3E4-1

マイクロブログにおける同意・反論関係を用いた情報信頼度推定

Assessing Information Credibility Based on
Agreement and Rebuttal Relationship Between Microblog Posts

佐藤雅宏*1　岡崎直観*1*2　乾健太郎*1
Masahiro Sato　Naoaki Okazaki　Kentaro Inui

*1東北大学 大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University

*2科学技術振興機構 さきがけ
Japan Science and Technology Agency (JST)

In the past few years, many people use microbrogs (e.g. twitter) to spread a wide variety of information. On the other hand, because the spreaded information is not always true, it would be helpful to develop a system that assists identifying reliable information. In this paper, we show how PageTrust algorithm, an extension of PageRank algorithm with negative edges for credibility assessment, can be used to estimate the credibility of posts in microblogs. In addition, we propose a method to create a graph representing semantic relation (i.e. agreement or rebuttal) between two posts in microblogs. We also evaluate the proposed method on a real dataset crawled Twitter.

1. 序論

近年、マイクロブログの普及に伴い、新聞やニュースの代わりにマイクロブログに投稿された文章から情報を得る人が増加している。マイクロブログとは Twitter や mixi のボイスなどのことで、誰でも気軽に投稿できるため大量の情報が存在する。しかし、その情報の真偽は不明なものが多いため、大量の情報の中から正しい情報を見極める必要がある。このように多くの情報の中から正しい情報を自動推定する手法としては、L. Page らの PageRank[1] や、Kleinberg らの HITS[3] などが有名である。これらは Web ページのリンク関係をグラフとして考え、それをグラフ解析することで各 Web ページの信頼度を求めるアルゴリズムであるが、リンクの有無だけで関係を判断するため、その Web ページがリンク先に対して肯定的な立場なのか、否定的な立場なのかを反映できない。つまり、これらのアルゴリズムではマイクロブログで見られる、ある投稿内容に対する反論意見を反映させることが出来ない。

投稿内容に対する否定的な態度を反映するには Web ページ間の反論関係を負のエッジとしてグラフに追加する必要がある。これを実装した手法として、Kerchove らの PageTrust[2] がある。これは、グラフ内の各 Web ページが他の Web ページをどの程度反論しているのかを考慮することで、PageRank[1] を負のエッジが組み込めるように拡張したものである。このアルゴリズムにより、反論関係を考慮した信頼度推定を行うことが理論上可能になったが、実際のデータを用いた検証はなされていない。

本研究では、代表的なマイクロブログである Twitter を用いて PageTrust[2] の実データ上での検証を行い、ツイートの信頼度推定の性能を評価する。具体的には、ツイート集合から同意・反論を考慮したグラフを作成する手法を提案し、PageTrust を用いて信頼度の高いツイートランキングを作成する。また、東日本大震災時に情報の信頼性が問題になったトピックに対して提案手法を適用し、負のエッジを考慮することで信頼度推定が改善することを実証する。

以降、2 章では信頼度推定アルゴリズムに関する関連研究を取り上げ、本研究で使用する PageTrust について詳しく説明する。3 章では PageTrust を Twitter に適用するためのグラフ作成手法について説明する。4 章では実験設定と結果、その考察について記述し、5 章では本研究で明らかになった点と今後の課題をまとめる。

連絡先: 佐藤雅宏，東北大学 大学院情報科学研究科，宮城県仙台市青葉区荒巻字青葉 6-3-09，022-795-7140，022-795-4285，sato-m@ecei.tohoku.ac.jp

2. 関連研究

信頼度を推定する手法には、初期値を伝搬させる手法 [7]、初期値なしでグラフ解析を行う手法 [1, 2, 3]、分類器を用いる手法 [5]、半教師ありでグラフの最適化問題を解く手法 [6] などが存在する。しかし、マイクロブログ上では、各情報に対する正しい真偽値を予め定めることは困難であるため、今回は初期値や教師の必要としない手法に焦点を当てた。また、分類器による手法は一般的にグラフ解析による手法より信頼度推定の精度が低いため、本研究では初期値なしでグラフ解析を行う手法 [1, 2, 3] を用いた。

以下では、初期値なしでグラフ解析を行う手法として PageRank アルゴリズム [1] を説明し、その上で本研究で用いる PageTrust アルゴリズム [2] を詳しく説明する。

2.1 PageRank

PageRank[1] は主に Google が Web ページの重要度を計算するために使用される教師なしグラフ解析アルゴリズムで、確率を用いて信頼度を計算するため、入力できるグラフは正のエッジのみで構成されたグラフとなる。計算方法としては、グラフに存在する各ノード $i \in N$(N は全ノードの集合) の信頼度 x_i が収束するまで式 (1) を反復計算する。ただし、信頼度 x_i は $[0, 1]$ の実数で、各ノード $i \in N$ の初期値は入力するグラフの全ノード数 n を用いて $x_i^{(0)} = 1/n$ で計算される。

$$x_i^{t+1} = \alpha \sum_{j,(j,i)\in L^+} x_j^{(t)}/d_j + (1-\alpha)z_i \qquad (1)$$

L^+ は正のエッジの集合を表し、d_j はノード j から張られたエッジの総数を表わす。ここで、パラメータ α は各グループごとに決められた $[0, 1]$ の定数で親ノードから得られる信頼度の割合を定めている。また、z_i は親ノード以外から得られる信頼度の値 (ランダム成分) を示している。z_i は $\sum_{i=1}^{n} z_i = 1$ を満たし、通常は $z_i = 1/n, i \in N$ で与えられる。

ベクトル x は式 (2) のように行列 G の固有ベクトル π に収束することが証明されている。ここで行列 G の各要素は $G_{ij} = \alpha A^+_{ji}/d_j + (1-\alpha)z_i, i, j \in N$ で計算される。

$$\lambda\pi = G\pi \tag{2}$$

2.2 PageTrust

PageTrust[2] は正のエッジのみを扱う PageRank に負のエッジを組み込めるように拡張したアルゴリズムである。正のエッジだけでなく負のエッジを組み込むことによって、Web ページ A が Web ページ B を反論しているなどのネガティブな意見を取り入れることができる。計算方法としては、各ノード $i \in N$ の信頼度 x_i が収束するまで式 (3) を反復計算して信頼度 x_i を求める。ただし、x_i は $[0, 1]$ の実数である。

$$x_i^{(t+1)} = (1-\tilde{P_{ii}})^\beta \cdot \left(\alpha \sum_{j,(j,i)\in L^+} x_j^{(t)}/d_j + (1-\alpha)z_i\right) \tag{3}$$

$\tilde{P_{ii}}$ は負のエッジを考慮した場合にノード $i \in N$ が信頼できない確率を表し、式 (4) で計算される。従って、$(1-\tilde{P_{ii}})$ はノード $i \in N$ が信頼出来る確率を表している。また、β は負のエッジの影響力を決めるパラメータである。

$$\tilde{P}^{(t+1)} = T^{(t)} \cdot P^{(t)} \tag{4}$$

T と P はどちらも $n \times n$ の行列であり、T を遷移行列、P を反論行列と呼ぶ。それぞれ式 (5)、(6) で計算する。

$$T_{ij}^{(t)} = \frac{\alpha A^+_{ji} x_j^{(t)}/d_j + M \cdot (1-\alpha) z_i x_j^{(t)}}{\alpha \sum_{k,(k,i)\in L^+} x_k^{(t)}/d_k + (1-\alpha)z_i} \tag{5}$$

$T_{ij}^{(t)}$ は時間 t の時のノード $i \in N$ の信頼度のうち、ノード $j \in N$ から得られた信頼度の割合を示す。また、A^+ は正のエッジに関する隣接行列で、z_i は $z_i = 1/n, i \in N$ で与えられる。M はランダム成分に関して負のエッジの情報を記憶する $(M = 1)$ かしない $(M = 0)$ かを表すパラメータで、$M = 1$ の方が $M = 0$ の場合より負のエッジの影響力が高くなる。

$$P_{ij}^{(t+1)} = \begin{cases} 1 & (\text{if } (i,j) \in L^-) \\ 0 & (\text{if } i = j) \\ \tilde{P}_{ij}^{(t+1)} & (\text{otherwise}) \end{cases} \tag{6}$$

$P_{ij}^{(t)}$ は時間 t でノード $i \in N$ がノード $j \in N$ を反論している割合を示す。そのため、遷移行列 T と反論行列 P の内積をとることで $\tilde{P}$ の対角成分 $\tilde{P_{ii}}$ がノード $i \in N$ の信頼出来ない確率を表すことができる。初期値としては $P^{(0)} = \tilde{P^{(0)}} = A^-$ (A^- は負のエッジに関する隣接行列) を用いる。

PageTrust の原理を分かりやすく説明するために、random walker というモデルを用いる。random walker とは、PageRank や PageTrust のもととなったモデルで、グラフの各ノードに walker と呼ばれる人が存在すると仮定し、各ノードに存在する walker の数が多いほどそのノードの信頼度が高いとする考え方である。初期状態では各ノードに一定の walker が配置され、各反復でエッジが張られたノードへランダムに移動する。そのため、信頼度の高いノードから多くのエッジが張られているノードにはたくさんの walker が集まることになり、そのノードの信頼度は高くなる。つまり、正のエッジのみのグラフを考えた場合に、random walker によって導かれたアルゴリズムが、2.1 節で説明した PageRank である。

ではここで、図 1 に示すような負のエッジを追加したグラフを考える。この時、PageTrust では random walker に次の 3 つの制約を加える事で負のエッジの影響力を計算する。

- 負のエッジは walker が通れない道である
- 時刻 t で負のエッジ $(i, j) \in L^-$ を持つノード i に辿り着いた walker は、時刻 $t+1$ でノード j が信頼出来ないノードだという意見を持ち、隣接ノードに移動する
- 時刻 t でノード $k \in N$ に辿りついた walker は、時刻 $t+1$ でノード k は信頼できるノードだという意見を持ち、隣接ノードに移動する

これらの制約により、負のエッジは walker に意見を与える情報の役割を果たす。また、P_{ij} はノード i にいる walker のうちノード j を信頼出来ないという意見を持つ walker の割合、T_{ij} はノード i にいる walker のうちノード j から遷移してきた walker の割合を表わす。さらに、各ノード $i \in N$ に存在する walker のうち、ノード i が信頼出来ないという意見を持つ walker の割合がノード i の信頼出来ない確率 $\tilde{P_{ii}}$ で表されている。これにより、PageRank の計算に加えて負のエッジの影響力を計算できるため、PageTrust では負のエッジを組み込むことが可能となる。

3. PageTrust の Twitter への適用

本章では、Web ページを対象としたアルゴリズムである PageTrust を、Twitter への投稿に適用するためのグラフ作成手法を提案する。

3.1 同意・反論エッジの張り方

本研究では、東日本大震災時のツイートの中から真偽が問題になった事例を鍋島らの手法 [4] を用いて集めた。真偽が問題となる事例には「〜はデマです」のような訂正表現が多く見られたため、訂正表現に着目して同意・反論関係を判断した。例えば、(1) のようなツイートを考える。

(1) 被曝を予防するためにイソジンを飲めというデマが流れているので止めてください。

(1) は「というデマ」という表現から、特定の情報を訂正するツイートであることが分かる。このとき (1) が訂正しているのは、「というデマ」の前の部分である「被曝を予防するためにイソジンを飲め」という情報である。つまり、(1) は「被曝を予防するためにイソジンを飲め」という情報に反論の立場を表明しており、この情報を支持するツイートと (1) が反論関係であると判断できる。また、同意関係を考える場合は、ツイートの内容がどれだけ似ていても「〜はデマ」のような否定表現があるだけで正反対の内容になるという問題が存在する。そのため、否定表現が含まれるツイートと含まれないツイートにカテゴリ分けし、同じカテゴリ内で内容の似ているツイート同士が同意関係にあると判断した。以下では具体的なグラフの作成手法について説明する。

まず、鍋島らの手法 [4] を用いて訂正ツイートを検索し、各ツイートを訂正カテゴリとその他カテゴリの 2 つに分類する。鍋島らの手法は正規表現によって表 1 のような表現を含むツイートを検索し、マッチしたものを訂正ツイートと判断するもので、検索時には「というデマ」のように、表 1 に書かれた「接続パターン」+「5 文字までの任意の文字列」+「訂正パターン」がツイートの本文内に存在するかどうかを調べる。この手

表 1: 正規表現リスト

接続パターン	訂正パターン
は、的な、などの、 との、とか、って、 なんて、という、 のような	デマ、誤報、誤り、誤情報、 嘘、虚偽、チェーンメール、 事実はない、事実はありません、 うそ、まちがい、ウソ、ガセ

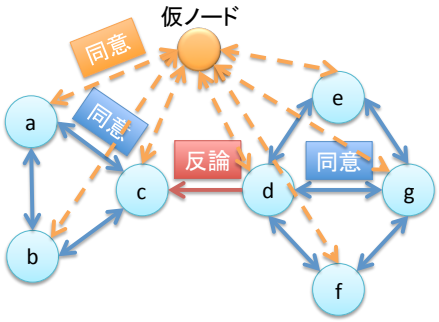

図 1: 作成されるグラフの例

法により訂正ツイートと判断されたツイートを訂正カテゴリ、それ以外のツイートをその他カテゴリに追加する。

次に、各ツイート間の類似度を計算し、類似度の高いツイート間に同意もしくは反論の関係を付与する。類似度とは 2 つの文章がどれだけ似ているかを表す指標で、本研究では文章に出現する単語のコサイン類似度を用いて計算した。類似度の高いツイート同士は内容が似ていると判断できるが、「被曝を予防するためにイソジンを飲め」と「被曝を予防するためにイソジンを飲めという情報はデマ」のように、類似度が高くとも内容は反論関係にあるものが存在する。そこで本研究では、同じカテゴリ内のツイート間で類似度が高かった場合は同意関係、異なるカテゴリ間で類似度が高かった場合は反論関係と見なした。この時、訂正カテゴリのツイートは「被曝を予防するためにイソジンを飲め」のように訂正パターンの前部分を本文として類似度を計算した。同じカテゴリ内では、類似度を計算した 2 つのツイート両方に否定表現が含まれている、もしくは両方に否定表現が含まれていないため、類似度の高いツイートの内容は同意関係にあると判断できる。また、異なるカテゴリ間ではどちらか片方のツイートにのみ否定表現が含まれるため、類似度の高いツイートは反論関係にあると判断できる。これにより、否定表現の有無だけで内容が正反対になるツイートにも対応することができる。

最後に、同意・反論関係にあるツイートをグラフのノードとして用い、同意関係にあるツイート間に両向きの正のエッジ、反論関係にあるツイート間に訂正ツイートからその他ツイートに向けて負のエッジを張ることで、各ツイート間の同意・反論関係をグラフとして表現する。この時、類似度は 2 つの文章がどれだけ似ているか、つまり 2 つのツイートがどの程度同意関係にあるかを表しているため、エッジの重みに類似度の値を用いた。この手法により作成されるグラフの例を図 1 に示す。

3.2 仮ノードの追加

PageTrust は入力として負のエッジを含むグラフを用いることができる。これは、負のエッジは通れない道と定め、負のエッジ $(i, j) \in L^-$ が張られたノード i に辿り着いた walker が、ノード j が信頼出来ないノードだという意見を持ちながら移動を続けるからである。しかし、図 1 のようにお互いに同意関係にあるグループ間に反論のエッジを貼った場合は、同意関係にあるノード間だけで walker が移動し、信頼出来ないノードの情報を持つ walker が他のグループに移動することができないため、反論のエッジが存在しない場合と同じ動きをしてしまう。そこで本研究では、この現象を防ぐために新たなノードとして仮ノードを追加した。仮ノードは全てのノードに対して重み $1/n$(n は仮ノードに張られたエッジの総数) の同意エッジが張られたノードで、これにより、図 1 の左右の同意グループ間を仮ノードを経由することで walker が移動できるようになるため、反論エッジが正しく作用されると考えられる。

表 2: 実験に用いたデータセット

データセット名	キーワード	ツイート数	制限 RT 数
コスモ石油	コスモ石油	72887 件	50
イソジン	イソジン	24883 件	10
尾田	尾田	21097 件	10
トルコ	トルコ	27421 件	20

3.3 ソース URL の追加

作成したグラフに新たなノードとしてソース URL を追加し、ソース URL を持つツイートとの間に同意のエッジを張った。ソース URL とは (2) のようにツイートの本文に記載された情報の提供元となる URL のことである。

(2) コスモ石油の爆発で有害物質の雨が降る件はデマ。
http://ow.ly/4cYQ9

これにより、同じソース URL が記載されたツイート同士の結びつきが強くなるため、ソース URL が記載されたツイートの信頼度が高くなることが予想される。また、一般的にソース URL が記載されたツイートは、記載されてないツイートに比べ信頼度が高いため、ソース URL を追加することで信頼度推定の精度を向上させることができると考えられる。さらに、ソース URL を追加することで、ツイートの信頼度だけでなくソース URL の信頼度も同時に推定できるという利点もある。今回はどのドメインの Web ページから情報を得ているのかが重要と考え、ドメイン名までをソース URL としてグラフに追加した。

4. 実験

4.1 実験設定

実験に用いるデータセットを表 2 に示す。各データセットは、東日本大震災時 (2011/3/11 ～ 2011/3/29) のツイートのうち特定のキーワードを含むもので、計算時間の短縮と雑談のようなツイートを取り除くため、RT 数で制限を加えた。実験では各データセットを入力し、算出された PageTrust のスコアが高いツイートのランキングを作成する。ベースライン手法として、各ノードに張られた同意・反論のエッジの数をそれぞれ求め、(同意エッジの数) – (反論エッジの数) の値が高いツイートでランキングを作成する。

正解データは、人手で真、偽、その他の 3 つのラベルをつけたものを用意した。真のラベルは実際に起こったイベント等、正しい内容のツイートであることを表し、偽のラベルは実際に起こったイベントとは異なる事象のように、間違っている内容のツイートを表す。また、その他のラベルは雑談やユーザーの個人的な意見のように真偽を判断することができないツイートを表す。

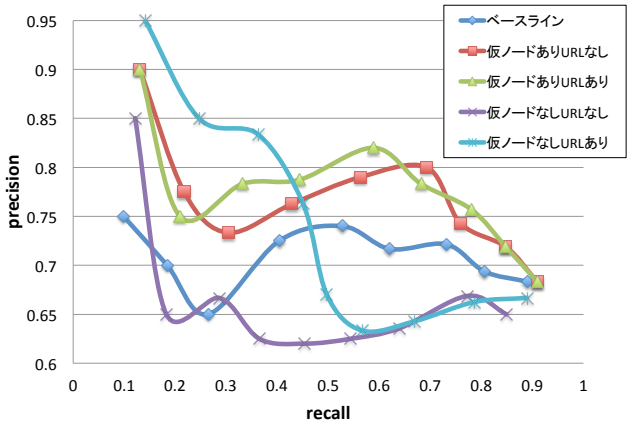

図 2: 実験結果

4.2 評価方法

信頼度の高いツイートランキング上位 m 件について、閾値 m を変化させながら PR 曲線を描くことによって評価を行う。また、*precision* と *recall* は次式で計算する。

$$precision = \frac{\text{上位 } m \text{ 件までに含まれる真のツイート数}}{m} \quad (7)$$

$$recall = \frac{\text{上位 } m \text{ 件までに含まれる真のツイート数}}{\text{正解データ内の真のツイートの総数}} \quad (8)$$

ソース URL を追加した場合は、グラフのノード数が追加していない場合よりも増加してしまうため、ソース URL を追加することで増えたノードを無視して評価を行う。

4.3 実験結果

各データセットの実験結果を平均したグラフを図 2 に示す。図 2 のグラフは全て一番上の点の閾値が $m = 5$ であり、上から下に向かって閾値 m を 5 ずつ増加させながら *precision* と *recall* を計算している。また、閾値 m の最大値はデータセットにより異なるため、最も小さいデータセットの値に統一した。

図 2 より、追加要素を考慮しない場合はベースラインよりも PageTrust の信頼度推定精度が低いのに対し、仮ノードもしくはソース URL を追加することで推定精度が大きく向上していることが分かる。仮ノードを追加した場合については、3.2 節でも述べた通り、図 1 のようなグラフにおいても反論のエッジの影響力を正しく反映させることが出来たため、推定精度が向上したと考えられる。また、ソース URL を追加した場合では、特に仮ノードを追加しない場合におけるランキング上位 20 件までの precision の高さが顕著である。これは、ソース URL の記載されたツイートの信頼度が高くなっただけではなく、同意・反論関係の判別ミスによって異なるカテゴリに分類された同じ内容のツイート間に、ソース URL を介した同意のエッジが張られたためと考えられる。これにより、異なるカテゴリ間に同意のエッジが張られるため、仮ノードを追加した場合と同様の効果が得られると同時に、同意・反論関係の判別精度をフォローすることが可能となる。

5. まとめ

本研究では、PageTrust を Twitter データに適用するためのグラフ作成手法を提案し、PageTrust の Twitter データによる性能評価を行った。これにより、PageTrust はグラフの作成手法を工夫することで、Twitter データにも適用可能であることが分かった。また、グラフ作成時には追加要素として仮ノードとソース URL を追加したが、これらの要素は PageTrust による信頼度推定において有用であることが分かった。今後の課題としては、同意・反論関係の判別精度を向上させることが挙げられる。今回は正規表現による判別手法を用いたが、正規表現では予め想定していた表現が出現しなければ判別することが出来ず、特定のドメインではうまく判別することが出来ない。そのため、より複雑な否定表現に対応した同意・反論関係の判別手法を取り入れることで、信頼度推定の精度を更に向上させることができると考えている。

謝辞

本研究は、文部科学省科研費 (23240018)、文部科学省科研費 (23700159)、および JST 戦略的創造研究推進事業さきがけの一環として行われた。

参考文献

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, 1999.

[2] De Kerchove, Cristobald and Dooren, PV. The PageTrust algorithm: how to rank web pages when negative links are allowed. Proc. SIAM Int. Conf. on Data Mining, 346–352, 2008.

[3] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 604–632, 1999.

[4] 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. マイクロブログからの誤情報の発見と集約. 言語処理学会 第 19 回全国大会 発表論文集, 2013.

[5] Castillo Carlos, Mendoza Marcelo and Poblete Barbara. Information credibility on twitter. Proceedings of the 20th international conference on World wide web, 675–684, 2011.

[6] Yin Xiaoxin, Tan Wenzhao. Semi-supervised truth discovery. Proceedings of the 20th international conference on World wide web, 217–226, 2011.

[7] Yin Xiaoxin, Han Jiawei and Yu Philip S. Truth discovery with multiple conflicting information providers on the web. Knowledge and Data Engineering, IEEE Transactions on, 796–808, 2008.