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　　６　月　１３　日現在

機関番号：　１６１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３０９

研究課題名（和文）コミュニティと用語の同時獲得手法に関する研究

研究課題名（英文）A study on simultaneous extraction of SNS communities and terms

研究代表者
吉田　稔（Yoshida, Minoru）

徳島大学・大学院社会産業理工学研究部（理工学域）・講師

研究者番号：４０３６１６８８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、SNSにおいて、類似の傾向を持つユーザーの集まりを「コミュニティ」と定義し、類似のコミュニティが類似した用語を用いるという仮定のもと、コミュニティ抽出と用語抽出を並行して行う手法についての研究を行った。具体的には、マイクロブログからランダムサンプリングされたプロフィールと発言のペアから、単語の共起関係を取得するシステムの作成や、単語分散表現ベクトルの学習において、プロフィール中の単語と本文中の単語の共起関係を反映した学習を行うモデルの提案を行った。

研究成果の概要（英文）：We studied a method for extracting communities (i.e., the users that have similar characteristics) and the terms used in the communities simultaneously, based on the assumption that similar communities use similar terms. We focus on the profiles written by the users on Twitter, and we developed a system that extract the frequently-used terms by the users with the profiles that include the given query word. We also developed a method for word distributed representations that reflect the co-occurrence of the words in profiles and tweets, and cluster users and tweets based on the learned vector representation.

研究分野：　自然言語処理

キーワード：　SNS　分散表現　コミュニティ抽出

１．研究開始当初の背景

本研究提案では、WWW 上の「コミュニティ抽出」と、そのコミュニティに関する「用語抽出」を同時に行う手法の研究を行う。

ユーザーが興味のある事柄について調べるとき、関連するキーワードを用いて WWW 文書を検索し、知識を得ることが一般的となってきている。しかしながら、WWW 上の文書は、様々な思想や趣味を持つ記述者によって記述されており、記述者の立場によって、書かれている内容（視点）も異なっていることが普通である。このような記述者の立場や視点を把握することは、メディアリテラシーの観点から重要であるが、そのような判断を、一つの文書を短時間見ただけで行うことは容易ではない。

一方、近年、Twitter や facebook をはじめとするソーシャルネットワーキングサービス（SNS）の普及により、同一の分野に属する人物同士のコミュニティ（クラスタ）が WWW 上で形成されている。WWW 上の発言や文書が、どのようなコミュニティに属するユーザーに記述されているかを知ることにより、例えば、あるテーマについて、特定の観点からではなく、多様なコミュニティ（観点）からの考え方を得ることに役立つ。

SNS ユーザーのコミュニティ抽出に関しては、国内外でこれまで様々な研究が行われている。これらコミュニティ抽出技術を用いることで、SNS 上の発言に関しては、発言者の属するコミュニティを発見することができる。しかしながら、そのコミュニティがどのような性質を持っているのかは、コミュニティの所属者の発言をより詳細に読む必要がある。さらに、WWW 上では、HTML 文書、SNS 上の発言、blog の文書、テキストファイル等、様々な種類の電子テキストが存在し、それぞれの形態がそれぞれの特徴を持ち活用されている。このため、特定の形態に依らない解析手法が、偏りのない視点を得るためには不可欠であるが、SNS 以外のテキストについては、上記のコミュニティ発見手法の適用は難しい。

これに対し、本研究提案では、用語を抽出することによる、コミュニティの性質の簡潔な要約、および、抽出された用語を利用することによる、SNS に限らない文書を対象としたコミュニティ抽出を目標とする。

単語よりも長い複合語や、分野特有の言い回しは、単なる単語と比べ、分野特定力が強いことが期待できる。クラスタリングによるコミュニティ抽出と、特徴文字列抽出技術を組み合わせることにより、「同姓同名人物に限らない、同一分野に属する人物のまとまりの抽出」及び、「単語に限らない、クラスタ特有の文字列、およびその類義語の抽出」を実現することができると考え、本研究提案の着想に至った。

２．研究の目的

WWW 上の文書の内容から類似するユーザーをまとめる「コミュニティ抽出」を、各コミュニティで良く使われる用語を体系的にまとめた「用語抽出」と並列に行うための手法について研究を行う。既存研究では主にＳＮＳのリンク構造を中心に研究されてきたコミュニティ抽出を、文字列抽出を介することにより、一般の文書でも可能とする。また、従来手法では主に単語に限定されていた「コミュニティに特徴的な表現の抽出」を、複合語等の、多様な文字列表現に対応させるため、専門用語抽出手法を組み合わせた新たな手法を提案する。

ある分野の文書を選択するために、WWW 上の全文書を調べ分類することは極めて困難であるが、「特徴的な文字列」を手掛かりとすることにより、将来的には、Web 検索、SNS 検索等、既存の検索サービスを利用し、分野に属する文書を網羅的に取得することが可能になる。

３．研究の方法

研究タスクをコミュニティ発見・用語抽出と、関連語抽出に切り分け、それぞれについて研究を行う。Twitter のデータに関しては、公式に提供されている API を用い、データを収集する。

（１）コミュニティと用語の同時発見に関する研究：

本研究提案では、一方にコミュニティのクラスタをモデル化し、もう一方にコミュニティを識別する特徴をモデル化したモデルを用いる。これにより、本研究提案のモデルでは、単語、複合語、固有名詞を「特徴文字列」として統一的に扱う。特に、SNS 上で高頻度で記述される、ユーザーの公開プロフィール情報を手がかりとして、コニュニティ抽出を試みる。具体的には、プロフィール中に出現する単語と、本文中に出現する単語の共起関係を計算し、「この単語をプロフィールに記述するユーザーは、この単語を発言しやすい」「この単語を発言するユーザーは、この単語をプロフィールに記述しやすい」といった関係をマイニングする。また、単語に限らない、任意の文字列でこのような共起関係を取得する手法についても研究を行う。提案者は、これまで、単語単位に限らない特徴文字列抽出に関する研究を行ってきており、この技術を応用する。

また、単純な共起関係のみならず、近年研究の盛んな単語分散表現に関しも研究を行う。すなわち、単語ベクトルの学習において、プロフィール中の単語と本文中の単語の共起関係を反映した新たな学習方法について研究を行う。

（２）シソーラス構築に関する研究：

発見された特徴文字列どうしを関連付ける、あるいは、新たな関連文字列を発見する手法を開発する。提案手法では、単語抽出に依らない特徴量として、通常の単語に加え、

「単語切り分けを必要としない文脈抽出」「数値情報を文脈として抽出」等を有力な特徴量として用いる。

特に、専門的な話題においては、特徴的な文字列には、数値的表現が含まれることが多い。(例：(ＣＰＵクロック)「3GHz」(赤字)「700 億円」等）しかしながら、数値的表現は、値の表記ゆれ（漢数字、カンマの有無、等）や、値そのものの揺れ（3.1GHz を 3GHz と表記、等）が通常の専門用語より多いため、数値を表現するための適切な手法を研究することで、この問題に対処する。

（３）抽出クラスタ提示システム構築：

コミュニティと用語だけでなく、関連する文書を簡潔に表示する要約モジュールを開発する。実際にユーザーの入力した検索語から、コミュニティと関連文書を取得し、用語リストをリアルタイムに構築するシステムの実装を行う。検索システムを実装することによって、処理速度等、実装に関する問題を明らかにすることができる。特に処理速度に関して、リアルタイム性を維持するためのデータ構造等について検討を行う。

４．研究成果

（１）コミュニティ抽出：

Twitter を対象に、ランダムサンプリングで、プロフィール文字列と発言文字列のペアを取得する方針での収集を行った。収集したコーパスのサーベイを行った結果、プロフィール文字列が、ユーザーの特性を取得するのに極めて有効であるという見通しを得た。また、プロフィール文字列における特定のキーワードの有無を区別することで、プロフィールの違いによって異なる抽出文字列を得ることができ、これが、対義語等の取得に有用であることを確認した。

コミュニティ抽出については、この、発言者のプロフィール情報を手がかりとした手法を発展させ、コミュニティを「視点」として捉え、発言者の視点を分類する手法を提案した。

具体的には、単語分散表現の獲得において、プロフィール欄に記述された単語と、本文中に記述された単語を別々のベクトルで表現することで、より視点抽出に効果的なベクトルを獲得するための手法を開発した。実際に、いくつかのクエリにおいて、複数の視点が得られることを確認し、提案手法の精度を測定した。

また、具体的なコミュニティ抽出の一例として、音楽アーティストのファンを対象に、プロフィール中に共起するアーティスト名を取得することでアーティスト推薦を行う手法についても研究を行った。その他、ジオタグの付与された発言を対象に、発言者の所在地をコミュニティの一種と考え、発言者の位置に特徴的な文字列の発見と、特徴的な文字列を得られる地理的な境界を交互に取得する手法について研究を行った。

（２）用語抽出：

Twitter の API を用いたコーパス取得について、指定したアカウントから発言ログを取得するシステムを開発し、実際にいくつかのアカウントでログを取得した。各アカウントのログから、用語の抽出を行うためのアルゴリズムを開発し、用語抽出を行った。

また、用語抽出アルゴリズムの改良を行った。用語抽出の中間計算結果の保存手法を改良することで、厳密かつ省メモリで計算を行えるようになったほか、処理も高速化された。これにより、中間結果を用いて動的に用語候補スコアを更新することが可能になり、結果として、すでに抽出された用語をその後回避することで、より精度の高い用語リストを得られるようになった。また、この応用として、クエリ文字列に関連する用語を動的に取得する、高速な関連文字列抽出手法を実装し、実際に妥当な文字列が得られることを確認した。

（３）関連語抽出：

抽出された文字列を用いて文書を検索し、関連するツイート（発言）からなる部分集合を形成し、そこからあらためて用語抽出を行うことで、各用語の関連語を取得するためのアルゴリズムを開発した。実データに適用したところ、ある程度の関連語を取得できるほか、元の用語が不完全な文字列だった場合にこれを補完する効果もあることが確認できた。

また、その他、文字列を対象とした分散表現獲得の可能性についても検討を行い、文字列対象でもある程度妥当な分散表現を得ることが可能であるという知見が得られた。

（４）数値文字列の表現：

数値文字列の表現に関しては、有効数字の概念を用いた抽象化手法と、数値範囲を設定し量子化する手法の両者について開発を行い、それぞれの特性を比較した結果、用途に応じた使い分けが必要であるという知見を得た。

また、数値文字列を分散表現により表現する手法の研究を行った。桁数と有効数字を用いた文字列化の手法と、数値を連続値として捉え、動的に文脈ベクトルを取得できる手法の両者を検討し、それぞれの手法である程度妥当な文脈が取得できることを確認した。

図１：発言クラスタと、対応するプロフィール単語クラスタの俯瞰システム

（５）システム実装：
抽出結果を俯瞰するためのシステムの実装を行った。具体的には、異なる視点からの発言どうしを同一のクラスタにまとめ、プロフィール中の単語クラスタ及び発言クラスタを効率的に一覧するためのシステムを実装した。（図１）

（６）その他の研究成果：
そのほか、Twitter データのマイニングに対する文字列抽出とは別の方向性として、トピックモデルによる俗語の分析、アカウントの性格推定、特定分野を対象としたツイート（発言）のカテゴリ分類に関する研究も行った。また、テキスト中の数値表現の取り扱いに関して、その意味付けを行うための知識を、Wikipedia 上の表形式を利用して学習するアルゴリズムを開発した。
その他、顔文字の分類に関する研究及び、動画投稿サイトのコメントに使われる用語に関する研究も行った。また、用語抽出に関して、単語分散表現を利用した類義語検索において、従来の検索を高速化するための手法についても研究を行った。

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２件）
① Kazuyuki Matsumoto, Akira Fujisawa, Minoru Yoshida and Kenji Kita : Emotion Recognition of Emoticon Based on Character Embedding, Journal of Software, Vol.12, No.11, 849-857, 査読有, 2017.

②松本 和幸, 土屋 誠司, 芋野 美紗子, 吉田 稔, 北 研二 : 感性を考慮した日本語俗語の標準語変換, 人工知能学会論文誌, Vol.32, No.1, WII-A_1-12, 査読有, 2017 年.

〔学会発表〕（計１１件）
①Minoru Yoshida, Kazuyuki Matsumoto and Kenji Kita : Distributed Representations for Words on Tables, Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2017), Part-I (LNAI 10234), pp. 135-146, 査読有, 2017.
②Minoru Yoshida, Kazuyuki Matsumoto and Kenji Kita : Acceleration of Similar Word Search in Distributed Representation, Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2017) (poster), 査読有, 2017.

③吉田 稔, 松本 和幸, 北 研二 :
プロフィール情報を用いたテキストの視点分類, 情報処理学会研究報告, Vol.2017-NL-234, No.11, 1-5, 査読無, 2017.

④ Hirokimi Fukuda, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita : Index Generation of BGM Video Based on Distinctive Comments, Proceedings of The 12th International Conference on Natural Language Processing and Knowledge Engineering(NLP-KE'17), pp. 179-184, 査読有, 2017

⑤ Akira Fujisawa, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita : Facial Expression Classification Based on Shape Feature of Emoticons, Proceedings of 1st International Conference on Machine Learning and Data Engineering (iCMLDE2017), pp. 29-34, 査読有, 2017.

⑥松本 流星, 吉田 稔, 松本 和幸, 北 研二 : Twitter を用いた感染症発生動向の視覚化, 人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会(第 15 回), 48-53, 査読無, 2017

⑦Minoru Yoshida, Kazuyuki Matsumoto and Kenji Kita : Table Topic Models for Hidden Unit Estimation, Proceedings of the 12th Asia Information Retrieval Societies Conference (AIRS2016), LNCS 9994, 302-307, 査読有, 2016.

⑧吉田 稔, 松本 和幸, 北 研二 : 表形式からの分散表現獲得, 情報処理学会自然言語処理研究会研究報告, Vol.2016-NL-229, No.19, 1-6, 査読無, 2016.

⑨吉田 稔, 松本 和幸, 北 研二 : 表形式のトピックモデルとその数値単位推定への応用, 情報処理学会自然言語処理研究会研究報告, Vol.2016-NL-226, No.16, 1-6, 査読無, 2016.

⑩岩朝 史展, 松本 和幸, 吉田 稔, 北 研二 : Twitter ユーザの属性別感情推定の検討, 言語処理学会第 22 回年次大会講演論文集, 査読無, 2016.

⑪松岡 雅也, 松本 和幸, 吉田 稔, 北 研二 : トピック変動の分析による俗語の特徴抽出, 情報処理学会研究報告, Vol.2016-NL-225, No.4, 1-5, 査読無, 2016.

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
吉田　稔　　（Yoshida Minoru）
徳島大学・大学院社会産業理工学研究部
（理工学域）・講師
研究者番号：40361688