

contextHashtag の提案とそれを用いた Twitter ユーザ向けイベント推薦システム

海江田 隆博^{†1} 黄 宏 軒^{†2} 川 越 恭 二^{†2}

Twitter の Hashtag によりイベント情報を得る際に、Hashtag 間に相互関連がないため適切なイベントが得にくい。そこで、Hashtag を構造化した多次元空間を持つ contextHashtag を提案する。contextHashtag の多次元空間内で既存 Hashtag を領域として自動的に位置付けることで、過去イベントと類似したユーザが望む未来イベントを領域間類似性により求めることが可能となる。contextHashtag を用いたイベント推薦システムの試作について説明する。

Proposal of contextHashtag and An Event Recommendation System for Twitter Users Using the contextHashtag

TAKAHIRO KAIEDA,^{†1} HUNG-HSUAN HUANG^{†2}
and KYOJI KAWAGOE^{†2}

When Twitter users can obtain event information using Hashtags on Twitter, Hashtag based event retrieval can not meet the user expectation due to Hashtag property where they are no mutual relation between Hashtags. In order to solve this problem, we propose contextHashtag in its a structured Hashtags space with the multidimensional space. Each of event Hashtags can be automatically located as a region in the contextHashtag multidimensional space. Similarity between Hashtags can be easily calculated in the contextHashtag multidimensional space. As a result, Twitter users can get new events based on the contextHashtag. We also explain an event recommendation prototype of system using the contextHashtag.

^{†1} 立命館大学大学院 理工学研究科

Graduate School of Science and Engineering, Ritsumeikan University

^{†2} 立命館大学 情報理工学部

College of Information Science and Engineering, Ritsumeikan University

1. はじめに

現在、Twitter^{*1}を代表とするマイクロブログサービスが爆発的に普及し、それに関する研究も盛んに行われている¹⁾²⁾³⁾⁴⁾⁷⁾¹⁰⁾。マイクロブログは投稿における文字数制限があるため、利用者は短い文章でマイクロブログサービス Web サイトに気軽に投稿することができる。従来のブログと比べ、投稿の頻度が多くなり、リアルタイムな情報が投稿される特性がある。

Twitter ユーザは、その一機能である Hashtag からコンサートなどのイベント情報を得ることが多い。Hashtag は Twitter ユーザがメッセージを投稿する際、「#文字列 (#summersonic, #地震など)」をメッセージ中に付加したものである。Hashtag により、同じ Hashtag が付加されているメッセージをグループ化できるため、同じ話題のメッセージを他のユーザから見つけやすくなる。Hashtag を検索できる Web サイトも多数存在する。例えば、hashtagsjp^{*2}では 26000 件 (2011 年 8 月 23 日現在) を超える Hashtag と Hashtag に対応するイベント情報を含めた様々な情報が登録されている。このような Web サイトによって、Hashtag に対応するイベントを検索し、イベント情報を取得することが可能である。

Hashtag に関する研究も活発に行われている⁵⁾⁶⁾⁹⁾。Hashtag は Twitter ユーザなら誰でも生成可能であり、生成した Hashtag は Hashtag 検索サイトにイベント情報とともに登録可能である。そのため、Hashtag 検索サイトは「ぶらぶら^{*3}」などの既存のイベント情報サイトに記載されていないイベントも登録されている。また、Hashtag は Twitter のようなマイクロブログの特性であるリアルタイム性や投稿の頻度の多さの影響を受けている。しかし、既存の Hashtag は独立に宣言されるため Hashtag 間に相互関連がなく、同様のイベントでありながら異なる Hashtag が宣言されることが多い。また、Hashtag 検索サイトには膨大なイベントが管理されている。このため、Twitter ユーザが参加したいイベントを検索することは困難である。従って、Hashtag を用いたイベント検索では Twitter ユーザが望むイベントを得ることができない。

そこで本研究では、Hashtag を構造化した多次元空間を持つ contextHashtag を提案する。本稿では contextHashtag の説明とそれを用いたイベント推薦システムの試作について述べる。contextHashtag では、その多次元空間内で既存 Hashtag を領域として自動的に位置付ける。その結果、過去イベントと類似したユーザが望む未来イベントを領域間類似性により求

*1 <http://twitter.com/>

*2 <http://hashtagsjp.appspot.com/>

*3 <http://www.pla2.net/>

めることが可能となる。

2. Hashtag の問題点と contextHashtag

2.1 Hashtag の問題点

Twitter ユーザは Hashtag からユーザが参加したいコンサートなどのイベントを発見することがある。この目的のために、ユーザは直接既存の Hashtag 検索サイトを用いて Hashtag に対応するイベント情報を取得する。例えば、ある Twitter ユーザのメッセージに「#mbs10」が付加されていたとする。それを閲覧した Twitter ユーザは、Hashtag 検索サイトにアクセスし、「#mbs10」を検索することにより、「#mbs10」に対応するイベント情報を取得することが可能である。ここで、「#mbs10」は 2010 年 8 月 21 日、22 日に国営讃岐まんのう公園で開催された MONSTER baSH 2010 と呼ばれる音楽フェスティバルである。しかし、既存の Hashtag に同様の Hashtag が用いられたとして独立に宣言することができる。すなわち、イベント情報の類似した Hashtag が存在する可能性がある。例えば、2011 年 8 月 20 日、21 日に開催された MONSTER baSH 2011 に対応する Hashtag として、「#mbs11」が存在する。ユーザが MONSTER baSH 2011 に関連した Hashtag を検索したいとき、この Hashtag を用いて MONSTER baSH 2011 に関連する Hashtag は検索できる。しかし、イベントの開催場所や出演者などのイベント情報が類似した Hashtag(#mbs10, #summersonic など) は検索できない。これは、これらの Hashtag に関連するイベント情報が整理されていないと考えることができる。このように、既存の Hashtag を扱う環境では、同一のイベントに対応する複数の Hashtag が生成されるが、それらに関連性がなく、イベント情報の類似したイベントを検索しづらい原因となっている。なお、hashtagsjp で管理されている情報を用いて「#mbs11」と類似したイベントを抽出することができるが、上記のイベント情報が整理されていないという問題があり現実的には類似したイベントを得ることが困難である。

2.2 contextHashtag

contextHashtag とは Hashtag を基本に生成された多次元構造を持つ仮想的 Hashtag である。既存の Hashtag は 2.1 節で説明したように Hashtag 構造を持たないため、Hashtag 間同士に相互関連を見出すことは困難である。そこで、contextHashtag はイベントに対応する Hashtag をイベントのジャンルを基にグループ化し多次元構造を生成する。イベントのジャンルは Wikipedia のイベント欄^{*1}を用いてシステム管理者が生成する。イベントに対応する個々の

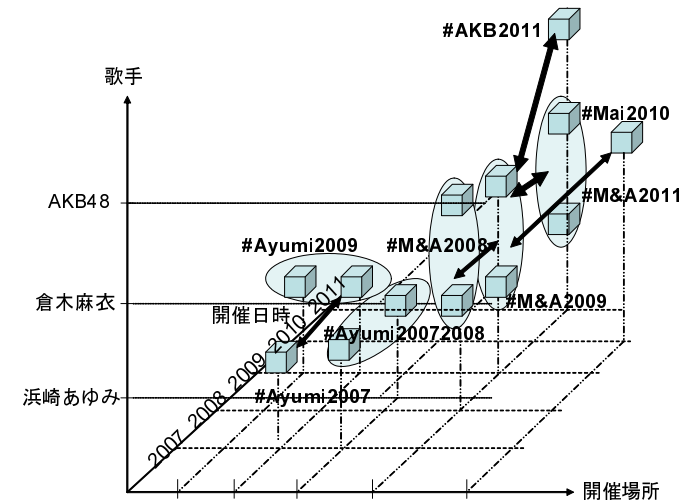


図 1 コンサート contextHashtag の概念
Fig. 1 Concept of concert ContextHashtag

Hashtag をグループ化された contextHashtag の多次元構造化された空間内に位置付ける。

contextHashtag をコンサート contextHashtag と呼ぶ contextHashtag を用いて説明する。コンサート contextHashtag の概念のイメージを図 1 に示す。図 1 のコンサート contextHashtag は、説明を簡単にするため、三次元空間内で定義している。個々の Hashtag は領域 (正確には点の集合) で表現している。図 1 のコンサート contextHashtag は開催日時 (年) と歌手、開

表 1 コンサート contextHashtag の具体的要素

Table 1 Concrete element of concert ContextHashtag

イベント Hashtag	空間を構成する要素	イベント Hashtag 基本データ
#M&A2011	倉木麻衣, 浜崎あゆみ, 東京, 2011	{{(倉木麻衣, 東京, 2011), (浜崎あゆみ, 東京, 2011)}}
#M&A2009	倉木麻衣, 浜崎あゆみ, 東京, 2009	{{(倉木麻衣, 東京, 2009), (浜崎あゆみ, 東京, 2009)}}
#M&A2008	倉木麻衣, 浜崎あゆみ, 東京, 2008	{{(倉木麻衣, 東京, 2008), (浜崎あゆみ, 東京, 2008)}}
#Ayumi2009	浜崎あゆみ, 広島, 大阪, 2009	{{(浜崎あゆみ, 広島, 2009), (浜崎あゆみ, 大阪, 2009)}}
#Ayumi2007	浜崎あゆみ, 大阪, 2007	{{(浜崎あゆみ, 大阪, 2007)}}
#Mai2010	倉木麻衣, 札幌, 2010	{{(倉木麻衣, 札幌, 2010)}}
#Ayumi20072008	浜崎あゆみ, 名古屋, 2007, 2008	{{(浜崎あゆみ, 名古屋, 2007), (浜崎あゆみ, 名古屋, 2008)}}
#AKB2011	AKB48, 東京, 2011	{{(AKB48, 東京, 2011)}}

*1 <http://ja.wikipedia.org/wiki/イベント>

催場所(地名)を軸にした空間である。また、図1内に記述されているイベントに対応するHashtagの具体的要素を表1に示す。

図1に示すように、表1のイベントHashtag「#M&A2009」は(歌手,開催場所,開催日時)={(倉木麻衣,東京,2009),(浜崎あゆみ,東京,2009)}のように表現できる。これは、2009年東京で浜崎あゆみと倉木麻衣が一緒に出演したコンサートを意味する。また、イベントHashtag「#M&A2008」は{(浜崎あゆみ,東京,2008),(倉木麻衣,東京,2008)}で、「#Mai2010」は{(倉木麻衣,札幌,2010)}として表現できる。このとき、「#M&A2009」は「#M&A2008」の方が「#Mai2010」よりも類似性が高いことがユーザの立場からは明らかである。さらに、図1に示す空間表現においても図1の「#M&A2008」や「#Mai2010」のように各々のイベントHashtagを領域間の類似性を得ることが可能である。

また、利用者が過去に前述の「#M&A2008」というHashtagを使用していたとする。その後、コンサートcontextHashtagの空間内に新たに「#AKB2011」と「#M&A2011」が同時に登録されたとする。「#AKB2011」に対応するイベントは2011年東京で開催されたAKB48のコンサートとし、「#AKB2011」は(AKB48,東京,2011)で表現する。「#M&A2011」に対応するイベントは2011年東京で開催された浜崎あゆみと倉木麻衣が一緒に出演したコンサートとし、「#M&A2011」は(浜崎あゆみ,倉木麻衣,東京,2011)で表現する。「#M&A2008」と「#AKB2011」、「#M&A2008」と「#M&A2011」の類似度を比較する際、利用者が過去に使用した「#M&A2008」にはAKB48を要素として含んでおらず、開催場所の東京は同一だがイベント内容のAKB48が異なるため類似度が低く「#AKB2011」は推薦されない。一方で、「#M&A2011」は開催場所の東京が同一であり、イベント内容の浜崎あゆみ、倉木麻衣も同一である。そのため、「#M&A2011」は類似度が高く利用者に推薦可能である。

このように、contextHashtagは既存のHashtagには存在しない多次元構造を持つため、空間内にHashtagを位置付けることが可能になる。contextHashtagの空間内であれば各々のHashtag間の類似度を算出することが可能になる。その結果、利用者が過去に使用したHashtagに対応するイベント情報と新たに生成されたHashtagに対応するイベント情報の類似度を算出し、類似度が高ければ利用者に推薦することが可能になる。

3. contextHashtagによるイベント推薦手法

本研究では、Twitterユーザが参加したいイベントを推薦するシステムと方式を提案することを目的とする。その際、contextHashtagによりイベントHashtagを管理し、contextHashtagの空間内にイベントHashtagを領域として位置付ける方式をポイントとする。その結果、

イベントの内容だけでなく開催場所、開催日時を含めたイベント情報の類似性を算出できる。

本研究におけるイベントとは、開催場所と開催時間が決定し、利用者が参加可能である出来事とし、それに対応するHashtagをイベントHashtagと定義する。contextHashtagはイベントHashtagを基本に生成する。ここで、イベントを特定するためには、開催場所、開催日時、イベント内容の3つの要素が重要であると考えられる。

(1) 開催日時が重要な理由は、開催日時の類似度を取り入れることにより利用者の嗜好の変化を表現することができると考えたためである。例えば、開催場所とイベント内容が同じだが、1年前に参加したイベントと20年前に参加したイベントでは、1年前に参加したイベントの方が利用者の嗜好が強いと考えられる。

(2) 開催場所とイベント内容が重要な理由は、一般的にこの2つを基準として利用者がイベントに参加するか判断すると考えたためである。イベント内容とは、イベントに関連する名詞の集合で表現し、イベントのタイトルやイベントの説明、Twitterに投稿されるイベントHashtag付きのメッセージから抽出する。例えば、コンサートでは出演者をイベント内容とする。

イベント内容と開催場所と開催日時の3つの要素をイベントHashtag基本データと呼ぶ。本稿では、イベントHashtagをhashtagsjpやハッシュタグクラウド^{*1}と呼ばれるHashtag検索サイトのイベントカテゴリに属しているイベントHashtagを対象とする。これら既存のHashtag検索サイトは、Hashtagとイベント情報が管理され、イベントに関連する情報を容易に取得できる。

イベントHashtagは、過去イベントHashtagと未来イベントHashtagの2種類に分けることができる。過去イベントHashtagとは、利用者がシステムログイン時に開催日時がログイン時より以前でかつ、contextHashtagの空間内に位置付けられているイベントHashtagである。また、未来イベントHashtagとは、利用者がシステムログイン時に開催日時がログイン時より以後でかつ、contextHashtagの空間内に位置付けられているイベントHashtagである。

3.1 基本的考え方

本研究では、イベントHashtagに関連するイベントHashtag基本データをもとに、イベントのジャンルが同じイベントHashtagから構成されるcontextHashtagを提案する。既存のHashtagは各々が独立に宣言されるため、Hashtag間に関連性が存在しないという問題がある。そこで、既存のイベントHashtagをイベントのジャンルをもとにグループ化する。各々

*1 <http://hashtagcloud.net/>

のグループにおいてイベント Hashtag 基本データをもとに多次元構造を生成し、過去イベント Hashtag 全てをこの空間内の領域として位置付けする。ここで、領域(イベント Hashtag)とは点(イベント Hashtag 基本データ)の集合であると定義する。その結果、未来イベント Hashtag と過去イベント Hashtag 間の領域間類似度を算出することが可能となる。なお、単純化のために本稿では既存のイベント Hashtag のグループ化は人手によって行うものとする。

3.2 contextHashtag

contextHashtag を情報検索分野に頻繁に用いられるベクトル空間モデルを用いて構造的に表現する。ベクトル空間モデルを用いることにより、管理している contextHashtag の空間内に過去イベント Hashtag と未来イベント Hashtag を統一的に扱うことができる。なお、各々の contextHashtag の空間内において構成される要素が異なるので、はじめの分類ならびに構造は予め設計する。contextHashtag を以下の構造で表現する。

Hashtag は任意の文字列である。Hashtag を h_i とすると N 個の Hashtag の全体集合 H を $H = \{h_1, h_2, \dots, h_N\}$ と表現する。同様に、contextHashtag を g^k とすると M 個の contextHashtag の全体集合 G は $G = \{g^1, g^2, \dots, g^M\}$ と表現できる。contextHashtag g^k は、構成する n_k 個の要素 $E^k = \{E_1^k, \dots, E_{n_k}^k\}$ で作られる空間である。ここで、 E_j^k は文字列の集合である。Hashtag の全体集合 H は関連する contextHashtag によって $H = H^1 \cup H^2 \dots H^N$ に分解できる。その結果、 g^k と H^k の関係は次のように表現できる。 $\forall h_j^k \in H^k$ について、

$$h_j^k = \{d_{ji}^k = (e_1^k, \dots, e_{n_k}^k) \mid e_i^k \in \tilde{E}_i^k \subset E_i^k (i = 1, \dots, n_k)\} \quad (1)$$

$$h_j^k \subset g^k \quad (2)$$

ここで、 d_{ji}^k をイベント Hashtag 基本データと呼ぶ。 h_j^k は contextHashtag g^k に管理されているひとつの Hashtag である。 \tilde{E}_i^k はある条件で絞り込まれた E_i^k の部分集合で文字列の集合である。また、contextHashtag g^k は識別子 Ig^k によって名前付けされるとする。 Ig^k は任意の文字列である。イベント Hashtag h_j^k の要素 e_i^k は、管理されている contextHashtag g^k 内の要素 E_i^k に含まれる。

3.3 イベント Hashtag を構成する点の特徴ベクトルの生成

contextHashtag に管理されているイベント Hashtag 特徴ベクトルを生成する。(1)式において、要素 E^k を開催場所、開催日時、イベント内容の3つに分解する。すなわち、contextHashtag の特徴ベクトルは位置特徴ベクトル、時間スカラー、イベント内容特徴ベクトルから構成されるものとする。したがって、contextHashtag g^k に管理されているひとつのイベント Hashtag の特徴ベクトル h_j^k は、そのイベントの開催された位置特徴ベクトル p_j と時間特徴スカラー e_j とイベント内容特徴ベクトル f_j を用いて次のように表現する。

$$h_j^k = \left(p_j^t, e_j, f_j^t \right)^t \quad (3)$$

t は、ベクトルの転置を示す。

3.3.1 位置特徴ベクトル

開催場所から Geocoding API^{*1}を用いて緯度、経度を取得する。ここで、Geocoding API とは住所やランドマーク名から緯度、経度を検索して XML 形式で返すサービスである。例えば、パラメータとして『東京タワー』で問い合わせると、座標値『35.661913, 139.700943』が返ってくる。これを用いて、contextHashtag g^k に管理されているひとつのイベント Hashtag の位置特徴ベクトル p_j は、その開催場所の緯度 p_{jx} 、経度 p_{jy} を用いて次のように表現する。

$$p_j = \left(p_{jx}, p_{jy} \right)^t \quad (4)$$

3.3.2 イベント内容特徴ベクトル

イベントのタイトルや内容、Twitter に投稿されるメッセージからイベントに関連するキーワードを抽出しイベント内容特徴ベクトルを生成する。したがって、contextHashtag g^k に管理されているイベント Hashtag のイベント内容特徴ベクトル f_j は、そのイベント内容 $f_{ji} (i = 1, \dots, n_k)$ を用いて次のように表現する。

$$f_j = \left(f_{j1}, \dots, f_{jn_k} \right)^t \quad (5)$$

f_{ji} はイベント Hashtag において、イベント内容 i が存在するかどうかを示す値である。

$$f_{ji} = \begin{cases} 1 & (\text{イベント Hashtag にイベント内容 } i \text{ が存在するとき}) \\ 0 & (\text{存在しないとき}) \end{cases} \quad (6)$$

ここで、コンサート contextHashtag, \$concert に管理されているイベント Hashtag, #UH_WL を用いて説明する。\$concert は、イベント内容を歌手名とし、予め $f_j = \left(\text{浜崎あゆみ, AKB48, 宇多田ヒカル} \right)^t$ という構造を持つとする。#UH_WL に対応するイベントは、宇多田ヒカルが2010年12月8日に横浜アリーナで行ったコンサートである。これをイベント Hashtag 特徴ベクトルで表現すると(35.512228, 139.620165, 2010/12/8, 0, 0, 1)^t になる。

3.4 イベント Hashtag 間の類似度

利用者が使用した過去イベント Hashtag に類似した未来イベント Hashtag を抽出する。こ

*1 <http://www.geocoding.jp/>

ここで、利用者が使用した過去イベント Hashtag に類似した未来イベント Hashtag に対応するイベントが利用者が望むイベントであるとする。

過去イベント Hashtag と未来イベント Hashtag の領域間類似度を次のように算出する。ここで、領域間類似度とは各々の領域 (イベント Hashtag) に含まれる点 (イベント Hashtag 基本データ) 同士が最も類似度の高い値となる点 (イベント Hashtag 基本データ) の組み合わせと定義する。このような類似度を用いた理由は、計算量と精度の観点から上記の類似度を採用した。

同一 contextHashtag 内に管理されている過去イベント Hashtag を h_i と未来イベント Hashtag を w_j とすると、2 つの h_i, w_j 間の類似度 $Sim(h_i, w_j)$ を (7) 式で定義する。

$$Sim(h_i, w_j) = \max_{\{u,v\}} \{Sim_p(p_{iu}, p_{jv}) \times Sim_f(f_{iu}, f_{jv}) \times Sim_e(e_{iu}, e_{jv})\} \quad (7)$$

ただし、 $0 \leq Sim(h_i, w_j) \leq 1$ である。また、過去イベント Hashtag h_i は $\{(p_{iu}, e_{iu}, f_{iu})\}$ ($u = 1, \dots, u_i$) である。 $\{(p_{iu}, e_{iu}, f_{iu})\}$ は Hashtag h_i の基本データであり、 p_{iu} は位置特徴ベクトル、 e_{iu} は時間スカラー、 f_{iu} はイベント内容特徴ベクトルから構成される。同様に、未来イベント Hashtag w_j は $\{(p_{jv}, e_{jv}, f_{jv})\}$ ($v = 1, \dots, v_i$) であり、 (p_{jv}, e_{jv}, f_{jv}) は Hashtag w_j の基本データである。

ここで、 $Sim_p(p_{iu}, p_{jv})$ を (8) 式で $Sim_f(f_{iu}, f_{jv})$ を (9) 式で定義する。 $Sim_p(p_{iu}, p_{jv})$ は位置特徴ベクトル間の距離に 1 を足し、その逆数により求める。 $Sim_f(f_{iu}, f_{jv})$ は、イベント内容特徴ベクトル間のなす角の余弦値により求める。

$$Sim_p(p_{iu}, p_{jv}) = \frac{1}{1 + d(p_{iu}, p_{jv})} \quad (8)$$

$$Sim_f(f_{iu}, f_{jv}) = \frac{f_{iu} \cdot f_{jv}}{\|f_{iu}\| \|f_{jv}\|} \quad (9)$$

ただし、 $d(p_{iu}, p_{jv})$ はヒュベニの公式*1(基準値を 100km とする)である。さらに、 $Sim_e(e_{iu}, e_{jv})$ を (10) 式で定義する。 $F_x(e_{iu}, e_{jv})$ は、 x (年, 月, 日などの単位を示し contextHashtag ごとに予め設定される) に基づく e_{iu}, e_{jv} との間の時間を算出する関数である。 $Sim_e(e_{iu}, e_{jv})$ は $F_x(e_{iu}, e_{jv})$ に 1 を足し、その逆数により求める。

$$Sim_e(e_{iu}, e_{jv}) = \left(\frac{1}{F_x(e_{iu}, e_{jv}) + 1} \right)^{\frac{1}{\alpha}} \quad (10)$$

α は任意の自然数であり、極端に時間差がある場合にのみ時間スカラーの類似度の影響を考

#clrh60

CLR/第60回勉強会 <http://1.co/Un19fjx> [d2011/07/16] [札幌市中央区北5条西2丁目5 JRタワーオフィスプラザさっぽろ20F マイクロソフト北海道支店]

タグレベル:9 | つぶやき数:515件 | ユーザー数:50人 | リンク数:149件
登録日:2011/06/24 12:58 | 解析日:2011/08/05 09:57

イベント日付:
2011/07/16 ~ 2011/07/16

イベント場所:
札幌市中央区北5条西2丁目5 JRタワーオフィスプラザさっぽろ20F マイクロソフト北海道支店

図 2 ハッシュタグ検索サイトに記載されているイベント情報 (hashtagsjp)
Fig. 2 Event information described in Hashtag retrieval site (hashtagsjp)

慮するために用いる定数である。以降で述べるシステムでは $\alpha = 4$ を使用する。

ここで、類似度算出の例を表 1 の「#M&A2009」、「#M&A2011」、「#AKB2011」を用いて説明する。予め $f_j = (\text{浜崎あゆみ, 倉木麻衣, AKB48})^t$ という構造を持つとする。「#M&A2009」をイベント Hashtag 特徴ベクトルで表現すると、 $(35.689506, 139.691701, 2008, 1, 1, 0)^t$ となる。同様に、「#M&A2011」をイベント Hashtag 特徴ベクトルで表現すると、 $(35.689506, 139.691701, 2011, 1, 1, 0)^t$ となり、「#AKB2011」をイベント Hashtag 特徴ベクトルで表現すると、 $(35.689506, 139.691701, 2011, 0, 0, 1)^t$ となる。 $Sim(X)$ は「#M&A2009」と「#AKB2011」の類似度とすると、 $Sim(X) = 0$ になる。また、 $Sim(Y)$ は「#M&A2009」と「#M&A2011」の類似度とすると、 $Sim(Y) = 0.707$ になる。従って、 $Sim(X) < Sim(Y)$ であるので、この方式でも類似度の高い「#M&A2011」を推薦することが可能である。

4. contextHashtag を用いた Twitter ユーザ向けイベント推薦システム

4.1 contextHashtag への自動位置付け手法

著者は 3 章の手法を用いてシステムを作成し、その報告を行った⁸⁾。予備評価実験の結果、contextHashtag によるイベント推薦が有効であることを確認した。しかし、前回の報告の際はイベント Hashtag を contextHashtag の空間内に位置付ける作業は人手で行った。そこで、本章はイベント Hashtag を自動的に contextHashtag の空間内に位置付ける手法を提案する。

システム管理者が contextHashtag を予め生成しておくとする。図 2 に Hashtag 検索サイトに記載されているイベント情報を示す。3 章で述べた通り、contextHashtag は位置特徴ベクトル、時間スカラー、イベント内容特徴ベクトルから構成される。あるイベント Hashtag が新規に登録されたとき、contextHashtag に位置付けるために位置特徴ベクトル、時間スカラーはハッシュタグ検索サイトから生成できる。しかし、イベント内容特徴ベクトルは Hashtag

*1 <http://yamadarake.web.fc2.com/trdi/2009/report000001.html>

検索サイトからだけで生成することは情報の観点から十分でない．そこで，Twitter に投稿されるイベント Hashtag 付きメッセージを用いてイベント内容特徴ベクトルを生成する．

以下に Hashtag 検索サイトに新しいイベント Hashtag が登録された際，そのイベント Hashtag が該当する contextHashtag に位置付けされるまでの流れを示す．

【STEP1】 システムは Hashtag 検索サイトからイベント Hashtag に関する基本データと Twitter からイベント Hashtag 付きメッセージを取得する．

【STEP2】 システムは STEP1 で取得したイベントのタイトル，イベント情報，メッセージからキーワードを抽出する．

【STEP3】 STEP2 で抽出されたキーワードの TF 値を算出し，TF 値の低いキーワードをフィルタリングする．

【STEP4】 キーワードを Wikipedia 辞書とストップリストを用いてフィルタリングする．

【STEP5】 STEP4 でフィルタリングされたキーワードが contextHashtag に存在すればイベント Hashtag として該当する contextHashtag の空間に位置付ける．

4.1.1 キーワード候補抽出

キーワードの抽出には，Yahoo!デベロッパーネットワーク^{*1}で提供されているキーフレーズ抽出の API を利用する．上記のキーフレーズ抽出を利用して，名詞だけを抽出する．ここで，一般的に用いられる形態素解析を用いない理由は，キーワードが細かく区切れ精度が悪くなると考えたためである．

4.1.2 スコア付け

4.1.1 節で抽出された空間に位置付けするために必要なキーワードを出現頻度を用いてスコア付けを行う．キーワード t がメッセージの全体集合 D に高い頻度で現れるなら， t は D を特徴付けると考えられる．以下の式で求める．

$$tf_j^i = freq(i, j) \quad (11)$$

ただし， $freq(i, j) =$ 文書 D_j におけるキーワード t_i の出現頻度である．出現頻度の代わりに，キーワードの重要度として頻繁に用いられる tfidf 値が考えられる．しかし，例えば同一の出演者が多く出演するコンサートが存在するとする．そうすると，tfidf 値を用いた場合には，イベントを特徴付けるスコアが小さくなる．そこで本提案では，単純な出現頻度を用いることにした．TF 値が低いキーワードはそのイベントを特定する要素ではないと考えられる．

*1 <http://developer.yahoo.co.jp/>

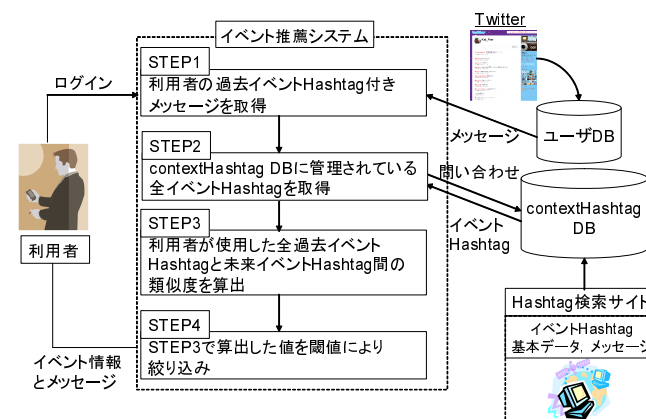


図3 イベント推薦システムの処理手順

Fig. 3 Processing procedure of event recommendation system

4.1.3 フィルタリング

4.1.1 節で抽出されたキーワードは，イベントを特定するためのキーワード以外のキーワードが数多く存在する．そこで，Wikipedia 辞書とストップリストでフィルタリングする．Wikipedia 辞書を用いることで，一般的なキーワードだけを抽出できる．また，ストップリストを作成することで，より精度を上げることができる．ストップワードは以下の事項を満たすキーワードとする．

- 固有名詞以外のキーワード
- 全イベント Hashtag に頻繁に使用されるキーワード

4.2 イベント推薦システムとその流れ

本稿では，利用者の参加したイベントに類似したイベントを推薦するシステムを提案する．本システムの処理手順を図3に示す．また，本システムの出力例を図4に示す．本研究における利用者とは，Twitter ユーザでありイベント Hashtag 付きのメッセージを頻繁に Twitter に投稿するユーザとする．ここで，予め contextHashtag を生成し，イベント Hashtag 全てを対応する contextHashtag の空間内に位置付けされている状態とする．contextHashtag データベースには位置付けされたイベント Hashtag を格納する．ユーザデータベースには利用者が Twitter に投稿したメッセージを格納する．

【STEP1】 利用者はシステムにログインすると，システムはユーザデータベースから利用

者が使用した過去イベント Hashtag 付きメッセージを取得する。

【STEP2】 システムは contextHashtag データベースから管理されている全イベント Hashtag を取得する。

【STEP3】 システムは、STEP1 で取得した利用者が使用した contextHashtag の空間内で管理されている過去イベント Hashtag と未来イベント Hashtag との類似度を算出する。

【STEP4】 システムは STEP3 で算出した値を閾値により絞り込み、閾値以上であれば利用者に未来イベント Hashtag に対応するイベント情報とメッセージを出力する。

4.3 contextHashtag を用いたイベント推薦システム

4.3.1 前提

本システムは、contextHashtag により利用者が望む未来イベントを検出し、ユーザに対してイベント推薦を行うシステムである。

表 2 に示すように利用者は過去にコンサート contextHashtag に管理されているイベント Hashtag を使用したとする。それらのイベント Hashtag と他の N 個のイベント Hashtag を対象に類似度を算出し、類似度の高い上位 K 件 ($K = 3$) のイベントを推薦する。なお、予めコンサートに関連するイベント Hashtag を基に contextHashtag を生成しておく。イベント内容特徴ベクトルの次元を歌手の総数とする。時間スカラーは年単位である。

4.3.2 イベント推薦システムの具体例

具体例を表 3、表 4、表 5 に示す。表 3 は利用者が過去に使用したイベント Hashtag を #1996MaiAyu とした場合の他のイベント Hashtag との類似度の高い上位 K 件 ($K = 3$) の計算結果を示す表であり、表 4 は利用者が過去に使用したイベント Hashtag を #1992Sazan とした場合の他のイベント Hashtag との類似度の高い上位 K 件 ($K = 3$) の計算結果を示す表である。さらに、表 5 は利用者が過去に使用したイベント Hashtag を #1991Smmap とした場合の他のイベント Hashtag との類似度の高い上位 K 件 ($K = 3$) の計算結果を示す表である。

表 3 では、コンサートの開催場所が東京で歌手に浜崎あゆみや倉木麻衣に関連するイベントを出力する。表 4 では、コンサートの開催場所が京都付近で歌手がサザンオールスターズに関連するイベントを出力する。表 5 では、コンサートの開催場所が札幌で歌手が SMAP に関連するイベントを出力する。このように、利用者が過去に使用したイベント Hashtag 基本データの類似性の高いイベントを出力することが可能である。また、表 3、表 4、表 5 から、過去に使用した Hashtag のイベント内容に関連のないイベントは出力されていない。また、表 4 からイベント内容が類似していれば京都に比較的距離の近い大阪のイベントも出力されたが、札幌や福岡などさらに離れているイベントは出力されていない。実行結果から

開催場所が著しく離れているイベントやイベント内容が異なるイベントなど、利用者に適していないイベントは出力されていない。

従って、これらの表 3、表 4、表 5 に示すように、contextHashtag を用いることで利用者に適したイベントを提供することが可能である。

5. 実験

5.1 実験条件

本稿で提案した contextHashtag とそれを用いたイベント推薦システムの有効性を確認するため、実データを用いて実験を行った。本システムは Java 言語を用いて実装した。また、Twitter API の Java ラップである Twitter4J を利用してメッセージを取得した。まず、イベント内容を出演者としたコンサート contextHashtag を作成した。既存の Hashtag 検索サイトから取得したイベント情報を基に 2010 年 5 月から 2010 年 8 月中頃の間に開催されたイベント

表 2 過去に使用したイベント Hashtag

Table 2 Event hashtag used in the past

Hashtag	開催場所	歌手	開催日 (年)
#1996MaiAyu	東京	浜崎あゆみ, 倉木麻衣	1996
#1992Sazan	京都	サザンオールスターズ	1992
#1991Smmap	札幌	SMAP	1991

表 3 具体例の実行結果 1

Table 3 Execution result of concrete example1

開催場所	歌手	開催日時 (年)	類似度
東京	浜崎あゆみ, 倉木麻衣, EXILE	2000	0.546
東京	浜崎あゆみ	2006	0.388
東京	倉木麻衣	2010	0.359

表 4 具体例の実行結果 2

Table 4 Execution result of concrete example2

開催場所	歌手	開催日時 (年)	類似度
京都	サザンオールスターズ	1998	0.615
大阪	サザンオールスターズ	1997	0.468
大阪	サザンオールスターズ, 倉木麻衣	1993	0.286

表 5 具体例の実行結果 3

Table 5 Execution result of concrete example3

開催場所	歌手	開催日時 (年)	類似度
札幌	SMAP	2001	0.549
札幌	SMAP, 倉木麻衣, 浜崎あゆみ	1993	0.439
札幌	SMAP, 倉木麻衣	1999	0.408

推薦されるイベント一覧

イベント Hashtag	開催場所	開催日時	イベント内容(出演者)	URL
#mb11	国営清見丸の公園(茨城)	2011/08/20 2011/08/21	# Road of circle,back number,zembon,Gocco,Dragon Ash,豊 基博,MY,毛皮のブーツ&mm Parabolian,BasketONE,OK ROCK,サッポロ音楽系-SHAKALABETS,スキズ&スイッチ,THE BAWDIES,the HATUSU,the telephones,ACIDMAN&KIMAMA,[Chamaqura]FACT,KREVA,九州男,たごご,Nothing's Carved in Stone,MAN WITH A MISSION,レズンダム,ザ・ホルモン,MONDRIGHT,難波歌麿,Perfume,佐野元春,THE GOVYTE BAND,西島秀孝,カワタニマサアツ	http://tco/098Ssf
#mb11に関するメッセージ				
投稿者	メッセージ	送信日時		
usj0215	誰かに送ったのね、#mb11 RT @AK012 毛皮のブーツ/MAN WITH A MISSIONのグッズのね、を壊さない我が家のMAN WITH A MISSION。 #mb11 http://tco/4fDnG98	Sat Aug 27 00:23:54 JST 2011		
usj0215	RT @unicormnaly: モノノシ語録「心のT1バックをカッしちゃえ」「今までの君は間違」心や「泣」「佐野バイゼン」「俺等」にこのカステープもったい、泣、と最後は涙は名を言えやな。 #mb11	Sat Aug 27 00:21:07 JST 2011		
Jairotochi	モノノシ回顧。正式名称は#mb11と、mb11もっすごぶっかたの船、もっすご、または、もの、もっす、(ものすごの)意味ですな！ #mb11 http://tco/hFvdlMC	Sat Aug 27 01:44:08 JST 2011		
Jairotochi	モノノシ回顧。これがモノノシとんだん…！ 二回観ました。 #mb11 http://tco/4Yw0GnG	Sat Aug 27 01:28:16 JST 2011		
isawakojo	みくし、日記書きました(´▽`)興味のある方は「tsoco」で検索どうぞ！ #mb11	Thu Aug 25 13:44:20 JST 2011		
kurutama_jivego	※、NHKBSで佐野バイゼンのライブしてるけど、かっくも、な、あ、 #mb11	Wed Aug 24 22:58:09 JST 2011		
silachuschu	RT @unicormnaly: モノノシ語録「心のT1バックをカッしちゃえ」「今までの君は間違」心や「泣」「佐野バイゼン」「俺等」にこのカステープもったい、泣、と最後は涙は名を言えやな。 #mb11	Wed Aug 24 22:10:00 JST 2011		
masaquarius	#mb11 #mb11 行ったよー！ 誰かに京都大作戦に参加したぞ。しかも2日目のモノノシとタオラ軍には笑いましたw	Wed Aug 24 21:51:50 JST 2011		
rock.tiger	RT @hambart: ホルモンは何が楽しいって、モックしてもみんながもうニコニコ。運気、体ぶつりあって、も相性があつと、初対面、誰べつとめっちゃいい笑顔、いやー、笑ってよかった！ #mb11	Wed Aug 24 21:01:36 JST 2011		
rock.tiger	RT @7564ch: ほぼクッキーロック好きです！ #サンホロロックロール！ ロックロール！ 難波の観客！ ロックロール！ パンクロック！ 電話でデスコ！ デスコ！ デスコ！ デスコ！ デスコ！ 根付ウカウ！ ホルモン！ 難波！ コッパリ！ ヤッダー！ 根付さんな観です！ #mb11	Wed Aug 24 20:52:01 JST 2011		

過去に使用した類似イベント

イベント Hashtag	開催場所	開催日時	イベント内容(出演者)	URL
#mb2010	国営清見丸の公園(茨城)	2010/08/21 2010/08/22	阿部真央,andy,mort,avengers in sc,FB&EAT CRUSADERS,BJOMMAMA,チャットモンチー-ET-KING,カズノリ,TK,AKAMOTO SONE OK ROCK,ORANGE RANGE,SEAKO,スライヤーター,THE BACK HORNETS,THE BANWIES,THE PREDATORS,Base Ball Bear,志保,天ガガガSP,Galileo Galilei,GOOD 4 NOTHING,HAN-KUN,クワメンシ,小林大助,黒猫,チェルシー,LOVE PSYCHEDELIC,DOMON,GOOL,800,NICO Touches the Walls,Perfume,世界の絆,かりSuperfly,西島秀孝	http://www.monsterbash.jp/

図 4 出力例
Fig. 4 Example of output

に対応するイベント Hashtag20 個を過去イベント Hashtag とし、コンサート contextHashtag の空間内に位置付けを行った。同様に、2010 年 8 月中旬頃から 2011 年 8 月中旬に開催されたイベントに対応するイベント Hashtag30 個を未来イベント Hashtag とし、コンサート contextHashtag の空間内に位置付けを行った。未来イベント Hashtag と過去イベント Hashtag の類似度を算出し、閾値以上であれば推薦する。なお、本実験では閾値を 0.05 とした。

予備実験の被験者は学生 6 名とした。被験者には予め過去イベント Hashtag のイベント情報を提示し、被験者はその中から「参加したかったイベント」を選択する。同様に、未来イベント Hashtag に対して、被験者は「参加したいイベント」を選択し、これを正解集合と定める。評価尺度には、情報検索の評価に頻繁に使用される再現率、適合率、F 値を用いる。比較手法には、一般的なベクトル空間モデルを用いた手法を用いた。ベクトル空間モデルを

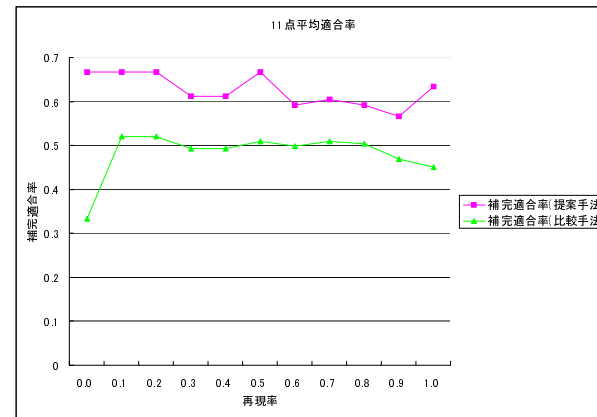


図 5 11 点平均適合率による評価結果

Fig. 5 Evaluation result of average precision of 11 points

用いた手法は、contextHashtag を使用せず、イベントに出演している出演者のみのイベント情報からイベント Hashtag の特徴ベクトルを生成する手法である。上記の比較評価に加えて提案手法と同様に、過去イベント Hashtag と未来イベント Hashtag 間の特徴ベクトル間の類似度を算出し、閾値以上であれば推薦する。

被験者に提示されるイベント情報に関して被験者による主観の評価を行った。以下の 3 つの各評価項目について、被験者が、1(悪い)~4(良い)点で評価した。

- (1) 自分の嗜好に適したイベントが推薦されたか
- (2) 本システムを利用したいか
- (3) 見やすいか

なお、表 7 では、0(悪い)~1(良い)で示す。

5.2 実験結果ならびに考察

実験結果は理想的に contextHashtag の空間内にイベント Hashtag が位置付けされた際のシステムの有用性について述べる。情報検索の評価方法による実験結果を表 6、主観的評価に

表 6 情報検索における評価結果

Table 6 Evaluation result in information retrieval

	比較手法	提案手法
11 点平均適合率	0.481	0.625

よる実験結果を表 7, 11 点平均適合率の評価結果を図 5 に示す.

表 6 と図 5 より, 提案手法は比較手法と比べて 11 点平均適合率は高いことがわかる. 11 点平均適合率が高い理由として, 比較手法より推薦の対象となるイベントが絞り込まれたと考えられる. その結果, システムはより利用者の望むイベントだけを出力することが可能になった.

表 7 に示す主観的評価により, 本システムを使用したいという評価項目が高く本システムの高い有用性を示すものとする. しかし, 推薦されたイベントが利用者の嗜好に適合しているかという評価ではあまり高くないことがわかる. その理由として, 利用者の嗜好にあまり適していなくても, システムとしてれば便利と感じる利用者が多いと考えられる. また, 本システムの出力は見やすいかという評価はあまり良いとは言えない. 本システムは推薦されるイベント情報, イベント Hashtag が付加されているメッセージ, そして過去に使用した類似イベントを利用者に出力している. 開催場所が地名だけだとわかりづらいという意見があったため, 地図や現在値からイベント開催場所までの距離を出力するなどインタフェースを改良していく.

以上より, 本システムは以下のことが言える.

- (1) イベント推薦に関する評価結果から, contextHashtag を用いたイベント推薦は有効である
- (2) 主観的評価より, イベント推薦システムとして有用性がある

6. 関連研究

Twitter に関する研究が盛んに行われている. Miles⁶⁾ は, マイクロブログにおいて Hashtag 検索の言語モデル化手法を提案した. その結果, ユーザに役立つ Hashtag を推薦することを可能にした. 本岡⁹⁾ は, Hashtag とその Hashtag が示すイベントの対応関係に基づいて, 与えられた Hashtag に対する類似イベントを示す Hashtag を発見する手法を提案した. Hashtag を用いて類似イベントを抽出するという点は同じだが, ユーザ集合から類似イベントを抽出する方法が本研究の contextHashtag を用いてイベント情報から類似イベントを抽

表 7 主観的評価結果
Table 7 Subjective evaluation result

	嗜好に合うか	利用したいか	見やすいか
平均値	0.75	0.833	0.5
標準偏差	0.33	0.22	0.67

出する方法とは異なる. 山口¹⁰⁾ らや Wei³⁾ らは, Twitter ユーザに対してタグ付けを行っている. Twitter ユーザから投稿されるメッセージには何らかの特徴がある. その特徴を考慮し Twitter ユーザに対してタグ付けを行うことで, 膨大な Twitter ユーザの中から有益な情報を得ることが可能なユーザを検索しやすくなる.

7. おわりに

本研究では, contextHashtag の提案とそれを用いた Twitter ユーザ向けイベント推薦システムを提案した. contextHashtag により, 過去イベント Hashtag と未来イベント Hashtag との領域間類似度が算出可能となった. また, contextHashtag を自動的に生成し, 未来イベント Hashtag を contextHashtag の多次元空間内に自動的に位置付けを行う方式を提案した. その結果, Twitter の一機能である Hashtag を使用することによりリアルタイムなイベントや規模に関係なくイベントを推薦することが可能になった.

今後は, ユーザごとに適した類似度算出式方法を考える. さらに, システムをインターネット上に公開することで実際に Twitter ユーザに利用してもらい, 継続的な評価, 改善を行いながら精度を上げていく.

謝辞 本論文では hashtagsjp, ハッシュタグクラウドからイベント情報を取得している. ここに記して感謝の意を表す.

参考文献

- 1) Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng, “Why we twitter: understanding microblogging usage and communities”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56-65, 2007.
- 2) Owen Phelan, Kevin McCarthy, Barry Smyth, “Using Twitter to Recommend Real-Time Topical News”, RecSys '09 Proceedings of the third ACM conference on Recommender systems, pp. 385-388, 2009.
- 3) Wei Wu, Bin Zhang, Mari Ostendorf, “Automatic Generation of Personalized Annotation Tags for Twitter Users”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 689-692, 2010.
- 4) Jeff Huang, Katherine M. Thornton, Efthimis N. Efthimiadis, “Conversational Tagging in Twitter”, Proceedings of the 21st ACM conference on Hypertext and hypermedia, pp. 173-178, 2010.
- 5) Simon Carter, Manos Tsagkias, Wouter Weerkamp, “Twitter hashtags: Joint Translation and Clustering”, ACM WebSci' 11, pp. 1-3, 2011.

- 6) Miles Efron, “Hashtag Retrieval in a Microblogging Environment”, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 787-788, 2010.
- 7) Sandra Garcia Esparza ,Michael P. O’Mahony ,Barry Smyth, “Towards tagging and categorization for micro-blogs”, 21st National Conference on Artificial Intelligence and Cognitive Science (AICS) , 2010.
- 8) 海江田隆博, 黄宏軒, 川越恭二, “contextHashtag による Twitter ユーザ向けイベント推薦システム”, DEIM Forum 2011, F5-4, 2011.
- 9) 本岡亮, 湯本高行, 新居学, 高橋豊, 角谷和俊, “Twitter ハッシュタグを用いた類似イベント検索”, DEIM Forum 2011, A1-5, 2011.
- 10) 山口祐人, 天笠俊之, 北川博之, “リストを用いた Twitter ユーザのタグ付け手法”, DEIM Forum 2011 , A1-1, 2011.