共起と潜在トピックを考慮したハッシュタグの構造化手法

木村　輔† 宮森　恒†

† 京都産業大学コンピュータ理工学部インテリジェントシステ学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1044390,miya}@cse.kyoto-su.ac.jp

あらまし　Twitter には，自身の投稿にタグを付ける，ハッシュタグという機能がある．これは，SBS(注1)などで発展したフォクソノミーという分類法の亜種である．ハッシュタグを用いて話題が明示されたツイートを検索・共有することで，ユーザは有益な情報を得ることができる．しかし，既存の手法によって推薦されたハッシュタグは，必ずしもユーザが求めていたものではない場合がある．そこで我々は，推薦精度向上のため，関連する他のハッシュタグの意味や利用の差異が分かるよう提示し補間することが重要であると考えた．本稿では，フォクソノミーの 3 部グラフ構造と，ハッシュタグの潜在トピックを用いて，ハッシュタグの構造化を行う手法を検討する．

キーワード　Twitter，ハッシュタグ，folksonomy，フォクソノミー，構造化，情報推薦

1. はじめに

スマートフォンやタブレット端末の普及に伴い，企業・団体だけではなく，より多くの個人がインターネットから情報を収集したり，発信したりできるようになった．消費者でしかなかった個人が，情報の生成者にもなりやすい環境になったことで，様々な手段を用いて非常に多様な情報が世界へと発信されている．例えば，SNS のカテゴリで言うと Twitter(注2) や Facebook(注3) などが挙げられる．

Twitter では，140 文字という制限内で書かれたテキストを投稿することができ，これをツイートという．タイムラインでは，フォローしたユーザ（以下，フォロイーと称す）と自身のツイートが投稿時間の新しいものから順に表示される．また Twitter には，自身のツイートに“#”を付けた文字列でタグを付与する，ハッシュタグという機能がある．これは，SBS などで発展したフォクソノミーという，コンテンツに対してユーザが自由にタグ付けを行い，検索できるようにする分類法である．ハッシュタグを用いて話題が明示されたツイートを検索・共有することで，ユーザは有益な情報を得ることができる．

しかし，ハッシュタグには以下のようにいくつかの問題を抱えている．

（1） 一見しただけではどのような利用をされてきたのか分かりにくい点

（2） 自身のツイートにどのハッシュタグを付与すれば適切なのか分からない点

（3） どのようなハッシュタグが存在しているのか分かりにくい点

このように，利用する上で情報の発信者・受信者共に問題を抱えていることもあり，ハッシュタグの利用率は低く，ハッシュタグ検索による有益な情報の共有などの，本来のポテンシャルを活かしきれているとは言い難い．

(1)(3) の問題を解決するために，ハッシュタグクラウド(注4) や hashtagsjp(注5) という Web サイトでは，ハッシュタグの説明をサイト利用者に登録・編集してもらうことによって，ハッシュタグの意味を調べ易くたり，検索し易くしたりするなど，ハッシュタグを便利にするサービスが提供されている．

また (2) の問題を解決するために，ツイートの内容に一致する適切なハッシュタグを推薦する研究が行われている．

しかし，既存の手法によって推薦されたハッシュタグは必ずしもユーザが求めていたものではない場合があるなど，また別の問題を抱えている．

そこで我々は，推薦精度向上のため，関連する他のハッシュタグの意味や利用の差異が分かるよう提示し補間することが重要であると考えた．本稿では，丹羽ら [2] が提案したフォクソノミーの 3 部グラフ構造を用いた手法をベースとし，トピック(注6) に基づく統計的言語モデルである，LDA（Latent Dirichlet Allocation）[6] という手法によりハッシュタグから抽出した潜在トピックを考慮したハッシュタグの構造化する手法を検討する．

本論文の構成は以下の通りである．まず，2 章にて関連研究をまとめ，3 章にて提案手法についての説明する．4 章で実験に用いるデータセットと評価実験の手法・目的をまとめ，5 章にて評価実験の結果と考察を述べる．6 章で本論文のまとめ，最後に 7 章で今後の課題を述べる．

2. 関連研究

ツイートへのハッシュタグ推薦を対象とした研究には以下のようなものがある．例えば，竹中 [3] は，ハッシュタグ毎にベイジアンフィルターを構築し，ツイートがどのハッシュタグに属しているのかを推定するシステムを考案している．小寺 [4]

(注1)：SBS : Social Bookmark Service の略称
(注2)：Twitter : https://twitter.com/
(注3)：Facebook : https://ja-jp.facebook.com/
(注4)：ハッシュタグクラウド : http://hashtagcloud.net/
(注5)：hashtagsjp : http://hashtagsjp.appspot.com/
(注6)：ここでいうトピックとは，単語の出現確率を変動させる複数の要因をまとめたものである．例えば，分野，話題，時期，文書，共起などがある．

は，ハッシュタグごとの特徴語を抽出し RandomForest 等の分類木を作成し，それを用いてハッシュタグを持たないツイートの本文内容の推定を目的としたシステムを考案している．また，フォクソノミーの構造化の研究として，丹羽ら [2] はドキュメント-タグ-ユーザの 3 部グラフ構造から得られるそれぞれの共起関係のみからタグを構造化する手法を提案している．谷田川ら [5] は，丹羽ら [2] の手法を Twitter のハッシュタグに適用させるために，類似したツイートを 1 つのドキュメントとみなす手法を用いている．

Twitter を対象とする LDA の手法を用いた研究として，Weng ら [7] は，影響力の強いユーザを発見するために，LDA を用いて推定したトピックを考慮した PageRank アルゴリズムを提案している．Zhao ら [8] は，1 つのツイートに 1 つのトピックを割り当てるという仮説を立て，Twitter 用に LDA を改良した Twitter-LDA を提案している．古賀ら [1] は，潜在的トピックに着目した，Twitter 上のユーザ推薦システムを提案している．

本研究は，ツイート本文中の単語-ハッシュタグ-ユーザの三部グラフ構造から得られる共起に関する特徴量に加え，ハッシュタグが付与されたツイートから LDA により抽出したトピックを用いて，ハッシュタグを構造化し適切な推薦を目指すものである．

3. 提 案 手 法

3.1 2 つの共起率を用いた構造化の仮説

ハッシュタグの構造化する手法として，谷田川ら [5] と同様に，ベースとして丹羽ら [2] の提案手法を用いる．ここでは，彼らが提案した構造化の手法と，それを Twitter のハッシュタグ構造化に用いる上で解決すべき問題点とその改善案について説明する．

3.1.1 2 つのハッシュタグ間の共起率

丹羽ら [2] の仮説を展開する上で必要な 2 つの共起率，ドキュメントベースの共起率とユーザベースの共起率の定義について説明する．ドキュメントベースの共起率を，“1 つのドキュメントに 2 つのタグが同時にラベリングされる確率”と定義する．共起率は AEMI(Augmented Expected Mutual Information) [9] を用いて以下の式で計算される．図 1 にて例題を示す．また計算で使用されているそれぞれの記号は表 1 にて解説する．

$$
\begin{aligned}
MI_D &= P_D(a,b)log\frac{P_D(a,b)}{P_D(a)P_D(b)} \\
AEMI_D &= MI_D(A=a,B=b) + MI_D(A=\overline{a},B=\overline{b}) \\
&\quad - MI_D(A=a,B=\overline{b}) - MI_D(A=\overline{a},B=b)
\end{aligned}
$$

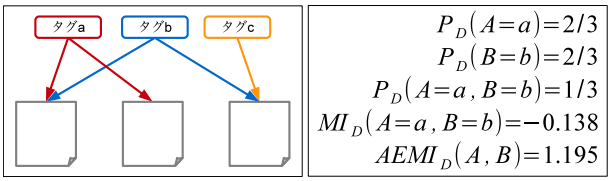

図 1 ドキュメントベースの共起率の計算例

表 1 $AEMI_D$ の計算に用いる記号の定義

記号	定義
A,B	あるタグの集合
a,b	タグの集合のある要素
$P_D(a)$	任意のドキュメントにタグ a がラベリングされる確率
$P_D(\overline{a})$	任意のドキュメントにタグ a がラベリングされない確率
$P_D(a,b)$	任意のドキュメントにタグ a,b が同時にラベリングされる確率
MI_D	タグ a,b 間の相互情報量
$AEMI_D$	1 つのドキュメントに 2 つのタグが同時にラベリングされる確率

同様に，ユーザベースの共起率を，“1 人のユーザが 2 つのタグを同時にラベリングする確率”と定義すると，以下の式で表せる．図 2 にて例題を示し，計算で使用されているそれぞれの記号は表 2 にて解説する．

$$
\begin{aligned}
MI_U &= P_U(a,b)log\frac{P_U(a,b)}{P_U(a)P_U(b)} \\
AEMI_U &= MI_U(A=a,B=b) + MI_U(A=\overline{a},B=\overline{b}) \\
&\quad - MI_U(A=a,B=\overline{b}) - MI_U(A=\overline{a},B=b)
\end{aligned}
$$

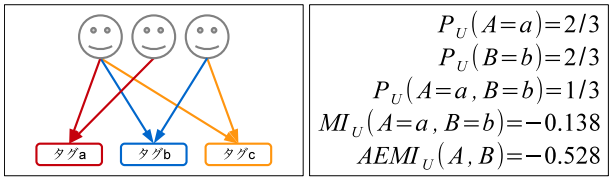

図 2 ユーザベースの共起率の計算例

表 2 $AEMI_U$ の計算に用いる記号の定義

記号	定義
A,B	あるタグの集合
a,b	タグの集合のある要素
$P_U(a)$	任意のユーザがタグ a をラベリングする確率
$P_U(\overline{a})$	任意のユーザがタグ a をラベリングしない確率
$P_U(a,b)$	任意のユーザがタグ a,b を同時にラベリングする確率
MI_U	タグ a,b 間の相互情報量
$AEMI_U$	1 人のユーザが 2 つのタグを同時にラベリングする確率

3.1.2 仮 説

丹羽ら [2] は 2 つの共起率を用いてタグ間の関係性を推定するために，次のような仮説を展開した（図 3）．

（1） Synonym 関係：2 つのタグが Synonym 関係の場合，それらのユーザベース共起率（$AEMI_U$）はドキュメントベース共起率（$AEMI_D$）に比べて大幅に低い傾向にある

（2） Conflict 関係：2 つのタグが Conflict 関係の場合，それらのユーザベース共起率（$AEMI_U$）はドキュメントベース共起率（$AEMI_D$）に比べて高い傾向にある

（3） Relevant 関係：2 つのタグが Relevant 関係の場合，それらのユーザベース共起率（$AEMI_U$）とドキュメントベース共起率（$AEMI_D$）は共に高い傾向にある

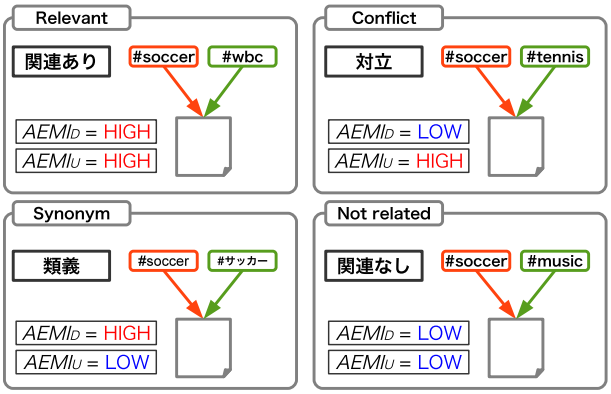

図 3　構造化する上での仮説

しかし，これはフォクソノミーを対象とした構造化のための仮説である．Twitter は完全なフォクソノミーではなく，いうなれば制限付きフォクソノミーである．本来，フォクソノミーは，任意のユーザが，任意のドキュメントに，自由にタグを付けし情報を統率する分類手法のことをいう．

しかし，Twitter では，ドキュメントにあたるツイートを投稿した本ユーザのみが，ハッシュタグを付与することができる（図 4）．

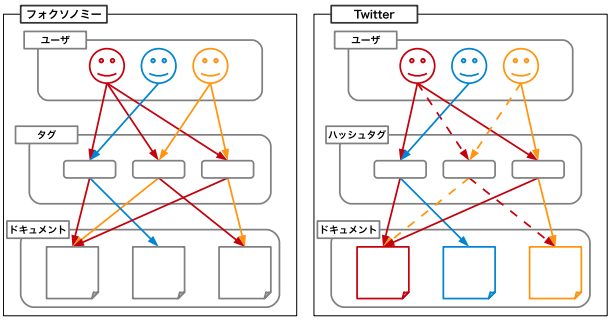

図 4　Twitter と一般的な Folksonomy の違い

そのため，本問題に丹羽らの提案手法・仮説をそのまま適用できるかどうかは必ずしも明らかではない．そこで我々は，この問題を解決するため，Twitter のハッシュタグの使われ方の特徴・傾向を調べた．そして，以下のようなことが確認できた．

（1）　ハッシュタグは，ツイートで言及している話題そのものを指すことが多い

（2）　ハッシュタグは，ツイート本文によって何らかの説明をされていることが多い

この調査から，ハッシュタグが先行研究でのドキュメントの役割を担い，ツイートの本文がタグの役割を担っていると考えられる（図 5）．以上をふまえ我々は，ハッシュタグ-ツイート本文中の単語-ユーザの三部グラフ構造を考えることで，Twitter に適応可能な新たな仮説を提案する．

（1）　Synonym 関係：2 つのタグが Synonym 関係の場合，それらのユーザベース共起率（$AEMI_U$）は単語ベース共起率（$AEMI_W$）に比べて大幅に低い傾向にある

（2）　Conflict 関係：2 つのタグが Conflict 関係の場合，それらのユーザベース共起率（$AEMI_U$）は単語ベース共起率（$AEMI_W$）に比べて高い傾向にある

（3）　Relevant 関係：2 つのタグが Relevant 関係の場合，それらのユーザベース共起率（$AEMI_U$）と単語ベース共起率（$AEMI_W$）は共に高い傾向にある

本研究では，Synonym 関係，Conflict 関係，Relevant 関係の仮説を用いて，ハッシュタグ間の関係性を図 3 に示す 4 つのクラスに分類し構造化を目指す．

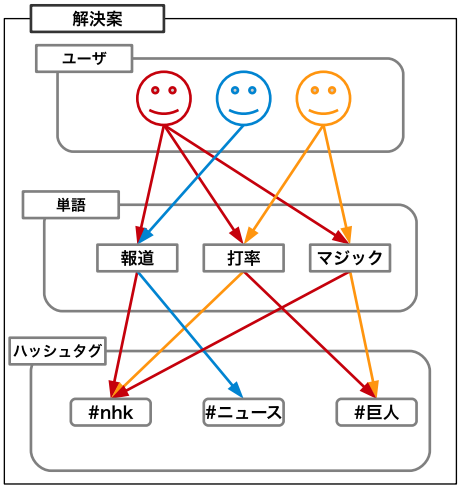

図 5　ハッシュタグ-ツイート本文中の単語-ユーザの三部グラフ構造

なお，新たな仮説で用いた $AEMI_W$ の定義を“1 つの単語に 2 つのタグが同時に使用される確率”とし，以下の式で計算するものとする．また計算で使用されているそれぞれの記号は表 3 にて解説する．

$$
\begin{aligned}
MI_W &= P_W(a,b)log\frac{P_W(a,b)}{P_W(a)P_W(b)} \\
AEMI_W &= MI_W(A=a,B=b)+MI_W(A=\overline{a},B=\overline{b}) \\
&\quad - MI_W(A=a,B=\overline{b})-MI_W(A=\overline{a},B=b)
\end{aligned}
$$

表 3　$AEMI_W$ の計算に用いる記号の定義

記号	定義
A,B	あるタグの集合
a,b	タグの集合のある要素
$P_W(a)$	任意の単語でタグ a が使用される確率
$P_W(\overline{a})$	任意の単語でタグ a で使用されない確率
$P_W(a,b)$	任意の単語でタグ a,b で同時に使用される確率
MI_W	タグ a,b 間の相互情報量
$AEMI_W$	1 つの単語に 2 つのタグが同時に使用される確率

3.2　LDA(Latent Dirichlet Allocation)

ここでは，LDA について簡単な解説をする．

LDA は，文書が潜在的な複数のトピックからなると仮定した，確率的な文書生成モデルである．本稿では，1 つのハッシュタグについて全てのツイート本文を連結し，1 つの文章とみなすことで，ハッシュタグ毎に潜在トピックの分布を推定する．この特徴量から潜在トピック分布の類似度が算出でき，類似度

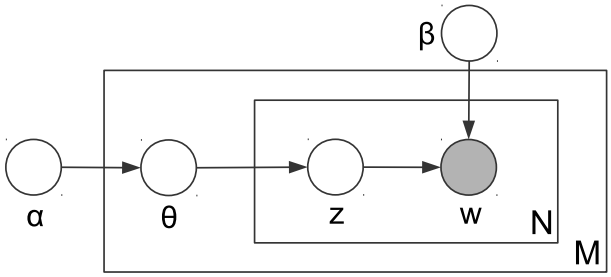

図 6　LDA のグラフィカルモデル

が高いと予想される Synonym 関係，類似度が低いと予想される Not related 関係の推定精度を向上させると考えられる．

具体的な生成過程を説明する上で，グラフィカルモデルにより LDA を示す（図 6）．グラフィカルモデルとは，確率変数またはパラメータを頂点とし，それらの依存関係を有向辺で接続したものである．矩形内に書かれた文字は，処理の繰り返し回数である．また表 4 にて記号の解説する．

表 4　LDA の説明に用いる記号の定義

記号	定義
α	トピック分布を生成するためのディリクレ分布のパラメータ
β	単語分布を生成するためのディリクレ分布のパラメータ
θ	文書のトピック分布
z	各単語のトピック
w	単語
N	ある文書の単語数
M	対象となる文書数

LDA における文書生成過程は，以下の通りである．

- 各トピックについて

（1） ディリクレ分布に従ってトピック別の単語の多項分布をサンプリング

- 各文書 d = 1,…,M について

（1） ディリクレ分布に従ってトピックの多項分布（θ）をサンプリング

（2） 文書 d における各単語 n = 1,…,N について

（a） 多項分布（θ）に従ってトピック（z）をサンプリング

（b） 多項分布に従って単語（w）をサンプリング

トピックの集合 Z を推定するために，変分ベイズ法 [Blei+ 2001,2003]，Gibbs サンプリング [Griffiths&Steyvers 2004]，期待値伝搬法 [Minka&Lafferty 2002]，Collapsed 変分ベイズ法 [Teh+ 2006]，固有値計算 [Anandkumar+, arXiv 2012] などさまざまな手法が考案されている．本論文では以上の推定方法の説明は割愛する．

3.3　システム構成

本研究の提案システムにおけるハッシュタグ構造化の処理概要を図 7 に示す．

3.3.1　前 処 理 部

- クローリング：Twitter から Streaming API を用いて，あらかじめ人手で選択・収集したハッシュタグが付与されたツイートを収集する．収集したツイートは表 5 の様に DB に登

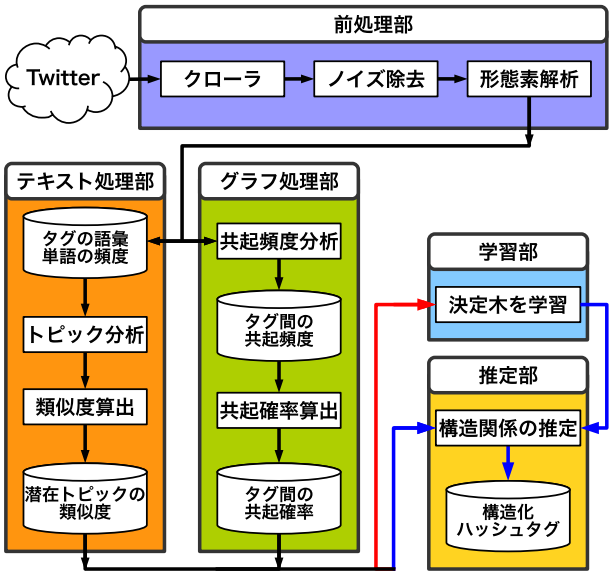

図 7　提案システムの処理概要

録している．また，曜日毎の学習ができるように，表 6 の様に DB に登録している．なお表中の‘X’は，0-9 のいずれかを意味し，ツイート ID は 16 桁の整数が用いられている（2010 年 11 月 7 日時点）．

表 5　ツイート本文の登録例

ツイート ID	JSON 形式のツイート情報
XXXXXXXXXX	{”description”:”日本経済が一日でも早く,…}
XXXXXXXXXX	{”description”:”日本は素晴らしい,…}
XXXXXXXXXX	{”description”:”E テレのお願い！編集長にて,…}

表 6　ツイート作成日時情報の登録例

ツイート作成日	ツイートに付与されたハッシュタグ	ツイート ID
1390230000000	nhk	XXXXXXXXXX
1390230000000	nhk	XXXXXXXXXX
1390230000000	nhk	XXXXXXXXXX

- ノイズ除去：ツイート本文の解析のノイズになると考えられる，ツイートの話題を明示するハッシュタグ（#hashtag），返信などに用いられるメンション（@username），外部サイトへのリンクである URL（http://...）を本文から全て削除する．
- 形態素解析：入力されたツイートの本文を形態素解析し，各ハッシュタグ毎の単語の出現頻度を調べ，ハッシュタグ毎の単語の出現頻度表，ハッシュタグ全体の語彙表を表 7, 表 8 の様に DB に登録する．

表 7　ハッシュタグ毎の単語の出現頻度の登録例

ハッシュタグ	単語	出現頻度
nhk	法	44
nhk	報道	34
nhk	ハーブ	33

表 8　ハッシュタグ全体の語彙の登録例

単語
一ノ蔵
一休
一口

3.3.2　テキスト処理部

- トピック分析：作成した DB（表 7, 表 8）を入力とし LDA を用いて，各ハッシュタグ毎のトピック分布を求める
- 類似度算出：ハッシュタグ毎のトピック分布の類似度を，KLD（Kullback-Leibler Divergence）を拡張した JSD（Jensen-Shannon Divergence）で算出し，表 9 の様に DB に記録する．

表 9　潜在トピックの類似度の登録例

ハッシュタグ$_A$	ハッシュタグ$_B$	類似度
nhk	天気	2.447e-08
nhk	NHK	6.111e-08
nhk	ニュース	2.845e-07

3.3.3　グラフ処理部

- 共起頻度分析：入力されたツイートからハッシュタグのみを抽出し，ハッシュタグを利用したユーザを記録したハッシュタグ別利用ユーザ表を表 10 の様に DB に登録する．なお表中の‘X’は，0-9 のいずれかを意味し，ツイート ID は最大 15 桁の整数が用いられている（2011 年 5 月 11 日時点）．

表 10　ハッシュタグ別利用ユーザ表の登録例

ハッシュタグ	ユーザ ID
nhk	XXXXXXXXX
nhk	XXXXXXXXX
nhk	XXXXXXXXX

- 類似度算出：作成された DB（表 10）とハッシュタグ毎の単語の出現頻度表（表 7）を入力とし，各ハッシュタグ間の共起確率を AEMI を用いて算出し，表 11 の様に DB に記録する．

表 11　共起頻度の類似度の登録例

ハッシュタグ$_A$	ハッシュタグ$_B$	$AEMI_W$	$AEMI_U$
nhk	天気	-0.0427	-0.0887
nhk	地震	-0.0273	-0.0940
nhk	news	-0.0306	-0.1180

3.3.4　学　習　部

テキスト処理，グラフ処理によって得られた DB（表 9, 表 11）を入力とし決定木を作成する．

3.3.5　推　定　部

作成された分類木と，関係を推定したいハッシュタグ間の 3 つの特徴量を用いて，Synonym 関係（類似），Conflict 関係（対立），Relevant 関係（関連あり），Not related 関係（関連なし）のハッシュタグ間の関係を推定し，その結果を DB に登録する．

4.　評 価 実 験

4.1　データセット

評価実験に用いるデータセットとして Twitter の検索 API(注8) を用いて，任意のハッシュタグ 98 件の 2014/1/22 2014/1/29 までのツイート（748,426 件）を収集した．

4.2　実 験 目 的

本稿は，推薦されたハッシュタグの精度を改善するために，ハッシュタグを構造化する重要性を示したい．そのためにはまず，提案した手法を用いて高い推定精度でハッシュタグを構造化できることを示さなけらばならない．よって本実験では，ハッシュタグ間の関係性の推定を従来手法の精度を保ちつつ向上できるのかを評価することが目的である．

4.3　評 価 方 法

提案手法によりハッシュタグ間の関係性を適切に推定できているかを評価するためには，実験に用いたハッシュタグ間の関係についての正解をあらかじめ用意する必要がある．よって我々は，ハッシュタグクラウドや hashtagsjp を用いて，人手により実験対象となったハッシュタグ間の関係性を定義した．用意した学習データを用いて，提案手法によって作成された決定木の適合率の分類精度を評価した．形態素解析には，MeCab(注9) を用い，デフォルト辞書として JUMAN(注10) で提供されている辞書を設定した．決定木の生成には，データマイニングツールである Weka3(注11) に実装されている J48 のアルゴリズムを用いた．

5.　結果と考察

5.1　分類精度の結果

本研究の提案手法の実験結果を図 8 に示す．

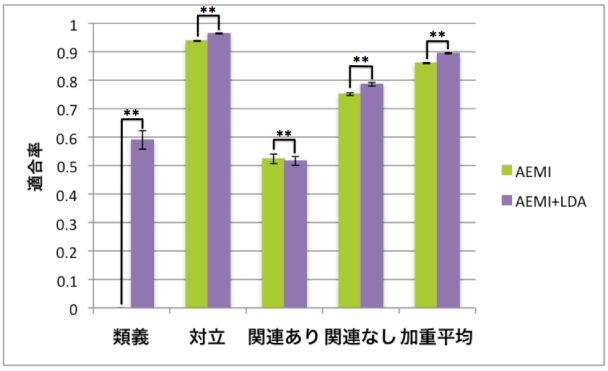

図 8　各評価項目の平均と標準偏差

5.2　考　　察

前説の評価実験の結果に対して，有意水準 1%として t 検定を行った．Synonym 関係，Relevant 関係，Conflict 関係，Not related 関係，総合精度，全てのついて精度に差があると結論

(注8)：TwitterSearch：http://search.twitter.com/
(注9)：MeCab：http://mecab.sourceforge.jp/
(注10)：JUMAN：http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
(注10)：Weka3：http://www.cs.waikato.ac.nz/ml/weka/

づけることができた．

図 8 より，従来手法（AEMI）に，特徴量として潜在トピックの類似度（LDA）を考慮した決定木（AEMI+LDA）の Synonym 関係の精度が大幅に向上しているのが分かる．また，潜在トピックの類似度の特徴量を用いることで，Relevant 関係の精度が低下するが，Synonym 関係，Conflict 関係，Not related 関係については精度が向上することが分かった．

5.3 誤り解析

実験では，Relevant 関係の適合率が特に低かった．これは仮説では“2 つのタグが Relevant 関係の場合，それらのユーザベース共起率 （$AEMI_U$）と単語ベース共起率（$AEMI_W$）は共に高い傾向にある”となっていたが，実験に用いたデータセット内では全ての関係において $AEMI_W$ が $AEMI_U$ より低い値になっており，仮説とは異なり 2 つの共起率間で大小関係が発生してたためであると考えられる．同様に，前述のとおり全ての関係において $AEMI_W$ が $AEMI_U$ より低い値になっていたため，他の関係についての仮説も上手く機能しておらず，先行研究と比べ，Synonym 関係・Relevant 関係共に適合率は低い値になった．ただ，Conflict 関係の精度のみ，先行研究より良い結果となったが，これは，学習データの偏りが Conflict 関係について大きくなっていたためであると考えられる．

全体的に AEMI_W が $AEMI_U$ より低い値になった原因として，仮説で扱っていたノードである“ドキュメント”を，本実験では“ツイート本文中の単語”という異なる粒度のノードへ置き換えてしまったことや，ハッシュタグの利用のされ方が，先行研究の実験対象であった SBS(注1)でのタグの利用のされ方と異なっていたためであると考えられる．

6. まとめ

本論文では，フォクソノミーの 3 部グラフ構造と，ハッシュタグの潜在トピックを用いて，ハッシュタグを構造化する手法の検討を行ってきた．評価実験により，LDA を用いてハッシュタグから抽出した潜在トピック間の類似度を用いることで，Synonym 関係，Conflict 関係，Not related 関係の抽出について有効であることが分かった．

しかし，Relevant 関係については精度が低下することが分かった．今後の課題としては，Relevant 関係の精度を向上しつつ，その他の関係の精度を保てるような新たな特徴量を考案することや，ハッシュタグの利用のされ方を考慮した新たな仮説の検討すること，より詳細な評価実験を行うことが挙げられる．

7. 今後の課題

現在，新たな特徴量として，ツイートから取得できる時系列情報の類似度に関する特徴量を考慮したいと考えている．これは，Synonym 関係にあるタグが，Conflict 関係や Not related 関係と比べて，同時間帯に，同周期で用いられやすいという傾向・仮説から分類する上で有力であると考えられるからである．また，ツイート本文からの単語の抽出精度向上のため，形態素解析器のユーザ辞書の作成を試みたいと考えている．Twitter では未知語（新語）が誕生しやすく，形態素解析器のデフォルト辞書では，十分に対応しきれているとは言い難い．そこで，Wikipedia などの未知語の登録・更新が早い Web リソースを用いて，ユーザ辞書の作成し，単語の抽出精度向上を目指す．本実験では作業の関係で限られたデータセットしか用意できなかったが，データを整えた上で実験に臨みたいと考えている．

文　献

[1] 古賀裕之，谷口忠大 潜在トピックに着目した Twitter 上のユーザ推薦システムの構築 ヒューマンインタフェースシンポジウム 2010

[2] 丹羽智史，土肥拓生，本位田真一 Folksonomy の 3 部グラフ構造を利用したタグクラスタリング 人工知能学会 セマンティックウェブとオントロジー研究会，(2007)

[3] 竹中姫子，古宮嘉那子，小谷善行 ベイジアンフィルターを用いた Twitter におけるツイートのハッシュタグ分類 情報処理学会研究報告. 情報学基礎研究会報告 vol.2011, no.1, pp.1-6, 2011-03-21

[4] 小寺英爾 ハッシュタグを用いた RandomForest による tweet 分類精度の向上 南山大学大学院 数理情報研究科 2012 年度 修士論文・OJL 報告書要旨集

[5] 谷田川将之，永森光晴，杉本重雄 Twitter ハッシュタグの構造化に関する研究，全国大会講演論文集 2011(1), 693-695, 2011-03-02

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan,“Latent Dirichlet Allocation,”Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.

[7] Weng, J., Lim, E.-P., Jiang, J., and He, Q.,“TwitterRank: Finding Topic-sensitive Influencial Twitterers,” Proc. of WSDM ’10, pp.261—270, 2010.

[8] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.- P., Yan, H., and Li, X.,“Comparing Twitter and Tra- ditional Media using Topic Models,”Proc. of ECIR ’11, pp.338—349, 2011.

[9] Chan, P.K. : A non-invasive learning approach to building web user profiles, KDD-99 Workshop on Web Usage Analysis and User Profiling, (1999).