Vol.2015-DBS-162 No.8
2015/11/26

閲覧期間を考慮したトピックグラフに基づくTwitterの見落とし情報抽出手法

大原 啓詳[1,a)]　鈴木 優[2,b)]　灘本 明代[3,c)]

概要：代表的なマイクロブログの一つである Twitter では，閲覧者が興味のある情報発信者をフォローし，ツイートを閲覧者のタイムラインに表示することで，閲覧者自身が比較的興味のある情報を集約することができる．しかしフォロー数の多い閲覧者のタイムラインには膨大な量のツイートが表示されるため，タイムラインを閲覧できない期間が生じると，有益な情報が見落されてしまうことが考えられる．そこで本研究ではタイムラインの閲覧期間と，ツイートのトピック，そしてツイートのトピックから構築されるトピックグラフに基づき，見落とした情報をわかりやすい形式で提示する手法を提案する．

1. はじめに

近年，様々な情報を手軽に共有する手段として，マイクロブログが普及している．特に世界中に多くのユーザを持つ代表的なマイクロブログの一つである Twitter には，多様かつ膨大な量の情報が存在している．Twitter ではある情報の閲覧者が，自分の興味のある情報の発信者をフォローすることにより，その情報の発信者の投稿したツイートを閲覧者自身のタイムライン上で閲覧することができる．本研究ではこの閲覧者によりフォローされている情報の発信者をフォロイーと呼ぶ．閲覧者はより多くのフォロイーをフォローするほど，多くのツイートを自身のタイムライン上で閲覧することが可能となる．しかしながら閲覧者がタイムラインを常に閲覧しているということは困難であり，特にフォロイーの多いユーザは，タイムラインを閲覧していなかった期間に多くのツイートがタイムライン上に表示され，これらを見落としてしまうと考えられる．また，見落としたツイートには閲覧者にとって重要な話題や目新しい話題に関するツイートが存在する一方で，あまり興味のない話題についてのツイートやすでに知っている情報についてのツイートも混在していると考えられる．見落としたツイートの数が膨大であった場合，興味のあるツイートをこれらの中から再発見するには，見落とした期間のタイムラインを遡り，ツイートの内容を確認する必要があり，閲覧者にとって負担の大きい作業であると考えられる．そこで我々は，閲覧者が閲覧できなかった期間中に投稿されたタイムライン上の見落としたツイート群から，有用な情報抽出し提示する手法を提案する．本研究では閲覧できなかった期間を見落とし期間と呼び，見落とし期間中に含まれる閲覧者にとって有用な情報を見落とし情報と呼ぶ．本論文では見落とし情報抽出のはじめの一歩として，見落とし期間におけるツイートの話題について，閲覧者が既に知っている話題であるか，全く知らない話題であるか，またその話題の粒度構造に着目し，これらを閲覧者にわかりやすい形でグラフ化し提示する手法を提案する．これにより閲覧者は見落とし期間中のツイートから，見落とし情報を発見する際の負担の軽減が可能となる．

以下本論文では 2 章で関連研究を，3 章で見落とし情報と話題の粒度の定義を行い，4 章で話題の粒度基づくトピックグラフの生成と閲覧期間にも基づく話題の分類手法について 5 章で話題の粒度についての実験について述べ，6 章でまとめと今後の課題について述べる．

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りそ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，本研究ではより詳細な情報を取得するためにフォロ

[1] 甲南大学大学院　自然科学研究科
Konan University, Kobe, Hyogo 658–0072, Japan
[2] 奈良先端科学技術大学院大学 情報科学研究科
NAIST, Ikoma,Nara 630–0192, Japan
[3] 甲南大学　知能情報学部
Konan University, Kobe, Hyogo 658–0072, Japan
[a)] m1424003@center.konan-u.ac.jp
[b)] ysuzuki@is.naist.jp
[c)] nadamoto@konan-u.ac.jp

イー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．本研究ではツイート群の文章による要約ではなく話題をトピックグラフとして提示することで，ユーザが話題の構造などを理解しやすくすることを目的としている点が異なる．

またツイートの話題の分類に関して西田ら [4] は，閲覧者が着目している話題に関するツイートを，情報圧縮を行った際に着目している話題と，その他の話題のどちらに分類されやすいかという点に着目し，単語や文章に依存しないツイートの特定の話題への分類を行っている．西田らの手法は閲覧者が着目している話題が明確に定まっている場合に，着目している話題に関する有益なツイートを集約する際には有効である．一方，本研究ではフォロイーの話題の多様さにも着目し，フォロイーのツイート全体に含まれる複数の話題へと分類することを試みている．これにより本研究では，閲覧者の興味が明確ではない場合などでも，閲覧者が有益な情報を容易に発見できる仕組みが実現できると考える．Michelson ら [6] は 1 アカウントを対象として，その投稿内容のトピック抽出を Wikipedia のカテゴリを利用し行っている．しかしながら Michelson らの研究では時系列的な話題の変化等には着目していない点が我々の研究とは異なる．

ツイートの話題と時系列的な話題の変化に関する研究として，糸川ら [9] は特定のトピックに対する特徴語の時間経過による変化などから，話題追跡を行っている．糸川らの研究は対象を特定トピックに限定しているが，本研究は複数のトピックの混在する見落としたタイムラインを対象とし，話題の抽出を行っている点で異なる．藤野ら [3] はダイナミックトピックモデル [7] によるツイートのトピック抽出と，投稿内容の時系列変化に着目しユーザの相関関係の発見を行っている．藤野らの研究ではツイートから得られたユーザの特徴をもとに，ユーザのタイプ分類とユーザ間の相関関係を発見することを目的としている．それに対し本研究は時系列的な話題変化や，話題の多様性から，見落とされた話題，話題の持つ概念構造を発見することを目的としている点が異なる．また，時系列を考慮したトピックモデルに関する研究として佐々木ら [5] は Twitter-LDA[8] を話題の時間発展を考慮することで拡張した手法により，ツイートのトピックモデル化を試みている．本研究では，時間，ツイートのトピックモデル抽出のみならず，トピック間の概念構造に対しても着目している点が佐々木らの研究とは異なる．

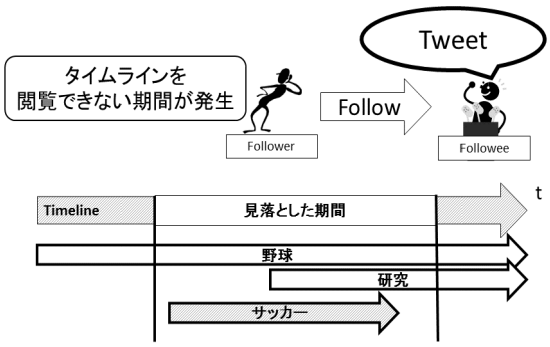

図 1: 既知の話題と未知の話題の分類

3. 見落とし情報と話題の粒度の定義

3.1 閲覧期間に基づく見落とし情報

閲覧できなかった期間中の話題について，本研究では “未知の話題” と “既知の話題” という 2 種類のタイプに分類する．以下にそれぞれのタイプの定義を示す．

- 既知の話題
  見落とし期間中に存在する話題のうち，閲覧者が閲覧していた期間中にもその話題のツイートが存在しており，その話題について閲覧者が部分的に閲覧したことのある話題.
- 未知の話題
  見落とし期間中に存在する話題のうち，見落し期間にのみ出現する話題であり，閲覧者が閲覧したことのない話題.

既知の話題，未知の話題の分類について図 1 に示す．図 1 では，“野球”，“研究” という話題は閲覧していた期間と見落とした期間の両方に含まれる話題であるため，“野球”，“研究” に関する話題は既知の話題となる．一方で “サッカー” に関する話題は見落とした期間にのみ含まれる話題である．この時，話題 “サッカー” は未知の話題となる．

3.2 話題の粒度

話題の粒度は見落とし期間中に投稿されたツイートの話題構造の把握と，閲覧期間に基づく話題の分類において重要な要素の一つである．

例えば見落とし期間に投稿された話題が図 2 に示すように，“イチロー”，“田中将大”，“坂本勇人”，“鳥谷敬”，“マツダスタジアム”，“ベイスターズ” についてのものであったとする．この時，“イチロー”，“田中将大” についての話題はいずれも “日本人メジャーリーガー” という，より大きな粒度の話題でまとめることが可能である．また，これらの話題に “坂本勇人”，“鳥谷敬” といった話題も含めて考えると，さらに大きな “野球選手” という粒度でまとめることも可能である．さらに “イチロー”，“田中将大”，“

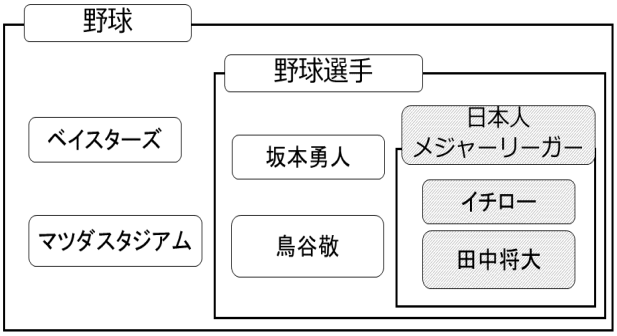

図 2: 話題の粒度構造

坂本勇人”，“鳥谷敬”，“マツダスタジアム”，“ベイスターズ” のすべての話題はいずれも “野球” という粒度でまとめることも可能である．粒度構造を提示することで，閲覧者は複数の話題間の結びつきを理解することができる．

本研究では既知の話題，未知の話題への分類を考える際に，話題の持つ粒度構造にも着目する必要があると考える．例えば上記の例において閲覧者にとっての未知の話題が “イチロー”，“田中将大” だけであった場合，閲覧者にとって “日本人メジャーリーガー” の話題全体が未知の話題であるといえる．一方，“野球選手” や “野球” といった粒度の話題については，“坂本勇人”，“鳥谷敬” といった既知の話題が含まれるため，これらも既知の話題であると判断できる．

4. 見落とし情報の抽出手法

本研究では，フォロイーのツイートの話題構造と，閲覧期間に基づき分類したタイプを，見落とし情報抽出のためのトピックグラフとして生成し提示する．以下に提案手法の具体的な手順を示す．

(1) 見落とし期間とその前後のフォロイーのツイートを抽出する
(2) フォロイーのツイートを話題ごとにクラスタリングする
(3) 各クラスタの話題の粒度構造に基づきトピックグラフを生成する
(4) タイムラインの閲覧期間に基づき，話題を既知の話題と未知の話題に分類する
  4–1 各ツイートの投稿時間を抽出する
  4–2 それぞれの話題に含まれるツイートの投稿時間から，話題を未知の話題と既知の話題に分類する
(5) トピックグラフに既知の話題，未知の話題の分類を反映する

提案手法のシステムフローについて図 3 に示す．

4.1 ツイートのクラスタリング

まず見落とし期間とその前後のツイートを取得し，ツイートを話題ごとにクラスタリングするために各ツイートから特徴語の抽出を行う．本研究では各ツイート中の名詞と未知語を特徴語として抽出し用いる．ここで未知語を用いる理由としては，Twitter における新語への対応を行うためである．また，取得した名詞と未知語のうち，“試合”，“選手” のような，一般的名詞や，意味の判断ができない未知語などを取り除くため，固有名詞以外の語については，Wikipedia の全記事情報に対して検索を行い，ヒットした記事数が 1 件以上，100 件未満の単語のみを用いる．また，名詞や未知語が連続して用いられる場合には，それらの語を一つの複合名詞として用いる．各単語ごとの重みについてはツイート 1 件を 1 文書とした場合の IDF 値を用いるものとする．

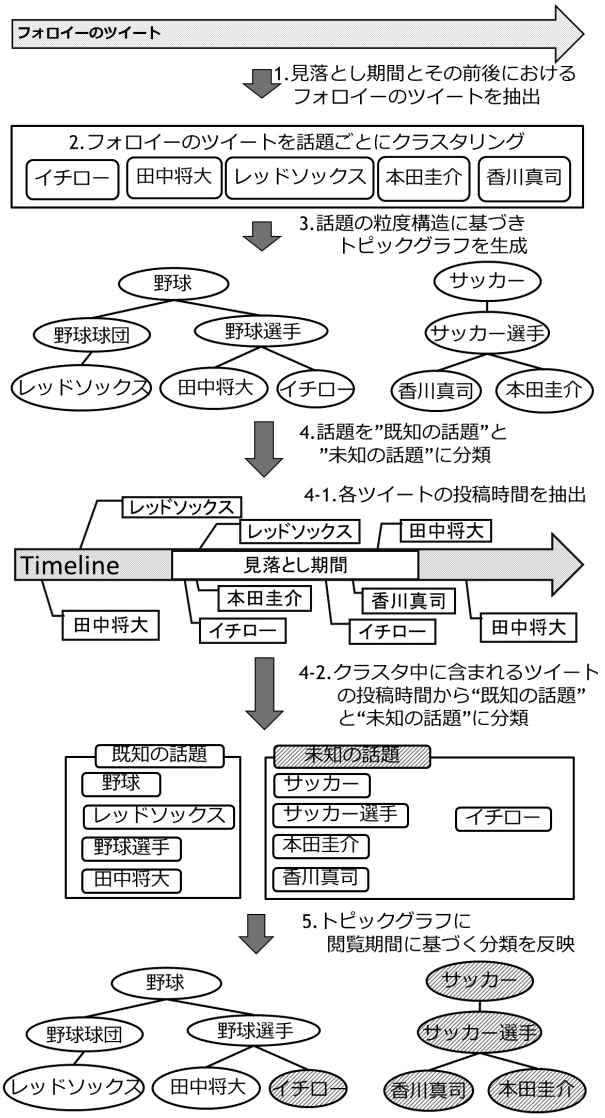

図 3: 提案手法のシステムフロー

ツイートのクラスタリング手法には CLUTO[11] や Bayon[*1] といったクラスタリングツールで用いられているクラスタリング手法である Repeated Bisection 法 [12] を

[*1] https://code.google.com/p/bayon/

用いる．この手法は K-means 法の拡張手法の一つであり，短文に対してもある程度の対応が可能である [10].

また Repeated Bisection 法はハードクラスタリングの手法であるため，本来いずれのクラスタにも属さないようなツイートが不適切なクラスタに分類されることや，内容が極端に疎なクラスタが生成される場合がある．そこで，本研究ではクラスタの中心ベクトルと，ツイートの特徴ベクトルのコサイン類似度の値が小さいツイートについては，所属度の低いツイートとしてクラスタリング結果から削除する処理を行う．本論文ではクラスタリングの結果として得られた各クラスタを “話題クラスタ” と定義する．

4.2 トピックグラフの生成

4.2.1 最小トピックグラフの生成

本研究では各話題クラスタの中心ベクトルを構成する特徴語を，話題クラスタのトピックとして，Wikipedia のカテゴリの上位下位概念構造を用いることによりトピックグラフの生成を行う．

まず，各話題クラスタとその上位概念から成るトピックグラフを生成する．本研究ではこのトピックグラフを最小トピックグラフ STG_j と定義する．ここで，j は話題クラスタの id である．

最小トピックグラフ STG_j を生成するために，まず各話題クラスタの中心ベクトルを構成する特徴語を話題クラスタのトピック c_i とする．この時，ある話題クラスタのトピックが複数の特徴語であった場合，その特徴語それぞれを話題クラスタのトピックとして，トピックグラフを生成する．例えばある話題クラスタ j が “イチロー”，“メジャーリーグ” の 2 語によって構成された中心ベクトルを持つクラスタが存在したとする．この場合，話題クラスタ j については “イチロー”，“メジャーリーグ” の両方を話題クラスタのトピックと考え，それぞれの最小トピックグラフを生成する．次に全話題クラスタのトピック集合 $C = \{c_1, c_2, \ldots, c_n\}$ の要素 c_x において，Wikipedia の c_x をタイトルとする記事に付与されたカテゴリを Wikipedia のカテゴリリンクデータベースより取得し，c_x の上位概念 $s_{x1}, s_{x2}, \ldots, s_{xi}, \ldots, s_{xm}$ とする．この時，“Wiki” や “〜のスタブ”，“〜であるユーザ” といった Wikipedia の内容編集に関するカテゴリについては上位概念になりえないため削除する．また “存命人物” や “〜語圏の人名” といった，重要な意味を持たないカテゴリについても削除する．また，各カテゴリ s_{xi} の上位カテゴリは，s_{xi} よりもさらに大きな粒度における c_x の上位概念として扱う．そして c_x をトピックとする話題クラスタ j をリーフノードとして，話題クラスタ j と，c_x の上位概念から成る最小トピックグラフ STG_j を生成する．本論文では各 c_x について 2 階層上位のカテゴリまでを用いて STG_j を生成する．この手順をトピック集合 C に含まれるすべての要素に対して行うことで，すべての話題クラスタに対応した最小トピックグラフの生成を行う．図 4 に最小トピックグラフ STG_j について示す．

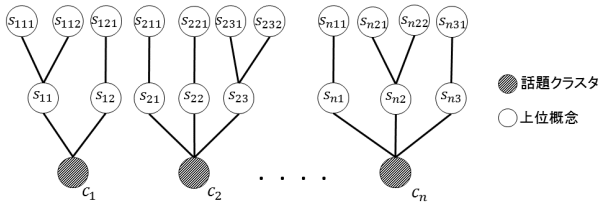

図 4: 最小トピックグラフ

最小トピックグラフ生成の具体例として，“イチロー”，“田中将大” というトピックの話題クラスタが 2 つ存在した場合，c_1 は “イチロー”，c_2 は “田中将大” となる．そこでまず，c_1 に対して STG_1 を生成するために Wipkipedia 上の “イチロー” の記事に付与されたカテゴリを取得する．そしてその結果 “盗塁王 (MLB)” や “日本人メジャーリーガー” といったカテゴリが抽出された場合，これらを s_{11}, s_{12} として c_1 の上位概念として付与する．本論文では 2 階層上位の概念を取得するため，さらに s_{11} “盗塁王 (MLB)” のさらに上位のカテゴリを抽出することで得られた “MLB のタイトル・表彰” を s_{111} とする．同様に c_2 の “田中将大” についても上位カテゴリとして抽出された “日本人メジャーリーガー”，“オリンピック野球日本代表選手” を上位概念 s_{21}, s_{22} として付与する．

4.2.2 トピックグラフの結合

次に生成された最小トピックグラフ同士を，共通するトピックまたは上位概念により結合する．本研究で考えられる結合のタイプには上位概念同士による結合と，上位概念とトピックの結合が考えられる．図 5 にトピックグラフの結合について示す．(a) の例として STG_1 の各ノードについて c_1 が “イチロー” で s_{12} が “日本人メジャーリーガー”，STG_2 の各ノードについて c_2 が “田中将大” で s_{21} が “日本人メジャーリーガー” であった場合，STG_1 と STG_2 を共通した上位概念ノードである s_{12} と s_{21} により結合するまた (b) の例として，STG_1 の各ノードについて c_1 が “プロ野球選手” で s_{12} が “プロ野球” であり，STG_2 の各ノードについて c_2 が “プロ野球” であった場合，STG_1 と STG_2 を上位概念ノード s_{21} とリーフノードのトピック c_2 により結合する．

結合処理を行ったトピックグラフのノードのうち，話題クラスタのトピックによりラベル付けされたノードを “トピックノード”，話題クラスタのトピックではない上位概念語によりラベル付けされたノードを “概念ノード” と定義する．本論文では上記の手法により生成されたトピックグラフにより，見落とし期間に投稿された話題の粒度構造を可視化する．

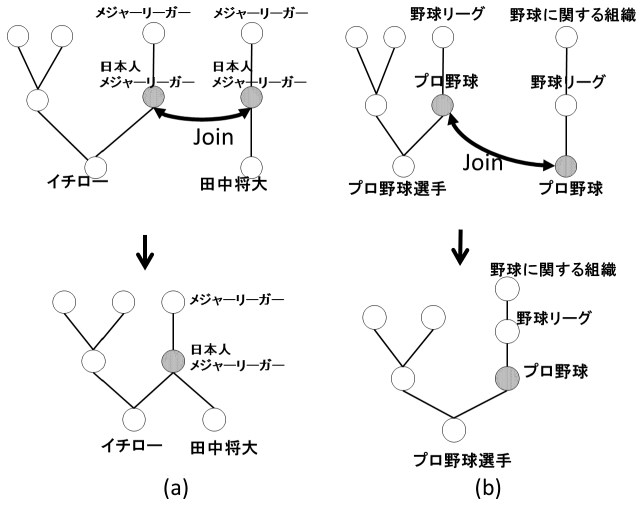

図 5: トピックグラフの結合

4.3 投稿時間に基づく話題のタイプ分類

次に，各話題クラスタのトピックについて，ツイートの投稿時間とタイムラインの閲覧期間から，未知の話題と既知の話題の 2 つのタイプへの分類を行う．

まず閲覧者の見落とし期間の時間情報と，各話題クラスタに含まれるツイートの投稿時間を取得する．そして，各ツイートが見落とし期間に投稿されたツイートであるか，タイムラインを閲覧していた期間に投稿された閲覧済みのツイートであるかを分類する．ここで，ある話題クラスタに見落とし期間中に投稿されたツイートと，閲覧済みのツイートの両方が含まれる場合，この話題クラスタのトピックを既知の話題とする．一方，ある話題クラスタに含まれるツイートが，見落とし期間に投稿されたツイートだけである場合，この話題クラスタのトピックを未知の話題とする．

例えば図 6 の場合，トピックが “イチロー” の話題クラスタは閲覧済みのツイートと見落とし期間に投稿されたツイートの両方を含んでいるため，“イチロー” は既知の話題となる．一方で，トピックが “田中将大” の話題クラスタは，クラスタ中のツイートがすべて見落とし期間に投稿されたツイートであるため，“田中将大” は未知の話題となる．

この手法により各トピックについて既知の話題であるか未知の話題であるかを分類し，4.2 節で生成したトピックグラフ上のトピックノードに反映する．

5. 実験

本論文では見落とし情報の抽出に必要な，クラスタリング手法とトピックグラフの結合手法について，提案手法が適切であるかを確かめるために実験を行った．

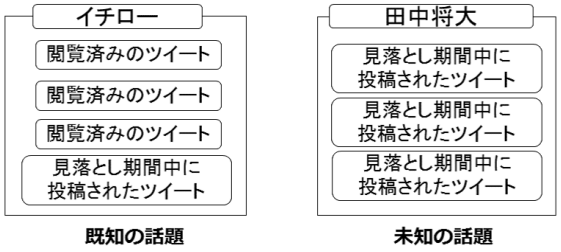

図 6: 投稿時間に基づく話題のタイプ分類

5.1 ツイートのクラスタリングの実験

実験条件

以下の条件にて実験を行った．

- データセット：
  対象アカウント数 : 5 ユーザ
  ツイート件数 : Twitter REST API[*2] により取得した，1 ユーザあたり 1000 件のツイート，計 5000 件．

- 各ユーザの特徴：
  ユーザ A : 主に “アニメ”，“ゲーム” について投稿．
  特定のタイトルではなく，多様なアニメやゲームについての投稿を行っている．
  他のユーザに比べて，短いツイートや重要な意味を持たないネタツイートなどが多い．
  ユーザ B : 主に “アニメ”，“アイドル” についての投稿．
  特に興味のある特定のアニメやアイドルグループについての投稿が多い．
  ユーザ C : 主に “サッカー” について投稿．
  特に試合情報や選手情報に関する投稿が多い．
  他のユーザに比べて 1 ツイートあたりの文字数や単語数が多い．
  ユーザ D : “実体験”，“日常生活”，“ゲーム”，“野球”，“サッカー” など様々な話題について投稿．
  “ゲーム” “野球”，“サッカー” に関するツイートが比較的多い．
  ユーザ E : “音楽”，“ラジオ” についての投稿が比較的多い．
  地名や施設名などがツイート中に多く含まれる．

- 諸条件：
  クラスタリングツール : Bayon
  クラスタ数 : Bayon の分割ポイントにより自動で決定
  Bayon の分割ポイント : 1.0
  クラスタリング結果に対する cos 類似度の閾値 : 0.5

[*2] https://dev.twitter.com/rest/public/

形態素解析器:汎用日本語形態素解析エンジン MeCab[*3]
辞書データ:IPA 辞書を日本語版 Wikipedia のページタイトルと，はてなキーワード[*4] の単語を固有名詞として追加した辞書

実験内容

提案手法により，ユーザごとにクラスタリングを行い，クラスタリング結果の各クラスタに含まれるツイートに対して，適切なクラスタに分類されているかを人手で判断し適合率を求めた．

結果と考察

表 1 に各ユーザについて，提案手法で分類可能であったツイート数，クラスタ数，含まれるツイート数が最大となったクラスタ中のツイート数，適合率を示す．また各ユーザのクラスタリング結果について，クラスタの中心ベクトルを構成する特徴語のポイントが特に大きい上位 5 クラスタについて，クラスタの ID，中心ベクトルを構成する特徴語 2 件とそれぞれの特徴語のポイントを表 2 に示す．なお，表中のクラスタ番号が最終結果のクラスタ数を上回っているものが存在するのは，cos 類似度によるツイートのフィルタリングを行う前の数値を用いているためである．

表 1: クラスタリングの結果

	ツイート数	クラスタ数	最大クラスタ中のツイート数	適合率
ユーザ A	325	97	14	0.708
ユーザ B	353	47	23	0.912
ユーザ C	583	171	7	0.762
ユーザ D	487	124	14	0.891
ユーザ E	469	121	12	0.938
平均	443.4	112	14	0.842

実験の結果，表 1 からいずれのユーザについても 0.7 以上の比較的高い適合率が得られていることがわかる．特にユーザ B，D，E について高い適合率が得られることが確認された．表 2 から，これらのユーザはツイート中に人名や作品名，楽曲名，地名といった特徴的な固有名詞が多く含まれていたため，正確にクラスタリングを行うことができたと考えられる．しかしながらユーザ A とCについては適合率が他の 3 ユーザに比べて低くなった．これは，まずユーザAについてはツイートが短く，略語やスラングなどがツイート中で多用されていたことにより，クラスタリングが困難であったことが原因として考えられる．またユーザCについてはツイートの分量は他のユーザより長いが，ツイートの内容については 1 ツイートの話題が 1 つに定まっていない場合や，“興奮” や “関心” のように，Wikipedia の全文検索を用いた一般語の削除では取り除ききれなかった一般的な名詞が多く含まれていたためクラスタリングが困難であったことが原因として考えられる．

[*3] https://code.google.com/p/mecab/
[*4] http://d.hatena.ne.jp/keyword/

表 2: 各ユーザの話題クラスタの特徴語

(a) ユーザ A の各クラスタ

クラスタ番号	特徴語 1	ポイント	特徴語 2	ポイント
61	スピコア	0.999	トパゾ	0.023
36	ファミマ	0.978	寝冷え	0.211
65	コミケ	0.963	お疲れ様	0.103
53	??	0.962	ティア	0.157
41	ラーメン	0.962	ラーメン二郎	0.160

(b) ユーザ B の各クラスタ

クラスタ番号	特徴語 1	ポイント	特徴語 2	ポイント
51	ISUCA アニメ化楽しみッ	0.986	コヒメ	0.121
39	胸開きタートルネック	0.986	エリザベス	0.067
27	anisama	0.985	客席カメラ	0.087
13	オトメ	0.969	ISUCA アニメ化楽しみッ	0.201
74	ISUCA	0.959	イスカ-」怒涛	0.186

(c) ユーザ C の各クラスタ

クラスタ番号	特徴語 1	ポイント	特徴語 2	ポイント
10	興奮	0.876	エロチシズム	0.340
14	bundesliga	0.856	役目	0.333
17	大阪	0.839	ヒトゴト	0.164
19	gekisaka	0.836	ゼニト	0.176
21	関心	0.831	新天地	0.287

(d) ユーザ D の各クラスタ

クラスタ番号	特徴語 1	ポイント	特徴語 2	ポイント
32	タイガース	0.974	セ・リーグ	0.175
15	デール	0.971	ツムツム	0.239
93	甲子園	0.971	島袋	0.0962
151	ゴメス	0.970	打点	0.136
113	バイト	0.969	休憩時間	0.068

(e) ユーザ E の各クラスタ

クラスタ番号	特徴語 1	ポイント	特徴語 2	ポイント
148	ミナホ	0.980	カルボナーラ	0.099
123	新潟	0.962	ドシテ	0.084
129	福岡	0.962	fmfukuokafm	0.136
93	名古屋	0.959	きれいな空	0.113
14	ツイート	0.958	愛してる	0.285

これらの結果から，本論文で提案するツイート中の特徴語と，クラスタリング時に与えたパラメータを用いる，Repeated Bisection 法によるクラスタリングは，特徴語の選択などに改良が必要な点はあるが，ツイートを話題ごとにクラスタリングする際にはある程度有効であることが確認された．

5.2 トピックグラフの結合に関する実験

トピックグラフの生成について，最小トピックグラフの結合手法が適切であるかを確かめるために実験を行った．

実験条件

以下の条件で実験を行った．

- ツイート数：話題の判別が容易なツイート 2335 件
- 話題クラスタ数：152 クラスタ
- 評価を行う被験者：20 代の男女 9 名
- 諸条件：

表 3: 実験 5.2 の結果の一例

共通する概念語	概念語の下位カテゴリ数	STG_A のリーフノードのトピック	STG_B のリーフノードのトピック	評価値の平均
J リーグクラブ	54	ベルマーレ, 湘南 BMW スタジアム	浦和レッズ, urawareds	1.89
アニメ	95	アニメ, NeversayNever	声優パラダイス, NeversayNever	1.78
マクロ経済学	79	プライマリバランス, セカンドレポート	マクロ経済学 , 市政報告会	1.56
ヘルプ	107	サッカー選手, 野球選手	Mamers, スケジュール	0.00
社会	149	テニス, コンディション	エネルギー, 山雅メンバー	0.00
商標登録	414	ソフトバンク, 国道 1 号線	キャンキャン, フラペチーノ	0.00

クラスタリング時のツールのパラメータなどは.1.1 節の諸条件と同様のものを用いた.
話題クラスタの中心ベクトルが複数の特徴語により構成される場合，特徴量が高い特徴語上位 2 件を話題クラスタのトピックとして用いた.

実験内容

具体的な実験の手順について以下に示す.

(1) 各トピックについて 2 階層上までの Wikipedia カテゴリを検索し，最小トピックグラフを生成する
(2) 提案手法に基づき，最小トピックグラフ同士を結合する．今回の実験では 2 つの最小トピックグラフを STG_A，STG_B と置き，2 つずつ比較を行い，トピックグラフの結合を行う
(3) 結合した 2 つのトピックグラフについて，それぞれのリーフノードのトピックと，2 つの最小トピックグラフに共通する概念ノードやトピックノードのラベルとなっているトピックを表示し，トピックグラフの結合が適切であるかを被験者 9 名がそれぞれ 3 段階（0，不適切である；1，少し適切である；2，適切である）で評価する

結果と考察

今回の実験では 299 件のトピックグラフの結合に対して評価を行った．実験結果のうち，評価の平均値が高かった結果と，評価の平均値が低かった結果について，それぞれの例を表 3 に示す．今回の実験では J リーグクラブ，アニメ，マクロ経済学といった概念語により結合されたトピックグラフについて，比較的高い評価が得られた．一方でヘルプ，社会，商標登録といった概念語で結合されたトピックグラフについては非常に評価が低かった．この結果から，J リーグクラブやアニメ，マクロ経済学のように具体的な内容がわかる概念語かつ，Wikipedia のカテゴリ構造において下位概念となるカテゴリの少ない概念語については比較的高い評価が得られ，ヘルプや社会，商標登録のように内容が漠然としており，Wikipedia のカテゴリ構造においても下位概念に多くのカテゴリを持つ概念語については，評価が低くなることが確認された．これらの特徴をもとに，今後は粒度構造を考慮する際に適切な上位概念となるような語がトピックグラフの生成時に用いられるように調節を行う予定である.

6. まとめと今後の課題

本論文では Twitter における有益な情報の見落としを解決することを目的として，タイムラインの閲覧期間とフォロイーの投稿したツイートの話題と，その話題の粒度構造に着目し，フォロイーのツイートの話題が未知の話題であるか既知の話題であるかを考慮したトピックグラフとして閲覧者に提示することにより，見落とした期間中の話題について閲覧者が容易に確認できる仕組みについての提案を行った．特に，トピックグラフ生成のために必要な，フォロイーのツイートの話題ごとへのクラスタリングと，話題の粒度を考慮したトピックグラフの結合手法について詳細の説明と，提案手法が適切かを確かめる実験を行った．実験の結果，クラスタリング手法については人名や地名など，話題が理解しやすい特徴語を含むツイートについては提案手法がある程度有効であることが確認された．また，トピックグラフの結合手法に関しては，具体的なないようがわかる概念語による結合については，適切に結合が行われていたが，漠然とした概念語による結合については，不適切な結合が行われていることが確認された.

今後の課題は，ツイートからの特徴語抽出部分の改善によるクラスタリング精度の向上や，トピックグラフの結合時に用いる上位概念語を適切に選択するためのロジックの再考．そして実際にトピックグラフを提示するインタフェースの実装と，適切な提示方式などについても考慮する予定である.

謝辞 本研究の一部は JSPS 科研費 26330347 及び，私学助成金 (大学間連携研究補助金) の助成によるものです．ここに記して謝意を表します.

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”. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.513-522, 2013.
[3] 藤野 巌, 星野 祐子, “ダイナミックトピックモデルを用いた Twitter ユーザーの時間相関特徴の発見手法について”.

第 7 回 Web とデータベースに関するフォーラム (WebDB Forum 2014)，2014.
[4] 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.
[5] 佐々木謙太朗, 吉川大弘, 古橋武, “Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案”, 情報処理学会論文誌 数理モデル化と応用（TOM）,Vol.7, No.1, pp.53-60, 2014.
[6] Matthew Michelson, and Sofus A. Macskassy. “Discovering users’ topics of interest on twitter: a first look”, Proceedings of the fourth workshop on Analytics for noisy unstructured text data, pp.73-80, 2010.
[7] D. M. Blei, J. D. Lafferty, “Dynamic topic models”, Proceedings of the 23rd international conference on Machine learning, pp.113-120, 2006.
[8] W. X. Zhao and J. Jiang and J. Weng, J. He and E. Lim and H. Yan and X. Li, ”Comparing Twitter and Traditional Media using Topic Models”, In Proceedings of the 33rd European Conference on Information Retrieval, 2011.
[9] 糸川翔太 , 白松俊 , 大囿忠親 , 新谷虎松, “時系列的話題追跡のためのツイートの特徴語を用いた探索的閲覧支援システムの開発”, 第 76 回全国大会講演論文集 2014(1), 107-108, 2014.
[10] 花井俊介, 灘本明代, “食材名をクエリとしたレシピ検索における酷似レシピクラスタリング”, 信学技報, vol. 114, no. 204, DE2014-31, pp. 47-52, 2014.
[11] G. Karypis, “CLUTO - A Clustering Toolkit”, Dept. of Computer Science, 2002.
[12] M. Steinbach and G. Karypis and V. Kumar, “A comparison of document clustering techniques”, In 6th ACM SIGKDD,World Text Mining Conference, 2000.