4l1-3

分散表現を用いたコミュニティにおける単語使用傾向の分析

Analysis on the Word Usage in Communities by Using Distributed Expression

丸井 淳己[*1] 則 のぞみ[*2] 榊 剛史[*1*3] 森 純一郎[*1]
Junki Marui Nozomi Nori Takeshi Sakaki Junichiro Mori

[*1]東京大学大学院工学系研究科 [*2]京都大学大学院情報学研究科
University of Tokyo, School of Engineering Kyoto University, Graduate School of Informatics

[*3]株式会社ホットリンク
Hottolink, Inc.

The popularization of social media exposes the structure of people's conversation - with whom, on what topics and with what kinds of words people speak. In this paper, we conducted empirical analyses on the relation between social network (with whom people speak) and language network (on what topics and with what kind of words people speak) using a large dataset from Twitter, which covers more than 7M people. By qualitative and quantitative analyses, we revealed that there is a distinct difference in the use of language among communities extracted from the social network. Our findings include (1) we can extract a community composed of people who use the same kinds of jargon by exploiting information from both the social network and word usage, (2) when we focus on similarity among communities in terms of both interaction and word usage, we can find specific patterns based on the people's profile information including attributes and interests.

1. はじめに

「類が友を呼ぶ」ということわざもある通り、人は自分と似た人につながる傾向にあると信じられている。この性質をHomophily と呼ぶが [McPherson 01]、一体彼らはどのように似ているのだろうか。Homophily については社会ネットワーク上の距離と人同士の類似度の観点から様々な研究が行われてきた。しかし、彼らがつながっているから似ているのか、似ているからつながっているのかという問いは未だに残ったままである [Shalizi 11]。

本論文では Twitter 上の 700 万ユーザ以上を網羅した大規模データを用いた分析を行うことで、コミュニティにおけるHomophily について調べた。そのために (a) プロフィールに書かれたその人の興味や属性、(b) どのような言葉を用いているか、という 2 点からコミュニティ内の類似性を分析した。本研究におけるコミュニティとはその外よりも中で相互のやりとりが多い集団を指している。また Homophily についてより探るために、コミュニティ同士がどう類似しているかを、会話ネットワークでの距離と言葉遣いの近さの両方の観点から調べた。

まずデータから会話ネットワークを切り出し、コミュニティ抽出を行う。それぞれのコミュニティの特徴を知るためにそれぞれのコミュニティに属するユーザのプロフィールからラベル付けを行う。その後それぞれのコミュニティの言葉遣いを見るために、コミュニティごとにユーザの書き込みをまとめ、それをコーパスとしてニューラルネットワークを用いた言語モデルを学習させる。それぞれのコミュニティでの言葉の使い方の違いをいくつかの例で確かめ、言葉遣いの差の定量的な評価とコミュニティ間のやりとりの量の 2 つがどのような関係を持つか、コミュニティの性質と合わせて分析する。

我々の知る限り、本研究は (1) プロフィールと言葉遣いという観点から見た Homophily、(2) 大規模データからみたコミュニティ間のやりとりの多さと言葉遣いの関係の 2 点を調べた最初の研究である。

連絡先: 丸井淳己, 東京大学大学院工学系研究科技術経営戦略学専攻 坂田・森研究室, marui@ipr-ctr.t.u-tokyo.ac.jp

2. 関連研究

社会ネットワーク分析において、人のつながりと類似性の関係について多くの研究がなされてきた。Romero らはフォロー関係のネットワークとハッシュタグを用いて社会ネットワークと興味の類似性について研究した [Romero 13] 。また情報伝播については盛んに研究されており [Bakshy 12, Hoang 12, Agrawal 12]、特に Bakshy らの研究は、“つながっているから似ているのか、似ているからつながっているのか” という本質的な問いに取り組んだものである [Bakshy 12]。人々の情報拡散行動に SNS のつながりの強さがどう影響するか対照実験によって調べ、やりとりの頻度の高い人からの影響は強いが、やりとりの低い人からの情報伝播の方が新しい情報を得るには重要な役割を演じることを観測し、“弱い紐帯” と “強い紐帯” が情報拡散において違った作用をした事を発見した。本研究で扱う会話ネットワークについても最近研究が進んでおり、Sofus A. らは Twitter の会話について解析し、ソーシャルメディアにおける会話行動について観察している [Macskassy 12]。

最も関連している研究は Bryden らによる Twitter のコミュニティの研究であろう [Bryden 13]。20 万ユーザを Twitter からサンプリングした上で、フォロー関係を用いてコミュニティ分割を行い、コミュニティ内の会話のキーワードを取り出すことでコミュニティの特徴付けを行っている。さらにこのキーワードでどのコミュニティに属しているかの予測もしている。しかし我々の研究はさらに多くのユーザ数を対象としており、言葉遣いについても語の頻度だけでなくニューラルネットワークを用いた言語モデル (以下 NPLM と呼ぶ) で踏み込んだ分析を行い、コミュニティの特徴付けだけでなくコミュニティごとにどのように違うかを観察している。

NPLM が最初に導入されたのは Bengio らによってである [Bengio 03]。n-gram モデルの次元の呪いを回避するために、単語に分散表現、すなわち単語ベクトルを導入した。そのモデルはいくつかの単語列 (コンテクストと呼ぶ) の次の単語を予測す

るモデルとなっている。彼らは、コンテクストの単語ベクトルを結合させたものをフィードフォワード・ニューラルネットワークに入力し、次に来る単語の非正規化対数事後確率を出力するものを考えた。しかし事後確率を正規化するためにソフトマックス関数が必要となり、モデルの学習の際にその勾配を計算すると項数が全単語数分となるためにその計算時間が問題となった。計算手順を減らすために階層化ソフトマックス [Morin 05, Mnih 08] や Noise contrastive estimation[Mnih 12] といった手法が提案されたが、それらの手法が Mikolov らにより単純化・高速化された [Mikolov 13]。Mikolov らはニューラルネットワークを単純化し、ターゲットの単語の周りの単語を語順と関係なしにコンテクストとして与えた。大規模なデータに適応することで生成される単語ベクトルの “質” が改善され、似た文脈で使われる単語がコサイン距離の近いベクトルとして表現され、語同士の関係もベクトルの加減演算で得られるようになった。現時点で NPLM によって多義語や同表記語の問題を完全に解決することはできていないが、我々は単語ベクトルがその単語が使われるコンテクストを表現し、似た単語が近いベクトルとして表現されるという性質に着目して今回の分析を行っている。

3. 分析フレームワーク

3.1 コミュニティ抽出

Twitter での友人関係を取り出すために会話ネットワークを使う。本研究ではネットワークを構成するために、相互にメンションをしているユーザにエッジを張ることにした。多くの分析ではフォロー関係を用いているが、本研究では Homophily についての調査であるため相互に会話をしたということが重要だと考えた。コミュニティ抽出には大規模なネットワークに対して効率の良い Louvain 法を用いる [Blondel 08]。貪欲法で Modularity を最適化する手法でボトムアップにコミュニティ分割を行う手法であり、実装は公開されているものを用いる*1。

3.2 コミュニティの特徴付け

それぞれのコミュニティに属するユーザのプロフィールを集め、特徴語の計算で最もよく使われる TF-IDF スコアを用いてコミュニティごとのキーワードを計算する。スコアの高い上位 20 単語をそのコミュニティを表している特徴的な単語とみなし、それぞれのコミュニティにラベルを付けた。

3.3 コミュニティごとの単語ベクトルの学習

コミュニティごとに書き込みをサンプリングして集め、それぞれのコミュニティに対して単語ベクトルを学習させる。先述の通り単語ベクトルはその単語が現れるコンテクストを表しているので、それぞれのコミュニティでどのようにある単語が使われているか調べることができる。そのために、全体から1%サンプリングしたツイート群を作り、ベースラインのコーパスとする。その上でそれぞれのコミュニティのツイート群のコーパスを作る。コミュニティごとのコーパスの大きさを揃えるため、ユーザ数に応じたサンプリング (100 万人で 1%、1 万人で 100%) を行う。

単語ベクトルの学習には Mikolov らの NPLM[Mikolov 13] を実装した “word2vec”*2 を用いた。まずベースラインのコーパスを学習させ単語ベクトルを得てから、それぞれのコミュニティのツイート群のコーパスを使って単語ベクトルを再学習させる。word2vec は階層化ソフトマックスと負例サンプリングの 2 つを実装しているが、追加的な学習をさせるために負例サンプリングを用いる。

3.4 コミュニティごとの言葉遣いの違いの分析

コミュニティ同士の近さを見るために、会話ネットワーク上の近さと言葉遣いの近さの 2 つの類似度の定義をする。

定義：会話ネットワークにおけるコミュニティ間の類似度
コミュニティ i と j の会話ネットワーク上の近さ $Sim_{social}(i,j)$ は次の通り。

$$Sim_{social}(i,j) = \frac{|E_{i,j}|}{|V_i||V_j|}, \tag{1}$$

ここで $|E_{i,j}|$ はコミュニティ i と j の間のエッジの数、$|V_i|$ はコミュニティ i におけるノード (ユーザ) 数を指している。

定義：言葉遣いにおけるコミュニティ間の類似度
コミュニティ i と j の言葉遣いの近さ $Sim_{word}(i,j)$ は次の通り。

$$Sim_{word}(i,j) = \frac{1}{N}\sum_{a=1}^{N}\frac{|S_{ia} \wedge S_{ja}|}{|S_{ia} \vee S_{ja}|} \tag{2}$$

$$S_{ia} = \{w_b | \text{top 30 of } v_{ia} \cdot v_{ib}\} \tag{3}$$

ここで w_a は a 番目の単語 w、v_{ia} は w_a のコミュニティ i における単語ベクトル、“·”はベクトルの内積を指し、N はコミュニティ i、j 両方に 100 回以上出現する単語の総数を指している。コミュニティごとに違うツイート群を与えられて単語ベクトルが計算されるので、w_a に対応する単語ベクトルはコミュニティごとに違うことになる。S_{ia} はコミュニティ i の書き込みにおいて w_a に近い単語群を表し、$Sim_{word}(i,j)$ はコミュニティ i と j の間で、その近い単語群がどれだけ似ているかの Jaccard 係数を取ったものである。

これら 2 つの類似度でコミュニティの関係をプロットし、これら 2 つの類似度の相関関係やコミュニティの種類による違いを見る。

4. 実験と結果

4.1 データセット

ツイートは 2012/1/1 から同年 12/31 に渡って、Twitter API で日本語で書き込んでいると判定されたユーザを対象に取得した。取得された 49 億ツイートから返信をしていると判定されたツイートを取り出してネットワークを作った所、含まれるユーザ数は約 700 万であり、404 百万リンクのうち 125 百万リンクが取得期間内に相互に返信をしているものだった。

4.2 会話ネットワークのコミュニティ

Louvain 法を適用した所、34835 のコミュニティが抽出された。1 万人以上のユーザを含むコミュニティのみに絞ったところ 38 のコミュニティがあることが分かり、それらで全体の 97.7%のユーザを占めることがわかった。図 1 はこのコミュニティを可視化したものである。それぞれのノードはコミュニティとその ID を表している。ノードの大きさはコミュニティのユーザ数を表し、エッジの太さはそれぞれのコミュニティのリンク数である。中央のコミュニティが他のコミュニティと多くのやりとりをしている様子がわかる。

4.3 プロフィールから見える特徴

コミュニティのラベル付けを行うために、プロフィールから TF-IDF を計算しキーワードを取り出した。するとはっきりとコミュニティごとの特徴をつかむことができ、容易にラベル付

*1 https://sites.google.com/site/findcommunities/
*2 https://code.google.com/p/word2vec/ に公開されている

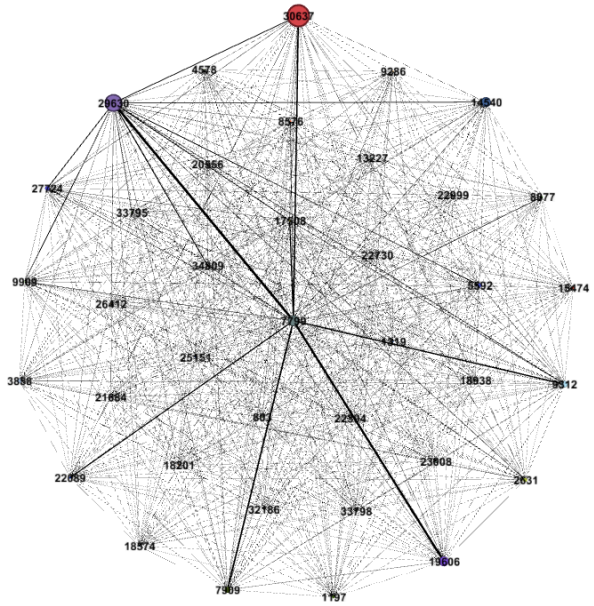


図 1: 会話ネットワークから抽出したコミュニティネットワーク

けができた。表 1 にコミュニティごとのキーワードとラベルを載せた。紙面の都合上 38 コミュニティから幾つか選んだ。ほとんどのコミュニティは (a) 同じか地域が近い高校、(b) 同じか地域が近い大学、(c) 趣味・興味によるコミュニティの 3 種類に分かれた。

4.4 言葉遣いの特徴

言葉の使われ方がコミュニティによってどのように違うか見るために、同じ単語がどのように違うコンテクストで使われているかを調べた。ここでは異なる意味に用いられた単語を紹介する。先行研究では単語の頻度分布のみに着目して、単語の分布が統計的に違う事を示していた [Bryden 13]。それに対して本研究は NPLM を用いてそれぞれの語のコンテクストを捉えた分析を行っている。

異なった意味に使われる単語を得るために、単語ごとにそれぞれのコミュニティでの単語ベクトルの差をコサイン距離で計算し、コミュニティ間の差が大きい単語を取り出した。4 番目にコミュニティ間の差異が大きい「ミート」という単語を例に取る。この単語はコミュニティ26029(オンラインゲームファン) とコミュニティ23008(ディズニーファン) との間で最も差異が大きい単語であった。

表 2 にそれぞれのコミュニティで、コンテクストを共有している度合いの高い語上位 10 語をコサイン類似度とともに載せた。オンラインゲームファンの方では「ミート」は基本的には肉の意味で用いられ、「早食い」「フローズン」といった単語とコンテクストを共有しているという結果となった。しかしこのコミュニティにおける肉とはオンラインゲームでのアイテムも指していて、モンスターハンターというゲームのアイテムである「ホットドリンク」「シモフリトマト」といった単語とも近いという結果となった。一方でディズニーファンの方は「イン」「ぜひぜひ」といった単語とコンテクストを共有し、一見するとどのような意味であるか不明確だが、顔文字や絵文字、「するする」といった書き込みから女子らしい書き込みが垣間見える。ツイートの中身を見てみると、「イン」「ミート」とはディズニーファンの間では「ディズニーランド/シーに入る」「ディズニーランド/シーで会う」ことを意味していることがわかった。

このようにコミュニティ間で単語ベクトルのコサイン類似度を取る単純な手法だけで、こういった複数の意味を持つ単語を取り出すことが可能であることが分かった。特に、このような手法でコミュニティ間でのジャーゴンを比べることができるので、カルチュラル・スタディーズに有用だと考えられる。

コミュニティにおける単語の使われ方の差異を分析することで、それぞれのコミュニティが特有の言葉遣いをしていることがわかった。Bryden らによる研究でも示唆されている通り [Bryden 13]、この分析もソーシャルメディア上の友人は似た言葉遣いをしているという Homophily があることを示唆している。

表 2: コミュニティごとの「ミート」に近い語

コミュニティ26029 オンラインゲームファン		コミュニティ23008 ディズニーファン	
単語	類似度	単語	類似度
早食い	0.821099	イン	0.844532
フローズン	0.802880	タイミング合え	0.803420
ドスヘラクレス	0.800420	次いつ	0.801293
クランチ	0.792865	ぜひぜひ	0.801245
ラード	0.791672	(＞﹏＜)!!!	0.792765
ホットドリンク	0.782958	するする	0.790120
アサイ	0.778490	まなみん	0.789295
銀シャリ	0.778290	??♡	0.787898
シモフリトマト	0.775939	それでもよければ	0.785390
レッグ	0.775758	きんかん	0.782860

4.5 コミュニティ間の類似度

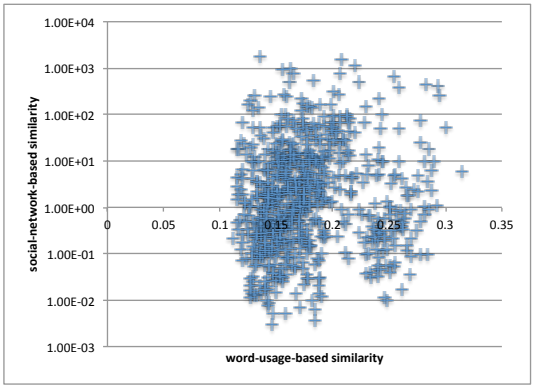
次にネットワークの近さと言葉遣いの近さに相関関係が見られるか、38 コミュニティのペア全てに対してネットワークの近さと言葉遣いの近さを前章の定義に従って計算しプロットした (図 2-a)。ネットワークの近さはコミュニティ間のやりとりの多さを示し、前節ではやりとりがある友達であると言葉遣いが似ることが示唆されていたので、ネットワークで近いコミュニティ同士は言葉遣いが近いことが期待されたが、図 2-a では無相関であるように見える。しかしコミュニティには高校、大学、趣味・興味の 3 種類があり、それぞれで違う特徴を示すかどうか調べるために、それぞれの種類別でのみ距離をプロットした (図 2-b)。するとそれぞれのコミュニティの種類では違う分布を示すことが分かった。高校コミュニティはネットワークでは比較的遠いが言葉遣いの類似度が高く、大学コミュニティはネットワークで近い上に言葉遣いの類似度が高い。一方で趣味・興味コミュニティ同士はネットワークで近いペアも遠いペアもあるが言葉遣いの類似度は相対的に低いということが見て取れる。ここから、高校や大学といった属性ベースでのコミュニティ同士はやりとりの量と関係なく言葉遣いが似て、興味・趣味ベースでのコミュニティ同士もやはりやりとりの量と関係なく言葉遣いが似ていないと言うことができる。これは Homophily は単純な性質ではなく、属性が近いために友達でなくても似ているという場合と、趣味・興味が近いから友達になる場合の 2 種類があることを示唆していると考えられる。

5. まとめ

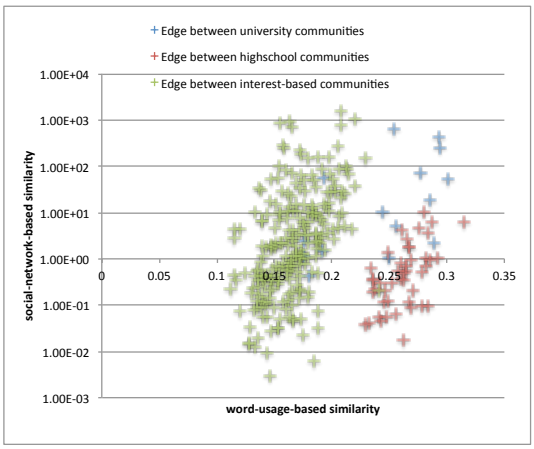
本論文では Twitter の大規模データを用いて社会ネットワークと言語、プロフィール情報の関係について探った。今回は日

表 1: Examples of Labels and Keywords in the top-38 communities

種類	ラベル	キーワード					
高校	東北地方の高校	磐城	湯本	松陵	富谷	勿来	勿工
高校	東京・神奈川・埼玉の高校	松が谷	大宮西	瀬谷	氷取沢	座間	荏田
大学	関西地方の大学	関大	同志社	オリター	武庫	関西大学	近大
大学	東京都の大学	立教	立教大学	法政	新潟大学	学習院大学	英和
趣味・興味	ヴィジュアル系ファン	TERU	ドエル	ハイヲタ	DEARS	SuG	TICK
趣味・興味	FPS ゲームファン	SuddenAttack	osu	サドンアタック	Clan	LoL	CyAC
趣味・興味	バイク、ツーリングファン	ニコツー	新居浜高専	レースシム	西条	iRacing	車載



(a) 全てのコミュニティ間での相関関係



(b) コミュニティの種類別での相関関係

図 2: ネットワーク上の近さと言葉遣いの類似度の相関関係

本語のデータを用いたが、この分析手法は他の言語でも適応可能であると考えている。また今回の分析はカルチュラル・スタディーズにも有用であると考えている。本研究が、社会ネットワークとそこで使われる言葉、そしてプロフィールについてのより深い関係について探るきっかけとなれば幸いである。

参考文献

[Agrawal 12] Agrawal, R., Potamias, M., and Terzi, E.: Learning the Nature of Information in Social Networks, ICWSM '12 (2012)

[Bakshy 12] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L.: The Role of Social Networks in Information Diffusion, WWW '12, pp. 519–528 (2012)

[Bengio 03] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C.: A Neural Probabilistic Language Model, *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155 (2003)

[Blondel 08] Blondel, V., Guillaume, J., Lambiotte, R., and Mech, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, pp. 10008–10019 (2008)

[Bryden 13] Bryden, J., Funk, S., and Jansen, V. A. A.: Word usage mirrors community structure in the online social network Twitter, *EPJ Data Science*, Vol. 2, No. 1 (2013)

[Hoang 12] Hoang, T.-A. and Lim, E.-P.: Virality and Susceptibility in Information Diffusions, ICWSM '12 (2012)

[Macskassy 12] Macskassy, S. A.: On the Study of Social Interactions in Twitter., ICWSM '12 (2012)

[McPherson 01] McPherson, M., Smith-Lovin, L., and Cook, J. M.: Birds of a Feather: Homophily in Social Networks, *Annual Review of Sociology*, Vol. 27, No. 1, pp. 415–444 (2001)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, NIPS '13 (2013)

[Mnih 08] Mnih, A. and Hinton, G. E.: A Scalable Hierarchical Distributed Language Model, NIPS '08, pp. 1081–1088 (2008)

[Mnih 12] Mnih, A. and Teh, Y. W.: A fast and simple algorithm for training neural probabilistic language models, ICML '12, pp. 1751–1758 (2012)

[Morin 05] Morin, F. and Bengio, Y.: Hierarchical probabilistic neural network language model, AISTATS '05, pp. 246–252 (2005)

[Romero 13] Romero, D. M., Tan, C., and Ugander, J.: On the Interplay between Social and Topical Structure, ICWSM '13 (2013)

[Shalizi 11] Shalizi, C. R. and Thomas, A. C.: Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, *Sociological Methods and Research*, Vol. 17, pp. 211–239 (2011)