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 4 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、ソーシャルメディア上で発信される言語及び話題の多様性に着目しこれを頑健に扱うため、語の意味を言語横断的に扱うための基礎研究、並びに情報カスケードから社会的影響力を持つもののみを検出する手法の研究開発を行った。(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を汎用フレームワークとして実装し、前年度比で約 10 倍となる 21 億辺のグラフデータでスケーラビリティを実証した。(3) については、大都市で日々発生するイベントの影響や時空間的な広がりを理解可能とする 3 次元可視化手法の研究開発を行った。

【平成 27 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な利活用がなされている。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なされており、書かれた内容を高速かつ頑健に処理可能な自然言語処理技術が求められている。平成 27 年度は、ソーシャルメディア上で発信される言語及び話題の多様性に着目しこれを頑健に扱うため、語の意味を言語横断的に扱うための基礎研究、並びに情報カスケードから社会的影響力を持つもののみを検出する手法の研究開発を行った。

多言語情報の情報検索や情報推薦等のアプリケーションにおいて語の意味表現及びその翻訳は重要な基礎技術となる。語の意味は共起する語のベクトル (単語ベクトル) により表現可能であるが、言語が異なれば共起する語に共通性がないため比較不可能である。多言語アプリケーションにおいてはある言語の単語ベクトルを他の言語の単語ベクトルに翻訳することが不可欠であるが、この単語ベクトルの変換は翻訳行列と単語ベクトルの積で表現することができる。翻訳行列を学習する際に文脈後ペアの対訳関係等を活用可能とすることで翻訳精度を向上する手法を提案し、日、英、中、スペイン語間での翻訳精度を最大 23.5% 改善することに成功した (図 1)。

ソーシャルメディアのひとつであるマイクロブログには多様な話題が投稿されており、情報カスケードと呼ばれるユーザ間での情報拡散現象を経て企業不祥事の発覚等、現実社会に大きな影響を与えるものも少なくない (図 2)。しかし、大きく拡散した情報カスケードでも社会的影響力を持つものは 20% 程度と少なく、多様な話題に埋もれて発見が難しいことが問題となっている。本研究では、投稿内容、拡散経路、反応ユーザを特徴量とする分類手法を開発し、Twitter 上でのリツイートからなる情報カスケードから、拡散初期の 50 リツイート時点において、F 値 0.66 の比較的高い精度で検出することに成功した。

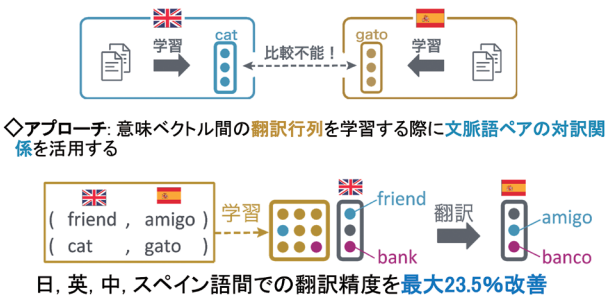

図 1　単語の意味ベクトルの翻訳手法

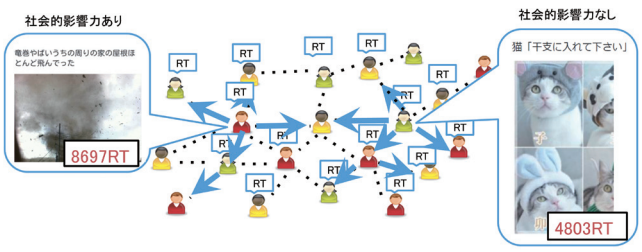

図 2　社会的影響力を持つ情報カスケード検出

（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。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。平成 27 年度は、前年度開発した GraphSlice 手法を、現在のデファクトスタンダードである Apache Giraph 及び Hadoop 上に汎用フレームワークとして実装し、グラフデータの偏りを検知してパラメタを自動調整する機構を実現した。前年度比で約 10 倍となる 21 億辺のグラフデータでスケーラビリティを実証している（図 3）。

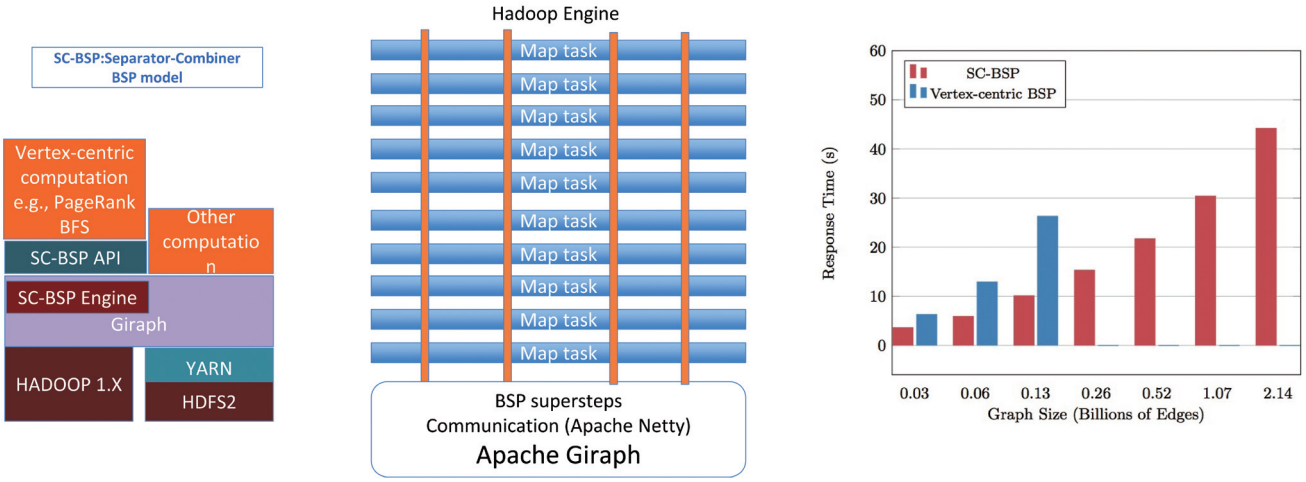

図 3　GraphSlice：スケーラブルな分散グラフ DB

（3） 大規模情報統合可視化技術の研究開発

本年度はソーシャルビッグデータとして得られるテキスト・非テキストデータより、大都市で日々発生するイベントの影響や時空間的な広がりを理解可能とする 3 次元可視化手法の研究開発を行った。マイクロブログストリーム中の位置参照表現に着目し、位置座標情報が付加されていないつぶやきについても、地名や施設名などの位置参照表現に基づいて位置に関連付け、局所的なイベント、広範囲イベントを認識して 3 次元空間に多層的に可視化し、その時間変化をアニメーション可能とする可視化システムのプロトタイプを実現した（図 4）。

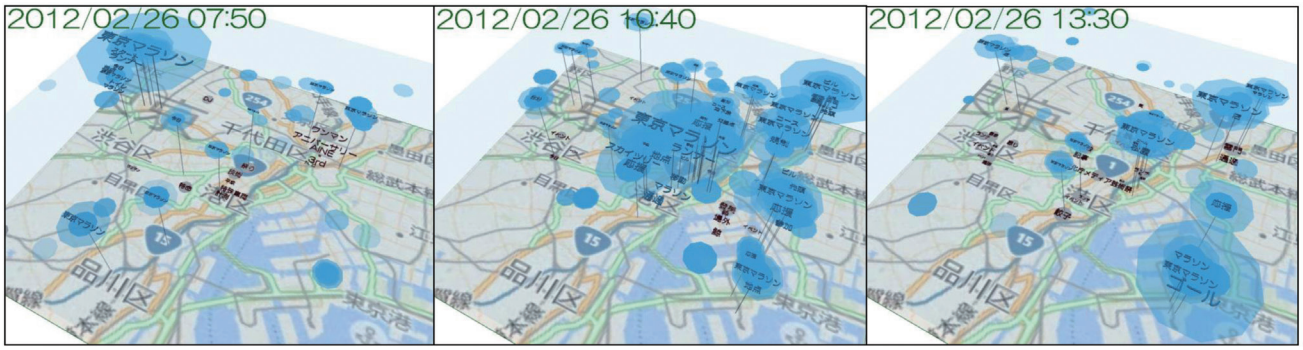

図 4　時系列トピック 3 次元可視化統合基盤フレームワーク