

閲覧期間を考慮した Twitter 上の意外な情報の抽出

Extracting of Unaware Information from Tweets Based on Browsing Time

大原 啓詳† 鈴木 優‡ 灘本 明代††
 Hiromitsu Ohara Yu Suzuki Akiyo Nadamoto

1. はじめに

近年, Twitter に代表されるマイクロブログの普及により, 誰もが手軽に情報の収集や発信を行うことが可能となった. 特に代表的なマイクロブログの一つである Twitter には膨大な量の情報が日常的に投稿されている.

Twitter では閲覧者が, あるユーザをフォローすることにより, そのユーザのツイートを自身のタイムライン上で管理することができる. この時多くのユーザをフォローしていた場合, タイムライン上のツイートが増加し, 管理が困難となることが考えられる. また, ユーザが常にタイムラインを閲覧するという事は困難であり, フォローしているユーザのツイートを見落とす期間が発生することも考えられる. そのため, 膨大な量のツイートの中から有益なツイートを抽出し閲覧者に提示することが必要であると考えられる.

有益なツイートに含まれる情報は, 日常的に繰り返されるようなありふれたツイートではなく, 閲覧者にとってこれまで知らなかった目新しい情報を含むツイートや, 興味のある事柄についての何らかの変化に関する情報を含むようなツイートであると考えられる. このような情報の目新しさなどに関する有益さは, 閲覧者の過去の経験や, 周囲からの伝聞, 文書の閲覧, 既知の情報に基づく推測や予測等, 閲覧者の予備知識に依存すると考えられる. そこで本論文では特にツイートの閲覧期間と, ツイートに含まれる話題に着目し, 有益なツイートの抽出手法の提案を行う.

ここで, Twitter における情報の見落としやすさに影響する要素として, 情報の発信者の社会における認知度や発言の影響力の大きさなどが考えられる. 具体的には企業やイベント運営によるプロモーションアカウントやニュースサイトのアカウント, 有名人のアカウントなどは, 多くのユーザにフォローされており, そのツイートがリツイートなどで拡散される機会が多い. また一般人ユーザの中にも, キュレーターと呼ばれる非常に多くのユーザにフォローされている影響力の強いユーザが存在する. このようなユーザのツイートは, 拡散されやすく, たとえあるユーザがそのツイートを見落とすとしても, リツイートなどで拡散されることにより情報が埋没する可能性は低い. 一方で, 少数の共通の趣味を持つユーザにのみフォローされているような一般人ユーザについては, フォロワーにツイートが見落とされてしまうと, 有益な情報を含んでいた場合でも, 拡散されずにそのままそのツイートが埋没してしまうことが考えられる. そこで本論文では特に後者のような一般人ユーザによる投稿を対象とし有益なツイートを抽出する手法の提案を行う.

本論文では, 閲覧者が見落としたツイート中に含まれる, ある一般人情報発信ユーザの投稿における, 新鮮味のある情報や変化のあった情報を意外な情報と定義し抽出を行う.

2. 関連研究

本研究と同様に, 閲覧者の有益な情報の見落としの解消を目的とした研究は行われている. 辻ら[1]は, 閲覧者のタイムライン上の頻出語から興味を推定し, 新着ツイート中に含まれる閲覧者にとって興味のあるような話題のツイートの抽出を行っている. 辻らは, フォローしているユーザ全体を対象として興味分析とツイートの部類を行っているが, 本研究では特定の興味のあるユーザのみを対象として有益な情報の抽出を試みている点が異なる.

Ren ら[2]は, 時系列変化とユーザ間のつながりに基づくトピックモデルを提案しトピック抽出を行い, 有益と判断されたツイートに対し Wikipedia の概要部分を利用し要約と説明を行っている. 本研究では要約ではなく, 新規性を持った変化や, 既知の情報に対しての差分情報を取得することを目的としているため異なる.

またツイートの話題の分類に関して西田ら[3]は, ツイートのデータ圧縮のしやすさに基づき, あるツイートが着目した話題に関するツイートかどうかの分類を行っている. 西田らの研究では何らかの着目した話題に対しての分類であるのに対し, 本研究では, Twitter を利用する一般ユーザの多面性にも着目し, 情報発信者のツイート全体に含まれる複数の話題へと分類しようと試みている点が異なる.

3. 提示する情報の分類

本研究で抽出の対象とする意外な情報は, 閲覧者が知らなかった新しい話題に関する情報と, 既知の情報に対して変化のあった差分情報の両方を対象としている. しかし, 閲覧者が何らかの話題に興味を持ち情報の発信者をフォローしているということを考えると, 前者のように全く新しい話題については閲覧者が興味を持つかどうかの判断は困難である. たとえば, ある情報発信者について, その発信者の投稿した「野球」の話題に関するツイートに興味を持った閲覧者がその発信者をフォローしたとする. ここで閲覧者が見落とした期間中のみ, 発信者が「サッカー」の話題に関するツイートを投稿していたとする. この時, この「サッカー」の話題に関するツイートに閲覧者が興味を持つか持たないかの判断を Twitter 上の情報のみから行うことは困難である.

一方, 既知の情報に対する差分情報は, 閲覧者が閲覧していた期間と, その期間に情報発信者がどのような情報を発信していたかにより, 抽出する意外な情報の粒度が異なると考えられる. たとえば図 1 のように「野球」の話題についてツイートを投稿している情報発信者について, あるユーザ U_1 が閲覧していた期間においては「球団 A」「球団 B」といった話題に関するツイートが投稿されていたとする. ここで, U_1 が見落とした期間中に, これまでの閲覧期間には出現しなかった「球団 C」という話題に関するツ

† 甲南大学大学院自然科学研究科 Konan University

‡ 名古屋大学大学院情報科学研究科 Nagoya University

†† 甲南大学知能情報学部 Konan University

ツイートがあったとする。ここで、 U_1 に対しては「球団C」に関するツイート全体を差分情報として提示する必要がある。一方で同じ情報発信者をフォローしているユーザ U_2 が閲覧していた期間においては、「球団A」「球団B」「球団C」のいずれの話題についても、ツイートが投稿されていたとする。ここで、 U_2 が見落とした期間には「球団C」に関する話題ではあるが、より詳細な「選手」や「試合」といった話題について、 U_2 の閲覧期間中に出現しなかった新しい情報や、変化のあった情報が存在したとする。この時、 U_2 には「球団C」に関する、より詳細な差分情報を提示する必要があると考えられる。

これらのことから、各話題、および提示する情報の粒度について、提示方法を考慮する必要があると考えられる。

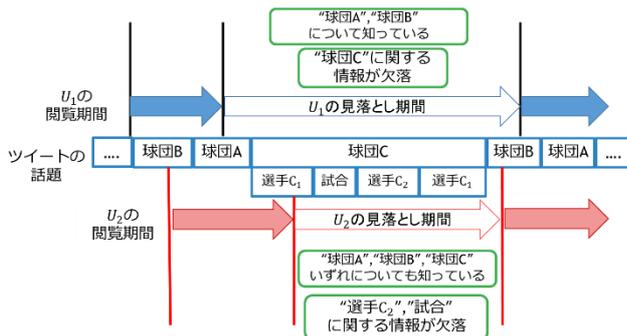


図1. 閲覧者ごとの提示すべき情報の粒度の違い

4. 抽出手法と提示方法

我々の提案する意外な情報の抽出手法について説明する。ここで、ツイートのもつ話題をトピック、あるトピック中におけるさらに詳細な話題（“野球”というトピックにおける“球団名”，“選手名”など）をサブトピックとして本研究では扱うものとする。

以下 (1) ~ (7) と図2に、意外な情報の抽出から提示までの流れを示す。

- (1) 情報発信者のツイートを収集する。
- (2) 収集されたツイート群に対し、ツイートを話題ごとにクラスタリングし、クラスター毎のトピックを抽出する。
- (3) 閲覧者の入力した閲覧期間におけるツイートと、見落としたツイートに (1) で収集したツイートを分類する。
- (4) 分類したツイートのトピックを比較し、見落とした期間中のツイートのトピックを、既知のトピックと、新規トピックとに分類する。
- (5) 新規トピックについてはトピックそのものを選択肢として表示する。
- (6) 既知のトピックについては閲覧期間外の変化や新規性のあるサブトピックを選択肢として表示する。
- (7) 閲覧者が各選択肢を選択した際、詳細なツイートを提示する。

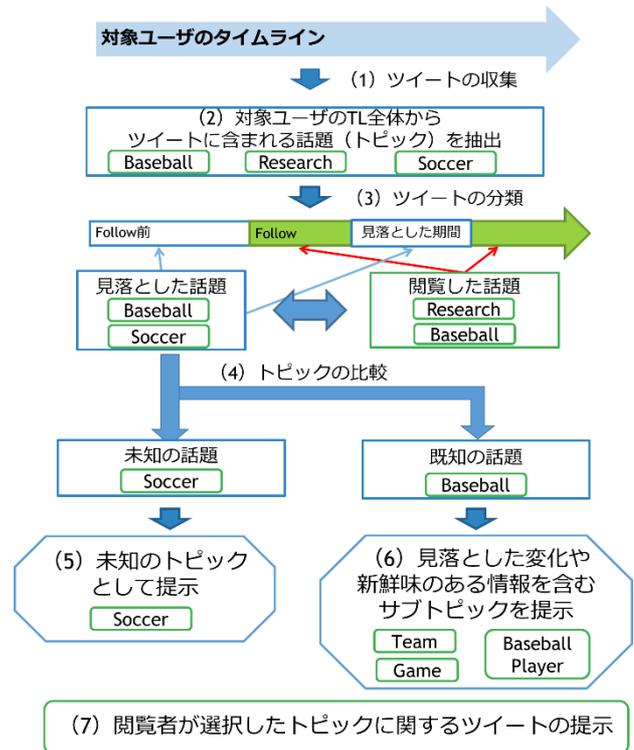


図2. システムの全体図

5. まとめと今後の課題

本論文では、閲覧者の閲覧期間に基づく、情報発信者の意外な情報を抽出する必要性について示し、意外な情報の抽出システムの提案とその概要を示した。

今後は実装に向けてどのようなクラスタリング手法が、ツイートを話題ごとに分類する際に妥当であるかの検討、ツイートのトピックの構造の分析、そして閲覧者にとって有効な表示方法の検討などを行う予定である

6. 参考文献

- [1] 辻 一明, 宝珍 輝尚, 野宮 浩揮. “新着ツイート群からの興味をひくツイートの抽出に関する考察”. 情報処理学会関西支部平成23年度支部大会, C-7, 2011.
- [2] Zhaochun Ren, Shangsong Liang, Edgar Meij, Maarten de Rijke. “Personalized Time-Aware Tweets Summarization”. In SIGIR2013, 2013.
- [3] 西田京介, 坂野遼平, 藤村考, 星出高秀. “データ圧縮によるTwitterのツイート話題分類”. DEIM Forum 2011, A1-6, 2011