テキストセグメンテーション手法を用いたマイクロブログポストの情報源推定

齋藤 正樹　　乾 孝司　　山本 幹雄
筑波大学大学院 システム情報工学研究科
masaki@mibel.cs.tsukuba.ac.jp, {inui,myama}@cs.tsukuba.ac.jp

1　はじめに

近年、マイクロブログと呼ばれる、文書長の短い投稿からなるウェブログサービスが普及している。代表的なマイクロブログサービスである Twitter [1] では、投稿当たりの最大文字数が 140 文字と定められており、投稿される文書には、速報性が高く、投稿の総量が極めて多いといった特徴がある。従来のブログサービスと同様に、マイクロブログに投稿される文書には、意見やレビューなどの価値のある情報が多く含まれている。しかし、マイクロブログでは、投稿内容の記述が簡略化されがちであり、投稿された文書を単独で読むだけでは投稿内容の理解が難しい場合がある。例えば、以下のような Twitter における投稿 (以降、ツイートと呼ぶ) の例が挙げられる。

- (例 1) 自分は東京の治安が心配
(東京五輪の決定に対して)
- (例 2) 声かけたら芸能人が止まってくれた！
(大阪マラソンの観戦)
- (例 3) 決め球が無いにしろ、登板機会が少なすぎ
(野球選手の解雇に対して)

例に挙げたようなツイートをユーザーが投稿する背景には、何らかの情報源が存在している。本研究において情報源とは、投稿者がツイートを行う際に、直接参照したテレビ番組や新聞記事などの情報を指す。情報源の内容をユーザーに提示することは、当該ツイートを補完し、ユーザーの読解を支援する上で有益である。しかし、既存のトピック分類やイベント検知などの先行研究では、例に挙げた「東京五輪」や、「大阪マラソン」など、比較的大きな社会的イベントを推定することは可能であっても、「ある野球選手の解雇」という粒度の細かいイベントの推定を行うことは想定されていない。また、推定されたトピックやイベントは一般に当該のツイートに比べて抽象的であり、投稿者が直接参照した情報自体を示すものではない。

本稿では、このような背景から、新たな情報源推定課題の設定を行う。そして、文書を話題ごとのまとまりに分割する技術である、テキストセグメンテーション手法による情報源推定課題の解決手法を提案する。評価実験では、テキストセグメンテーション手法のひとつである Hearst の TextTiling 法 [1] を本課題に適用した結果を報告する。さらに、Web データに特有の素性を用いた TextTiling 法の拡張について併せて報告を行う。

2　ツイートの情報源

一般に、ツイートの情報源には、テレビ放送から、新聞記事、インターネット、日常生活、交友関係まで様々な種類が考えられる。その中で、本研究では実現性を考慮し、インターネット上のテキスト (Web ページ) が情報源である場合について議論を行う。つまり、あるツイートに対し、その内容を補完する内容が記述された Web ページを推定することが目的となる。ツイートと Web ページの対応を推定することにより、ユーザー支援の他に、Web ページに対する意見ツイートの抽出、検索支援、ユーザープロファイリングなどの応用が期待できる。また、一般的なトピックやイベント、ハッシュタグが持つ情報の粒度に対して、Web ページの持つ情報の粒度は比較的細かく、単独で読んだ際に理解できないようなツイートの情報補完という点で有用であると考えられる。

3　関連研究

研究の目的自体は異なるが、ツイートと Web ページの対応推定を行う研究として、Abel ら [2] がある。Abel らは TF-IDF に基づいて、ツイートと Web 上のニュース記事間の類似度に着目し、両者の関連付けを行うことで、多様なユーザープロファイリングが可能になることを明らかにした。しかし、TF-IDF を用いる手法では、1 件のツイートに対し、全ての Web ページとの類似度の計算を行う必要がある。そのため、現実問題として、推定を行うことができる Web ページの数 (種類) が限定されてしまう。本研究では、後述するように、URL を含むツイートとその近辺のツイートに着目することで、効果的に計算コストを低減させた手法を提案する。

[1] Twitter https://twitter.com/

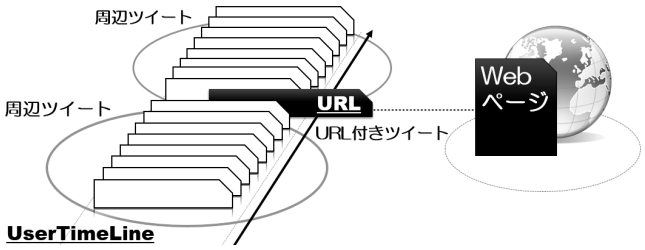

図 1: URL 付きツイートと周辺ツイート

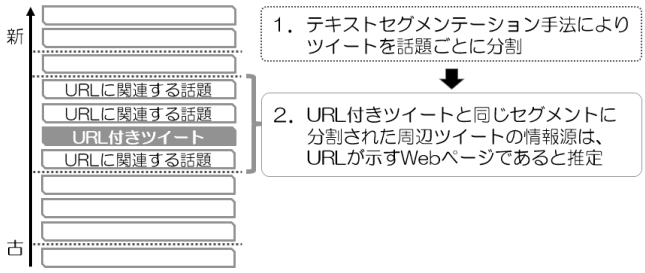

図 2: テキストセグメンテーションを用いた推定

4 情報源推定課題

4.1 URL 付きツイートと周辺ツイート

あるツイートに対し、その情報源である Web ページを推定する上で、二点の問題がある。まず一点目に、先に挙げた Web ページの総量が大きいという問題がある。2012 年 12 月時点での全世界の Web サイトの総数はおよそ 6 億 3400 万件 [2] に上り、これら全てと 1 件のツイートの対応関係を推定することは現実的ではない。また、二点目として、Web 上に情報源を持たないツイートの存在が挙げられる。例えば朝夕の挨拶のツイートなど、そもそも情報源を持たないようなツイートや、ユーザーの日常生活に関するツイートなど、情報源の推定が不可能なツイートに対しては、あらかじめ推定の対象から除外することが望ましい。

提案手法において、上に挙げた問題を回避するため、先立って URL 付きツイートと周辺ツイートという 2 つの概念の導入を行う。あるユーザーが投稿した全てのツイートを投稿日時の順に並べたツイートの列を、ユーザータイムラインと呼ぶ。ユーザータイムライン上に存在する、投稿テキストに URL を含むようなツイートを、URL 付きツイートと定義する (図 1)。また、ユーザータイムライン上の任意の URL 付きツイートを選んだとき、選ばれた URL 付きツイートの周辺に位置する前後 N 件ずつのツイートを、周辺ツイートと定義する。

4.2 問題設定

ここで、情報源推定の対象を URL 付きツイートと周辺ツイートに限定した場合について考える。これによって、一部のツイートに対する情報源の推定を行うことができなくなる一方で、Web ページの総量による計算コストの問題が回避できる。つまり、URL 付きツイートに含まれる URL が示す Web ページが、周辺ツイートの情報源であるか否かの推定のみを行えば良く、他の Web ページは推定に必要としないためである。また、Twitter では、URL 付きツイートを投稿したユーザーが、その URL が示すウェブページについての意見を述べるような利用法がしばしば見受けられる。周辺ツイートに、そのようなウェブページを情報源とするツイートが多く含まれているならば、情報源を推定する対象として好ましい。

[2]RoyalPingdom http://royal.pingdom.com/2013/01/16/

そこで、URL 付きツイートと周辺ツイートに対する予備調査を行った。まず、無作為に抽出した、投稿者の異なる URL 付きツイート 1,067 件について、それぞれ $N = 5$ の周辺ツイートの収集を行った。ある URL 付きツイート 1 件に対応する周辺ツイート 10 件のいずれかのうちに、URL 付きツイートに含まれる URL が示す Web ページが、情報源であるようなものが含まれる件数を集計した。その結果、URL 付きツイートのおよそ半数となる約 47%(502/1,067 件) において、URL 付きツイートに含まれる URL が示す Web ページを情報源とするような周辺ツイートが存在することを確認した。予備調査の結果から、周辺ツイートには推定可能なツイートが高い割合で含まれており、同時に、情報源の推定が不可能なツイートの多くを除外することが可能であることがわかる。そこで、本研究では情報源推定課題を URL 付きツイートに含まれる URL が示す Web ページが、周辺ツイートの情報源であるか否かを推定する課題と定義する。

5 提案手法

5.1 テキストセグメンテーションによる情報源推定

本研究では、4 節で述べた情報源推定課題に対して、文書中の話題の境界を推定する技術であるテキストセグメンテーション手法を要素技術とした推定手法を提案する。処理の概要を図 2 に示す。URL 付きツイートとその周囲 N 件の周辺ツイートを投稿時間順に並べたツイートの列が、ひとつの処理単位となる。このツイート列を 1 つの文書と見なしてテキストを結合する。結合した文書に対して、テキストセグメンテーションを行うことで、ツイートを話題ごとに分割する。ただし、出力されるセグメントの境界はツイートとツイートの境界のみに存在するものとして、1 件のツイートの内部でセグメントの境界を作らないものとする。分割処理により、URL 付きツイートと周辺ツイートのツイート列は、いくつかの話題ごとにまとまったセグメントとして分割される。ここで、URL 付きツイートと同じセグメントに分割された周辺ツイートは、URL 付きツイートと同じ話題の内容を含むものと考えられる。そのため、URL 付きツイートと同じセグメントに分割された周辺ツイートの情報源については、URL

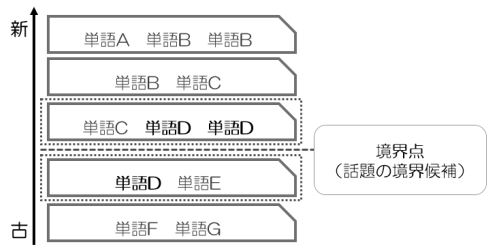

図 3: TextTiling 法の窓枠

付きツイートの URL が示す Web ページであると推定する。

5.2 TextTiling 法の適用

本稿では、テキストセグメンテーション手法として、Hearst の TextTiling 法 [1] を使用する。TextTiling 法では、まず、文書内の文と文の境界となる、ある一点を境界点に設定する。境界点 (話題の境界候補) の左右の文脈から、それぞれ同数の単語 (形態素) n 個を含むような窓枠 w_l, w_r を設ける。左右の窓枠内に出現する全ての単語 t について出現回数を数え上げ、両窓枠の結束度スコア $sim(w_l, w_r)$ を求める。結束度のスコアは左右の窓枠内に同じ単語が出現する傾向がある境界点において大きくなり、逆に同じ単語があまり出現しない傾向がある場合に小さくなる。最大値は左右の窓枠で単語が全く同じように出現する場合の 1 であり、最小値は同じ単語が全く出現しない場合の 0 である。なお、素性として用いる単語は、形態素解析により名詞一般あるいは未定義語と解析された形態素のみを用いる。TextTiling 法は、文書中における話題の境界を文レベルでセグメントに分割する手法である。しかし、ツイートの情報源を推定するという本研究の目的を考えたとき、分割されるセグメントはツイートレベルであることが望ましい。そこで、本研究では、1 件のツイート中で話題の変化は起こらないという仮定をおき、ツイート内の文同士の境界を無視し、セグメンテーションを行う際の分割単位をツイート単位とする。また、同様の理由で、窓枠について単語数を基準とせず、ツイート数を基準とした窓枠を考える。窓枠がツイート 1 件である場合の例を図 3 に示す。左右の窓枠に含まれる単語数が異なる場合、N_{w_l} と N_{w_r} をそれぞれ左右の窓枠に出現する単語の数として、以下の式により、正規化されたスコアを求める。

$$sim(w_l, w_r) = \frac{\sum_{k=1}^{m} \frac{f(t_k, w_l)}{N_{w_l}} \frac{f(t_k, w_r)}{N_{w_r}}}{\sqrt{\sum_{k=1}^{m} (\frac{f(t_k, w_l)}{N_{w_l}})^2 \sum_{k=1}^{m} (\frac{f(t_k, w_r)}{N_{w_r}})^2}} \quad (1)$$

ここで、m は左右の窓枠に含まれる単語の種類数であり、$f(t_k, w_l)$ と $f(t_k, w_r)$ は単語 t_k が左窓、右窓内それぞれに出現する頻度である。このスコアを文書の頭から 1 文ごとに境界点を移動させながら求めていく。スコアの値が極小となるような境界点のスコアと、その左右に存在する極大点におけるスコアの値の差の合計 d が以下の閾値 d_{th} を超えた場合、その境界点を話題の境界と判断する。ただし、$\bar{S}$ は結束度スコアの平均であり、 σ は標準偏差である。

$$d_{th} = \bar{S} - \sigma/2 \quad (2)$$

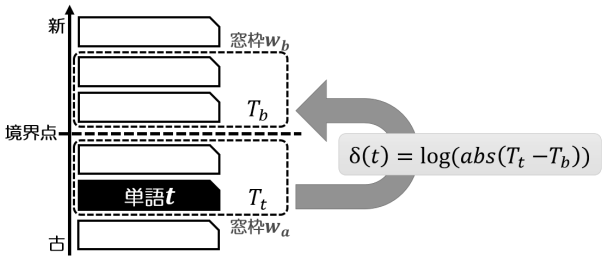

図 4: Web ページの文書内情報の利用

5.3 TextTiling 法の拡張

5.3.1 Web ページの文書内情報の利用

ツイートの本文は一般にテキスト長が短く、テキストセグメンテーションを行うために必要な単語の出現数が不十分である可能性が考えられる。そこで、ユーザータイムラインの文書情報に加え、URL 付きツイートに含まれる URL が示す Web ページの文書を用いたテキストセグメンテーションを行う。Web ページに含まれる文書内情報として、Web ページのタイトルと、Web ページの記事本文を抽出し、抽出したテキストを URL 付きツイートに含まれる文書の後方に順に結合する。その後、結合したテキストをテキスト長の長いひとつのツイートと見なし、テキストセグメンテーションを行う。

5.3.2 投稿時間情報の利用

境界点において隣接するツイート同士の投稿時間の差を投稿間隔として、結束度スコアに組み込む。一般に、投稿間隔が離れているほどツイート同士の話題の結束性は低くなることが考えられる。そこで、投稿間隔によるペナルティ δ を単語の出現回数から減算する。図 4 に例を示す。あるツイートに含まれる単語 t が含まれる窓枠を w_a、w_a に対して境界点の反対側に位置する窓枠を w_b とおく。窓枠 w_b に含まれるツイートのうち最も境界点に近いツイートの投稿時間を T_b とおくと、単語 t の出現回数に課されるペナルティ $\delta\ (t)$ は、単語 t が含まれるツイートの投稿時間 T_t を用いて、以下のように表すことができる。

$$\delta\ (t) = \log(abs(T_t - T_b)) \quad (3)$$

ただし、単語の出現回数の下限を 0 とするため、ペナルティの減算により単語の出現回数が負の値を取る場合、0 の値をとるものとする。

表 1: 評価データの概要

種類	件数
URL 付きツイート	318 件
周辺ツイート	3,180 件
情報源を持つ周辺ツイート	317 件/3180 件

6 実験

6.1 評価データ

Yahoo!Japan の運営するスポーツニュースサイト「スポーツナビ」[3] のドメイン名「sportsnavi.yahoo.co.jp」がツイート内に記述されている URL 付きツイート 318 件を無作為に抽出した。抽出した URL 付きツイートについて、$N = 5$ の周辺ツイートを収集した。収集した評価用データに対し、人手によるラベル付け作業を行った。2 名の作業者が独立にラベル付けを行い、両名が「情報源である」と判定した 317 件を URL 付きツイートに含まれる URL が示す Web ページが情報源であるものとして採用した。一致率を示す統計量である κ 値は 0.782 であった。評価データの概要を表 1 に示す。

6.2 比較手法

提案手法と比較を行う 3 つの手法について説明を行う。まず、全ての周辺ツイートの情報源が、URL 付きツイートの URL が示す Web ページであると見なす手法を、ナイーブな手法と呼ぶ。ナイーブな手法では、再現率が 1 になる。次に、Abel ら [2] の、TF-IDF を用いる手法を TF-IDF 法と呼ぶ。TF-IDF 法では、あるツイート内に出現する全ての単語について、Web ページとの TF-IDF を求め、その合計をツイートと Web ページの類似度スコアとする。最後に、ツイートに付与されたハッシュタグを利用する方法をハッシュ法と呼ぶ。ハッシュ法では、周辺ツイートが URL 付きツイートと同一のハッシュタグを含むとき、周辺ツイートの情報源が、URL 付きツイートの URL が示す Web ページであると推定する。

6.3 評価・実験結果

実験の結果を図 5 に示す。図から、テキストセグメンテーション手法を用いた情報源推定手法がベースラインの手法をいずれも上回っていることが確認できる。また、Web 文書情報の拡張と、投稿時間を用いた拡張手法の両方で、「TextTiling」と比較して、さらなる精度の改善が見られる。Web 文書の情報と投稿時間の素性が情報推定課題に有用であるという結果が得られたと言える。「Web 文書+投稿時間」は、Web 文書情報の拡張と、投稿時間の両方を単純に組み合わせた拡張手法である。Web の文書情報をツイートに結合し、さらに投稿間隔によるペナルティを結束度スコアに組み込む。しかし、この統合手法は、Web 文書情報と投稿時間のそれぞれ単独の拡張手法と比べて、精度の改善が見られていない。組み合わせの際にパラメータの調整など、何らかの最適化が必要であると考えられる。

[3] スポーツナビ http://sportsnavi.yahoo.co.jp/

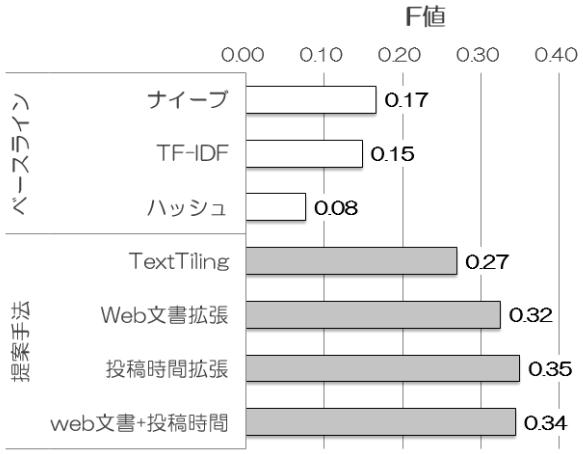

図 5: 性能の評価

7 おわりに

本研究では、URL を含むツイートとその周辺のツイートに着目し、テキストセグメンテーション手法を用いた情報源推定課題の解決手法を提案した。評価実験においては、提案手法の有効性を示したが、実用レベルでの満足な精度を達成したとは言い難い。提案手法は、同様の話題のツイートは連続して投稿されるという仮定に基づいたものである。しかし実際は、同じ話題のツイートの間に、全く関係のない話題のツイートや、ユーザー同士の会話などが挟まれる場合もあり、テキストセグメンテーション手法の応用部分に改善が必要である。投稿時間などの素性は有用であれど、有効に活用できたとは言えず、情報源推定課題に適する手法の更なる検討が必要である。また、今回は教師あり学習を用いず計算コストの少ない TextTiling 法を用いたが、教師データの用意と機械学習によるアプローチも検討課題である。

参考文献

[1] Hearst, Marti A. "TextTiling: Segmenting text into multi-paragraph subtopic passages." Computational linguistics 23.1(1997), pp33-64, 1997.

[2] Abel, Fabian, et al. "Semantic enrichment of twitter posts for user profile construction on the social web." The Semanic Web: Research and Applications. Springer Berlin Heidelberg, pp375-389, 2011.