1T-08

Twitterにおいて関連付け可能なツイートをスレッド化するシステムの検討と開発

山崎　颯太†　　寺澤　卓也†
東京工科大学　メディア学部†

1. はじめに

Twitter[1]の機能の用法を独自に解釈して用いるユーザーがいる。例えば、タイムライン上に会話の一部になるであろうツイートを、リプライ機能を用いずにそのまま投稿するユーザーが挙げられる。この場合、ツイートはリプライなどの形で関連付けされていないため、その会話に加わっていなかったユーザーがそのツイートを見た際に、何に関して言及しているのかが不明瞭になる。これは、Twitterを情報収集ツールとして使用しているユーザーにとって、情報収集を阻害する問題点となる。

そこで、関連付けすることが可能だがされていないツイート群を、リプライで繋げられているような、1 つのスレッドのような形として扱うことが出来れば、他のユーザーから見たときにタイムライン上で混乱を招くことなく、正しい情報を時系列順に得ることが出来るのではないかと考えた。

この問題を解決するため、本研究ではツイートごとの投稿時間やツイート本文などの特色から、リプライなどの形で関連付けされていないツイート群が関連付け可能かを分析し、関連付け可能であればスレッドとして扱う方法を考案して、実装した。

2. ツイート関連付け手法

ツイート本文、ツイート自体の id 番号、アカウント id、投稿時刻、リプライの有無、投稿元のアプリケーションの URL の情報を、Twitter API を用いて取得した後にデータベースに格納する。次に格納したツイートを、データを基に分析し、関連付け可能か否かを判断する。最後に、関連付け可能なツイート群は、時系列順に 1 つのスレッドとしてまとめる。

ツイートを関連付けするにあたり、ツイート本文や周辺情報を分析する必要がある。そこで、関連研究[2][3]で行っていた連結手法を一部改変し、さらに本研究で新たに考案した手法を加えて、関連付けが可能かを分析する。

関連研究で用いていたツイートの連結手法のうち、共通の固有名詞、一般名詞、リツイートによる話題の差し込み、話題の転換や継続を表す語によるものは変更せず、ツイートの時間間隔については、ユーザーごとのツイートの頻度によって基準を変更する。

本研究で新たに追加する手法として、TF-IDF 法を用いて複数ツイート内の各単語の重要度を測る手法、コサイン類似度を用いてツイートどうしの類似度を測る手法の 2 つを挙げる。加えて、各ツイートに共通した一般動詞が含まれているかも関連付けの判断基準とする。

本研究では、コサイン類似度を求める際に用いるベクトルの成分を TF-IDF で求めた値に置き換えて、各ツイート間の類似度を求める。

3. ツイート関連付けシステムの実装

分析システムでは、会話の最初になるツイートを起点ツイートとし、起点ツイートと関連付ける可能性があるツイート群を関連付け候補ツイート群とする。関連付け候補ツイート群を分析システムで削っていき、最終的にはその内の 1 件のツイートを関連付けツイートとする。関連付け候補ツイートを検出しなくなった場合は、リプライの終了ツイートであるとみなす。

関連付け候補ツイート群は、以下の条件と順番に照合し、適合しないものは除外される。

- 起点ツイートの投稿日時から 1 日以内に投稿されていること
- ツイートの投稿元 URL が広告などの連携アプリケーションなどでないこと
- ツイート本文が URL のみでないこと

Design and implementation of the system that makes a thread of associable tweet on Twitter
†Ryuta Yamazaki, Takuya Terasawa
School of Media Science, Tokyo University of Technology

- 起点ツイートと、起点ツイート投稿ユーザーの投稿頻度から算出した3時間あたりのツイート投稿平均数までの投稿であること
- 共通する名詞及び一般動詞が含まれていること
- TF-IDF とコサイン類似度により算出された類似度が0.1以上であること（類似度0.1以上の関連付け候補ツイートが無ければ類似度の平均値が0.03以上かつ平均値より高いものとする）

また、下記のような場合は特例として一部の処理を無視し、関連付け候補ツイートとする。

- リツイート直後の同一ユーザーのよるツイート
- ツイート本文の先頭に接続詞が含まれているツイート
- 「わかる」「りょ」など Twitter で頻繁に扱われる単語が含まれているツイート

4. 評価

実際にタイムラインに単体で存在する関連付け可能なツイートは、関連付け可能か否かの判断が困難であるため、本研究では、評価対象を、実際にリプライの形で繋がっているツイート群とする。それらを分解した後に、リプライが始まるツイートから分析を開始し、リプライで繋がったツイートをスレッドとして再現できるか、という手法で評価を行う。

リプライの再現率 R は、T を 1 つのリプライで繋がったツイート群のツイート数、C を連続で関連付けが成功した箇所のツイート数とするとき、以下の式で表す。

$$\mathrm{R} = \sqrt{\frac{1}{T^2}\{(C_1-1)^2+(C_2-1)^2+\cdots+(C_n-1)^2\}}$$

$$0 \le \mathrm{R} \le \left(\frac{T-1}{T}\right)^2$$

分析システムで関連付けを行った結果は以下のとおりである。

リプライで繋がった対象ツイート群	46
リプライをすべて再現したツイート群	0
リプライを一部再現できたツイート群	32
再現箇所が無かったツイート群	14
平均再現率	0.159

ほとんどのツイートにおいて関連付けに失敗した。

関連付けができたツイートには、比較的単純な日本語でやりとりが行われ、共通する名詞・一般動詞などが明確であるなどの特徴があった。

関連付けができなかったツイートには、リプライが終了したツイートの後に余分にツイートを関連付けしたケース、リプライとなる一部のツイートを無視して関連付けを行ったケース、リプライの中間地点でリプライとは異なるツイートを関連付けしたケースが多く見られた。

原因は、特例のツイートを除き関連付け候補ツイートの選択基準が同じであるからと考える。評価対象のリプライはほぼ1対1で行われていることが確認出来た。よって、最初の起点ツイートに関連付けされたツイート投稿ユーザーの、その後のツイートが関連付け候補ツイート群に残る条件を時間間隔のみにすると、上記の無視するケース、異なるツイートを関連付けるケースは避けることが出来ると考える。

5. おわりに

ツイートの関連付けをするにあたり、1 つの起点ツイートを定めた後に、起点ツイートのツイート本文及び投稿時刻などの周辺情報を関連付けられる可能性があるツイートと比較した。比較する際に用いた条件は、一定の時間内に投稿されている、ツイート本文の文頭に接続詞がある、ツイート本文中に共通の名詞・一般動詞がある、TF-IDF 法とコサイン類似度を用いて算出した類似度が一定の値以上であることとした。

結果は大半の関連付けに失敗し、失敗したものは3つのケースに大別された。各ケースに対応した分析システムに修正することで、再現率を高めることが出来る可能性がある。

今後はそれらの修正を行った上で、再度検証を行う必要がある。

参考文献

[1] Twitter: http://twitter.com/
[2] 星皓介、山田剛一、絹川博之、“Twitter における話題のツイートの連結と話題抽出”、情報科学技術フォーラム講演論文集、pp.159-160、2013
[3] 星皓介、山田剛一、絹川博之、“Twitter タイムラインからの話題の抽出とその評価”、情報処理学会　第 76 回全国大会講演論文集、pp.499-500、2014