

ニュース記事に対するツイートの可視化システムの提案

水澤 千尋[†] 白川 真一^{††} 大原 剛三^{††}

[†] 青山学院大学大学院理工学研究科

^{††} 青山学院大学理工学部

1. はじめに

近年、マイクロブログと呼ばれる Twitter から発信されたツイートの分類に関連する研究が進んでいる。例えば、ツイートがある特定のコミュニティ内のみでしか意味を持たないツイートか他のユーザにも有用な情報であるかの判別[1]や、ニュース、イベント、意見、掘り出し物、プライベートメッセージなどツイートの種類の分類などが行われている[2]。

これに対して本研究では、特定のトピックに対してどのような意見がツイートされているのかを俯瞰できるシステムの実現を目指す。具体的には、ツイートボタン[3]などを利用してツイートされた URL を含むツイートのテキスト情報に着目した。URL が記述されていることで、そのツイートがどのような話題に対して言及しているかがわかる。今回はニュース記事に対するツイートを対象とした。本稿では、ユーザがある特定の記事に対してコメントしているツイートから、コメントのみを自動で抽出し、その分類方法と可視化システムについて検討を行う。

2. 提案システム

提案システムについては、コメントの抽出、ツイートから抽出したコメントの分類方法、インタフェース部に分けて概説する。

2.1. コメントの抽出

本研究で扱うツイートは、図 1 で示すような形式のツイートとなる。ツイートボタンによって、あらかじめ用意されているそのウェブサイト名やニューストップック、URL などが記述されている。これらに加えて、ユーザのコメントを含むツイートが対象となる。コメント以外の情報を除去することで、コメントを抽出し、それを対象データとする。

2.2. コメントのクラスタリング

抽出したコメントに対して形態素解析を行い、名詞(一般名詞, 固有名詞), 動詞, 形容詞, 感動詞をキーワードとして抽出する。なお, 活用のある品詞については原形に戻す。式(1)に示す Jaccard 距離を用いて, 一つのコメントに含まれるキーワード群 S を別のコメントのキーワード群 T と相対的に比較する。これをすべてのコメントの組み合わせに対して行い距離行列を得る。

$$JD(S, T) = 1 - \frac{|S \cap T|}{|S \cup T|} \quad (1)$$

次に, 得られた距離行列を用いてコメントをクラスタリングする。さらに, 各クラスタ内のコメント間の Jaccard 距離を算出し, クラスタ内の他のコメントとの平均距離が最小

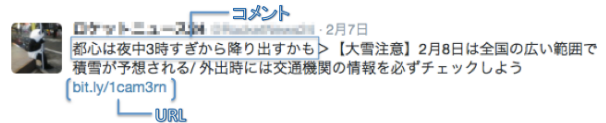


図1 対象とするツイート形式の例

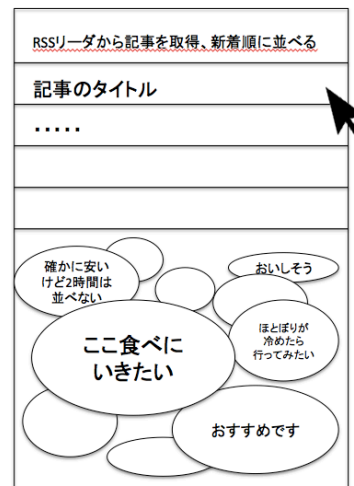


図2 提案システムのインタフェースイメージ図

のコメントをそのクラスタの代表コメントとして決定する。

2.3. インタフェース

図 2 は提案システムのインタフェースのイメージ図である。具体的には, RSSリーダのように登録した Web サイトに対して, 新着記事のタイトルと記事に対するコメントや評判などを視覚的に提示することで, ユーザがさまざまなトピックのニュースにアクセスするきっかけとなるようなシステムを目指す。

3. 今後の課題

今後は, 各種の既存クラスタリング手法を適用した場合の詳細な性能評価, ならびにより効果的な可視化方法を検討していく予定である。

参考文献

- [1] 田中 淳史, 田島 敬史, “twitter のツイートに関する分類手法の提案”, DEIM Forum, A5-4, 2010.
- [2] B. Sriram, D. Fuhry, and M. Demirbas, “Short text classification in twitter to improve information filtering,” Proceedings of 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.841-842, 2010.
- [3] Twitter, “Tweet Button”, <https://dev.twitter.com/docs/tweet-button>