八戸に関するツイートの ソーシャルセンサーとしての活用に関する一考察 — 潜在的ディリクレ配分法による分析 —

長谷川 魁飛[†]・島内 宏和[††]・武山 泰[†††]

A Consideration for using Tweets referred to Hachinohe as a Social Sensor — Analysis by Latent Dirichlet Allocation —

HASEGAWA Kaito[†], SHIMAUCHI Hirokazu[††]and TAKEYAMA Yasushi[†††]

ABSTRACT

In this paper, the tweets referred to Hachinohe were analyzed, that were tweeted in October 2020. The total 35,459 tweets that include the word “Hachinohe” were collected by Twitter API with Python. For analyzing the user base, the topics in user profiles are extracted by Latent Dirichlet Allocation, a probabilistic language model. Latent Dirichlet Allocation was adapted to the tweets for analyzing the topics in the tweets referred to Hachinohe. The results conclude that the direction of utilizing the tweets could be considered as a social sensor for Hachinohe.

Key Words : *Hachinohe, Twitter, Latent Dirichlet Allocation, Social Sensor, Neural Language Model*
キーワード : 八戸, *Twitter,* 潜在的ディリクレ配分法, ソーシャルセンサー, ニューラル言語モデル

1. ソーシャルセンサーとしての Twitter

近年、SNS は社会に広く浸透し、その代表的なもののひとつである Twitter の月間利用者数は 2017 年時点で 4,500 万人を超えている ([15] を参照)。Twitter は非常に速報性の高いプラットフォームであり、利用者が移動中や外出中にスマートフォンの Twitter アプリを利用してリアルタイムに投稿を行う形態が広まっている。そのため、頻繁に投稿されるツイートをセンサー情報の一種と捉え、実世界を観測するために活用しようとする研究が活発に進められている。(活用事例は、[9] などの論文を参照)。

本論文では、八戸に関するツイートを行うユーザやツイートに含まれる話題について分析し、そのソーシャルセンサーとしての活用の方向性について考察する。具体的には、Twitter が提供している API を用いて「八戸」を本文内に含むツイートを収集する。

収集したツイートの数や、つぶやいたユーザ数、頻度等の基本的な代表値を確認する。また、八戸に関するツイートを行ったユーザのプロフィール欄のテキストを潜在的ディリクレ配分法で分析することで、ユーザの傾向を調べる。同様に、八戸に関するツイートに含まれる話題についても調査する。以上の分析結果を踏

令和 2 年 12 月 7 日受付
† 工学部システム情報工学科・4 年
†† 工学部システム情報工学科・講師
††† 工学部システム情報工学科・教授

まえ、八戸に関するツイートのソーシャルセンサーとしての活用について考察する。

2. 潜在的ディリクレ配分法による分析

2.1 分析に用いるツイート

Python および Twitter API の「Standard Search API[14]」を用い、2020 年 10 月 1 日から 10 月 31 日までに投稿されたキーワードとして「八戸」を含むツイートのうち 35,459 件を収集した[1]。収集したツイートに含まれるユーザの数は 13,652 人であり、この期間における一人当たりの平均ツイート数は約 2.6 件である。最もツイート数が多いユーザのツイート数は 820 件であり、次に多いアカウントは 726 件となっているものの[2]、それ以降のアカウントのツイート数は 250 件以下となっている。ツイート数が上位のアカウントには、八戸に関する天気や鉄道の運行情報、スマートフォンの販売店、野球・サッカーに関する情報を提供するアカウント等がある。他方、ツイート数が 10 件以下のユーザの数は 13,224 人であり、ユーザ全体の 97% を占める形となっている。

2.2 潜在的ディリクレ配分法による分析

3. 以降、八戸に関するツイートを行うユーザと、そのツイート内の話題を把握するために、トピックモデルによる分析を行う。トピックモデルとは、文書集合における各文書が潜在的な話題（トピック）に基づいて生成される過程を確率的に表現した確率的言語モデルの一種であり、情報推薦 [5] やアンケートの自由記述欄の分析 [11] 等、様々な分野において応用されてきた。そのひとつである潜在的ディリクレ配分法は、各文書におけるトピックの分布および各トピックにおける単語の分布がディリクレ分布により生成されると仮定したモデルであり、2002 年に Blei ら [2] により提唱されて以来広く活用されている[3]。潜在的ディリクレ配分法における 1 文書の生成過程は以下の通りである。

1. ディリクレパラメータ α を用い、トピックベクトル θ を求める。
2. 以下を文書の単語の総数 N 回繰り返す。
   （a）トピックベクトル θ からトピック z を選ぶ。
   （b）トピック z と単語生成確率ベクトル β から単語 w を 1 つ選ぶ。

文書の生成過程を式で表現すると、

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{i=1}^{N} p(z_i|\theta)p(w_i|z_i, \beta)$$

となる。ここで、z_i はトピック、w_i は文書の単語である。また、1 文書における事後確率の周辺分布 $p(\mathbf{w}|\alpha, \beta)$ は、

$$\int p(\theta|\alpha)\prod_{i=1}^{N}\sum_{z_i} p(z_i|\theta)p(w_i|z_i, \beta)d\theta$$

となる。

本研究では、潜在的ディリクレ配分法が実装されているパッケージ Gensim[10] を用いて分析を行う。潜在的ディリクレ配分法は教師なし学習の手法の一種であり、適用にあたっては文書集合（ここではツイート本文の集合およびツイートを行ったユーザのプロフィールの集合）に含まれるトピックの数をあらかじめ指定しておく必要がある[4]。ここでは、最適なトピック数を推定するために、Coherence を用いてグリッドサーチを行った。Coherence は、抽出されたトピックが人間にとってわかりすいかどうかを表す指標を目指し開発されたものであり、

[1] Twitter API の「Standard search API」では、Twitter 社にてサンプリングされたツイートの中から特定のキーワードを含むものを収集できるが、該当期間における八戸を含むすべてのツイートが取得できるわけではないことに注意する。

[2] 上位 2 件のアカウントはいずれも八戸の天気に関する情報についてツイートをするアカウントとなっている。

[3] 実際、Google Scholar によると、[2] の引用件数は 11 月 25 日時点で 34,803 件となっている。

[4] その他、重要なパラメータとして事前分布に関するパラメータがあるが、ここでは Gensim の auto の設定を利用している。

現時点で複数の定義が提唱されているが、ここでは Röder ら [12] によるものを採用した。なお、他にもユーザのロケーションなどの項目も利用しうるが、登録されている件数は 6,083 件であり、また必ずしも地名が登録されているわけではないこともあることから、今回は分析の対象からは除外した。

3. 八戸に関するツイートの分析結果

3.1 ユーザプロフィール内のトピックの抽出

Twitter のユーザのプロフィール欄には、出身地や趣味などの情報が、ユーザ自身により入力されている。ここでは、「八戸」を本文に含むツイートを行った 13,224 人のユーザの傾向を把握するために、ユーザプロフィール欄に記載のテキストを文書集合とし、潜在的ディリクレ配分法によりそのトピックを抽出した。

MeCab[7] を用いて、収集したデータ内における各ユーザのプロフィール欄のテキストの形態素解析を行った。潜在的ディリクレ配分法適用の際に考慮する名詞・動詞・副詞・形容詞を抽出した後、クリーニングおよび正規化の前処理を行った。具体的には、URL 等の不要な情報および Slothlib の日本語のストップワードのリスト [13] に記載の単語を除去し、ひらがな・カタカナ・英数字の一文字からなる語は正規表現を用いて除いた。また、数字はすべて 0 に統合した。前処理後のユーザのプロフィール欄全体に含まれる単語数は 38,935 である。前処理の後、トピック数 k を $k = 11, 12, \ldots, 70$ とし、潜在的ディリクレ配分法を適用し、抽出されたトピックの Coherence を算出した。各トピック数ごとの Coherence のグラフを**図 1** に示す。

ここでは、Coherence が最大となる 24 を採用した[5]。抽出された 24 トピックの内、例として比較的鮮明に判定できる 4 つのトピックのワードクラウドを**図 2** に示し、残りのトピックは APPENDIX にまとめる。ユーザプロフィールの中には、旅行に関する話題や野球、スケートに関する話題、ゲームなどの趣味に関する話題などが含まれていた。

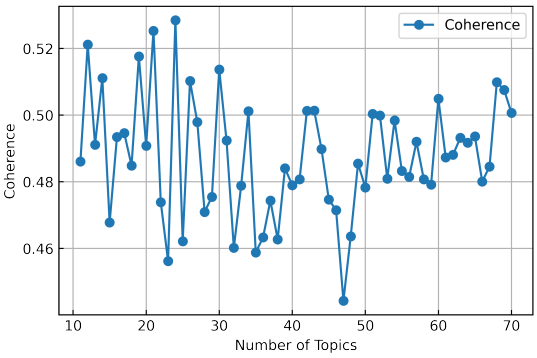

図 1　ユーザのプロフィール欄に対し潜在的ディリクレ配分を適用したときの Coherence

図 2　ユーザのプロフィール欄における 4 つのトピックのワードクラウド（他は APPENDIX を参照）

3.2 ツイートに含まれるトピックの抽出

八戸に関するツイートに含まれる話題を把握するために、10 月 1 日から 10 月 31 日までに投稿された 35,459 件のツイートに対し、潜在的ディリクレ配分法を適用した。**3.1** にて行ったユーザのプロフィールのトピック抽出と同様に、ツイート本文に前処理を行った上で潜在的ディリクレ配分法を適用した。トピック数については、Coherence を指標とし

[5] トピック数が 30, 68 等の場合にも Coherence は相対的に高い値を示しているが、これらを採用する場合には各トピックがより細かなサブトピックに分かれる等、より詳細な分析ができる可能性があることに注意する。ここでは、八戸に関するツイートを行うユーザの概要を把握することを目的としているため、コヒーレンスが最大となる 24 を用いて分析を進めている。**3.2** のツイートに含まれるトピックの抽出についても同様である。

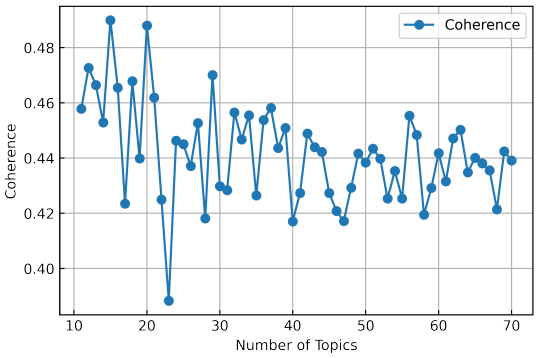

図 3　ツイートに対し潜在的ディリクレ配分法を適用したときの Coherence

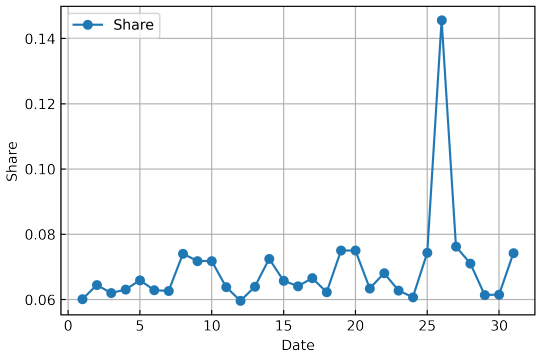

図 5　野球に関するトピックの割合の一日ごとの平均値

図 4　ツイートにおける 4 つのトピックのワードクラウド（他は APPENDIX を参照）

て $k = 11, 12, \ldots, 70$ の範囲でグリッドサーチを行った。トピック数ごとの Coherence の値の推移のグラフを**図 3** に示す。

ここでは、Coherence が最大となる 15 を採用した。抽出された 15 のトピックの内、鮮明な 4 つのワードクラウドを**図 4** に示す。八戸の観光地や交通に関する話題、スポーツ関連の話題、天気に関する話題などが含まれていると考えられる。

4. 八戸に関するツイートのソーシャルセンサーとしての活用可能性

3. までの結果に基づき、八戸に関するツイートのソーシャルセンサーとしての活用可能性について考察する。

ツイート数が 10 件以下のユーザ数は 13,224 人であり、ユーザ全体の 97% を占めていた。ユーザのプロフィール欄から観光や野球をはじめとする複数の鮮明なトピックが抽出できることが確認できてることから、対象とする期間においてどのような Twitter 上のユーザが八戸に関心を持っているか、概要を把握することができると考えられる。また、今回はトピックの抽出のみを行ったが、ユーザのプロフィール欄の各トピックの割合を要素とするベクトルを分散表現と見なす、もしくは Doc2Vec[8] などのニューラルネットワークを用いた言語モデル等を適用することでその分散表現を得ておくことにより、*k*-means 法などを用いてユーザのクラスタリングを行うことができる。クラスタリングの結果次第では、対象の期間におけるクラスタごとの興味・関心の傾向をより詳細に分類したり、クラスタ内においてリツイート数が多いユーザを抽出したりすることで、八戸に関心をもつユーザの傾向をより詳細に把握することができる可能性がある。関連研究の [4] では、分散表現の尺度を用いて各文書の類似度を計算することにより、各トピック内の文書集合の話題集約の粒度をサブトピック単位へと詳細化している。

ツイート本文には、八戸に関する観光や野球、サッカーなどの地域に関する話題が含まれていた。一日、一時間等の単位で、ツイートに含まれる各トピックの割合を見ることにより、Twitter において八戸のどのようなことが話題になっているか、概観を把握することができる。また、各トピックの割合の推移の時系列を扱うことで、八戸に関するツイートを行うユーザの関心の推移を確認することができると考え

られる。具体例として、**図 4** の右上の八戸の野球に関するトピックの各ツイートにおける割合の、一日ごとの平均値をプロットしたグラフを**図 5** に示す。10 月 26 日にこのトピックのシェアは他の日の 10 倍以上の値をとっており、八戸の野球に関する話題に対しユーザの関心が集まったと考えられる。実際、10 月 26 日は八戸学院大学の硬式野球部の選手 2 名がドラフトで指名を受けた日となっていた。このように、ツイートにおける各トピックの割合の時系列を元にした分析が考えられるが、例えば Kleinberg のアルゴリズムによりトピックのバーストを検出した研究 [6] の手法などを応用することで、八戸において急速に注目が高まったトピックを検出することなどが考えられる[6]。さらに、注目が高まったトピックの詳細を把握するために、Doc2Vec 等による分散表現を併用することも考えられる。Cos 類似度等を用いて該当トピックに関連が強いツイートの分散表現と似たツイートを抽出し、その中でもリツイート数が多いもの等を提示することで、該当のトピックで起こった出来事の把握が可能となるかもしれない。近年、転移学習を活用した深層学習による言語モデルである BERT[3] が自然言語処理の分野で注目を集めているが、これを応用することで Twitter における八戸の話題の要約文の生成ができる可能性もある。

5. おわりに

本論文では、「八戸」が含まれるツイートを行うユーザおよびその話題について分析し、そのソーシャルセンサーとしての活用について考察を行った。データの基本的な統計量を確認した上で、ユーザのプロフィール欄およびツイートの本文に対し潜在的ディリクレ配分法を適用し、両者のトピックを抽出した。ユーザのプロフィール欄においては、旅行、趣味、スポーツに関するものなど 24 のトピックが抽出され、ツイートの本文からは八戸に関する観光、天気、野球をはじめとする 15 のトピックが抽出された。抽出されたトピックを用い、八戸に関するツイートを行うユーザの概要や、八戸のツイートに含まれる話題について確認し、「八戸」を含むツイートの活用の方向性について考察した。

[6] 時系列を考える場合には、時間が経過するごとにトピックを特徴づけるキーワードが変化するという仮定も含めた Dynamic Topic Model [1] などを用いた方が良い結果が得られると考えられる。

APPENDIX

3. の**図 2** に記載した 4 つのユーザのプロフィール欄におけるトピック以外のもののワードクラウド、および**図 4** に記載した 4 つのツイート本文のトピック以外のもののワードクラウドを、それぞれ**図 A-1**、**図 A-2** に示す。

参考文献

[1] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113 – 120.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993 – 1022.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)*, 4171 – 4186.

[4] 丁易, 趙辰, 川畑修人, 宇津呂武仁, 河田容英. (2018). トピックモデル・分散表現の併用によるウェブ検索結果話題集約におけるサブトピック化. 第 10 回データ工学と情報マネジメントに関するフォーラム.

図 A-1　**図 2** を除くプロフィール欄のワードクラウド

図 A-2　**図 4** を除くツイート本文のワードクラウド

[5] Iwata, T., Watanabe, S., Yamada, T., & Ueda, N. (2009). Topic tracking model for analyzing consumer purchase behavior. In *Twenty-First International Joint Conference on Artificial Intelligence*, 1427 – 1432.

[6] Koike, D., Takahashi, Y., Utsuro, T., Yoshioka, M., & Kando, N. (2013). Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 917 – 921.

[7] Kudo, T. (n.d.). MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Retrieved December 1, 2020 from `https://taku910.github.io/mecab/`

[8] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, 1188 – 1196.

[9] 大島裕明, 中村聡史, 田中克己. (2007). Slothlib: web サーチ研究のためのプログラミングライブラリ. 日本データベース Letters, 6(1), 113 – 116.

[10] Řehůřek, R. (2020). GENSIM: Topic modelling for humans. Retrieved December 1, 2020 from https://radimrehurek.com/gensim/

[11] Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. Paper presented at *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

[12] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399 – 408.

[13] 榊剛史, 松尾豊. (2012). ソーシャルセンサとしての Twitter: ソーシャルセンサは物理センサを凌駕するか?. 人工知能, 27(1), 67 – 74.

[14] Twitter, Inc. (n.d.). Standard search API — Twitter Developers. Retrieved December 1, 2020 from https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

[15] Twitter Japan [TwitterJP]. (2017, Oct 27). いつも、そして何年もの間、Twitter をご利用いただきありがとうございます。おかげさまで日本での月間利用者数が 4500 万を超えました。安心してサービスをご利用いただけますように、一層の努力を行います。引き続きのご指導、ご支援のほど、よろしくお願い申し上げます。[Twitter moment]. Retrieved December 1, 2020 from https://twitter.com/TwitterJP/status/923671036758958080

要　旨

本論文では、「八戸」を含むツイートを収集し、そこに含まれる話題やユーザのプロフィールについて分析を行った上で、そのソーシャルセンサーとしての活用可能性について考察した。具体的には、Twitter API を用いて「八戸」を本文内に含むツイートを収集し、そこに含まれるユーザおよびツイートの数、頻度等のユーザに関する基本的な統計量を確認した。また、八戸に関するツイートを行ったユーザのユーザプロフィール欄のテキストおよびツイートの本文に対し、潜在的ディリクレ配分法を適用することで、ユーザのプロフィール欄やツイートに含まれるトピックを調べ、その活用について考察した。

キーワード： 八戸, *Twitter*, 潜在的ディリクレ配分法, ソーシャルセンサー, ニューラル言語モデル