Twitter 上の話題の元となったツイートを発見する手法に関する研究

豊田　万葉

Twitter では拡散や推薦によって大量の情報が流れてくる。そのため、Twitter の中で話題が盛り上がることがあり、1 つの話題を完全な形で見るのが難しくなる。例えば、キーワード X について最初にされたツイートをツイート A、それについてリプライや引用リツイートをせずに意見や感想を述べたツイートをツイート B とする。ツイート B にあたるツイートが大量になされると、ツイート B のみを見た人は、なぜキーワード X についてのツイートがいくつも投稿されているのか、つまり話題になっているのかを疑問に思う。このような疑問は、キーワード X が話題になっている理由を突き止めることができれば解決できると考える。

そこで本研究では、あるツイートが原因で Twitter の中で話題が盛り上がる場合において、話題の元となったツイートを発見するための手法を提案する。ここで、本研究でいう話題の元となったツイートとは、何らかのトピックが話題になったとき、その話題が盛り上がる原因となったツイートのことを指す。このような研究をする意義としては、個人の情報探索や企業のマーケティングを支援することや、バーストに関する研究領域に役立つことが期待できる。本研究では、ニュースや天気のような Twitter の外の世界が原因である話題は対象外とした。なぜなら、このような場合は人手で話題の元となったツイートを発見するのが容易だからである。

提案手法として、まず特定のトピックに関するツイート群を収集し、その中から Kleinberg の手法でバーストを検出し、バーストした日時より n 日前を対象として、リツイート数が多いツイート上位 5 個のうち最も投稿日時が早いものを話題の元となったツイートとして出力した。話題の元となったツイートを発見する際には、細部に変更を加えて、シンプルな手法 A、いいねやリツイートの数を重視する手法 B、バーストした日時前後に着目する手法 C、バーストした日時の情報を用いない手法 D の 4 つの手法を提案した。

手法を評価する際には、本研究のために作成した独自のデータセットを用いた。TwitterAPI を使って作成した。システムの評価のために、人力で判断した正誤ラベルを付した。このようなデータセットを 20 件のトピックについて作成した。

4 つの手法のうち最も正解率が高かったのは、n の値を n=3 と大きくし、いいねとリツイートの数が多いツイートを優先的に出力する手法 B で、20 件のトピックのうち 16 件で正解できた。それ以外の手法については、対象とする期間が長い方が正答率が高くなる傾向があった。

今後の課題として、表記揺れに対応すること、本研究で対象とするような事象を容易に見つける方法を確立すること、より正確にバーストを検出できるように改善を行うことが挙げられる。

（指導教員　高久雅生）