情報処理学会論文誌　データベース　Vol.6　No.1　1–15 (Jan. 2013)

推薦論文

大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法

中島 伸介[1,a)]　張 建偉[2]　稲垣 陽一[3]　中本 レン[3]

受付日 2012年6月20日, 採録日 2012年10月11日

概要：本研究では，小さいコミュニティから徐々に広がり，最終的に多くの人々に知れわたるような流行語を拡張型流行語と定義し，ブログ記事を時系列解析することで拡張型流行語の早期発見する手法に関する検討を行った．具体的には，すでにメジャーな流行語となったトピックに対し，ブログ上でどのように拡散していったのかを分析することで，早期発見に必要な分析手法について検討した．kizasi.jp で扱っている 3,776,154 ブログサイトで過去 2 年間に投稿された 81,922,977 件のブログ記事データの分析の結果，流行語候補がメジャーな流行語に発達する過程において，総発言数に占める，対象トピックと関連の深いコミュニティからの発言割合が減少しつつ，関連の薄いコミュニティからの発言割合が増加する状況を確認した．また，対象トピックと関連の深いコミュニティの特定手法を検討するとともに，総発言数に占めるこのコミュニティからの発言数の割合の減少状況について分析を行った．さらに，ライバル関係にある複数の流行語候補のランキングに基づく，提案手法の妥当性の検証を行った結果，良好な結果を得た．

キーワード：流行語発見，ブログ記事，時系列分析，ブロガコミュニティ

Early Detection of Gradual Buzzwords Based on Large-scale Time-series Analysis of Blog Entries

Shinsuke Nakajima[1,a)]　Jianwei Zhang[2]　Yoichi Inagaki[3]　Reyn Nakamoto[3]

Received: June 20, 2012, Accepted: October 11, 2012

Abstract: In this paper, we focus on “gradual buzzwords” that begin from a restricted community, spread little by little to other communities, and finally become widely known to most people, and discuss a method for their early detection by analyzing time-series data of blog entries. We observe the process in which certain topics grow to become major buzzwords and determine the key indicators that are necessary for their early detection. From the analysis results based on 81,922,977 blog entries from 3,776,154 blog websites posted in the past two years, we find that as topics grow to become major buzzwords, the percentages of blog entries from the blogger communities closely related to the target buzzword decrease gradually, and the percentages of blog entries from the weakly related blogger communities increase gradually. We also discuss how to identify the blogger communities which are closely related to these buzzwords, and conduct a slope analysis of percentage variation of blog entries from these closely related blogger communities. Moreover, we verify the effectiveness of the proposed method through experimentation that compares the rankings of several buzzword candidates with popularity competition.

Keywords: buzzword detection, blog entries, time-series analysis, blog community

[1] 京都産業大学
Kyoto Sangyo University, Kyoto 603–8047, Japan
[2] 筑波技術大学
Tsukuba University of Technology, Tsukuba, Ibaraki 305–8520, Japan
[3] 株式会社きざしカンパニー
kizasi Company, Inc., Chuo, Tokyo 103–0015, Japan
[a)] nakajima@cse.kyoto-su.ac.jp

1. はじめに

世間の流行語は，テレビや雑誌で紹介されるなど，世間

本論文の内容は 2011 年 11 月の WebDB2011 にて発表され，同シンポジウムプログラム委員会により情報処理学会論文誌データベースへの掲載が推薦された論文である．

に知れわたった後から知ることが多く，流行語の先駆けを発見することは容易でない．しかしながら，マーケティングの観点から見ても，有望な流行語候補を素早く検出することは重要といえる．そこで我々は流行語候補の早期発見に着目した．

近年，インターネットの普及にともない，Twitter などを含めたブログサービスが急激に普及している．ブログサービスは，マスメディアが発信する情報とは異なり，ユーザが自分の意思や，趣向，興味に基づいて，リアルタイムで発信される情報源である．すなわち，ブログコンテンツは，人々の関心がリアルタイムに反映されたコンテンツであるともいえる．事実，企業が自社製品の評判分析を行うために，ブログなどの分析を行っている例もあり，解析対象としてのブログコンテンツの価値が高まっているといえる．そこで我々は，このブログコンテンツを適切に分析することで世間に広まる前の流行語の先駆けを発見できる可能性があると考え，ブログ記事を時系列解析することで流行語候補を早期発見する手法に関する検討を行う．

ここで，本研究で扱う流行語のタイプについて説明する．流行語の生まれ方としてはいくつかのパターンが存在すると考えられる．1 つ目の流行語のタイプとして，「テレビ，ニュース，雑誌などで取り上げられたことで，いっせいに広がり，様々なコミュニティで一時的に話題になる流行語」を突発型流行語と呼ぶ．「小惑星探査機 “はやぶさ”」はこのタイプといえる．もう 1 つの流行語のタイプとして，「小さいコミュニティから徐々に広がり，最終的に多くの人々に知れわたるような流行語」を拡張型流行語と呼ぶ．「AKB48」や「女子会」はこのタイプといえる．2 タイプの発言者数の時間推移曲線を**図 1** に示す．これまで突発型流行語の抽出を対象とした研究 [1], [2], [3], [4], [5], [6] があるが，拡張型流行語に着目した研究はまだ少ない．本研究では，早期発見を目指す流行語のタイプとして 2 つ目の拡張型流行語を対象とする．

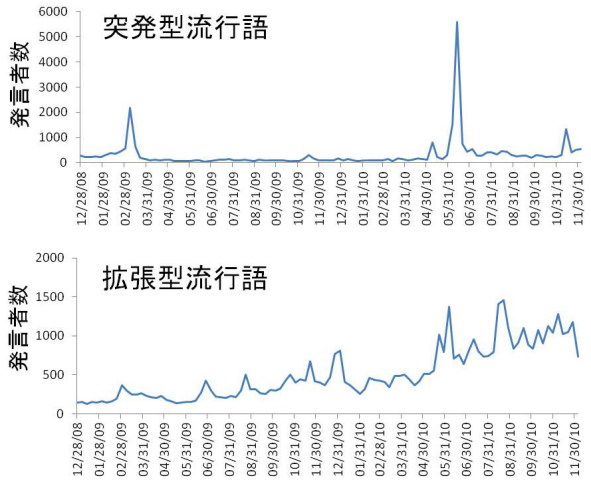

図 1　流行語のタイプ
Fig. 1　Buzzword types.

上記のような拡張型流行語を早期に発見する手法として，あるトピックに関して書かれたブログ記事投稿状況の時系列変化を分析することで，そのトピックがメジャーな流行語になりうるのかどうかを判定することを目指す．この際，ブログ記事数の増加のみに着目するだけではなく，コミュニティ間の話題の広がりに着目することにより，効率的に発見することができると考える．すなわち，あるトピックがメジャーな流行語に発達するかどうかを判定するためには，投稿数のみに着目するのではなく，投稿ブロガの幅の広がり（一部の限られたブロガから一般的なブロガへの広がり）に注目するべきであると考えている．

我々はこれまでの研究において，流行語候補の早期発見に関する基本的方針についてすでに提案 [7] を行っているが，コミュニティ間での流行語というよりは，年代別のブロガグループ間での流行語の伝搬に関する議論にとどまっている．したがって本稿では，趣味や興味に基づいたコミュニティ（ブロガグループ）間での話題の伝搬について，より詳細に分析し，コミュニティ間の話題の広がりの分析することに基づく流行語候補の早期発見手法について検討したので報告する．具体的には，すでに流行語大賞などでも取り上げられてメジャーな流行語となったトピックをいくつかピックアップし，これらのトピックがどのように世の中に拡散していったのかを大規模な実データを分析することで，早期に発見するために必要な分析手法について検討する．

コミュニティ（ブロガグループ）の特定方法としては，我々がすでに行っているブロガ分析に関する先行研究 [8] の成果を利用する．この研究では，ブロガの体験熟知度に基づくブログランキングシステムの開発を行っており，その中で，小分類として 11,090 領域，大分類として 122 グループのコミュニティを対象とし，そのコミュニティに属するブロガ判定を行っている．これらを暗黙的なコミュニティと考えることができるので，この 122 グループに対するコミュニティ間の話題の広がりを分析することが可能となる．

なお，世の中の嗜好や関心の時系列変化を分析するためにはある程度大規模なデータを確保する必要があるが，本研究では，kizasi.jp で扱っている，3,776,154 サイトで過去 2 年間に投稿された 81,922,977 件のブログ記事データを扱うことで，世の中の流行語に関する分析を可能にしている．

以下，2 章で関連研究について述べる．3 章で拡張型流行語の特徴分析について述べる．4 章でコミュニティの特定方法について述べる．5 章で拡張型流行語の早期発見を目指したブログ記事の時系列分析について述べる．6 章で流行語候補のランキングに関する検証実験を行い，提案手法の妥当性を示す．最後に 7 章でまとめと今後の課題について述べる．

2. 関連研究

流行語抽出は，バーストネス（burstness）の概念と深く関係している．Kleinberg [9] は無限状態のオートマトンを用いて活動のバースト（burst of activity）をモデル化した．状態遷移にコストを付与することによって，短いバーストを防ぎ，バーストの持続期間を識別できる．Yi [6] は Kleinberg のモデルを採用するうえ，バーストの勢い（momentum）および相対的な持続期間を考慮したアルゴリズムを提案した．Araujo ら [2] は Kleinberg の提案モデルに対して複数のコスト関数を検討し，状態の最良分布を調査した．Parikh ら [5] は大規模な電子商取引システムにおいて，ユーザクエリのバーストを検出する手法を提案した．Lappas ら [10] は指定したタームに対してバーストの時間間隔を識別できるような，パラメータフリーで線形の処理時間のアプローチを提案し，バーストの情報を検索プロセスに利用した．

従来技術として，流行語やトレンドを抽出するための実用システムやデモシステムがある [1], [3], [4]．Yahoo! Buzz Index [1] は検索ユーザ数の割合に基づいて検索語のスコアを算出し，leader（スコアの最大のもの）と mover（前日より検索ユーザの割合の増加が最大のもの）を抽出している．BlogPulse [3] はフレーズまたは人名の出現頻度の割合を計算することにより，ブログからキーフレーズやキーマンを抽出している．TwitterMonitor [4] はツイッターにおいて，流行語の発見とグルーピングによってトレンドの検出を提案した．これらの研究は，単語やフレーズの出現頻度，発言者の人数や，時間推移パターンを分析することで，突発型流行語の抽出を対象としたものである．一方，著者らの研究は，ブロガコミュニティ間で話題の広がりに着目し，拡張型流行語の抽出を目指すものである．

Kumar ら [11] はブログコミュニティ内のバースト活動を抽出する手法を提案した．Gruhl ら [12] はブログコミュニティ内で個人から個人への情報伝達を分析した．1 つのブログコミュニティ内でブロガ間の情報伝達ではなく，著者らはコミュニティから他のコミュニティへの広がりに着目している．また，これらの研究でのコミュニティはブログ間のリンクを用いて識別したものであるが，本研究ではブロガの興味や関心に基づいてブロガの潜在的なコミュニティを抽出する．

奥村 [13] は，ブログ記事中でのキーワードの出現頻度の推移を調べることで，そのキーワードが，いつ，どの程度広がったかを検出し提示するシステムを開発した．長谷川らは，類似度と新規性の両方を考慮した文書のクラスタリング手法 [14] を用いた，ニュース記事のトレンド可視化システム [15] を開発した．金澤ら [16] は，検索エンジンを用いて将来情報が含まれる文書を効率的に収集し文書中の将来情報を抽出するとともに，情報の信頼性に基づいてクエリに関する将来情報を集約しグラフを用いて可視化する方式を提案している．これらの研究ではユーザがトレンドを把握しやすいように可視化することを目的としているため，流行語そのものの発見という著者らの研究の目的と異なっている．

3. 拡張型流行語の特徴分析

本稿では，「ニュースや雑誌などで取り上げられいっせいに広がり，様々なコミュニティで一時的に話題になる突発的流行語」ではなく，「小さいコミュニティのみで語られていたものが徐々に別のコミュニティでも話題になり広まっていくような流行語」を拡張型流行語と呼び，このような流行語を早期に発見するために必要な分析手法について検討を行う．この拡張型流行語の特徴としては，単純に発言者数が増えること，また発言者の幅（たとえば世代）が拡大すること，があげられる．これらの特徴に関して，既知の拡張型流行語を対象とした分析を行った．

3.1 拡張型流行語の分析 1（発言者数の増加）

流行語の拡張という観点でいえば，発言者の増加は不可欠である．したがって，少なくとも発言者数の増加が見られなければ，拡張型流行語にはなりえない．**図 2** に，2008 年 12 月 28 日から 2010 年 12 月 5 日の期間中に「AKB48」について発言した発言者（ブロガ）数と年代別発言数割合の時間推移を，**図 3** に，同期間中に「女子会」について発言した発言者数と年代別発言数割合の時間推移を示す．

なお，各図の上部のグラフの横軸は時間，縦軸は各キーワードを発言したブロガ数である．「AKB48」「女子会」ともに，2010 年までに流行し一般的にはよく知られたキーワードである．図 2，図 3 の上グラフからも分かるとおり，これらのキーワードは 2008 年 12 月当初ではそれほど多くの発言者数はないが，徐々に発言者数が拡大している様子が確認できる．したがって，これらは典型的な拡大

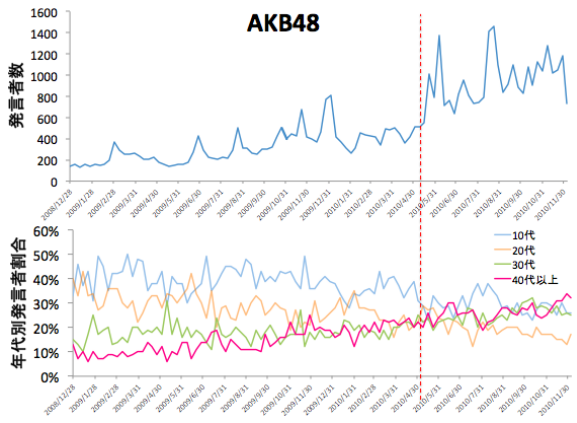

図 2　発言者数と年代別発言数割合の時間推移（「AKB48」の場合）
Fig. 2　Numbers of bloggers and percentages of bloggers in different age groups (Topic: AKB48).

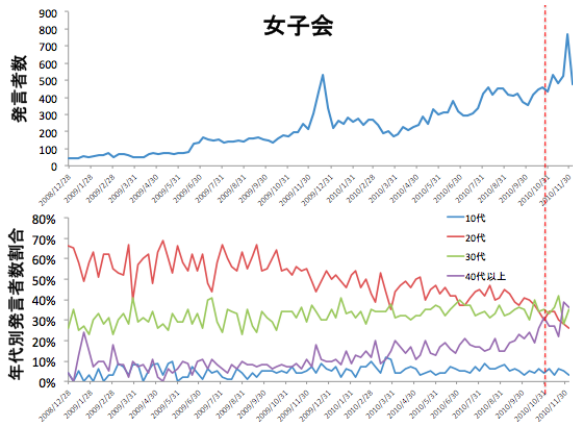

図 3　発言者数と年代別発言数割合の時間推移（「女子会」の場合）
Fig. 3　Numbers of bloggers and percentages of bloggers in different age groups (Topic: Joshikai).

型流行語であるといえる．これらを早期に発見するためには，この漸増状況を検出するという方法が考えられるが，この発言者数の時間推移をミクロに見れば，小さな増減を繰り返しているため，発言者数の増加のみから拡大型流行語を早期に発見することは容易ではない．また，たとえば「AKB48」に関する発言者数が増えたとしても，そのほとんどがアイドルファン（つまり，限られた趣向の人々）であれば，多くの国民から認知されるようなメジャーな流行語になるかどうかは不明である．したがって，次節で世代間の発言者の拡大に関して分析した結果を示す．

3.2　拡張型流行語の分析 2（世代間での発言者の拡大）

3.1 節で説明したとおり，発言者数の増加のみから拡大型流行語を早期に発見することは容易ではないため，コミュニティ間での発言者の拡大に注目した．なお，ここでいうコミュニティとしては，興味のあるキーワードに代表されるようなもの（たとえば，「政治」「ダイエット」「サッカー」「株式」など）はもちろんのこと，年代別や男女別，都道府県別なども適用可能と考えている．一部のコミュニティで話題になっていたものが，徐々に多くのコミュニティで話題になることは，拡張型流行語の典型的な特徴であると考えられる．すなわち，このような他のコミュニティへの伝播を検知することが，拡張型流行語の兆しを見つける際の重要であると考えられる．

図 2 および図 3 の下グラフではともに，10 代，20 代，30 代，40 代以上の 4 つのグループに分類している．そして，この 4 つのグループの合計を 100%とした際の各年代の割合を時系列でプロットしたものである．図 2 の下グラフにおいては，2008 年，2009 年は 10 代を中心に話題になっている．しかし，2010 年の 5 月下旬になると最も話題となっている 10 代に 20 代，30 代，40 代以上の年代の人々が追いついていることが分かる．「AKB48」は元々アキバ系アイドルであり，若い世代（10 代）を中心に認知されつつあったが，他の年代にまで認知されるようになってきたことが推測できる．図 3 の下グラフに関しては，2008 年，2009 年は 20 代を中心に話題になっているが，2010 年 10 月下旬を見ると，30 代，40 代以上の年代の人が 20 代に追いついていることが分かる．すなわち，新しい言葉であった「女子会」は若い世代を中心に認知されていたが，他の年代でも認知が広まっていった様子が推測できる．10 代にはほとんど広まっていないのは，そもそも「女子会」は飲み会的な要素があるために，10 代にとっては主要な話題にならなかったものと考えられる．これらの図から，ある特定の年代において支配的であったトピックが，他の世代に伝播していっていることが分かる．これが拡張型流行語の典型的な特徴であり，1 つのコミュニティのみで支配的であった流行語が，他の多くのコミュニティ間共通の流行語となるターニングポイントになるのではないかと考えている．

図 2，図 3 の上下のグラフより，拡張型流行語の分析 1（発言者数の増加），特徴 2（世代間での発言者の拡大）をふまえて，「AKB48」では 2010 年 5 月下旬ごろ，「女子会」は 2010 年 10 月下旬ごろが，あるコミュニティ限定の流行語から，より広いコミュニティでの流行語へ変化する際のターニングポイントであるといえる．確かに，それ以降の発言者数を見てみるとさらに増加しているようにも判断できる．

以上より，1 つのコミュニティのみで支配的であった流行語が，他の多くのコミュニティ間共通の流行語へと拡大する状況をいち早く検知することが，拡張型流行語を早期に発見するための鍵になると考えている．しかしながら，他のコミュニティへの拡大を分析するうえで，世代間での伝搬のみを解析するだけでは不十分である．そこで，ブロガの興味や関心を表す潜在的なコミュニティ間での流行語の拡大について分析を行う必要があると考えている．次節で，まずブロガの興味や関心を表すブロガの潜在的コミュニティの特定手法について述べる．

4.　ブロガの潜在的コミュニティの特定

本研究で扱う「潜在的コミュニティ」とは，ブロガの意識にかかわらず，同じトピックについて興味や関心を持つブロガ集合を指す．したがって，お互いに面識があるかどうかは問わない．すなわち，SNS などでユーザが明示的に登録するようなコミュニティとは異なる．

たとえば「女子会」を例にとると，「飲み会」や「ガールズトーク」などに興味があるブロガ集合のみで語られていたものが，「政治」や「自動車」や「サッカー」など「女子会」とは直接的な関係が薄いトピックに興味があるブロガ集合においても認知度が上がったとすれば，それはメジャーな流行語に近づいていると本研究では判断しようとしている．したがって，コミュニティ内での面識は考慮す

る必要はないということになる．

また，“明示的には「自動車」や「サッカー」コミュニティのみに登録しているが，潜在的には「飲み会」や「ガールズトーク」に興味があるブロガ” がいたとする．このブロガが「女子会」に関するブログ記事を投稿した場合，「女子会」と関係が薄いブロガにも「女子会」が認知され始めたと判定される可能性があるが，これは誤りである．したがって，やはり本研究では，明示的なコミュニティのみに限定して分析を行うのではなく，潜在的なコミュニティも含めた分析が必要となる．

本研究では，潜在的なコミュニティを特定するために，過去に投稿したブログ記事から，そのブロガの興味度の強さを分析することで，ブロガの潜在的コミュニティを特定する．なお，この潜在的コミュニティ特定手法は，著者らが過去に行った技術「ブロガの熟知度分析」に基づいている [8]．以下に本手法を簡単に説明する．

4.1　潜在的コミュニティおよび共起語辞書の作成

潜在的コミュニティ名は，ブログでよく言及されるトピックから抽出する．これらは，自動抽出により作成したものと，独自のシソーラスにより拡充したものと，2 つの部分からなる．2011 年 10 月 12 日現在で，小分類として 11,090 領域，大分類として 122 グループとしている．

次に，各潜在的コミュニティに対して，共起語辞書を構築する．各潜在的コミュニティに対して，一定期間内のブログエントリを対象とし，その潜在的コミュニティ名であるキーワードとの共起度が高い 400 個の語句を抽出する．なお，共起語の選定は半自動で行っている．共起語候補を自動的に抽出した後に，最終的にはすべての辞書に対して人間の目で確認し，共起語として不適切な語句を排除しているため，共起語辞書の選定精度は 100%と考えてよい．

共起語の選定に関しては，自らの生活体験を表すような語句を優先的に採用することで，実体験に即したブログエントリを記述するブロガを分類しやすくしている．具体的には，商用ブログの多い特定のカテゴリに対して，生活者の実体験を示す語彙を増やすために，共起条件を手動で追加した．たとえば，「温泉」のカテゴリに対して，「温泉＋行った」「温泉＋宿泊した」などの共起条件を用いることで，実際に温泉に行ったブロガが書いたエントリから実体験に即する語彙を抽出した．手動で設定した共起条件の数は，カテゴリごとに大きく異なるが，商用ブログやスパムブログの多い「温泉」カテゴリのケースでは，50 個の共起条件を設定している．

4.2　ブロガ興味度スコアの算出および潜在的コミュニティの特定

潜在的コミュニティに対する各ブロガの興味度スコアの算出方法を説明する．基本的な考え方としては，対象となる潜在的コミュニティに関連するトピックを含んだエントリの投稿数に基づいて算出する．なお，各ブロガは潜在的コミュニティごとに異なる複数の興味度スコアを有する．つまり，あるブロガが「経済」と「政治」に関する潜在的コミュニティに属する場合，このブロガは「経済」に関する興味度スコアと「政治」に関する興味度スコアを別々に有することになる．

ここで，対象潜在的コミュニティ g_i に対する，あるブログエントリ e_k の関連度スコアを $relevance_{g_i}(e_k)$ とすると，以下のように表すことができる．

$$relevance_{g_i}(e_k) = \sum_{j=1}^{n} \alpha_{ij} \cdot \beta_{ij} \cdot \gamma_{ij} \quad (1)$$

ただし，n はこの潜在的コミュニティ g_i の共起語数であり，今回は $n = 400$ である．α_{ij} は潜在的コミュニティ g_i の共起度順位 j 番目の共起語 w_{ij} の重みであり，$\alpha_{ij} = (n - j + 1)/n$ で表される．これは，各共起語の共起度以上に，共起順位の高い語句の重みを大きくするためのものであり，共起順位 1 位の重みは 400/400，2 位の重みは 399/400 となり，400 位の重みは 1/400 となる．β_{ij} は潜在的コミュニティ g_i の j 番目の共起語 w_{ij} の共起度である．そして，γ_{ij} は順位 j 番目の共起語 w_{ij} が当該エントリ e_k 内に存在するかどうかを表現する変数であり，存在する場合 1，存在しない場合 0 の値をとる．

共起度の算出法としては，単純頻度，t スコア，MI スコア，LogLog スコアなど多くの尺度が提案されている [17], [18]．単純頻度では，常識的な語を抽出するのに対して，特徴的な語を上位におく t スコアや MI スコアでは，納得できる語がなくなる傾向がこれまでに行った実験で見られた．そのため，本手法では，それらの中間の尺度 LogLog スコアを採用している．ブログエントリの総語数を N とし，キーワード x と周辺語 y の出現回数をそれぞれ N_x と N_y とする．x と y の共起回数を N_{xy} とすると，LogLog スコアの算出式は下記である．

$$LogLog\ score = \log \frac{N_{xy} * N}{N_x * N_y} * \log N_{xy} \quad (2)$$

次に，対象潜在的コミュニティ g_i に対するブロガ b の興味度スコアを $knowledge_{g_i}(b)$ とすると，以下のように表すことができる．

$$knowledge_{g_i}(b) = \frac{l}{n} \cdot \frac{\log(m)}{m} \cdot \sum_{k=1}^{m} relevance_{g_i}(e_k) \quad (3)$$

ただし，e_k はブロガ b が投稿したエントリである．m はブロガ b が対象期間内に投稿したエントリ数である．l はブロガ b が対象期間内に投稿したエントリに出現した共起語数である（$l \leq n$）．したがって，l/n はブロガ b が使用した共起語の全共起語に対する網羅率である．$\log(m)/m$ では，関連性の低いエントリを大量に投稿した場合に，その

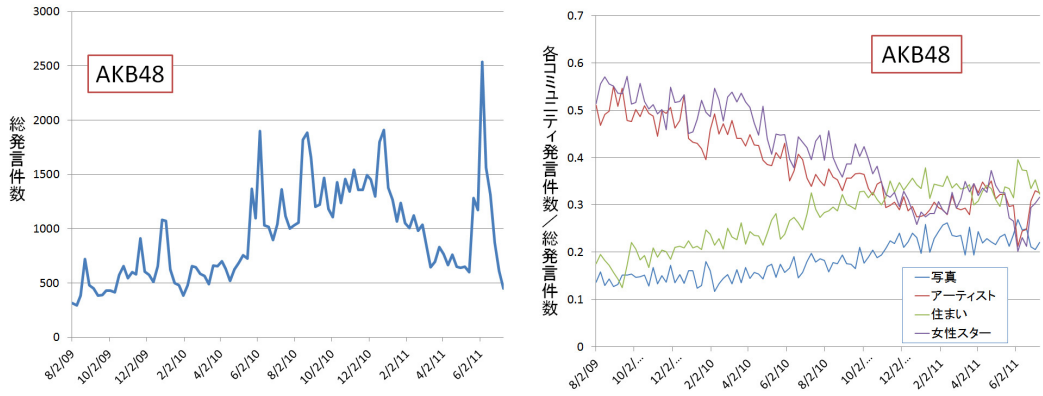

図 4　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（AKB48）
Fig. 4　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: AKB48).

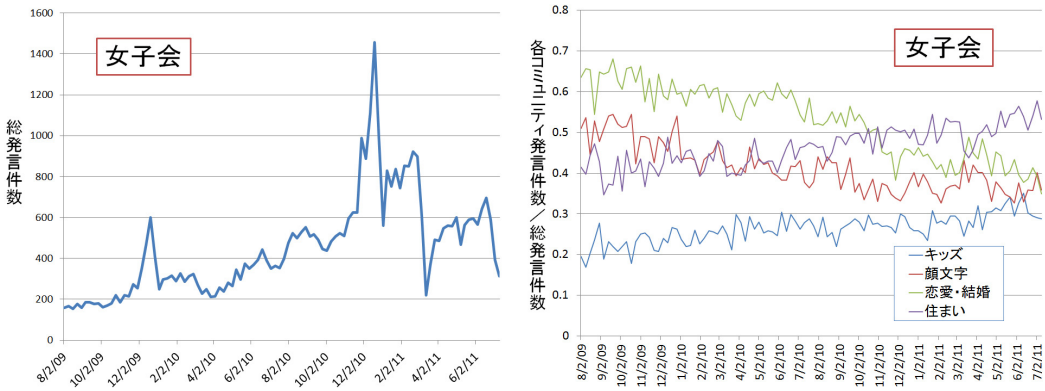

図 5　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（女子会）
Fig. 5　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: Joshikai).

ブロガの興味度が高くなってしまうという問題に対して，エントリ数の増加の影響を緩和させている．

最終的に対象となる潜在的コミュニティに対する興味度スコアが，設定した閾値を超えれば，その潜在的コミュニティはそのブロガが属するものと判定する．

5.　拡張型流行語の早期発見を目指したブログ記事の時系列分析

本章では，拡張型流行語の早期発見を目指したブログ記事の時系列分析を行う．具体的には，潜在的コミュニティ間での拡張型流行語の拡大に関する分析を行う．

5.1　コミュニティ間での流行語の拡大状況の分析

本節では，潜在的コミュニティ間での流行語の拡大状況の分析を行う．具体的には，すでに流行語大賞などでも取り上げられてメジャーな流行語となったトピックをいくつかピックアップし，これらのトピックがどのように世の中に拡散していったのかを分析することで，早期に発見するために必要な分析手法について検討する．

本稿で分析対象とした流行語は，「AKB48」「女子会」「スマフォ」「Android」「Facebook」「K-POP」である．**図 4** に「AKB48」，**図 5** に「女子会」，**図 6** に「スマフォ」，**図 7** に「Android」，**図 8** に「Facebook」，**図 9** に「K-POP」，それぞれの対象トピックに対する 1 週間ごとの総発言件数（投稿ブログ件数）と各コミュニティの発言割合の時間推移を示す．すなわち各図の左図は，対象トピックに対する 1 週間ごとの総発言数（投稿ブログ件数）の時間推移を示しており，右図は，対象トピックに関する総発言件数に占める，各コミュニティからの発言割合を示している．このコミュニティは，4 章で説明した手法により特定された潜在的コミュニティであり，ここでは 122 グループに分類された大分類を採用している．各右図には，4 つのコミュニティの値を表示しているが，これらはこの「各コミュニティの発言割合の時間推移」のうち，増加傾向もしくは減少傾向を顕著に示しているコミュニティを著者らが目視で確認し，手動で抽出したものを表示している．また，1 人のブロガが複数のコミュニティに属することを許しているため，各期間の発言割合は和は 1 を超えることもありうる．なお，分析期間は，2009 年 8 月から 2011 年 7 月までの 2 年間である．各トピックともに，投稿ブログ件数は増加傾向にあり，拡張型流行語の特徴を示していることが分かる．

図 4 の「AKB48」に対する「女性スター」や「アーティ

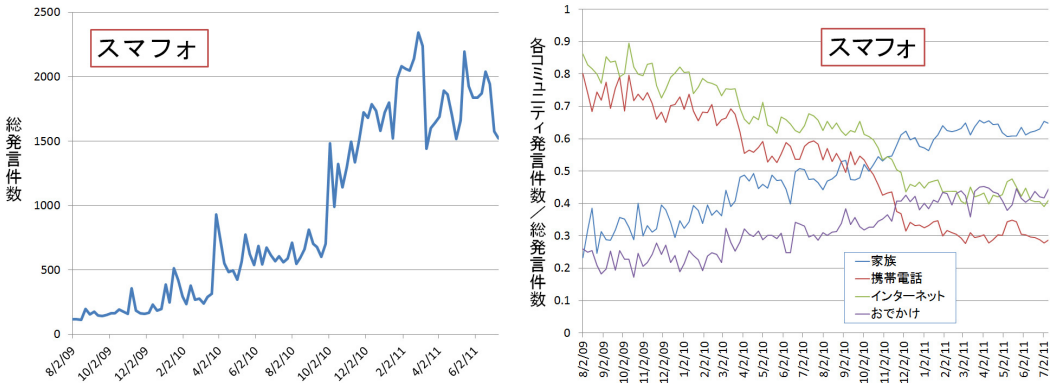

図 6　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（スマフォ）
Fig. 6　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: Smartphone).

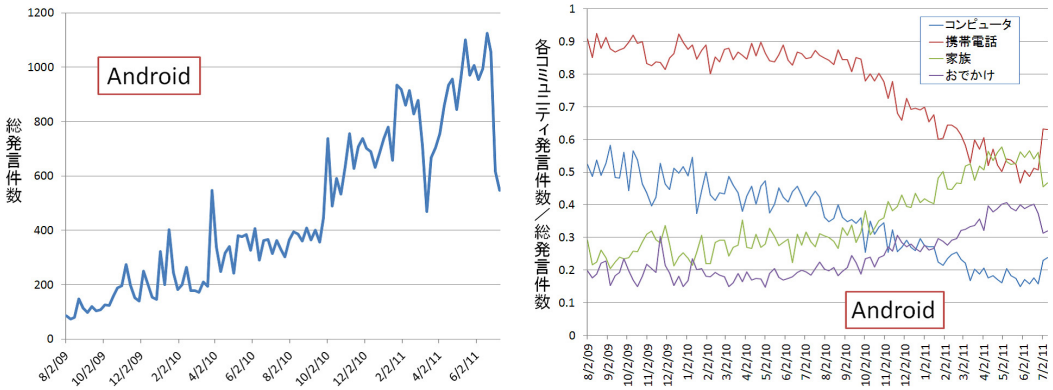

図 7　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（Android）
Fig. 7　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: Android).

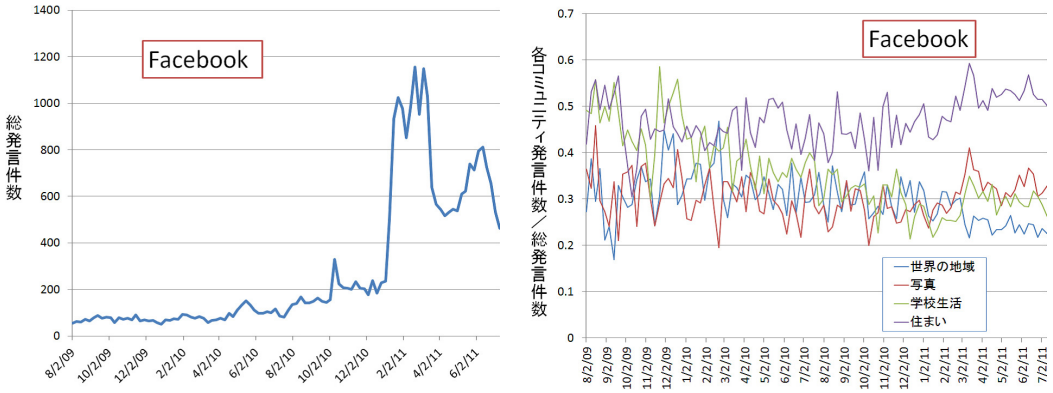

図 8　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（Facebook）
Fig. 8　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: Facebook).

スト」,

図 5 の「女子会」に対する「恋愛・結婚」や「顔文字」,

図 6 の「スマフォ」に対する「携帯電話」や「インターネット」,

図 7 の「Android」に対する「コンピュータ」や「携帯電話」,

図 8 の「Facebook」に対する「世界の地域」,

図 9 の「K-POP」に対する「韓流スター」や「アーティスト」

など，対象トピックに関連の深いコミュニティからの発言割合は，高い値から低い値へ推移していき，それに応じて，各トピックとは直接的な関係が薄いコミュニティからの発言割合が徐々に増加していく様子が確認できる．これは各トピックがメジャーな流行語になる以前は，対象トピック

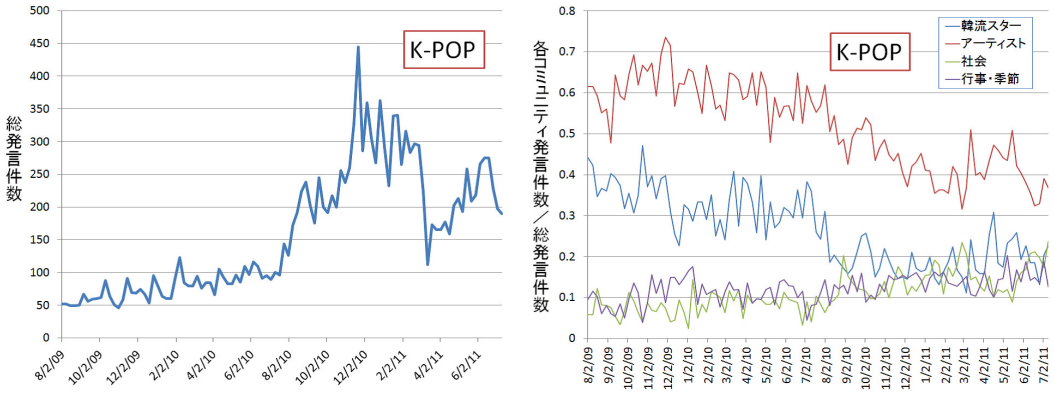

図 9　対象トピックに対する総発言件数と各コミュニティの発言割合の時間推移（K-POP）
Fig. 9　Numbers of blog entries including a buzzword and percentages of blog entries from different communities (Topic: K-POP).

に関連の深い一部の限られたブロガからの発言件数が大部分を占めていたが，その後，対象トピックがメジャーな流行語に近づくにつれて，他の一般的なコミュニティへの認知が進み，コミュニティの枠を超えて流行語が広がっていくという状況を表しているものと考えられる．すなわち，このように潜在的なブロガコミュニティごとの発言割合の時間推移を分析することで，流行語候補となる各トピックがメジャーな流行語に近づいていく状況が分析可能であることが確認できた．

ただし，今後，流行語候補の早期発見手法の実運用を目指すうえで，これら分析手法の実装が必要となるが，対象トピックに関連の深いコミュニティを自動でシステムに発見させる必要がある．したがって，対象トピックに関連が深いコミュニティを特定する手法について，次節で述べる．

なお，各トピックとの直接的な関係が薄いコミュニティとしては，「写真」や「住まい」「家族」などの重複が見られる．これらのコミュニティは多くのブロガが分類されやすい一般的なコミュニティである．したがって，この直接的な関係が薄いコミュニティからの発言割合の増加の分析に関しては，注目すべき一般的なコミュニティを事前に選定できる可能性がある．

5.2　流行語候補と関連の深いコミュニティの特定

各トピックに関連の深いコミュニティの特定方法としては，前節の図 4〜9 にも示しているような各コミュニティの発言割合の高さを使う方法が考えられる．しかしながら，各コミュニティは属するブロガ数に大きな隔たりがあり，「家族」のような一般的なコミュニティはブロガ数がかなり多いため，たとえ発言割合が高いとしても，必ずしも対象トピックとの関連が深いことを表しているとは限らない．したがって，総発言数に対する発言割合とは別の手法が必要となる．

そこで我々は，コミュニティ別総人数に対するコミュニティ別発言ブロガ数の割合を使って，対象トピックと関連の深いコミュニティを特定することを試みる．すなわち，コミュニティの大きさ（属するブロガの多さ）にかかわらず，コミュニティ内の全メンバのうち，どのくらいの割合のブロガが対象トピックについて発言しているのかを示している．この割合が大きければ，対象トピックとの関連が深いと判断できると考えている．

図 10 に，コミュニティ別総人数に対するコミュニティ別発言ブロガ数の割合を示す．ここでは，各トピックに関して発言数の多いコミュニティ 30 個を対象に，結果を表示している．図 10 が示すとおり，「AKB48」に対しては「アーティスト」，「スマフォ」に対しては「携帯電話」，「Android」に対しては「コンピュータ」，「facebook」に対しては「世界の地域」，「K-POP」に対しては「韓流スター」のように，確かに関連の深いコミュニティが，他のコミュニティに比べて高い値を示している．しかしながら，「女子会」に対しては関連の深いコミュニティ「恋愛・結婚」と「顔文字」の割合は高くなかった．これより，対象トピックと関連が深いすべてのコミュニティを特定できるわけではないが，関連の深いコミュニティを効率的に特定するための手法として有効であることが分かった．今後は他の手法についても検討して，複合的に処理することで，さらに効率的に関連の深いコミュニティを自動で特定できるようなシステムの構築を目指す．なお，他の手法としては，対象トピックと各コミュニティの類似度（共起語の重なりなど）を使って，関連の深さを見積もる手法についても検討する予定である．

メジャーな流行語に発達するまでの状況を確認するためには，対象トピックと関連の深いコミュニティを特定することと，この関連の深いコミュニティの発言割合が減少していく様子を観察する必要がある．したがって，発言割合の減少状況の観察について次節で述べる．

5.3　各コミュニティの発言割合の変動状況の分析

前節で述べたとおり，流行語候補がメジャーな流行語に

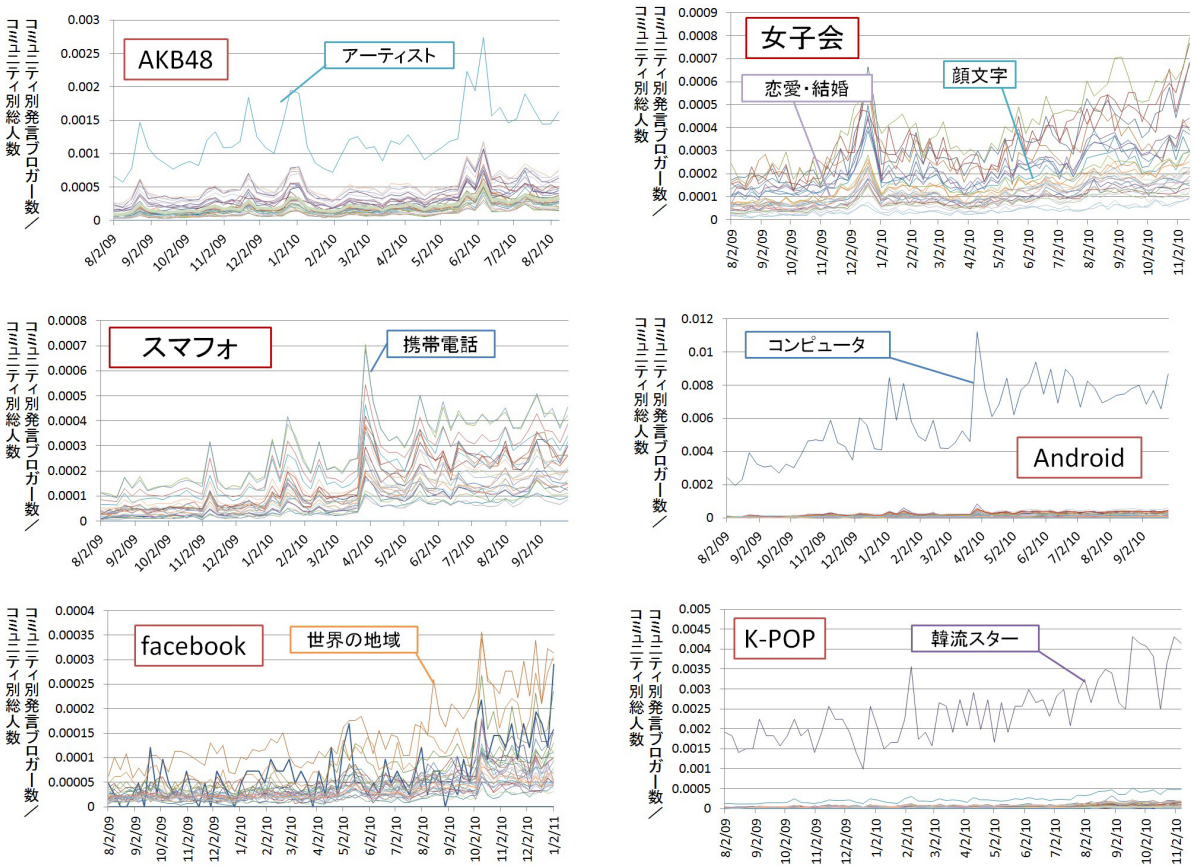

図 10　コミュニティ別総人数に対するコミュニティ別発言ブロガ数の割合

Fig. 10　Ratios of bloggers in a community who talked about the target topic to all the bloggers in the community.

発達するまでの特徴として，流行語候補と関連の深いコミュニティの発言割合の減少があげられる．したがって，過去のメジャーな流行語が発達する過程において，どのような挙動を示していたのか，さらには，メジャーな流行語に発達する前段階で，どのような挙動が確認できれば，その後メジャーな流行語になる可能性が高まるのか，ということについて検討する．

ここで，**図 11** に，各トピックと関連の深いコミュニティの発言件数割合の減少とその近似式を示す．各グラフは，それぞれ「AKB48」「女子会」「スマフォ」「Android」「Facebook」「K-POP」に関するものである．各グラフの上図は，総発言数の時間推移であり，このグラフ内の縦棒は発言数が急増する直前の時期を示している．将来，流行語候補の早期発見システムを構築する場合には，発言数が急増する前に流行語候補を特定できなければ，早期発見システムとしての価値は低い．したがって，発言数が急増しメジャーな流行語へと発達するまでの挙動から，流行語候補の早期発見を行う必要がある．そこで，各グラフの下図に，上図の縦棒で示している時期までの期間を対象として，各トピックと関連の深いコミュニティの発言割合の減少状況をその近似線とあわせて示している．なお，近似線は最小二乗法 [19] で求められた．

各グラフにおける各コミュニティの発言割合の減少状況を見ると，12 個のコミュニティのうち，10 個のコミュニティの近似線の傾き（経過日数に対する発言割合の変化）が −0.0003 以下となっている．今後さらなる検証実験を行っていく必要があるが，この傾きの大きさが 1 つの目安になるものと考えている．また，この発言割合の減少の期間や減少幅も考慮すべきである．

なお，流行語候補と関連の薄いコミュニティのうち，流行語候補がメジャーな流行語に発達するにつれて，発言割合が増加していくものに関しては，今後その特定方法についての検討が必要であるが，変動状況の分析については本節の手法を適用することが可能である．

6. 提案手法に基づく流行語候補のランキングに関する評価実験

6.1 評価実験の目的および流行語の定義

本研究の最終目標は，ブログ記事の時系列分析に基づく流行語候補の早期発見システムの実現である．アプリケーション例として，同カテゴリのライバル関係にあるキーワードについて，どのキーワードがメジャーな流行語に近づいているか，ということをランキングで示すというものがあげられる．本評価実験ではライバル関係にある複数の

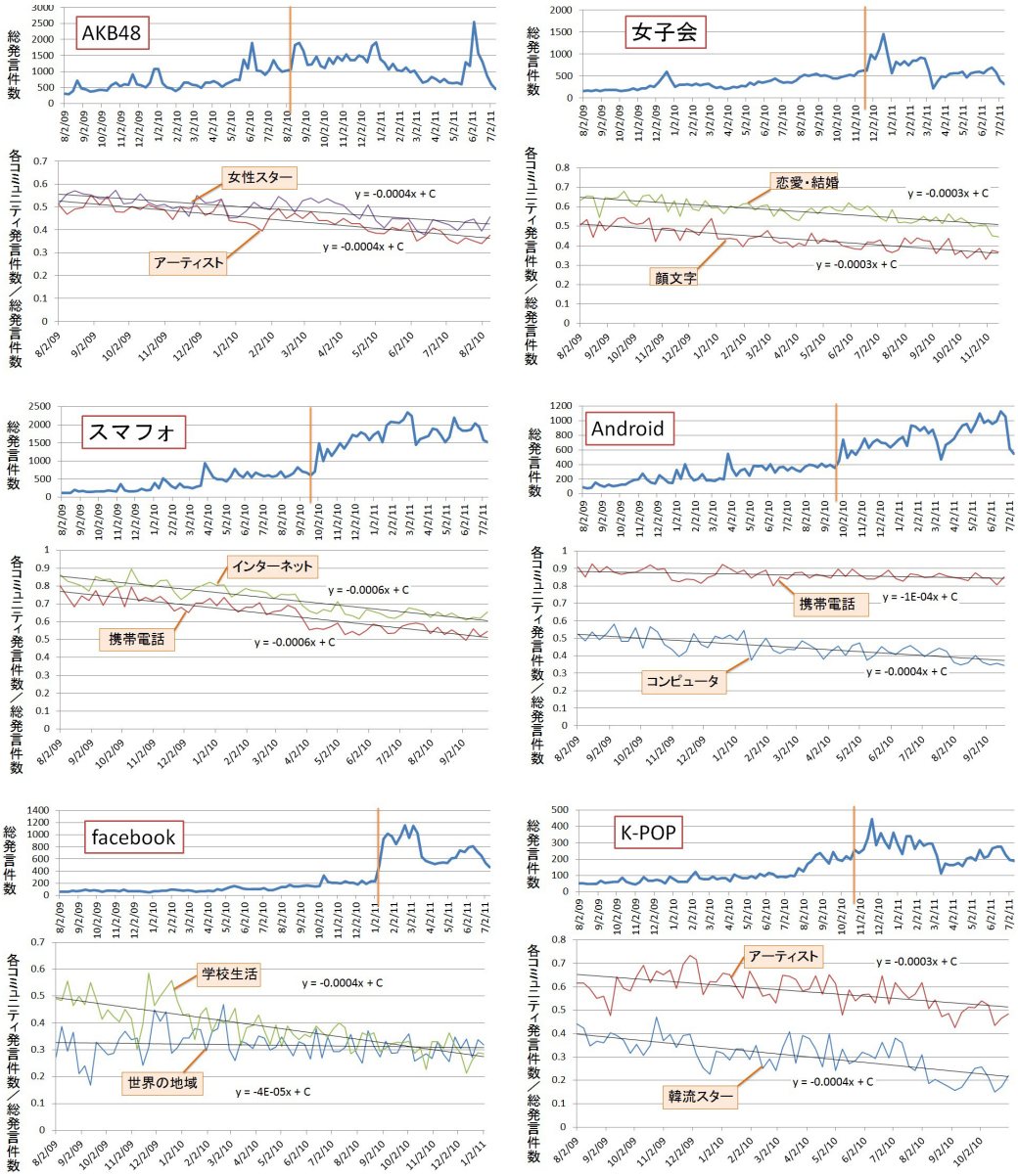

図 11　各トピックと関連の深いコミュニティの発言件数割合の減少とその近似式

Fig. 11　Approximate straight lines of percentage decrease of blog entries from closely related communities.

流行語候補に対するランキングにおいて，ベースライン手法との比較を行うことで，提案手法の妥当性の評価を行うことを目的とする．

本評価実験においては，「流行語」を以下のように定義する：「ある語 X の社会的認知度が過去の時点よりも大きく高まった場合，キーワード X は流行している」．

また，相対的な定義としては，以下のように定義することができる：「ある語 X が別の語 Y よりも “流行している” とは，“アイテム X の順位躍進度がアイテム Y より大きい”」．

なお，順位躍進度は，以下のように算出するものとする：「順位躍進度 = アイテムの過去の順位/現在の順位」．

6.2　評価実験にて扱う対象データ

本評価実験では，対象データとして，アイドルグループ「AKB48」の第 2 回総選挙結果（2010 年 6 月 9 日実施），第 3 回総選挙結果（2011 年 6 月 9 日実施）およびテレビドラマ（2011 年 7〜9 月期）の視聴率を採用した．なお，AKB48 総選挙にて分析対象としたメンバ 10 名は，第 3 回選挙におけるトップ 10 に入ったメンバ 10 名とした．

解析対象期間としては，AKB48 の第 2 回総選挙では，第 1 回から第 2 回選挙までの期間である 2009 年 8 月 2 日〜2010 年 6 月 6 日，AKB48 の第 3 回総選挙では，第 2 回から第 3 回までの期間である 2010 年 6 月 13 日〜2011 年 6 月 5 日とした．テレビドラマ（2011 年 7〜9 月期）の視聴率では，各番組で異なるが初回放送日から最終回放送日の前日までを解析対象期間とした．また，正解データとして

表 1　評価実験対象データ：AKB48 総選挙
Table 1　AKB48 election.

メンバ	第 1 回順位	第 2 回順位	第 3 回順位	第 2 回総選挙での順位躍進度	第 2 回総選挙での順位躍進度の順位	第 3 回総選挙での順位躍進度	第 3 回総選挙での順位躍進度の順位
前田	1	2	1	0.50	10	2.00	3
大島	2	1	2	2.00	2	0.50	9
柏木	9	**8**	**3**	1.13	5	**2.67**	**1**
篠田	3	3	4	1.00	6	0.75	8
渡辺	4	5	5	0.80	9	1.00	6
小嶋	6	7	6	0.86	7	1.17	4
高橋	5	6	7	0.83	8	0.86	7
板野	**7**	**4**	8	**1.75**	**3**	0.50	10
指原	27	**19**	**9**	1.42	4	**2.11**	**2**
松井	**29**	**11**	10	**2.64**	**1**	1.10	5

表 2　評価実験対象データ：ドラマ視聴率（2011 年 7〜9 月期）
Table 2　Drama ratings (July〜September, 2011).

ID	ドラマ名	初回視聴率	初回順位	最終回視聴率	最終回順位	順位躍進度	順位躍進度の順位
A	全開ガール	14.6	3	12.6	6	0.50	10
B	絶対零度 2	15.4	1	16.3	2	0.50	9
C	チーム・バチスタ 3	14.2	4	15.4	4	1.00	4
D	**ブルドクター**	**13.9**	**5**	**16.4**	**1**	**5.00**	**1**
E	それでも，生きてゆく	10.6	9	10.1	9	1.00	5
F	美男ですね	10.9	8	11.5	7	1.14	3
G	**ドン★キホーテ**	**11.7**	**7**	**12.8**	**5**	**1.40**	**2**
H	華和家の四姉妹	13.5	6	10.5	8	0.75	7
I	花ざかりの君たちへ 2011	10.1	10	7.3	10	1.00	6
J	新・警視庁捜査一課 9 係 3	14.8	2	15.9	3	0.67	8

は，AKB48 の総選挙結果では，その前回選挙からの順位躍進度を，テレビドラマの視聴率では，初回視聴率順位から最終回視聴率順位への順位躍進度を採用した．

表 1 に，AKB48 総選挙の第 1 回から第 3 回までの順位，第 2 回総選挙，第 3 回総選挙時の順位躍進度と，順位躍進度の順位を示す．表 1 より，第 1 回選挙から第 2 回選挙にかけては，松井 および 板野 が，“29 位から 11 位”，“7 位から 4 位” と大きく躍進していることが分かる．また，柏木 および 指原 が，“8 位から 3 位”，“19 位から 9 位” と大きく躍進している．すなわち，これら順位躍進度の高いメンバを上位に予測することが，提案手法としては重要となる．

表 2 に，2011 年 7〜9 月期の主要ドラマ 10 件の初回視聴率および最終回視聴率とその順位躍進度を示す．初回視聴率の順位と比べた最終回視聴率の順位が大きく躍進しているのが，ブルドクター および ドン★キホーテ であることが分かる．すなわち，これら順位躍進度の高い番組を上位に予測することが，提案手法としては重要となる．

6.3　提案手法および Baseline 手法

本節では，評価実験にて採用した提案手法および Baseline 手法の説明を行う．

提案手法としては，5.3 節で例を示したように，流行語候補と関連の深いコミュニティの発言割合の時系列変化の近似式を分析し，この一次近似の傾きの大きさに基づいてランキングを行う手法を採用する．すなわち，あるカテゴリに属するライバル関係のキーワード群に対して，一次近似のマイナスの傾きが大きいキーワードがより流行しているものと判定する．

Baseline 手法としては，頻度ベースの手法として，対象期間における流行語候補に関する総発言件数に基づいたランキング手法（Baseline1）および，流行語候補に関する発言件数の時系列変化の一次近似の傾きに基づいたランキング手法（Baseline2）を採用する．すなわち，Baseline1 では，対象期間における総発言件数が多ければ，より流行しているものと判定する．Baseline2 は，Baseline1 に対して時系列変化を考慮したものであり，対象期間における発言件数の増加率が高い場合には，より流行しているものと判定する．

6.4　評価実験結果における Baseline 手法との比較

表 3 に，AKB48 第 2 回選挙（AKB48-2）での順位躍進度，AKB48 第 3 回選挙（AKB48-3）での順位躍進度，2011 年 7〜9 月期の主要ドラマ（Drama）の初回視聴率に対す

表 3　各手法により予測された順位躍進度によるランキング結果
Table 3　Ranking results of rapid rise predicted by each method.

AKB48-2	前田	大島	柏木	篠田	渡辺	小嶋	高橋	板野	指原	松井
正解データ	10 位	2 位	5 位	6 位	9 位	7 位	8 位	**3 位**	4 位	**1 位**
提案手法	3 位	8 位	7 位	10 位	5 位	9 位	6 位	**2 位**	4 位	**1 位**
Baseline1	1 位	4 位	7 位	5 位	8 位	3 位	2 位	**6 位**	10 位	**9 位**
Baseline2	8 位	3 位	10 位	1 位	6 位	2 位	4 位	**7 位**	9 位	**5 位**

AKB48-3	前田	大島	柏木	篠田	渡辺	小嶋	高橋	板野	指原	松井
正解データ	3 位	9 位	**1 位**	8 位	6 位	4 位	7 位	10 位	**2 位**	5 位
提案手法	5 位	6 位	**1 位**	7 位	3 位	10 位	9 位	8 位	**2 位**	4 位
Baseline1	1 位	3 位	**7 位**	6 位	8 位	4 位	2 位	5 位	**9 位**	10 位
Baseline2	1 位	5 位	**2 位**	9 位	8 位	7 位	10 位	3 位	**4 位**	6 位

Drama	A	B	C	**D**	E	F	**G**	H	I	J
正解データ	10 位	9 位	4 位	**1 位**	5 位	3 位	**2 位**	7 位	6 位	8 位
提案手法	5 位	9 位	4 位	**3 位**	1 位	6 位	**2 位**	8 位	7 位	10 位
Baseline1	2 位	3 位	8 位	**5 位**	4 位	1 位	**9 位**	7 位	6 位	10 位
Baseline2	4 位	8 位	5 位	**6 位**	1 位	10 位	**3 位**	7 位	9 位	2 位

表 4　DCG と nDCG 計算例（AKB48-2 を例）
Table 4　Calculation example for DCG and nDCG (based on AKB48-2).

Ranking	1	2	3	4	5	6	7	8	9	10	DCG	nDCG
正解データ	松井	大島	板野	指原	柏木	篠田	小嶋	高橋	渡辺	前田	(IDCG)	
(rel_i)	2.64	2.00	1.75	1.42	1.13	1.00	0.86	0.83	0.80	0.50	7.116	1.000
提案手法	松井	板野	前田	指原	渡辺	高橋	柏木	大島	小嶋	篠田		
(rel_i)	2.64	1.75	0.50	1.42	0.80	0.83	1.13	2.00	0.86	1.00	6.766	0.951
Baseline1	前田	高橋	小嶋	大島	篠田	板野	柏木	渡辺	松井	指原		
(rel_i)	0.50	0.83	0.86	2.00	1.00	1.75	1.13	0.80	2.64	1.42	5.159	0.725
Baseline2	篠田	小嶋	大島	高橋	松井	渡辺	板野	前田	指原	柏木		
(rel_i)	1.00	0.86	2.00	0.83	2.64	0.80	1.75	0.50	1.42	1.13	5.701	0.801

る最終回視聴率の順位躍進度の正解データおよび，提案手法，Baseline 手法による予測結果の一覧を示す．

このランキング結果より，AKB48-2 における 松井 および 板野 の躍進，AKB48-3 における 柏木 および 指原 の躍進，Drama における D（ブルドクター）および G（ドン★キホーテ）の躍進を，提案手法では非常によく予測できていることが分かる．

ここで，正解データのランキングに対して，提案手法および Baseline 手法のランキングがどれだけ一致するのかを評価する．評価尺度としては，DCG および nDCG [20] を採用した．

DCG とは，Discounted Cumulative Gain（減損累積利得）の略であり，「上位 n 件に正解データを含めることができたか」というような指標とは違い，順位を含めて正解データのランキングをどれだけ再現できるのかを評価するものである．すなわち，単に正解ランキングの上位 5 件を，提案手法のランキングにおけるランキングの上位 5 件に含めればよい，ということではなく，そのランキング順位についても合致する際には高く評価することができる手法である．なお，nDCG とは，正規化減損累積利得であり，DCG を理想的な DCG の値（IDCG）で正規化したものである．DCG および nDCG の算出式を以下に示す．

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)} \tag{4}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{5}$$

ただし，rel_i は，ランキング i 番目のアイテムの利得スコアである．また，p は，ランキングを考慮する件数を示しており，今回の実験では $p = 10$ である．なお，$IDCG_p$ は，理想的な DCG_p の値を示す．すなわち，ランキング上位であれば，高い重みでスコアを加算することができるため，スコアの高いランキング上位のアイテムを上位にランキングできれば DCG の値は高くなる．

ここで，AKB48-2 を例とした DCG と nDCG の計算例を**表 4** に示す．表 4 では，正解データおよび各手法によるランキング結果と各アイテム（メンバ）の利得スコア rel_i（今回の実験では順位躍進度である）を表示している．正解データでは，rel_i の大きい順にランキングされているため，DCG は最大となり，nDCG が 1 となっている．提案

表 5　DCG および nDCG の比較による評価結果
Table 5　Evaluation results based on DCG and nDCG.

検証トピック	AKB48-2		AKB48-3		Drama	
評価尺度	DCG	nDCG	DCG	nDCG	DCG	nDCG
正解データ	7.116	1.000	7.105	1.000	8.384	1.000
提案手法	6.766	**0.951**	6.932	**0.976**	6.328	**0.755**
Baseline1	5.159	0.725	5.916	0.833	5.609	0.669
Baseline2	5.701	0.801	6.608	0.930	5.543	0.661

表 6　流行語早期発見時期
Table 6　Time for earlier detection of buzzwords.

検証トピック	分析期間	提案手法順位	流行語早期発見時期
AKB48-2	約 10 カ月	1 位：松井	4 カ月 3 週間前
		2 位：板野	6 カ月前
AKB48-3	約 1 年間	1 位：柏木	9 カ月 1 週間前
		2 位：指原	9 カ月 2 週間前
Drama	11 週間	1 位：ブルドクター	6 週間前
		2 位：ドン★キホーテ	2 週間前

手法では，そのランキングにおいて正解データと近いため，DCG および nDCG の値が比較的大きな値を得ている．これに対して，Baseline 手法では，そのランキングにおいて正解データと比較的ずれが大きいため，提案手法に比べると DCG および nDCG の値がそれほど大きな値を得ることができていない．

表 5 に，DCG および nDCG の比較による評価結果を示す．対象としたすべてのトピックにおいて，提案手法が Baseline 手法を上回る結果となっている．特に AKB48-2 および AKB48-3 の nDCG においては，0.951 および 0.976 と非常に高い値を示している．したがって，対象データにおける各アイテムの躍進の予測において提案手法の有効性を示せたと考えている．

6.5　流行語早期発見能力の評価

本節では，流行語の早期発見能力について検討する．6.4 節で示した実験結果の正解データであるアイテムを，提案手法ではどれほど早期に検知できるかを検証した．

表 6 に，流行語早期発見に関する実験結果を示す．ここでは，各トピックに対して，提案手法にて 1 位および 2 位としてランキングしたアイテムを，実際どれほどの期間で予測可能であったかと検証した．なお，AKB48-2 および AKB48-3 においては，関連の深いコミュニティの発言割合の 1 次近似の傾きが，5 週間連続で −0.0003 を下回った場合には，その時点を流行語発見時期としている．Drama においては，関連の深いコミュニティの発言割合の 1 次近似の傾きが，3 週間連続で −0.0003 を下回った場合には，その時点を流行語発見時期としている（全体の分析期間が短いため）．

AKB48-2 では，全体の分析期間は約 10 カ月であったが，1 位の 松井 に関しては，第 2 回選挙の 4 カ月 3 週間前の時点で予測可能であり，2 位の 板野 に関しては，第 2 回選挙の 6 カ月前の時点で予測可能であった．

AKB48-3 では，全体の分析期間は約 1 年間であったが，1 位の 柏木 に関しては，第 3 回選挙の 9 カ月 1 週間前の時点で予測可能であり，2 位の 指原 に関しては，第 3 回選挙の 9 カ月 2 週間前の時点で予測可能であった．

Drama では，全体の分析期間は 11 週間であったが，1 位の ブルドクター に関しては，最終回放映日の 6 週間前の時点で予測可能であり，2 位の ドン★キホーテ に関しては，最終回放映日の 2 週間前の時点で予測可能であった．

以上より，必ずしも対象となる分析期間のすべてを調べる必要はなく，適切な閾値を設定することにより，早期に流行語候補を予測することが可能となると考えている．なお，このときの閾値の設定は容易ではなく，対象トピックや分析期間に依存するため，実運用する際にはこれらのチューニングが必要である．

7.　おわりに

本研究では，すでにメジャーな流行語となったトピックに対し，これらにトピックがブログ上でどのように世の中に拡散していったのかを，大規模な実データを時系列的に分析することで，拡張型流行語を早期に発見するために必要な分析手法について検討した．kizasi.jp で扱っている，3,776,154 サイトで過去 2 年間に投稿された 81,922,977 件のブログ記事データの分析の結果，流行語候補がメジャーな流行語に発達していく過程において，総発言数に占める，対象トピックと関連の深いコミュニティからの発言割合が減少しつつ，関連の薄いコミュニティからの発言割合が増加する状況を確認できた．また，対象トピックと関連の深いコミュニティの特定手法を検討するとともに，総発言数に占めるこのコミュニティからの発言数の割合の減少状況

について分析を行った．そのうえで，拡張型流行語の早期発見システムに必要な機能について検討した．さらに，ライバル関係にある複数の流行語候補のランキングに基づく，提案手法の妥当性の検証を行った結果，良好な結果を示し，本提案手法の大きな可能性を示せた．

なお，提案手法である，ブログ記事時系列分析に基づく流行語候補の早期発見システムを実現するために必要なものとしては，以下の機能があげられる．

- 対象トピックと関連の深いコミュニティおよび関連の薄いコミュニティの自動判別
- 関連の深いコミュニティおよび関連の薄いコミュニティの発言割合の変動状況の評価（数，傾き，期間，変動幅など）

したがって，今後はこれらの機能を有するプロトシステムの実装と評価実験に基づく精度向上を目指しつつ，実運用に向けた取り組みを進めていく予定である．

謝辞　本研究の一部は，文部科学省科学研究費助成事業（学術研究助成基金助成金）基盤研究（C）（課題番号：#23500140）による．ここに記して謝意を表します．

参考文献

[1] Yahoo! Buzz Index, available from ⟨http://buzzlog.yahoo.com/overall/⟩.
[2] Araujo, L. and Guervos, J.J.M.: Automatic Detection of Trends in Time-stamped Sequences: An Evolutionary Approach, *Soft Computing*, Vol.14, No.3, pp.211–227 (2010).
[3] Glance, N.S., Hurst, M. and Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs, *WWW 2004 workshop* (2004).
[4] Mathioudakis, M. and Koudas, N.: TwitterMonitor: Trend Detection over the Twitter Stream, *SIGMOD 2010*, pp.1155–1158 (2010).
[5] Parikh, N. and Sundaresan, N.: Scalable and Near Real-Time Burst Detection from eCommerce Queries, *KDD 2008*, pp.972–980 (2008).
[6] Yi, J.: Detecting Buzz from Time-Sequenced Document Streams, *EEE 2005*, pp.347–352 (2005).
[7] 山岡千夏，中島伸介，張　建偉，稲垣陽一，中本レン：ブログ記事の時系列解析に基づく流行語候補「兆し」の早期発見手法，第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2011），A4–1 (2011).
[8] 稲垣陽一，中島伸介，張　建偉，中本レン，桑原　雄：ブロガーの体験熟知度に基づくブログランキングシステムの開発および評価，情報処理学会論文誌：データベース，Vol.3, No.3 (TOD47), pp.123–134 (2010).
[9] Kleinberg, J.M.: Bursty and Hierarchical Structure in Streams, *SIGKDD 2002*, pp.91–101 (2002).
[10] Lappas, T., Arai, B., Platakis, M., Kotsakos, D. and Gunopulos, D.: On Burstiness-aware Search for Document Sequences, *SIGKDD 2009*, pp.477–486 (2009).
[11] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the Bursty Evolution of Blogspace, *WWW 2003*, pp.568–576 (2003).
[12] Gruhl, D., Guha, R., L-Nowell, D. and Tomkins, A.: Information Diffusion Through Blogspace, *WWW 2004*, pp.491–501 (2004).
[13] 奥村　学：blog マイニング—インターネット上のトレンド，意見分析を目指して，人工知能学会誌，Vol.21, No.4, pp.424–429 (2006).
[14] Khy, S., Ishikawa, Y. and Kitagawa, H.: A Novelty-based Clustering Method for On-line Documents, *World Wide Web Journal*, Vol.11, No.1, pp.1–37 (2008).
[15] 長谷川幹根，石川佳治：T-Scroll：時系列文書のクラスタリングに基づくトレンド可視化システム，情報処理学会論文誌：データベース，Vol.48, No.SIG 20 (TOD 36), pp.61–78 (2007).
[16] 金澤健介，Adam Jatowt，小山　聡，田中克己：Web 上の将来情報の集約的提示，Web とデータベースに関するフォーラム（WebDB Forum 2009），4A-1 (2009).
[17] 相澤彰子：共起に基づく類似性尺度，〈特集〉自然言語とコンピュータ），オペレーションズ・リサーチ：経営の科学，Vol.52, No.11, pp.706–712 (2007).
[18] 松尾　豊，石塚　満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌，Vol.17, No.3, pp.217–223 (2002).
[19] Kariya, T. and Kurata, H.: *Generalized Least Squares*, Wiley (2004).
[20] Jarvelin, K. and Kekalainen, J.: Cumulated Gain-based Evaluation of IR Techniques, *ACM Trans. Information Systems*, Vol.20, No.4, pp.422–446 (2002).

中島 伸介　（正会員）

京都産業大学コンピュータ理工学部准教授．博士（情報学）．1997 年神戸大学大学院自然科学研究科博士前期課程修了．2004 年京都大学大学院情報学研究科博士後期課程修了．情報通信研究機構専攻研究員，奈良先端科学技術大学院大学助教を経て，2008 年より現職．主にブログマイニングおよび情報推薦の研究に従事．電子情報通信学会，日本データベース学会，ACM，IEEE-CS 各会員．

張 建偉　（正会員）

筑波技術大学産業技術学部助教．2005 年筑波大学大学院システム情報工学研究科博士前期課程修了．2008 年筑波大学大学院システム情報工学研究科博士後期課程修了．博士（工学）．埼玉大学情報メディア基盤センター産学官連携研究員，京都産業大学コンピュータ理工学部特定研究員を経て，2012 年より現職．Web マイニング，Web 情報システム，情報保障の研究に従事．日本データベース学会会員．

稲垣 陽一

1990 年東京大学文学部言語学科卒業．（株）シーエーシー入社，技術研究室に配属．スタンフォード大学コンピュータサイエンス学科客員研究員（1996〜1998）．きざしサーチエンジンの研究開発を経て，2007 年 1 月より（株）きざしカンパニー代表取締役専務 CTO をつとめる．

中本 レン

（株）きざしカンパニー技術研究員．2003 年オレゴン州立大学工学部卒業．2008 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．

（担当編集委員　河合 英紀）