投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討

新谷　歩生† 　関　洋平†† 　佐藤　哲司††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学大学院図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s0711605@u.tsukuba.ac.jp, ††{yohei,satoh}@slis.tsukuba.ac.jp

あらまし　簡潔で思いつきの文章が次々と投稿されるマイクロブログでは，一つの話題が複数の記事に分割されて記述されることが多い．本研究では，マイクロブログからある話題について記述された複数の記事の塊を話題チャンクとして抽出する手法を提案する．提案法は，単純なテキスト処理では話題の連続性を抽出することが困難な，短い文章が連続的に投稿されるマイクロブログに対して，投稿の時間間隔を指標として適用する．提案手法の有効性を確認するために，Twitter から選定した 6 名を対象に，*Jaccard* 係数を基本とする評価手法を適用した．その結果，3 名のユーザにおいて単語の共起関係を指標とした抽出より投稿間隔を指標とした抽出が高い精度を示すことを確認した．また，共起語と投稿間隔を併用した抽出手法では，3 名のユーザにおいてそれぞれの手法を個別に適用した手法より高い精度を示すことを確認したので報告する．

キーワード　マイクロブログ，話題チャンク

A study of Topical Chunks Extraction from Micro-blogs based on the Contribution Interval

Ayumu SHINTANI†, Yohei SEKI††, and Tetsuji SATO††

† College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
1-2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
†† Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
E-mail: †s0711605@u.tsukuba.ac.jp, ††{yohei,satoh}@slis.tsukuba.ac.jp

Abstract　You contribute concise sentences to micro-blogs one after another. In micro-blogs, one topic is often divided into some articles. In this study, we propose the method of extracting Topical Chunks, the chunks of some articles described about same topic, from micro-blogs. It is difficult to extract continuousness of the topic only by simple text processing because so short sentences are contributed in a row, so we adopt the contribution interval as the key. We tested our method by experiments based on the *Jaccard* coefficient on six persons who use the Twitter. As a result, the method of extraction by contribution interval were higher extraction accuracy than that of by Co-occurrence word in three persons. Moreover, the method of extraction by both of Co-occurrence word and contribution interval were higher extraction accuracy than individual method.

Key words　Micro-blogs, Topical Chunk

1. はじめに

近年, Twitter(注1)に代表されるマイクロブログが注目を浴びている．国内の Twitter ユーザ数は，2010 年において前年比 1900 %増との報告(注2)もあり，日本国内で Twitter が急速に普及しているといえる．

従来のブログと異なり，マイクロブログは文字数制限が設定されている．このため，まとまった長い文章を記述することは

(注1)：http://twitter.com/

(注2)：http://www.netratings.co.jp/New_news/News06302010.htm

できないが，今していることや，考えていることを短い言葉で即座に手軽に書き込むことができるという特徴がある．また，ユーザは関心を持った他ユーザを選択し，フォロー関係を生成することで，その投稿記事を時系列順に次々と表示させ，自身のタイムラインとして閲覧することができる．そのためユーザ間でリアルタイムに知識や体験を共有することができる．

しかし, 簡潔で思いつきの文章が次々と投稿されるために, 一つの話題が複数の記事に分割されて記述されることが多い．例えば，あるレストランに行くことを述べ，次の記事で誰とどこのレストランに向かっているか述べ，その次の記事でレストランに着き店の内装について述べ，次に料理を食べていることを述べ，その後料理の感想を投稿する，ということがある．このような場合，話題を十分に理解するために関連する前後の記事を閲覧する必要が生じる．そのため, ユーザの話題の理解を容易にするために，同一話題の複数記事を集約して提示する手法が必要とされてきた．また，マイクロブログの記事を検索するサービスも活用されているが，検索条件に合致した記事だけでなく，その記事と関連する記事を出力したいという要求も高まっている．

目　　的

本論文では, マイクロブログからある話題について記述された複数の記事の塊を話題チャンクとして抽出し, ユーザに提示することで，話題を理解する支援を行う．マイクロブログから話題チャンクを抽出するために，記事間が同一の話題であるか否か判断する指標が必要となる．指標としてまず考えられるのは，連続して投稿される前後の記事間に共起する単語の有無で記事間の話題の連続性を判断することである．しかし，文字数制限のあるマイクロブログでは文章が極めて短い記事が投稿されるため，単純なテキスト処理のみを指標として抽出することは困難である．そこで，本論文ではマイクロブログの即時的な投稿スタイルに着目し，ユーザが短期間に連続して投稿する記事は同一話題である可能性が高いとし，投稿の時間間隔に基づく抽出手法を提案する．共起語に加えて記事投稿の時間間隔を話題チャンクの抽出指標として適用することで，抽出精度を向上させることが本研究の目的である．

2. 先 行 研 究

マイクロブログからあるトピックに関する複数の記事をまとめて抽出する研究には, Maxim ら [1] の提案した検索システムがある．ユーザが入力したキーワードに対し，バースト投稿が発生した日のマイクロブログ記事をまとめて提示する．

マイクロブログ記事の投稿時間に着目した研究には, 高村ら [2] がある．高村らは，スポーツの試合のテレビ中継中等に大量発生する実況や感想の書き込みの集合をマイクロブログストリームと称した上で, その要約を自動生成する手法を提案している．その際, テキスト上は類似した書き込みであっても時間的に離れた記事は異なるイベントに対する記事である可能性が高いことを検証している．また, 青島ら [3] は, 書き込み時間が近い記事間は共通点があると仮定し, 戸田ら [4] らの提唱した時間類似度をクラスタリングの指標として適用している．

その他, マイクロブログに関する研究は盛んに行われている．藤坂ら [5] は，マイクロブログサイトを実空間を観察するネットワークであると定義し，ユーザの移動パターンモデルを提唱し，マイクロブログを位置，時間及びメッセージの指標に基づき分析し，地域イベントの影響範囲を推定している．松村ら [6] は，場所キーワードを用いて Twitter から有用な記事を自動収集するシステムを提案している．岩木ら [7] は，ユーザと記事との近接度を過去の投稿履歴や返信回数に基づき計算した上で感性辞書を適用することで，有用な記事の発見支援を提案している．吉本ら [8] は，ツィート数とフォロー数を用いて Twitter のユーザの重要度を推定した上で，算出した重要度を用いてフォローすべきユーザを推薦する手法を提案している．桑原ら [9] は投稿者のメッセージから共通の話題を持つ投稿者の推薦を行っている．吉田ら [10] はリンクを含む日本語記事を分析した．その結果，娯楽サービスの URL が投稿されやすく，また，リツィートは URL の投稿に影響を及ばさない，ボット(注3)以外の投稿では 10〜20 文字の一言の記事が多いとの知見を得ている．

本研究の位置づけ

マイクロブログから同一話題のチャンクを抽出する研究は, 抽出されるチャンクは複数のユーザによる記事によって構成されるため，非常に大きな集合体となる．そのため，本論文のように一人のユーザが次々と即時的に書き込むことで形成される話題チャンクを抽出対象としていない．本研究は記事の投稿時間を利用する点で高村らと類似した手法ではあるが，特定イベントに付随するストリームを対象としていない点で異なる．

本研究では, 一人のユーザが共通する話題について連続して書き込むことで形成される記事の塊を話題チャンクとし，単語の共起関係と投稿の時間間隔から抽出する手法を提案する．

3. 話題チャンク抽出手法

マイクロブログから話題チャンクを抽出し，ユーザに提示することで，話題の理解を支援する．本論文では，ある話題について記述された複数の記事の塊を話題チャンクと定義する．マイクロブログ記事は，従来のブログ記事と比較すると極めて短い文章で記述されるため，単純にテキスト内容を比較するだけで記事間の話題の連続性を判断することは難しい．そこで，本論文では，単語の共起関係に加えて記事間の投稿間隔を利用した話題チャンク抽出手法を提案する．話題チャンクには他ユーザとのリプライ関係により複数ユーザによって形成される話題チャンクと，単一ユーザがある話題について複数の記事に分けて投稿することで形成される話題チャンクの 2 種類が考えられる．現在，マイクロブログにはリプライによって形成される会話を前後の記事含めまとめて表示する機能があるため，複数ユーザによる話題チャンクを閲覧することは可能である．そのため，複数ユーザによる話題チャンクは本研究の対象外とし，単一ユーザによって形成される話題チャンクのみを本研究における抽出対象とした．

(注3)：定期的にニュースを配信したり，映画やアニメのキャラクターを模倣して自動的に投稿を行うプログラム

本論文で提案する話題チャンク抽出システムの位置づけを図 2 に示す．提案システムは，ユーザが閲覧しているマイクロブログのタイムラインより記事集合を取得し，本論文で提案する指標を用いて記事間の関連性を比較する．記事間に関連性があるとされたら同一話題とし，図 3（a）のようなタイムラインを閲覧しているユーザに対し，(b）のように話題チャンクとして提示する．以下，3.1 節においてタイムラインの構成を述べた上で，その特徴と問題点を明らかにする．3.2 節では話題チャンク抽出に関する予備調査の結果を示す．予備調査では，あるユーザのタイムライン上に形成されている話題チャンクを手作業で抽出し，その実態を明確にした．3.3 節，3.4 節において，それぞれ共起語と投稿間隔を指標とした抽出手法について述べる．

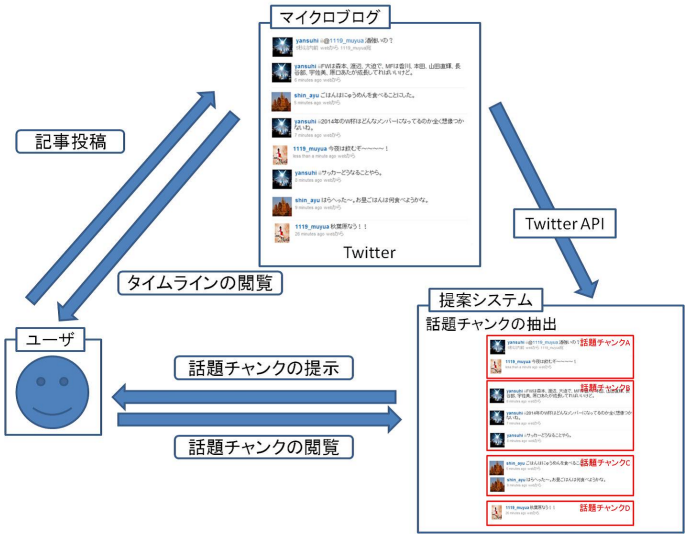

図 1　提案システムの位置づけ

図 2　タイムライン表示と話題チャンク表示

3.1　タイムラインの構成

マイクロブログでは，ユーザは好みの他ユーザを選択しフォロー関係を生成することで，他ユーザが投稿した記事を自分のタイムラインに取り込み，次々と表示させることができる．図 3（a）にあるように，タイムラインには自身が投稿した記事と，自身がフォロー関係にあるユーザの記事が時系列順に表示され，新規の記事が投稿されるたびに更新されていく．マイクロブログには他ユーザの書き込みに返信を行うリプライや，他ユーザの投稿記事をタイムライン上に引用するリツィート等の機能がある．

リプライやリツィートを駆使して他ユーザとのコミュニケーションを楽しむユーザもいれば，自身の周辺で起こった出来事について頻繁に書き込みを行うユーザもおり，タイムライン上では多数のユーザによって展開される様々な話題が混在している状態にある．先述のように，マイクロブログではその投稿の手軽さゆえに，1 回の書き込みで話題が完結するとは限らない．そのような場合，関連する前後の記事を閲覧しなければ，その話題を十分に理解できない可能性が高い．しかし，タイムライン上には多数のユーザによって様々な話題の記事が投稿されるため，関連する記事の発見が困難な場合もある．高いリアルタイム性はマイクロブログの大きな特徴の一つであるが，即時的な投稿形態によりタイムラインが次々と更新され，同一話題の記事が寸断されやすいことが問題であるといえる．

3.2　話題チャンク抽出に関する予備調査

タイムライン上で話題チャンクを形成している記事がどの程度あるのか調査を行った．対象データは，Twitter における，あるユーザのタイムライン上で 2010 年 8 月 21 日 14 時 58 分から 8 月 22 日 16 時 10 分までの約 25 時間に投稿された，178 件の記事である．これらのデータに対し，手作業による話題チャンクの抽出を行った．抽出対象とした話題チャンクは，単一ユーザが同一話題について連続して記述することで形成される話題チャンクと，複数ユーザがリプライやリツィート等の関係を構築することで形成される話題チャンクの 2 種類である．同一話題であるか否かの指標は，実際に記事内容を閲覧することによる比較のみとした．その抽出結果を表 1 に示す．

表 1　手作業による話題チャンク抽出結果

	話題数	記事件数
単一ユーザによるチャンク	12	46
複数ユーザによるチャンク	6	24
チャンクを形成していない記事	108	108
合計	126	178

本論文が対象としている単一ユーザによる話題チャンクは 12 チャンク抽出され，取得した記事 178 件のうち約 4 割にあたる 70 件の記事が何らかの話題チャンクを形成していた．また，話題チャンクを形成している記事の多くは，他の話題チャンクの記事やチャンクを形成していない記事と混在しており，ユーザが自らタイムラインを閲覧しているだけでは見つけにくい状況にあった．

また，抽出した話題チャンクを構成している記事件数を調査したところ図 4 のような結果が得られた．2 件の記事で構成される話題チャンクが最も多い数値を示しているが，約 4 割の話題チャンクは 5 件以上の記事で構成されていることが明らかになった．話題チャンクを構成する記事件数が多くなるほど，タイムライン上でユーザが話題を理解するために多くの記事を閲

覧する必要があり，話題を十分に理解するためには多くの時間を要する可能性が高い．

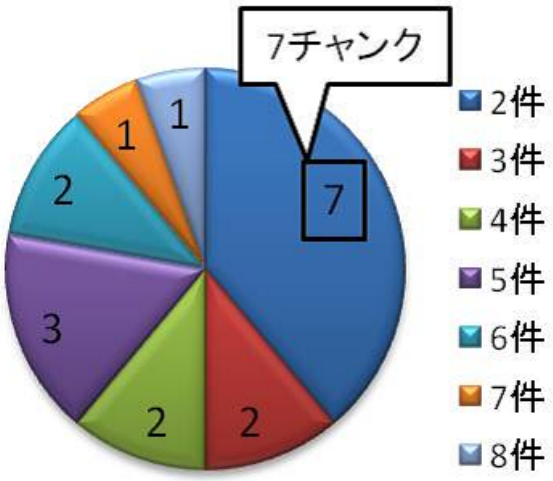

図 3　話題チャンクを構成する記事数

3.3　共起語に基づく抽出手法

マイクロブログでは極めて少ない文字数で記述された記事が投稿される．青島ら [4] が Twitter の日本語記事約 2500 万件を対象に行った調査によると，Twttter における一つの記事あたりの平均文字数は 43.6 文字である．Twitter の記事には 140 文字以内の文字数制限が設定されているが，多くのユーザは制限よりもはるかに少ない文字数で書き込みを行っていることが分かる．そのため，TF-IDF のような単語の出現頻度を用いた計算を行ってもスコア差が生まれにくいと考えられる．

しかし，極めて短文の記事に連続して同じ単語が登場すれば，話題継続の手掛かりとなる．そこで本論文では，連続して投稿される前後の記事間に共起する単語の有無で，話題の連続性を判断する手法を提案する．まず，抽出対象となるユーザの記事集合に対し，形態素解析を行い，名詞，動詞，形容詞を抽出する．次に，記事間で連続して登場する単語を検出する．共起している単語があれば，話題継続と定義し，チャンクとして結合する．

3.4　投稿間隔に基づく抽出手法

上述した共起語に基づく話題チャンク抽出手法では，記事間で共起している単語が無ければ話題チャンクとして抽出されない．そのため，共起語のみを指標としては抽出できない話題チャンクに対し，別のアプローチが必要となる．

そこで，本論文では，マイクロブログでは短い文章で即時的な投稿が行われる点に着目し，短期間で連続して投稿された記事間は同一話題である可能性が高いとする投稿間隔に基づく抽出手法を提案する．

提案手法では，ユーザがマイクロブログに記事を投稿する時間間隔を算出した上で，与えられたパラメータより短い間隔で投稿された記事間を同一話題と判断しチャンクとして結合する．

パラメータとして与えた投稿間隔に対し，それぞれ抽出精度を算出することで，チャンク抽出に適切な投稿間隔を明らかにする．チャンク抽出に適切な投稿の時間間隔はユーザの平均投稿間隔により異なるため，全てのユーザを同条件で比較するための指標が必要となる．

そこで，抽出指標として与える投稿間隔を，平均投稿間隔で正規化したパラメータ t を式 1 のように定義する．t の値を変動させながら話題チャンクを抽出し，抽出精度の変化を分析することで，抽出に適切なパラメータを明らかにする．

$$t = \frac{\text{抽出指標とする時間間隔}}{\text{平均投稿間隔}} \tag{1}$$

4.　話題チャンク抽出に関する評価実験

提案手法に対し，評価実験を行うことで有効性を確認する．以下，4.1 節では，評価対象として取得したユーザの記事集合について述べる．4.2 節では，抽出した話題チャンクの精度算出手法について述べる．

4.1　評価対象ユーザ

Twitter から，表 2 に示す 6 名を評価対象ユーザとして選定した．表の特徴欄に示すように，この 6 名は平均投稿間隔や投稿形態から 2 名ずつの 3 タイプに分類される．

表 2　評価対象ユーザの特徴

	平均投稿間隔（分）	投稿形態
ユーザ A	15.7	投稿頻度が高い．普段は短文だが
ユーザ B	19.2	長文で持論を展開することもある．
ユーザ C	161.1	総じて投稿頻度は高く，短文で日常の
ユーザ D	218.4	些細なことを次々と書くことが多い．
ユーザ E	1513.1	大衆に向けた情報発信源として利用．
ユーザ F	1919.8	投稿頻度は低いが，長文が多い．

評価対象ユーザ 6 名に対し，それぞれ約 200 件(注4)の連続する投稿記事を収集し，人手で話題チャンクを抽出する．これを正解チャンクとし，提案手法による抽出チャンクと比較し，抽出精度を算出する．ここで，話題チャンクとは同一話題について記述された複数の記事の塊であることから，同一の話題について 2 件以上の記事が言及している場合を話題チャンクとして抽出することとした．

4.2　話題チャンク抽出精度の評価手法

提案手法の有効性を議論するには，抽出された話題チャンクと，正解データとして与えられる話題チャンクとの間で抽出精度を算出するための評価手法が必要となる．本論文では，複数の要素を含む集合間の類似度計算で一般的な *Jaccard* 係数を基本とする評価手法を用いることとした．*Jaccard* 係数とは，2 つの集合 P，Q があった時，

$$Jaccard\,(P, Q) = \frac{|P \cap Q|}{|P \cup Q|} = \frac{|P \cap Q|}{|P| + |Q| - |P \cap Q|} \tag{2}$$

(注4)：Twitter API により一度に取得できる記事数が制限されることによる

で与えられる指標である．すなわち，2 つの集合間で共通する要素が多いほどその値が大きくなる．

本論文では正解データの話題チャンクと提案手法により抽出した話題チャンクの間で *Jaccard* 係数を算出する．比較対象とする正解チャンクは，抽出チャンク中の記事が 1 件以上含まれる正解チャンクし，それぞれのチャンク間で図 5 のように *Jaccard* 係数を算出していく．以下に，具体的な算出の手順について述べる．

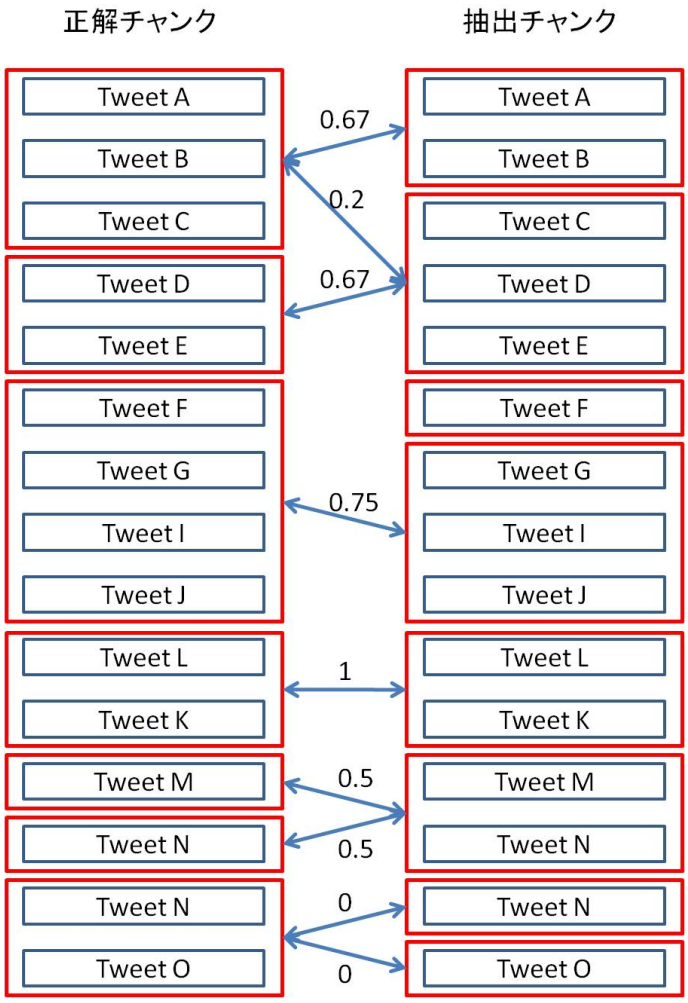

図 4　抽出精度計算処理の流れ

（1）　比較対象となる正解チャンクが 1 チャンクの場合

抽出チャンクにおける TweetA，B のチャンクように，抽出チャンクの比較対象となる正解チャンクが一つに決まるときは，その正解チャンクとの間で *Jaccard* 係数を算出する．

（2）　比較対象となる正解チャンクが複数ある場合

抽出チャンクにおける TweetC，D，E のチャンクは，正解チャンクにおける TweetA，B，C のチャンクと TweetD，E のチャンクの 2 つのチャンクが比較対象となる．このような場合は，比較対象となる全ての正解チャンクとの間で *Jaccard* 係数を算出し，複数のスコアを与える．

（3）　抽出チャンクが単一記事で構成される場合

本研究では，一つの話題について記述された複数の記事の塊を話題チャンクと定義しているため，抽出チャンクにおける TweetF のように，チャンクが 1 件の記事で構成されている場合は抽出精度算出の対象外とする．そのため，*Jaccard* 係数の値は算出しない．ただし，正解チャンクがあるにも関わらず，提案手法による抽出では記事が全て分割されてしまった場合は（4）のように処理する．

（4）　正解チャンクを全く検出できなかった場合

抽出チャンクにおける TweetN，O のように，正解チャンクがあるにも関わらず，提案手法による抽出では，記事が全て分割されてしまった場合は，分割された記事それぞれにペナルティとしてスコア値 0 を与え，抽出精度が低下するようにする．

上記のように算出した値の平均値を話題チャンクの抽出精度とする．

5.　結果と考察

5.1　共起語に基づく抽出

形態素解析による単語抽出の際に，正解チャンクデータと比較しながら話題継続の手掛かりとならないと判断した単語およびその活用形を不要語として削除した．これを表 3 に示す．

表 3　不要語リスト

@ こと い する せる なる
いる れる よう ある もの

上記の不要語を削除した上で，抽出した単語の共起関係に基づき話題チャンクの抽出を行った．その抽出精度を表 4 に示す．

表 4　共起語に基づく話題チャンク抽出精度

	抽出精度
ユーザ A	0.511
ユーザ B	0.294
ユーザ C	0.478
ユーザ D	0.459
ユーザ E	0.347
ユーザ F	0.323

表 4 より，共起語による話題チャンク抽出手法では，評価対象ユーザ 6 名に対し，平均約 40 %の抽出精度を示した．最も高い数値を示したのがユーザ A で，約 50 %の抽出精度を示した．一方でユーザ B を対象とした場合の抽出精度が最も低く，抽出精度は 30 %にも満たなかった．

5.2　投稿間隔に基づく抽出

それぞれの平均投稿間隔に応じて，パラメータを変動させながらチャンク抽出精度を算出した．その結果を図 6 に示す．横軸は式 1 のパラメータを，縦軸は抽出精度をそれぞれ示す．また，表 5 に各ユーザの最大抽出精度を示す．

図 6 より，ユーザ F を除き，いずれユーザも投稿間隔のパラメータ t が約 0.1〜0.2 のときに，精度がピークを迎えていることが分かる．このことから，平均投稿間隔の約 0.1〜0.2 倍の時間より短い間隔で投稿された記事間は同一話題である可能性が比較的高いといえる．ユーザ E とユーザ F は同タイプのユーザであるが，投稿間隔の分散を分析したところ，ユーザ F の方

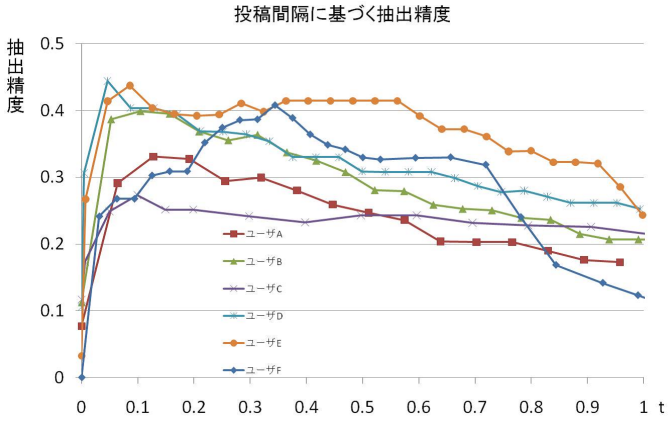

図 5　投稿間隔に基づく話題チャンク抽出精度

表 5　投稿間隔による抽出の最大抽出精度

	平均投稿間隔（分）	抽出精度
ユーザ A	15.7	0.331
ユーザ B	19.2	0.395
ユーザ C	161.1	0.273
ユーザ D	218.4	0.444
ユーザ E	1513.1	0.437
ユーザ F	1919.8	0.408

が投稿間隔のバラつきが大きいことが分かった．これがユーザ F のみが異なる傾向を示した原因であると考えられる．また，いずれのユーザも投稿間隔が短すぎると抽出精度は低く，その後ピークを迎えた後に，抽出精度は緩やかに低下していく傾向を示している．これは，平均投稿間隔より極端に短い時間間隔で話題転換点を抽出すると，ほとんどのチャンクが単一記事によって形成されてしまい，2 件以上の記事で構成されるチャンクが抽出できず，また，あまりにも平均投稿間隔より長い時間間隔で話題転換点を抽出すると，不要な記事までチャンクに結合してしまい，ノイズが多くなるためであると考えられる．

各ユーザの平均投稿間隔に着目して分析すると，平均投稿間隔が長いユーザが抽出精度が高くなる傾向にある．これは，普段から極めて短い投稿間隔で書き込みを行うユーザは，継続している話題でも別の話題でも連続して投稿する傾向があるため，短い間隔で投稿が行われても同一話題であると判断できないためであると考えられる．一方でユーザ E やユーザ F のように普段は頻繁に投稿を行わないユーザが，短い投稿間隔で書き込んだ記事は継続した話題である可能性が高いと考えられる．このことから，投稿間隔に基づく話題チャンク抽出は，平均投稿間隔が長いユーザにより有効であるといえる．

共起語に基づく抽出手法と比較すると，ユーザ B，ユーザ E，ユーザ F において投稿間隔に基づく抽出手法が共起語に基づく抽出を上回る精度を示している．このことから，共起語と投稿間隔を併用した抽出手法を採用することで，更に精度を高めることができると考えられる．

5.3　共起語と投稿間隔を併用した抽出

5.2 節で行った実験では，投稿間隔に基づく抽出では平均投稿間隔の約 0.1 から 0.2 倍の時間間隔で抽出した時に抽出精度のピークを迎える傾向があることを示した．例外としてユーザ F はパラメータ t が約 0.3 から 0.4 の時、最も高い精度を示している．そこで，共起語による抽出を行ったうえで，チャンクを検出できない記事間に対し，投稿間隔のパラメータを 0.1〜0.4 で変動させ，パラメータ以下の間隔で投稿されていればチャンクとして結合する手法を提案する．共起語と投稿間隔を併用した話題チャンク抽出手法を適用した上で，それぞれの手法を個別に採用したときと比較し，抽出精度の変化を分析した．その結果を表 6 に示す．

表 6　共起語と投稿間隔を併用した際の抽出精度

	0.1	0.2	0.3	0.4
ユーザ A	0.531	0.510	0.494	0.480
ユーザ B	0.448	0.490	0.472	0.438
ユーザ C	0.415	0.415	0.381	0.381
ユーザ D	0.474	0.418	0.413	0.407
ユーザ E	0.367	0.370	0.370	0.370
ユーザ F	0.331	0.331	0.331	0.335

ユーザ A からユーザ E はいずれも投稿間隔のパラメータを 0.1〜0.2 に設定したときに抽出精度のピークを迎えている．一方で最も平均投稿間隔が長いユーザ F はパラメータの変動に対し，あまり抽出精度の変化は見られなかった．共起語と投稿間隔を併用した抽出手法を適用することで最も顕著な変化が見られたのはユーザ B で，共起語のみを指標とした手法による抽出精度は約 29.4 %，投稿間隔のみを指標とした手法による抽出精度は約 39.5 %であったのに対し，共起語と投稿間隔を併用した抽出手法では約 49 %の抽出精度を示しており，それぞれの手法を個別に適用した場合より大きく精度が上がった．また，ユーザ A とユーザ B は共起語と投稿間隔を併用して抽出したときに最も高い精度が得られた．このことから提案手法はユーザ A，ユーザ B のように平均投稿間隔が短く，長文で特定話題について連続で書き込むことがあるユーザに対し有効であると考えられる．一方で，投稿間隔のみを指標として抽出したときには上位の精度を示したユーザ E，ユーザ F は共起語による抽出手法と併用することで精度は低下した．このようなユーザには，先に投稿間隔による手法で抽出した上で，未検出チャンクを共起語で抽出する手法が必要であるといえる．

6.　おわりに

本論文では，マイクロブログからある話題について記述された複数の記事の塊を話題チャンクとして抽出する手法を提案した．単純に単語の共起関係を比較するだけでは抽出が困難なマイクロブログ記事に対し，投稿の時間間隔を指標とした抽出を行い，*Jaccard* 係数を基本とする評価実験を Twitter から選定した 6 名のユーザを対象に行った．

その結果，共起語による抽出手法で平均約 40 %の精度を示したのに対して，投稿間隔に基づく抽出手法は，3 名のユーザが共起語に基づく抽出精度を上回る結果を示した．また，平均投稿間隔の約 0.1 倍の時間間隔で抽出したときに，最も精度が高くなる傾向を明らかにした．さらに，共起語と投稿間隔を併

用した抽出手法を提案し，3 名のユーザにおいて，それぞれの手法を個別に適用した際の抽出精度を上回る結果を示すことを確認した．これらのことから，投稿間隔を指標とした話題チャンク抽出の有効性を確認できたといえる．

今後の課題として，前後の記事間だけでなくチャンク間で話題の関連性を比較することにより，離れた記事間をチャンクとして結合すること，ユーザの投稿形態や投稿頻度に応じて適切な投稿間隔のパラメータを採用し，抽出精度の向上につなげること，抽出した話題チャンクを記事検索に適用することなどが挙げられる．

謝　　辞

本研究の一部は科研費（21500091）の助成を受けたものである．ここに記して謝意を示す．

文　　献

[1] Maxim Grinev, Maria Grineva, Alexander Boldakov, Leonid Novak, Andrey Syssoev, and Dmitry Lizorkin. Sifting micro-blogging stream for events of user interest. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009.

[2] 高村大地, 横野光, 奥村学. Summarizing microblog stream. 人工知能学会研究会資料, 2010.

[3] 青島傳隼, 福田直樹, 横山昌平, 石川博. マイクロブログを対象とした制約付きクラスタリングの実現. 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 Forunm 2010）, 2010.

[4] 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. 電子情報通信学会 第 18 回データ工学ワークショップ (DEWS2007), 2007.

[5] 藤坂達也, 李龍, 角谷和俊. 実空間マイクロブログ分析による地域イベントの影響範囲推定. 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 Forunm 2010）, 2010.

[6] 松村飛志, 安村通晃. 街に着目した twitter メッセージの自動収集と分析システムの提案と試作. 2006.

[7] 岩木祐輔, アダムヤトフト, 田中克己. マイクロブログにおける有用な記事の発見支援. 第 1 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 Forum 2009）, 2009.

[8] 吉本和紀, 鈴木優, 吉川正俊. マイクロブログにおける他者への影響を考慮した投稿者の重要度推定手法. 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 Forum 2010 ）, 2010.

[9] 桑原雄, 稲垣陽一, 草野奉章, 中島伸介, 張建偉. マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式. 情報処理学会データベースシステム研究発表会, 2009.

[10] 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 Forum 2010）, 2010.