ユーザ体験指向のTwitter検索手法

有光　淳紀† 　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: †arimitsu@db.soc.i.kyoto-u.ac.jp, ††{qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし　本稿では，Twitterにおいて断片化されているユーザ体験（就職活動など）に関する情報を体系化して検索する手法を提案する．複数行動からなるユーザ体験をキーワード列で表現し，つぶやき（ツイート）の時系列性や語の関連性およびユーザ体験を構成する行動の遷移を考慮して，ユーザ体験単位でつぶやきをまとめて検索・組織化する手法を提案する．また，実験結果についても述べる．

キーワード　マイクロブログ, Twitter, 時系列データ

A User Experience-oriented Microblog Retrieval Method

Junki ARIMITSU†, Qiang MA††, and Masatoshi YOSHIKAWA††

† Undergraduate School of Informatics and Mathematical Science Science, Faculty of Engineering, Kyoto University　Yoshida-honmachi, Sakyo, Kyoto 606-8501 Japan
†† Graduate School of Informatics, Kyoto University　Yoshida-honmachi, Sakyo, Kyoto 606-8501 Japan
E-mail: †arimitsu@db.soc.i.kyoto-u.ac.jp, ††{qiang,yoshikawa}@i.kyoto-u.ac.jp

Abstract　In this paper, we propose a novel retrieval method to search for information about user experience (job-hunting, etc.) from the Twitter. We represent the user experience as keyword sequence and propose a novel method to search for and organize Tweets(posts on Twitter) by per-user experience considering time series feature of Tweets, term relatedness, and transition pattern of actions which constitute user experience. This paper also describes the experimental results.

Key words　Microblogging, Twitter, Time-series data

1. はじめに

近年，インターネット上の電子掲示板やブログ，SNS（Social Networking Service）などを通じて，ユーザ自身が情報の作成・発信・伝達を行うメディア，いわゆるCGM (Consumer-Generated Media) がネット社会に爆発的に広まってきている．SNSの代表格であるFacebook(注1)や画像共有サービスのFlickr(注2) などにおいて，ユーザが自身の体験に関する日記や写真を投稿することも当たり前のことになっており，今後もユーザの発信するコンテンツはますます増えていくと予想される．とりわけ，その中でもTwitter(注3)を始めとするマイクロブログ（ミニブログ）と呼ばれるサービスは，ユーザの身近にあったことや考えていることを簡単に投稿することができ，リアルタイムに情報が発信できるという特性から，日本や欧米で急速な普及を見せている．Twitterでの投稿（つぶやき）は，2010年9月時点で1日9000万件にも達し，様々な個人の意見や体験がリアルタイムで蓄積されていっている．こうした膨大な情報を整理・分析することで，例えばウェブ上で注目される話題の提示や，商品の評判検索，トラブル事例の検索など，ユーザの意思決定や学習につながる様々な活用法の登場が期待されている．

Twitterの特徴としては，簡単に投稿ができる一方，140字という字数制限があるため，投稿されたつぶやきは断片的な情報となり，そのままでは経験の共有には向いてない．一般的な日記のような形式を取るブログと比較すると，ブログではユーザは日に何度も更新せず情報をある程度まとめて書くことが多いが，字数制限のあるTwitterではブログのような自由な記述はできないため，まとめるといった整理・分類が必要になる．

そこで，本研究ではTwitterにおけるユーザのつぶやきから，ユーザ体験を検索・組織化する手法を提案する．ユーザ体験とは，ユーザのつぶやきの中で関連性の強いイベント・行動の系列・集合のことであり，各ユーザの体験をまとめることができれば，自身のつぶやきを体験ごとに組織化したり，他のユーザ

(注1)：http://www.facebook.com/
(注2)：http://www.flickr.com/
(注3)：http://twitter.com/

の体験に関するつぶやきをまとめて検索したりすることが可能である．例えば，あるユーザの「就職活動」に関する情報をまとめることで，ユーザ自身が自分の「就職活動」の体験をまとめて閲覧でき，また他のユーザはそれを参照することができる．行動は,「エントリー」「説明会」といったように，ユーザ自信に関連している活動のことであり，Twitter の投稿上のキーワードはその行動を行った結果に登場する．体験を行動の系列・集合とすることで,「エントリー」「説明会」「面接」などの「就職活動」を表すキーワードによって体験を特徴づけることができる．

既存のキーワードベースの手法では，マイクロブログのような断片的なコンテンツからユーザの欲しい情報を獲得することが困難な場合がある．例えば,「就職活動」で検索しても，ユーザ体験が一連の行動からなるため，就職活動を明記してない多くのつぶやき（面接や内定などに関するもの）が検索されない．「面接」というキーワードで検索をしても,「アルバイトの面接」などの就職活動に関係ない情報が検索されたり,「適正」「キャリア」「グループディスカッション」といった就職活動に関わりがありそうな情報は検索結果に現れない場合もある．また，140 字以内という制限により,「就職活動の合同説明会に行ってきた.」「A 社の企業理念や社風は〜な印象だ.」というように体験談が分けてつぶやかれることも多く，情報が断片化することで体験の文脈を追うことが難しくなるという難点もある．そこで，我々はユーザ体験を表す一連の行動の時系列性と，また Twitter のつぶやきにも時系列性があるという点に着目した．検索対象に含まれる語の関連性や時系列性，ユーザ体験を構成する行動の遷移という観点を用いることで，前後の文脈を含んだ検索結果が得られると考えている．

提案手法は，以下の二つのステップからなる．

（1） ユーザ体験を記述している可能性のあるつぶやきの候補の抽出

ユーザのつぶやきの中から，検索したいユーザ体験を記述している可能性のあるつぶやきを抜き出す．まず，複数行動からなるユーザ体験をキーワード列で表現することで体験を特徴づける．例として,「就職活動」は「エントリー」,「説明会」,「面接」,「内定」という行動から特徴づけられる．一般的にこれらのキーワード列は「エントリー」が済んだら「説明会」に参加し,「面接」を受けて,「内定」に至るという大まかなプロセスがあるため，これらのキーワードが，この順で登場するつぶやきを抽出することで，目的のユーザ体験の候補が抽出できる．こうして，体験を系列で特徴づけることで同じキーワードを含む別の体験との差別化を図る.「アルバイトの面接」「大学（院）の説明会」といった他の体験では，キーワードはこの順序関係に基づかないため，同じキーワードを含む「就職活動」の体験と差別化することができ,「就職活動」に関係が深いと思われる候補が抜き出せると考えられる．キーワードを検索し，その中である一定の期間に体験を特徴づけるキーワードをすべて含む部分を見つけ出し，キーワードを含むつぶやきとその前後のつぶやきを候補とする．

（2） 候補のつぶやきとユーザ体験との関連度を用いたユーザ体験コンテンツの抽出

ステップ 1 で抽出された候補の中から目的のユーザ体験を判別するステップである．ここでは内容関連度と，内容関連度に時系列の影響を導入したコンテキスト関連度を用いて，つぶやきとユーザ体験の関連性を図っている．内容関連度は，つぶやきに含まれる各単語とユーザ体験のキーワード列との共起頻度をそれぞれ計算し求める．ここで着目するのはキーワードの遷移に応じて，共起しやすい単語も変わってくる点である．例えば「就職活動」において,「エントリー」というキーワードが頻出する周辺では,「締切」「応募」といった単語が共起しやすく,「面接」「内定」というキーワードが頻出する周辺では,「落ちた」「通った」といった語が共起する場合が多いと考えられる．この考えに基づいて，関連度を計算する際に重要視するキーワードをかえることで，より適切に内容関連度を求められると思われる．コンテキスト関連度とは，内容関連度につぶやき同士の相互影響を考慮した関連度である．Twitter 上の投稿は初めにも述べたように内容が断片化していることが多いが，つぶやき同士の時間が近ければ近いほど話題が類似しているという特徴がある．従って，ユーザ体験との関連性が高いつぶやきの内容関連度を投稿時間が近いつぶやきに伝搬させることで，コンテキスト関連度を計算することができる．

上記手順により，内容関連度からコンテキスト関連度を計算し，数値が高いものをユーザ体験コンテンツと考えられるつぶやきだと判定する．

実験では，あらかじめ与えられた「就職活動」に関するキーワード列を検索質問として，キーワードを含むつぶやきだけを検索した場合と提案手法を用いて検索した場合を比較し，関連度に対する閾値を変化させることで再現率と適合率の違いを観察した．結果，既存の手法では検索できない「就職活動」に関するつぶやきを検索することができた．

以下，2 節では Twitter の特徴と関連研究を紹介する．3 節と 4 節では，それぞれ提案手法とその評価実験について述べる．最後に 5 節でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 Twitter に関する関連研究

Twitter に対する研究は近年盛んに行われている．既存の研究では，Twitter におけるつぶやきやユーザのクラスタリングや分析などの研究 [1] [2] [3] が行われている．Java らは Twitter の分析を行い，ユーザが互いにどのようにつながっているかについて提示した [4]．Kwak らはフォローする理由やつぶやきの内容など Twitter の利用方法について様々な統計分析を行った [5]．O ’Connor らは Twitter に対する検索手法を提案した．この検索手法は，クエリを含むつぶやき (エントリ) の集合と共に，その集合内の頻出単語を返し，また得られたエントリ集

合を各頻出単語を含むか否かでグループ化するという特徴を持つ [6]．Sakaki らは，Twitter のユーザをセンサーと見立て，地震などのリアルタイムイベントを発見する手法を提案している [7]．高村らは本研究と同じく Twitter の時間的特徴に着目し，時間軸上にある短文書要約手法を提案している [8]．また，Togetter は Twitter 上の話題や体験を手動でまとめることができ，我々の研究と目的を同じくしているサービスである [9]．我々の研究は，個々のユーザを対象にしており，体験を表す検索質問を入力として与えることで，キーワードを明記していないつぶやきに対しても自動的にまとめることができるという点で，このサービスと異なっている．

2.2 経験マイニングに関する関連研究

文書からの個人の経験の抽出については経験マイニングという研究が行われている．Inui らは個人の経験を述語構造に基づく表現形式で構造化し，経験クラスに分類することで経験データベースを構築するプロジェクトを行っている [10]．倉島らは経験を状況，行動，主観からなる情報と捉え，「興味深さ」を評価する尺度を用いてデータの分析を行った [11]．いずれも，大規模データからのマイニング手法であり，全体的な傾向や相関性を分析するには効果的であるが，断片化した個々人の体験の組織化には適していない．

2.3 時系列データに関する関連研究

本研究で扱うデータは投稿時間を持った時系列文書である．時系列文書に対する代表的な研究では，Topic Detection and Tracking（TDT）プロジェクトなどがある [12]．このプロジェクトでは，時系列文書からトピックを持つ文書を特定する研究が行われている．提案された手法はニュース記事などの大量の語が含まれる文書においては有効ではるが，Twitter のような単語の出現頻度が少ないデータに対しては効果は期待できない．崔らは時系列文書のトピックに対する関連性と到着頻度を考慮し，文書ストリームに対する話題の活性度分析を行った [13]．戸田らは同様に関連性と時間的近さを考慮し，話題の構造を抽出する手法を提案している [14]．いずれも到着頻度や時間的近さを考慮しており，我々の研究と着眼点が似ているが，我々は行動の遷移という点にも着目している点が異なる．平野らはウェブサイトの利用パターンやカテゴリ同士の関連性より，ウェブアクセスの時系列相関性を調べている [15]．また，文書データに特化しない一般の時系列データに対する検索手法としては，Sakurai ら [16] や豊田ら [17] の研究がある．

3. 提 案 手 法

提案手法では，Twitter におけるつぶやきに対し，検索したい体験のキーワード列を入力とし，目的のユーザ体験に関連性の高いつぶやきの列を出力とする．本来ならユーザ体験を自動的にキーワード列に展開することが必要だが，今回はユーザ体験を表すキーワード列を入力とする．また，ユーザ体験は複数の行動からなり，これらの行動には順序関係があるとする．ここで，ユーザ体験を e，キーワードを k_i とすると，キーワード列によるユーザ体験は以下のように表される．

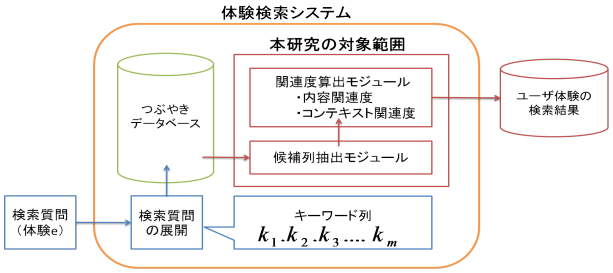

図 1　提案手法の概要

$$e = k_1.k_2....k_m$$

またこのキーワード列は $k_1 \rightarrow k_2 \rightarrow \cdots \rightarrow k_m$ という順序関係も持っている．例えば「就職活動」というユーザ体験は (エントリー．説明会．面接．内定) で表現でき，エントリー→説明会→面接→内定という順序関係を持っている．

図 1 は提案手法の概要を示す．検索質問の展開は今回の研究には含まず，ユーザ体験を表すキーワード列をつぶやきデータに対する入力とし，候補列の抽出と関連度の計算をすることで，ユーザ体験を検索結果に出力する．

ユーザ体験の検索は次の手順で行う．

(1) ユーザ体験コンテンツの候補となるつぶやき列を抽出
つぶやきからユーザ体験について記述している可能性がある候補を抽出する．k_i を含むつぶやきとその前後のつぶやきを候補列とする．

(2) 内容とコンテキスト関連度を用いてユーザ体験コンテンツを判別
候補のつぶやきとユーザ体験の関連度を求めて，ユーザ体験に関連するつぶやき（列）を抽出する．本研究では，以下の二つの側面からつぶやきとユーザ体験（検索質問）の関連を計算している．

(2-1) 内容関連度
つぶやきとユーザ体験（検索質問）との関連度を，キーワード列の類似に基づいて計算する．

(2-2) コンテキスト関連度
つぶやきの時系列特徴から，つぶやき同士の相互影響を考慮したコンテキスト関連度を計算する．

以下にその詳細を記す．

3.1 ユーザ体験コンテンツ候補の抽出

まず，ユーザ体験と関連性の高いつぶやきはユーザ体験を表すキーワード列の周辺にあると考え，与えられたつぶやきデータから候補列を抽出する．Twitter での対象ユーザのつぶやきを $T = t_1.t_2.t_3...t_n$ とし，検索したい体験を $e = k_1.k_2.k_3....k_m$ とする．ただし，これらのキーワード列は与えられるものとする．つぶやき t_i から単語を取り出したとき，検索したいユーザ体験のキーワード列の項目 k_i を含むつぶやきとその前後 N 時間のつぶやきを候補として抽出する．また，会話関係にあるつぶやきが同じ話題に関して言及している可能性が高いため，対象ユーザのフォローとフォロワーのつぶやきから，上記の候補

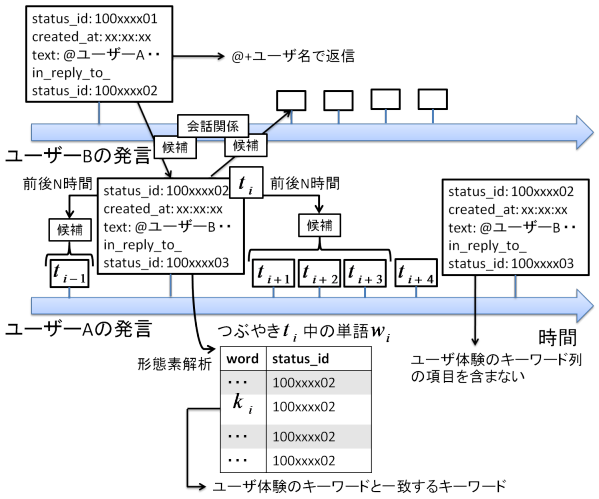

図 2　ユーザのタイムライン

との会話関係にあるものも候補列とする．

図 2 の例では，t_i が検索したいユーザ体験 e $(= k_1.k_2.k_3....k_m)$ のキーワード k_i を含むため，t_i の前後 N 時間のつぶやき $t_{i-1}, t_{i+1}, t_{i+2}, t_{i+3}$ を候補としている．また，つぶやきには図 2 のように，つぶやきを区別する固有の ID（$status\_id$）や，投稿時間（$created\_at$）が必ず付加されており，返信先の ID（$in\_reply\_to\_status\_id$）が付加されている場合もある．対象のつぶやきと会話関係にあるつぶやきも同じ話題について言及していると考えられるので，候補とする．なお，投稿中に返信の記述が登場するのに，返信先の ID がない場合は直近の発言への返信とみなして，前処理している．

3.2　共起語と状態遷移からの内容関連度の計算

つぶやき t_i とユーザ体験 e の内容関連度を，つぶやきに含まれているキーワードと，ユーザ体験のキーワード列にあるキーワードとの共起関係を用いて求める．また行動の遷移を考慮してキーワードとの内容関連度に重み付けをすることを考える．ユーザの行動は時間や状況と共に変化し，それに伴って登場するキーワードも変化していく．その時間や状況をまとめて状態と定義すると，状態によって登場しやすいキーワードも遷移していくことになる．このキーワードの登場しやすさを状態に応じた重みの値を大きくすることで計算式に表現することができる．

i 番目のつぶやき t_i とキーワード w との共起度を $Rc(t_i, w)$(式 (2) で定義する) とすると，体験 e との内容関連度 $Rc(t_i, e)$ は次のように表される．

$$Rc(t_i, e) = \alpha_1 Rc(t_i, k_1) + \alpha_2 Rc(t_i, k_2) + \cdots + \alpha_m Rc(t_i, k_m) + \xi Rc(t_i, k_e) \quad (1)$$

式 (1) は体験 e と t_i の内容関連度をキーワード列との関連度の和で表している．ただし，k_e はユーザ体験 e を表すキーワードである（前述の例でいう「就職活動」というキーワード）．キーワード k_e 自体もユーザ体験 e をよく特徴づけていると考えられるので，k_e と t_i の関連度も式 (1) に加算している．

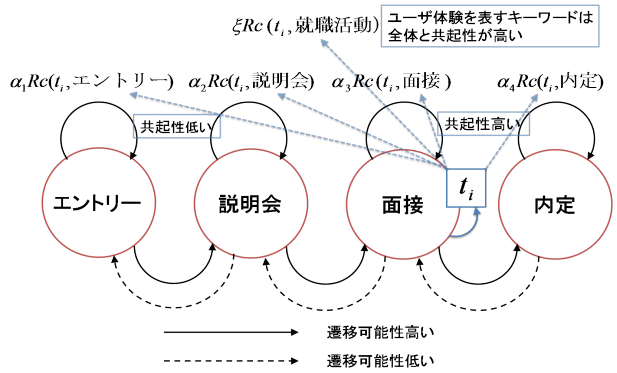

図 3　就職活動における行動の遷移

$\alpha_1, \alpha_2, \cdots, \alpha_m, \xi$ は状態係数である．状態係数や関連度の式に関しては以下で詳しく説明する．

状態係数は，t_i とその直前のつぶやき候補に含まれているユーザ体験 e のキーワード列の項目で決められる．例えば，直前のつぶやきが「面接」との関連性が高ければ，現在のつぶやきが「面接」および「内定」との関連性が高いと考えられ，「面接」や「内定」との関連性を計算するための状態係数を高く設定する．ユーザ体験 e のキーワード列には順序関係があることから，ユーザの行動が図 3 のように遷移していくと考えられる．つぶやき t_i が図 3 のように「面接」というキーワードを含むつぶやきの直後に投稿された場合，もう一度「面接」という行動が行われるか，次の「内定」という行動を行う実線の方向に遷移する可能性が高く，ひとつ以上前の行動にさかのぼるような破線の方向には遷移する可能性は低い．従って，「エントリー」「説明会」というキーワードとの関連度より，「面接」「内定」というキーワードとの関連度を重視した方が，より文脈に沿った値に補正できると考える．図 3 の場合は，$\alpha_1 > \alpha_2 > \alpha_3 \geqq \alpha_4$ のように値を設定することで，内容を重視した関連度の計算を行うことができる．一般的には，t_i の直前のつぶやき t_{i-1} に k_i が含まれていれば，$\alpha_i \geqq \alpha_{i+1} > \alpha_{i+2} > \cdots$，かつ，$\alpha_i \geqq \alpha_{i+1} > \alpha_{i-1} > \alpha_{i-2} > \cdots$ のように，値を設定する．ξ はユーザ体験 e を表すキーワード k_E の係数であり，全体的に共起性が高いと考えられるため，最も高い値を設定させる．また，キーワードと t_i の関連度は以下の式で定義している．

$$Rc(t_i, w) = \frac{\sum_{k \in t_i} cooc(k, w)}{N_k} \quad (2)$$

$$cooc(k, w) = \frac{df(k, w)}{df(k) + df(w) - df(k, w)} \quad (3)$$

k はつぶやき t_i 中に登場する単語であり，N_k はつぶやき中の単語の総数である．(2) 式は t_i 中に登場する各単語とキーワード w の共起度，(3) 式は個々の単語とキーワード w の共起度であり，(2) 式では個々の共起度の合計を語句の数で割って平均をとることで，単語数の大小による値の格差を抑えた値を出すことができる．また，$df(k)$ は k を含むつぶやきの数であり，$df(k, w)$ は k と w が同時に登場するつぶやきの数である．(3) 式は 2 つの語のうちどちらか一方が登場した数のうち，2 語が同時に登場した割合を表し，一般的にジャッカード (Jaccard) 係数という名称で集合の共起を表すのによく用いられる．

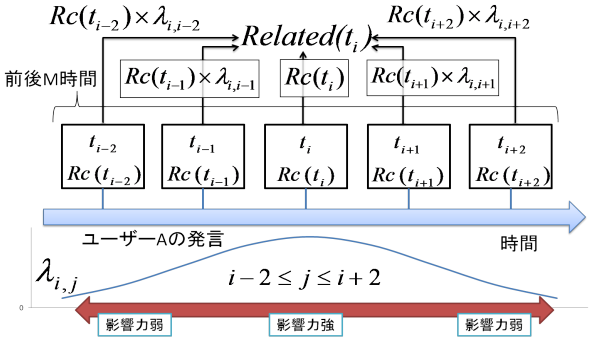

図 4　他からの影響を考慮したコンテキスト関連度の計算

3.3　時系列特徴を考慮したコンテキスト関連度の計算

つぶやきの話題はつぶやき同士の時間が近ければ，内容も似ていると考えられる．3.2 節で計算された内容関連度は，つぶやきと体験 e の共起情報と直前のつぶやきの状態を考慮した尺度であるが，つぶやきに含まれる単語に状態に応じた共起の高いキーワードが出現しない場合を考慮しきれていない．例えば，同時間系列の話題の中でたまたま共起の低い単語しか含まれていなかった場合，内容関連度は低い値が計算されてしまう．そこで本節では，話題の影響を他のつぶやきに伝搬させたコンテキスト関連度を考える．

コンテキスト関連度とは，内容関連度に他のつぶやきの内容関連度の影響を合計した値であり，他のつぶやきからの影響は時間が近くなればなるほど大きくなり，離れれば離れるほど小さくなると考える．情報の影響力の時間逓減の割合は，既存の研究 [13] でもしばしば用いられている指数関数逓減モデルから求める．つぶやきの影響力は t 時間経過ごとに，$\lambda = \mathrm{e}^{-\mu t}$ のように逓減するとみなす．ここで μ は時間的逓減の割合を決定するパラメタである．ユーザ体験 e との内容関連度が $Rc(t_i, e)$ となるつぶやき t_i が投稿されてから時間 t 経過後につぶやき t_j が投稿されるすると，そのつぶやきに $Rc \times \lambda_{i,j}$ を加えることで，時系列を考慮したコンテキスト関連度を再計算できる．$\lambda_{i,j}$ はつぶやき t_i に対する t_j の逓減率である．以下に具体的な式を記す．

$$Related(t_i, e) = Rc(t_i, e) + \sum_{j \in around_i} Rc(t_j, e) \times \lambda_{i,j}$$

$$\lambda_{i,j} = \mathrm{e}^{-\mu|time(t_i) - time(t_j)|}$$

ただし，$around_i$ は t_i を除いた，t_i の前後 M 時間のつぶやきである．また，$time(t_i)$ はつぶやき t_i の投稿時間である．M をある程度大きく取ることで時間的に離れたものの影響は微差とあると考えられるので影響範囲を制限している．図 4 はコンテキスト関連度計算のモデルである．t_i の前後 M 時間の範囲のつぶやきを $t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}$ とすると，t_i の内容関連度に，時間逓減率をかけたものの合計がコンテキスト関連度となる．

t_i と検索対象のユーザ体験 e とのコンテキスト関連度 $Related(t_i, e)$ が高ければ，t_i がユーザ体験について記述していると判断する．

4.　実　　験

4.1　実験環境

提案手法の評価実験を行った．評価実験では，Twitter API を用いて，Twitter 上から対象ユーザとユーザの following/followers のつぶやきに関する情報をすべて取得した．取得した情報は，つぶやきの ID（status_id），投稿時間（created_at），つぶやきの本文（text），返信先（in_reply_to_status_id）などである．MeCab(注4) によって，取得したつぶやきの本文を形態素解析し，助詞や接続詞などを除いて単語共起を計算するためのデータベースを作成する．作成されたデータベースは，452 人のユーザ，627,789 件のつぶやき，収録語数は 7,208,565 語となった．@junki0526 から作成されたデータセットは 2501 件となった．

検索したいユーザ体験は (エントリー．説明会．面接．内定) とし，ユーザ体験と判断できる正解集合は手動で選別し，データセットで使用したつぶやき 2501 件中，134 件あると集計した．各つぶやきの関連度 $Rc(t_i$, 就職活動), $Related(t_i$, 就職活動) の値が閾値 θ を超えるときユーザ体験と判断し，判定結果に対して，再現率と適合率，F 値を比較する．再現率は検索結果に含まれる正解の件数を正解集合の件数で割って計算し，適合率は検索結果に含まれる正解の件数を検索結果の件数で割って計算した．F 値は再現率を r と適合率を p とすると以下の式で計算される．

$$F = \frac{2r \times p}{r + p}$$

最後に，実験結果の適合率，再現率，F 値などの値から状態遷移を考慮する妥当性，時系列を考慮する妥当性について考察を述べる．

提案手法を適用する前に，使用するデータセットに対して，就職活動に関連するキーワードの出現回数を計測し，図 5 のようにグラフ化した．図 5 は横軸に時刻を 1 週間刻みで設定し，縦軸にその 1 週間の間に発言されたキーワードの件数を置いている．図 5 を見ると，各キーワードにより登場回数に差はあるが，概ね 就職活動=(エントリー．説明会．面接．内定) の順にキーワードが登場していっているのがわかる．これにより，就職活動にキーワードの遷移があることが裏付けられた．今回の実験では，就職活動が盛んに行なっていた時期（2 月から 4 月まで）とその前後 1ヶ月（1 月と 5 月）のデータを対象とした．

4.2　共起語と状態遷移を考慮した内容関連度に関する実験

ここでは，内容関連度を計算するために用いた共起という指標と状態遷移という指標が検索結果に及ぼす影響を検証する．内容関連度は 3.2 で定義したように語の共起から計算したスコアに，状態係数を乗じた項の和で計算される．実験では以下の 2 点に関して検証を行う．

実験 1　内容関連度の適用が検索結果に及ぼす影響

キーワードベースで検索した場合（ベースライン）と内容関連度を求めた場合で，再現率・適合率を比較する．

(注4)：オープンソースの形態素解析エンジン http://mecab.sourceforge.net/

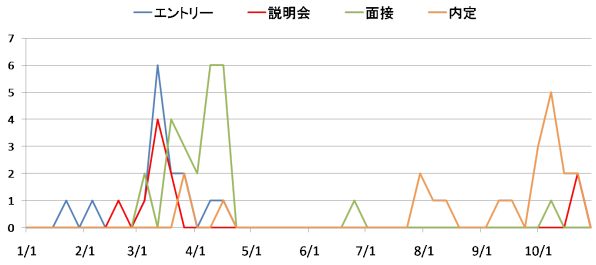

図 5　ユーザ junki0526 におけるキーワード出現回数

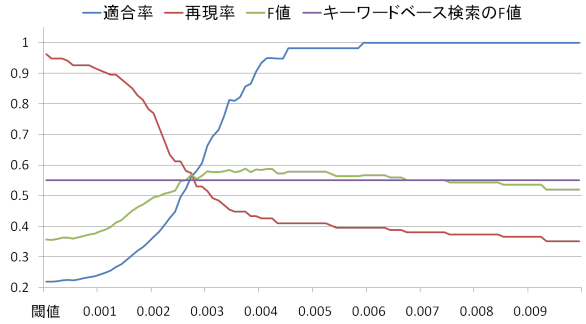

図 6　内容関連度を用いた場合とベースラインの比較

表 1　内容関連度を用いた場合の再現率と適合率

実験	閾値	再現率	適合率	F 値
ベースライン	—	0.3805	1	0.5513
提案手法	0.0038	0.8571	0.4477	0.5882

実験 2　状態遷移の考慮が検索結果に及ぼす影響

内容関連度を求める際，状態係数 $\alpha_i (1 \leqq i \leqq 4)$ を均一にした場合と状態遷移を考慮し状態係数を変化させた場合で，再現率・適合率を比較する．

まず，内容関連度の有効性を検証する実験 1 を行った．4.1 節で抽出された候補列から，就職活動=(エントリー．説明会．面接．内定) のキーワード列を含むつぶやきを検索すると，検索結果は 51 件となった．そのうち正解集合に含まれるつぶやきは 51 件だった．適合率は 1，再現率は 0.3805，F 値は 0.5513 であった．次に，3.2 節の計算式から内容関連度を計算し，閾値を変化させて適合率と再現率，F 値の変化を以下の図 6 にまとめた．上記で計算したベースラインの F 値も図 6 上に表示している．閾値が 0.003 から 0.007 の辺りにかけて，ベースラインの F 値を，提案手法の F 値が上回っている．

また，ベースラインの F 値と提案手法の F 値で最も高かった結果を表 1 のようにまとめた．提案手法の F 値は閾値が 0.0038 の時最も高く 0.5882 だった．これはベースラインの F 値を 0.0369（約 3.7%）上回る結果となった．

次に，状態遷移について検証する実験 2 を行った．状態とは，3.2 節で定義されるように，そのつぶやきがどのキーワードの直後にあるかという指標であり，例えば α_1 はキーワード「エントリー」をどれだけ重視するかを表しており，エントリーが現れた直後のつぶやきで α_1 の値を大きくすれば，状態遷移を考慮していることになる．

表 2　状態係数値を一定

キーワード	α_1	α_2	α_3	α_4	ξ
エントリー	2.5	2.5	2.5	2.5	5.0
説明会	2.5	2.5	2.5	2.5	5.0
面接	2.5	2.5	2.5	2.5	5.0
内定	2.5	2.5	2.5	2.5	5.0
就職活動	2.5	2.5	2.5	2.5	5.0

表 3　状態係数値を変化

キーワード	α_1	α_2	α_3	α_4	ξ
エントリー	4.5	4.5	0.5	0.5	5.0
説明会	1.0	4.5	3.5	1.0	5.0
面接	0.5	1.0	4.5	4.0	5.0
内定	0.5	0.5	4.0	5.0	5.0
就職活動	2.5	2.5	2.5	2.5	5.0

表 4　状態係数を変化させた場合の再現率，適合率，F 値の比較

実験	閾値	再現率	適合率	F 値
状態係数の影響を考慮しない	0.0038	0.8571	0.4477	0.5882
状態係数の影響を考慮する	0.0042	0.8857	0.4626	0.6078

キーワード列の項目に対する状態係数の値 $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \xi$ は，3.2 節のモデルに従い，t_i とその直前のつぶやき候補に含まれているユーザ体験のキーワード列の項目（表 3 の一番左の列）によって設定している．例えば，α_3 は「面接」というキーワードをどれだけ重視するかを表しており，「面接」が直前に現れる場合，次に遷移するのは「面接」か「内定」という状態の可能性が高いという推定から，「面接」に対応する係数 α_3 と「内定」に対応する係数 α_4 の値を大きくすることで，状態遷移の影響を考慮していることになる．体験を表すキーワード「就職活動」に対応する係数 ξ は常に高く設定するため，「面接」の直後にあるつぶやきの状態係数は $\xi > \alpha_3 \geqq \alpha_4 > \alpha_2 > \alpha_1$ となるように値を設定している．そこで，状態係数の影響を評価するために状態係数を一定にした場合と状態係数を変化させた場合について実験を行った．状態係数の値は表 2，表 3 のように設定をした．表 2 は実験 1 で使用した係数値であり，いずれのキーワード列も均等に重要視するため $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ となっている．表 3 は状態の重みを変動させた場合の係数値である．表 3 では，値が大きいほど対応する状態を重視しているということを意味する．

状態係数を固定した場合と，変化させた場合の再現率，適合率曲線を描いた結果を図 7 に示す．再現率，適合率曲線は x 軸に再現率，y 軸に適合率をプロットしたグラフで，適合率が同じ時に再現率を比較したり，逆に再現率が同じ時に適合率を比較したりすることができ，性能の比較に役立つ．この図 7 を見ると，両端では値は変わらないが，再現率が 0.5〜0.8 のときなどは状態係数を変化させた場合の方が全体的に適合率の値が改善されることがわかった．そのうち，それぞれ F 値が最大値を取ったときの値を表 4 のようにまとめた．F 値の最大値を比較したところ，状態係数を考慮した場合の方が 0.0196（約 2%）上回っていた．

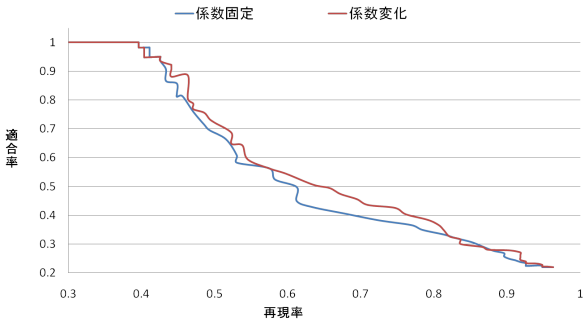

図 7　実験 2 の再現率，適合率曲線

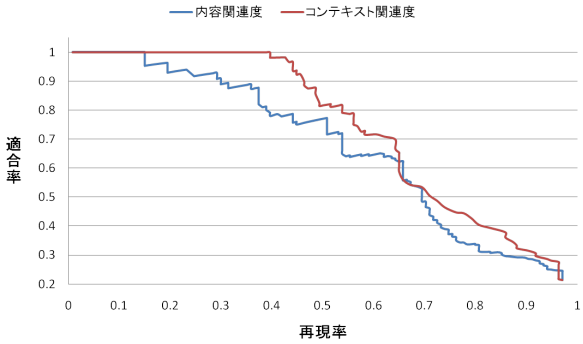

図 8　実験 3 の再現率，適合率曲線

表 5　実験 3 の再現率，適合率，F 値の比較

実験	閾値	適合率	再現率	F 値
内容関連度（実験 2）	0.0038	0.8571	0.4477	0.5882
コンテキスト関連度（実験 3）	0.0042	0.6991	0.6417	0.6692

実験結果としては，実験 1 で約 3.7%，実験 2 で約 2% の F 値の改善が見られ，キーワードベースで検索した手法と比較し，状態遷移を考慮した内容関連度を用いた手法では，F 値が約 5.7% 改善することがわかった．実験で検索されたデータを確認すると，「就職活動」に関連するつぶやきなのに，内容関連度が低く計算されてしまう場合が見受けられた．この原因としては，以下の点が考えられる．

- 断片化された会話中で登場するため，共起する単語が含まれていない．
- 辞書の単語の偏りのため就職活動に関する共起単語の値が低く計算されている．
- Twitter のつぶやきには口語などの崩れた表現やネットスラング（インターネット上で使用される特有の隠語）などが含まれており，同じ意味を持つ単語が違う単語として辞書登録されてしまう．

特に Twitter は気軽に投稿ができる分，崩れた表現が多数混入してしまいがちだ．精度向上は全体的な課題でもあり，共起表現が正しく取られることで，内容関連度を用いた手法の再現率も向上していくと考えられる．

4.3　コンテキスト関連度に関する実験

ここでは，内容関連度につぶやきの相互影響を考慮したコンテキスト関連度の効果を検証するために次の実験を行う．

実験 3　時系列特徴が検索結果に及ぼす影響

実験 2 で得られた状態遷移を考慮した内容関連度を用いた場合と，内容関連度に時間的影響を考慮したコンテキスト関連度を用いた場合で，再現率と適合率を比較する．

情報の影響力は内容関連度につぶやき間の時間から求まる逓減率をかけることで，計算できる．情報の影響範囲は前後 $M = 24$ 時間とし，指数関数逓減のパラメタ値を $\mu = 2$ とした．実験 3 の結果は図 8 のようになった．表 5 のようになった．

実験結果としては，内容関連度を導入した実験 1 や状態遷移を考慮した実験 2 よりさらに F 値の改善が見られた．F 値だけを比較すると，ベースラインの F 値が 0.5513 だったのに対し，状態遷移と時系列特徴を用いた方法では，F 値は 0.6692 と大幅に改善した．図 8 や表 5 を見ると，F 値が最大になった周辺では再現率の値が大幅に上昇しており今まで検索されなかった体験のコンテキストが検索結果に現れてきているとわかる．

実際に抽出されたデータを検証すると，就職活動に関して断片化している話題が組織化できていることが判明した．内容の関連度とコンテキスト関連度の 2 つの視点を用いることで，断片化しているつぶやきの組織化・検索が可能であるとわかった．しかし，一方で提案手法でも検索結果に現れない例があった．例えば，つぶやきの単語数が少なく，内容関連度が極小化する場合や，また，ユーザ体験との関連性は高くないのに，関連度計算の結果が高くなってしまう例もあった．コンテキスト関連度では時間逓減率のパラメタ μ を極端に大きくして影響範囲を広げすぎたことで逆に精度が下がってしまう結果も現れた．他人への返信や引用などのつぶやきは，自分の状態に関係なく現れることも多く，単純に時間的近さだけを考慮してコンテキスト関連度を計算すると，ノイズとして現れてしまうのである．

これらの実験では，提案手法により全体的な検索結果の改善が見られ，特に時系列特徴によるコンテキスト関連度を用いた手法で，大きな F 値の改善が見られた．1 節で述べたように，ユーザ体験を検索するには，従来のキーワードベースの手法では，適合率は非常に高いが，再現率が低いという問題があった．実験 3 の結果から，提案手法は適合率として約 70% の精度をもちながらも，再現率を 1.7 倍近くまで改善させることができたことがわかった．

5.　ま　と　め

本研究では，Twitter のつぶやきをユーザ体験単位でまとめる手法について提案した．提案手法では，語の関連性や時系列性，ユーザ体験を構成する行動の遷移に着目し，ユーザ体験とつぶやきとの内容関連度，コンテキスト関連度という尺度でユーザ体験を検索した．つぶやきの内容や状態遷移に着目した内容関連度を用いた実験を実験 1・2 で行い，時系列特徴に着目したコンテキスト関連度を用いた実験を実験 3 で行った．その結果，キーワードベースの検索手法が適合率 100%，再現率 38%，F 値 55% だったのに対し，実験 3 では適合率 64%，再現率 70%，F 値 67% という結果を得た．つぶやきのような断

片化したデータをまとめるためには，キーワードベースの検索手法では困難であり，状態遷移によるキーワードの重要度の変化や時系列性を考慮した関連度を用いる手法が有効であることが検証された．

今後，以下のような課題について検討し提案手法の改良を行う予定である．

(1) ユーザ体験のキーワード列の自動発見
今回行った予備実験では，ユーザ体験を表すキーワード列を人手で定義した．人手で定義することでユーザ体験を表すキーワードを正しく設定することができるが，本来ならこのシステムを使用するユーザは「就職活動」というキーワードを入力するだけで，自分でキーワード列を設定することはしない．したがって，ウェブ上のシソーラスやWikipedia などから (エントリー．説明会．面接．内定) というキーワード列を自動発見することが必要である．

(2) ユーザ体験候補の自動抽出
予備実験でキーワード列を含むつぶやきデータを候補として使用した．キーワードの出現間隔が一定時間以上ならそこまでのデータを体験候補とするよう設定したのだが，キーワードの出現頻度の境が曖昧な場合や体験の時間的範囲が狭い場合には個別に抽出する部分を選択しなければならない．したがって，キーワードの抽出範囲を設定する手法が必要である．キーワードの密度などを計算し，キーワード系列が順序関係を持ちながら密集している部分を自動で抽出するなどの手法が考えられる．また，キーワードが (エントリー．説明会．面接) までしか登場しない場合，つまり内定というキーワードが登場しなかった場合にどう判断するかも設定しなければならない．

(3) 状態遷移の隠れマルコフモデル化
今回，状態遷移は登場したキーワードによって状態遷移図として定義していたが，これを隠れマルコフモデルに基づく確率過程に拡張することを考える．隠れマルコフモデルとは内部状態がわからない場合に外部から観測される単語の系列から，内部の状態遷移を確率的に類推できるモデルである．特に，表には状態がはっきりと出現しない時系列解析手法などで用いられる．隠れマルコフモデルはビタビアルゴリズムという手法によって，キーワードを出力した可能性が最も高い状態列を求めることができる．また，確率分布を粒子と呼ばれるサンプル点で近似し状態推定を行う粒子フィルタなど，状態推定には様々な方法があるため，実験でどの推定が有効か検証してみる必要がある．

(4) 追加実験
まず，今回は「就職活動」というユーザ体験に関して，@junki0526 を対象に実験を行ったが，他のユーザに関しても同様に組織化することができるか確認する必要がある．また,「就職活動」以外の体験に関しても同様の組織化が可能かも検証しなければならない．追加実験の対象としては，関連研究でも紹介した togetter [9] でまとめられているつぶやきを対象に実験を行い，精度を検証することで，より実利的なシステムになると考える．

謝　辞

本研究の一部は，科研費（20700084 と 20300042) の助成を受けたものです．

文　献

[1] 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.

[2] 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 (DEIM2010) 論文集, 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 (DEIM2010) 論文集, 2010.

[4] Java, A., Song, X., and Tseng. ,T. F. B. Why we twitter:understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, 2007.

[5] Kwak, H., Lee, C., Park, H. and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web(WWW 2010)*, pp. 591–600, 2010.

[6] O'Connor, B., Krieger, M. and Ahn, D. Tweetmotif: Exploratory search and topic summarization for twitter. In *In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 384–385, 2010.

[7] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of 18th International World Wide Web Conference (WWW2010)*, pp. 851–860, 2010.

[8] 高村大也, 横野光, 奥村学. Summarizing microblog stream. 人工知能学会第 22 回 SWO 研究会 SIG-SWO-A1001-03, 2010.

[9] Togetter. http://www.togetter.com/.

[10] Inui, K., Abe, S., Morita, H., Eguchi, M., Sumida, A., Sao, C., Hara, K., Murakami, K. and Matsuyoshi, S. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[11] 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.

[12] Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[13] 崔春花, 北川博之. 到着頻度と関連性を考慮した時系列文書のトピック分析. *DBSJ Letters, Vol.3, No.2, September 2004*, pp. 37–40, 2004.

[14] 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.

[15] 平野真太郎, 成凱, 岩井原瑞穂. 階層型カテゴリを用いたウェブサイトのアクセス履歴の時系列相関性解析. 電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005) 論文集, 2005.

[16] Sakurai, Y., Faloutsos, C. and Yamamuro, M. Stream monitoring under the time warping distance. In *Proceedings of the 23th IEEE International Conference on Data Engineering (ICDE)*, pp. 1046–1055, 2007.

[17] 豊田真智子, 櫻井保志, 市川俊一. データストリームのための部分シーケンスマッチング. 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 (DEIM2009) 論文集, 2009.