D-032

Twitter における同一話題のツイートの連結と話題抽出

Topic Detection Based on Concatenation of Tweets on Same Topic

星 皓介† 山田 剛一† 絹川 博之†
Kosuke Hoshi Koichi Yamada Hiroshi Kinukawa

1. はじめに

近年，ソーシャルメディアサービスの発展により人々の情報発信をする場が急速に増えてきている．特に，そのひとつである Twitter[1]が大きな成長を見せている．

Twitter のサービスの特徴の 1 つに，タイムラインと呼ばれる，自身の発言およびフォローしているユーザの発言が表示される場の存在がある．興味のあるユーザをフォローすることで，ユーザ独自のタイムライン(ホームタイムライン)を作ることができる．

タイムラインは積極的に情報を取得することができ，有用であると考えられる．しかしながら，ユーザの興味の拡がりとフォローするユーザの増加に伴い，タイムラインには多様な情報が現れるようになる．同時に，一般的な話題を表す語の割合が高くなり，現れる語の多くは，ユーザの興味から離れたものとなっている．

そこで我々は，一般的な話題ではなくホームタイムラインに固有な話題に着目し，ユーザにとって価値のある話題を抽出できるか調査検討を行なってきた[2].

しかしながら，話題の抽出を行う際，ツイートあたりの情報量が少なく，タイムライン上の話題が集約されない問題が現れた．この問題を解決するには，ツイートの連結を行う必要があると考えた.

本論文では，ホームタイムラインではなく，ユーザの発言(ユーザタイムライン)を対象とした同一話題のツイートの連結を行ういくつかの手法を提案する.

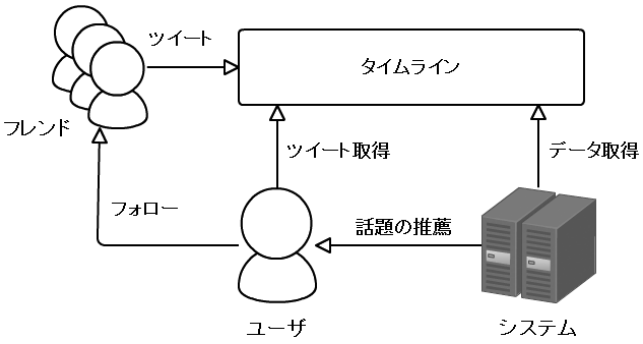

図 1 システムフロー図

2.話題抽出

話題とその抽出について説明する.

話題は話の内容を表すため，名詞で表現される．本研究では名詞の抽出がベースとなる．特に名詞の中でも，より具体的な表現である語ほど話題としてふさわしいものであるとする．これは，抽象的な語である場合，別により具体的な話題を表す語が存在することが多いためである.

Twitter には，ハッシュタグという話題を明示的に表す機能がある．これは複数のユーザで話題を共有するために使われることが多いが，個人が複数回ハッシュタグを付けて発言することも多いため，ユーザタイムラインにおいても話題をまとめるのに利用できると考えられる.

これらを基に話題抽出を行った場合，次の問題が発生した.

†東京電機大学大学院　未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

3. 話題抽出時のツイート分割問題

ツイートの内容が分割されると，話題が正しく抽出できないことがある．例えば，話題の主題を表す言葉を含むツイートと話題の詳細を表す言葉を含むツイートに分かれるような場合が挙げられる．これらは話題を抽出する際に問題があると言える.

ツイートが分かれて現れるのは，2 つのケース存在する．ひとつは，ユーザの書きたい内容が 140 字を超えるとき．もうひとつは，ユーザが思いついたことをすぐに投稿し，あとから訂正，追記をするときである．Twitter の特性上，後者のケースが多く存在する.

以上を踏まえ連結手法の提案を行う.

4.連結手法の提案

ここでは，同じ話題のツイートの連結手法をいくつか提案する．基本的に連結はユーザタイムラインを対象とする．また，連結を行うツイートは連続している必要はない.

4.1 ツイートの時間間隔

ツイートの時間間隔が短いと，ユーザが同じ話題を発言している可能性が高いと考えられる.

例.

「急須でいれたお茶が飲みたい」

(2013-04-13 02:29:14)

「ふけたな」

(2013-04-13 02:30:47)

しかし、実際には短い時間間隔においても話題の転換が起きることがある。

これには 2 種類の原因が考えられる。

1) ユーザの頭の切り替えが早い。
2) 複数のツイートの内容があらかじめ準備されている。

1) は通常よりも移り気なユーザであり、前の話題を打ち切った直後に次の話題のツイートを行う。

例.

「R ときいて、真っ先に統計処理が思い浮かんだ」

(2013-07-01 01:41:08)

「定期的にクリーンインストールすると、必要なやつと必要じゃないやつがよくわかっていいですよね。（以下略）」

(2013-07-01 01:41:56)

2) はあらかじめ準備しておいた内容を連続してツイートするケースである。一部のユーザは Twitter が利用できない間に複数の話題を投稿しようと考え，利用可能となった時に一度に複数の話題を投稿するのである．

また，ユーザによってツイートの間隔は様々であり，時間間隔をひとつに定め，連結を行うのは難しいと言える．投稿する文字数によっては，投稿に時間がかかることも考えられる．

このように，ユーザによるツイートの行動の違いが見られたため，ユーザをいくつかに分類する必要がある．

4.2 会話

日常の会話と違い，Twitter での会話はある 1 つの事柄について話すことが多く，話題が転換することはあまりない．そのため，会話の一連のやり取りをひとつにまとめることが可能であると考えられる．

また，会話からは，前項で紹介した時間間隔を用いる連結手法と比較し，他のユーザの発言ごとまとめることができるためタイムラインの話題をまとめるのに有用である．

しかし，会話中に別の話題に替わることがないとは言えない．その場合には，次の手法を用いる．

4.3 話題の転換や継続を表す語

日本語には話題の転換や継続を表す語が存在し，これらはツイートの連結における手がかりとなる．例えば以下のようなものである．

話題の転換を表す語：
「そういえば」「さて」
話題の継続を表す語：
「でも」「あと」「そして」

これらは Twitter 上でもよく見られる．以下に例を示す．

例.

「サッカーでは絶対勝てないからねえ」

(2013-03-02 22:10:09)

「**でも**ブラジルは親日なイメージだから好き セナ然りカカ然り」

(2013-03-02 22:11:15)

しかし，注意すべき点も存在する．それは，あるユーザの発言以外に対してこれらの語が用いられた時である．例えば，ホームタイムライン上のツイートに対してリツイートなしで話題の継続を表す語が使われることが考えられる．

4.4 共通の固有名詞

ツイートを比較する時，共通する固有名詞が含まれていると連結可能なことが多い．ただし，固有名詞でも一般的によく使われる語は連結できるとは限らない．例えば，地域名がそれに当たる．地域名は非常に多く用いられ，また，何度も使われることが多いため，特別に扱う必要がある．

例.

「スライドファクトリー始まりました！**シュルツ**先生の基礎レッスン。誰でも参加できるよ！」

(2013-03-29 17:32:36)

「**シュルツ**先生、質問受け付けてます。これ、積極的に前のポジション取るべき！」

(2013-03-29 17:38:16)

「**シュルツ**先生のマスタークラスの聴講、始まります。」

(2013-03-29 18:04:19)

4.5 考察

それぞれの手法は単独で連結できるとは限らず，いくつかの手法を組み合わせて利用する必要がある．特に時間は大きな手がかりとなる．また，ユーザごとの Twitter の利用方法は様々であるので，ユーザをいくつもの観点から分類をすることが重要であると考えられる．

5. おわりに

本論文では，同一話題のツイートの連結手法をいくつか提案した．今後，提案手法の詳細な分析と，手法の組み合わせを考慮した連結を行う．また，タイムライン上の話題の抽出に有効であるかを調査する．

参考文献

[1] Twitter：http://twitter.com/
[2] 星皓介, 山田剛一, 絹川博之, “Twitter におけるタイムライン固有の話題の抽出”, 情報科学技術フォーラム(2012)