3.14.2　ソーシャル ICT 推進研究センター　ソーシャルビッグデータ ICT 連携センター

連携センター長（事務取扱）　　木俵　豊　ほか 5 名

ソーシャル・ビッグデータのリアルタイム蓄積・解析基盤の開発

【概　要】

ソーシャルビッグデータ ICT 連携センターでは、ソーシャル・ビッグデータのリアルタイム蓄積・解析基盤の開発を目指し、(1) 超高速・頑健自然言語処理技術、(2) 高度データマイニング技術、及び (3) 大規模情報統合可視化技術の研究開発を推進している。(1) については、大規模・大流量となるソーシャルメディアストリームの解析の際に問題となる、崩れた表現の正規化、新エンティティの検出、ユーザ位置推定に関する研究開発を実施し、ソーシャルメディアを多様な観点から解析する基盤手法を構築した。(2) については、ソーシャルグラフ等の大規模なグラフデータを効率的に処理可能な分散グラフデータベースエンジン、並びにイベント時系列から長期間高頻度に発生するパタンを抽出する手法を開発した。(3) については、様々なソーシャル・ビッグデータ解析から得られる解析結果を 3 次元空間に統合的に可視化するフレームワークの開発を行った。

【平成 26 年度の成果】

(1)　超高速・頑健自然言語処理技術の研究開発

ソーシャルメディアには、実世界で起こった災害、事故、イベント等の情報がリアルタイムに流れるようになっており、災害対策、事故・イベント等による状況把握、トレンド解析等、様々に利活用されている。Twitter を代表とするリアルタイムソーシャルメディアには 1 日に何億件もの投稿があり、その大半はスマートフォン等のモバイル端末からリアルタイムに投稿されており、書かれた内容を高速かつ頑健に処理可能とする自然言語処理技術が求められている。今年度は、ソーシャルメディアでなされる崩れた日本語表記に適応する形態素解析手法、及び新しく現れたエンティティの検出、並びにユーザ位置の推定に関する研究開発を行った。

ソーシャルメディア上に投稿される文章には文法に従わない崩れた表現が多く、そのまま形態素解析を行うと精度が大きく悪化することが知られている。そこで、東京大学生産技術研究所において整備された形態素・正規化情報付きのコーパス（辞書）を用い、形態素解析と表記の正規化処理を同時に実行可能とする新たなモデルを提案し、崩れた文を解析可能とした（図 1）。

ソーシャルメディア解析の際には、新製品の名称等、新たに出現するエンティティを正しく検出することが重要となる。また、投稿の大半には位置の情報が付加されておらず、イベントの発生位置を把握することは難しい。ソーシャルメディアに新たに出現するエンティティの検出に関しては、文字列としての表層的手がかりと、エンティティ出現位置の周辺に現れる文脈の履歴を共に素性として用いた分類器を提案した。これにより適合率 67%、再現率 78% で新エンティティをリアルタイムに検出することが可能となった（図 2 上）。ユーザ位置の推定に関しては、対象とするユーザの過去の投稿から得られる移動予定などの手がかりを用いることで、都道府県レベルでの位置推定精度を 7% 向上することに成功した（図 2 下）。

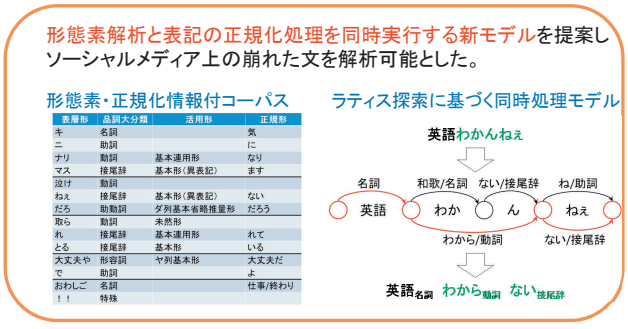

図 1　崩れた表記に適応する形態素解析手法

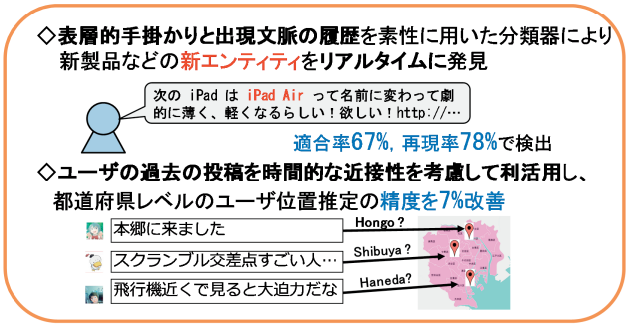

図 2　新エンティティ検出及びユーザ位置推定

(2) 高度データマイニング技術の研究開発

ソーシャルメディアにおけるユーザ間のつながりを表すソーシャルグラフや、検索・購買ログのようなトランザクションデータ等の非テキストデータに関する高度データマイニング技術に関しても研究開発を行った。グラフデータマイニングに関しては、クラウド環境に適したスケーラブルな分散グラフデータベースエンジンの開発を行った(図 3)。ソーシャルグラフ等、多くのグラフデータは次数分布に偏りがあり、通常の分散グラフデータベースエンジンでは効率的な処理が難しい。提案する GraphSlice 手法は、次数が大きいノードを効率的に分散処理可能なデータ構造を導入し、著名なオープンソース分散グラフデータベースの Apache Giraph と比較して 32 倍のスケーラビリティを実現した。トランザクションデータマイニングに関しては、スケーラブルな長期間高頻度パタンの抽出手法を開発した(図 4)。検索ログのような多種イベントが発生する時系列データにおいて、短い間隔で長期間継続するトレンドを表すパタンを高速に抽出する方法を提案し、データ量に対して処理時間がほぼ線形となるスケーラビリティを実現した。

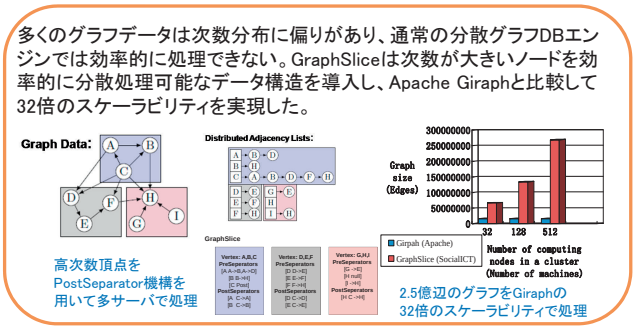

図 3　GraphSlice：スケーラブルな分散グラフ DB

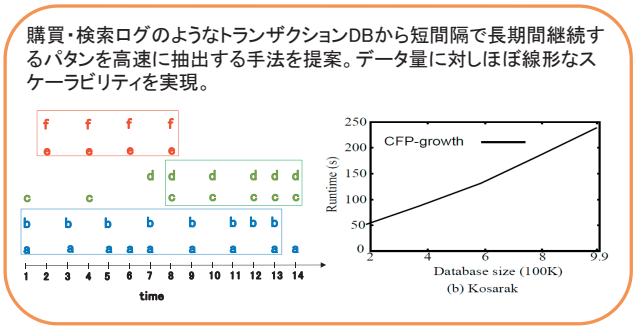

図 4　スケーラブルな長期間高頻度パタン抽出手法

(3) 大規模情報統合可視化技術の研究開発

ソーシャル・ビッグデータとして得られるテキスト・非テキストデータの解析結果を統合的に可視化するフレームワークに関する研究開発を行った(図 5)。テキストデータ、画像データ、時系列データ等、多様なデータの解析結果を 3 次元空間を活用して統合的に可視化するもので、様々な 3 次元可視化部品を整備し、組み合わせることで時系列的な話題(トピック)可視化を実現するフレームワークを提案した。本フレームワークを用いて、複数のメディアから得られる画像とテキストから構成されるトピックが解析可能であることを示した。

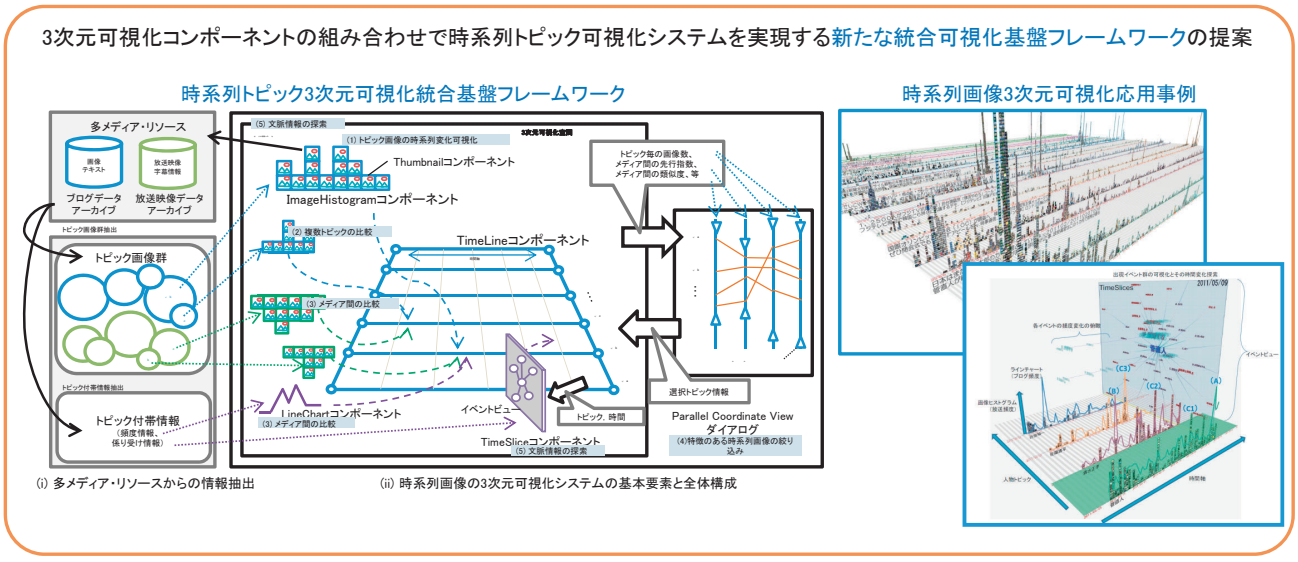

図 5　時系列トピック 3 次元可視化統合基盤フレームワーク