D-010

Twitter におけるタイムライン固有の話題の抽出

Extraction of Home Timeline Topics Different from Public Timeline Topics on Twitter

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が情報発信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 が大きな成長を見せている．

Twitter のサービスの特徴として，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．

これらの問題の解決方法として，タイムラインにおける重要な語を提示することが考えられる．システムの流れは，タイムラインから話題語を抽出し，ユーザにとって重要な語の推薦を行うものである．（図 1）

本論文では，その一部である話題の抽出，特に，ユーザ固有の話題に着目し，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う．

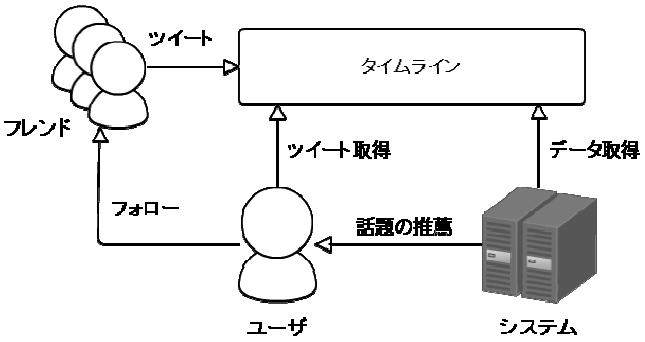

図 1 システムフロー図

2. Twitter における話題

2.1 タイムラインにおける話題の特徴

タイムラインにおける話題には，単一文章の話題とは異なる以下の特徴がある．

- 発言者が複数なので，多数の話題が混在する．
- 1 ユーザの発言に限っても，1 つの話題の持続期間が短く，また突然話題が変わる．

細切れの話題を扱う必要があるため，話題の一貫性，結束性に基づいて話題抽出をすることは難しい．語自体の話題になりやすさといった，語の特性が重要となる．

2.2 話題となる語の特性

まず，話題は名詞で表現されることが多い．また，抽象的な名詞よりも，具体的な名詞が話題として認識されやすい．例を挙げると「食べ物」のように曖昧なものより，「くだもの」更には「りんご」と言った語のほうが，より具体的になり話題として認識しやすい．このような語がツイート中に同時に出現した場合，より抽象度の低い語を優先し話題とみなすことが必要と考えられる．また，複合語は複合してより具体的な概念を表すものであるため，複合語は全体として 1 つの話題と認識されやすい．

タイムラインでは，発言者間のつながりのない複数の発言者が，同一の話題についてツイートすることがある．このような場合，語としては異なるものが用いられることがあるため，同義語を同義として扱う仕組み，略語と元の語を同義として扱う仕組み，表記の揺れを扱う仕組みが必要となる．

また Twitter には，ハッシュタグと呼ばれるツイートにタグ付けを行う機能がある．これは，複数のユーザ共通の話題について発言を行う場合に用いられる．そのため，話題として扱うことができると考えられる．

3. ユーザ固有の話題抽出

Twitter 全体で流行っている話題は，Web 上における一般的なニュースとしても取り上げられることが多い．そのため，積極的に情報を集めるユーザにとっては，情報の価値が低いと考えられる．また，ユーザは特徴的な話題を求めてフォローすることも多いため，ユーザ固有の話題は価値のある情報であるといえる．

3.1 固有な話題

固有の話題とは，日常的な話題や抽象度の低い話題を指す．また，ツイートの主題となり得るかでも判断する．

3.2 ユーザの発言分類

ユーザの発言に含まれる語は，特徴的なものと一般的なものに分けられる．タイムラインはユーザの発言の集合とも言え，その集合の拡大により一般的な語が多く現れるようになる．そのため，一般的な語が特徴的な語を隠すノイズとなっている．

†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

3.3 Twitter 全体の話題

固有の話題の抽出を行うにあたり，Twitter 全体の話題をパブリックタイムラインと呼ばれる，Twitter すべての公開ユーザの発言が現れる場を用いる．ユーザのホームタイムラインから，全体で多く出現する語を比較することで，固有の話題が抽出できると考えられる．

4.固有話題抽出実験

4.1 実験環境・対象データ

パブリックタイムラインとホームタイムラインの抽出した語の違いから，固有な語を抽出可能か検証する．また，話題ではないような不用語が除けるのかも検証する．

検証に用いるデータは，ユーザ 3 人のホームタイムラインから 12 時間分のツイートを集め，同様の時間範囲でパブリックタイムラインからも集める．Twitter には，多くのボットが存在するが，一定に短い期間でツイートをしたり，定型の文を多くツイートをしたりするため，予めデータから除いておく．

まず，ホームタイムライン，パブリックタイムラインの双方からすべての名詞を取り出す．ただし，複合語は分割せず複合語のまま取り出す．

ホームタイムラインの語は以下に分類される．

1) 話題語
   1a) 固有語：ホームタイムライン固有の話題を表す語
   1b) 固有でない話題語：Twitter 全体の話題を表す語
2) 不用語：通常話題になり得ない一般的な語

パブリックタイムラインには 1a) の固有語が含まれていないので，ホームタイムラインの語からパブリックタイムラインの語を除けば，1a) の固有語が得られるはずである．

ホームタイムラインの語から，パブリックタイムライン上で出現回数の多い語上位 50，100，200，500 件を除く．残った語のうち，出現回数 3 回以上の語のものを対象として，話題語の占める割合，固有語の占める割合をそれぞれ求めた．また，不用語をあらかじめ除いておいた場合の固有語の割合も求めた．

結果を表 1，表 2 に示す．ホームタイムラインから除くパブリックタイムラインの語を増加させると，話題語，固有語，不用語を除いた場合の固有語，すべてにおいて出現割合が高くなった．

表 1 比較データにおける話題出現数の割合

	比較なし	50	100	200	500
話題語	0.690	0.734	0.747	0.857	0.874
固有語	0.343	0.419	0.455	0.502	0.581
不用語除いた場合の固有語	0.495	0.568	0.606	0.607	0.665

表 2 比較データにおける話題出現数の再現率

	比較なし	50	100	200	500
話題語	1.00	0.828	0.741	0.637	0.503
固有語	1.00	0.935	0.903	0.806	0.677

4.2 考察

実験より，提案手法によって固有語の割合が高まることを確認した．また，話題語の割合が高まることから，不用語の除去に効果があることも確認できた．パブリックタイムラインに多く現れる語を除くほど，精度は上昇していると分かるが，再現率も大きく低下している．話題語はパブリックタイムラインにも多く現れるため，差分を除いた場合に話題語の出現回数が減少するのは特に問題にはならない．固有語が減少する理由は，固有語の定義に問題があるのではないかと考えられる．また，固有な話題であると認識していた語が，一般でも日常的に現れているのではないかと思われる．固有の話題を人に依存しない形で定義することで，解決を図りたい．

5. おわりに

本論文では，ユーザに固有な語に着目し，提案手法が，ホームタイムラインの固有語の割合を高め，不用語の割合を低下させることを確認した．今後，精度と再現率の向上を行う．

参考文献

[1] Twitter：http://twitter.com/