行動の時空間連続性とコンテンツの共有価値を考慮した観光ツイートの組織化

長谷川馨亮† 馬 強† 吉川 正俊†

† 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., qiang@, yoshikawa@}i.kyoto-u.ac.jp

あらまし Twitter 上には観光など多くのユーザ体験コンテンツが公開されているが，それらを整理するための技術は不十分である．そこで本研究では，観光体験における行動の時空間連続性とコンテンツの共有価値を考慮して Twitter 上に断片的に投稿されるユーザの観光体験をまとめて整理する手法を提案する．提案手法では，時空間連続性を考慮して構築した共起辞書を用いて，ツイートの観光体験との関連性及び共有価値を時間・空間・内容の側面から計算して組織化を行う．また，本稿では評価実験を行い，提案手法の精度を評価した．

キーワード マイクロブログ，Twitter，ユーザ体験，共起辞書，時空間連続性，コンテンツの共有価値

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観光であり，観光中のユーザが，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．

Twitter では，ユーザが過去の自分の投稿を全てダウンロードできるようになるサービスが提供されている．(注2)．こうしたデータが広く利用可能になることにより，Twitter に投稿されたコンテンツを整理，検索，共有したいといったニーズがより高まるものと考えられる．

しかし現状では，Twitter に投稿されたコンテンツを有効活用するための技術は発展途上である．そのため，例えば以前の観光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，観光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．

また，ある観光体験についての投稿を検索しようとしても，1 つの投稿あたりの文字制限がある 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 [1] [2] [3].

さらに，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 に書いたことを再度ブログや SNS に投稿しない傾向がある．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い一因である．

こうした問題を解決するために，我々は Twitter 上に投稿されたコンテンツの中から，ある特定の観光体験を表すツイート群をまとめて整理できる手法について研究している [2] [3] [4]．ツイート群をまとめて整理することを，本研究では組織化と呼ぶ．

提案手法では，ユーザが検索したい観光体験を一つ以上の地名や観光地の名前によって指定し，その地名列とユーザのツイートのデータを入力として以下の 2 つの観点から検索したい観光体験とツイートとの関連度を計算する(注3)．

- 内容関連度

ツイートの本文にある観光地との関連が強い単語が含まれていれば，そのツイートはその観光地に関して述べている可能性が高い．そこで，地名とツイート中に含まれる単語の共起度からツイートと観光体験の関連度を計算する．共起度の計算のために，時間と空間の連続性を考慮して作成した共起辞書を使用する．

- コンテキスト関連度

Twitter では 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ため，求める観光体験に関するツイートがあれ

（注1）：http://twitter.com
（注2）：http://blog.twitter.com/2012/12/your-twitter-archive.html
（注3）：実際には，取得できたツイート全てに対して処理を行うわけではなく，観光が行われた期間の前後に投稿されたツイートを指定して取得し，処理を行う．

ば，その前後のツイートも同じ観光体験について記述している可能性が高い．そこで，コンテキストを考慮してツイートと観光体験の関連度を計算する．

我々の先行研究では，これらの関連度により観光体験との関連性の高いツイートをまとめる手法を提案している．しかし，得られた結果の中には，観光体験に関連しているが他人と共有する価値はそれほどないものもある．例えば，「清水寺なう」というツイートは，そのときに清水寺にいたことを表しており京都を観光した体験に関連してはいるが，清水寺にいたという以上の情報は含まれていないため，観光地の感想やそこで体験したことなどを詳細に記したツイートに比べると他人と共有する価値は低いといえる．

そこで本稿では，こうしたユーザが自分の体験を整理する場面においてより他人と共有したいようなツイートであるかを表す概念として共有度を提案する．共有度はツイート本文の特徴や添付画像の有無などにより計算される．

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 [5] 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 [6] 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 [7] 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 [8] 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 [9] 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．しかし Twitter をはじめとするマイクロブログでは，従来のブログと異なり一つの投稿が短く内容が断片化されているため，マイクロブログのコンテンツを整理するためにはブログとは異なる手法が必要となる．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 [10].

CGM の投稿から特定の地域に関する情報を抽出する手法もいくつか提案されている．Yin ら [11] は Flickr 上の位置情報つきの写真から，LGTA(Latent Geographical Topic Analysis) によりテキストの特徴とジオタグの情報を考慮して地域ごとに特徴的な話題を抽出するための手法を提案している．Hong ら [12] は Twitter 上の位置情報付きの投稿から，地域ごとに特徴的な話題を抽出するための手法を提案しており，Yin らの手法よりもより高い精度でのトピックの抽出が可能となっている．これらの手法では投稿に GPS による位置情報が付加されていることが前提となっているが，実際に Twitter 上に投稿されているツイートのうち位置情報が付加されているのはわずか 0.77% (注4)

(注4)：http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

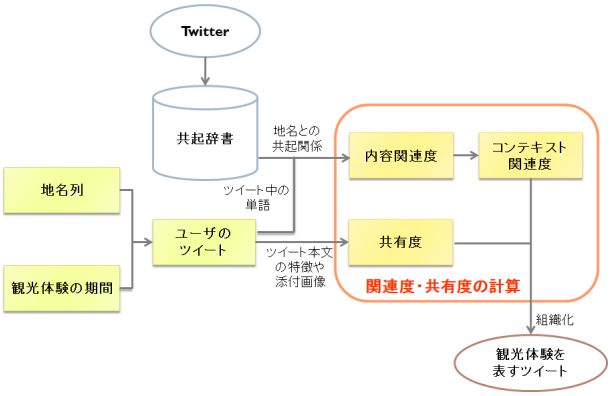


図 1　提案手法によるツイート組織化の流れ

である．我々の手法では GPS による位置情報が付加されていないツイートについても，ツイートの本文や投稿時間などの情報から観光体験を表すツイートを取得することができる．また，Twitter 全体からではなく，あるユーザ 1 人の投稿を対象として観光体験を表すツイートを組織化している点も我々の手法の特徴である．

Twitter からユーザ体験を検索するための手法は有光ら [1] 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遷移していくものとしているが，本研究では観光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，観光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. 提案手法の概要

観光体験は，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．この観光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，本稿における観光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．システムには予め，4. 節の方法で作成した共起辞書を保持させておくものとする．ユーザがシステムに与える入力は検索したい観光体験を表す地名の系列，観光体験が行われた期間，整理する対象とする Twitter のユーザアカウントとし，システムは最終的に検索したい観光体験を表すツイートの系列を出力する．

例えば，あるユーザ A が自分が以前訪れた京都観光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」と観光した期間“(2012.01.01，2012.01.02)”を入力とし，京都観光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，必ずしもその観光体験を表す地名が全て含まれる必要はない．ユーザが探したい観光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観光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，本研究での提案手法により，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ができる．

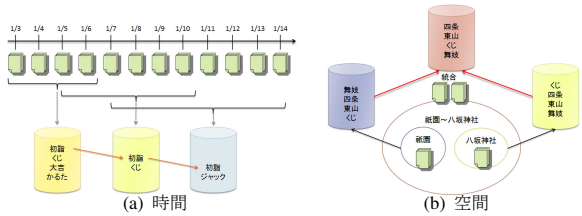
図 1 に示したように，まずシステムは候補となるツイートを収集する．ユーザは観光体験の前後数日間にも観光体験に関するツイートを投稿している場合があるため，候補となるツイートとして観光体験が行われた期間の前後数日間のツイートを対象に検索を行う．候補となるツイートについては，現在は Twitter API を用いて収集しているが，Twitter アーカイブが広く普及してくれば，Twitter アーカイブのデータを入力とする．候補となるツイートが決まれば，各ツイートに対して関連度や共有度を計算し，関連度及び共有度が高いツイートをまとめて最終的な出力結果としてユーザに提示する．

4. 共起辞書

内容関連度を計算するために，地名ごとに他の単語との共起関係を表す共起辞書を作成する．共起辞書を利用することで，その地名との関連が強い単語を判別することができる．ツイートに含まれる単語と観光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観光体験とより強く関連している，という仮定に基づいた内容関連度の計算が可能となる．

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」を問合せ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．

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．京都市内の代表的な地名や観光地の名前をキーワードとして Twitter API(注5) 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から計算される地名とそれぞれの単語の共起度を記録する．この単語と共起度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．



(a) 時間　(b) 空間

図 2　共起辞書における時空間の連続性

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6) 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．

4.1 共起度の計算

まず，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．

$$co(p_k, w) = \frac{T_{p_k \cap w}}{T_{p_k \cup w}} = \frac{T_{p_k \cap w}}{T_{p_k} + T_w - T_{p_k \cap w}} \tag{1}$$

$T_{p_k \cap w}$ は p_k と w を共に含むツイートの個数，$T_{p_k \cup w}$ は p_k と w のどちらか一方のみを含むツイートの個数，T_{p_k}，T_w はそれぞれ p_k，w を含むツイートの個数を表している．

4.2 共起辞書の類似性

次に，2 つの共起辞書を時間及び空間の連続性を考慮して統合するする方法について述べる．

ここで，2 つの共起辞書間の類似度について考える．単語 $w_1, w_2, ..., w_n$ について地名 p_k との共起度 $co(p_k, w_l)$ が記録されている地名 p_k についての共起辞書 d_k をベクトル $\vec{d_k}$ で表すものとする．

$$\vec{d_{p_k}} = (co(p_k, w_1), co(p_k, w_2), \cdots, co(p_k, w_n)) \tag{2}$$

2 つの共起辞書 d_i, d_j の類似度 $S(d_i, d_j)$ は，コサイン類似度を用いて以下のように表される．

$$S(\vec{d_i}, \vec{d_j}) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}||\vec{d_j}|} \tag{3}$$

類似度 $S(d_i, d_j)$ が高いほど二つの共起辞書の内容が類似しているといえる．二つの共起辞書における時間，空間が連続しているとき，その内容も類似していると考えられるので，時間及び空間が近い二つの共起辞書の類似度が一定の閾値以上であればその二つの共起辞書は時間的，及び空間的に連続しているといえるので，その二つを統合した新たな共起辞書を用いて組織化を行うことができる．

(注5)：http://dev.twitter.com/

(注6)：http://code.google.com/intl/ja/apis/maps/

4.2.1 時間の連続性を考慮した共起辞書の統合

観光地ではある特定の時期にイベントが行われたり，その季節ならではの名物が存在したりというように，時間の経過に伴ってその場所での体験は変化していく．そこで，共起辞書の作成に利用するツイートの期間を変化させることで時間の経過による観光地の様子の変化を共起辞書に反映させることができ，より精度の高い組織化が可能になると考えられる．共起辞書における時間の連続性のイメージを表したものが図 2(a) である．

ある地名 p_i についての共起辞書を作成する際に用いるツイートの期間を，1 日，1 週間，1ヶ月といったようにある一定の期間で区切る．ツイートの期間が連続した複数の共起辞書 $(\vec{d^{u_1}_{p_i}}, \cdots, \vec{d^{u_m}_{p_i}})$ を作成する．ここで $u_1, \cdots, u_m$ は，1 日目から m 日目，あるいは 1 週目から m 週目というように，それぞれの共起辞書に含まれるツイートの期間を表す．つまり，$\vec{d^{u_m}_{p_i}}$ は地名 p_i についての期間 u_m における共起辞書を表す．

そして，内容が類似しており，かつ期間が近い共起辞書を統合して新たな共起辞書を作成する．その際に用いる共起辞書間の時間の連続性を考慮した類似度は以下のように計算される．

$$St(\vec{d^{u_j}_{p_i}}, \vec{d^{u_k}_{p_i}}) = int(u_j, u_k) \times S(\vec{d^{u_j}_{p_i}}, \vec{d^{u_k}_{p_i}}) \tag{4}$$

$int(u_j, u_k) = (k - j) \times \delta$ はそれぞれの共起辞書に含まれるツイートの期間の間隔を表し，δ は 1 日，1 週間などの期間の単位を表す．

2 つの共起辞書 $\vec{d^{u_j}_{p_i}}$，$\vec{d^{u_{j+1}}_{p_i}}$ の類似度が一定以上であれば，2 つの共起辞書をマージして新たに共起辞書 $\vec{d^{u'_j}_{p_i}}$ を作成する．この新しい共起辞書における各単語の共起度は以下のように更新される．

$$Tco(p_i, w) = \frac{1}{2}\left(Co^{u_j}(p_i, w) + Co^{u_k}(p_i, w)\right) \tag{5}$$

4.2.2 空間の連続性を考慮した共起辞書の統合

時間の連続性に基づいて統合された各地名における共起辞書 $(\vec{d_{p_1}}, \cdots, \vec{d_{p_n}})$ に対して，さらに空間の連続性を考慮した共起辞書の統合を行う．2 つの地名における共起辞書について，それぞれの場所が近接しており，さらに内容も類似していれば，両共起辞書を統合し 1 つの大きなエリアにおける共起辞書を作成して共起度を計算することで，隣接する地域との空間の連続性を考慮した関連度が求められるようになる．共起辞書を統合する際のイメージを表したものが図 2(b) である．

まず，2 つの地名 p_i, p_j についての共起辞書 d_{p_i}, d_{p_j} 間の類似度を以下のように計算する．

$$Ss(\vec{d_{p_i}}, \vec{d_{p_j}}) = dist(p_i, p_j) \times S(\vec{d_{p_i}}, \vec{d_{p_j}}) \tag{6}$$

ここで, $dist(r_i, r_j)$ は p_i と p_j の空間的連続性を考慮した重み付けであり, 戸田らの研究 [13] や崔らの研究 [14] など多くの研究で用いられている指数関数逓減モデルを p_i と p_j の距離 $d(p_i, p_j)$ に適用したものである．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．

$$dist(p_i, p_j) = e^{-\mu_d d(p_i, p_j)} \tag{7}$$

2 つの共起辞書間の空間の連続性を考慮した類似度が一定以上であれば，これらの共起辞書を統合して新たに共起辞書 $\vec{d_{r_{i,j}}}$ を作成する．この新しい共起辞書における各単語の共起度は以下のように計算される．

$$\begin{aligned} co(p_i, w) &= \frac{1}{2}\left(co(p_i, w) + dist(p_i, p_j) \times co(p_j, w)\right) \\ co(p_j, w) &= \frac{1}{2}\left(co(p_j, w) + dist(p_i, p_j) \times co(p_i, w)\right) \end{aligned} \tag{8}$$

5. 関連度の計算

地名列 $(p_1, \cdots, p_n)$ で表される観光体験 $E = (p_1, \cdots, p_n)$ に関連するツイートを組織化するために，以下で述べる内容関連度，コンテキスト関連度によってツイートと観光体験との関連度を計算する．

5.1 内容関連度

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．共起度の計算には，4. 節の方法で作成した共起辞書を用いる．検索したい観光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，観光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観光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観光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．

ツイート t_i に含まれる単語を $w_1, \cdots, w_m$ とすると，ツイート t_i と観光体験 E との内容関連度は以下のように計算される．

$$Rc(t_i) = \sigma + \sum_{j=1}^{n}\sum_{k=1}^{m} co(p_j, w_k) \tag{9}$$

σ は，共起辞書に存在する単語を含まないツイートなどにおいて $Rc(t_i)$ の値が 0 となるのを防ぐために加える値である．

5.2 コンテキスト関連度

Twitter では一つの話題が連続した複数のツイートに分割されて投稿される場合が多い．この特徴を利用して，以下で述べるコンテキスト関連度により周辺のツイートの内容を考慮してツイートと観光体験の関連度を求める．

実際には求める観光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ために，そのツイート中の単語のみを用いて内容関連度を計算すると，内容関連度の値が極めて低くなる場合がある．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 と検索したい観光体験とのコンテキスト関連度 $Rx(t_i)$ は以下の (10) 式で求められる．

$$Rx(t_i) = \sum_{x=-X}^{X} \tau_{i+x} Rc(t_{i+x}) \tag{10}$$

$$\tau_{i+x} = e^{-\mu_t |time(t_i) - time(t_{i+x})|} \quad (11)$$

(11) 式は，計算対象となるツイート t_i の投稿時間 $time(t_i)$ とその X 件後のツイート t_{i+x} の投稿時間 $time(t_{i+x})$ の差から計算される時間の連続性を考慮した重みである．

同じ話題について連続して述べた複数のツイートは，投稿時間が近接している場合が多い．観光体験においても同様に，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，観光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，観光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，前後のツイートの投稿時間が計算対象のツイートの投稿時間と近いほどその前後のツイートは計算対象のツイートの内容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，計算対象のツイートと前後のツイートとの時間差に適用した (11) 式により，時間の連続性を考慮した関連度を求める．

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．

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した場合のコンテキスト関連度 $Rx(t_i|X = 1)$ と，それに加えて 2 つ前，2 つ後の影響も考慮した場合のコンテキスト関連度 $Rx(t_i|X = 2)$ を利用するものとする．

6. 共　有　度

6.1 共有度の計算

関連度により観光体験に関連するツイートを決定することができるが，他人と共有する価値の低いツイートも含まれる可能性が高い．

例えば，先に述べたように，観光体験について述べたツイートの中には「京都なう」のように単純にユーザがいた場所のみを記述しているもののように 1 つのツイートだけで見るとあまり情報量が多いとはいえないものもあれば，観光地を訪れた感想などを詳細に記述しているものや観光地の写真を添付したものなど 1 つのツイートあたりの情報量が多いものもある．

観光体験を共有・整理する場合においてはより情報の多いツイートの方がより過去の体験の振り返りやコミュニケーションのきっかけ作りにおいて有用であると考えられるため，ここでは以下のようなツイートが共有・整理する価値の高いツイートであると仮定し，その抽出手法を提案する．

- 観光の中での体験や感想などが詳細に記述されたツイート
- 観光地で撮影された写真や動画などが添付されたツイート

各ツイートがどの程度共有・整理する価値があるかを示すための指標として共有度を提案する．共有度の計算の際には，ツイート本文の特徴とツイートへのメディアファイルの添付の有無を考慮する．ツイート本文の特徴としては，ツイート本文中の単語数が多いほど，またその中でも特に形容詞や副詞，形容動詞，連体詞のように感情や様子などを表す単語がより多く含まれているほど共有度が高くなるものとする．また，ツイートに写真や動画などのメディアファイルが多く添付されているものほど共有度が高くなるものとする．

ツイート t_i の共有度 $S(t_i)$ は以下のように計算される．

$$S(t_i) = \log(words(t_i) + \phi\, pic(t_i)) \quad (12)$$

$$words(t_i) = \alpha\, mod(t_i) + other(t_i) \quad (13)$$

(13) 式における $mod(t_i)$ はツイート t_i 中に含まれる形容詞，形容動詞，副詞，連体詞の語数の合計であり，$other(t_i)$ はそれ以外の品詞である単語の語数である．形容詞，形容動詞，副詞，連体詞は感情などを表す単語であり，こうした単語が多いツイートはより共有する価値が高いと仮定する．そのための重みづけが α である．(12) 式における $pic(t_i)$ はツイート t_i に添付された画像の件数である．ϕ は画像 1 枚がツイート中の単語何語分に相当するかを表す値である．

6.2 関連度と共有度の統合

観光体験との関連が高く，なおかつ共有する価値が高いツイートを決定するために，5.2 節で求めたコンテキスト関連度と 6. 節で求めた共有度を統合して最終的な関連度の値 $RS(t_i)$ を計算する．

$$RS(t_i) = Rx(t_i) \times S(t_i) \quad (14)$$

$RS(t_i)$ の値の高い順にツイートを表示することで，ユーザに観光体験との関連性が高く，なおかつ他人との共有価値も高いツイートを提示することができる．

7. 実　　験

提案手法によるツイートの組織化の性能を評価する実験を行った．

実験で用いるデータには，Twitter API を用いて京都の地名や観光地の名前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 などを取得して利用した．

7.1 提案手法における組織化性能の評価

7.1.1 実 験 方 法

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観光体験を検索することを想定した．MeCab を用いてツイートの本文に対し形態素解析を行い，単語や地名を各ツイートから抽出した．

今回の評価実験では，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中から手動で選定したユーザ U_a，U_b，U_c の観光体験 E_a，E_b，E_c 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．各ユーザのツイートのうち，観光体験が行われた日とその前後 3 日ずつ，計 7 日間に投稿されたツイートを実験のテストデータとして用いた．共起度の計算には各ユーザの観光体験が行われた日を含む期間である 2012 年 1 月 2 日から 17 日にかけて投稿されたツイートから作成した共起辞書を用いた．Twitter API を用いてユーザ U_a，

U_b，U_c のツイートを取得し，取得した各ツイートについて節で述べた基準に基づき手動で正解，不正解の判断を行った．正解かどうかの判断の方法は 7.1.2 節で述べる．

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．ユーザ U_a，U_b，U_c のツイートについて $Rc(t_i)$，$Rx(t_i)$，$S(t_i)$，$RS(t_i)$ 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．

7.1.2 正解データ

今回の評価実験で用いたテストデータのうち，正解となるツイートを決定するために，テストデータ中の各ツイートが求める観光体験と関連しているか，及び共有する価値がどの程度あるかの判断を，人手により 5 段階で行った．観光体験に関連する度合いを表す関連ランクと，ツイート自体に供する価値の度合いを表す共有ランクを設定し，それぞれ評価が最高のものを 5，最低のものを 1 とした．

観光体験に関連しているかどうかは，組織化したい観光体験に依存する，つまりクエリ依存であるが，共有する価値があるかどうかは，クエリに関わらずツイートが与えられると一意に決まる，つまりコンテンツ依存である．

7.1.3 関連度を用いた組織化性能の評価

3 ユーザの観光体験を表すツイートの関連度を求める提案手法による組織化とベースラインであるキーワード検索による組織化について，適合率-再現率曲線を用いて組織化性能を比較した．この実験では，関連ランクが 4 以上であるツイートを正解データとして適合率，再現率，F 値を計算した．

提案手法としては，(9) 式による内容関連度 $Rc(t_i)$，コンテキスト関連度 $Rx(t_i)$ については直前直後のツイートの影響のみを考慮する (10) 式において $X = 1, \mu_t = 1$ としたもの，$X = 1, \mu_t = 10$ としたもの，2 つ前，2 つ後のツイートの影響も考慮する $X = 2, \mu_t = 10$ としたものの 4 つを比較した．

この実験では，(9) 式，(12) 式，(13) 式におけるパラメータをそれぞれ $\sigma = 0.01, \phi = 100, \rho = 0.1, \alpha = 5$ としている．

表 1 は 4 つの提案手法において F 値が最大となる時の各関連度の閾値，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 値をまとめたものである．

ベースラインとの比較ではどの観光体験においても，内容関連度については F 値が平均して 3 倍程度に，コンテキスト関連度については F 値が平均して 4 倍程度になっている．また，適合率も 4 倍から 10 倍程度向上していることがわかる．

次に，4 つの提案手法についてそれぞれ比較する．図 3, 4, 5 はそれぞれ観光体験 E_a，E_b，E_c における 4 つの提案手法の適合率–再現率曲線を表している．内容関連度 $Rc(t_i)$ とコンテキスト関連度 $Rx(t_i)$ を比較すると，F 値ではどの観光体験においても平均して 2 割程度向上している．適合率については，再現率

表 1　関連度 : F 値が最大の時の適合率と再現率

観光体験	閾値	適合率	再現率	F 値
E_a $Rc(t_i)$	0.1007	0.5287	0.4792	0.5027
E_a $Rx(t_i \mid X = 1, \mu_t = 1)$	0.5355	0.6484	0.6146	0.6310
E_a $Rx(t_i \mid X = 1, \mu_t = 10)$	0.5204	0.6860	0.6146	0.6484
E_a $Rx(t_i \mid X = 2, \mu_t = 10)$	0.6901	0.7582	0.7188	0.7380
E_a ベースライン	—	0.7273	0.0833	0.1495
E_b $Rc(t_i)$	0.1085	0.5952	0.4167	0.4902
E_b $Rx(t_i \mid X = 1, \mu_t = 1)$	0.2148	0.4607	0.6833	0.5616
E_b $Rx(t_i \mid X = 1, \mu_t = 10)$	0.2135	0.4659	0.6833	0.5541
E_b $Rx(t_i \mid X = 2, \mu_t = 10)$	0.4874	0.4940	0.6833	0.5734
E_b ベースライン	—	1.0000	0.1000	0.1818
E_c $Rc(t_i)$	0.0000	0.3874	1.0000	0.5584
E_c $Rx(t_i \mid X = 1, \mu_t = 1)$	0.0484	0.4111	0.8605	0.5564
E_c $Rx(t_i \mid X = 1, \mu_t = 10)$	0.0474	0.4805	0.8605	0.6167
E_c $Rx(t_i \mid X = 2, \mu_t = 10)$	0.0837	0.5263	0.9302	0.6723
E_c ベースライン	—	1.0000	0.0930	0.1702

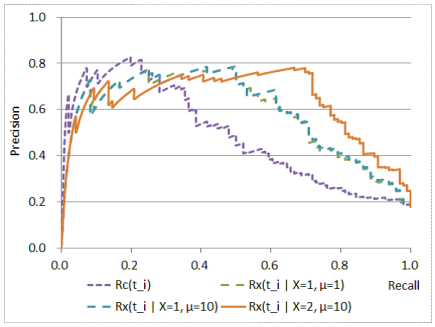


図 3　関連度 E_a : 適合率-再現率曲線

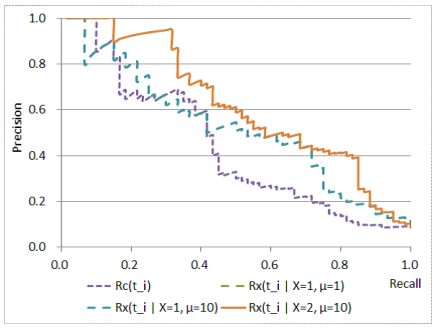


図 4　関連度 E_b : 適合率-再現率曲線

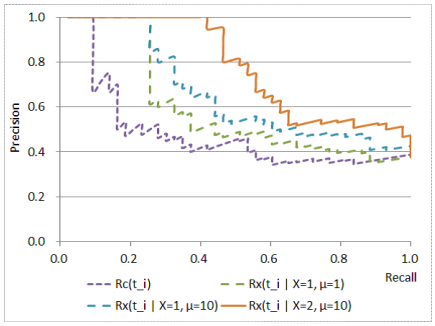


図 5　関連度 E_c : 適合率-再現率曲線

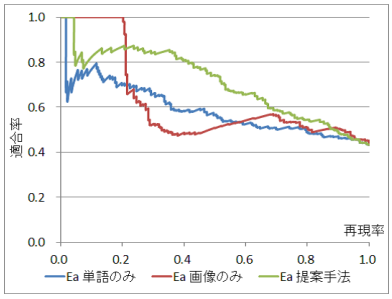


図 6　共有度 Ea : 適合率-再現率曲線

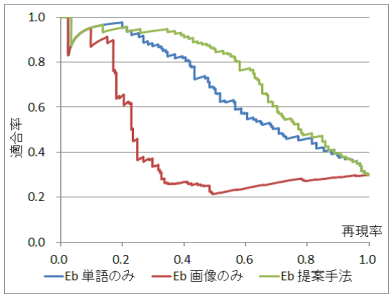


図 7　共有度 Eb : 適合率-再現率曲線

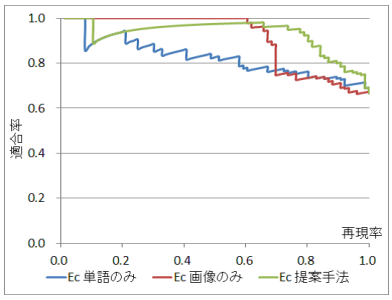


図 8　共有度 Ec : 適合率-再現率曲線

が 0.4 以上になるとどの観光体験でも適合率が向上していることがわかる．

コンテキスト関連度 $Rx(t_i)$ について (11) 式におけるパラメータを変化させたときの影響について検討してみると，μ_t の値を 1 から 10 に変化させると，観光体験 E_c において F 値と適合率が向上した．

これらの結果から，内容関連度とコンテキスト関連度を用いた提案手法により観光ツイートの組織化の性能が大きく向上したことが確認できた．

7.1.4　共有度を用いた組織化性能の評価

三ユーザの観光体験を表すツイートの共有度を求める提案手法による組織化について，適合率-再現率曲線を用いて組織化性能を比較した．共有度 $S(t_i)$ を求める提案手法では，(12) 式で示したようにツイート本文の単語の特徴から計算された $words(t_i)$ と添付画像の有無により計算された $pic(t_i)$ から共有度を計算しているが，ベースラインとして $words(t_i)$ の値のみを用いた組織化の結果と $pic(t_i)$ の値のみを用いた組織化の結果を用いる．この実験では，共有ランクが 3 以上であるツイートを正解データとして適合率，再現率，F 値を計算した．またこの実験では，(12) 式，(13) 式におけるパラメータは $\phi = 50, \alpha = 5$ としている．

表 2　共有度 : F 値が最大の時の適合率と再現率

観光体験	閾値	適合率	再現率	F 値
E_a 提案手法 : $S(t_i)$	1.3222	0.5447	0.8405	0.6610
E_a 単語のみ : $words(t_i)$	1.0000	0.4581	0.9655	0.6214
E_a 画像のみ : $pic(t_i)$	0.0000	0.5099	0.8922	0.6489
E_b 提案手法 : $S(t_i)$	1.5052	0.7665	0.6244	0.6882
E_b 単語のみ : $words(t_i)$	1.3222	0.5305	0.6780	0.5953
E_b 画像のみ : $pic(t_i)$	0.0000	0.2993	1.0000	0.4607
E_c 提案手法 : $S(t_i)$	1.3979	0.9242	0.8026	0.8592
E_c 単語のみ : $words(t_i)$	0.8451	0.7143	0.9868	0.8287
E_c 画像のみ : $pic(t_i)$	0.0000	0.6726	1.0000	0.8042

図 6, 7, 8 はそれぞれ観光体験 E_a，E_b，E_c における提案手法とベースラインの適合率–再現率曲線を表している．観光体験 E_b については，再現率が約 0.3 以下の範囲では提案手法とツイート本文の単語の特徴のみを考慮した，つまり $words(t_i)$ のみを用いた手法の適合率が，それ以外の範囲では提案手法の適合率が最も高くなっている．

観光体験 E_a については，再現率が約 0.2 以上の範囲では提案手法の適合率が，約 0.2 以下の範囲では $pic(t_i)$ のみを用いた，つまり添付画像の有無のみを考慮した手法の適合率が最も高くなっている．これは，今回は画像が添付されているデータは全て正解データとなるため，画像が添付されており $pic(t_i)$ の値が高いツイートが多く含まれる再現率 0.2 以下の範囲では適合率が極めて高くなり，再現率を上げていき画像が添付されていないツイートが多く含まれるようになると適合率が大幅に低下する．一方，提案手法では再現率を上げていっても適合率が 0.8 前後で安定している．

同様に，観光体験 E_c についても提案手法と添付画像の有無のみを考慮した手法の適合率の値が類似している．これは，観光体験 E_c の最中にユーザが投稿したツイートの大半に画像が添付されていたためである．

表 2 は提案手法とベースラインにおいて F 値が最大となる時の各関連度の閾値，適合率，再現率，F 値をまとめたものである．提案手法と二つのベースラインを比較すると，三つの観光体験全てにおいて提案手法により適合率は平均約 1.5 倍に，F 値は平均約 1.1 倍に向上したことがわかる．

これらの結果から，ユーザの投稿の特徴にかかわらず共有度の高い投稿を組織化するためには，ツイート本文の単語の特徴と添付画像の有無を同時に考慮した提案手法が最適であることが分かった．

7.2　共起辞書の作成法

内容関連度の計算に用いる共起辞書について，時空間連続性を考慮して作成した共起辞書を用いた組織化と時空間連続性を考慮せずに作成した共起辞書を用いた組織化による組織化の性能を F 値の最大値により比較した．

時空間連続性を考慮して作成した共起辞書としては，観光体験 E_a，E_b，E_c について各体験が行われた日のツイートから共起辞書を作成し，八坂神社と祇園，清水寺と高台寺についてそれぞれ共起辞書を統合したものを用いた．時空間連続性を考慮せずに作成した共起辞書としては，2012 年 1 月 2 日から 17 日

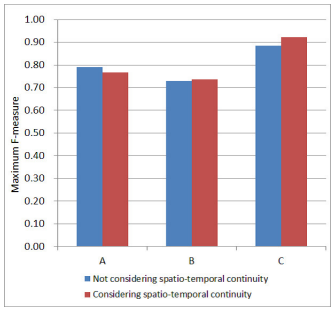


図 9　共起辞書作成の際の時空間連続性の考慮の有無による F 値の最大値の変化

にかけた投稿されたツイートから作成された共起辞書を用い，八坂神社と清水寺についてそれぞれの周辺の場所における共起辞書との統合は行わなかった．三つの観光体験についてそれぞれ時空間連続性を考慮して作成した辞書とそうでないものを用意し，それぞれの辞書を用いて組織化を行い性能を比較した．

図 9 は共起辞書構築時の時空間連続性の考慮の有無による F 値の最大値の変化を示したものである．三つの観光体験それぞれについて結果を比較すると，観光体験 C では時空間連続性を考慮した方が約 4%F 値の最大値が高かったが，観光体験 B ではほとんど差がみられず，観光体験 A では時空間連続性を考慮しない方が約 3%F 値の最大値が高かった．

以上の結果からは，共起辞書構築時の時空間連続性の考慮の有無による組織化性能の変化についての明確な結論は得られなかった．より正確な結果を得るためには，データ数を増やしたさらに大規模な実験が必要となるが，それは今後の課題である．

もし共起辞書構築時に時空間連続性を考慮する方法を改善することで関連度の計算の精度が向上するならば，より高精度なツイートの組織化が行えるようになる．共起辞書構築の方法の改善策の一つに，共起辞書を統合する際にそれぞれの辞書を作る際に使用したツイートの数を考慮して共起度を正規化したうえで新しい共起辞書での共起度を計算することが挙げられる．各地名における共起辞書の作成には，地名をキーワードとして Twitter 上を検索して得られたツイートを利用しているが，清水寺や祇園といったように有名な地名では 1 日に数百件のツイートが得られるのに対し，高台寺などのようにそれほど有名でない地名では 1 日に数十件のツイートしか得られないというように，場所によって投稿されているツイートの数に差が見られた．そのため，Jaccard 係数により共起度を計算すると，ツイート数が少ない地名の共起度が高くなるといった現象が確認された．こうした場所によるツイート数の差は本稿では考慮していないため，各場所のツイート数を考慮して正規化を行うことでより正確に共起度を計算できる可能性がある．

8. ま と め

本稿では，Twitter 上に投稿されたユーザ体験に関するコンテンツの検索や共有を支援するために，ある観光体験に関するツイートを組織化し，共有価値の高いツイートを優先的にユーザに提示するための手法を提案した．

提案手法の最大の特徴は，観光体験における行動の時空間連続性を考慮して組織化を行う点である．Twitter の投稿には文字数制限があるため，観光体験に関連しているが観光体験を表す地名などを含んでいない投稿も多数存在する．そこで提案手法では，ツイートと観光体験との関連度を，ツイート本文中の単語と地名との共起関係と，前後のツイートの影響を考慮して計算している．また，地名とツイート本文中の単語の共起関係を求める際に用いる共起辞書を地名ごとに作成し，その際に時空間連続性を考慮してその場所の特徴を共起辞書に反映させるための手法も提案した．さらに，ツイート本文の特徴などを考慮して共有価値の高いツイートを求める手法も提案した．

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，キーワード検索による組織化よりもより高い精度での組織化が行えることが検証された．また，周辺のツイートの影響を考慮することでさらに組織化の性能が向上することも確認された．

今後の課題としては，共起辞書作成時に時空間連続性を反映させることによる組織化性能への影響を評価するための追加実験や，関連度及び共有度計算のための手法の全体的な改良，提案手法によるツイートの組織化を利用したアプリケーションの作成などが挙げられる．

文　献

[1] 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.
[2] 長谷川馨亮，馬強，吉川正俊．行動の時空間連続性を考慮した旅行ツイートの組織化．第 4 回データ工学と情報マネジメントに関するフォーラム（DEIM2012）論文集，2012.
[3] Hasegawa, K., Qiang M., Yoshikawa M.: Trip Tweets Search by Considering Spatio-temporal Continuity of User Behavior. In: DEXA, (2012) 141–155
[4] 長谷川馨亮，馬強，吉川正俊．時空間連続性を考慮した共起辞書構築による旅行ツイートの組織化．第 22 回 Web インテリジェンスとインタラクション研究会（SIG-WI2）論文集，2012.
[5] 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.
[6] Fujisaka, T., Lee, R., Sumiya, K.: Discovery of user behavior patterns from geo-tagged micro-blogs. In: ICUIMC, (2010) 246–255
[7] Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: WWW, (2011) 705–714
[8] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: WWW, (2011) 675–684
[9] Kurashima, T., Fujimura, K., Okuda, H.: Discovering Association Rules on Experiences from Large-Scale Blog Entries. In: ECIR, (2009) 546–553
[10] Ushiama, T., Watanabe, T.: An Automatic Indexing Approach for Private Photo Searching Based on E-mail Archive. In: KES(2), (2006) 1111–1118
[11] Yin, Z., Cao, L., Han, J, Zhai, C., Huang, T.,: Geographical topic discovery and comparison. In: WWW, (2011) 247–256
[12] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In: WWW, (2012) 769–778
[13] Toda, H., Kitagawa, H., Fujimura, K., Kataoka, R.: Topic structure mining using temporal co-occurrence. In: ICUIMC, (2008) 236–241
[14] Cui, C., Kitagawa, H.: Topic activation analysis for document streams based on document arrival rate and relevance. In: SAC (2005) 1089–1095