情報処理学会論文誌　データベース　Vol.9　No.1　20–30 (Mar. 2016)

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

吉田 光男[1,a)]　荒瀬 由紀[2,b)]

受付日 2015年9月27日, 採録日 2016年1月6日

概要：ソーシャルメディアでの言及量やウェブ検索エンジンでの検索頻度をもとに，トレンドキーワードを発見する研究開発が広く行われている．また，注目されているキーワードに対して情報を付与し，そのキーワードの理解を促すような試みもある．しかし，それらのトレンドキーワードが様々なウェブリソースでどのように振る舞うのかは必ずしも明らかではない．そこで本研究では，トレンドをとらえうるウェブリソースを対象に，収集したトレンドキーワードがどのように振る舞うのかを横断的に調査する．この調査により，大半のトレンドキーワードがオンライン辞書サービスに登録されていないこと，検索のトレンドは 2 日で 50%未満の頻度に収束すること，ソーシャルメディア（Twitter）がほかのウェブリソースよりもトレンドに敏感であることなどを明らかにする．

キーワード：トレンドキーワード，検索エンジン，頻度分析，トレンド分析

Trend Query Analysis on Heterogeneous Web Resources

Mitsuo Yoshida[1,a)]　Yuki Arase[2,b)]

Received: September 27, 2015, Accepted: January 6, 2016

Abstract: Many researchers work on studies for discovering trend keywords and queries on the web, i.e., search frequency and social media. Moreover, studies on trend query classifications are being conducted. However, the behavior of trend queries for various web resources is unclear. In this study, we investigate how trend queries appear in different resources on the web. We clarify the following. (1) Most trend queries are not registered with online dictionary services. (2) The trend converges in approximately two days. (3) Social media websites (such as Twitter) are responsive to trend queries.

Keywords: trend query, search engine, frequency analysis, trend analysis

1. はじめに

ウェブ検索エンジンは利用者がなにかを知りたいときに利用されることから，検索エンジンに入力されたキーワード（クエリ）の頻度は利用者の興味関心を反映している可能性が高い．実際，この仮説に基づいた研究開発がさかんに行われている [1], [2]．これらの研究開発では，ニュースの流行や車の販売台数などの予測を試みている．このような予測を行う以外にも，検索事業者は**図 1** のように，集中的に検索されているキーワード，すなわち，利用者からの興味関心が高まっているキーワードを**トレンドキーワード**として提供している．

検索頻度データは一般的に入手するのが困難であることから，検索頻度によらず，トレンドキーワードを獲得する試みもなされている [3], [4]．しかしながら，獲得されたトレンドキーワードがウェブリソースでどのように振る舞うのかは必ずしも明らかになっていない．トレンドキーワードはどれほどの期間，検索され続けるのだろうか．あるいは，ソーシャルメディアや掲示板で言及され続けるのだろうか．そもそも，検索頻度をもとにしたトレンドキーワードは，ソーシャルメディアに出現するのだろうか．本研究では，このような疑問に答えるべく，ウェブリソースを横

[1] 豊橋技術科学大学
Toyohashi University of Technology, Toyohashi, Aichi 441–8580, Japan
[2] 大阪大学
Osaka University, Suita, Osaka 565–0871, Japan
[a)] yoshida@cs.tut.ac.jp
[b)] arase@ist.osaka-u.ac.jp

図 1　Yahoo! JAPAN が提供しているトレンドキーワード
Fig. 1　Example of trend queries provided by Yahoo! JAPAN.

断したトレンドキーワードに関する調査を行い，トレンドキーワードの特徴を明らかにする．特徴を明らかにすることにより，今後，トレンドキーワードを取り扱う研究開発において，その方針を制定することを支援する．

本稿では，焦点を次の 3 点に絞り，ウェブリソースを横断した調査を行う．

(1) トレンドキーワードそのものの特徴．
(2) トレンドキーワード発生源である検索エンジンとの関係．
(3) 検索エンジン以外のウェブリソースでの出現状況．

これらの調査により，大半のトレンドキーワードがオンライン辞書サービスに登録されていないこと，検索のトレンドは 2 日で 50%未満の頻度に収束すること，ソーシャルメディア（Twitter）がほかのウェブリソースよりもトレンドに敏感であることなどを明らかにする．

2.　関連研究

検索頻度データを分析することで，トレンドの予測を試みる研究開発がさかんに行われている．Radinsky ら [1] は，ニュースの流行を予測するために，検索頻度を利用している．Choi と Varian [2] は，自動車販売台数や失業率などの経済指標の予測を試みている．一般的に検索頻度データは公開されていないことから，いずれの研究においても検索頻度データとして Google トレンド[*1]から得られる頻度データを利用している．検索事業会社であるヤフー株式会社は，自社で保有する検索頻度データをもとにした「Yahoo! JAPAN ビックデータレポート」を公開している[*2]．ここでは，選挙の議席数や経済指標，感染症患者数などの予測および分析が公開されている．検索頻度データを利用したトレンドキーワードの抽出は，筆者らが調査した限りにおいては研究されておらず（公知になっておらず），検索事業者によって提供されているにとどまる．

検索頻度データの入手が困難であることから，検索頻度によらず，トレンドキーワードを獲得する試みもなされている．古川ら [3] はブログ空間上の話題の移り変わりに着目し，重要語を抽出する手法を提案している．Benhardus ら [4] はソーシャルメディア（Twitter）のデータに対し，従来から用いられている TF-IDF [5] などに基づく手法を適用し，高精度にトレンドキーワードを抽出できることを示した．また，データに依存せず，頻度パターンからバースト性を検出し，キーワードを抽出する手法もある [6]．

検索頻度データそのものを別のデータで再現する試みもある．Huss ら [7] は遺伝子工学に関するコミュニティ活動のために，その分野がどれほど需要があるかを検索頻度と Wikipedia のページビューとのそれぞれのデータで確認している．Kristoufek [8] はビットコインの市場価格を予測するための基礎的な調査として，ビットコインの市場価格と検索頻度および Wikipedia のページビューとの類似性を調査している．これらの調査をもとに，吉田ら [9] は検索頻度の推定を目的とし，その基礎的な調査として検索頻度と Wikipedia のページビューとの関係を多面的に調査している．この調査により，少なくとも検索頻度の高いキーワードであれば，推定が可能であると示唆されている．

トレンドキーワードを対象とする研究としては，キーワード分類があげられる．Yoshida と Arase [10] は集中的に検索されるクエリ（トレンドキーワード）を分類するために，即時性のある Twitter のデータを利用している．その手法および条件を拡張し，トレンドキーワードが出現した当日の検索結果データが分類に有効であることも報告している [11]．トレンドキーワードは特定の期間に集中的に検索されたキーワードであるが，検索キーワード一般としては，恒常的に検索されるクエリ，周期的に検索されるクエリも存在することが知られている [12]．

本研究では，トレンドキーワードがウェブリソースでどのように出現し，どのような特徴があるのかを調査する．トレンドキーワードの抽出や活用には，ニュースやソーシャルメディア（Twitter）のデータを利用する傾向があるが，本研究では，それらに加えて，掲示板（2 ちゃんねる）のスレッドタイトルや各種辞書サービスのページビューデータも利用するなど，多様なウェブリソースを横断した分析を行う．本調査により，大半のトレンドキーワードがオンライン辞書サービスに登録されていないこと，検索のトレンドは 2 日で 50%未満の頻度に収束すること，ソーシャルメディア（Twitter）がほかのウェブリソースよりもトレンドに敏感であることなどを明らかにする．

3.　調査対象トレンドキーワード

本研究で調査の対象とするトレンドキーワードは，一

*1　https://www.google.co.jp/trends/ (cited 2016–01–25)
*2　http://docs.yahoo.co.jp/info/bigdata/ (cited 2016–01–25)

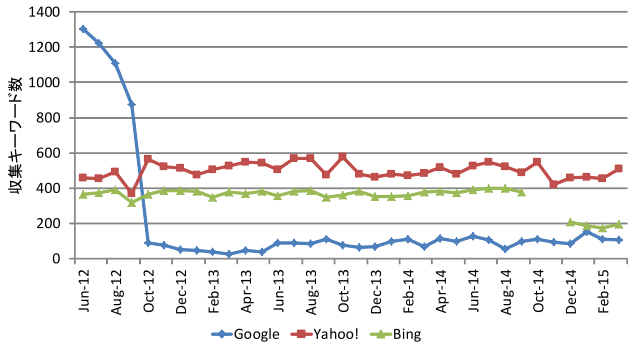


図 2　月別のトレンドキーワード収集数

Fig. 2　The number of trend queries collected per month.

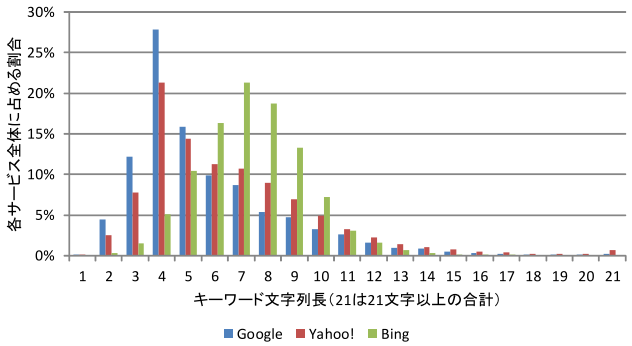


図 3　トレンドキーワード収集元別のキーワード字列長の分布

Fig. 3　The character length of trend queries collected from Google, Yahoo! JAPAN, and Bing.

般に公開されている「Google トレンド 急上昇ワード[*3]」「Yahoo! 検索データ デイリーランキング 急上昇ワード[*4]」「Bing 話題の検索ワード[*5]」の 3 サービスから収集する．毎時 0 分および 30 分に各サービスにアクセスし，未収集であるキーワードを新規のトレンドキーワードとして収集する．また，その収集時点を**トレンド日時**とし，その日付を**トレンド日**とする．

2012 年 6 月 1 日から 2015 年 3 月 31 日までの期間で収集を行い，計 35,192 キーワードを収集した．**図 2** に収集したトレンドキーワードの数を月別に示す．2012 年 10 月に「Google トレンド 急上昇ワード」の仕様変更により該当サービスからの収集量が大幅に低下し，2014 年 10 月には「Bing 話題の検索ワード」の仕様変更により収集システムの対応までの間，新規トレンドキーワードの収集を停止していた．「Yahoo! 検索データ デイリーランキング 急上昇ワード」に関しては，特段の仕様変更はなく，安定的に収集できている．収集したトレンドキーワードの内訳は，Google が 7,027 キーワード（20.0%），Yahoo! が 16,991 キーワード（48.3%），Bing が 11,174 キーワード（31.6%）である．なお，本研究で調査の対象とするトレンドキーワードおよび一部のウェブリソースは参照できる[*6]．

4.　トレンドキーワードの特徴

4.1　トレンドキーワードの表層

本節では各サービスから収集したトレンドキーワードがどのようなキーワードであるか，その表層を調査する．まず，トレンドキーワードが複数の語からなるか否か，つまりトレンドキーワードに空白文字が含まれるか否かを調査した．その結果，Google から収集したトレンドキーワードのうち 87.5%が 1 つの語からなるのに対し，Yahoo! および Bing から収集したトレンドキーワードの場合，それぞれ 75.4%および 55.3%と比較的低い値であった．サービスごとに入力される単語数が著しく変化するとは考えづらく，Yahoo! および Bing は実際のトレンドキーワードに加え，トレンド要因に関係するキーワードを付与している可能性がある．筆者らによる観察においても，編集されている可能性を確認した．

次に，トレンドキーワードの文字列長を調査した．Google，Yahoo!，Bing から収集したトレンドキーワードの平均文字列長はそれぞれ 5.8 文字，6.6 文字，7.3 文字であった．文字列長の分布を**図 3** に示す．Google および Yahoo! では 4 文字からなるトレンドキーワードが多数提供されており，Bing では 7 文字からなるトレンドキーワードが多数提供されている．Bing は複数の語からなるトレンドキーワードを多数提供しており，その結果，全体的にトレンドキーワードの文字列長が大きくなったと考えられる．

4.2　オンライン辞書での登録状況

これまでのトレンドキーワードの分類研究 [10], [11] は，多種多様なキーワードを人手で分類するには非常にコストがかかることを理由に，自動分類手法を検討していた．一方，Wikipedia[*7]やはてなキーワード[*8]など，ユーザ参加型のオンライン辞書サービスが提供されており，そこには多種多様な見出し語（キーワード）が登録されている．本節では，トレンドキーワードがどのタイミングでオンライン辞書に登録されているのかを調査する．

トレンド日まで（当日は含まない）にオンライン辞書に登録されているトレンドキーワードの割合を**表 1** に示す．Wikipedia においてトレンド日よりも前に登録されているトレンドキーワードは 27.9%であり，同様にはてなキーワードでは 25.4%であった．トレンドキーワードの収集元別に見ると，Google から収集したトレンドキーワードはオンライン辞書双方ともに 60%程度登録されているものの，Yahoo! や Bing から収集したキーワードの大半が登録されていない．Yahoo! および Bing に関しては，前節で述べた

*3　http://www.google.co.jp/trends/hottrends (cited 2016–01–25)
*4　http://searchranking.yahoo.co.jp/burst_ranking/ (cited 2016–01–25)
*5　http://www.bing.com/ (cited 2016–01–25)
*6　http://doi.org/10.5281/zenodo.45056
*7　https://ja.wikipedia.org/ (cited 2016–01–25)
*8　http://d.hatena.ne.jp/keyword/ (cited 2016–01–25)

表 1　トレンド日までの辞書登録状況
Table 1　Percentage of trend queries having been registered before their emerging days.

	Wikipedia	はてなキーワード
Google	63.1%	59.9%
Yahoo!	26.1%	21.8%
Bing	8.3%	9.0%
マイクロ平均	27.9%	25.4%

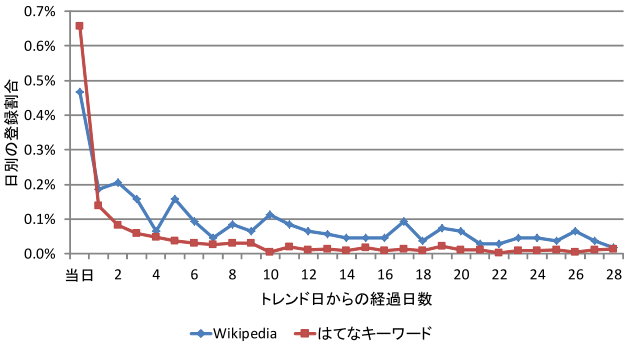


図 4　トレンド日以降の辞書登録状況
Fig. 4　The distribution of days taken for a trend query to get registered into an online dictionary.

ような Google とは異なる基準によってトレンドキーワードを提供していることが，ここでも示唆されている．なお，Wikipedia に登録されているか否かの調査では，照合時点のリダイレクトや曖昧性回避のページも照合対象としており，たとえば照合時点で「環太平洋パートナーシップ協定」から「環太平洋戦略的経済連携協定」へのリダイレクトが存在すれば，「環太平洋パートナーシップ協定」も辞書に登録されているとみなす．

トレンド日以降，4 週間（28 日間）に登録されたトレンドキーワードの推移を**図 4** に示す．横軸はトレンド日からの経過日数，縦軸は辞書に登録されたトレンドキーワードの割合である．トレンド日より 4 週間以内に Wikipedia に登録されたトレンドキーワードは 2.57%，はてなキーワードにおいては 1.36%であり，Wikipedia のコミュニティの方がトレンドに敏感であるものの，トレンドキーワードの大半はオンライン辞書に登録されていない．このことから，トレンドキーワードに対して情報を付与する際には，オンライン辞書サービスの情報に頼るだけでは不十分であり，自動推定手法の適用を検討する必要があるといえる．

5.　検索エンジンとの関係

5.1　トレンドの継続状況

トレンドキーワードは各検索サービスにおいて突発的に検索されるようになったキーワードである．本節ではトレンドキーワードがどれほどの期間，集中的に検索され続けるのかを調査する．検索頻度データが公開されていないため，本調査では，Google トレンドの検索ボリュームデータを検索頻度データとみなした．Google トレンドは高検索頻度キーワードに関し，調査期間の最大検索頻度日を 100 とした 100 段階による頻度データを提供している．今回，トレンド日を中心とする 29 日間（前後 14 日間）の頻度データを取得したところ，9,009 キーワード（25.6%）に関して頻度データを取得することができた．トレンド日の検索頻度に対し，頻度が $\{50, 30, 20\}$%未満になるまでの日数をトレンドが継続した日数であるとみなす．なお，トレンド日前から徐々に頻度が上昇していることも考えられるため，継続日数はトレンド日前も合算する．

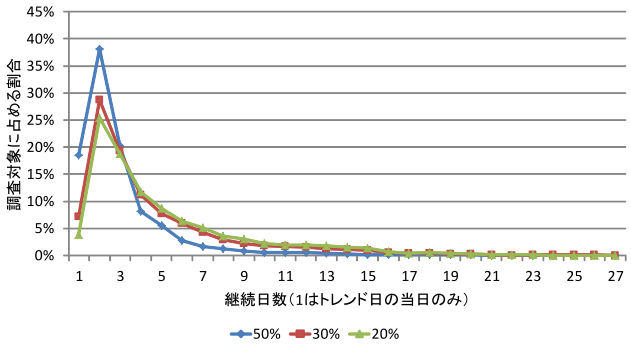


図 5　検索トレンドの継続日数
Fig. 5　The distribution of query's trend period.

継続日数の計算にあたり，トレンド日を中心とする 29 日間（前後 14 日間）にトレンドの開始と終了を観測できたキーワード数を調査した．つまり，トレンド日よりも前 14 日間に頻度が $\{50, 30, 20\}$%未満になる日があり，かつ，トレンド日よりも後 14 日間に頻度が $\{50, 30, 20\}$%未満になる日があるキーワード数を調査した．その結果，頻度 50%では 7,346 キーワード，30%では 6,658 キーワード，20%では 6,002 キーワードに関してトレンドの開始と終了を観測できた．

図 5 にトレンドが継続した日数の分布を頻度条件ごとに示す．横軸はトレンド日の当日を含む継続日数（1 の場合はトレンド日当日のみを表す），縦軸はその継続日数を有するキーワードの割合である．割合の算出にあたっては，先のトレンドの開始と終了を観測できたキーワード数を分母としている．図 5 のグラフから分かるとおり，いずれの頻度条件においても，トレンドの継続日数が 3 日以下のキーワードが過半数を占め，トレンド日を含めた 3 日間に集中的に検索されていることが分かる．最頻値に着目すれば，いずれの頻度条件においても，トレンド日を含めた 2 日間で収束する傾向があることも分かる．このことから，トレンドキーワードを検索するユーザに対するコンテンツの準備はリアルタイムに行う必要があると示唆される．

5.2　検索結果の変化

トレンドキーワードを分類する素性として，検索結果に含まれるページのコンテンツが効果的であると報告されて

表 2　検索種別ごとの共通出現単語統計量（トレンド日時と 240 時間ごとの共通単語数）
Table 2　The statistics of common terms between the trend day and 240 hours after for different search functions.

n	検索種別	平均値	中央値	最頻値*	最小値	最大値	標準偏差
100	ウェブ検索	70.85	73	73	0	100	16.00
	ブログ検索	53.96	53	53	0	100	15.21
	ニュース検索	58.97	60	68	0	100	16.92
	フレッシュ検索	35.17	35	36	0	98	12.06
1000	ウェブ検索	670.45	677	631	0	1000	160.88
	ブログ検索	415.59	385	347, 376	0	1000	143.99
	ニュース検索	440.97	430	413	0	1000	135.97
	フレッシュ検索	321.43	317	305	6	947	86.70

*最も頻度の高い値が複数ある場合，それらのすべてを記述している．

いる [11]．その報告では，検索結果ページが時間経過により変化する，具体的には出現する URL の残存率が 7 割程度であることを報告しているものの，コンテンツの内容がどのように変化しているかは明らかではなかった．本節では，コンテンツの内容変化に関する調査を行う．

検索結果のデータは，トレンドキーワードを収集した時点（トレンド日時）で，ウェブ検索，ブログ検索，ニュース検索を利用して Google，Yahoo! JAPAN，Bing からそれぞれ上位 10 件の検索結果 URL を取得し，その URL のコンテンツ[*9]をそれぞれ取得した．ただし，Bing はブログ検索を提供していないため，ブログ検索には Bing の結果は含まれない．また，ウェブ検索は適合順，ブログ検索およびニュース検索は日付順で取得した．Google に関してはフレッシュ検索（ウェブ検索を日付順で取得する機能）の結果も取得した．同様に，収集時点からの 48 時間後（2 日後），96 時間後（4 日後），…，240 時間後（10 日後）の検索結果も取得した．なお，特段の断りがない限り，それぞれの検索事業者の検索結果を区別せずに，ウェブ検索，ブログ検索，ニュース検索，フレッシュ検索を取り扱うものとする．そのため，それぞれの検索結果集合には，重複した検索結果ページが含まれる場合もある．

本稿では，収集対象トレンドキーワード集合を K（$|K| = 35192$），トレンドキーワード k に関するウェブ検索結果集合を S_k として次のように表現する．

$$S_k = \{S_k^0, S_k^{48}, S_k^{96}, ..., S_k^{240}\}$$

ここで，S_k^t はトレンド日時から t 時間経過したときに検索したトレンドキーワード k に関するウェブ検索結果集合とする（$t = 0$ はトレンド日時を表す）．すべてのウェブ検索結果集合 $\{S_k|k \in K\}$ に出現した単語集合 $W = \{w_1, w_2, w_3, ..., w_L\}$（$L$ はタイプ数）をもとに，S_k^t を次のような bag-of-words モデルによってベクトル $\boldsymbol{V_k^t}$ として表現する．

$$\boldsymbol{V_k^t} = (C_k^t(w_1), C_k^t(w_2), C_k^t(w_3), ..., C_k^t(w_L))$$

ここで，$C_k^t(w_i)$ は単語 w_i がウェブ検索結果集合 S_k^t に出現した回数とする．ブログ検索，ニュース検索，フレッシュ検索も同様とする．

まず，トレンド日時と 240 時間後（10 日後）とで共通して出現する単語を調べた．ここでは，各トレンドキーワードに対する検索結果集合の中から，頻出 n 単語を調査する．つまり，トレンドキーワード k に関する $\{C_k^t(w)|w \in W\}$ の上位 n 単語を抽出し，$t = \{0, 240\}$ の各々で共通する単語数を計算する．この計算をトレンドキーワード集合 K に含まれるキーワードすべてに対して行い，それらの統計量を計算する．$n = \{100, 1000\}$ としたときの統計量を**表 2** に示す．ウェブ検索においては平均的にそれぞれ 71 単語，670 単語が共通している一方，フレッシュ検索においてはそれぞれ 35 単語，321 単語と小さくなっており，検索種別によって共通する単語数に大きな差があることが分かる．

次に，検索結果集合のベクトル $\boldsymbol{V_k^t}$ を用い，トレンド日時（$t = 0$）と任意の経過時間（$t = \{48, 96, ..., 240\}$）との検索結果の類似性をコサイン類似度によってはかる．t 時間後におけるトレンドキーワード k に関する類似度は次のように計算できる．

$$Sim(\boldsymbol{V_k^0}, \boldsymbol{V_k^t}) = \frac{\boldsymbol{V_k^0} \cdot \boldsymbol{V_k^t}}{||\boldsymbol{V_k^0}||||\boldsymbol{V_k^t}||}$$

この計算をトレンドキーワード集合 K に含まれるキーワードすべてに対して行い，それらの統計量を計算する．**表 3** は計算した統計量，**図 6** は統計量のうち平均値の変動を示したものである．48 時間後，96 時間後，…，240 時間後と時間が経過するに従って類似性は下がるものの，240 時間後においてもウェブ検索では 0.84，ニュース検索では 0.74 であり，依然として高い類似性を保つ検索種別も存在する．

以上より，収集時点から徐々に検索結果の内容が変動するものの，その変動幅は検索種別によって大きく異なることが分かった．フレッシュ検索に関しては 48 時間を経過した時点で類似性が比較的低下しているものの，ウェブ検

*9 HTML ファイルから抽出したテキストノード（HTML タグ以外の部分）を指す．

表 3　検索種別ごとのコサイン類似度の詳細
Table 3　Details of cosine similarities after the trend day for different search functions.

t	検索種別	平均値	中央値	最頻値*	最小値	最大値	標準偏差
48	ウェブ検索	0.90	0.95	1.00	0.00	1.00	0.13
	ブログ検索	0.75	0.77	1.00	0.00	1.00	0.18
	ニュース検索	0.83	0.87	0.92	0.01	1.00	0.13
	フレッシュ検索	0.51	0.52	0.54, 0.56	0.01	1.00	0.17
96	ウェブ検索	0.87	0.93	0.99	0.00	1.00	0.15
	ブログ検索	0.70	0.72	0.73	0.01	1.00	0.18
	ニュース検索	0.80	0.83	0.91	0.01	1.00	0.14
	フレッシュ検索	0.49	0.50	0.54	0.01	1.00	0.17
144	ウェブ検索	0.86	0.91	0.99	0.00	1.00	0.16
	ブログ検索	0.68	0.69	0.72	0.00	1.00	0.18
	ニュース検索	0.77	0.81	0.89	0.01	1.00	0.15
	フレッシュ検索	0.48	0.48	0.47	0.02	1.00	0.16
192	ウェブ検索	0.85	0.90	0.99	0.00	1.00	0.16
	ブログ検索	0.66	0.68	0.68	0.00	1.00	0.18
	ニュース検索	0.75	0.79	0.87	0.00	1.00	0.15
	フレッシュ検索	0.47	0.47	0.47	0.01	1.00	0.16
240	ウェブ検索	0.84	0.89	0.99	0.00	1.00	0.16
	ブログ検索	0.65	0.67	0.66	0.00	1.00	0.18
	ニュース検索	0.74	0.77	0.87	0.00	1.00	0.16
	フレッシュ検索	0.46	0.47	0.43	0.01	1.00	0.16

*最も頻度の高い値が複数ある場合，それらのすべてを記述している．

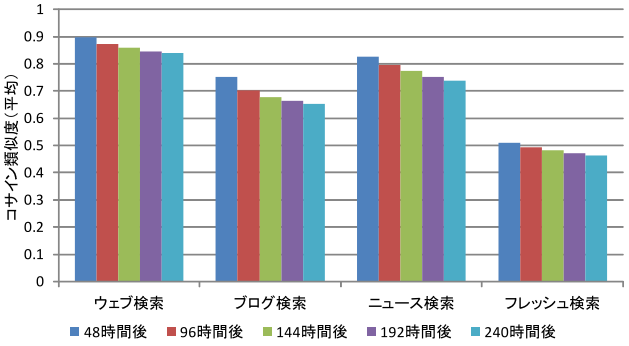


図 6　検索種別ごとのコサイン類似度の変動
Fig. 6　Average cosine similarities after the trend day for different search functions.

索およびニュース検索に関しては 240 時間を経過しても依然として高い類似性を保っている．そのため，トレンド日における状況を分析するにしても，ウェブ検索やニュース検索を利用する場合，必ずしもトレンド日当日の検索結果データを準備する必要はないと推察される．なお，本稿では，基礎的な分析を行う観点から，単語の出現頻度を用いて類似性を調査している．そのためトレンドキーワードとは関係のない一般的な語が類似性の調査に影響している可能性がある．このような一般的な語を除外した調査は今後の課題とする．

6.　ウェブリソースでの出現状況

6.1　利用可能な言語リソース

4.2 節で述べたように，トレンドキーワードに情報を付与するにはオンライン辞書サービスの情報を参照するだけでは不十分である．そのため，自動的に情報を付与することとなり，その素性を抽出するためにウェブのリソースが用いられる．しかしながら，5.1 節で示したように，トレンドキーワードは集中的に検索されはじめてから 2 日程度で 50%未満の頻度に収束する傾向がある．そのため，同様に即時性のあるウェブのリソースでなければ，トレンドキーワードの出現をとらえ，その情報付与に寄与するのは困難であると考えられる．本節では，トレンド日当日において各言語リソースにそのトレンドキーワードが一度以上出現したか否かを調査することで，どの程度の言語リソースがトレンド日当日に利用可能であるかを明らかにする．対象とする言語リソースは，Twitter[*10]，2 ちゃんねるのスレッドタイトル[*11]，ニュース記事[*12]の 3 リソースである．

トレンド日当日の言語リソースにトレンドキーワードが出現するかどうかを調査し，トレンドキーワードに対する言語リソースの利用可能性を検証した．その結果，Twitter に関しては 33,470 キーワード（95.1%），2 ちゃんねるに関しては 9,268 キーワード（26.3%），ニュースに関しては 13,390 キーワード（38.0%）がトレンド日当日に利用可能であった．基本的な統計量を**表 4** に示す．統計量の算出

[*10] 2015 年 6 月に一括して収集したため，ツイートの一部が削除されるなどにより，リアルタイムに収集した場合よりも少なくなっている可能性がある．
[*11] Ceek.jp Open Data (http://open.ceek.jp/) が提供する「2 ちゃんねる掲示板のスレッドデータ」を対象とする．
[*12] ニュースポータルサイト Ceek.jp News (http://news.ceek.jp/) が収集したニュース記事を対象とする．

表 4　言語リソースごとの出現統計量
Table 4　Appearance of trend queries in different language resources on the web.

	対象数	平均値	中央値	最頻値	最小値	最大値	標準偏差
Twitter	33,470	665.2	19	3	1	1104596	10344.2
2 ちゃんねる	9,268	11.2	2	1	1	1962	42.3
ニュース	13,390	17.1	4	1	1	1986	51.2

表 5　ウェブリソースごとの対象数ならびに平均継続日数および標準偏差
Table 5　Average days of trend for different frequency thresholds.

		頻度条件 50%		頻度条件 30%		頻度条件 20%	
	収集数	対象数	継続日数	対象数	継続日数	対象数	継続日数
Wikipedia	10,690	10,310	2.57 (2.15)	9,764	3.33 (2.63)	9,169	4.06 (3.09)
はてなキーワード	7,999	7,204	3.64 (3.40)	6,427	5.01 (4.11)	5,599	6.10 (4.54)
Twitter	30,011	28,264	3.00 (2.43)	27,014	3.65 (2.89)	25,803	4.21 (3.17)
2 ちゃんねる	1,114	1,015	3.76 (3.45)	939	4.90 (3.87)	851	5.81 (4.31)
はてなダイアリー	1,392	1,178	4.30 (3.78)	1,027	5.73 (4.32)	856	6.82 (4.52)
ニュース	2,438	2,303	3.72 (3.17)	2,199	5.04 (3.83)	2,083	6.09 (4.17)
検索頻度	9,009	7,346	3.52 (3.02)	6,658	4.59 (3.82)	6,002	5.15 (4.01)

に関して，利用可能なキーワードにのみ限定し，平均的にどの程度言及されているかを算出した．表 4 のとおり，いずれのリソースにおいても平均値と最頻値との間に大きな乖離がある．たとえば，Twitter においては平均値は 665.2 であるのに対し，最頻値は 3 であり，その差は大きい．つまり，一部のトレンドキーワードに対して著しく言及されるケースがあるものの，大半のトレンドキーワードに対してはほとんど言及されていないと考えられる．このことから，集中的に検索されるキーワードであっても，そのことに対して必ずしもウェブで言及されているとはいえず，検索行動をウェブの言語リソースからとらえることは困難であることが示唆される．

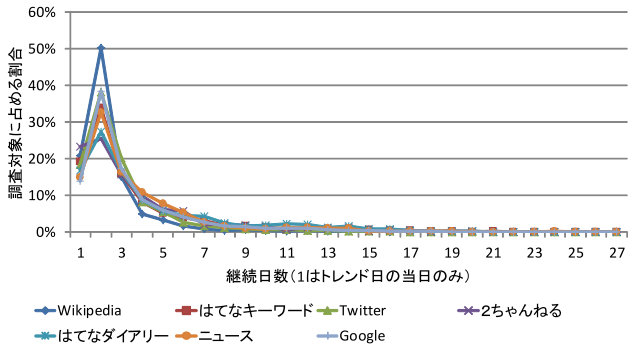
6.2　頻度パターンの相違

前節ではトレンド日に言語リソースがどの程度利用可能であるかを調査したが，本節では言語リソースに限定せず，頻度パターンの相違に着目して調査を行う．たとえば，Wikipedia のページビューデータは言語リソースではないものの，キーワード（見出し語）の注目度を過去に遡って調査できることから，トレンドを理解するための有益なリソースになりうる．

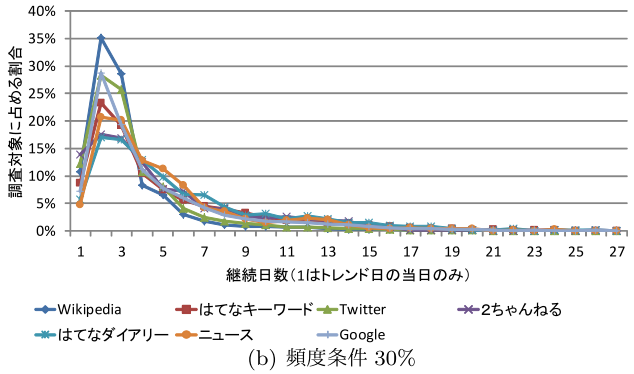
まず，5.1 節と同様にリソース別の頻度パターンにおいて，どの程度，トレンドが継続するのかを調査する．この際，トレンドが継続した日数を，トレンド日の頻度に対して頻度が $\{50, 30, 20\}$%未満になるまでの日数であるとして，それぞれ調査した．対象とする頻度データは，Wikipedia のページビュー（Wikipedia），はてなキーワードのページビュー（はてなキーワード），Twitter での言及量（Twitter）[*13]，2 ちゃんねるスレッドタイトルでの言及量（2 ちゃんねる），はてなダイアリーでの言及量（はてなダイアリー），ニュースでの言及量（ニュース）の 6 リソースである．トレンド日を中心とする 29 日間（前後 14 日間）の頻度データ（言及量データ）を取得し，その期間での言及量（ページビューも言及量とみなす）の合計が 100 を超えるものを調査対象とした．調査にあたり，比較対象として Google トレンドによる検索頻度も用意した．調査対象キーワード数およびトレンド継続日数の平均値を**表 5** に示す[*14]．頻度データを準備できたトレンドキーワードに限定し（表中の「収集数」にそのキーワード数を示す），さらに，5.1 節と同様にトレンド日をトレンド日を中心とする 29 日間（前後 14 日間）にトレンドの開始と終了を観測できたキーワードに限定しているため，表中の「対象数」は頻度条件およびリソースごとに異なる．2 ちゃんねるやはてなダイアリーのようなユーザ作成型のコンテンツでの言及は，トレンドの継続日数が長い傾向がある．ただし，同じユーザ作成型のコンテンツであっても，Twitter では検索頻度よりもトレンドの継続日数が短い傾向にあった．なお，表 5 に示した平均値の差については，ウェルチの t 検定により有意水準 1%で有意差を確認した．**図 7** にウェブリソースごとのトレンド継続日数の日別分布を頻度条件別に示す．頻度条件 50%においては，いずれのリソースもトレンド日当日を含めた 2 日間でトレンドが収束し，頻度条件 30%，20%においても 4 日間程度で収束していることが分かる．

次に，ウェブリソースの即時性に着目した調査を行う．調査対象期間（トレンド日を中心とする 29 日間）の全言

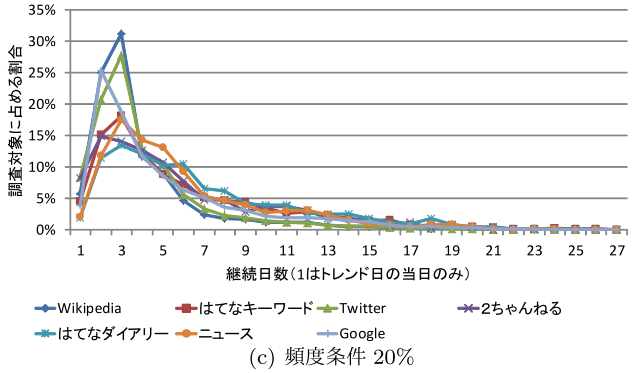
[*13] Twitter 分析サービスである Topsy (http://topsy.com/) が提供する Twitter 言及量を収集した．なお，Topsy は 2015 年 12 月にサービスが終了した．
[*14] 6.1 節では 1 回以上の出現のあるトレンドキーワードを対象としたが，本節では調査期間中に 100 回以上の出現のあるトレンドキーワードを対象としたため，同じリソースであっても表 5 の収集数および対象数と表 4 の対象数は異なる．



(a) 頻度条件 50%



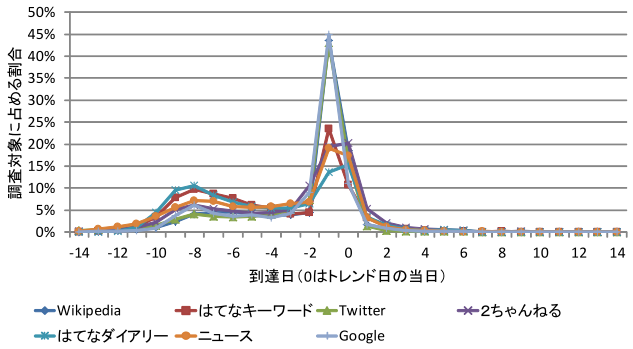
(b) 頻度条件 30%



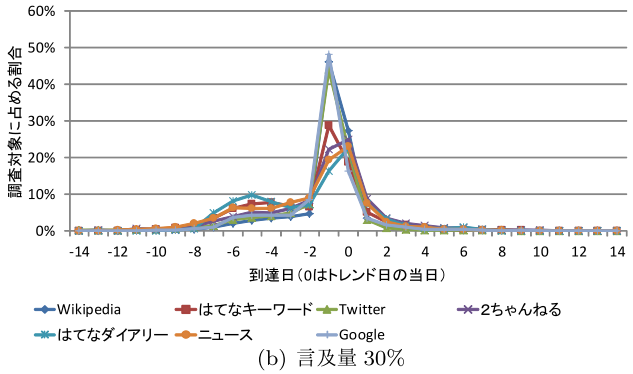
(c) 頻度条件 20%

図 7　ウェブリソースごとのトレンド継続日分布

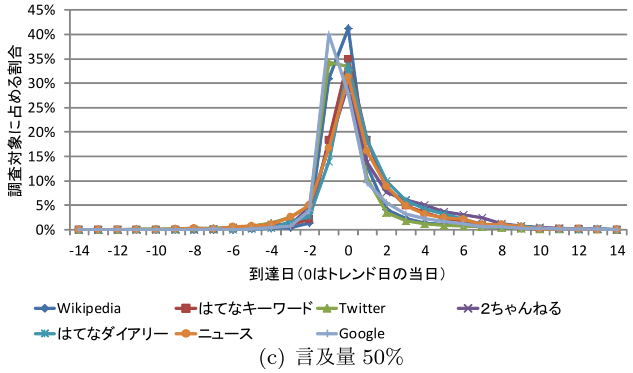
Fig. 7　The distribution of trend period for different web resources.



(a) 言及量 20%



(b) 言及量 30%



(c) 言及量 50%

図 8　ウェブリソースごとの即時性分布

Fig. 8　The distribution of days taken to reach specific ratios of total frequencies.

表 6　ウェブリソースごとの言及即時性（平均到達日および標準偏差）
Table 6　Average days taken till a trend query appears a sufficient number of times in a web resource.

	言及量 20%	言及量 30%	言及量 50%
Wikipedia	−2.28 (9.05)	−1.11 (2.18)	0.25 (1.95)
はてなキーワード	−3.95 (13.89)	−1.90 (2.85)	0.75 (2.22)
Twitter	−2.49 (10.04)	−1.53 (2.52)	−0.20 (2.32)
2 ちゃんねる	−2.77 (13.67)	−1.22 (3.02)	1.10 (2.81)
はてなダイアリー	−4.07 (14.36)	−1.87 (2.99)	1.04 (2.32)
ニュース	−3.64 (14.29)	−1.90 (3.08)	0.62 (2.61)
検索頻度	−2.53 (9.49)	−1.39 (2.30)	0.20 (2.19)

及量に対し，言及量が $\{20, 30, 50\}$%に達した日を「すでに十分に言及された日」とし，どのウェブリソースが最も早くその日に達したかを調査した．**表 6** にそれぞれの言及量に達した，トレンド日からの経過日数の平均値と標準偏差（括弧書き）を示す．1 であればトレンド日よりも 1 日後に特定の言及量に達したことを示し，−1 であればトレンド日よりも 1 日前に特定の言及量に達したことを示す．この調査により，Twitter での言及に関し，平均してトレンド日よりも前に言及量 50%に達するなど，言及の即時性が明らかになった．一方，はてなダイアリーのように一定の分量を記述すると考えられるユーザ作成型のコンテンツでは，そのコンテンツの準備までに時間を要するためか，トレンド日よりも 1 日程度遅れて言及量 50%に達することが分かった．なお，表 6 に示した平均値の差については，ウェルチの t 検定により有意水準 1%で有意差を確認した．**図 8** にウェブリソースごとの即時性（トレンド日を中心とする相対的な到達日）の日別分布を言及量別に示す．言及量 50%においては，いずれのリソースもトレンド日当日（グラフ中の 0）または前日（グラフ中の −1）に最頻値

表 7　ウェブリソース間の共通出現キーワード数
Table 7　The number of common queries found between web resources.

	Wikipedia	はてなキーワード	Twitter	2 ちゃんねる	はてなダイアリー	ニュース
はてなキーワード	8163					
Twitter	11045	9578				
2 ちゃんねる	8050	7341	18251			
はてなダイアリー	7752	8954	9018	7198		
ニュース	8274	7380	22510	14953	7098	
検索頻度	4851	4320	9105	6608	4239	6721

表 8　ウェブリソース間の類似性（平均値および標準偏差）
Table 8　Average and standard deviation of (a) correlation coefficients and (b) UDCRs between web resources.

(a) 相関係数

	Wikipedia	はてなキーワード	Twitter	2 ちゃんねる	はてなダイアリー	ニュース
はてなキーワード	0.59 (0.35)					
Twitter	0.72 (0.30)	0.51 (0.35)				
2 ちゃんねる	0.59 (0.36)	0.42 (0.36)	0.68 (0.33)			
はてなダイアリー	0.55 (0.34)	0.42 (0.34)	0.57 (0.34)	0.49 (0.36)		
ニュース	0.54 (0.38)	0.39 (0.37)	0.62 (0.36)	0.57 (0.37)	0.46 (0.37)	
検索頻度	0.76 (0.27)	0.51 (0.35)	0.73 (0.28)	0.65 (0.33)	0.55 (0.33)	0.55 (0.37)

(b) 増減一致率

	Wikipedia	はてなキーワード	Twitter	2 ちゃんねる	はてなダイアリー	ニュース
はてなキーワード	0.51 (0.15)					
Twitter	0.55 (0.15)	0.47 (0.15)				
2 ちゃんねる	0.25 (0.18)	0.30 (0.18)	0.33 (0.21)			
はてなダイアリー	0.32 (0.18)	0.36 (0.18)	0.35 (0.17)	0.58 (0.20)		
ニュース	0.30 (0.19)	0.33 (0.18)	0.42 (0.23)	0.70 (0.22)	0.58 (0.20)	
検索頻度	0.45 (0.20)	0.41 (0.16)	0.47 (0.19)	0.59 (0.24)	0.53 (0.19)	0.58 (0.24)

があることが分かる．言及量 20%または 30%においては，トレンド日よりも 1 週間ほど前になだらかな山ができており，一部の早期言及者による言及の表れである可能性がある．しかしながら，今回の調査では単純なキーワード出現（ページビューに関してはアクセス）を「言及」とみなしており，トレンド日における言及とのトピックの差異は考慮されていない．そのため，異なるトピックで言及されている可能性もあり，このようなトピックを考慮した言及量の調査は今後の課題とする．

最後に，頻度パターンの類似性に着目した調査を行う．高頻度キーワードに関し，Google トレンドによる検索頻度と Wikipedia ページビューとの相関が認められている [9]．本調査では，Wikipedia ページビュー以外の頻度データも対象とし，同様の調査を行った．ここでは，先行研究 [9] と同様，類似性の評価にピアソンの積率相関係数（以下，単に「相関係数」という）と増減一致率を用いる．増減一致率は相関係数ではとらえることが困難である上昇および減少の変動を評価する指標である．2 つの時系列データ $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ および $\boldsymbol{Y} = \{y_1, y_2, ..., y_n\}$ が与えられたとき，相関係数（$Correlation$）および増減一致率（$UDCR$）を次のように計算する．

$$Correlation(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$UDCR(\boldsymbol{X}, \boldsymbol{Y}) = \frac{||\{t \in \{2, 3, ..., n\} \mid x'_t = y'_t\}||}{n - 1}$$

$$x'_t = \begin{cases} 1 & (x_t - x_{t-1} > 0) \\ 0 & (x_t - x_{t-1} = 0) \\ -1 & (x_t - x_{t-1} < 0) \end{cases}$$

$\bar{Y}$ に関しても $\bar{X}$ と同様に計算し，また y'_t に関しても x'_t と同様に計算する．

表 7 にリソース間で共通に出現したトレンドキーワード数を示す．今回，調査対象となる 2 つのリソースに共通して出現するトレンドキーワードのみを類似性を計算する対

象とし，どちらか一方にしか出現していないトレンドキーワードに関しては類似性を計算していないため，キーワード数にはばらつきが生じている[*15]．**表 8** にリソース間の平均相関係数および平均増減一致率を示す．表中の値は，各トレンドキーワードについてリソース間における相関係数および増減一致率を計算し，それぞれの平均値と標準偏差（括弧書き）を計算したものである．相関係数に関し，先行研究 [9] のとおり検索頻度と Wikipedia ページビューとの間に高い相関（0.76）が認められるほか，Twitter の言及量とも同等の相関が認められる．増減一致率をみると，2ちゃんねるとニュースとの間に高い類似性が認められるものの，実データを観察すると，大半の日の頻度が 0 になっており，横ばいが評価された結果，高い数値が出ていることが分かった．表 8 をもとにリソース間の類似性をみると，Wikipedia と Twitter のように一部のリソース間に一定の類似性が見受けられるものの，多くのリソース間ではほとんど類似していないことが分かる．頻度パターンに関しては，一部のリソースは別のリソースで代替できる可能性があるものの，ほとんどのリソースはそれぞれ異なる特性をとらえており，代替が難しいと示唆される．

7. おわりに

本研究では，検索エンジンにおいて集中的に検索されたトレンドキーワードに対し，それらがウェブリソースでどのように出現するかの分析を行った．この分析により，大半のトレンドキーワードが Wikipedia などのオンライン辞書サービスに登録されていないこと，検索のトレンドは 2 日で 50%未満の頻度に収束すること，ソーシャルメディア（Twitter）がほかのウェブリソースよりもトレンドに敏感であることなどを明らかにした．一方で，トレンド日当日に利用可能なウェブリソースは依然として少量であり，従来より用いられている検索結果データを利用することが好ましいことも示唆された．また，ウェブリソースごとの言及パターンに関しても，相関係数および増減一致率をもとに調査した限りでは，それぞれに差異があり，どれか 1 つのリソースでほかのリソースを代替することは困難であると示唆された．今後，本研究で得られた知見をふまえ，高度なトレンド分析システムを構築し，トレンドキーワードを取り扱う研究開発において，その方針を制定することを支援する．

参考文献

[1] Radinsky, K., Davidovich, S. and Markovitch, S.: Predicting the News of Tomorrow Using Patterns in Web Search Queries, *Proc. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol.1, pp.363–367 (2008).

[2] Choi, H. and Varian, H.: Predicting the Present with Google Trends, *Economic Record*, Vol.88, pp.2–9 (2012).

[3] 古川忠延，松尾　豊，大向一輝，内山幸樹，石塚　満：ブログ上での話題伝播に注目した重要語判別，知能と情報，Vol.21, No.4, pp.557–566 (2009).

[4] Benhardus, J. and Kalita, J.: Streaming trend detection in Twitter, *International Journal of Web Based Communities*, Vol.9, No.1, pp.122–139 (2013).

[5] Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company (1984).

[6] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.91–101 (2002).

[7] Huss, J.W., Lindenbaum, P., Martone, M., Roberts, D., Pizarro, A., Valafar, F., Hogenesch, J.B. and Su, A.I.: The Gene Wiki: community intelligence applied to human gene annotation, *Nucleic Acids Research*, Vol.38, No.suppl 1, pp.D633–D639 (2010).

[8] Kristoufek, L.: BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era, *Scientific Reports*, Vol.3, No.3415 (2013).

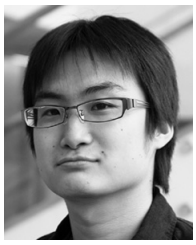
[9] 吉田光男，荒瀬由紀，角田孝昭，山本幹雄：検索頻度推定のための Wikipedia ページビューデータの分析，第 29 回人工知能学会全国大会 (2015).

[10] Yoshida, M. and Arase, Y.: Exploiting Twitter for Spiking Query Classification, *Proc. 8th Asia Information Retrieval Societies Conference*, pp.138–149 (2012).

[11] 吉田光男，荒瀬由紀：ラベル伝搬によるトレンドクエリのカテゴリ推定，人工知能学会論文誌，Vol.30, No.1, pp.161–171 (2015).

[12] Kulkarni, A., Teevan, J., Svore, K.M. and Dumais, S.T.: Understanding Temporal Query Dynamics, *Proc. 4th ACM International Conference on Web Search and Data Mining*, pp.167–176 (2011).

[*15] 先の 2 つの調査では調査対象期間において言及量が 100 を超えるものを調査対象としたが，本調査では 1 以上のすべてのトレンドキーワードを対象としたため，表 7 の値は表 5 および表 6 の「対象数」を超える場合がある．



吉田 光男　（正会員）

2009 年筑波大学第三学群情報学類卒業．2011 年同大学院システム情報工学研究科博士前期課程修了，2014 年同博士後期課程修了．博士（工学）．同年より豊橋技術科学大学大学院工学研究科（情報・知能工学系）助教．ウェブ工学，自然言語処理，計算社会科学に関する研究に従事．言語処理学会，人工知能学会，日本データベース学会各会員．

荒瀬 由紀　（正会員）

2006年大阪大学工学部電子情報エネルギー工学科卒業．2007年同大学院情報科学研究科博士前期課程修了，2010年同博士後期課程修了．博士（情報科学）．同年，Microsoft Research Asiaに入社し，Natural Language Computingグループ研究員となる．2014年より大阪大学大学院情報科学研究科准教授．言い換え表現抽出，統計的機械翻訳，ウェブデータマイニングに関する研究に従事．ACL，言語処理学会，日本データベース学会各会員．

（担当編集委員　張 建偉）