays a list of news items with categories like '政治' (Politics) and '台風' (Typhoon). The detailed view shows a tweet about Typhoon 18 (台風18号) with a matching score and a 'フォロー' (Follow) button.

図6 オンラインニュースと Twitter のリアルタイムマッチングシステム

今後の展開

去る3月11日(金)、慶応義塾大学矢上キャンパスで開催された情報処理学会第78回全国大会において、本事業に関する特別セッションを企画し、今年度までの研究成果を報告しました。会場には立ち見が出る程の多くの参加者が集まり、ビッグデータ活用技術に対する社会の関心の高さを示していました。今後、これまでに研究開発したシステムや技術を、実証研究に適用し、評価・改良を進めると共に、実社会ビッグデータの連携活用を進める上で必要とされている様々な高度な技術の研究開発に取り組んでいく予定です。

主要な論文発表

1. Twitter, センサーデータ等の多様なストリームデータを処理するための高性能基盤システムの開発

[1] Salman Ahmed Shaikh, Yousuke Watanabe, Yan Wang and Hiroyuki Kitagawa, "Multi-Query Optimization for the Smart Query Execution Scheme", 第8回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), 2016年2月29日.

[2] Salman Ahmed Shaikh, Yousuke Watanabe, Yan Wang, Hiroyuki Kitagawa, "Smart Query Execution for Event-driven Stream Processing", Proc. 2nd IEEE International Conference on Multimedia Big Data (IEEE BigMM 2016), Taipei, Taiwan, April 20-22, 2016 (発表予定).

2. ソーシャルメディア利用者の位置情報, 年齢等の属性を世界最高水準の精度で推定する新たな手法を開発

[3] Yuto Yamaguchi, Toshiyuki Amagasa, Hiroyuki Kitagawa, and Yohei Ikawa, "Online User Location Inference Exploiting Spatiotemporal Correlations in Social Streams," Proc. 23th ACM International Conference on Information and Knowledge Management (CIKM 2014), pp. 1139-1148, Shanghai, China, November 2014.

[4] Yuto Yamaguchi, Christos Faloutsos, and Hiroyuki Kitagawa, "OMNI-Prop: Seamless Node Classification on Arbitrary Label Correlation", Proc. 29th AAAI Conference on Artificial Intelligence (AAAI 2015), Austin, USA, January 2015.

3. オンラインニュースとTwitterを高精度・高性能でリアルタイムマッチングするシステムを開発

[5] Sei Onishi, Yuto Yamaguchi, and Hiroyuki Kitagawa, "Real-Time Relevance Matching of News and Tweets", Proc. 23th International Conference on Cooperative Information Systems (CoopIS 2015), pp. 109-126, Rhodes, Greece, October 28-30, 2015.

[6] 大西 誠, 山口 祐人, 北川 博之, "オンラインニュースとツイートのリアルタイムマッチング手法", 第8回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), 2016年3月1日.

問い合わせ先

<研究に関すること>

北川 博之(きたがわ ひろゆき)
筑波大学 システム情報系 教授

<文部科学省「未来社会実現のためのICT基盤技術の研究開発」に関すること>

文部科学省研究振興局参事官(情報担当)付