

ビッグデータに潜在している 有益な情報を探し出す

ユニット名

データ工学・知識発見

ユニット代表者 システム情報系 教授 北川 博之

◆ユニット構成員 総数 10名 (教員7名/ポストク0名/他機関3名)



キーワード データ工学、知識発見、データベース、データマイニング、ビッグデータ

<http://kde.cs.tsukuba.ac.jp/>

モバイル・ユビキタスコンピューティングをはじめとする社会情報基盤の高度化や、多様な情報ビジネス・情報応用の展開に伴い、人類がこれまで体験したことのない超大規模なデータを扱うビッグデータ時代が到来しています。地球規模でビッグデータがリアルタイムで飛び交う中で、必要なデータをいかに入手し有効な情報として活用するかが重要となっています。北川博之教授が代表のリサーチユニット「データ工学・知識発見」では、ビッグデータを適切に管理し、その中に潜在する有益な情報を活用するための技術に主眼を置いて、研究活動に取り組んでいます(図1)。



図1:「データ工学・知識発見」リサーチユニット

範な技術が必要となります。本リサーチユニットでは、データ工学的なアプローチにより、ビッグデータの管理や利活用に関わる様々な技術の研究開発を推進しています。研究テーマの例としては、近年急増しているセンサー等からのストリームデータの利活用基盤の構築、誤差やノイズを含むようなデータからの異常データの発見、GPUを用いてデータマイニング処理を何十倍も高速化するための手法、Twitter やソーシャルブックマークからの有用情報のマイニング等があります。より具体的には、Twitter を用いて実世界のイベント情報をより正確に把握することを目的に、Tweetの発信者の位置情報を推定する新たな手法を研究開発していますが、これまで知られているどの手法よりも精度面で優れているという実験結果を得ています(図2)。

■他分野、学外組織との連携で技術を現場に生かす

本リサーチユニットのコアとなるメンバーは、計算科学研究センターにも所属し、素粒子物理学、気象学、天文学等の他分野の研究者と連携して研究を進めています。その中で得られた研究開発成果の一端としては、XML 技術を活用した素粒子シミュレーションデータ検索システム、気象データデータベース(図3)、X線天文データの類似検索技術、等があります。また、学外の研究機関や民間企業との共同研究や受託研究も積極的に進めています。

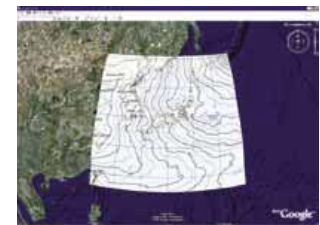


図3: GPU/JMA 気象データベース研究開発

■ビッグデータの利用が可能になった現代だからこそ

膨大なデータを管理するための技術としてデータベースがあります。データベースが最初に登場した時代は、コンピュータに蓄積可能なデータ容量も小さく、格納コストもはるかに高かったため、業務処理に必要なデータだけを厳選し形を整えて処理していました。現在は、蓄積できるデータ容量はほぼ無制限で格納コストも安く、また誰でもがインターネットで新たな情報を発信できます。その結果、多様性・流動性が極めて高くノイズを多く含むデータが大量にインターネット上にビッグデータとして存在し、真に有用な情報を見つけるのが極めて難しくなっています。したがって、いかにビッグデータから潜在的な価値をもつ情報を探し出すかという課題が非常に重要になっています。

ビッグデータの利活用のためには、データの生成からアプリケーション・利用者による利活用に渡るデータのライフサイクル全体をカバーする広

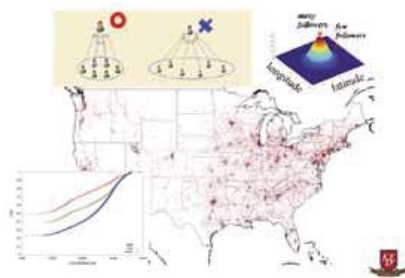


図2: ランドマークに基づく Twitter ユーザの位置推定

社会への貢献・実績

- スマートコミュニティの実現に必要な多様構造情報統合・分析技術に関する企業との共同研究
- 時系列データの分析基盤技術に関する企業との共同研究
- ストリームデータからの類似データ検出や異常検知に関する企業との共同研究
- データプライバシー保護技術に関する企業への技術指導
- 実践的 ICT 人材育成に関わる産学連携
- 大規模データ分析分野の人材育成に関わる産学連携
- ソフトウェア保守支援技術に関する企業との連携
- データ工学、機械学習分野における国内外の学会活動、学術振興

取材:平成25年9月5日