行動の時空間連続性を考慮した旅行ツイートの組織化

長谷川馨亮† 馬 強†† 吉川 正俊††

† 京都大学工学部情報学科 〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科 〒 606–8501 京都府京都市左京区吉田本町
E-mail: †k.hasegawa@db.soc.i.kyoto-u.ac.jp, ††{qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし 本研究では，Twitter 上に断片的に投稿されるユーザの旅行体験をまとめて検索できる手法を提案する．旅行体験は一連の行動から構成されることが多い．このような旅行体験における時空間の連続性を考慮して，旅行体験とツイートとの関連性を時間・空間・内容から計算し，体験ごとに整理してユーザに提示する．また，本稿では提案手法の評価実験を行い，時空間の連続性を考慮した関連度を用いることにより，キーワード検索の結果による組織化よりも高精度の組織化が行えることが確認された．

キーワード マイクロブログ，Twitter，ユーザ体験，時系列データ，空間の連続性

1. はじめに

近年，インターネット上ではマイクロブログや SNS などのユーザが自ら情報を発信するメディアである CGM（Consumer Generated Media）が急速に普及してきている．CGM にはユーザ体験についての情報が大量に蓄積されているが，蓄積された情報を整理・活用するための技術の開発は十分に進んでいないという現状がある．

特に，代表的なマイクロブログである Twitter(注1)は，1 つの投稿が 140 文字以内と短く，ユーザが体験したことをリアルタイムで気軽に投稿できる点が特徴である．多くのユーザがリアルタイムに体験を投稿する代表的な場面が旅行であり，旅行中のユーザが，ある場所を訪れた感想をその場で撮った写真とともに Twitter に投稿する，というような利用が一般的となってきている．

一方で Twitter に投稿されたコンテンツの共有や整理，検索などの技術は発展途上である．そのため，例えば以前の旅行について振り返るために過去の Twitter の投稿を見返そうとしても，Twitter で過去の投稿を閲覧するには最新の投稿から順に遡っていくしかなく，それにも制限があるため，一度投稿した書き込みを後から見返したり整理したりすることが困難である．また，Twitter では 1 日に 1 人のユーザが何回も投稿することが多いため，旅行などの体験コンテンツの共有がされにくいといった問題点もある．さらに，ある旅行体験についての投稿を検索しようとしても，1 つの投稿が短い Twitter では，同じ体験が複数の投稿に分けて記述されていたり，その体験を表すキーワードが省略された投稿が多かったりするので，従来のキーワードによる検索だけでは漏れが生じるという問題も起こっている．

また，Twitter の普及に伴い，従来はブログや SNS の日記などをメインとして投稿していたユーザが，Twitter を中心に普段の出来事を投稿するようになった，というケースも存在する．こうしたユーザは一度 Twitter に書いたことをもう一度ブログや SNS に投稿しない，またはブログなどを使う場合でも Twitter とは違った視点からの投稿となるという場合が多い．こうした状況も情報の共有がされにくくなる一因となっている．

こうした問題を解決するために，本研究では Twitter 上に投稿されたコンテンツの中から，ある旅行体験を表すツイートだけをまとめて整理できる手法を提案する．一般的な旅行体験では，数時間や数日などある一定の時間内に移動しながらいろいろな場所を訪れていくため，旅行体験を表すツイートには時間の連続性と空間の連続性が存在する．そこで，従来の検索のようにキーワードのみを用いて整理するのではなく，時間の連続性と空間の連続性も考慮して旅行体験を表す投稿を完全に網羅した整理を行い，その検索結果によって Twitter 上のコンテンツを旅行体験という観点から組織化する手法を提案するのが本研究の目的である．

2. 関連研究

近年，Twitter に関する研究が盛んに行われている．青島ら [1] はマイクロブログの特徴を考慮した Twitter 上の投稿に対する制約付きクラスタリングの手法を提案している．藤坂ら [2] は Twitter の投稿に付加されたジオタグを用いて特定の地域で発生したイベントを発見し，その影響範囲を推定する手法を提案している．Wu ら [3] は Twitter 上でマスメディアや有名人などから一般ユーザへどのように情報が伝わっていくかを調査している．Castillo ら [4] は Twitter 上の情報が信頼できるものであるかどうかを教師あり学習に基づく方法を利用して判別する手法を提案している．

文書から個人の経験を抽出する経験マイニングという研究も行われている．倉島ら [5] はブログに書かれた経験の状況，行動，主観の関係をルールとして抽出する手法を提案している．また，個人の経験をライフログとして整理する研究も行われている [6].

Twitter からユーザ体験を検索するための手法は有光ら [7] によっても提案されている．有光らの手法ではユーザ体験をいくつかの行動によって定義し，それぞれの行動は決められた順に

(注1)：http://twitter.com

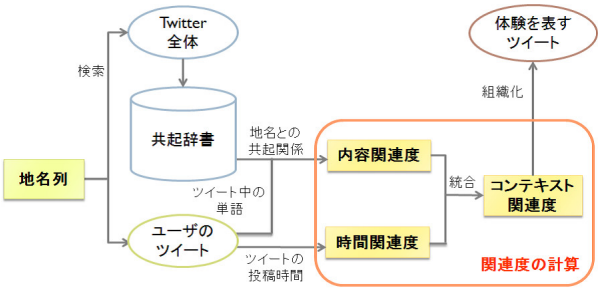


図 1　提案手法によるツイートの組織化の流れ

遷移していくものとしているが，本研究では旅行体験をいくつかの地名によって定義し，その際に空間の連続性を考慮している点，また，旅行体験を定義する地名列に含まれるそれぞれの地名が出現する順番は特に指定していない点が異なる．

3. 提 案 手 法

旅行体験は，地名の系列で表すことができる．例えば，京都を観光する場合には，最初に八坂神社を見学し，それから高台寺に寄り，その後清水寺に向かうといったコースが考えられる．この旅行体験を表すツイートの中にも八坂神社や清水寺といった地名を含むツイートが存在する可能性が高いため，このコースを観光した体験を「八坂神社，高台寺，清水寺」という地名の系列で表現することができる．このように，本稿における旅行体験を表すツイートとは，一連の観光体験を表すツイートの系列で，いくつかの地名の系列で表されるものとする．

図 1 に提案手法によるツイートの組織化の流れの概略を示した．以下でそれぞれの段階で行う処理について述べる．

提案手法では，Twitter 上のツイートに対して，整理したい旅行体験を表す地名の系列を入力とし，旅行体験を表すツイートの系列を出力とする．例えば，あるユーザ A が自分が以前訪れた京都旅行について書いたツイートを検索する場合は，ユーザ A のツイート全体に対して地名列「八坂神社，高台寺，清水寺」を入力とし，京都旅行で八坂神社や清水寺などを訪れた体験を表すツイートを組織化し，整理したい体験を表すツイートの系列としてまとめてユーザに提示する．

なお，ユーザが入力として与える地名列には，必ずしもその旅行体験を表す地名が全て含まれる必要はない．ユーザが探したい旅行体験で訪れた地名や名所の名前を全て覚えていなくても，1 つでも訪れた場所を覚えていれば，それを地名列として入力すればよい．例えば,「清水寺」という地名だけを地名列として与えても，空間の連続性などを考慮してその旅行で他に訪れた八坂神社や高台寺などに関するツイートも検索することができる．つまり，本研究の目的は，与えられた不完全な地名列に含まれる地名から，与えられた地名に関する情報だけでなくそれ以外の地名に関する情報も補完して検索を行うことであるともいえる．

また，本来であれば旅行体験から地名列への展開，ここでは「京都」という体験全体を表す地名から「八坂神社，高台寺，清水寺」という具体的に体験を表す地名列への変換作業が必要となるが，本稿では最初から地名列を入力として与えるものとし，旅行体験から地名列への自動的な展開については今後の検討課題とする．

整理したいユーザのツイートのうち実際に関連度を計算するツイートの範囲については，地名列によるキーワード検索や，件数や投稿時間の差などを利用して絞り込んでおく方法も考えられるが，本稿では与えられたユーザのツイート全てを関連度の計算対象とした．

空間の連続性や内容の関連性を求めるために，候補となるツイートの本文に対して MeCab(注2) を用いて形態素解析を行う．形態素解析の結果や，それ以外にツイートから得られる情報を利用して，検索によって得られた体験を表す投稿の候補となるツイートに対し検索したい旅行体験との関連度を計算する．そして，関連度が一定値以上のものを求める旅行体験を表すツイートとする．候補となるツイートの旅行体験との関連度は，ツイートの時間の連続性，空間の連続性，内容の関連性を考慮して決定する．それらを考慮した関連度の指標として以下の指標を用いる．

- 内容関連度 (3.1 節)

主に内容の関連性を考慮した指標として，地名とツイート中に含まれる単語の共起度を用いる．ある場所との関連が強い単語は，その地名との共起度が高いという特徴に着目している．

- 時間関連度 (3.2 節)

主に時間の連続性を考慮した指標として，地名を含むツイートとの時間間隔を用いる．旅行体験を表す一連のツイートは投稿時間が近接しているという特徴に着目している．

- コンテキスト関連度 (3.3 節)

周辺のツイートの影響を考慮した指標である．1 つの話題が複数のツイートに断片化して投稿されることが多いという特徴に着目し，内容関連度と時間関連度の結果を統合した結果に前後のツイートの影響を加味して最終的な関連度を決定する．

これらの指標により関連度を計算する．一定の値以上の関連度を持つツイートを求める体験を表すツイートとし，組織化を行う．

以下でそれぞれの指標の詳細について述べる．

3.1 内容関連度

内容の関連性と空間の連続性を考慮した指標として，地名とツイート中に含まれる単語の共起度を用いる．検索したい旅行体験を表す投稿の候補となるそれぞれのツイートについて，ツイートに含まれる地名以外の単語と，旅行体験を表す地名との共起度を求める．ツイートに含まれる単語と旅行体験を表す地名との関連性が高いほど，そのツイートは求める旅行体験とより強く関連している，という仮定に基づいている．

例えば，八坂神社でおみくじを引いた体験について記述したツイートに「おみくじ」という単語のみが現れ「八坂神社」という単語は現れていない場合，従来のキーワード検索によって「八坂神社」をクエリとして検索を行っても，おみくじについて記述したツイートは検索結果に含まれない．しかし,「おみく

(注2)：http://mecab.sourceforge.net/

じ」という単語が「八坂神社」という地名と強く関連しているという情報を考慮して「おみくじ」を含むツイートに高い関連度を与えることによって，八坂神社についての検索結果におみくじを引いた体験について記述したツイートを含めることができる．

ツイート中に含まれる単語と地名の共起度を求めるために，まず，共起を求める対象となる地名について共起度計算を行うための共起辞書を作成する．そして，その共起辞書を利用して各ツイート中の単語について対象となる地名との共起度計算を行い，その結果をもとにツイートと旅行体験を表す地名との関連度を決定する．

3.1.1 共起辞書の作成

本研究で扱う共起辞書は，Twitter 上の投稿から作成する．地名と関連性の高い単語を幅広く収集するために，共起辞書の作成においては，一般に公開されているツイート全体からツイートを収集する．

単語との共起度を求める対象となる地名には，次の 2 種類がある．

（1） クエリとなる地名列に含まれる地名

（2） 地名列に含まれる地名が表す場所の周辺の場所を表す地名

例えば,「八坂神社，高台寺，清水寺」という地名列に対して単語との共起度を求める場合には，地名列に含まれる 3 つの地名に加えて，祇園や二年坂，円山公園といった 3 つの地名の周辺にある地名についても共起を計算する．

これらの対象となる地名それぞれについて，地名をキーワードとして Twitter API(注3)による検索を行い，検索結果として得られたツイートを地名ごとに蓄積しておく．地名ごとに蓄積されたツイートの本文に対して形態素解析を行い，出現した単語と，それぞれの単語が出現したツイートの数を記録する．この単語と出現回数の組を地名ごとに記録したものが，本研究で扱う共起辞書となる．

地名列に含まれる地名の周辺の地名は，本来ならば Google Maps API(注4) などを用いて自動的に取得するのが望ましいが，本稿ではあらかじめ指定しておいた地名列について，その周辺の地名もいくつか与えておくものとする．そして，本稿では地名列に含まれる地名と，与えておいた周辺の地名について共起辞書をあらかじめ作成しておき，それを利用して共起度の計算を行うものとする．

3.1.2 共起度の計算

まず，地名列の地名 $p_1, p_2, ..., p_n$，およびそれらの近くの地名 $q_1, q_2, ..., q_m$ において，それぞれの地名について地名 p_k と求める単語 w との共起度 $Co(p_k, w)$ を Jaccard 係数により求める．Jaccard 係数とは集合の共起を表すためによく用いられる指標であり，2 つの語のうちどちらか一方が出現した集合のうち，2 つの語が同時に出現した集合の割合を表すものである．

$$Co(p_k, w) = \frac{|p_k \cap w|}{|p_k \cup w|} = \frac{|p_k \cap w|}{|p_k| + |w| - |p_k \cap w|} \quad (1)$$

(注3)：http://dev.twitter.com/

(注4)：http://code.google.com/intl/ja/apis/maps/

$|p_k \cap w|$ を p_k と w を共に含むツイートの個数，$|p_k \cup w|$ を p_k と w のどちらか一方のみを含むツイートの個数，$|p_k|$，$|w|$ をそれぞれ p_k，w を含むツイートの個数とすると，$Co(p_k, w)$ は Jaccard 係数の定義から (1) 式ように表される．(1) 式は地名 p_k について共起度を計算したものであるが，周辺の地名 q_l についても同様に計算できる．

次に，(1) 式によって求められた地名と単語の共起度に基準となるツイートに含まれる地名と地名列に含まれる地名との距離によって計算される重みを掛け合わせることで，空間の連続性を関連度の計算に反映させる．例えば，あるツイートの前後に「円山公園」という地名を含むツイートが存在したとする．円山公園は八坂神社のすぐ近くにあるため，前後に「八坂神社」を含むツイートが存在しなかったとしても，円山公園と八坂神社の位置関係からある程度は八坂神社に関連したツイートであることが推測できる．そして，その関連の度合いは，ツイートに含まれる地名が表す場所とクエリとなる地名列に含まれる場所の距離が近いほど求める旅行体験と強く関連している，という仮定に基づいて，地名列の地名との距離が近いほど値が大きくなるような重み付けによって表すことができる．

ここで利用するのが，戸田らの研究 [8] などで提案されている指数関数逓減モデルである．戸田らの研究では，タイムスタンプを持つ文書について，文書間のタイムスタンプの差が大きくなるほど話題の類似度が減少する，という仮定に基づき文書間の話題の類似度を指数関数により求める時間類似度という指標が提案されている．この考え方を地名間の距離に応用することにより空間の連続性を考慮した重み付けを行う．

この考え方に基づいて，地名列に含まれる地名 p_k と単語 w との，周辺の地名との共起も考慮した共起度 $Rco(p_k, w)$ を以下の (2) 式により求める．

$$Rco(p_k, w) = \frac{1}{m+1}\left(Co(p_k, w) + \sum_{l=1}^{m} dist(p_k, q_l) \times Co(q_l, w)\right) \quad (2)$$

$$dist(p_k, q_l) = e^{-\mu_d d(p_k, q_l)} \quad (3)$$

(2) 式において，地名列に含まれる地名 p_k とその周辺の地名 $q_l (l = 1, ..., m)$ について，p_k と q_l の距離 $d(p_k, q_l)$ を反映した重み付けが，(3) 式の $dist(p_k, q_l)$ となっている．μ_d は，内容の類似度が距離が離れるにつれて逓減していく割合を決定するパラメータである．周辺の地名 q_l については，単語 w との共起度 $Co(q_l, w)$ とこの重み付けを掛けたものを各地名についてそれぞれ計算し，その合計を計算に用いた地名の個数で割ったものが $Rco(p_k, w)$ となる．

そして，これを地名列に含まれるすべての地名 $p_1, p_2, ..., p_N$，求めるツイート t_i に含まれる単語 $w_1, w_2, ..., w_M$ について合計することで，地名とツイート t_i 中に含まれる単語の共起度による関連度 $Rc(t_i)$ を計算することができる．

$$Rc(t_i) = \frac{1}{MN} \sum_{w=1}^{M} \sum_{k=1}^{N} Rco(p_k, w) \quad (4)$$

3.2 時間関連度

時間の連続性と空間の連続性を考慮した指標として，地名を

含むツイートとの時間間隔を用いる．検索したい旅行体験を表す投稿の候補となるそれぞれのツイートについて，地名列に含まれる地名，及びその地名と近い場所を表す地名を含むツイートとの投稿時間の差を計算する．

旅行体験においては，徒歩やバスなどで移動しながらいくつかの場所を順に訪れていくため，旅行体験を表すツイートの投稿時間は近接しているものが多いと考えられる．例えば，旅行体験を表す一連のツイートの中に「清水寺」という地名を含むツイートが存在すれば，そのツイートと投稿時間が近い前後のツイートも清水寺について述べている可能性が高い．

そこで，候補となるツイートの投稿時間が地名を含むツイートの投稿時間と近いほど候補となるツイートはその地名により強く関連していると仮定して，前節でも用いた指数関数逓減モデルを，候補となるツイートと地名を含むツイートとの時間差に適用することで，時間の連続性を考慮した関連度を求める．

一連のツイートの中に同じ地名を含むツイートが複数存在する場合は，その地名を含むツイートの中で最もその地名と関連が強いものを時間間隔を求める基準とする．本稿では，手動でツイートの内容を判断し，関連度の高いツイートを基準となるツイートとした．自動的に関連度の高いツイートがどれか判断して基準となるツイートを決定する手法の提案については今後の課題とする．

クエリとなる地名列に含まれる地名を $p_1, p_2, ..., p_n$，その周辺の地名を $q_1, q_2, ..., q_m$，候補となるツイート t_i の投稿時間を $time(t_i)$，地名列の地名 p_k の近くの地名 q_l を含む基準となるツイート t_j の投稿時間を $time(t_j)$ とし，候補となるツイート t_i の前後に地名を含み基準となるツイートが N 個存在したとする（t_j において $j = 1, ..., N$）と候補となるツイート t_i の前後の地名を含むツイートとの時間間隔による関連度 $Rt(t_i)$ は以下の (5) 式で求められる．

$$Rt(t_i) = \sum_{j=1}^{N} dist(p_k, q_l) \times e^{-\mu_t |time(t_i) - time(t_j)|} \tag{5}$$

(5) 式においても，(3) 式で定義した基準となるツイートに含まれる地名と地名列に含まれる地名との距離による重み付け $dist(p_k, q_l)$ を時間間隔による関連度 $e^{-\mu_t |time(t_i) - time(t_j)|}$ に掛け合わせることで，$Rt(t_i)$ は時間の連続性に加えて空間の連続性も考慮した関連度の指標となる．μ_t は，時間間隔が離れるにつれて内容の類似度が逓減していく割合を決定するパラメータである．

3.3 コンテキスト関連度

3.1 節で求めた地名とツイート t_i に含まれる単語の共起度による関連度 $Rc(t_i)$ と，3.2 節で求めたツイート t_i と地名を含む前後のツイートとの時間間隔による関連度 $Rt(t_i)$ の 2 つの指標の結果を統合して，ツイート t_i と検索したい旅行体験との関連度 $R(t_i)$ を求める．

まず，地名を含むツイートとの時間間隔による関連度 $Rt(t_i)$ と地名と単語の共起度による関連度 $Rc(t_i)$ からツイート t_i の前後のツイートの影響を考慮せずに算出されるツイート t_i の関連度 $Rtc(t_i)$ は以下の (6) 式で求められる．

$$Rtc(t_i) = Rt(t_i) Rc(t_i) \tag{6}$$

(6) 式では，共起度による関連度 $Rc(t_i)$ に対して，時間間隔による関連度 $Rt(t_i)$ を重み付けとして掛け合わせている．これにより，2 つの指標の両方を考慮した関連度を求めることができる．

求める旅行体験を表すツイートの中にも，地名との共起度が極めて低い単語が含まれているものや，地名と共起する単語が全く含まれないものも存在している．そのようなツイートについては周辺のツイートの内容から判断できるため，関連度を計算する際にも前後のツイートが持つ関連度の影響を加味する．ツイート t_i の関連度を求める際に前後の X 件のツイートの影響を考慮するとき，ツイート t_i の影響をどの程度加味するかを表す重み付けを τ_i とすると，ツイート t_i と検索したい旅行体験との関連度 $R(t_i)$ は以下の (7) 式で求められる．

$$R(t_i) = \sum_{x=-X}^{X} \tau_{i+x} Rtc(t_{i+x}) \tag{7}$$

本稿では，ツイート t_i の直前のツイートと直後のツイートの影響のみを考慮するものとする．本稿で求める関連度 $R(t_i)$ は，(7) 式において $X = 1$ とした以下の (8) 式により計算される．

$$R(t_i) = \tau_{i-1} Rtc(t_{i-1}) + \tau_i Rtc(t_i) + \tau_{i+1} Rtc(t_{i+1}) \tag{8}$$

$R(t_i)$ の値が一定以上のツイートを求める体験を表しているツイートとして，ツイートの組織化を行う．

4. 実　　験

提案手法における各指標について評価するために，Twitter 上から検索して得られた旅行体験を表すツイートの系列について提案手法による組織化を行った．組織化の性能の評価には主に再現率，適合率，F 値といった尺度を用い，3. で定義した各指標による組織化と，従来のキーワードベースの検索結果による組織化の性能比較，及び各指標間の組織化の性能比較を行った．

4.1 実験方法

各地名ごとの共起辞書の作成と組織化の対象となるツイートの収集のために，Twitter API を用いて対象とする地名をキーワードとした検索を行い，検索結果として得られたツイートについてツイートの ID，投稿したユーザのアカウント名，投稿時間，ツイートの本文，返信先のツイートの ID などを取得した．

今回の評価実験では，「八坂神社，清水寺」という地名列で表される旅行体験を検索するものとした．3.1.2 節における $Rco(p_k, w)$ を算出するために，八坂神社の周辺の地名を祇園，清水寺の周辺の地名を高台寺とし，八坂神社，祇園，清水寺，高台寺のいずれかの地名を含むツイートの中から $Rco(\text{八坂神社}, w)$ と $Rco(\text{清水寺}, w)$ を求めるための共起辞書をそれぞれ作成した．今回は，共起辞書の作成と実験対象ユーザの選定についてはともに 2012 年 1 月 2 日から 2012 年 1 月 17 日にかけて投稿されたツイートを対象とした．

八坂神社についての共起辞書を例に挙げて共起辞書の作成法を述べる．まず，八坂神社を含むツイート，祇園を含むツイートのそれぞれについて MeCab を用いてツイートの本文に対し形態素解析を行い，手動で設定したストップワードを除く名

詞，形容詞，動詞について各単語が出現するツイートの数を記録する．次に，記録された出現頻度を用いて (2) 式における $Co(\text{八坂神社}, w)$ 及び $dist(\text{八坂神社}, \text{祇園}) \times Co(\text{祇園}, w)$ を計算する．今回は，$d(\text{八坂神社}, \text{祇園})$ は八坂神社と祇園間の直線距離とし，その算出には Google Maps API を用いた．そして，八坂神社についての計算結果と祇園についての計算結果を統合し，単語 w と $Rco(\text{八坂神社}, w)$ の組を記録したものを共起辞書とした．

提案手法の評価に用いる組織化の対象となるツイートの収集は，八坂神社，清水寺を含むツイートの中から，八坂神社を含むツイートと清水寺を含むツイートの両方を投稿したユーザのツイートを検索し，その中から八坂神社と清水寺のどちらか一方，もしくは両方を訪れた旅行体験についてまとまった量のツイートを投稿しているユーザを手動で選定した．

その中から今回の評価実験では 2 人のユーザの旅行体験である体験 A，B を表すツイートについて提案手法による組織化を試みた．Twitter API を用いてそれぞれのユーザのツイートを取得し，取得したツイートの中から「八坂神社，清水寺」の地名列で表される京都旅行の体験について述べたツイートを正解として，取得した各ツイートについて手動で正解，不正解の判断を行った．

また，3.2 節における基準のツイートについても手動で選定し，それぞれのユーザのツイートのうち八坂神社及び清水寺を含むものの中から，最もその地名と関連が強いツイートをその地名の基準とした．祇園を含む投稿を行ったユーザについては，祇園についても基準となるツイートを設定し，$Rt(t_i)$ の計算に反映させた．

なお，(3) 式，(5) 式において関連度を計算する際のパラメータである μ_d，μ_t の値については，今回の実験では $\mu_d = \mu_t = 1$ とした．

4.2 評価実験

提案手法による検索およびツイートの組織化の性能を評価するために，3. で提案した以下の 4 つの手法について，テストデータを用いて評価実験を行った．

- 内容関連度 $Rc(t_i)$ のみを用いた組織化 (3.1 節)
- 時間関連度 $Rt(t_i)$ のみを用いた組織化 (3.2 節)
- 内容関連度と時間関連度を統合した $Rtc(t_i)$ を用いた組織化 (3.3 節)
- 時間・内容に加えてコンテキストも考慮した関連度 $R(t_i)$ による組織化 (3.3 節)

テストデータとして用いるユーザ A，B の投稿を Twitter API を用いて収集し，収集できた全てのツイートを求める旅行体験を表している可能性のあるツイートとして，各指標による関連度の計算を行った．

取得できたツイートについて，手動で求める旅行体験を表しているものかどうかを判断し，検索における正解集合に含まれるツイートを決定した．今回の実験では旅行中にその場で訪れた場所について投稿したツイートのみを正解と判断し，京都を訪れたその日の旅行体験全体を求める旅行体験とした．

今回扱うユーザ A，B による京都旅行はどちらも 1 日かけて行われたものである．ユーザ A による旅行体験 A については，ユーザ A は前日にも京都旅行の計画のために京都の名所に関するツイートを投稿しているが，それらは不正解とした．一方，ユーザ B による旅行体験 B では，ユーザ B は京都を訪れた帰りに大阪に立ち寄っているが，大阪に立ち寄ったことも含めて 1 日の旅行が形成されているため，それに関するツイートは正解とした．

表 1　正解及び他の話題について述べたツイートの件数

ユーザ	正解 件数	他の話題 件数	旅行日
A	99	6	2012/1/6
B	69	29	2012/1/2

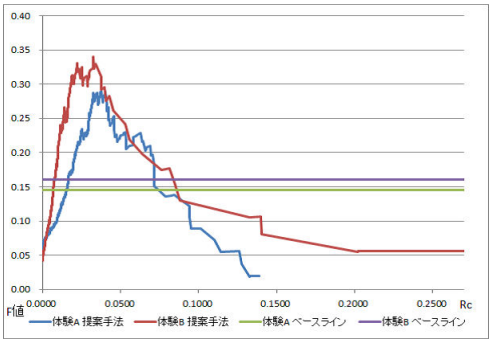


図 2　$Rc(t_i)$ における提案手法とベースラインの F 値の変化

また，旅行体験に関するツイートは短期間に連続して投稿されるが，体験 A，B のどちらも各名所を訪れながら，旅行に全く関係ない話題についても投稿している．不正解集合に含まれるツイートのうち，これらの旅行中に投稿された他の話題について述べたツイートの件数も，正解集合に含まれるツイートの件数と合わせて表 1 に示してある．ユーザ A よりユーザ B の方が旅行中に投稿された他の話題に関するツイートの件数が多く，このことが関連度の計算に与える影響についても調べる．

提案手法を評価するための指標としては，適合率，再現率，F 値を用いた．評価対象とするユーザ A，B のツイートについて $Rc(t_i)$，$Rt(t_i)$，$Rtc(t_i)$，$R(t_i)$ をそれぞれ求め，それぞれの指標が閾値以上であるツイートを求める体験 A，B を表すツイートとして組織化を行い，閾値を変化させていきながらそれぞれの結果について適合率，再現率，F 値を計算した．比較対象とするベースラインには，各ユーザのツイート本文から今回検索対象とする地名列により「八坂神社 OR 清水寺」という OR 検索を行ったものを用いる．ベースラインについても，得られた結果に対して適合率，再現率，F 値を計算した．

4.2.1 内容関連度による組織化の評価

地名とツイート t_i 中に含まれる単語の共起度に基づいた内容関連度 $Rc(t_i)$ による組織化の評価を行う．

まず，提案手法とベースラインの比較，及び体験 A と体験 B の比較を行った．閾値とする $Rc(t_i)$ を変化させたときの体験 A と体験 B における提案手法とベースラインの F 値の変化をまとめたものが図 2 である．体験 A，B ともに $Rc(t_i)$ の閾値が 0.02 〜0.07 程度のときはベースラインよりも提案手法の方が F 値が高くなっている．また，体験 A と体験 B を比較すると，$Rc(t_i)$

表 2　$Rc(t_i)$ において F 値が最大の時の再現率と適合率

体験	Rc(t_i)	適合率	再現率	F 値
A 提案手法	0.0375	0.2871	0.2929	0.2900
A ベースライン	—	0.7273	0.0808	0.1455
B 提案手法	0.0324	0.6400	0.2319	0.3404
B ベースライン	—	1.0000	0.0870	0.1600

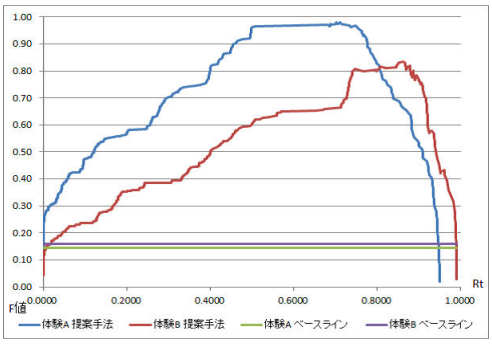


図 3　$Rt(t_i)$ における提案手法とベースラインの F 値の変化

の閾値が 0.055〜0.065 程度のときを除きおおむね体験 B の方が F 値が高くなっている．

両体験について提案手法において F 値が最大となる時の $Rc(t_i)$ の閾値，適合率，再現率，F 値と，ベースラインにおける適合率，再現率，F 値をまとめたものが表 2 である．提案手法とベースラインを比較すると，両体験ともに提案手法における F 値がベースラインにおける F 値の約 2 倍程度になっている．適合率と再現率についても，ベースラインより適合率の低下を 4〜6 割ほどに抑えながら，再現率が約 3 倍にまで向上されている．

体験 A と体験 B で性能に差が出た原因を検証するために，ユーザ A とユーザ B について $Rc(t_i)$ が高い順にツイートを並べ，上位のツイートの内容を比較してみたところ，体験 A では「行く」や「笑」など，ストップワードに設定していなかったが共起する地名にかかわらず頻出する単語を含む関係のないツイートの $Rc(t_i)$ が，正解集合に含まれるツイートよりも高くなっていることがわかった．今回の実験では，形態素解析に用いる MeCab の辞書の修正を行っていないので，「二年坂」などの地名が「二」「年」「坂」と別の単語に分解されてしまうというように，本来ひとまとまりである単語がまとまって認識されないケースが存在する．このような場合通常ならばストップワードとなりうる「二」などの単語も関係のある単語を推測する手がかりとなりうるため，今回はあえて設定するストップワードを最小限にとどめた．そのため，今後辞書の修正やストップワードの追加を行うことで，共起を用いた関連度による組織化の性能はさらに向上するものと考えられる．

4.2.2　時間関連度による組織化の評価

ツイート t_i の前後の地名を含むツイートとの時間間隔に基づいた時間関連度 $Rt(t_i)$ による組織化の評価を行う．

4.2.1 節と同様に，提案手法とベースラインの比較，及び体験 A と体験 B の比較を行った．閾値とする $Rt(t_i)$ を変化させたときの体験 A と体験 B における提案手法とベースラインの F 値の変化をまとめたものが図 3 である．両体験ともに $Rt(t_i)$ の

表 3　$Rt(t_i)$ において F 値が最大の時の再現率と適合率

体験	Rt(t_i)	適合率	再現率	F 値
A 提案手法	0.7020	0.9798	0.9798	0.9798
A ベースライン	—	0.7273	0.0808	0.1455
B 提案手法	0.8626	0.7416	0.9565	0.8354
B ベースライン	—	1.0000	0.0870	0.1600

閾値にかかわらず，提案手法ではベースラインよりも F 値が大きく向上している．

提案手法における体験 A と体験 B の F 値について比較すると，$Rt(t_i)$ の閾値が 0.8 前後〜1.0 のときは体験 B の方が F 値がわずかに高く，$Rt(t_i)$ の閾値が 0.8 以下のときは体験 A の方が F 値が高くなっている．これは，$Rt(t_i)$ が 0.8 以上のときについては，時間間隔を求める際に基準となるツイートそのものの $R(t_i)$ の値がユーザ B の方が高くなっているためであると考えられる．基準となるツイートの $Rt(t_i)$ の値は体験 A では 0.95 前後，体験 B では 0.99 前後となっているが，これは体験 B を表す投稿の中に「祇園」やその他の八坂神社や清水寺の近く地名を含むものがなかったために，$Rt(t_i)$ の計算の際に体験 B では周辺の地名を含むツイートの影響を考慮できなかったことが原因として挙げられる．別のテストデータを用いて検証したり，体験において八坂神社や清水寺から少し離れている地名についても $Rt(t_i)$ の計算の際に考慮に入れるようにすることによって，この範囲における体験間の F 値の差は解消されるものと考えられる．

一方，$Rt(t_i)$ の閾値が 0.8 以下のときに体験 B の方が F 値が体験 A よりも低い原因としては，旅行中に他の話題について述べたツイートが体験 B の方が多いことが挙げられる．両体験を表すツイートのうち $Rt(t_i)$ が高いツイートの内容を検証してみたところ，体験 A では $Rt(t_i)$ が高いツイートのほとんどが正解ツイートであり，旅行中に他の話題について述べたツイートの $Rt(t_i)$ の値はどれも 0.7 前後で他の正解ツイートよりも低い値であった．ところが体験 B では他の話題に述べたツイートについても $Rt(t_i)$ の値が 0.9 前後と他の正解ツイートよりも高い値となっており，こうした不正解ツイートに閾値以上の $Rt(t_i)$ の値が与えられることが体験 B において $Rt(t_i)$ による組織化の性能を低下させている原因であると考えられる．

これらの結果から，旅行中に投稿されたツイートの中に他の話題について述べたツイートが少ない場合は $Rt(t_i)$ だけでも高い精度で組織化が行えるが，旅行中に他の話題についてのツイートも多く投稿されている場合には，$Rt(t_i)$ のみによる組織化では精度が低下してしまうことがわかる．

4.2.1 節と同様に，提案手法において F 値が最大となる時の $Rc(t_i)$ の閾値，適合率，再現率，F 値と，ベースラインにおける適合率，再現率，F 値をまとめたものが表 3 である．この結果からも提案手法によりベースラインよりも性能が向上していること，及び体験 B よりも体験 A の方がより高い精度での組織化が行えることが確認できる．

4.2.3　内容関連度と時間関連度を統合した組織化の評価

前後のツイートの影響は考慮せずに，内容関連度と時間関連度を統合して求めたツイート t_i の関連度 $Rtc(t_i)$ による組織化の

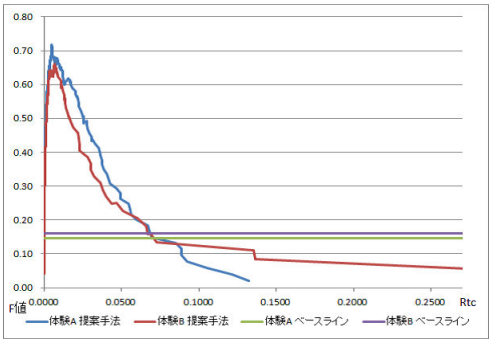


図 4　$Rtc(t_i)$ における提案手法とベースラインの F 値の変化

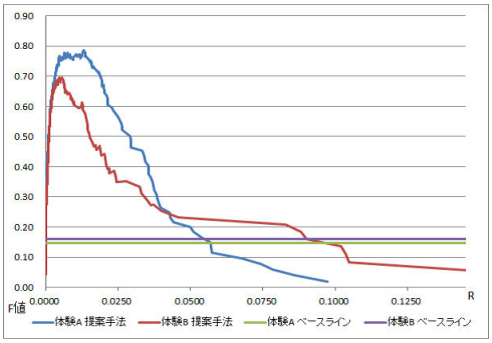


図 5　$R(t_i)$ における提案手法とベースラインの F 値の変化

表 4　$Rtc(t_i)$ において F 値が最大の時の再現率と適合率

体験	Rtc(t_i)	適合率	再現率	F 値
A 提案手法	0.0050	0.6726	0.7677	0.7170
A ベースライン	—	0.7273	0.0808	0.1455
B 提案手法	0.0070	0.7959	0.5652	0.6610
B ベースライン	—	1.0000	0.0870	0.1600

表 5　$R(t_i)$ において F 値が最大の時の再現率と適合率

体験	R(t_i)	適合率	再現率	F 値
A 提案手法	0.0131	0.9189	0.6869	0.7861
A ベースライン	—	0.7273	0.0808	0.1455
B 提案手法	0.0056	0.6957	0.6957	0.6957
B ベースライン	—	1.0000	0.0870	0.1600

評価を行う．

4.2.1 節と同様に，提案手法とベースラインの比較，及び体験 A と体験 B の比較を行った．閾値とする $Rtc(t_i)$ を変化させたときの体験 A と体験 B における提案手法とベースラインの F 値の変化をまとめたものが図 4 である．両ユーザともに，$Rtc(t_i)$ の閾値が約 0.7 以下のときはベースラインよりも提案手法の方が F 値が高くなっている．また，体験 A と体験 B の F 値の比較についても，体験 A の方が体験 B の F 値よりも多少高くなっているものの，閾値の変化による F 値の変化の傾向は両体験ともに類似している．

体験 A と体験 B のツイートのうち $Rtc(t_i)$ の値が高いツイートについて内容を調べてみると，$Rc(t_i)$ による組織化の際に体験 A で起きていた特定の語を含むツイートに高い値が与えられてしまう問題や，$Rt(t_i)$ による組織化の際に体験 B で起きていた旅行中に投稿された他の話題についてのツイートに高い値が与えられてしまう問題がどちらもある程度解消されていることがわかった．この結果から，$Rc(t_i)$ や $Rt(t_i)$ を単独で用いていたときに生じていた問題がツイートの組織化に与える影響が，2 つの指標を組み合わせた $Rtc(t_i)$ を用いることによって低減され，体験 A，体験 B のどちらについても同程度の精度で組織化が行えるようになったことがわかる．

4.2.1 節と同様に，提案手法において F 値が最大となる時の $Rc(t_i)$ の閾値，適合率，再現率，F 値と，ベースラインにおける適合率，再現率，F 値をまとめたものが表 4 である．この結果から，体験 A，B ともに提案手法ではベースラインと比べて F 値が 4〜5 倍程度向上し，適合率の低下をわずかに抑えながら再現率を大きく向上させることができたことがわかる．

4.2.4　コンテキスト関連度も加味した組織化の評価

コンテキスト関連度により前後のツイートの影響を考慮して求めたツイート t_i の関連度 $R(t_i)$ による組織化の評価を行う．今回の実験では，(8) 式において $\tau_i = 0.5$，$\tau_{i-1} = \tau_{i+1} = 0.25$ として $R(t_i)$ の計算を行った．τ_i の値を変化させることで関連度 $R(t_i)$ における組織化の性能も変化することが予想されるが，τ_i の値の変化が関連度 $R(t_i)$ に及ぼす影響の評価は今後の課題とする．

4.2.1 節と同様に，提案手法とベースラインの比較，及び体験 A と体験 B の比較を行った．閾値とする $R(t_i)$ を変化させたときの体験 A と体験 B における提案手法とベースラインの F 値の変化をまとめたものが図 5 である．

ベースラインとの比較では，体験 A は $R(t_i)$ の閾値が 0.06 前後より小さいとき，体験 B は $R(t_i)$ の閾値が 0.09 前後より小さいときに提案手法の方が F 値が高くなっている．

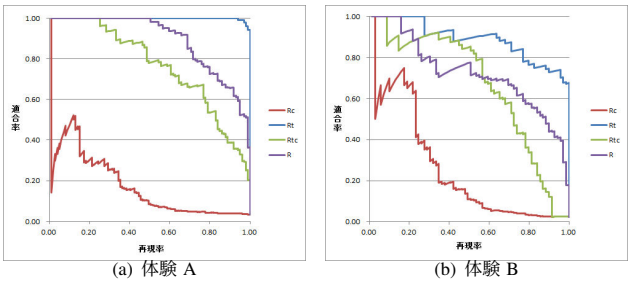
4.2.1 節と同様に，提案手法において F 値が最大となる時の $Rc(t_i)$ の閾値，適合率，再現率，F 値と，ベースラインにおける適合率，再現率，F 値をまとめたものが表 5 である．図 4 と図 5 との比較，及び表 4 と表 5 の比較からは $Rtc(t_i)$ による組織化に比べ $Rt(t_i)$ による組織化の方が F 値が体験 A では約 10%，体験 B では約 5%向上していることが読み取れる．また，このことに伴い $Rtc(t_i)$ による組織化に比べ $R(t_i)$ による組織化では体験 A と体験 B の性能の差もより大きくなっている．

$R(t_i)$ の値が大きいツイートの内容についても検証してみたところ，体験 A では $R(t_i)$ の値が大きいツイートはほとんどが正解となるツイートであり，$Rtc(t_i)$ を用いた組織化よりも性能が向上していることがツイートの内容からも読み取れた．ところが体験 B では，正解ツイートにより高い $R(t_i)$ の値が与えられるようになった点は体験 A と共通しているが，それと同時に $Rtc(t_i)$ では低い値が与えられた旅行中に他の話題について述べたツイートに対しても，前後のツイートの影響を加味することで $R(t_i)$ に再び高い値が与えられるようになってしまっていることがわかった．これらの現象が両体験間の組織化の性能差を生む原因となっていることが考えられる．

4.2.5　指標間の比較

前節までは 4 つの指標による組織化のそれぞれについて性能を評価してきたが，ここでは各指標間の性能を比較する．

体験 A と体験 B の投稿を $Rc(t_i)$，$Rt(t_i)$，$Rtc(t_i)$，$R(t_i)$ の指標



(a) 体験 A　　(b) 体験 B

図 6　適合率-再現率曲線

表 6　適合率が 0.23 以上 0.55 以下の範囲に含まれるツイートの分類

手法	範囲内 件数	正解 件数	他の話題 件数
$Rtc(t_i)$	31	23	2
$R(t_i)$	37	23	7

を使って組織化を行った時の体験ごとの適合率と再現率の変化を示した適合率-再現率曲線が図 6 である．

体験 A については，$Rt(t_i)$ による組織化の性能が極めて高くなっていることがわかる．これは，旅行中の他の話題の投稿が少ないために，時間の連続性を考慮するだけでも求める体験を表すツイートを高い精度で組織化できるためである．また，$Rtc(t_i)$ と $R(t_i)$ については，再現率にかかわらず $R(t_i)$ の方が高い適合率を示している．このことから，体験 A においては関連度を決定する際に前後のツイートも考慮したほうがよい結果が得られるということがわかる．

体験 B については，$Rt(t_i)$ の適合率が全体的に体験 A よりも低くなっており，$Rtc(t_i)$ や $R(t_i)$ との適合率の差も体験 A より少ない．これは，旅行中に他の話題が多く投稿されているために，時間の連続性を考慮するだけでは検索結果に関係ない話題についてのツイートも含まれてしまうためである．

また，$Rtc(t_i)$ と $R(t_i)$ については，再現率が約 0.23〜約 0.55 のときには $Rtc(t_i)$ の方が適合率が高く，それ以外の範囲の再現率においては $R(t_i)$ の方が適合率が高くなっている．その原因を検証するために，$Rtc(t_i)$ による組織化と $Rc(t_i)$ による組織化の結果それぞれに対して，前後のツイートの影響を考慮しない $Rtc(t_i)$ の方が適合率が高くなっている再現率が約 0.23〜約 0.55 のツイートにおいて正解ツイート，旅行中の他の話題に述べたツイートが何件存在するかを調べた．

その結果，表 6 のように $R(t_i)$ の方に旅行中の他の話題についてのツイートがより多く含まれており，本来不正解となるべきツイートについても，直前や直後のツイートの関連度が高いとそのツイートにも高い関連度が与えられて組織化による検索結果に含まれてしまうことがわかった．$R(t_i)$ で再現率がこの範囲に含まれる他の話題についてのツイートの内容を調べると，内容自体は旅行に全く関係のない他のユーザの投稿への返信や，リツイートなどが多く，そのツイート自体の内容の関連性が低くても直前や直後のツイートが関連度の高い正解ツイートであるために $R(t_i)$ の値も高くなってしまったものと考えられる．これを解決するための方法としては，$R(t_i)$ を導出する (8) 式においてより適切な τ_i の値を与える，また，その重み付けを内容に応じて動的に変化させるなどの方法が挙げられる．

$Rc(t_i)$ の性能が他の指標よりも低くなっている点は，4.2.1 節で述べたように辞書の修正などにより形態素解析の精度を向上させることで解消されると考えられる．$Rc(t_i)$ の値は $Rtc(t_i)$ や $R(t_i)$ の値に大きな影響を及ぼすため，$Rc(t_i)$ による組織化の性能向上に伴って $Rtc(t_i)$ や $R(t_i)$ による組織化の性能も向上するものと思われる．

5. ま と め

本研究では，Twitter 上のあるユーザの投稿を旅行体験ごとに組織化する手法について提案した．提案手法では，旅行体験が持つ時間と空間の連続性や内容の関連性，前後のツイートの影響などに着目し，内容関連度，時間関連度，コンテキスト関連度の 3 つの指標を用いて検索したい旅行体験とツイートとの関連度を求め，組織化を行った．

評価実験では，提案手法におけるツイートの組織化の性能について評価した．キーワード検索による組織化と比較した結果では，提案手法の方が高い F 値を示していることが検証され，適合率の低下を最小限に抑えながら高い再現率が得られた．また，最終的な関連度を決定する際に直前や直後のツイートの影響も加味することでより高い精度で組織化することができた．

今後の課題としては，共起度計算の精度向上のための形態素解析の設定の見直し，時空間連続性を考慮した共起辞書の作成，コンテキスト関連度の計算における重み付けの検討，内容関連度の計算におけるリプライやリツイートの利用，地名列への展開の自動化などが挙げられる．

謝　　辞

本研究の一部は，科研費 (No.20300042，No.20300036) の助成を受けたものです．

文　　献

[1] 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（DEIM2010）論文集，2010.

[2] 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（DEIM2010）論文集，2010.

[3] Shaomei Wu，Jake M. Hofman，Winter A. Mason，Duncan J. Watts. Who Says What to Whom on Twitter. In Proceedings of the 20th International World Wide Web Conference（WWW2011），pp.705-714，2011.

[4] Carlos Castillo，Marcelo Mendoza，Barbara Poblete. Information Credibility on Twitter. In Proceedings of the 20th International World Wide Web Conference（WWW2011），pp.675-684，2011.

[5] 倉島健，藤村考，奥田英範．大規模テキストからの経験マイニング．電子情報通信学会 第 19 回データ工学ワークショップ（DEWS2008）論文集，2008.

[6] 牛尼剛聡，渡邉豊英．ライフログ検索における時間粒度を考慮した索引付け．夏のデータベースワークショップ，2005.

[7] 有光淳紀，馬強，吉川正俊．ユーザ体験指向の Twitter 検索手法．第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM2011）論文集，2011.

[8] 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）論文集，2007.