日本データベース学会和文論文誌
Vol. 18-J, Article No. 12, 2020年 3月

トレンドエンティティに関するWebメディア間横断調査

中村達哉[1]　藤田澄男[2]　原隆浩[3]

マイクロブログやQ&Aサイトといったメディア上のテキストを解析し，実世界で注目されている事象を獲得するトレンド分析に関する研究が盛んに行われている．実世界におけるトレンドは複数のWebメディアを横断して波及すると考えられるが，異なるメディア間でトレンドに対する注目の観点のようなトレンドの文脈（コンテキスト）が同一であるとは限らない．しかし，既存の複数メディアを対象とした横断的なトレンド分析に関する研究では，トレンドに関連するキーワードの出現頻度のようなトレンドの量的な観点にのみ着目しており，メディア間で共通のトレンドの文脈の差異に基づく分析は行われてこなかった．本論文では，マイクロブログ，Q&Aサイト，Web検索サイトの三つのWebメディアについて，各メディアで注目を集めているトレンドエンティティを対象としてメディア間のトレンドの文脈の差異を調査し，その調査結果からメディアを横断したトレンド分析におけるトレンドの文脈の差異を考慮した分析の必要性について議論する．

1　はじめに

Web上のテキストを解析し，注目を集めている人や組織，出来事といったエンティティに関する実世界のトレンドを分析する研究が数多く行われている．既存研究の多くでは，分析の対象となるWebメディア*1上のトレンドキーワード，すなわち，ある期間において集中的に出現する語句を対象とすることで注目されているエンティティを捉え，実世界のトレンドの分析を行っている．例えば，Web検索サイトで頻繁に検索される語句やマイクロブログにおいて多くのユーザにより言及される語句は，ユーザが興味や関心を持つエンティティを表している可能性が高いと考えられ，この仮説に基づいたトレンド情報の分析や話題抽出などへの応用が行われている [2, 3, 10].

実世界においてあるエンティティに関するトレンドが発生した場合，そのトレンドが複数のWebメディアにおいて同時に観測される場合がある．このような異なるWebメディア間のトレンドを比較分析する研究 [8, 13]や共通するトレンドを抽出する研究 [5]が行われている．これらの研究では，Webメディア上のトレンドキーワードについてメディア間の共通点や相違点をもとに，各メディアで観測されるトレンドの特徴の理解やメディア間の共通トレンドの抽出を試みている．しかし既存研究では，メディア間であるエンティティがどれくらいの度合で共通して注目されるかといったトレンドの量的な観点のみが着目されており，複数のメディアで共通のトレンドであってもトレンドに対する注目の観点や他のトレンドとの関連性がメディア間で異なるといったトレンドの文脈（コンテキスト）の観点については研究の対象とされてこなかった．例えばある企業の新製品について，あるメディアではその新製品についてのみ注目されている一方，別のメディアでは競合する他社製品についても同時に注目されている場合，その新製品は両方のメディア間で共通して注目されているがその文脈は異なるトレンドであると考えられる．このような注目を集めトレンドとなっているエンティティに対するユーザの注目の観点のようなトレンドの文脈の差異をメディア間で明らかにすることにより，実世界のトレンドに対してメディアの特徴やメディア間の関係に応じたより詳細な分析が可能になると考えられる．また，エンティティを対象とした調査を行うことで，“企業Xが企業Yを買収した”のような複数のエンティティが関与する実世界の話題や，エンティティ間の新しい関係の抽出への応用に貢献できると考えられる．

そこで本研究では，Webメディア上のトレンドエンティティについてメディアを横断した調査を行い，各メディアにおけるトレンドエンティティ間の関連性からメディア間のトレンドの文脈の差異を抽出することを目指す．具体的には，エンティティリンキングと単語に対する分散表現を用いて，各メディアにおけるトレンドエンティティの抽出とエンティティのベクトル化を行う．あるトレンドエンティティについて，異なる時点間あるいは異なるメディア間でベクトルを比較することにより，メディア内におけるトレンドの文脈の経時変化やメディア間のトレンドの文脈の差異を調査する．本論文では，マイクロブログ，Q&Aサイト，Web検索サイトの三つのWebメディアを対象として，各メディアにおけるトレンドエンティティの量的な基礎調査に加えて，以下の観点によるトレンドの文脈に関する調査を行い，メディアを横断したトレンド分析におけるトレンドの文脈を考慮した分析の必要性について議論する．

1. 分散表現に基づくトレンドの継続性：それぞれのメディアにおいて，異なる時点におけるトレンドエンティティのベクトルを比較することにより，トレンド発生日からのトレンドの文脈の経時変化を調査する．この調査により，メディア間で共通して出現するトレンドエンティティであっても，そのトレンドエンティティに対するトレンドの文脈の経時変化の傾向がメディア間で異なることを示す．
2. メディア間のトレンドの文脈の差異：メディア間で共通して出現するトレンドエンティティについて，メディア間でベクトルを比較しその類似性からメディア間のトレンドの文脈の差異を調査する．この調査により，複数のメディアで共通して出現するトレンドエンティティであっても，そのトレンドの文脈にはばらつきがあることを示す．

[1] 非会員　ヤフー株式会社
tatnakam@yahoo-corp.jp
[2] 非会員　ヤフー株式会社
sufujita@yahoo-corp.jp
[3] 非会員　大阪大学大学院情報科学研究科
hara@ist.osaka-u.ac.jp

*1 本論文では，Web上の各種サービスが日々提供するコンテンツを，サービス毎に総称してWebメディアと呼ぶ．

3. Wikipedia[*2]の意味情報との比較：それぞれのメディアにおいて出現するトレンドエンティティについて，他のエンティティとのベクトル間類似度とWikipediaの意味情報に基づくエンティティ間関連度を比較し，メディア間のトレンドの文脈の差異をWikipediaの意味情報に基づく観点から調査する．この調査により，複数のメディアに共通して出現しやすいトレンドエンティティであるほど，他のエンティティとの関連性がWikipediaの意味情報に基づくエンティティ間の関連性と類似した傾向にあることを示す．

2 関連研究

Webメディアから実世界のトレンドや話題を抽出する研究が盛んに行われている [3]．これらの研究の成果をもとに，最近では，複数のWebメディア間で共通するトレンドの抽出や，複数のWebメディアを横断したトレンドの分析に関する研究が行われている．

複数のWebメディアからメディア間で共通するトレンドを抽出するタスクは，Topic Detection and Tracking (TDT)に関する分野で研究が行われている． Wangら [9]は，実世界で何らかのイベントが発生した際に，そのイベントに関する情報が複数のメディアで発信される現象に着目し，複数のテキストストリームを入力として，ストリーム間で注目度の高いトピックを抽出する手法を提案している． Hongら [5]は，複数のテキストストリームから，ある時点においてストリーム間で共通するトピックに加えて，あるストリームに特有なトピックを抽出するトピックモデルを提案している．これらの研究は，複数のメディア間に共通する，あるいは，特定のメディアにのみ出現するトレンドといったメディア間のトレンドの量的な観点に基づく手法を提案しており，本研究で対象とするメディア間で共通して注目されていても注目の観点が異なるといった文脈が異なるトレンドは抽出できない．本研究では，複数のメディアを横断した調査により，メディア間で共通するトレンドであってもそのトレンドの文脈が異なる傾向にあることを示し，これらのメディア間を横断したトレンド分析においてトレンドの文脈を考慮する必要性について議論する．

吉田と荒瀬 [13]は，Web検索サイトが提供するトレンドキーワード（語句の検索頻度データ）について，Wikipediaやオンライン辞書サービスなどのWebリソースにおけるトレンドキーワードの登録状況やトレンドの継続期間，各リソースのトレンドに対する反応の速さについて横断的な分析を行い，各リソースに出現するトレンドキーワードの類似性を明らかにした．この研究では，Webリソース間のトレンドキーワードの類似性というトレンドの量的な観点に基づく調査を行っており，Webリソース間で共通するトレンドキーワードについて他のキーワードとの関連性のようなトレンドキーワードの文脈の観点に基づく調査は行われていない．

Bandariら [2]は，ニュースの人気度の予測を目的として，ニュース記事の内容とマイクロブログTwitter[*3]においてニュース記事が参照される頻度の関係について調査し，Twitterで参照されやすいニュースの特徴を明らかにした．また，調査結果に基づいてニュースの人気度を予測する手法を提案している．Yoshidaら [12]は，キーワードの検索頻度の予測を目的として，Web検索サイトにおけるキーワードの検索頻度とそのキーワードに対応するWikipediaの記事の閲覧回数について調査を行い，高頻度に検索されるキーワードについてはWikipediaの記事の閲覧回数を用いて検索頻度を予測できる可能性を示した．これらの研究では，異なるWebメディア間のトレンドの量的な観点の類似性に基づいてあるメディアのトレンドの予測を実現している．メディア間で共通のトレンドであっても，その文脈が異なるトレンドは各メディアで特有な特徴を持って発展していくと考えられるため，本研究により得られたトレンドの文脈に関する知見を用いることで，トレンド予測などのタスクにおける有用な特徴の一つとしての利用が期待できる．

3 データの前処理

3.1 調査対象のWebメディア

本研究で調査の対象とするトレンドエンティティは，以下に述べる三つのWebメディアから抽出する．

- **Twitter**：代表的なマイクロブログサービスの一つであり，ユーザは自身の興味関心をツイートと呼ばれる140文字以内の短いテキストの形式で投稿できる[*4]．他のメディアと比較して，実世界の話題に関連したエンティティが得られやすいと考えられる．本調査ではTwitter Streaming API[*5]で収集した日本語のツイートを用いた．
- **Yahoo!知恵袋**[*6]：日本国内最大級のQ&Aサイトの一つであり，ユーザは自身が持つ疑問について質問したり，他のユーザの質問に対して回答できる．疑問の解消を目的として利用されるため，実世界の話題からユーザの生活まで幅広いジャンルにおいて疑問を持たれているエンティティやWeb上で情報が不足しているエンティティが得られると考えられる．本調査では，Yahoo!知恵袋に投稿された質問文を用いた．
- **Yahoo!検索**[*7]：日本国内最大級のWeb検索サイトの一つである．実世界の話題からユーザの生活まで幅広いジャンルのエンティティが得られると考えられる．本調査では，Yahoo!検索のウェブ検索ログデータを用いた．

本調査では，上記の三つのWebメディアついて2016年10月1日から2017年9月31日まで収集したデータを用いた．

3.2 前処理の手順

本節ではWebメディア間のトレンドの文脈の差異を調査するための前処理について説明する．まず，エンティティリンキングにより各Webメディアのテキストに対応するWikipediaの記事を紐

[*2] https://www.wikipedia.org/
[*3] https://twitter.com/
[*4] 2017年11月7日より中国語・日本語・韓国語以外の言語では1ツイートあたり280文字まで利用可能である．
[*5] https://developer.twitter.com/en/docs/tweets/sample-realtime/overview
[*6] https://chiebukuro.yahoo.co.jp/
[*7] https://search.yahoo.co.jp/

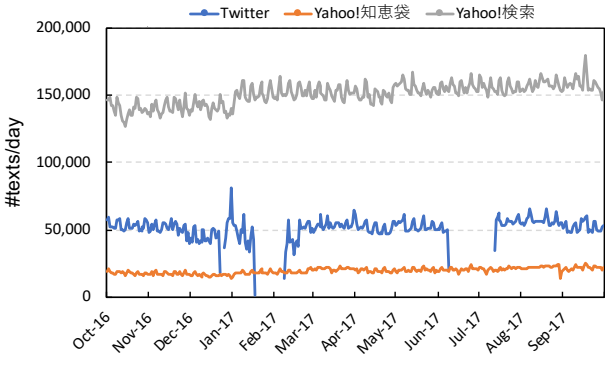


図1　各Webメディアの日別テキスト数

付ける．その後，各Webメディアにおけるトレンドエンティティの抽出と分散表現を用いたエンティティのベクトル化を行う．以下では，前処理の各手順について説明する．

3.2.1　データの加工

はじめに，Yahoo!検索のログデータについて加工を行う．Yahoo!検索のログデータは“(匿名化されたユーザID,時刻,検索キーワード)”のタプルにより与えられるため，そのままではユーザの一連の検索行動に関する情報を扱うことができない．そこで本調査では30分を1セッションとして，同一ユーザによるセッション中の連続した検索で利用されたキーワードを連結し，セッションの開始時刻に投稿された一つのテキストと見なした．その後，Twitterのツイート，Yahoo!知恵袋の質問文，Yahoo!検索の加工済みログに対してMeCabによる単語分割を行った[*8]．

3.2.2　エンティティリンキング

前処理を行った各Webメディアのテキストに対してTAGME [4]を用いたエンティティリンキングを適用する． TAGMEは，ツイートのような短いテキストを対象として，テキスト中にキーワードに対応するWikipediaの記事を紐付ける手法である．本調査では，2018年8月1日に公開された日本語版Wikipediaのダンプデータを用いてTAGMEを実装した．手法に関わるパラメータについては，TAGME [4]において短文のデータセットでの評価で用いられていた値を使用した． TAGMEにより，例えば，分かち書き済みのテキスト「アップル が 新しい iPhone を 発売」に対して，テキスト中のキーワード「アップル」と「iPhone」に対応するWikipediaの記事（エンティティ）が紐付けられ，テキスト「[[アップル_(企業)]] が新しい [[IPhone_XS]] を発売」が出力される．

図1にエンティティリンキングまでの処理を適用し，少なくとも一つのエンティティが紐付けられた各Webメディアの日別テキスト数を示す．なお，Twitterに関しては収集システムの問題によりツイートが収集できていない期間が存在する[*9]．そのため，調査対象の全Webメディアついて，Twitterの欠損期間とその前後7日については本調査の対象外とした．

[*8] MeCabの辞書にはipadic–2.7.0を用いた．
[*9] 次の三つの欠損期間が存在する．(1) 2016年12月24日～同月25日，(2) 2017年1月18日～同年2月7日，(3) 2017年6月10日～同年7月11日

表1　各Webメディアから抽出したトレンドエンティティ数

	Twitter	Yahoo!知恵袋	Yahoo!検索
トレンドエンティティ数	25,700	9,047	98,497
エンティティのユニーク数	10,832	5,145	42,267

3.2.3　トレンドエンティティの抽出

テキストに紐付けられたエンティティの出現回数に基づいて，各Webメディアのトレンドエンティティを抽出する．本調査では各Webメディアのデータを日別に分割し，あるエンティティeについて，ある調査対象日tのエンティティの出現回数$n(e,t)$とそれ以前の参考期間Rにおけるエンティティの出現回数の平均値$\mu(e,R)$と標準偏差$\sigma(e,R)$を用いて以下の式で定義される$zscore$を計算する．

$$zscore(e,t,R) = \frac{n(e,t) - \mu(e,R)}{\sigma(e,R)} \quad (1)$$

本調査では$zscore(e,t,R)$の値がしきい値θ以上のエンティティeを対象日tのトレンドエンティティとして抽出する．本調査では，参考期間Rとして対象日の直前7日間を利用し，$zscore(e,t)$のしきい値を$\theta = 5$とした．

表1に各Webメディアから抽出したトレンドエンティティ数とユニークエンティティ数を示す．トレンドエンティティ数については，Yahoo!検索が最も多く，次いでTwitterが多い． Yahoo!検索とTwitterはユニークエンティティ数に対してトレンドエンティティ数が2倍以上であり，同一のエンティティがトレンドエンティティとして複数回観測されやすいメディアであると考えられる．

3.2.4　分散表現を用いたエンティティのベクトル化

3.2.2項で説明したエンティティリンキングを適用後のテキストに対して単語の分散表現手法を適用し，エンティティのベクトル化を行う．分散表現により，各エンティティはテキスト中で互いに共起する単語やエンティティの情報に基づいたベクトルとして表現される．そのため，単語やエンティティとの共起パターンが類似している，すなわち，注目の観点が類似しているエンティティ同士であるほど互いに類似したベクトルとして表現される．また，異なるメディア間であってもエンティティ間の共起パターンが類似している場合，エンティティ間の関連性はメディア間で類似していると考えられる．そこで，あるトレンドエンティティについてメディア間でベクトルを比較することで，そのトレンドエンティティに対するメディア間のトレンドの文脈の差異を明らかにできると考えられる．

本調査では，各Webメディアのデータを日別に分割し，Yaoらが提案した時系列を考慮した分散表現手法 [11]を用いて各日におけるエンティティのベクトル化を行った．この手法はあるタイムウィンドウ幅で分割されたテキストの時系列集合を入力として，隣接するタイムウィンドウ間のベクトルの差が最小となるように分散表現を獲得する．これにより，時間経過により他の単語との共起に変化が少ない単語のベクトルは異なるタイムウィンドウ間でも似たベクトルとして表現される一方，共起の変化が大きい単

語はタイムウィンドウ間で異なるベクトルが得られる．本調査はトレンドエンティティというある期間においてのみ頻出するエンティティを調査の対象としているため，上記の特徴を持つYaoらの手法を用いた．本調査では，計算量の観点から分散表現による獲得するエンティティベクトルの次元数を10とした．また，ベクトルの計算に関わる条件をWebメディア間で共通とするため手法のパラメータには文献 [11]中で用いられている値を用いた．

各メディアからそれぞれ獲得した分散表現は，同一のエンティティのベクトルであっても，メディア間で直接比較することができないため，異なるメディア間で分散表現のアライメントが必要である．本調査では，異なるメディア間で類似した文脈を持つトレンドエンティティのベクトルは互いに類似したベクトルとして表現できればよいため，分散表現を用いた異なる言語間の単語翻訳タスク [1]で用いられている直交行列によるアライメント手法を用いた．メディアA，B間で共通して出現するエンティティのベクトル行列をそれぞれ$\mathbf{W}_A$および$\mathbf{W}_B$としたとき，メディアAからメディアBへのアライメント行列$\mathbf{R}$は次の最適化により得られる．

$$\underset{\mathbf{R}\ s.t.\ \mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\arg\min} \|\mathbf{W}_A\mathbf{R}-\mathbf{W}_B\|_F. \tag{2}$$

ここで$\mathbf{I}$は単位行列，$\|\mathbf{M}\|_F$は行列$\mathbf{M}$のフロベニウスノルムを表す．なお，式(2)による最適化については，特異値分解による効率的な解法が存在する [7]．本調査では，TwitterとYahoo!知恵袋の分散表現をそれぞれYahoo!検索の分散表現にアライメントした後，メディア間でトレンドエンティティのベクトルの比較を行った．なお，ベクトル間類似度としてコサイン類似度を用いた．

4 トレンドエンティティのメディア間横断調査

本章では，3章で述べた前処理により得られたトレンドエンティティとエンティティの分散表現を用いて，トレンドエンティティの特徴についてWebメディアを横断した調査を行う．はじめに，4.1節でトレンドエンティティについてメディア間の量的な差異について調査を行った後，4.2節でトレンドエンティティについて文脈の差異の調査を行う．

4.1 トレンドエンティティの量的な差異の調査

各メディアに出現するトレンドエンティティについて，吉田と荒瀬の研究 [13]を参考に，以下の観点に基づくメディア間を横断して出現するトレンドエンティティが持つ量的な特徴を調査する．

- メディア間に共通して出現するトレンドエンティティ数
- トレンド発生日のずれ
- 頻度情報に基づくトレンドの持続期間

4.1.1 共通トレンドエンティティ数

表2にメディア間に共通して出現するトレンドエンティティ数を示す．表2より，三つのメディア間で共通して出現するトレンドエンティティは1,584件あり，Twitter・Yahoo!検索間の7,150件の場合を除いて，本実験で抽出したトレンドエンティティのほとんどが単一のWebメディアにのみ出現することがわかった．これは，普段は各Webメディアでは内的なトレンドが発生しており，あるエンティティについて注目度の高い出来事が発生したときのみ，メディア間を横断してトレンドエンティティが観測されるためだと考えられる．実際に三つのメディア間で共通のトレンドエンティティとして，2017年8月末には北朝鮮によるミサイル発射やそれに伴う全国瞬時警報システムに関するエンティティ，また，2017年9月13日には当日に発表されたスマートフォンiPhone 8やiPhone Xのエンティティなど，メディアを問わず多くのユーザから注目されていたと考えられるエンティティが多く含まれていた．複数のWebメディアに横断して出現しやすいトレンドエンティティであるほど，多くのユーザから注目されており，また，その注目の観点も類似していると予想される． 4.2節の調査では，実際にメディア間を横断しやすいトレンドエンティティであるほどそのトレンドの文脈がメディア間で類似する傾向にあるが，トレンドエンティティによってその類似性にばらつきがあることを示す．

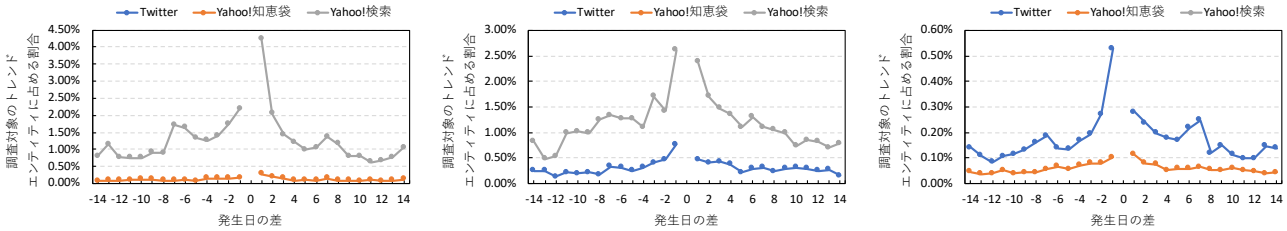
表2　メディア間の共通トレンドエンティティ集合のサイズ

	集合のサイズ $\lvert\mathcal{E}_X\rvert$
$\mathcal{E}_{all}$：全Webメディア	1,587
$\mathcal{E}_{t,c}$：Twitter・Yahoo!知恵袋のみ	301
$\mathcal{E}_{t,s}$：Twitter・Yahoo!検索のみ	7,150
$\mathcal{E}_{c,s}$：Yahoo!知恵袋・Yahoo!検索のみ	553
$\mathcal{E}_t$：Twitterのみ	16,665
$\mathcal{E}_c$：Yahoo!知恵袋のみ	6,609
$\mathcal{E}_s$：Yahoo!検索のみ	89,210

4.1.2 トレンド発生日のずれ

あるエンティティがTwitterでトレンドエンティティとなった翌日にYahoo!知恵袋でもトレンドエンティティとなるような，メディア間で同じエンティティがトレンドエンティティとして抽出されるタイミングに差が生じる可能性が考えられる．そこで表2に示した各Webメディアにおいてのみ抽出されたトレンドエンティティ（$\mathcal{E}_t$，$\mathcal{E}_c$，$\mathcal{E}_s$）が，別のWebメディアにおいて別日のトレンドエンティティとして出現していないかについて調査した．具体的には，あるメディアAの日tのトレンドエンティティeについて，別のメディアBの別日$t'(\neq t)$にトレンドエンティティとして出現している場合を列挙し，そのうち抽出日の絶対日数差$|t'-t|$が最小となる日t'を求めた． $t'-t>0$であれば，エンティティeはメディアAでトレンドエンティティとなった$|t'-t|$日後にメディアBでトレンドエンティティになっており，$t'-t<0$であればエンティティeはメディアAでトレンドエンティティとなる$|t'-t|$日前にメディアBでトレンドエンティティとなっていたエンティティである．

図2にWebメディア間のトレンドエンティティ抽出日のずれの調査結果について示す．図2より，Twitterの方で先にトレンドエンティティとなるエンティティが多い傾向にあることがわかる．また，この傾向はTwitter・Yahoo!検索間が顕著である．これは，あるエンティティがTwitter上で多くのユーザに言及されトレンドとなった後に，Yahoo!知恵袋やYahoo!検索でもそのエンティテ



(a) Twitterに対するずれ　(b) Yahoo!知恵袋に対するずれ　(c) Yahoo!検索に対するずれ

図2　メディア間のトレンドエンティティ抽出日のずれ

表3　頻度情報に基づくトレンドの継続期間の平均日数と標準偏差

頻度条件	Twitter	Yahoo!知恵袋	Yahoo!検索
50%	1.89 (2.99)	2.67 (4.09)	2.83 (6.91)
30%	2.86 (5.11)	4.75 (9.07)	4.78 (10.12)
20%	4.21 (8.13)	6.77 (11.77)	7.17 (17.52)

ィについて質問あるいは検索される回数が増えるといったWebメディアを横断したトレンドの伝播の存在を示唆している．また，Twitter・Yahoo!検索間では，前後7日を頂点とした小さな山ができており，1週間間隔の周期的なトレンドがTwitter・Yahoo!検索間で交互に出現している可能性が示唆される．一方，Yahoo!知恵袋はTwitterよりYahoo!検索との間に多くのトレンドの伝播があることがわかる．これは，ユーザが持つ疑問を解消を目的として利用されるYahoo!知恵袋では，情報発信が目的のTwitterよりも，情報の検索が目的のYahoo!検索の方がユーザの行動が類似しているためだと考えられる．また，Yahoo!知恵袋はYahoo!検索と同じ事業者により運営されているWebメディアであることも，Yahoo!知恵袋に対してYahoo!検索からのトレンドの伝播が多い要因の一つと考えられる．このようなメディア間で伝播するトレンドエンティティは，伝播する以前から類似したトレンドの質を持っていると考えられる．上記の点について，後述する4.2.2項の調査では，一方のメディアにのみ出現しているトレンドエンティティについてもメディア間のトレンドの質の差異を調査する．

4.1.3　頻度情報に基づくトレンドの継続期間

あるエンティティがトレンドエンティティとして観測された後，そのトレンドエンティティに対する注目の度合がどれくらいの期間維持されるかについてもメディア間で差があると予想される．例えば，Twitterは話題の移り変わりが早く，あるエンティティがトレンドエンティティとして抽出されたとしても翌日には言及数がトレンド発生以前に戻るといった場合が考えられる．そこで，トレンドの継続日数について調査し，トレンドエンティティの移り変わりの速度がメディア間でどのように異なるかを明らかにする．具体的には，表2に示した全てのWebメディアに共通して出現するトレンドエンティティ$\mathcal{E}_{all}$について，対象のトレンドエンティティの抽出日を起点として，トレンドの継続日数を起点日におけるエンティティの出現頻度が{50, 30, 20}%未満になるまでの日数として調査した．

表3に各Webメディアの頻度情報に基づくトレンドの継続日数の平均と標準偏差を示す．また，図3に各Webメディアのトレンドの継続日数の分布を頻度条件別に示す．なお，表3に示した各頻度条件における平均日数についてウェルチのt検定により，Yahoo!知恵袋・Yahoo!検索間を除き，有意水準1%の有意差を確認した．いずれのWebメディアにおいても，全ての頻度条件で3日以内に半数以上のトレンドエンティティが収束していることが確認できる．また，Yahoo!知恵袋およびYahoo!検索に対して，Twitter上のトレンドエンティティは当日または翌日に収束する傾向にあり，Twitterは短い期間でトレンドが変化しやすいメディアだと考えられる．一方，Yahoo!知恵袋およびYahoo!検索はトレンドが発生するとそのトレンドに対する注目度（質問数や検索頻度）が維持されやすいメディアであると考えられる．TwitterとYahoo!検索については，吉田と荒瀬の研究 [13]におけるトレンドキーワードを対象とした調査でも類似した結果が示唆されている．このようにYahoo!知恵袋およびYahoo!検索に対して，Twitterはトレンドに対する注目の度合が変化しやすい傾向にあることが示唆されたが，後述の4.2.1項の調査ではTwitterの方がトレンドに対する注目の観点のようなトレンドの質が長期間維持されやすいことを示す．

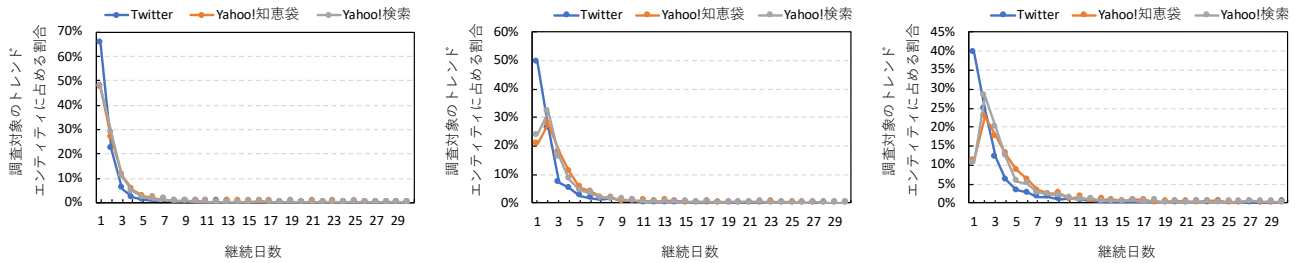
4.2　トレンドエンティティの文脈の差異の調査

本節では，メディア間のトレンドエンティティの文脈の差異を明らかにするために，3.2.4項で延べた手順により獲得したエンティティの分散表現（ベクトル）を用いて，1章で延べた以下の観点による調査を行う．

- 分散表現に基づくトレンドの継続期間
- メディア間のトレンドの文脈の差異
- Wikipediaの意味情報との比較

4.2.1　分散表現に基づくトレンドの継続期間

4.1.3項の調査ではトレンドエンティティの頻度情報に基づく継続期間の調査を行ったが，頻度が減少した後でも対象のエンティティに対する注目の観点のような文脈がトレンド発生時点から変化していない場合が考えられる．また，メディア間で共通のトレンドエンティティであってもトレンドの文脈の変化はメディアごとに異なると予想される．そこで，表2に示した全てのメディアで共通して出現するトレンドエンティティ$\mathcal{E}_{all}$を対象として，各メディアにおいてトレンド発生日からトレンドエンティティのベクトルの経時変化について調査した．具体的には，トレンドエンティティが抽出された日を起点として，分散表現に基づく



(a) 頻度条件50%　(b) 頻度条件30%　(c) 頻度条件20%

図3　頻度情報に基づくトレンドの継続日数

表4　分散表現に基づくトレンドの継続期間の平均日数と標準偏差

類似度条件	Twitter	Yahoo!知恵袋	Yahoo!検索
0.5	1.69 (0.95)	1.63 (0.89)	1.88 (1.07)
0.3	2.25 (1.38)	2.06 (1.25)	2.33 (1.62)
0.2	2.59 (1.64)	2.35 (1.53)	2.63 (1.93)

トレンドの継続期間を起点日におけるトレンドエンティティのベクトルに対するベクトル間類似度（コサイン類似度）が{0.5, 0.3, 0.2}未満になるまでの日数として調査した．

表4に各Webメディアの分散表現に基づくトレンドの継続日数の平均と標準偏差を示す．また，図4に各Webメディアのトレンドの継続日数の分布を類似度条件別に示す．なお，表4に示した各類似度条件における平均日数についてウェルチのt検定により，条件50%におけるTwiiter・Yahoo!知恵袋間および条件{30, 20}%におけるTwitter・Yahoo!検索間を除いて，有意水準1%の有意差を確認した． 4.1.3項の結果と比較して，分散表現に基づくトレンドの継続日数はメディア間で大きな差がないことがわかる．このことから，Twitterではトレンドエンティティに対する言及回数が減少したとしても，トレンド発生時のトレンドエンティティに対する注目の観点が維持されやすい傾向にあることが示唆された．また，Yahoo!知恵袋やYahoo!検索はトレンドエンティティに対する言及回数は維持されやすい一方，ベクトルの変化が早いことから，これらのメディアはトレンドエンティティに対する注目の観点が変化しやすいメディアであると考えられる．実際に，Yahoo!検索では台風のエンティティが2017年9月12日から17日にかけてトレンドとなっていたが，その検索内容が台風自体に関する検索から台風の被害に関する検索に変化する等の例が見られた．そのため複数のメディアから横断して共通するトレンドを抽出し追跡するようなアプリケーションでは，対象のトレンドの頻度情報のような量的な観点だけでなく，そのトレンドに対する注目の観点のようなトレンドの文脈を考慮することが重要であると考えられる．

4.2.2　メディア間のトレンドの文脈の差異

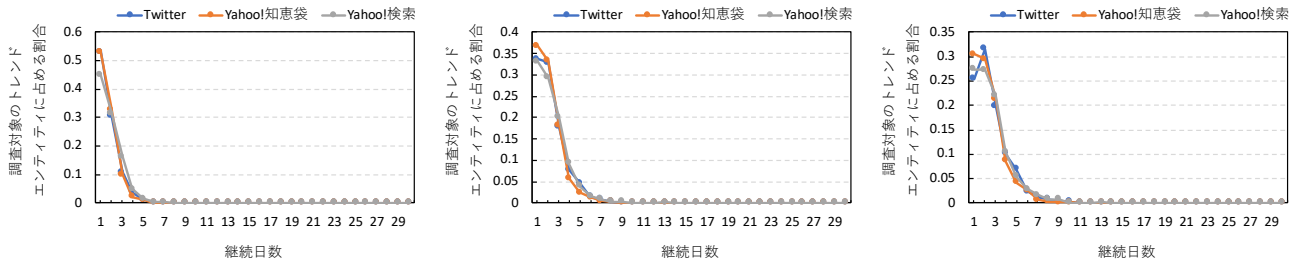
メディア間で共通のトレンドエンティティが出現したとしても，そのトレンドに対する注目の観点はメディアごとに異なる可能性がある．例えば，TwitterとYahoo!知恵袋間において企業の新製品がトレンドエンティティとして抽出された場合，Twitterで

表5　メディア間のベクトル間類似度の平均値と標準偏差

	Twitter・知恵袋間	Twitter・検索間	知恵袋・検索間
$\mathcal{E}_{all}$	0.55 (0.21)	0.63 (0.16)	0.65 (0.18)
$\mathcal{E}_{t,c}$	0.53 (0.27)	0.59 (0.20)	0.55 (0.24)
$\mathcal{E}_{t,s}$	0.47 (0.28)	0.64 (0.18)	0.53 (0.29)
$\mathcal{E}_{c,s}$	0.51 (0.24)	0.55 (0.24)	0.66 (0.19)
$\mathcal{E}_{t}$	0.34 (0.33)	0.44 (0.31)	0.35 (0.35)
$\mathcal{E}_{c}$	0.32 (0.34)	0.29 (0.35)	0.42 (0.34)
$\mathcal{E}_{s}$	0.19 (0.33)	0.29 (0.34)	0.27 (0.35)

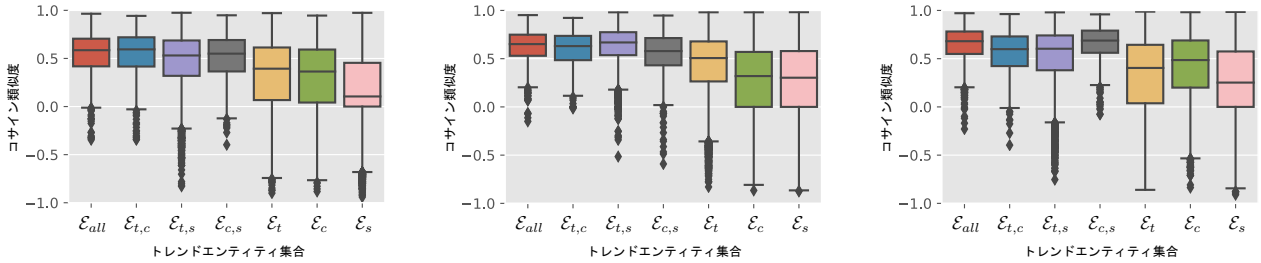
はその新製品についてのみが言及されやすい一方，Yahoo!知恵袋ではその製品の使い勝手や他の競合製品との比較に関する質問がされやすいといった違いが考えられる．そこで，トレンドエンティティのベクトルをメディア間で比較し，その類似度からトレンドエンティティに関するメディア間のトレンドの質の差異を調査した．具体的には，表2に示した各集合に含まれるトレンドエンティティについて，そのベクトルを全てのメディア間で比較し，ベクトル間類似度の分布を調査した．

表5にトレンドエンティティの異なるメディア間のベクトル間類似度の平均値と標準偏差を示す．また，図5にトレンドエンティティのベクトル間類似度の分布をそれぞれのメディアペアごとに示す．なお，表5に示した各メディアペア間におけるトレンドエンティティの平均類似度についてウェルチのt検定により，Twitter・Yahoo!知恵袋の$\mathcal{E}_{all}$・$\mathcal{E}_{t,c}$間および$\mathcal{E}_{t,c}$・$\mathcal{E}_{c,s}$間，Twitter・Yahoo!検索の$\mathcal{E}_{all}$・$\mathcal{E}_{t,s}$間および$\mathcal{E}_{c}$・$\mathcal{E}_{s}$間，Yahoo!知恵袋・Yahoo!検索間の$\mathcal{E}_{all}$・$\mathcal{E}_{c,s}$間および$\mathcal{E}_{t,c}$・$\mathcal{E}_{c,s}$間を除いて，有意水準1%の有意差を確認した．これらの結果より，対象のメディア間で共通して出現しているトレンドエンティティはそのメディア間のベクトル間類似度が高く，トレンド質の差異が小さい傾向にあることがわかる．しかし，メディア間に共通のトレンドエンティティであってもベクトル間類似度は幅広い範囲に分布しており，メディア間である共通のエンティティが注目されている場合でもその注目の観点が同じであるとは限らないことが示唆された．例えば，「大学入試センター試験」は2017年1月9日に全てのメディアで共通して出現しているトレンドエンティティであったが，メディア間のベクトル間類似度はYahoo!知恵袋



(a) 類似度条件0.5　(b) 類似度条件0.3　(c) 類似度条件0.2

図4　分散表現に基づくトレンドの継続日数



(a) Twitter・Yahoo!知恵袋間　(b) Twitter・Yahoo!検索間　(c) Yahoo!知恵袋・Yahoo!検索間

図5　トレンドエンティティの異なるメディア間のベクトル間類似度の分布

・Yahoo!検索間で0.549であるのに対して，Twitter・Yahoo!知恵袋間は−0.317であり，このようなトレンドエンティティはメディア間で異なる文脈を持つと予想される．そのため，メディア間で共通するトレンドを抽出・分析するようなアプリケーションでは，頻度情報などのトレンドの量的な観点だけでなく，本調査で対象としているようなトレンドの文脈も考慮する必要があると考えられる．

Twitter・Yahoo!検索間の$\mathcal{E}_t$，Yahoo!知恵袋・Yahoo!検索間の$\mathcal{E}_c$のように，一方のメディアにのみ出現するトレンドエンティティであってもメディア間のベクトル間類似度が高い場合も確認された．これは，4.1.2項の調査結果から，Twitterおよび Yahoo!知恵袋にのみ出現するトレンドエンティティ（$\mathcal{E}_t$，$\mathcal{E}_c$）が，異なる日にYahoo!検索のトレンドエンティティとして出現する傾向にあることが要因であると考えられる．このようなメディア間のトレンドの文脈の類似性は，メディアを横断したトレンドの伝播を早期に予測するアプリケーションなどに利用できる可能性がある．

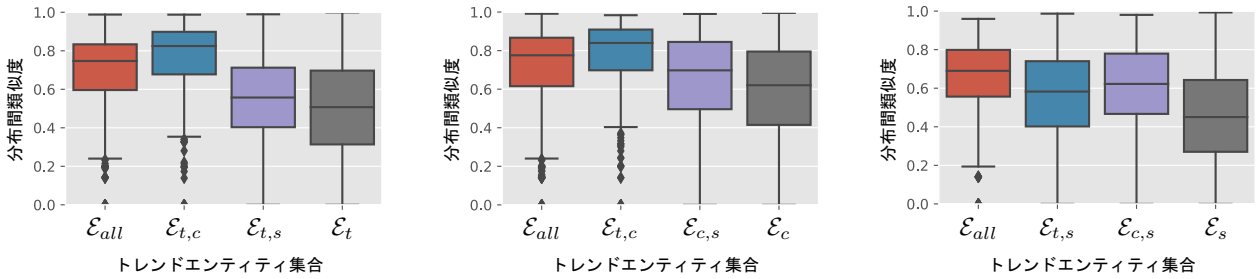
4.2.3　Wikipediaの意味情報との比較

あるWebメディアで何らかのトレンドが発生した場合，そのトレンドに対応するエンティティについて，他のエンティティ間と新しい関連性が生じることが考えられる．このようなトレンドを対象とした分析において，Wikipediaなどの知識源内で定義されているエンティティ間の関連性を直接用いると，トレンドに対して誤った解釈を導く可能性が考えられる．そこで，各メディアにおけるトレンドエンティティと他のエンティティの関連性とWikipediaの意味情報に基づくエンティティ間の関連性の関係を調査する．具体的には，各メディアに出現するあるトレンドエンティティに対してベクトル間類似度が高い上位50件のエンティティについて，対象のトレンドエンティティとそれらのエンティティのベクトル間類似度の分布とWikipediaにおけるエンティティ間関連度の分布の分布間類似度を算出し，その分布間類似度が対象のトレンドエンティティが出現するメディアの種類ごとにどのような傾向があるかを調査する． Wikipediaにおけるエンティティ間関連度の算出には，エンティティに対する被リンクの類似性に基づく手法 [6]を用いた．

表6　Wikipediaのエンティティ間関連度に対する分布間類似度の平均値と標準偏差

	Twitter	Yahoo!知恵袋	Yahoo!検索
$\mathcal{E}_{all}$	0.70 (0.18)	0.73 (0.19)	0.66 (0.17)
$\mathcal{E}_{t,c}$	0.75 (0.21)	0.77 (0.21)	–
$\mathcal{E}_{t,s}$	0.55 (0.21)	–	0.56 (0.22)
$\mathcal{E}_{c,s}$	–	0.66 (0.22)	0.61 (0.21)
$\mathcal{E}_t$	0.50 (0.24)	–	–
$\mathcal{E}_c$	–	0.59 (0.24)	–
$\mathcal{E}_s$	–	–	0.45 (0.24)

表6に各メディアのベクトル間類似度とWikipediaのエンティティ間類似度の分布間類似度の平均値と標準偏差を示す．また，図6に各メディアのベクトル間類似度とWikipediaのエンティティ間類似度の分布間類似度の分布を示す．なお，表6に示した各メディアの平均分布間類似度についてウェルチのt検定により有意水準1%の有意差を確認した．これらの結果から，Yahoo!知恵袋



(a) Twitter (b) Yahoo!知恵袋 (c) Yahoo!検索

図6 各メディアにおけるベクトル間類似度とWikipediaのエンティティ間類似度の分布間類似度の分布

に出現するトレンドエンティティ（$\mathcal{E}_{all}$，$\mathcal{E}_{t,c}$，$\mathcal{E}_{c,s}\cdot\mathcal{E}_{c}$）についてはメディアに関わらずWikipediaとの分布間類似度が高い傾向にあることがわかった．このことから，Yahoo!知恵袋ではあるエンティティについてWikipediaなどの知識源で整理されている他のエンティティとの関連性に基づいた質問が多くされる傾向にあると考えられる．全てのメディアに共通して出現するトレンドエンティティ$\mathcal{E}_{all}$についても，全てのメディアでWikipediaとの分布間類似度が高い傾向にあることがわかった． Wikipediaではエンティティに関する記事の編集方針として，エンティティの特筆性と信頼できる情報源の参照が挙げられており，社会的に注目度が高く，複数の報道機関に取り扱われるような情報ほどWikipediaの記事に反映されやすい．本調査で対象としているトレンドエンティティについても，複数のメディアに共通するトレンドエンティティであるほど社会的に注目度が高いものだと考えられるため，他のエンティティとの関連性がWikipediaのエンティティ間関連度と類似していたと考えられる．なお，今回の調査では調査対象のデータセットの収集期間よりも更新日時が新しいWikipediaのダンプデータを用いており，データセットの収集期間に新たに生じたエンティティ間の関係が整理された結果，その傾向がより顕著に出ていた可能性がある．トレンドエンティティが横断して出現するメディア数に着目することで，Wikipediaなどの知識源に整理されるようなエンティティ間の新たな関係を抽出できる可能性がある．また，トレンド分析においてエンティティ自身が持つ意味情報を用いる場合は，対象のトレンドが同時に観測されるメディア数や分析に利用する知識源のデータの更新日時を考慮する必要があると考えられる．今後はトレンド発生時点におけるWikipediaのダンプデータを用いて同様の調査を行い，上記の点について検証する予定である．

5 おわりに

本研究では，Twitter，Yahoo!知恵袋，Yahoo!検索の三つのWebメディアについて，トレンドエンティティに関する横断的な調査を行い，メディアを横断したトレンド分析においてトレンドの文脈の差異を考慮する必要性について議論した．調査の結果，メディア間で共通して出現するトレンドエンティティであっても，そのエンティティに対する注目の観点のようなトレンドの文脈が異なることがわかった．そのため，メディアを横断したトレンド分析においては，トレンドに関するキーワードやエンティティの出現回数といった量的な情報だけでなく，メディア間で共通して出現するトレンドの文脈の差異を考慮することが重要であると考えられる．今後の課題として，本調査により得られた知見をメディア間を横断するトレンドの抽出やモニタリングに関する手法に適用し，その有用性を検証することが挙げられる．

参考文献

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*, pages 2289–2294, 2016.
[2] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. The pulse of news in social media: Forecasting popularity. In *ICWSM*, 2012.
[3] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging topic detection for organizations from microblogs. In *SIGIR*, pages 43–52, 2013.
[4] Paolo Ferragina and Ugo Scaiella. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*, pages 1625–1628, 2010.
[5] Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD*, pages 832–840, 2011.
[6] David Milne and Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *WikiAI*, pages 25–30, 2008.
[7] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.
[8] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #twittersearch: A comparison of microblog search and web search. In *WSDM*, pages 35–44, 2011.
[9] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *KDD*, pages 784–793, 2007.
[10] Weu Xie, Feida. Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. Topicsketch: Real-time bursty topic detection from twitter. *TKDE*, 28(8):2216–2229, 2016.
[11] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *WSDM*, pages 673–681, 2018.
[12] Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, and Mikio Yamamoto. Wikipedia page view reflects web search trend. In *WebSci*, pages 65:1–65:2, 2015.
[13] 吉田光男 and 荒瀬由紀. トレンドキーワードに関するウェブリソースの横断的分析. 情報処理学会論文誌データベース *(TOD)*, 9(1):20–30, 2016.