Twitterのハッシュタグを用いた話題性を反映した動的カテゴリ生成

佐藤　和人† 　若林　 啓††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s1411530@u.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　近年，機械的な処理によって Web 上のコンテンツを分類するという研究課題は重要さを増している．SNS 上のコンテンツはタグを付与することで特定のトピックに関連するコンテンツの検索が容易になるという効果が期待できるが，タグによって検索できるのは特定のメディア上のコンテンツのみであることが多く，他のメディア上のコンテンツとの関連はない．特に，Web 上の話題は時間的な移り変わりが激しく，あらかじめ固定されたタグセットを用いて分類を行うことは現実的ではない．本研究では，Twitter で利用されているハッシュタグを動的に構成される話題のカテゴリとみなし，対象のテキストをこの動的カテゴリに分類する手法を提案する．具体的には特定のハッシュタグの現れるツイート全てを繋ぎ合わせたハッシュタグ文書を仮定し，ここに現れる語彙について TF-IDF によりベクトル化する．得られたハッシュタグの特徴ベクトルに対して k-means 法でクラスタリングすることで特定のトピックを表すハッシュタグを同一のクラスタにまとめ，これを動的カテゴリとして他のメディアのテキストを分類する．実験では，提案手法を用いて新聞記事に対して分類を行う．この結果から，提案手法により，当該の時期に特徴的なハッシュタグクラスタの割り当てを行うことができることを示す．

キーワード　Twitter, 新聞記事, ハッシュタグ, クラスタリング, タグ付け

1. 序　　論

近年，Twitter(注1)や Facebook(注2)などのソーシャルネットワーキングサービス (SNS) をはじめとする Web 上のサービスへの社会的関心が高まっている．Web では多くの情報をリアルタイムに収集できるため，この利用価値は高い．一方，Web に投稿されるテキストは整理されていない場合が多く，求める情報を選択的に収集することは困難になってきている．このため，機械的な処理によって Web 上のコンテンツを分類するという研究課題は重要さを増している．

SNS にはコンテンツを分類する手法として「タグ」が存在する．タグはユーザ自身によって投稿するコンテンツの属性を表現するために付与される．タグに使われる語は統制語ではなくユーザが決めた任意のキーワードである．タグに関する情報は複数のユーザで共有され，共通のトピックに対しては同一のタグが付与される．こうした分類形態をあらかじめ決められた統制語を元に分類する従来のタクソノミー (taxsonomy) と区別してフォクソノミー (folksonomy) と呼ぶ．

Twitter においては「ハッシュタグ」が導入されている．ハッシュタグは，ハッシュ記号 (#) の後にトピックを表すキーワードを続ける形で表される．これをツイート（Twitter への投稿）に含めることでそのツイートにハッシュタグを付与することができる．例えば，「#みちびき」というハッシュタグは，このツイートがみちびきの打ち上げ成功に言及していること，あるいはツイートの主題がみちびきの打ち上げ成功であることを表している．ハッシュタグはツイート中のどこにでも書くことができ，同一のハッシュタグの付与されたツイートを一覧できる機能がある．Twitter におけるハッシュタグに使うキーワードはあらかじめ決められた統制語ではなくユーザによって定義される．この意味でハッシュタグはフォクソノミーだと言える．ハッシュタグはツイートを投稿したユーザ以外が付与することができないため，完全なフォクソノミーではないとする議論もあるが [1]，本研究で重要な点はハッシュタグがユーザによって定義されたキーワードを用いる点なので，ハッシュタグをフォクソノミーの一種とみなすことにする．

タグは，付与することで特定のトピックに関連する投稿を検索することが容易になるという効果が期待できる．しかし多くの場合，タグによって検索できるのはタグの利用されているメディア上のコンテンツのみであり，そのトピックに関係する他のメディア上のコンテンツとの関連はない．こうした現状を改善するには実際に使われているハッシュタグを他のメディアのテキストに対しても付与できることの検証が必要である．

本研究では，Twitter 以外のメディアのテキストに対し，Twitter で利用されているハッシュタグを割り当てることで対象のテキストをカテゴリに分類する手法を提案する．対象は収集するハッシュタグと同時期の新聞記事とする．本研究では，ユーザの設定したタグを利用した動的なカテゴリ分類は可能か，あるいは異なるメディアにおいて，統一的なタグの利用は可能かを明らかにする．実験により，新聞記事の文書に対し，動的カテゴリを用いた効果的な分類が可能であることを示す．

2. 関 連 研 究

本研究に関連する研究として，フォクソノミーに関する研究

(注1)：https://twitter.com
(注2)：https://www.facebook.com

とハッシュタグの分類に関する研究がある．

フォクソノミーはソーシャルブックマークサービスである Delicious(注3) や写真共有サイトである Flickr(注4) といった SNS を発端として現在も多くのシステムで利用されている．フォクソノミーはエンドユーザによってタグが付与される仕組みで成立しており，多くのユーザが存在する Web のリソースと親和性が高い．また，Web のリソースはタグが付与されることで様々な意味づけがなされ，しばしば分析の対象になる．

馬場ら [2] は，タグと場所の共起情報と確率分布を用いてタグに関連する場所を推定する手法を提案した．タグと位置情報の共起の表現にはアスペクトモデルが用いられ，共起確率は混合ガウス分布を仮定している．また，タグと位置情報が関連する確率分布において，潜在変数の推定には EM アルゴリズム [3] により位置情報の推定を行なっている．この方法により，概念と地理的な情報との潜在的な関係をタグによって明らかにできることを示した．

丹羽ら [4] は，あるユーザのソーシャルブックマーク上のタグの分析することで，ユーザに対して Web ページを推薦する手法を提案した．丹羽らはまず，従来手法である協調フィルタリングにおいて，ユーザが全く同一の Web ページをブックマークしていなければユーザ間の嗜好の類似度が 0 になるという問題を指摘した．この問題を解決するため，ユーザの嗜好を各タグの親和度で表現し，嗜好類似度を算出した．タグをクラスタリングし，レベルを抽象化することで，推薦精度を下げることなく推薦ページの計算時間を短縮できることを示した．

こうした研究は，フォクソノミー上のタグにより意味付けされたリソースを利用して様々な分析が可能であることを示している．本研究により異なるメディア間において統一的なタグの利用を行うことができれば，こうした分析が可能となる範囲が増す．このため，フォクソノミー研究の文脈においても本研究の意義は大きい．

次にハッシュタグの分類に関する研究について述べる．Tsur ら [5] は，Twitter への投稿をはじめとする短い文書の分類を行う手法を提案した．Tsur らの提案した手法では短い文書の分類というタスクを，クラスタリングを行うフェーズと分類を行うフェーズの 2 段階に分ける．クラスタリングを行うフェーズでは，タグが現れる文書を全て繋げたタグ文書を仮定し，タグ文書を素性としてクラスタリングを扱う．これにより，短い文書において，スパース性のために TF-IDF ベクトルが機能しないという問題を解消する．また，このタグ文書に対し，k-means 法を適用し，クラスタリングを行う．以上を元に分類を行うことでオンラインのストリーム上においても短い文書を機械的に分類することが可能となる．

井上ら [6] は，Tsur らの手法の日本語に対する適用が可能であること，またユーザがイベントに参加しているか判定するという目的においてもこの手法が有効であること示した．

本研究ではこれらと同様のアプローチで，Twitter 以外ののメディアの文書のトピックを明らかにし，分類できることを示す．

木村ら [1] は，共起情報と潜在トピックを用いてハッシュタグ間の関係を構造化する手法を提案した．2 つのハッシュタグの共起度は相互情報量を用いたより精細な共起性の指標である AEMI(Augmented Expected Mutual Information) を用いて算出している．また LDA(Latent Dirichlet Allocation) を用いてハッシュタグの現れる文書を 1 つの文書とみなしたときのハッシュタグごとの潜在トピックを推定する．この潜在トピック分布の類似度によってハッシュタグ間の関係を推定している．木村らの手法は，Tsur ら，井上らと比較するとより多様な視点でハッシュタグ間の関係を分析している．本研究では木村らの示したハッシュタグの構造化までは行わないが，今後の課題として位置づけることができる．

3. 手　　法

3.1 ハッシュタグの特徴

本研究の目的は，端的に言えば Twitter で実際に使われているハッシュタグを他のメディアのテキストに付与することである．このためにまず，各ハッシュタグの特徴を抽出する必要がある．

ハッシュタグの特徴の抽出方法について説明する．ある期間に投稿された文書の全体集合 U とそこに現れるハッシュタグ集合 T を考える．文書集合 $D \subset U$ を $d \in D$ が少なくとも 1 つのハッシュタグ $t \in T$ を含むように選ぶ．さらに，ハッシュタグ文書集合 D' を仮定する．ここでハッシュタグ文書 $d_t \in D'$ は t が出現する文書すべてを繋げて作成した仮想文書である．本研究ではこのハッシュタグ文書をベクトル化したものをハッシュタグの特徴とみなす．

ハッシュタグ文書をベクトル化する手法として，本研究では TF-IDF を用いる．これは語彙の出現順序を考慮しない Bag of Words に基づく手法で，Salton ら [7] によって情報検索における単語の重み付けをする手法として提案された．TF-IDF を使うことで文書をベクトル空間上のベクトルとして表現できることからベクトル空間モデルともいわれる．TF-IDF は単語頻度である TF と逆文書頻度である IDF をかけることで重みの値を得る．この値は TF の値が大きくかつ文書頻度 DF の値が小さいほど大きくなる．直感的に言い換えると，文書内でよく出現する単語であり，かつ他に出現する文書の少ない単語ほど TF-IDF の値が大きくなる，ということである．

本研究における TF-IDF の具体的な定義について説明する．文書 d_t に語彙 w_i が出現する回数を表す単語頻度 $tf(w_i, d_t)$ を 1 のように定義する．

$$tf(w_i, d_t) = \sum_{w \in d_t} I_{w_i}(w) \tag{1}$$

ただし，$I_{w_i}(w)$ は式 (2) に示す指示関数である．

$$I_{w_i}(w) = \begin{cases} 1 & (w = w_i) \\ 0 & (w \neq w_i) \end{cases} \tag{2}$$

語彙 w_i が含まれる文書の総数を表す文書頻度 $df(w_i)$ を式

(注3)：https://del.icio.us

(注4)：https://www.flickr.com

(3) に示す．

$$df(w_i) = |\{d_t \in D' | w_i \in d_t\}| \tag{3}$$

語彙 w_i における逆文書頻度 (4) に示す．

$$idf(w_i) = log\frac{1+|D'|}{1+df(w_i)} + 1 \tag{4}$$

式 (4) に示した逆文書頻度は一般的な定義とは異なり，ゼロ除算を防ぐために全ての語彙が 1 回現れる文書 1 つを加えるように定義している．以上を元に文書 d_t に語彙 w_i における TF-IDF を式 (5) に示す．

$$\phi(w_i, d_t) = tf(w_i, d_t) \cdot idf(w_i) \tag{5}$$

$|D'|$ はハッシュタグ文書の総数であり，異なりハッシュタグ数と一致する．これによりハッシュタグ文書 d_t の特徴ベクトル $\boldsymbol{v}'_t \in \mathbb{R}^{|W|}$ は，語彙集合 W において式 (6) のように表せる．

$$\boldsymbol{v}'_t = \begin{bmatrix} \phi(w_1, d_t) \\ \phi(w_2, d_t) \\ \vdots \\ \phi(w_{|W|}, d_t) \end{bmatrix} \tag{6}$$

この $\boldsymbol{v}'_t$ を正規化した $\boldsymbol{v}_t$ は式 (7) のように与えられる．

$$\boldsymbol{v}_t = \frac{\boldsymbol{v}'_t}{\|\boldsymbol{v}'_t\|} \tag{7}$$

この正規化された特徴ベクトル $\boldsymbol{v}_t$ をハッシュタグ文書 d_t の特徴と定義する．

3.2 動的カテゴリの生成手法

動的カテゴリを生成する手法について述べる．本研究ではハッシュタグに対してクラスタリングを行い，生成されたクラスタを 1 カテゴリとする動的カテゴリを生成する．

クラスタリングとは，多次元空間上のラベルのない不完全なデータ群の持つパターンを，教師なし学習によって認識し，類似するパターンを持つデータ集合をクラスタとしてまとめる処理のことをいう．

タグは使われる語がユーザによって決められた任意のキーワードであるという特性から表記揺れが存在する．例えば，読売ジャイアンツに関する，実際に Twitter で使われているハッシュタグとして「#ジャイアンツ」,「#読売ジャイアンツ」,「#巨人」,「#読売巨人軍」,「#giants」,「#kyojin」などが存在する．これらのハッシュタグは同義のハッシュタグとして同時に割り当てられるべきである．また，タグは 1 つの文書に複数のタグをつけることができるという特性から，複数のタグで 1 つの出来事を表すという場合も考えられる．例えば，「鳥居みゆきと藤井ペイジがコンビで M1 に出場した」という出来事に言及した文書の場合，ハッシュタグは「#鳥居みゆき」と「#藤井ペイジ」の両方のタグがつけられるのが適当だろう．本研究ではこうした問題をクラスタという形で複数のハッシュタグを 1 カテゴリにまとめることで現実に利用されているハッシュタグの関係性を再現することができると考える．このためハッシュタグ文書をクラスタリングし，1 つのトピックを表すハッシュタグをハッシュタグクラスタにまとめ，これを 1 カテゴリとみなす．

またハッシュタグ文書をクラスタリングする方法としては，k-means 法を用いる．

3.3 *k*-means 法

k-means 法は，クラスタリングの最も代表的な手法である．k-means 法はクラスタ数 c を既知として，多次元空間上の各データ点に c 個のクラスタを割り当てる．クラスタ数は文字 k を使うのが一般的だが後述する k 近傍法のパラメータ k との混乱を防ぐために文字 c によってクラスタ数を表す．k-means 法のアルゴリズムは次のようになる：

（1） 各クラスタを代表するセントロイドを初期状態としてランダムに設定する．

（2） 与えられた各データ点について，このデータ点を距離が最小となるセントロイドのクラスタに割り当てる．

（3） 各クラスタについて，クラスタに属する各データ点の重心を求め，これを新たにこのクラスタのセントロイドとする．

（4） (3) でセントロイドに変化があれば (2) に戻り，変化がなければ収束したとみなして処理を終了する．

k-means 法は一般的には，c をそれほど大きい値に設定しない．しかし，井上ら [6] の研究を元に，c を文書数の 0.6-0.7 倍程度の大きい数として与えることで特定のトピックを表すハッシュタグが同一のクラスタとして得られることを仮定する．本研究では，こうした c を用いて k-means 法でクラスタリングを行うことで，他のメディアのテキストを適合した動的カテゴリに分類できると考える．

3.4 動的カテゴリの分類方法

コンテンツを動的カテゴリに分類する手法としては k 近傍法を用いる．まず，与えられた文書の特徴ベクトルとすでにクラスタリングされたハッシュタグ文書ベクトルのクラスタがあるとする．k 近傍法は，与えられた文書の特徴ベクトルと最も近傍な k の特徴ベクトルを参照し，これらのクラスタのうち最も多く見られたクラスタに分類するという方法である．

4. 実　　験

4.1 実験の概要

本研究で行う実験の概要を説明する．ある期間に Twitter で利用されているハッシュタグについてハッシュタグ文書コーパスを作成する．その後，新聞記事を動的カテゴリに分類する実験を行う．これにより，新聞記事は各ハッシュタグクラスタが一つのトピックを表す動的なカテゴリに分類される．また，生成された動的カテゴリとあらかじめ新聞記事が分類されている静的カテゴリのそれぞれに対し，カテゴリのまとまりのよさを比較する実験を行う．この実験により，既存手法である静的カテゴリを用いた分類と比較して動的に生成されたカテゴリが話題性のあるテキストに対し有効に分類を行えることを示す．

4.2 実験環境および実験データ

実験は CPU が Intel Xeon E5-2630 (2.40GHz) 8 core 2 機，OS が Ubuntu 16.04 という環境で行なった．実験データは，2012 年 4 月 7 日から 4 月 13 日に収集されたツイートと 2012 年 4 月 7 日から 4 月 13 日の毎日新聞の記事とした．毎日新聞の記事は「国際」や「経済」といったカテゴリに分類されており，このうち「1 面」や「解説」といったトピックの分類としてふさわしくないものを除いた 9 分類を静的カテゴリとして実験に用いた．ツイートおよび新聞記事に対する形態素解析には形態素解析機である MeCab [8] を使用した．文書の形態素解析，TF-IDF ベクトル化，k-means 法によるクラスタリングなどは全て Python によって実装を行った．文書の TF-IDF ベクトル化および k-means 法によるクラスタリングの実装は Python のオープンソースライブラリのひとつである scikit-learn [9] による．また，静的カテゴリと動的カテゴリの比較を行う実験にはクラウドソーシングサイトである Lancers(注5) を利用した．

4.3 実 験 方 法

2012 年 4 月 7 日から 4 月 13 日のツイートを対象に，ハッシュタグ文書コーパスを作成した．ハッシュタグ文書コーパスの作成には以下のような制限を設けた：

（1） $df(w) \geqq 20$ である語彙 w のみ用いる

（2） 「@ユーザ名」で表されるメンション，URL，ハッシュタグそのものをストップワードとして除く

（3） 品詞は一般名詞，固有名詞，サ変接続のみとする

（4） 100 回以上出現するタグについてのみハッシュタグ文書とする

（5） ハッシュタグに使われるキーワードの長さは 10 文字以下とする

各制限を設けた理由について述べる．(1) について，極端に出現頻度の少ないハッシュタグには，タイプミスのようなノイズや限られた人のみが使っているハッシュタグが含まれるため，これを除く．(2) について，メンションや URL は単なる記号列であり，共起する単語の特徴をもとに動的カテゴリを生成する目的と合致しないため，これを除く．(3) について，よりトピックの特徴が現れる品詞を選ぶ．(4) について，ハッシュタグ文書はある程度長いものでなければならないため，これを除く．(5) について，ハッシュタグには投稿する文書の属性やトピックではなく，「#○○だったら RT」，「#クリスマスプレゼントに欲しいもの」といった大喜利のお題のようなものや「#love」，「#enjoy」のように感情を表現したものも存在する．こうしたハッシュタグに使われるキーワードは比較的長い傾向にあるため，これを除くことにする．

以上のような方法で作成したハッシュタグ文書コーパスを元に，ハッシュタグ毎の TF-IDF ベクトルを作成する．その後，3.4 節で説明した k-近傍法を用いて新聞記事の TF-IDF ベクトルとハッシュタグ文書を類似度を計算し，ハッシュタグクラスタを用いて新聞記事を動的カテゴリに分類する実験を行う．k 近傍法のパラメータ k について井上ら [6] をもとに $k = 3$ として行うこととする．

(注5)：https://www.lancers.jp

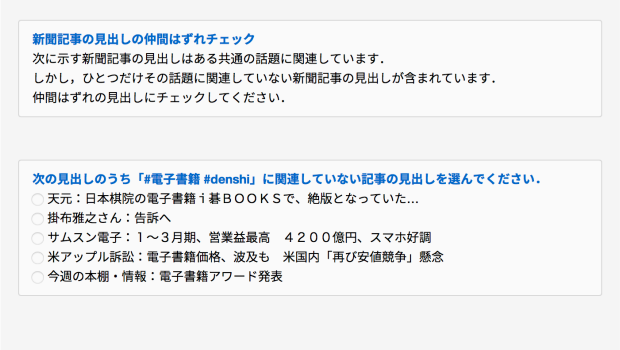

図 1　実際に行った Category intrusion のタスクの例

次に，生成された動的カテゴリとあらかじめ新聞記事が分類されていた静的カテゴリのまとまりのよさを比較する評価実験について具体的に説明する．

カテゴリのまとまりのよさを検証する目的と合致する手法として Chang ら [10] により考案された Word intrusion がある．Word intrusion は，トピックモデルにおいてトピックの Coherence（一貫性）を測定する手法として提案された．この手法では，あるトピックにおいて出現確率が高い複数の語と出現確率が低い 1 つの語を被験者に提示する．この上で，被験者は出現確率が低い 1 つの語がどれかを回答するタスクを行い，実験をもとに出現確率の低い語が選択された割合を算出する．この割合が高いトピックは人間にとってよくまとまったトピックであると評価することができる．

本研究ではこの Word intrusion を拡張した”Category intrusion”を用いて評価実験を行う．Category intrusion ではあるカテゴリに分類された記事 4 件の見出しと，このカテゴリに分類されていないランダムに選択された記事の見出し 1 件を被験者に提示する．この上で，被験者はこのカテゴリに含まれない 1 つの記事の見出しがどれかを回答するタスクを行い，実験をもとにカテゴリに含まれない記事の見出しが選択された割合を算出する．この実験を動的カテゴリと静的カテゴリのそれぞれについて行い，これらの結果に有意差が見られるかを χ^2 検定により確かめる．

本研究では，同一のカテゴリに含まれる新聞記事が 4 件以上であるカテゴリ 1 つに対し，このカテゴリに含まれる 4 件の新聞記事とこのカテゴリに含まれない 1 件の新聞記事をランダムに選択するという方法でタスクを作成する．各カテゴリに対し，動的カテゴリでは 5 タスク，静的カテゴリでは 15 タスクを作成する．図 1 に実際に行ったタスクの例を示す．本研究では図 1 のようにカテゴリのラベルであるハッシュタグを示した上でこれに関連していない記事の見出しを選ばせている．この例では「掛布雅之さん：告訴へ」という見出しの記事がカテゴリに含まれない記事である．

表 1　Category intrusion による実験の結果

	カテゴリに含まれる記事を選んだ回数	カテゴリに含まれない記事を選んだ回数	計
動的カテゴリ	74	476	550
静的カテゴリ	39	96	135
計	113	572	685

表 2　記事の見出しと分類されたカテゴリの例

新聞記事の見出し	カテゴリ
大阪・福島の変死：殺人容疑で男指名手配　暴行し高架下放置 器物損壊：元組長に依頼して、同僚警官が車損壊　容疑で逮捕－－岡山	#jiken #事件
訃報：方励之さん　７６歳＝中国の反体制派物理学者 シーラカンス：４億歳　中国で最古の化石発見	#中国 #china
ボクシング：アジア選手権　須佐が五輪へ ソフトボール：女子世界選手権　日本代表に上野ら	#olympic #オリンピック

4.4 実 験 結 果

ハッシュタグ文書コーパスを作成した結果，条件に当てはまる異なりハッシュタグ数は，11445 だった．したがって，ハッシュタグ文書集合 D' について，$|D'| = 11445$ である．井上らの結果から k-means 法におけるクラスタ数 c は $c = 0.6|D'|$ とした．以上より，この D' に対し，$c = 6768$ で k-means 法によるクラスタリングを行ない，カテゴリを生成した．1764 件の新聞記事を生成した動的カテゴリに分類した結果，実際に割り当てられたカテゴリは 765 種類だった．

また，Category intrusion のタスクは動的カテゴリは 4 件以上の新聞記事を含むカテゴリが 110 種類あったため 550 タスクを作成し，静的カテゴリは 9 種類であるため 135 タスクを作成した．動的カテゴリと静的カテゴリのまとまりのよさを比較する評価実験の結果を表 1 に示す．

表 1 より，検定統計量 T の実現値 t について $t = 18.7$ がわかる．自由度は 1 であり，$t > \chi^2_1(0.005) = 7.88$ であるため，有意水準 0.005 で有意差がある．すなわち，動的カテゴリは静的カテゴリと比較してカテゴリのまとまりがよく，被験者からはカテゴリが認識しやすかったことがわかる．これは 9 種類のみの静的カテゴリと比較して，動的カテゴリはより詳細かつ適当な分類を行えていたためだと考えられる．

また，記事の見出しと分類されたカテゴリの例を表 2 に示す．生成されたカテゴリの中には，「#jiken #事件」や「#中国 #china」といった普遍的なカテゴリが見られた．一方で，「#olympic #オリンピック」のようにカテゴリ生成に利用したツイートの投稿時期に特有のトピックを表すカテゴリも見られた．これは Twitter のハッシュタグを用いてカテゴリを生成することで話題性を反映したカテゴリを生成できるという意図が成功している例だと言える．

5. 結　　論

本研究では，Twitter で実際に使われているハッシュタグに対してクラスタリングを行い，得られた特徴を元に他のメディアのテキストを動的カテゴリに分類する手法を提案した．実験により，提案手法による動的カテゴリはあらかじめ決められた分類による静的カテゴリと比較してまとまりがよくなることを示した．

本研究では動的カテゴリのまとまりのよさについてのみ検証を行なった．本研究で生成した動的カテゴリはハッシュタグによる，いわばラベルが付いている．動的カテゴリのまとまりのよさを検証する実験では，ハッシュタグによるラベルを示すことはしていたものの，このハッシュタグがそのカテゴリに本当に適したハッシュタグであるかは未検証のままである．このため，今後はこのハッシュタグがそのカテゴリを表すラベルとして適しているか検証する研究が必要である．

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 16H02904）および筑波大学図書館情報メディア系プロジェクト研究の助成によって行われた．

文　　献

[1] 木村輔, 宮森恒. 共起と潜在トピックを考慮したハッシュタグ間関係の分類. 電子情報通信学会論文誌, Vol. J98-D, No. 8, pp. 1151–1161, 2015.

[2] 馬場雪乃, 石川冬樹, 本位田真一. Folksonomy 上のタグと関連する場所の抽出. 人工知能学会論文誌, Vol. 27, No. 1, pp. 1–9, 2012.

[3] A. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, Vol. 39, No. 1, pp. 1–38, 1977.

[4] 丹羽智史, 土肥拓生, 本位田真一. Folksonomy マイニングに基づく web ページ推薦システム. 情報処理学会論文誌, Vol. 47, No. 5, pp. 1382–1392, 2006.

[5] Oren Tsur, Adi Littman, and Ari Rappoport. Efficient Clustering of Short Messages into General Domains. *Proceedings of the 7th International Conference on Weblogs and Social Media, (ICWSM'13)*, pp. 1–10, 2013.

[6] 井上優作, 若林啓. 表記の多様性を考慮したハッシュタグ推薦. 第 14 回日本データベース学会年次大会, 2016.

[7] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*,

Vol. 24, No. 5, pp. 513 – 523, 1988.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. *Proc. of EMNLP-2004*, pp. 2–4, 2004.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

[10] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pp. 288–296. Curran Associates, Inc., 2009.