Twitterの多様なハッシュタグ間の同義・階層関係の推定

坂本　翼† 廣田　雅春†† 横山　昌平††† 石川　博†††

† 静岡大学大学院情報学研究科　〒 432–8011 静岡県浜松市中区城北 3–5–1
†† 静岡大学創造科学技術大学院　〒 432–8011 静岡県浜松市中区城北 3–5–1
††† 静岡大学情報学部　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †gs11022@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp,
†††{yokoyama,ishikawa}@inf.shizuoka.ac.jp

あらまし　近年，Twitter に代表されるマイクロブログが普及し，多くのユーザによってツイートが投稿されている．大量のツイートの検索を行う方法のひとつとして，ハッシュタグを用いた検索があげられる．ハッシュタグは，各ユーザがツイートを投稿する際に自由に定義してツイートに付与することができる．しかし，同じトピックに関するツイートについて,「# 紅白歌合戦」と「#kouhaku」のように表記が異なるが同義のハッシュタグが付与される場合や,「#NHK」と「# 紅白歌合戦」のように階層関係にあるハッシュタグが付与される場合がある．そのため，ひとつのハッシュタグを用いて検索しても，ユーザが求める全てのツイートを取得できないという課題がある．本論文では，このような課題を解決してクエリ拡張を行うことによって検索支援を行うことを目的とし，各ハッシュタグのバースト情報や共起情報を用いてハッシュタグ間の関係性を推定する手法を提案する．

キーワード　ハッシュタグ検索，マイクロブログ，Twitter，ハッシュタグ構造

Estimate Structure of Hashtags in Twitter

Tsubasa SAKAMOTO†, Masaharu HIROTA††, Shohei YOKOYAMA†††, and Hiroshi ISHIKAWA†††

† Graduate School of Informatics, Shizuoka University
3–5–1 Jouhoku, Naka–ku, Hamamatsu, Shizuoka, 432–8011 Japan
†† Graduate School of Science and Technology，Shizuoka University
3–5–1 Jouhoku, Naka–ku, Hamamatsu, Shizuoka, 432–8011 Japan
††† Faculty of Informatics, Shizuoka University
3–5–1 Jouhoku, Naka–ku, Hamamatsu, Shizuoka, 432–8011 Japan
E-mail: †gs11022@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp,
†††{yokoyama,ishikawa}@inf.shizuoka.ac.jp

Abstract　In recent years, many users submit tweets to microblogging service such as Twitter. In Twitter, we can use a hashtag to search tweets. Twitter allows users to annotate tweets with hashtag as metadata to specify the topic. However, Twitter contains synonym hashtag such as "#kouhaku" and "#kohaku" that are added to tweets on same topic. In similar case, users may add hierarchical hashtag such as "#kouhaku" and "#NHK" to tweets on same topic. In this paper, we propose a method to extract relations between hashtags and facilitate the search for Twitter.

Key words　hashtag retrieval, microblogging, Twitter, hashtag structure,

1. 背　　景

近年，Twitter [1] に代表されるマイクロブログが普及し，多くのユーザによって記事が投稿されている．Twitter のツイートの数は，2012 年 11 月 16 日の発表において，2.5 日間に 10 億件，月間アクティブユーザ数は 2012 年 12 月 18 日の発表に

おいて，2 億人を超えたと発表されている(注1)．マイクロブログに投稿される記事は，短いテキストであり，Twitter の記事であるツイートの場合，140 文字以内という制約がある．投稿記事の短さや，携帯端末から利用可能なアプリケーションの普及があり，更新が容易であることが特徴である．ユーザが投稿する記事のトピックは，日常生活に関するものからニュースや，ユーザが視聴しているテレビ番組，参加している行事など多岐に渡る．従来のメディアと比較してこれらの記事は即時性が高く，マイクロブログの記事は，ユーザにとって有益な情報源となっている．

Twitter において，膨大なツイートの中からユーザが求めるツイートを発見する主な手段として，キーワード検索，およびハッシュタグ検索が挙げられる．キーワード検索では，入力したキーワードを含むツイートの一覧を取得できる．ユーザはトピックに関連するキーワードを入力することで，検索結果としてそのトピックに関するツイートを得られる可能性があるが，トピックに関連していてもキーワードを含まないツイートは得られないという場合や，キーワードを含んでいるが関連のないツイートが含まれる場合がある．ハッシュタグ検索は，ツイートに付与されたハッシュタグを用いた検索である．ハッシュタグは，ユーザによってツイートに付与されるタグである．ユーザは，ツイートを投稿する際に「#(任意の文字列)」で構成されるタグをツイートに付与することができる．ハッシュタグは，ツイートがどのようなトピックに関するものかを表すために付与される場合が多い [2]．ハッシュタグ検索は，ユーザが興味のあるトピックに関するハッシュタグを入力することで，そのハッシュタグが付与されたツイートを取得することができる．一般的なキーワード検索は，部分一致検索であるが，ハッシュタグ検索は，完全一致検索である．例えば，「#NHK 紅白歌合戦」の検索結果に，「# 紅白歌合戦」や「# 紅白歌合戦」のハッシュタグが付与されたツイートは含まれない．Teevan ら [3] は，Twitter における検索のクエリを収集して Web における検索のクエリと比較し，Twitter においてハッシュタグによる検索が，ユーザがツイートを発見する手段として一般的であることを示している．

ハッシュタグは，ユーザが自由に定義するため，Folksonomy の一種と考えられる．Folksonomy は，多くのユーザがタグ付けを行うことによって分類がなされる．丹波らは，従来の Taxonomy による分類と比較し，Folksonomy による分類は，多くのユーザが分類を行うため分類速度が速い，多くのユーザにとって分類が実用的である，時代のパラダイムやユーザの意識に適応した分類構造が自動的になされるといった特長を挙げている [4]．これらの特長は，Twitter のハッシュタグも同様に持ち合わせていると考えられる [5]．一方，Twitter のハッシュタグに関しても， Folksonomy と同様に同義語，タグの抽象度に関する課題 [6] がある．

- 同義語について

タグの文字列は，ユーザが自由に決めることができるため，ひとつのトピックに関するツイートに対して，ハッシュタグの表記揺れが発生する場合がある．例として，NHK 紅白歌合戦のトピックに関するツイートに対して，「#NHK 紅白」，「# 紅白」，「# 紅白歌合戦」，「#kouhaku」，「#kohaku」などのように，日本語表記と英語表記などの言語の違いや，ローマ字表記，平仮名，カタカナの違いなどによって，同じトピックを表すが表記の異なるタグが存在し，共通のトピックを持つツイートに多様なハッシュタグが付与される場合がある．

- ハッシュタグの抽象度の違いについて

ユーザが自由にハッシュタグをツイートに付与するため，付与されるタグの抽象度が異なる場合がある．例として，NHK 紅白歌合戦のトピックに関するツイートに対して「#NHK」と「# 紅白歌合戦」のように抽象度が異なるタグが付与される．「#NHK」は「# 紅白歌合戦」が表すトピックを包含するハッシュタグであり，「#NHK」が上位，「# 紅白歌合戦」が下位の階層関係にある．ユーザの認識によって，これらのような階層関係にあるハッシュタグが選択されるため，共通のトピックを持つツイートに階層関係にある異なるハッシュタグが付与される場合がある．

- 同音異義語について

ユーザが自由にハッシュタグをツイートに付与するため，ひとつのタグが複数のトピックを表す場合がある．ハッシュタグの例として，「#ooo」というハッシュタグが存在し，もともとは OpenOffice.org に関するハッシュタグとして用いられていたが，テレビ番組「仮面ライダーオーズ/OOO」の放映とともに視聴者が「#ooo」を利用した．そのため，異なるトピックをに関するツイートがひとつのハッシュタグに関するツイートとして扱われる場合がある．

上記の同義語，ハッシュタグの抽象度の違いによって同一のトピックに関するツイートに，多様なハッシュタグが付与されている場合がある．そのため，ユーザがあるトピックに関するひとつのハッシュタグを用いて検索した際に，複数のハッシュタグが付与されているようなトピックに関するツイートを検索する場合，ひとつのハッシュタグのみを用いてそのトピックに関するツイートを全て取得することはできない．ハッシュタグのタイムラインの内容は時間とともに変遷するため，ある時間では同一のトピックに属するツイートが含まれていたハッシュタグ間であっても，時間変化によってそれぞれ異なるトピックに属するツイートが含まれる可能性がある．そこで，本手法では，指定された推定期間に対して，時系列的な内容に基づいたハッシュタグ間の関係を推定する．これによって，同一の時系列のトピックを持つハッシュタグを発見し，組み合わせて検索を行うことによって，ユーザが求める内容に関するツイートをより多く取得できるようになると考えられる．例として，テレビ番組の視聴率を計測しているビデオリサーチ社は，Twitter のツイートを用いて番組の視聴率を計測する試みを行なっている(注2)．現状では，テレビ局のハッシュタグのみを用いて検索を

(注1)：http://www.itmedia.co.jp/enterprise/articles/1212/20/news101.html

(注2)：http://www.itmedia.co.jp/news/articles/1210/23/news083.html

行なっており，本手法で推定されるハッシュタグ間の関係を用いることにより，テレビ局のハッシュタグだけでは取得できないツイートも取得できるようになると考えられる．本論文では，ハッシュタグの時系列的な内容の重複関係に基づくハッシュタグ間の関係を定義し，ハッシュタグ間の関係を推定する．ハッシュタグの時系列的なツイートの内容の変化を考慮し，バーストの被覆時間，およびツイートの内容に基づく類似度を用いた，ハッシュタグ間の関係推定手法を提案する．

2 章では，関連研究について記述する．3 章では，推定するハッシュタグ間のの関係について定義を行い，詳細な説明を記述する．4 章では，提案する手法について記述し，5 章では提案手法によるハッシュタグ間の関係の分類が可能かどうかを検証する．6 章では，本論文の内容についてまとめる．

2. 関 連 研 究

また，ツイートにはハッシュタグが付与されていないものもある．Zangerle ら [7] は，Twitter に投稿されたハッシュタグの付与されているツイートの集合を，入力ツイートとの $tfidf$ を用いた単語の出現情報に基づく類似度によってランク付けし，入力ツイートに対してハッシュタグを推薦する手法を提案している．

マイクロブログの記事検索に関して，ユーザの入力したクエリ拡張することで，ツイートの検索精度を向上させる研究がある．Efron [2] らは，トピックに関連するハッシュタグをランク付けしたリストを作成してフィードバックすることで，ハッシュタグのリストによるクエリ拡張に取り組んでいる．本研究では，ハッシュタグ間の同義・階層関係を用いて，リスト中のハッシュタグと組み合わせて同義・階層関係にあるハッシュタグをフィードバックすることで，更なるクエリ拡張を行うことが考えられ，ユーザのマイクロブログ記事の検索をより多様にできると考えられる．

Folksonomy のタグに関する関係の定義と推定に関して，様々な研究が行われている．丹波ら [4] は，従来のブログ記事へのソーシャルブックマークのタグについて，タグ間の関係を推定している．ソーシャルブックマークのタグは，ブログ記事がどのようなジャンルに属するのか，どのような内容を持っているのかといった情報を表す．丹羽らはタグ間の Synonym，Relevant，Unrelated，Conflicting の 4 つの関係について関係推定が行っている．これらは，タグの意味的な関係を表すものである．例えば，Synonym は 2 つのタグが同じ意味であることを表し,「スポーツ」と「sports」というタグは表記は異なるが意味的には同じであり，Synonym の関係にある．ここでのタグの関係は，そのタグの表記（単語）が示す意味による関係であり，WordNet [8] における単語間の関係と同じものであるとも考えられる．Laniado ら [9] は，Folksonomy のタグ間の意味的な関連や階層構造を WordNet の情報を用いて推定している．

本研究におけるハッシュタグ間の関係は，ハッシュタグの表記に関する意味の関係ではなく，そのハッシュタグが付与されたツイートの内容に関する関係であり，このようなタグの意味的な関係を推定するものではない．例えば，2 つのハッシュタグが意味的には同義であると考えられる場合でも，実際のハッシュタグのツイート内容は利用するユーザに依存しており，表記とは全く異なる内容のツイートが投稿されている場合がある．このような場合，ハッシュタグの表記の意味が同じでも，組み合わせて検索するとユーザが求めるトピックとは異なるツイートを取得してしまう可能性がある．そのため，本研究では，ハッシュタグの表記による意味的な関係ではなく，ハッシュタグの付与されたツイートの時系列的な内容によって，ハッシュタグ間の関係を推定する．

本手法では，ハッシュタグの時間変化に基づく特徴を用いてハッシュタグ間の関係を推定する．Vlachos ら [10] は，Web 検索に用いるクエリについて，時系列的な変化に着目し，バーストの被覆度よるクエリの類似度を求めることで，2 つのクエリの関連を抽出している．本手法においても，ハッシュタグ間の関係の推定のために，2 つのハッシュタグのバーストの被覆時間による類似度を用いる．

3. ハッシュタグ間の関係の定義

本研究では，ハッシュタグ検索の拡張を行うことを目的とし，ハッシュタグ間の関係の定義を行う．ここでの関係は，2 つのハッシュタグの時系列的な内容が同じであることに着目する必要がある．2 つのハッシュタグに関して以下の関係を定義する．

- 同義関係

2 つのハッシュタグが同義関係にある場合，それぞれのハッシュタグに含まれるツイートは，同じトピックに関する内容が記述されている．また，これらのハッシュタグは互いに，別のトピックに関する内容が記述されているツイートは含まない．図 1 は，同義関係にあるハッシュタグ A，B に関するトピックの範囲を示し，それぞれのトピックの範囲は一致する．ハッシュタグ A，B に含まれるツイートは共に，ハッシュタグ A，B に関するトピックに関するものである．例として,「# 紅白歌合戦」,「#kouhaku」は，共に NHK 総合テレビジョンの番組の 1 つである NHK 紅白歌合戦のトピックに関するハッシュタグであり，互いに同義関係にある．ユーザが興味のあるトピックに関する，あるハッシュタグを用いてハッシュタグ検索をする際に，そのハッシュタグと同義関係にあるハッシュタグを組み合わせて OR 検索を行うことで，そのトピックに関するツイートをより多く検索結果として取得できる．

- 階層関係

2 つのハッシュタグが階層関係にある場合，下位とされるハッシュタグに関するトピックは，その上位のハッシュタグに包含されるサブトピックである．また，上位のハッシュタグは，下位のハッシュタグが表すトピック以外のトピックに関するツイートも含む．図 2 は，階層関係にあるハッシュタグ A，B に関するトピックの範囲を示し，ハッシュタグ A が上位，ハッシュタグ B が下位のハッシュタグである．下位のハッシュタグ B に関するトピックの範囲は，上位のハッシュタグ A に関するトピックの範囲に包含される．すなわち，下位のハッシュタグ B に含まれるツイートは全て，上位のハッシュタグ A に含むことがで

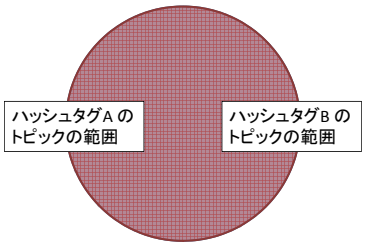

図 1　同 義 関 係

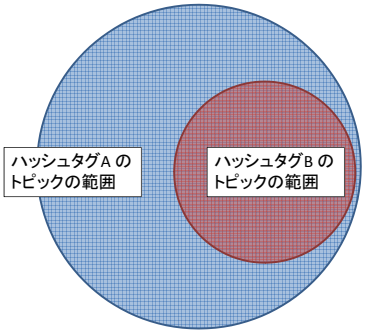

図 2　階 層 関 係

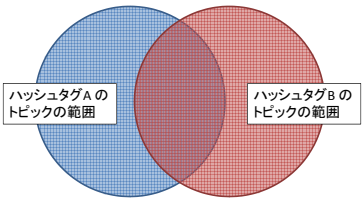

図 3　そ　の　他

きる．ハッシュタグ A，B に含まれるツイートは，重複するトピックの範囲においてはそれぞれのハッシュタグに含むことができるが，それ以外の範囲では互いに依存しない．例として，「#NHK」は NHK 総合テレビジョン,「# 紅白歌合戦」は NHK 紅白歌合戦のトピックに関するハッシュタグであり,「#NHK」が上位,「# 紅白歌合戦」が下位となる階層関係にあると考えられる.「#NHK」は,「# 紅白歌合戦」以外にも,「# 大河ドラマ」などの別のトピックに関するツイートも含む.「#NHK」の下位タグである「# 紅白歌合戦」や「# 大河ドラマ」のツイートは，上位のタグである「#NHK」を付与することができる．ユーザが興味のあるトピックに関する，あるハッシュタグを用いてハッシュタグ検索をする際に，そのハッシュタグの下位に当たるハッシュタグを組み合わせて OR 検索を行うことで，そのトピックに関するツイートをより多く検索結果として取得できる．また，上位のハッシュタグに関して，OR 検索を行うことで，そのトピックに関するツイートをより多く検索結果として取得できると考えられるが，他のトピックに関するツイートも検索結果に含まれる可能性がある．上位のハッシュタグが，同じトピックに関するツイートを含む時間をユーザが推測できる場合，検索を行う時間を限定して，上位のハッシュタグを組み合わせた OR 検索を行うことで，別のトピックのツイートを含まず，そのトピックに関するツイートを多く検索結果として取得できると考えられる．

- その他

2 つのハッシュタグが同義関係でも階層関係でもない場合，それぞれのハッシュタグの検索結果に含まれるツイートの一部は，同じトピックに関する内容が記述されている．また，それぞれのハッシュタグの検索結果は，互いに異なるトピックに関するツイートを含む．図 3 に関係がその他であるハッシュタグ A，B に関するトピックの範囲を示す．ハッシュタグ A，B に関するトピックの範囲は，互いに関連している一部のトピックにおいて重複しているが，それ以外のトピックに関しては依存しない．例として,「#NHK」と NHK 紅白歌合戦に出場する歌手に関するハッシュタグは，NHK 紅白歌合戦の番組に関して高い関連があり，それぞれ NHK 紅白歌合戦についてのツイートを検索結果に含む．また，それぞれのハッシュタグは，検索結果に別のトピックに関するツイートも含むと考えられる．本論文において，推定範囲の条件を満たしているが，推定を行う 2 つのハッシュタグ間の関係が同義・階層関係のどちらでもない場合をその他とする．

推定期間における 2 つのハッシュタグ間について，上記の関係を推定する．同義・階層関係にあるハッシュタグはある時間において同一のトピックに関するツイートを含んでおり，この時間においてこれらのハッシュタグを組み合わせて検索を行うことで，これらのハッシュタグが示すトピックに関するツイートを，1 つのハッシュタグで検索する場合と比較して，多く取得できると考えられる．

4. 提 案 手 法

ハッシュタグのタイムラインの時系列的な変化に着目し，ハッシュタグ間の関係を推定する．2 つのハッシュタグのタイムラインのある期間において，それぞれのタイムラインに対してバーストを検出し，それらのハッシュタグのバーストの時間情報と，その内容を用いたハッシュタグ間の同義・階層関係の推定手法を提案する．

ハッシュタグ間の関連を推定する手法において，各ハッシュタグのバーストする時間を求め，その類似度によってハッシュタグの関係を推定する．バーストの解析手法として，蛯名らが提案したバースト検出手法 [11] を用いる．はじめに，ノイズとなるツイートを除去するため，本手法を適用するツイートからリツイート，リプライ，メンション，および URL を含むツイートを除外する．

ハッシュタグのバースト時間に基づいて，ハッシュタグの同義・階層関係についての仮説を立てる．2 つのハッシュタグが同義関係にある場合，それらのハッシュタグは，同じトピックに関するツイートが同じような時間に投稿されると考えられる．そのため，それぞれのハッシュタグの検索結果は，同じような時間にバーストすると考えられる．図 4 にハッシュタグ「# 紅白歌合戦」と「#kouhaku」のバーストの例を示す.「# 紅白歌合戦」と「#kouhaku」は，共に NHK 紅白歌合戦のトピックに関するハッシュタグであり，番組中で歌手の登場などにユーザが反応し，多くのユーザがツイートを投稿するため，同じような時間にバーストが発生する.「# 紅白歌合戦」と「#kouhaku」のバースト解析結果は，ゴールデンボンバーの登場に関して,「# 紅白歌合戦」のバースト時間は 2012 年 12 月 31 日の 19:31:08〜19:37:38,「#kouhaku」のバースト時間は

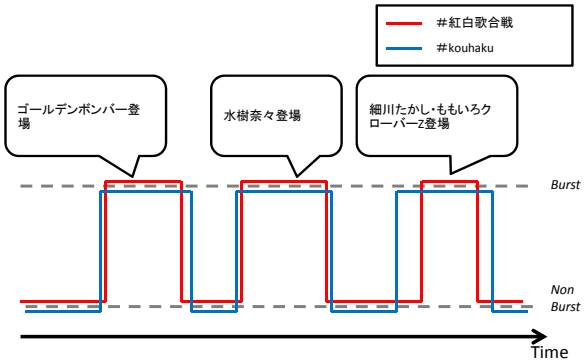

図 4　同義関係のバーストの例

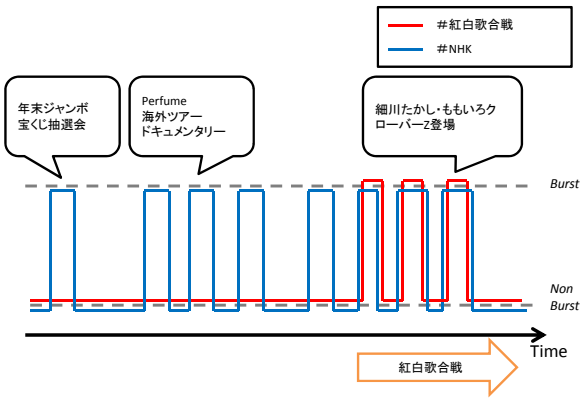

図 5　階層関係のバーストの例

2012 年 12 月 31 日の 19:31:03〜19:38:10，水樹奈々の登場に関して，「# 紅白歌合戦」のバースト時間は 2012 年 12 月 31 日の 19:45:23〜19:51:59，「#kouhaku」のバースト時間は 2012 年 12 月 31 日の 19:45:06〜19:52:48，細川たかしとももいろクローバー Z の登場に関して，「# 紅白歌合戦」のバースト時間は 2012 年 12 月 31 日の 20:18:40〜20:18:56，「#kouhaku」のバースト時間は 2012 年 12 月 31 日の 20:08:37〜20:22:49 である．これらのバースト時間は，バースト開始時間とバースト終了時間においてほとんど差がない．そのため，本手法では，バーストの開始時刻と終了時刻に基づいて，2 つのハッシュタグが同義関係である場合，推定可能であると考えられる．

2 つのハッシュタグが階層関係にある場合，下位のハッシュタグに関するトピックは，上位のハッシュタグに関するトピックに包含されるため，下位のハッシュタグに関するトピックのツイートが投稿される時間においては，上位のハッシュタグも同様のトピックのツイートが投稿されると考えられる．そのため，下位のハッシュタグに関するトピックが注目され，下位のハッシュタグがバーストする時間は，上位のハッシュタグもバーストすると考えられる．一方で，上位のハッシュタグは下位のハッシュタグに関するトピック以外のツイートも検索結果に含むため，下位のハッシュタグがバーストしていない時間についても，バーストする可能性がある．結果として，2 つのハッシュタグの検索結果の全体のバースト時間の被覆している時間の割合は低くなると考えられる．図 5 にハッシュタグ「#NHK」と「# 紅白歌合戦」のバーストの例を示す．下位のハッシュタグである「# 紅白歌合戦」に関するトピックである NHK 紅白歌合戦に関するツイートは，上位のハッシュタグである「#NHK」も含んでいる．どちらのハッシュタグも，NHK 紅白歌合戦の放送中は歌手の登場などにユーザが反応しているためツイートが集中し，同じような時間にバーストが発生すると考えられる．このような場合，これらのバースト時間の類似度は高くなると考えられる．一方で，「#NHK」は年末ジャンボ宝くじの抽選会や，Perfume 海外ツアードキュメンタリーなどの番組に関するツイートを含むため，これらの放送時間中もバーストしている．

このような仮説を用いて，ハッシュタグのバースト時間の類似度を求め，ハッシュタグ間の関係を推定する．

4.1　バーストの被覆時間に基づくハッシュタグ間の類似度

ハッシュタグ A，B の各バーストの集合を $Burst_A = \{b_{A1}, \cdots, b_{Ak}\}$，$Burst_B = \{b_{B1}, \cdots, b_{Bm}\}$ と表す．b_{Xy} は，開始時間と終了時間を持つ 1 つのバーストの要素を表し，ハッシュタグ X の y 番目のバーストの要素を表す．このとき，$|Burst_{(X)}|$ をハッシュタグ X のバースト時間の総和を表すものとする．2 つのハッシュタグ A，B のそれぞれのバースト時間に関して，バースト時間が被覆する割合が大きいほどハッシュタグ A，B のバースト時間の類似度が高いと考える．ハッシュタグ A の i 番目のバーストの要素とハッシュタグ B の j 番目のバーストの要素について，被覆する時間を，$overlap(b_{Ai}, b_{Bj})$ と表す．このとき，ハッシュタグ A，B のバーストの集合 $Burst_A$，$Burst_B$ のすべての要素の被覆する時間の総和 $|Burst_{A\cap B}|$ は，以下の式 (1) で表される．

$$|Burst_{A\cap B}| = \sum_{i=1}^{k}\sum_{j=1}^{m} overlap(b_{Ai}, b_{Bj}) \tag{1}$$

式 (1) を用いて，ハッシュタグ A，B バースト時間の類似度 $BSim\_divA$，および $BSim\_divB$ を，以下の式 (2)，(3) で表す．

$$BSim\_divA = \frac{|Burst_{A\cap B}|}{|Burst_A|} \tag{2}$$

$$BSim\_divB = \frac{|Burst_{A\cap B}|}{|Burst_B|} \tag{3}$$

$BSim\_divA$ はハッシュタグ A の総バースト時間に対するバーストの総被覆時間の割合，$BSim\_divB$ はハッシュタグ B の総バースト時間に対するバーストの総被覆時間の割合を表す．

ハッシュタグ間の関係に関する上記の仮説により，$BSim\_divA$，および $BSim\_divB$ について以下のことが成り立つと考えられる．

- 同義関係

ハッシュタグ A，B が同義関係にある場合，全てのバースト時間において高い類似度を示すと考えられる．そのため，バースト類似度 $BSim\_divA$，および $BSim\_divB$ は，共に高い値となると考えられる．

- 階層関係

ハッシュタグ A，B が階層関係にある場合，下位のハッシュタグがバーストしている時間は上位のハッシュタグもバーストしていると考えられる．しかし，下位のハッシュタグがバーストし

ていない時間において，上位のハッシュタグもバーストしている可能性がある．そのため，$BSim\_divA$，および $BSim\_divB$ に関して，ハッシュタグ A が上位，ハッシュタグ B が下位である場合，$BSim\_divA$ は低い値を，$BSim\_divB$ は高い値を示すと考えられる．ハッシュタグ A が下位，ハッシュタグ B が上位である場合はその逆である．

求めたバースト類似度 $BSim\_divA$，$BSim\_divB$ を用いて，ハッシュタグ間の関係を同義関係，階層関係，その他に分類する．

4.2 トピック類似度に基づくバーストの被覆時間の選択

4.1 の手法によって 2 つのハッシュタグの関係を推定した場合，それらのトピックの内容が異なるが，バーストの被覆時間による類似度が高くなる可能性が考えられる．これは，2 つのハッシュタグが，それぞれ別の要因によって同時刻にバーストした際に発生する．しかし，本研究は，時系列的な内容が同じハッシュタグの組み合わせを発見することが目的であるため，2 つのハッシュタグのトピックが同一である場合のみ，2 つのハッシュタグが類似していると考える．そのため，$overlap(b_{Ai}, b_{Bj})$ について，ツイートの内容を考慮したトピック類似度 $TopicSim(b_{Ai}, b_{Bj})$ を用いて，内容が類似している場合のみ，バーストが被覆したものとする．

トピック類似度 $TopicSim(b_{Ai}, b_{Bj})$ は，バーストの要素 b_{Ai}，b_{Bj} の開始時間から終了時間の間に存在するハッシュタグ A，B の付与されたツイートに含まれる，単語の出現割合を特徴量としたコサイン類似度によって求める．ツイートから単語を抽出するために形態素解析器である Kuromoji [12] を用いる．また，形態素解析辞書として UniDic [13] を用いる．各ツイートを形態素解析した結果の単語群から内容語 (名詞・動詞・形容詞) を抽出して用いる．また，形態素解析によって，1 つの熟語が複数の単語として抽出される場合があるため，名詞の連結を行った．名詞の連結ルールとして，前の形態素が「普通名詞」であり,「副詞可能」でない，かつ「形状詞可能」でないとき，後ろの形態素が「普通名詞」であり,「副詞可能」でない，かつ「形状詞可能」でない，かつ「サ変可能」でない場合に連結を行った．形状詞は，形容動詞の語幹を表す．

2 つのバーストの要素 b_{Ai}，b_{Bj} の開始時間から終了時間の間に存在するハッシュタグ A，B の付与されたツイートに含まれる，単語の集合を $W = \{w_1, w_2, \cdots, w_N\}$ とする．また，b_{Ai} の開始時間から終了時間の間に存在する，ハッシュタグ A の付与されたツイートに含まれる単語の集合を W_{Ai} と表す．ただし，閾値 γ を用いて b_{Ai} において出現数が γ 未満の単語は W_{Ai} から取り除いた．このとき，b_{Ai} の特徴量 $\vec{b_{Ai}} = \{t_1^{Ai}, t_2^{Ai}, \cdots, t_N^{Ai}\}$ を，以下の式 (4) によって求める．

$$t_n^{Ai} = \begin{cases} \frac{num_{Ai}(w_n)}{|D_{Ai}|} & (w_n \in W_{Ai}) \\ 0 & (otherwise) \end{cases} \quad (4)$$

ここで，$num_{Ai}(w_n)$ は b_{Ai} における単語 w_n の出現数，$|D_{Ai}|$ は b_{Ai} における単語の出現数の総数である．b_{Bi} の特徴量 $\vec{b_{Bj}}$ についても同様に求める．

バーストの要素 b_{Ai}，b_{Bj} のトピック類似度 $TopicSim(b_{Ai}, b_{Bj})$ を，それぞれの特徴量 $\vec{b_{Ai}}$，$\vec{b_{Bj}}$ を用いて，コサイン類似度に基づく以下の式 (5) によって求める．

$$TopicSim(b_{Ai}, b_{Bj}) = \frac{\vec{b_{Ai}} \cdot \vec{b_{Bj}}}{|\vec{b_{Ai}}||\vec{b_{Bj}}|} \quad (5)$$

求めたトピック類似度 $TopicSim(b_{Ai}, b_{Bj})$ と閾値 α によって，$TopicSim(b_{Ai}, b_{Bj}) < \alpha$ である場合は，$overlap(b_{Ai}, b_{Bj}) = 0$ とする．これにより，バースト時間が被覆していても，内容は異なる場合は，バースト時間が被覆していないものとして扱う．

5. 検　　証

本論文で提案したバースト類似度 $BSim\_divA$，$BSim\_divB$ によるハッシュタグ間の関係推定手法により，ハッシュタグ間の関係の同義，階層関係の分類が可能であるか，2 つの類似度による散布図を用いて検証を行う．また，2 つのハッシュタグ間の関係の時間的な変化についても検証する．

利用するツイートデータの収集には，Twitter Search API [14] を用いた．ツイートデータは，言語設定が日本語（ja）であるユーザのツイートに限定した．はじめに，2012 年 12 月 31 日の 9 時から 2013 年 1 月 1 日の 9 時までの期間について，# 紅白を含むツイートのデータを取得した．次に，# 紅白について取得したツイートにおいて共起しているハッシュタグについてもツイートのデータを取得した．# 紅白と共起するハッシュタグの組み合わせについて，2 つのハッシュタグ間の関係を推定する．

ハッシュタグ間の関係の推定期間は (1)1 日の期間 (2012 年 12 月 31 日の 9 時から 2013 年 1 月 1 日の 9 時まで)，(2) 紅白歌合戦の放送時間 (2012 年 12 月 31 日の 19 時 15 分から 23 時 45 分) の 2 つとして検証を行う．データセットのハッシュタグの組合せについて，それぞれ推定期間におけるバースト類似度 $BSim\_divA$，$BSim\_divB$ を求め，$BSim\_divA$ を縦軸，$BSim\_divB$ を横軸とする散布図を作成し，ハッシュタグ間の関係の分布を検証する．データセットのハッシュタグの組合せについて，ハッシュタグ間の関係をそれぞれの期間について人手で付与した．この際，2 つのハッシュタグが時間的に同じトピックを含むいるかどうかに基づいてハッシュタグ間の関係を付与した．

バースト検出の際のパラメータは，$N = 60$，$\beta = 0.90$，$W_{min} = 15000$(ミリ秒)，$Amin = 3$ とした．また，トピック類似度 $TopicSim(b_{Ai}, b_{Bj})$ に関する閾値 $\alpha = 0.30$，$\gamma = 3$ とした．

5.1 1 日の推定期間における検証

図 6 に，1 日の期間 (2012 年 12 月 31 日の 9 時から 2013 年 1 月 1 日の 9 時まで) を推定期間とする，バースト類似度 $BSim\_divA$，$BSim\_divB$ によるハッシュタグ間の関係の分布を示す．ここでは，ハッシュタグ A を # 紅白，ハッシュタグ B を # 紅白と共起するハッシュタグとし，# 紅白に対して共起するハッシュタグの分布を示している．4 章で示した $BSim\_divA$，$BSim\_divB$ に関する仮説が正しいならば，同義関係にあるハッシュタグは図の右上，上位にあるハッシュタグは図の右下，

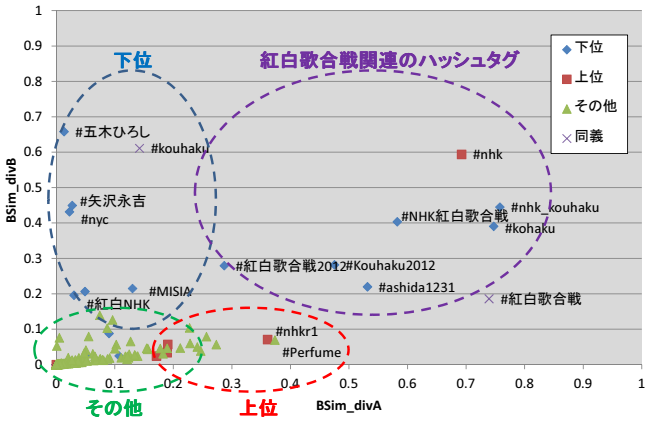

図 6　1 日の推定期間におけるハッシュタグ間の関係の分布

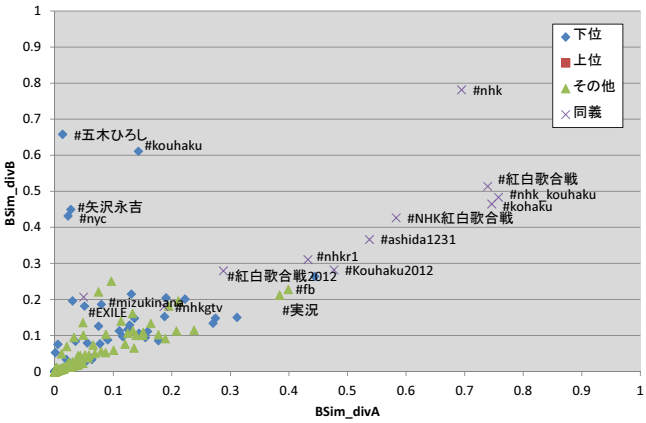

図 7　紅白歌合戦放送時間の推定期間におけるハッシュタグ間の関係の分布

下位にあるハッシュタグは図の左上，その他の関係にあるハッシュタグは図の左下に分布すると考えられる．図 6 では，その他の関係が図の左下，上位の関係が図の右下，下位の関係が図の左上と右上におよそ分布している．ここで，下位の関係の右上に分布しているハッシュタグは，#NHK 紅白歌合戦，#nhk_kouhaku，#kohaku，#Kouhaku2012 などの紅白歌合戦に関するハッシュタグである．また，#ashida1231 も紅白歌合戦に関するハッシュタグである．これらのハッシュタグについて，ハッシュタグの関係性は下位であると分類されているが，これは # 紅白が紅白歌合戦の開始時間前である 18 時頃から，紅白歌合戦の開始を待つユーザのツイートが出現しており，これらの共起ハッシュタグはこの内容を含まなかったため，人手による関係の付与で下位とされたものである．図 6 より，主に紅白歌合戦のトピックに関するハッシュタグは同義関係に近い分布を示している．

上位関係と判断されたハッシュタグについて，#nhkr1 は NHK 第一ラジオに関するハッシュタグであり，紅白歌合戦の放送時間中は紅白歌合戦についての放送を行なっており，ツイートも紅白歌合戦に関するものであるが，それ以外の時間は他の番組についてツイートが投稿されている．そのため，#nhkr1 は，上位のハッシュタグであると考えられ，分布から正しく関係が推定されたと考えられる．一方，同様に上位と考えられる #nhk が同義に近い分布を示している．これは，#nhk の中で特に注目された期間が紅白歌合戦の放送時間であり，紅白歌合戦の放送時間におけるバースト時間が，#nhk の総バースト時間の中でも大きな割合を占めていたことが原因だと考えられる．

また，紅白歌合戦に参加している歌手のハッシュタグについて，主に紅白歌合戦の時間中のみツイートが投稿されているハッシュタグは下位，それ以外の時間もツイートが投稿されているハッシュタグはその他と判断されている．例えば，# 五木ひろし，# 矢沢永吉などは，主に紅白歌合戦放送時間中のみツイートが投稿されているハッシュタグであり，下位関係の分布を示している．一方，#EXILE，#mizukinana などは，紅白歌合戦放送時間以外の時間もツイートが投稿されているため人手による関係の付与ではその他と判断されており，#EXILE が $BSim_divA = 0.023$，$BSim_divB = 0.051$，#mizukinana が $BSim_divA = 0.014$，$BSim_divB = 0.075$ とその他の分布を示している．

このように，バースト類似度 $BSim\_divA$，$BSim\_divB$ に基づいて，ハッシュタグ間の関係は特徴的な分布を示しており，$BSim\_divA$，$BSim\_divB$ によってハッシュタグ間の関係の分類が可能であると考えられる．

一方で，提案手法は，バーストの類似度に基づく手法であるため，ツイート数が少ない場合などにおいてバーストが検出されず，ハッシュタグ間の関係をうまく推定できない場合がある．#kouhaku は，# 紅白に対して同義関係であると人手で判断されたハッシュタグである．#kouhaku はツイート数が少ないため，# 紅白ではバーストが検出されるような期間に対して，特にその期間が少数のユーザが注目するような場合にツイートがあまり投稿されず，バーストが検出されない場合がある．そのため，ツイート数が少ない #kouhaku は，# 紅白と比較してバーストが検出される期間が少なくなり，結果として下位の分布を示している．また，紅白歌合戦の参加歌手に関するハッシュタグの中で，ツイート数が極端に少ないハッシュタグのいくつかは全くバーストが検出されず，$BSim_divA = 0$，$BSim_divB = 0$ に分布している．ハッシュタグのツイート数が少ない場合は，バースト検出手法が適用できないため，今後このようなハッシュタグに対して，バースト検出に依らない手法を検討する必要がある．

5.2　紅白歌合戦放送時間の推定期間における検証

図 7 に，紅白歌合戦の放送時間 (2012 年 12 月 31 日の 19 時 15 分から 23 時 45 分) を推定期間とする，バースト類似度 $BSim\_divA$，$BSim\_divB$ によるハッシュタグ間の関係の分布を示す．ここでは，推定期間の変化によるハッシュタグ間の関係の変化について検証する．# 紅白に対して，#nhk や #nhkr1 は 1 日の推定期間では上位関係であると考えられるが，推定期間を紅白歌合戦放送時間に限る場合では，これらのハッシュタグは主に紅白歌合戦のトピックに関するツイートを持つため，同義関係であると考えられる．図 7 では，#nhk と #nhkr1 は同義関係のハッシュタグと人手で判断されており，図 6 と比較して，#nhkr1 は同義関係に近い分布に変化している．また，図

7 より，推定期間を紅白歌合戦放送時間に限る場合では，# 紅白の上位関係であると人手で判断されたハッシュタグは存在せず，上位関係の分布を示すハッシュタグも存在しないことが確認できる．

紅白歌合戦に参加している歌手のハッシュタグについて，主に紅白歌合戦の時間中のみツイートが投稿されている # 五木ひろし，# 矢沢永吉などは，推定期間を紅白歌合戦放送時間に限る場合でも，推定期間を 1 日とした場合と同様に下位関係にあると考えられ，図 7 でも下位関係の分布を示している．また，紅白歌合戦放送時間以外の時間もツイートが投稿されている #EXILE，#mizukinana などは，推定期間が 1 日である場合はその他の関係であったが，推定期間を紅白歌合戦放送時間に限る場合では，紅白歌合戦のトピックに包含されるため下位関係に変化すると考えられる．図 7 では，#EXILE と #mizukinana は下位関係のハッシュタグと人手で判断されており，図 6 と比較して，これらのハッシュタグは下位関係に近い分布に変化している．

また，#fb，# 実況はともに紅白歌合戦放送時間に関して，紅白歌合戦のトピックに関するツイートと，日本テレビの番組であるが「ダウンタウンのガキの使いやあらへんで!!」などの紅白歌合戦以外のトピックに関するツイートの両方を含むハッシュタグである．その他のトピックに関するツイートを含むため，人手でその他の関係であると判断されていが，紅白歌合戦放送時間中は紅白歌合戦のトピックに関するツイートが多く出現していたため，図 7 では同義関係に近い分布を示している．このような同じ時間帯に複数のトピックに関するツイートを含むようなハッシュタグに関する対処として，トピック類似度の閾値を高くすることで同義関係の分布から排除することが考えられる．あるいは，ハッシュタグのツイートを，トピックごとのグループに分けることで対処が可能であると考えられる．

6. ま と め

本論文では，ユーザがマイクロブログにおいて，興味のあるトピックを検索する際に，同一のトピックに関するハッシュタグが複数存在する場合に，ひとつのハッシュタグによる検索では，これらの求めるトピックに関連するツイートが得られないという課題に取り組んだ．この課題について，ハッシュタグのトピックの重複関係に基づいて，ハッシュタグ間の関係を推定する手法を提案した．ハッシュタグ間の時間経過による関係の変化に対応するため，時系列的なトピックの重複によるハッシュタグ間の関係を定義した．これらのハッシュタグ間の関係を推定するために，バーストの被覆時間とトピックの類似度を組合わせて利用する手法を提案した．

今後の課題として，提案したバースト類似度に基づく分類手法の適用が挙げられる．ハッシュタグのツイート数が少ない場合にバースト検出手法が適用できないため，このようなハッシュタグに対してバースト検出に依らない手法を検討する必要がある．また，同じ時間帯に複数のトピックに関するツイートを含むハッシュタグに関して，このようなハッシュタグは，同義関係，階層関係であると推定される可能性があるが，ユーザの求めるトピック以外のツイートも含むため，望ましいとは限らない．このようなハッシュタグに対して，排除を行う，あるいはツイートをトピックごとに分けることで，ユーザの求めるトピックに関するツイートのみを取得することが考えられる．また，本研究の目標はユーザが求めるトピックに関するツイートをひとつのハッシュタグから自動的に取得することである．そのため，本手法により推定されるハッシュタグ間の関係を用いてツイート検索を自動的に拡張するシステムの実装を行う．

文　献

[1] Twitter. https://twitter.com/.

[2] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 787–788. ACM, 2010.

[3] J. Teevan, D. Ramage, and M.R. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 35–44. ACM, 2011.

[4] 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 (第 14 回), SIG-SWO-A602, 2006.

[5] L. Potts, J. Seitzinger, D. Jones, and A. Harrison. Tweeting disaster: hashtag constructions and collisions. In *Proceedings of the 29th ACM international conference on Design of communication*, pp. 235–240. ACM, 2011.

[6] S.A. Golder and B.A. Huberman. The structure of collaborative tagging systems. 2009.

[7] E. Zangerle, W. Gassler, and G. Specht. Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, Vol. 730, pp. 67–78, 2011.

[8] Wordnet. http://wordnet.princeton.edu/.

[9] D. Laniado, D. Eynard, M. Colombetti, et al. Using wordnet to turn a folksonomy into a hierarchy of concepts. In *Semantic web application and perspectives-fourth italian semantic web workshop*, pp. 192–201. Citeseer, 2007.

[10] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pp. 131–142. ACM, 2004.

[11] 蝦名亮平, 中村健二, 小柳滋 ebina2010. リアルタイムバースト検出手法の提案. 日本データベース学会論文誌, Vol. 9, No. 2, pp. 1–6, 2010.

[12] Kuromoji. http://www.atilika.org/.

[13] Unidic. http://www.tokuteicorpus.jp/dist/.

[14] Twitter search api. https://twitter.com/search.