

ニュース記事に関連するツイート収集手法の提案とその評価

邱 起仁^{1,a)} 櫛山 淳雄¹

概要: Twitter には、ニュース記事の話題に対する意見や感想を含む投稿が多数存在している。そこで本研究では、それらの意見や感想を抽出するため、ニュース記事に関連するツイートを収集する手法を提案する。ツイートはリンク付きツイート、リンクなしツイートに分類できる。提案手法では、まずリンク先のテキストを利用し、リンク付きツイートとニュース記事を関連づける。次に、リンク付きツイートがニュース記事を言及する際に、ニュース記事から引用した語を重要語とし、それらの語の重要度を考慮し、リンクなしツイートとニュース記事の関連づけを行う。また、ニュース記事の作成時刻とツイートの投稿時刻が近ければ近いほど、同じ話題に関連する可能性が高いと考えられる。そこで本研究では、ニュース記事とツイートの内容的類似性のみではなく、時間的類似性を合わせて考慮することで、関連づけの精度を向上させる。評価実験により、提案手法は 88.1% の適合率を保ちながら 98.1% の再現率でニュース記事に関連するリンク付きツイートを抽出することができ、65.2% の適合率を保ちながら 93.8% の再現率でニュース記事に関連するリンクなしツイートを抽出することができた。

Proposal and Evaluation of a Method for Collecting Tweets Related to News Articles

QIREN QIU^{1,a)} ATSUO HAZEYAMA¹

Abstract: Many tweets including opinions and comments regarding the topic of news articles exist in Twitter. In order to extract these opinions and comments, this paper proposes a method for collecting the tweets related to news articles. We classify the tweets into 'tweet with urls' and 'tweet without urls.' In the proposed method, at first, we associate news article with 'tweet with urls' by using texts that are linked to the tweet. Next, if 'tweet with urls' refers to a news article, the words quoted from the news article are regarded as important, and then the association of 'tweet without urls' with the news article is performed in consideration of the importance of those words. Moreover, we take not only the contents similarity between tweets and news articles, but also the time similarity into consideration to improve accuracy of the association. The evaluation experiment shows that our method can extract 'tweet with urls' related to news articles on the ratio of recall of 98.1% while maintaining the ratio of precision of 88.1%, and extract 'tweet without urls' related to news articles on the ratio of recall of 93.8% while maintaining the ratio of precision of 65.2%.

1. はじめに

近年、マイクロブログサービスの代表である Twitter[1] が急速に普及してきており、Web 上の重要な情報源のひとつになってきている。Twitter はニュースメディアと密な関係を持ち、Twitter では、85% の話題はニュースに関連するものであるとされている [2]。また、Twitter には、

ユーザの政治・社会に対する意見や商品に対する感想など、個人の主観的な評価情報を含む投稿が多数存在しており、Twitter はセンチメント分析や意見抽出の情報源として有効であることが検証されている [3]。すなわち、Twitter からニュース記事に対する評価情報を抽出することが可能であると考えられる。

ニュース記事に対する評価情報の抽出には多くの効果が期待できる。例えば、企業が提供する商品・サービスや、政党が推進する政策などがニュースとして報道されることは少なくない。企業や政党がニュース記事に対する世の中の

¹ 東京学芸大学大学院
Tokyo Gakugei University

^{a)} m123306m@st.u-gakugei.ac.jp

評価を把握することは、商品・サービスや政策の改善につながる可能性がある。また、商品・サービスの利用や政策の支持を考えるユーザは他のユーザの評価情報を参考にして、それらの商品・サービスや政策の良いところ、悪いところを知ることができる。これらの要求は、Twitter からニュース記事に対する評価情報を抽出することで実現可能である。そのため、まずニュース記事に関連するツイートを収集する必要がある。

ニュース記事とブログ記事の特徴ベクトルを作成し、その特徴ベクトル間のコサイン類似度によりニュース記事とブログの関連づけを行う研究 [4], [5] がある。しかし、従来のブログ記事と異なり、ツイートには 140 文字以内という文字数の制限があり、個々のツイートに含まれる情報量が極めて少ない。従ってツイートの特徴語を抽出し特徴ベクトルを作成することが困難である。Cao らはニュース記事に関連するマイクロブログの投稿を収集する手法を提案した [6]。彼らはニュース記事とマイクロブログの投稿に含まれる語の共起頻度に基づき、ニュース記事とマイクロブログの投稿の関連づけを行った。一部のツイートにはニュースサイト、ブログ等へのリンクが含まれている。本研究では、このようなツイートを「リンク付きツイート」と呼ぶ。それに対し、リンクを含んでいないツイートを「リンクなしツイート」とする。リンク付きツイートのリンク先テキストを併用し、情報量を増加させることが望ましいが、Cao らはツイート本文のみを考慮し、ツイートに含まれるリンク先の情報を利用してない。また、上記の研究はいずれも文書間の内容的類似性のみを考慮し、時間的類似性を考慮していない。戸田らは、関連する話題の記事は近い時期に発行されることが多いため、時間的な近さを考慮することにより、特定の話題に関する文書を精度良く集められると述べている [7]。マイクロブログの投稿とニュース記事の関連づけを行う際に、時間的類似性を考慮することにより、関連づけの精度を向上させることができると考えられる。

そこで、本研究ではニュース記事に関連するツイートを収集することを目的とし、ニュース記事とツイートの類似度計算の手法を提案する。特に本研究では、ツイートの投稿時刻はニュース記事の作成時刻に近ければ近いほど、そのツイートとニュース記事は同じ話題に関するものである可能性が高いという考えに基づき、マイクロブログの投稿とニュース記事の時間類似度を計算する。また、ツイートとニュース記事の内容的類似性を評価する際に、ツイートではそのニュース記事を言及する時によく使用され、他の場合にはあまり使用されないような語をニュース記事の重要語として抽出する必要がある。ツイートがこのような重要語を含む場合、そのニュース記事に関連するものである可能性が高いと考えられる。本研究ではツイートとニュース記事の内容類似度を計算する時に、従来の TF-IDF 法に加えて、このような重要語を抽出して適切な重みを与える

ことで、内容類似度の計算精度を向上させる。

以上のような時間類似度と内容類似度をもとに、ツイートとニュース記事の総合類似度を算出し、その総合類似度によりツイートとニュース記事の関連付けを行う。

以下、本稿の構成を示す。第 2 節では提案手法を説明する。第 3 節では総合類似度閾値及び時間半減期の決定手法を述べる。第 4 節では提案手法の有効性を検証するための評価実験を説明し、実験結果を考察する。最後に第 5 節で本研究をまとめ、今後の課題を述べる。

2. 提案手法

2.1 提案手法の概要

本研究では、Twitter のパブリックタイムラインから、ニュース記事について言及するツイートを抽出する手法を提案する。提案手法の概要を図 1 に示す。本研究ではリンク付きツイートのツイート本文とリンク先テキストを合わせて、拡張ツイートと呼ぶ。拡張ツイートとニュース記事のコサイン類似度を、リンク付きツイートとニュース記事の内容類似度とする。

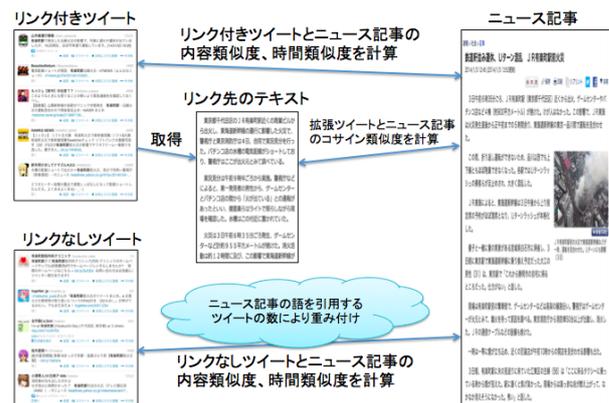


図 1 提案手法の概要

提案手法の流れを図 2 に示す。本研究では、まずツイートをリンク付きツイート、リンクなしツイートという 2 つの集合に分類する。次に、各集合のツイートとニュース記事の総合類似度を計算する。最後に算出した総合類似度が閾値 Threshold より大きい場合、ツイートとニュース記事を関連づける。リンク付きツイート、リンクなしツイートとニュース記事の内容類似度の計算手法を 2.2 項、2.4 項で述べ、時間類似度の計算手法を 2.5 項で述べる。最後に、内容類似度、時間類似度に基づき、ツイートとニュース記事の総合類似度を計算する手法を 2.6 項で説明する。

2.2 リンク付きツイートとニュース記事の内容類似度

リンク付きツイートとニュース記事の内容類似度 $sim_{link}(tweet, news)$ を計算する手法について述べる。手法の概要を図 3 に示す。ツイートの話題はリンク先のコン

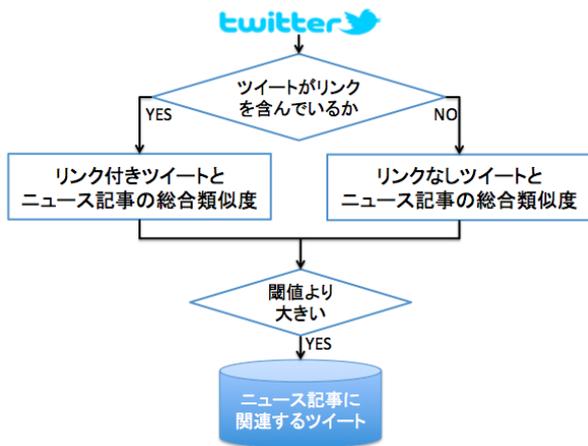


図 2 提案手法の流れ

テンツに強く関連していると述べられている [8]. 本研究では、文書間の内容類似度に注目するため、リンク先のコンテンツからテキストのみを抽出して利用する。

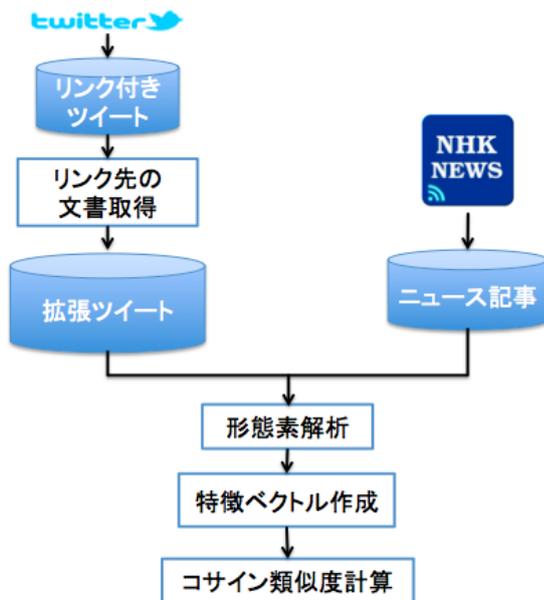


図 3 リンク付きツイートとニュース記事の内容類似度

まず、リンク付きツイートのリンク先テキストを取得する。ツイート本文とリンク先テキストを合わせて拡張ツイートを作成する。拡張ツイートに対して、形態素解析を行い、名詞、形容詞、動詞を抽出し、それぞれツイート情報、リンク先テキスト情報として登録しておく。なお、ツイート情報とリンク先テキスト情報を合わせて拡張ツイート情報とする。また、ニュースサイトから収集したニュース記事のタイトル、本文に対して、形態素解析を行い、名詞、形容詞、動詞を抽出し、タイトル情報、本文情報として登録しておく。タイトル情報、本文情報を合わせてニュース記事情報とする。

次に、拡張ツイートとニュース記事から抽出した各語の

TF-IDF 値を算出する。なお、IDF 値を計算するにあたって、全文書集合として、拡張ツイートとニュース記事集合の全体を用いる。そして、算出した TF-IDF 値により、各拡張ツイート、ニュース記事の特徴ベクトルを生成する。

最後に、拡張ツイートとニュース記事の特徴ベクトル間のコサイン類似度を計算する。算出したコサイン類似度をリンク付きツイートとニュース記事の内容類似度と見なす。

2.3 ニュース記事中の語の重要度

本項では、ニュース記事中の語の重要度を計算する手法について述べる。本研究では、ツイートがニュース記事について言及する際に引用しやすく、他の場合にはあまり引用しない語を重要語とするため、以下の 2 点を考慮し、ニュース記事中の語の重要度を計算する。

- (1) ツイートがニュース記事について言及する際に、繰り返し何度も引用するニュース記事中の語
- (2) 多くのツイートがニュース記事について言及する際に、引用するニュース記事中の語

140 文字以内という文字数の制約があるツイートがニュース記事について言及する際に、ニュース記事中の語を繰り返し何度も引用することが少ないと考えられるため、本研究ではニュース記事中の語 t を引用するツイート数のみを考慮し、語 t の重要度を算出する。本研究では、あるニュース記事 N に関連するリンク付きツイートを検出し、そのツイート集合を R_{link} とする。そして、ニュース記事中の語 t の重要度 $importance_t$ を式 (1) で計算する。

$$importance_t = \frac{|\{tweet : tweet \ni t\}|}{|R_{link}|} \quad (1)$$

なお、 $|\{tweet : tweet \ni t\}|$ は語 t を引用しているツイート集合のツイート数を意味する。ニュース記事中の語 t を引用するリンク付きツイートが存在しない場合、語 t の重要度 $importance_t$ が 0 になってしまう。また、式 (1) で算出する $importance_t$ の範囲は $[0, 1]$ である。従って、2.4 項でリンクなしツイートとニュース記事の内容類似度を計算する際に、 $importance_t$ をかけると、内容類似度が逆に小さくなってしまふ。これを避けるため、語 t の重要度として、 $importance_t$ に 1 を加えるものを使用する。

2.4 リンクなしツイートとニュース記事の内容類似度

Phuvipadawat と Murata は語の共起頻度を調査し、文字列の長さが短いツイート間の類似性の検出に成功している [9]。語 t が固有名詞 (例えば、日本、中国、ワールドカップ、オリンピック等がある) である場合、係数 $boost(t)$ をかけることでツイートの語に対する重み付けをする。しかし、ニュース記事に含まれる固有名詞は必ずしも重要語であるとは限らないため、本研究では 2.3 項で算出したニュース記事中の語の重要度を用いて、式 (2) でリンクなしツイートとニュース記事の内容類似度 $sim_{nolink}(tweet, news)$ を算

出する。IDF 値の計算にあたり全文書の集合としてニュース記事の集合を使用する。リンクなしツイートとニュース記事の内容類似度計算手法の概要を図 4 に示す。

$$sim_{nolink}(tweet, news) = \sum_{t \in tweet} [tf(t, news) \times idf(t) \times importance_t] \quad (2)$$

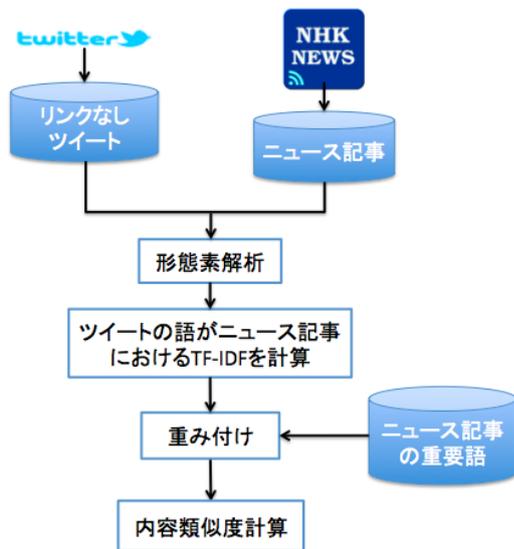


図 4 リンクなしツイートとニュース記事の内容類似度

2.5 ツイートとニュース記事の時間類似度

実世界で出来事が発生し、先に Twitter で議論され、その後ニュースで報道されることがある。また、実世界の出来事が先にニュースで報道され、その後 Twitter 上で話題になることもある。いずれの場合においても、ツイートの投稿時刻とニュース記事の作成時刻が近ければ近いほど、同じ実世界の出来事に関連する可能性が高いと考えられる。本研究では、ツイートの投稿時刻とニュース記事の作成時刻を考慮することで、偶然似ているような語を含んでいても、作成時刻が離れて実は異なる出来事に関連するツイートとニュース記事を関連づけてしまうことを防止する。また、内容類似度が比較的低いが、作成時刻が近く同じ出来事に関連する可能性の高いツイートとニュース記事を正確に関連づける。

時間類似度 $sim_{time}(tweet, news)$ は戸田らの提案手法を活用し、式 (3) で計算する。なお、 $time_i$ はツイートの投稿時刻であり、 $time_k$ はニュース記事の作成時刻である。また、 $t_{1/2}$ は時間類似度が 50% になる時の $time_i$ と $time_k$ の差であり、本研究では時間半減期と呼ぶ。

$$sim_{time}(tweet, news) = \exp\left(-\frac{0.693}{t_{1/2}} |time_i - time_k|\right) \quad (3)$$

戸田らの実験によると、ニュース記事を対象とする場合、

時間半減期 $t_{1/2}$ を 30~90 日に設定した時に有効な結果が得られる。本研究では、ニュース記事のみではなく、ツイートも対象とするため、改めて時間半減期を決定する必要がある。その決定手法は 3 節で述べる。

2.6 ツイートとニュース記事の総合類似度

ツイートとニュース記事の内容類似度、時間類似度を算出した後、式 (4) で総合類似度 $sim(tweet, news)$ を算出する。なお、本研究ではツイートとニュース記事の内容が類似するほど、同じ話題に関連する可能性が高いと考え、内容類似度に 0.6、時間類似度に 0.4 という重みを与えている。

$$sim(tweet, news) = 0.6 \times sim_{content}(tweet, news) + 0.4 \times sim_{time}(tweet, news) \quad (4)$$

リンク付きツイート、リンクなしツイートとニュース記事の総合類似度を計算する際に、ツイートとニュース記事の内容類似度 $sim_{content}(tweet, news)$ にそれぞれ $sim_{link}(tweet, news)$, $sim_{nolink}(tweet, news)$ を代入する。総合類似度が閾値より大きいツイートとニュース記事を関連づける。

3. 総合類似度閾値と時間半減期の決定

本研究では、提案手法の適合率と再現率のバランスを向上させることを目標としているため、適合率と再現率の調和平均である F 値 (F-measure) を利用する。ニュース記事のキーワードを含むツイート集合から、人手により対象となるニュース記事に関連しないものを削除したツイートの集合を正解データ集合 C とする。

戸田らの実験によると、ニュース記事を対象とする場合、時間半減期を 30~90 日に設定したときに有効な結果が得られる。本研究では、総合類似度閾値 Threshold と時間半減期 $t_{1/2}$ を決定するため、まず、 $t_{1/2}$ を 60 日に設定し、以下の手順で総合類似度閾値を決定する。

step 1. ツイートの集合を用いて、Threshold を変化させながら、2 節で提案した手法により、ニュース記事に関連するツイート集合 N を抽出する。N に含まれる正解集合を R とする。提案手法の再現率 (Recall) と適合率 (Precision) を以下のように算出する。

$$\text{再現率} = \frac{R}{C}$$

$$\text{適合率} = \frac{R}{N}$$

step 2. step 1 で算出した再現率と適合率により、F 値を以下のように計算する。

$$F \text{ 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

step 3. step 2 で算出した F 値が最大となる Threshold

を、ツイートとニュース記事の総合類似度閾値とする。次に、決定した総合類似度閾値を用いて、 $t_{1/2}$ を 30~90 日の範囲で変化させながら、総合類似度閾値の決定手順と同様に、提案手法の F 値を求め、F 値が最大となる $t_{1/2}$ を選定する。予備実験の結果により、リンク付きツイートとニュース記事の総合類似度閾値を 0.55、リンクなしツイートとニュース記事の総合類似度閾値を 0.6、時間半減期を 30~90 日に設定するとき最適の結果が得られた。

4. 評価実験

4.1 データセットと前処理

評価実験には NHK オンラインから収集したニュース記事 3 件を使用した。各ニュース記事から手でキーワードを選択し、Twitter を検索してツイートを 100 件ずつ収集した。

データ前処理として、本研究ではまず収集したツイートをリンク付きツイート、リンクなしツイートに分類し、リンク付きツイートのリンク先テキストを取得しておく。なお、リンク先テキストの取得に失敗したツイートをリンクなしツイートとする。次にツイート本文に含まれる RT @username、ハッシュタグを削除する。最後に MeCab を利用してツイート本文、リンク先テキスト、ニュース記事の形態素解析を行い、名詞、動詞、形容詞を抽出する。

4.2 評価実験の結果

評価実験結果を図 5 に示す。評価実験により、提案手法は 88.1%の適合率を保ちながら 98.1%の再現率でニュース記事に関連するリンク付きツイートを抽出することができ、65.2%の適合率を保ちながら 93.8%の再現率でニュース記事に関連するリンクなしツイートを抽出することができた。

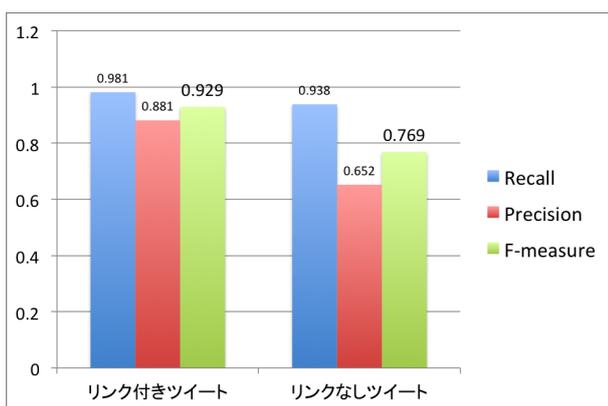


図 5 評価実験の結果

前述したように、リンクなしツイートは 140 文字以内という文字数の制限があり、各ツイートに含まれる情報量が極めて少ないため、ニュース記事との関連づけが難しい。本研究では、ニュース記事に関連するリンク付きツイートを利用し、ニュース記事の語の重要度を算出してリンクなし

ツイートとニュース記事との関連づけを行う。提案手法の有効性を検証するため、本研究では提案手法と、TF-IDF のみで重み付けをする手法の比較実験を行った。各手法を用いて、総合類似度閾値を変化させながら、ニュース記事に関連するツイートを抽出する。そして、抽出結果を評価するための F 値が最大となる再現率、適合率、F 値を図 6 に示す。

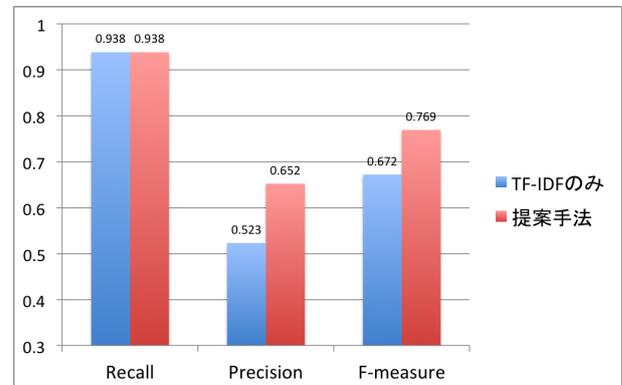


図 6 提案手法と TF-IDF のみで重み付けをする手法の比較

2.5 項で述べたように、ニュース記事の作成時刻とツイートの投稿時刻が近ければ近いほど、ニュース記事とツイートは同じ話題に関連する可能性が高いと考えられる。本研究では、ツイートの投稿時刻とニュース記事の作成時刻を考慮することで、偶然似ているような語を含んでいても、投稿時刻と作成時刻が離れていて実は異なる出来事に関連するツイートとニュース記事を関連づけてしまうことを防止する。また、内容類似度が比較的低いが、作成時刻が近く同じ出来事に関連する可能性の高いツイートとニュース記事を正確に関連づける。時間類似度を考慮する提案手法の有効性を検証するため、内容類似度のみを考慮する手法と提案手法の比較実験を行った。実験結果を図 7 に示す。

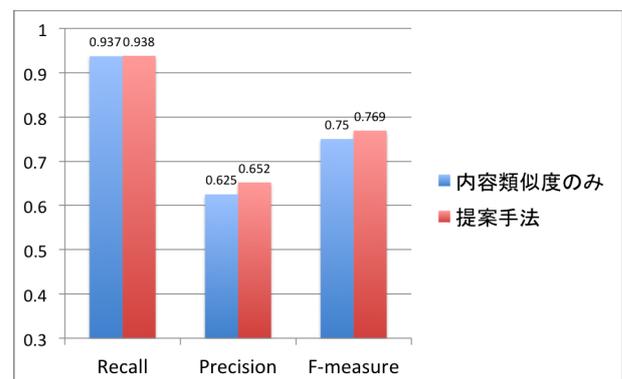


図 7 提案手法と内容類似度のみを考慮する手法の比較

4.3 評価実験の考察

4.3.1 評価実験の結果総評

図 5 からリンク付きツイートの抽出結果が適合率と再

現率ともに、リンクなしツイートの抽出結果を上回っている。評価実験に使用されたデータでは、リンク付きツイートのツイート本文の平均文字数は 20.3 であり、リンクなしツイートのツイート本文の平均文字数は 17.6 である。一方、リンク先テキストは平均 117.7 文字を持ち、提案手法ではリンク付きツイートのツイート本文とリンク先テキストを併用することで、より多くの特徴語が得られ、コサイン類似度を計算するための特徴ベクトルを正確に作成することができたと考えられる。

4.3.2 提案手法と TF-IDF のみで重み付けをする手法の比較

図 6 に示される実験結果を考察する。TF-IDF のみで語の重み付けをする手法により得られた適合率が 0.523 である。それに対して、提案手法で得られた適合率が 0.652 である。提案手法により、再現率を同水準に維持しながら、適合率を 24%、そして F 値をおよそ 14.4% 向上させることができた。TF-IDF のみで語の重み付けをする手法だと、ニュース記事とリンクなしツイートの類似度全体が低いいため、閾値を低く設定しなければならない。一方、本研究では、ニュース記事に関連づけられたリンク付きツイート集合を利用し、ツイートがそのニュース記事について言及する際に引用しやすいニュース記事中の語を重要語と見なし、適切な重みを与えた。それで、ニュース記事とそのニュース記事について言及するリンクなしツイート類似度が上がり、閾値を高く設定することができた。例えば、評価実験に使用したデータのうち、「侍ジャパン」ユニフォーム見つからず」というニュース記事と、「2002 年第 84 回大会 3 回戦、和歌山県代表の智弁和歌山と奈良県代表の智弁学園が対戦したわ。同系列の兄弟校であるため、ユニフォームのデザイン・色調もほぼ同一だわ。違いは、左腕の校章の一部と県名、胸の智辯の大きさのみで非常に稀な同一ユニホーム対決だったそうです。」というリンクなしツイートがあった。TF-IDF のみで語の重み付けをする手法により算出した類似度が 0.544 であり（閾値が 0.525 である）、提案手法で算出した類似度が 0.584 である（閾値が 0.6 である）。TF-IDF のみで語の重み付けをする手法だと、誤ってそのニュース記事とリンクなしツイートを関連づけてしまうが、提案手法はこのような誤抽出を防ぐことができる。

4.3.3 提案手法と内容類似度のみを考慮する手法の比較

図 7 の結果により、時間類似度を考慮した提案手法は従来手法と比べ、適合率を 4.3%、そして F 値を 2.5% 向上させることができたが、時間類似度を考慮することが有効であるとは言えない。その理由として、評価実験に使用したデータでは、ニュース記事の作成時刻とツイートの投稿時刻が近かったため、ツイートとニュース記事の時間類似度はほとんど 0.964 より大きかった。つまり、今回の評価実験では時間類似度の差は総合類似度にあまり影響を与えていないことがわかった。

5. おわりに

本研究はニュース記事に関連するツイートを収集する手法を提案した。また、提案手法の有効性を確認するため、評価実験を行った。特に、本研究ではツイートをリンク付きツイート、リンクなしツイートに分類し、ニュース記事との類似度を算出し関連づけを行った。情報量の少ないリンクなしツイートとニュース記事を正確に関連づけるため、ニュース記事に関連するリンク付きツイートを利用する語の重み付けを行い、その有効性を確認できた。

今後の課題として、まず本研究では語の表記揺れを考慮していない。ニュース記事、ツイートには似た意味で違う語が使用される可能性がある。また、「修論を書く」、「修論をかく」に含まれる「書く」、「かく」のように、漢字変換をしていない語がある。形態素解析では単語の意味を考慮していないため、そのような語が別の語として抽出されてしまう。今後は独立行政法人情報通信研究機構 (NICT) から公開されている日本語の意味辞書である「日本語 WordNet」等を利用し、語の表記揺れを解消する。

次に今回の評価実験では時間類似度がリンクなしツイートとニュース記事の関連付けに与える影響を確認できなかった。今後はより長期間にわたってデータを収集し、時間類似度を考慮する手法の有効性を確認する。

参考文献

- [1] Twitter : 入手先 (Twitter: <https://twitter.com/>) (2014.02.05).
- [2] Kwak, H., Lee, C., Park, H., and Moon, S.: *What is twitter, a social network or a news media?*, In WWW 2010, pp. 591-600, 2010.
- [3] Pak, A., and Patrick P.: *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, In LREC2010, pp.1320-1326, 2010.
- [4] 池田大介, 藤木稔明, 奥村学, blog とニュース記事の自動対応付け, 言語処理学会第 11 回年次大会論文集, pp.1030-1033, 2005.
- [5] 佐藤由紀, 横本大輔, 牧田健作, 宇津呂武仁, 福原知宏, ニュース記事中の話題に関連するブログ記事の収集手法, DEIM Forum 2011 A6-3, 2011.
- [6] Cao, X., Chen, K., Long, R., Zheng, G., and Yu, Y.: *News comments generation via mining microblogs*, In WWW 2012, pp. 471-472, 2012.
- [7] 戸田浩之, 北川博之, 藤村考, 片岡良治: 時間的近さを考慮した話題構造マイニング, 第 18 回データ工学ワークショップ (DEWS 2007 L6-4), 2007.
- [8] 吉田光男, 乾孝司, 山本幹雄: リンクを含むつぶやきに着目した twitter の分析, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM 2010 A5-1), 2010.
- [9] Phuvipadawat, S., and Murata, Y.: *Detecting a Multi-Level Content Similarity from Microblogs Based on Community Structures and Named Entities*, In JETWI 2011, Vol. 3, No.1, pp.11-19, 2011.