様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　１６　日現在

機関番号：　３２５０４
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３１４

研究課題名（和文）２億件超の東日本大震災ツイッターデータからの発言者の役割を反映した時系列話題解析

研究課題名（英文）Time Series Topic Extraction from Millions of Tweets after the East Japan Great Earthquake Considering Author's Role

研究代表者

橋本　隆子（HASHIMOTO, TAKAKO）

千葉商科大学・商経学部・教授

研究者番号：８０５５１６９７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では東日本大震災後に投稿された2 億件に及ぶTwitter データを対象とし，発言者の役割を反映した時系列話題解析とその評価を行った．発言者と単語の関係を2 部グラフで表し，発言者グループを抽出し，それを反映したクラスタリングを行うことで発言者の役割を推定し，解析精度を向上する手法を開発した．アルゴリズムはランダムウォークにより乱択化し，大規模データへの対応も行った．結果として，30分（あるいは1時間）ごとの投稿者と単語の二部グラフを生成し，そのデータに提案手法を適用することで，LDA等の従来手法より精度の高い話題を抽出することが可能であると確認できた．

研究成果の概要（英文）：This research has developed a time series topic extraction method from millions of Tweets that were posted after the East Japan Great Earthquake. Our method is using our original community detection technique in bipartite networks. Our method considers the relationship between the authors and the words as bipartite networks and explores the authors role by forming clusters from them as topics. We utilize the random walk algorithm to effectively apply to big data. As a result, by generating time series bipartite networks with authors and words (e.g. every 30 min) and employing our method, we found that our method could extract more precise topics compared with the conventional methods such as LDA.

研究分野：　ソーシャルメディア解析

キーワード：　データマイニング　東日本大震災　二部グラフ　話題抽出　クラスタリング　知識発見　著者の役割

１．研究開始当初の背景

東日本大震災発生後，SNS では，それ以前は観察されなかった「避難生活への不安」や「原発からの避難」といった話題が起こり，時間が経つに従って「被害の心配や原発への不満」等へ派生・変化していく様子が観察された．こうした「話題」がいつ頃，どのように発生し，変化していったかを解析することは，大震災の記録としても，災害発生後の時系列話題推移のモデルとしても重要である．大震災関連の SNS データの一つに「2 億件に及ぶ Twitter データ*1」がある．震災直前の 2011 年 3 月 9 日から 3 週間，約 100 万人のユーザーから投稿された Tweets による貴重なデータである．しかし，データ量が膨大（圧縮状態で 32GB）なため，これまで特定単語の数え上げや特定の言い回しパターン検索等が行われているのみであり，時系列話題解析は行われてこなかった．一般に SNS 上の話題構造は曖昧であり，1 つの単語（例：「避難」）が複数の話題（例：「避難生活への不安」，「原発からの避難」等）に重複して属する場合が多い．単語の意味（位置付け）は話題ごとに異なり，時間経過によりさらに変化していく．話題解析精度向上のためには，こうした「単語の位置付け」を考慮した解析が重要となると考えるが，既存手法の多くがこれを考慮していない．たとえ考慮していたとしても，その対象は数千～数万オーダーのデータであり，億を超える大規模データには対応できていない．

２．研究の目的

本研究では東日本大震災後に投稿された 2 億件に及ぶ Twitter データを対象とし，発言者の役割を反映した時系列話題解析とその評価を行う．SNS 上の話題構造は曖昧であり，同じ単語（例：「避難」）が複数の話題（「避難生活へ不安」，「原発からの避難」など）に異なった意味（位置付け）で属する場合が多く見られる．我々は，この「単語の位置付け」は「誰がどのグループで発言したか」という「発言者の役割」に依存すると考える．本研究では，発言者と単語の関係を 2 部グラフで表し，発言者グループを抽出することで発言者の役割を単語の位置付けに反映する．話題の解析精度を向上可能な手法を提案する．アルゴリズムはオンライン化・乱択化し，大規模データへの対応も行う．2 億を超える震災関連 Tweets の時系列話題解析は本研究が初めてであり，大震災の記録としての価値も高い

３．研究の方法

本研究では，以下のような方法で研究を実施した．

１）数億件規模のデータを対象とした発言者の役割を反映した時系列話題解析手法の開発

既存手法の多くは，発言者の役割を考慮しておらず，曖昧で複雑な話題構造を精度よく解析することができなかった．スケーラビリティも課題となっていた．本研究では，発言者と単語を 2 部グラフで表現し，時系列に発言者グループを抽出することで，発言者の役割を推定し，それを反映した高精度の時系列話題解析手法を開発する（「図 1 提案手法」を参照）．アルゴリズムをオンライン化・乱択化し，スケーラビリティも解決する．発言者の役割を考慮しつつ，大規模データに対応可能な既存手法は存在しておらず，高速・高精度な時系列話題解析手法を確立する．

２）東日本大震災後に投稿された 2 億件超の Twitter データに対する時系列話題解析

研究者向けに提供されている東日本大震災関連の「2 億件に及ぶ Twitter データ」に対して，１）で述べた本研究で考案する時系列話題解析手法を適用し，震災後に Twitter 上でどのような話題が発生し，変化していったかを示す．解析結果は，大震災の重要な記録として広く公開する．

３）災害発生後にソーシャルメディア上で発生する時系列話題推移の評価

上記２）の結果を元に，災害発生後にどのような話題が生まれたかを確認し，それらの話題のライフサイクル（生成・分割・統合・消失）を感が蹴る．LDA 等の既存の手法と比較することで，災害発生後の話題推移を表現するモデルとしての有効性を評価する．

４．研究成果

各年度の研究成果について以下に示す．

(1) 平成 27 年度

平成 27 年度は，大規模解析のための開発環境整備及び，発言者と単語の 2 部グラフ生成とそれに基づく発言者の役割推定アルゴリズムの開発・実験を行った．大規模解析のための開発環境整備としては，2 億件の Twitter データのデータクリーニングを行い，発言者と単語の関係を 1 時間ごとの時系列 2 部グラフ

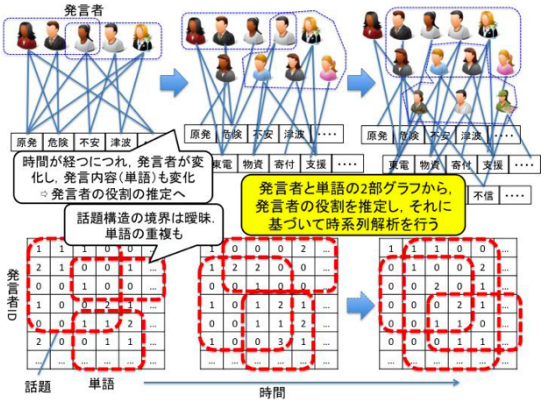

図 1 提案手法

で表現し，データベースに蓄積する処理を行った．これにより，以降のアルゴリズム開発・実験が非常に容易になった．さらに発言者と単語の 2 部グラフ生成とそれに基づく発言者の役割推定アルゴリズムの開発・実験においては，GENETIC ALGORITHMS の一つである BC ヒューリスティックや Median ヒューリスティックを用いて，発言者と単語の時系列 2 部グラフから話題抽出を行う手法の開発を行った．本アルゴリズムは既存手法（LDA）が話題抽出に数時間かかるところを数十秒で実現することを目指したものである．時系列話題解析としては，オリジナルの特徴抽出手法 CWC を開発し，それを適用することで，高速かつ高精度で時系列に話題を抽出できる手法の開発に取り組んだ．図 2 はオリジナルの特徴抽出手法を利用し，特徴量が大きく変化するところで話題の変化が発生したことを検知する提案手法を説明したものである．図 2 で一番上に表示されている折れ線グラフが特徴量の時系列変化を示している．その特徴量が大きく変化したタイミングで何か大きな出来事が

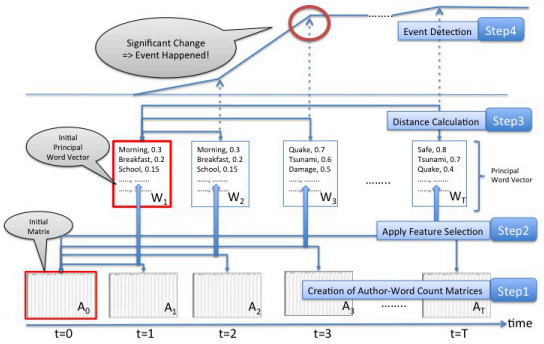

図 2 特徴抽出手法を活用した時系列話題解析

あったことが分かる．この時点ではまだ発言者の役割は十分に考慮されていないが，話題の時系列変化を特徴量の変化で表現できることを示すことができたと考える．

(2) 平成 28 年度

平成 28 年度は，平成 27 年度に開発した発言者の役割推定アルゴリズムの改良及び，時系列話題解析の詳細評価を行った．発言者の役割推定アルゴリズムにおいては，本研究で当初計画した通り，Twitter における発言者と発言ワードの関係を 2 部グラフで表現し(図 3)，2 部グラフの形状を保ったままクラスタリングを行える手法を提案し，それを用いた話題抽出手法の開発を行った．2 部グラフの特徴を活かしながらクラスタリングする手法は，世界的に見てもまだ確立されておらず，オリジナリティの高い効果的な手法である．実際にその手法を 2 億件強の Twitter データ（実データ）に適用することで，提案手法の効果を示すことも行った．図 3 は発言者と使用単語の関係を表現したものであり，発言者と単語間のエッジがその発言者がその単語を使用したことを示している．我々の提案クラスタリングは，発言者と単語の 2 部グラフ構造を

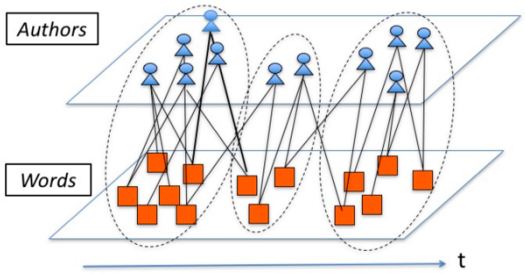

図 3 Twitter の著者，単語の 2 部グラフ表現

考慮しつつ，1 つのクラスター内に発言者と単語が含まれている手法である．従来手法は，発言者のみ，単語のみのクラスター生成であり，従来手法と比べ，より発言者の役割を小反映したクラスタリングが実現出来ている．

こうした話題抽出の評価は常に困難が伴うが，本研究では，Coherence というクラスタリングの意味的な品質を評価するパラメタを用いて，その評価を行った．Coherence パラメタを用いて各クラスターの質を数値で表現することで，定量的な評価も実施できている．さらに定性的な評価も実施しており，手法の有効性や改善すべき点などの検討が可能となっている．さらに東日本大震災のみならず熊本地震の Twitter データまで解析を広げ，手法の汎用性や災害後の Twitter 上の話題のモデル化にも開始した．

(3) 平成 29 年度

平成 29 年度は，平成 28 年度に得られたクラスタリング結果（発言者と単語の関係を 2 部グラフで表し，発言者グループを抽出し，それを反映したクラスタリングを行うことで発言者の役割を推定し，解析精度を向上する手法を用いたクラスタリング）の実験・評価を進めた．結果として，30 分（あるいは 1 時間）ごとの投稿者と単語の二部グラフを生成し，そのデータに対して提案手法によりクラスタリングすることで，LDA 等の従来手法より精度の高い話題を抽出することが可能であると確認できた．特に「石油コンビナート爆発」といった決め打ちの単語を用いれば，デマのような特徴的な話題と，それを訂正しようとする話題，さらにそれらの時系列推移を抽出できることを示すことができた．しかし一方で，全く情報が与えられていない段階で新たに発生する特徴的な話題を早期に発見し，その成長パターンを予測することが困難であることも確認できた（図 4 及び図 5）．

たとえば図 4 は，東日本大震災後に発生した「千葉の石油コンビナート爆発」のデマ話題の推移である．横軸が時間，縦軸が話題の意味的な品質を示す Coherence 値となる．また網掛けの丸がデマ話題，白丸がデマを訂正する話題を示している．図 4 から，デマが発生し拡散した後，デマ話題が品質を低めに保ちながら話題として発生していく様子を確認することができる．そしてデマを訂正する話

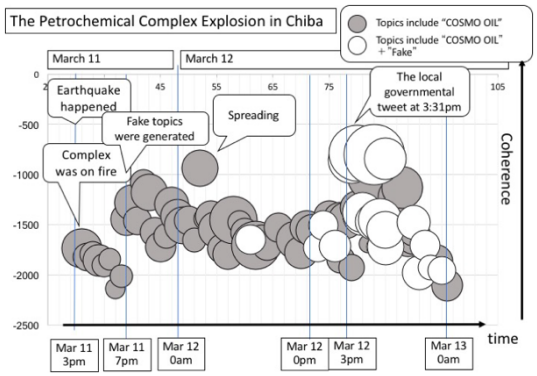

図 4 提案手法による「石油コンビナート爆発」の話題推移

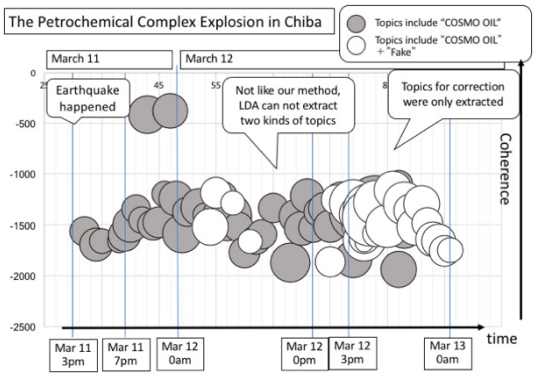

図 5 既存手法（LDA）による「石油コンビナート爆発」の話題推移

題（白丸）が，意味的に品質の高い話題として発生している様子も分かる．一方，図 5 はメジャーな話題抽出手法である LDA を利用して，同様の実験を行った結果である．デマとデマを訂正する話題の意味的な品質に区別が出来ず，ほぼ同じ品質のまま推移していることが分かる．デマとデマを訂正する話題は，共通の単語を利用しているため，従来の話題抽出（クラスタリング）手法では両者を区別した形で話題を抽出することが出来ない．一方，提案手法により，発言者の役割を考慮することで，デマとデマを訂正する話題を区別して抽出することが可能となる．デマを訂正する話題は多くの人が参照したものであり，意味的な品質が高くなるのが理解できる．発言者の役割を考慮することで，こうした意味的な品質の違いを表現することができ，時系列な変化を評価することができる．提案手法が Tweets を 2 部グラフとして捉え，著者の役割を考慮して話題抽出を行ったからこそ，デマを拡散する著者グループ，デマ訂正を行う著者グループを分けることができ，デマの発生から訂正までの時系列推移を表現できたと考える．

これまで述べた研究成果は ICDM，BIGDATA 等のトップカンファレンスで論文発表を行い，最先端の研究者との交流も実施できた．招待講演も国内外で多数行った．こうした結果は筆者の Web サイトでも紹介されている．

今後は，新たに発生する特徴的な話題をいち早く発見し，それがどのような経過をたどるかを予測する手法を開発していく予定である．それにより，社会への影響を測ることが可能となり，より的確なソーシャルメディア解析を行えると考える．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① Hashimoto, T., Shirota, Y., & Chakraborty, B. (2016). "Developing a framework for an advisory message board for female victims after disasters: A case study after east Japan great earthquake". Digital Scholarship in the Humanities, 31(4), 711-724 (2016).
② Shin, K., Kuboyama, T., Hashimoto, T., & Shepard, D., "sCwc/sLcc: Highly Scalable Feature Selection Algorithms", Information, 8(4), 159 (2017).
③ Shirota, Y., Hashimoto, T., & Chakraborty, B., "Visualization of Deductive Reasoning for Joint Distribution Probability in Simple Topic Model". Information Engineering Express 3 (1), 1-8.

〔学会発表〕（計 9 件）

① Hashimoto, T., Kuboyama, T., Okamoto H., & Shin K., "Topic Extraction on Twitter Considering Author's Role based on Bipartite Networks", Proc. of DS2017, pp.239-247 (2017).
② Hashimoto, T., Kuboyama, T., Okamoto H., & Shin K., "Topic life cycle extraction from big Twitter data based on community detection in bipartite networks", Proc. of BigData 2017, pp.2740-2745 (2017).
③ Hashimoto, T., Kuboyama, T., Okamoto H., & Shin, K., "Topic Extraction from Millions of Tweets based on Community Detection on Bipartite Networks" Proc. of the 27th International Conference on Information Modelling & Knowledge Bases, pp.409-424 (2017).
④ Pino Angulo, A., Shin. K. and Hashimoto, T., "New Hybrid Feature Selection Algorithm based on Consistency Measures and Simulated Annealing Search", Proc. of ITISE 2017 (International work-conference on Time Series), pp.575-584 (2017).

⑤ Hashimoto, T., Shepard, D., Kuboyama, T., & Shin, K. (2016, December), "Topic Extraction Method from Millions of Tweets Based on Fast Feature Selection Technique CWC", Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference, pp. 724-731 (2016).
⑥ D. Shepard, T. Hashimoto, T. Kuboyama, & K. Shin, "What Do Boy Bands Tell Us About Disasters? The Social Media Response to the Nepal Earthquake", Digital Humanities, pp.361-363 (2016).
⑦ Jotikabukkana, P., Sornlertlamvanich, V. S., Shirota, Y, and Hashimoto, T., "Disaster Consequence Analysis of Thailand's Severe Flood", 2016 International Workshop on Smart Info-Media Systems in Asia, pp. 395-408 (2016).
⑧ Shepard, D., Kawano, Y., and Hashimoto, T., "Betweenness Centrality As a Measure of User Influence During Disasters", 2016 International Workshop on Smart Info-Media Systems in Asia (2016).
⑨ Hashimoto, T., Shepard, D., Kuboyama, T., & Shin, K., "Event detection from millions of tweets related to the great east japan earthquake using feature selection technique", Data Mining Workshop (ICDMW), 2015 IEEE International Conference ,pp. 7-12 (2015).

〔その他〕
ホームページ等
http://www.cuc.ac.jp/~takako/research/

６．研究組織
(1)研究代表者
橋本隆子 (Hashimoto, Takako)
千葉商科大学 商経学部 教授
研究者番号：80551697

(2)研究分担者
白田由香利 (Shirota, Yukari)
学習院大学 経済学部 教授
研究者番号：30337901

久保山哲二 (Kuboyama, Tetsuji)
学習院大学 付置研究所 教授
研究者番号：30337901

チャクラボルティ バサビ (Chakraborty, Basabi)
岩手県立大学 ソフトウェア情報学部 教授
研究者番号：90305293