1M4-3

Community Vector: 分散表現を用いたソーシャルメディアのコミュニティ推定

Community Vector: Community Prediction with Distributed Representations on Social Media

丸井 淳己 [*1] 萩原 正人 [*2] 榊 剛史 [*3] 森 純一郎 [*4]
Junki Marui Masato Hagiwara Takeshi Sakaki Junichiro Mori

[*1*4]東京大学大学院工学系研究科 [*2]楽天技術研究所 [*3]株式会社ホットリンク
School of Engineering, the University of Tokyo Rakuten Institute of Technology Hottolink, Inc.

As the proverb 'birds of a feather flock together' indicates, do users in the same community tend to talk in the similar manner on social media? To answer this question, we extract communities from a conversational graph of Twitter, and conduct a task of predicting a user's community from her/his tweets. We also compare distributed representations and a bag-of-words model to see how they can deal with colloquial language. We use top 38 communities, and sampled 10,000 users from each community. We use 9,000 users per a community for training, 1,000 users for test. The Community Vector model, where users' tweets are embedded in distributed representations together with community information, can predict correct communities with an accuracy of 26.6%. The model outperforms simple bag-of-words model with TF-IDF, whose accuracy is 19.0%.

1. はじめに

古来より「類が友を呼ぶ」と言うように、人々は自分と似た人と親しくなる傾向にあると言われてきたが、一体彼らはどのように似ているのだろうか。ソーシャルメディアの進化によりこのような疑問に答えることができやすくなり、またマーケティングや広告の観点からも重要であるため、口コミを始めとした情報拡散と友人関係について盛んに研究されている。Bakshy らは Facebook の URL 共有を用いた実験を通じて、情報拡散が仲の良い友人でよく起こることを示した [Bakshy 12]。こうした研究ではハッシュタグや URL、他人の発言を引用するといった特徴的な行動から分析することが多いが、全てのユーザがこういった機能を用いるわけではない。これらを使わずにユーザの発言のみから似た性質をもち互いにつながるユーザ集合であるコミュニティを推測できるだろうか。

こうした分析を行うために、本稿では BOW モデルと分散表現モデルを比較する。従来、文書に対しては BOW(bag-of-words) 表現が多く用いられてきた。BOW モデルは各語に異なる次元を割り当て、疎ベクトルとして表現し、分類やクラスタリングに用いられる。しかしながら BOW モデルは語同士の類似性を無視して等しく異なる次元に対応させてしまう、語順が考慮できないという問題点がある。

そこで近年 Bengio らがニューラルネットワーク言語モデル [Bengio 06] を提案したのを皮切りに、単語や文に対してベクトルとして表された分散表現 (distributed representation) を割り当てるモデルが提案されている。Mikolov らは Skip-gram と CBOW モデルを提案し [Mikolov 13]、word2vec というツールに実装している。このモデルでは構文解析などに頼らずコーパスから高品質な分散表現を得ることに成功している。Le and Mikolov はこの考えを文や文書に拡張し、段落ベクトル (paragraph vector) モデルを提案した [*1][Le 14]。このモデルでは段落の分散表現が学習され、類似した意味の段落に対して類似した段落ベクトルが学習される傾向になる。Le and Mikolov は、分類・感情極性推定タスクにおいて、段落ベクトルモデルが従来の BOW やその他のニューラルネットワーク言語モデルを上回る性能を上げることを示した。著者らはこの段落ベクトルの拡張として、カテゴリを分散表現として表現するカテゴリベクトルモデルを提案し、大規模ショッピングサイトにおいて段落ベクトルモデルを上回る性能を得られることを示した [Marui 15]。

ソーシャルメディアの書き込みはネット独特の言い回しを多く含み、またコミュニティによって特徴的な単語使用法等が考えられる。本稿ではカテゴリベクトルモデルをコミュニティとユーザの書き込みに適応させ、これをコミュニティベクトルモデルとして書き込みからコミュニティを推測し、これを BOW モデルと比較する。実験では Twitter の大規模データを用い、分散表現を用いたモデルが BOW モデルを上回ることを示した。また分散表現を用いたモデルで 26.6%程の精度で書き込みからコミュニティが推測できることを示した。

2. コミュニティベクトルモデル

コミュニティベクトルモデルは、カテゴリベクトルモデル [Marui 15] と基本的に同じモデルで、カテゴリの代わりにコミュニティ、段落としてユーザを陽に与え、ツイートの単語 (群) を推定する (図 1)。このモデルは対象の単語をコミュニティベクトル、ユーザベクトルそして文脈ウインドウ内の単語ベクトルから推測する。この際それぞれのベクトルを中間層において結合もしくは平均し、対象の単語を予測する。訓練時の学習を効率的にするために、出力層では word2vec 同様に階層的ソフトマックスもしくは負例サンプリングを用いる。このようにしてコミュニティベクトル、ユーザベクトル、単語ベクトルを同時に学習している。学習アルゴリズムは word2vec や段落ベクトルモデルと類似しているが、初期化の方法を工夫し、SGD に加えて更新式に AdaGrad[Duchi 11] または Adam[Kingma 14] を用いることにより、ベクトルの収束速度を改善した点が異なる。

テスト時にツイートを与えてコミュニティを得るには、単語ベクトルを固定しコミュニティベクトルとユーザベクトルを求める。この際、2 つのベクトルには自由度があるため合計しか求めることができないが、この合計ベクトルをコミュニティの

連絡先: 丸井淳己, 東京大学工学系研究科, 東京都文京区本郷 7-3-1 工学部 3 号館 202 号室, marui@ipr-ctr.t.u-tokyo.ac.jp

*1 ここで言う段落とは、文や文の集合など、何らかの構造的まとまりを持った単語列一般を指す言葉であり、厳密な意味での「段落」ではないことに注意されたい。

推定に使うことができる。

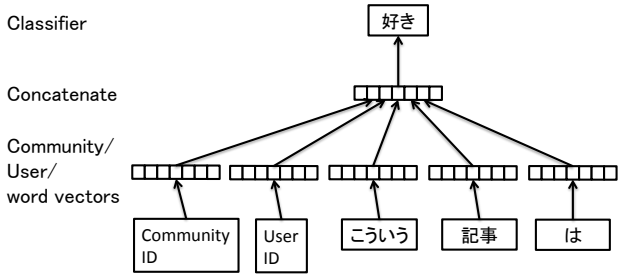

図 1: コミュニティベクトルモデル

3. 実験

本節では BOW モデルとコミュニティベクトルモデル (SGD, AdaGrad, Adam) を比較するため、評価実験を行う。

3.1 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、4 億リンクのうち 1.25 億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。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 法を用いた [Blondel 08]。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。ツイートに対しては、日本語の分かち書きに MeCab を用いている。

3.2 手順

抽出された上位 38 コミュニティから 1 万人ずつサンプリングし、それぞれのユーザについてデータセット内の全ツイートを用いる。それぞれ 9000 人を訓練データに、1000 人をテストに用いる。BOW モデルではユーザごとにツイート群の BOW ベクトルを作る。単語の重み付けを TF-IDF を用いて、ユーザ同士の発言傾向のコサイン類似度を計算する。テストデータのユーザに対するコミュニティとして、訓練データにある最近傍のユーザの所属コミュニティを候補として採用し、その精度を測る。コミュニティベクトルモデルでは 9000 人のツイートでコミュニティベクトル、ユーザベクトル、単語ベクトルを同時に学習させ、テスト時には単語ベクトルを固定しツイートのみからコミュニティベクトルとユーザベクトルの和を学習させる。ユーザの類似度をコミュニティベクトルとユーザベクトルの和同士のコサイン類似度を取ることで測り、訓練データ内の最近傍のユーザが所属するコミュニティを、テストデータのユーザの所属コミュニティ候補として推測し、その精度を測る。出力層は階層的ソフトマックスで近似し、10 回同じデータを繰返し与えて学習させた。

3.3 結果

表 1 に結果を示した。コミュニティベクトルモデルのうち、収束速度を改善した AdaGrad と Adam が BOW モデルを上回っていることが分かる。コミュニティベクトルモデルの中では AdaGrad が最も良く、26.6%の精度でコミュニティを推測できた。つまり書き込みの類似度のみで最も似ているユーザを取り出すと約 1/4 で同じコミュニティとなることが分かる。

BOW モデル	コミュニティベクトルモデル		
	SGD	AdaGrad	Adam
19.0%	15.7%	**26.6%**	19.3%

表 1: コミュニティ推定性能

また学習したコミュニティベクトルからコミュニティ同士の書き込みの類似性を調べることもできる。

4. おわりに

本稿では、(1) 書き込みの傾向のみからネットワーク構造から取り出したコミュニティが推測可能か、(2) 分散表現を用いることで既存の BOW モデルより良い精度で口語を扱えるか、という 2 点についてコミュニティ推測タスクを通じてアプローチした。分散表現を用いることにより、1/4 程度の精度で書き込みのみからコミュニティが推測可能であり、提案手法が BOW モデルを上回った。本稿で用いた手法はコミュニティが学習時に陽に与えられるため BOW モデルに比べて良い精度で推測できると考えられる。

なお、本手法は Twitter やネットワーク構造からのコミュニティに限らず用いることができ、あるユーザ集合をあらかじめ決めておき、それに対する書き込みの傾向を分散表現として取り出すことも可能である。また学習されたコミュニティベクトルを用いてコミュニティやユーザ集合同士の類似性を取り出すことも可能であるため、マーケティングに有用なツールとして使える可能性がある。

コミュニティの推測の精度は高くはないが、今後精度を高めるために局所的なネットワーク構造やプロフィールの情報等をモデルに組み込むといったことも考えられるだろう。

参考文献

[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 06] Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L.: Neural probabilistic language models, in *Innovations in Machine Learning*, pp. 137–186, Springer (2006)

[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)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159 (2011)

[Kingma 14] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Le 14] Le, Q. V. and Mikolov, T.: Distributed Representations of Sentences and Documents, *arXiv preprint arXiv:1405.4053* (2014)

[Marui 15] Marui, J. and Hagiwara, M.: Category2Vec 単語・段落・カテゴリに対するベクトル分散表現, 言語処理学会第 21 回年次大会発表論文集, pp. 680–683 (2015)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)