

Twitter データを対象としたマーケティングに活用可能な データサイエンス分野のトレンドトピック抽出方式の実現方式

黒木 友瑛[†] 岡田 龍太郎[†] 峰松 彩子[†] 中西 崇文[†]

武蔵野大学 データサイエンス学部 データサイエンス学科[†]

1. はじめに

現在, SNS 上において, 様々なユーザが自身の行動や感じたことを投稿し続けている. これらの SNS 上に投稿されたデータはライフログと見なすことができ, ユーザそれぞれの日々の生活や感じたことの記録として散在している. これらのデータを対象として, ユーザの感じたことを直接的に分析する手法が確立されれば, 世の中で起こっている事象や, 展開される商品やサービスに対する使い方や評判, 感じ方を知ることが可能になると考えられる.

特にマーケティング分野において, SNS データを用いて, 顧客・ユーザの動向や評判を抽出し活用することが増えている. これまで, マーケティング分野の調査は, アンケートによるものが主流であったが, 多くのユーザを対象として随時調査をすることに関してコストが大きいことが問題となっていた. SNS データを用いることにより, 顧客・ユーザの日々の行動や感じ方の率直な反応を取得することが可能になると考えられる.

我々は, SNS データとして, Twitter のツイートデータに着目する. Twitter では, ユーザのライフログとなりうるツイートに対して, いいねやリツイートなどの他のユーザの反応がデータとして保持されている. これらは, 他のユーザの共感の度合いを示していると仮定する. 我々は, Twitter において, トレンドとなるトピックを見つけるためには, いいねやリツイートといった他のユーザの反応の大きさが重要になると考えている.

本稿では, Twitter データを対象としたツイートへの反応回数の差分によるトレンドトピック抽出方式について示す. 本方式は, 同一分野に属する複数のアカウントのツイートデータを対象として, ツイート毎のいいね数の差分からツイートに含まれる単語の重要度を推定すること

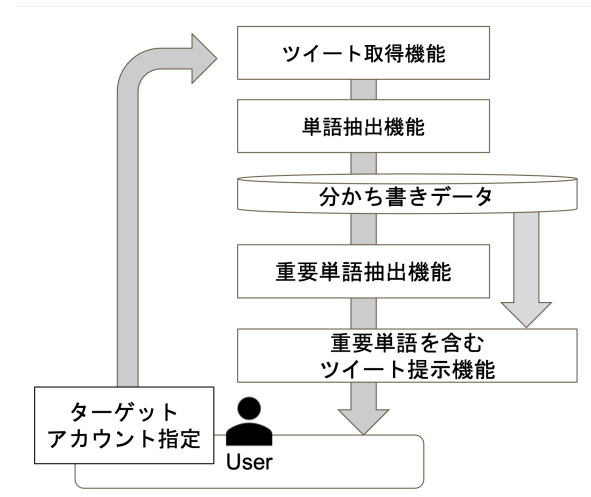


図1: 提案方式の全体図

により, その分野のトレンドトピックを抽出する. 本稿では, データサイエンス分野を対象として, 著名なアカウント群のツイートデータから, データサイエンス分野のトレンドを推定する. 本方式を実現することにより, データサイエンス分野を対象として, 着目されている話題, トレンドを把握することが可能となる.

2. 関連研究

Twitter におけるトレンドの予測を行う研究としては, ハッシュタグを用いるものが多数行われている[1][2].

田中ら[3]は, SNS における話題や流行の予測の手法として, ソーシャルグラフを用いて言葉の拡散を予測する手法を提案している.

本方式では, 分野を指定することでキーワードの候補を限定した上で, ツイートに付けられたいいねの数を世の中からの反応の基準として利用し, ツイートに含まれるキーワードに重み付けを行うことによってトレンドトピックを抽出する.

3. 提案方式

図1に提案方式の全体像を示す. 本方式は, ツイート取得機能, 単語抽出機能, 単語重み付け機能, 重要単語を含むツイート提示機能からなる.

Extraction Method of Trending Topics by Difference in Reaction Count to Tweets for Twitter Data

Tomoe Kuroki[†] Ryotaro Okada[†] Ayako Minematsu[†] Nakanishi Takafumi[†]

[†]Department of Data Science, Musashino University

3.1 ツイート取得機能

ツイート取得機能では、ユーザがターゲットアカウントを指定し、TwitterAPI を使用してアカウントのツイートを取得する。ここでは同一分野に属する複数のアカウントを指定することを想定している。

3.2 単語抽出機能

単語抽出機能では、取得したツイートから、MeCab を使用して、名詞・動詞・形容詞を抽出し、さらにその中から数詞・非自立語・接尾語を除去し、分かち書きにする。また、日本語のストップワードリストを使用して、分かち書きデータを作成する。

3.3 重要単語抽出機能

指定した一名のアカウントから抽出したツイート集合に対して単語抽出機能で分かち書きした文章を対象に YAKE! [3] を用いて重要単語抽出を行う。この際、抽出された重要単語には重要度として重みが付けられている。さらに、抽出された重要単語に対して、その単語を含むツイートを検索し、ツイートに付けられたいいねの数の合計を計算する。YAKE! によって付けられた重みといいねの数を乗算したものを、重要単語の重みとして設定する。

3.4 重要単語を含むツイート提示機能

3.3 節で抽出した重要単語とその重みを用いて、重みが大きい重要単語と、その単語が出現するツイートを出力する。

4. 実験

本研究では、データサイエンス分野を対象として、複数のアカウントのツイートデータを収集し、提案手法を適用して重要単語の抽出を行った。データサイエンス分野に関係するアカウントとして、有名なデータサイエンティストである「TJO」「マスクドアナライズ」「かめ」「やえり」の4名のtwitterアカウントからツイートを取得した。2022年1月6日23時に、各アカウントの最新200件のツイートを取得し、分析を行った。

上記4名について重要単語を抽出した結果を表1に示す。本実験では、重要単語の数は上位10件とした。

5. おわりに

本稿では、Twitter データを対象とした、キーワード抽出といいね数をかけ合わせた新たなトレンドトピック抽出方式を示した。本方式により、Twitter ユーザの反応を含んだ、ある分野のトレンドトピックを抽出することができる。

今後の課題としては、リツイートを考慮したトレンドトピック抽出手法の改良が挙げられる。

表1：データサイエンティスト4名のアカウントから抽出された重要単語

	TJO	マスクドアナライズ
1	真面目	タサイエンティスト
2	見える	エンジニア
3	タサイエンティスト	教科書
4	出来る	できる
5	わかる	オススメ
6	大きい	メタバ
7	取れる	ビジネスパ
8	モデル	わかる
9	できる	マスクドアナライズ
10	あけましておめでとうございます	入門書
	かめ	やえり
1	深層学習	目指す
2	学べる	タッチタイピング
3	タッチタイピング	深層学習
4	筋トレ	タサイエンティスト
5	ランニング	機械学習
6	超える	ドマップ
7	目指す	情報発信
8	わかる	学べる
9	コミュニティ	プログラミング
10	ドマップ	blog

参考文献

- [1] Z. Ma, S. Aixin, C. Gao. Will this# hash tag be popular tomorrow?, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012.
- [2] O. Tsur, R. Ari Rappoport, What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities, Proceedings of the fifth ACM international conference on Web search and data mining, ACM, 2012.
- [3] 田中勝也, 田島敬史. Twitter におけることばの流行予測. DEIM Forum 2017 D6-2, 2017.
- [4] R. Campos, V. Mangaravite, A. Pasquali, A. Mário, J. Nunes, A. Jatowt, YAKE! collection-independent automatic keyword extractor. In: European Conference on Information Retrieval. Springer, Cham, p. 806-810, 2018.