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

山岡　千夏† 　中島　伸介† 　張　建偉† 　稲垣　陽一†† 　中本　レン††

† 京都産業大学コンピュータ理工学部　〒 603–8555 京都市北区上賀茂本山
†† 株式会社きざしカンパニー　〒 103–0015 東京都中央区日本橋稲崎町 24–1 日本橋箱崎ビル 2F
E-mail: †{g0847344,nakajima,zjw}@cc.kyoto-su.ac.jp, ††{inagaki,reyn}@kizasi.jp

あらまし　流行語は，テレビや雑誌で紹介される等，世間に知れ渡った後から知る事が多く，流行語の先駆けを発見することは困難である．近年，インターネットの普及に伴い，ブログや Twitter といったブログサービスが急激に普及している．ブログコンテンツは，人々の関心を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ため，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トレンド語の先駆けを発見できる可能性がある．本研究ではブログ記事を時系列解析することで，記事の増加，コミュニティ間での発言者の拡大に着目し，流行語候補「兆し」の早期発見する手法の提案を行う．

キーワード　話題分析，ブログ分析，時系列解析

Earlier Detection of Buzzword Candidate ”KIZASI” Based on Time-series Analysis of Blog Articles

Chinatsu YAMAOKA†, Shinsuke NAKAJIMA†, Jianwei ZHANG†, Yoichi INAGAKI††, and Reyn NAKAMOTO††

† Faculty of Computer Science and Engineering, Kyoto Sangyo University　Motoyama, Kamigamo,Kita-ku, Kyoto, 603–8555 Japan
†† kizasi Company，Inc．24–1 Inazakityo,Nihonbashi ,Chuo-ku,Tokyo, 103–0015 Japan
E-mail: †{g0847344,nakajima,zjw}@cc.kyoto-su.ac.jp, ††{inagaki,reyn}@kizasi.jp

Abstract　Buzzwords are often known after they are introduced by TV or magazines and most of people have known them. It is difficult to detect the buzzwords at the early stage. Recently, blog services such as blogs and Twitter are rapidly becoming widespread as the Internet becomes popular. It is possible to detect the buzzwords before they become known by most of people by properly analyzing blog contents, which reflect people’s concerns in real time. We propose a method for early detecting buzzword candidates by conducting a time-series blog analysis and focusing on both the increase of blog articles and the growth of bloggers between communities.

Key words　Topic analysis，Blog analysis，Time-series analysis

1. はじめに

世間の流行語は，テレビや雑誌で紹介される等，世間に知れ渡った後から知る事が多く，流行語の先駆けを発見することは困難である．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である．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伴い，ブログや Twitter といっ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，人々の関心を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ため，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トレンド語の先駆けを発見できる可能性がある．したがって，本研究ではブログ記事を時系列解析することで流行語候補「兆し」の早期発見する手法の提案を行う．

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幾つかのパターンが存在すると考えられる．一つ目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渡るような流行語 ”を拡張型流行語と呼ぶ.「女子会」や「AKB48」はこのタ

イプといえる．別の流行語のタイプとして，“ テレビ，ニュース，雑誌等で取り上げられたことで，一斉に広がり，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．突発的な流行語は予測することが不可能であるため，本研究では，早期発見を目指す流行語のタイプとして 1 つ目の拡張型流行語を対象とする．

なお，ブログ記事の時系列解析においては，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，効率的に発見する事ができると考える．つまり，同じメンバー内での書き込み数が増えたケースでは，世の中への流行語の広がりとしては不十分だと考えており，書き込みをしているメンバーやコミュニティの数が広がる様子を検知することで流行語候補の早期発見精度の向上を目指している．我々は，先行研究 [1] にて，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している．その中で，12000 領域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．これを暗黙的なコミュニティと考える事ができるので，これらを利用する事で，コミュニティ間の話題の広がりを分析することが可能である．

2. 関 連 研 究

奥村らは，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，どの程度広まったかを検出し提示するシステムを開発している [2]．福原らは，感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出に関する研究を行っている [3]．長谷川らは，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に関する研究を行っている [4]．この研究ではトレンドの発見そのものではなく，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．灘本らは，ブロガーの注目情報を用いた株価変動予測に関する研究を行っている [5]．この研究では，ブログ記事中に表れる株価の変動と相関のあるキーワード群を抽出することで株価変動予測に取り組んでいる．金澤らは，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共に，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 [6]．内海らは，大規模テキストマイニングによる医療分野の社会課題・技術トレンド抽出に研究を行っている [7].

以上の通り，既に広まったキーワードの検出や可視化を目的とした研究は行われているが，コミュニティ間での流行語の広がりを分析することで流行語の早期発見手法の提案を目指したものはない．

3. ブログ記事の時系列解析に基づく拡張型流行語の早期発見手法

3.1 拡張型流行語のモデル

本稿では，“ ニュースや雑誌等で取り上げられ一斉に広がり，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，このような流行語の早期発見手法を提案することを目指している．したがって，この拡張型流行語の特徴としては，記事数が増えること，発言者コミュニティの幅が拡大すること，が挙げられる．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検証を行う．

なお，既知の拡張型流行語として,「2010 ユーキャン新語・流行語大賞」に選ばれた「AKB48」,「女子会」を例に挙げる．

3.1.1 拡張型流行語の特徴 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，拡張型流行語には成り得ない．

図 1 に，2008 年 12 月 28 日から 2010 年 12 月 5 日の期間中に「AKB48」について発言者数の時間推移を，図 2 に，同期間中に「女子会」について発言したブロガー数の時間推移を示す．

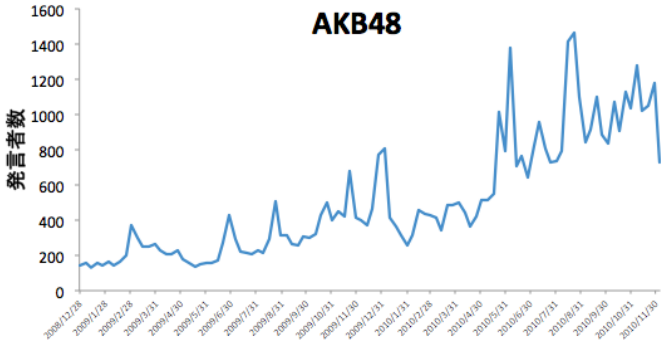

図 1　発言者数の時間推移 (「AKB48」の場合)

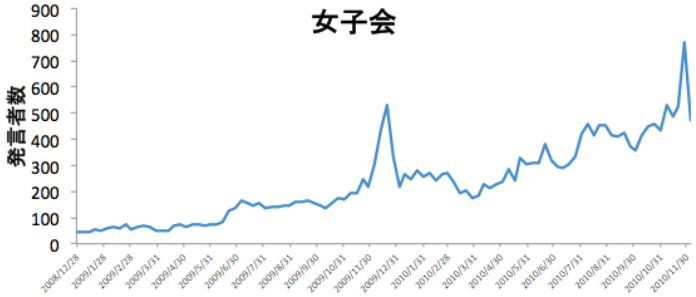

図 2　発言者数の時間推移 (「女子会」の場合)

なお，横軸は時間，縦軸は各キーワードを発言したブロガー数である.「AKB48」「女子会」共に，2010 年までに流行し一般的には良く知られたキーワードである．図 1，図 2 からも分かる通り，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居ないが，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型流行語であるいえる．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．したがって，次節にてコミュニティ間の発言者の拡大に関して説明する．

3.1.2 拡張型流行語の特徴２（コミュニティ間での発言者の拡大）

3.1.1 節にて説明したとおり，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（例えば,「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．

ここで，図 3 に，2008 年 12 月 28 日から 2010 年 12 月 5 日の期間中に「AKB48」について語ったブロガーの年代別発言割合の時間推移を，図 4 に，同期間中に「女子会」について語ったブロガーの年代別発言割合の時間推移を示す．

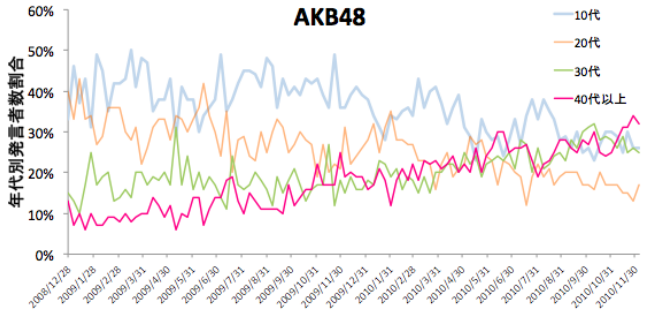

図 3　年代別発言割合の時間推移（「AKB48」の場合）

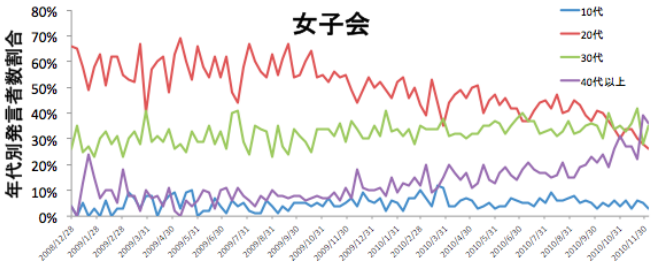

図 4　年代別発言割合の時間推移（「女子会」の場合）

図 3 および図 4 では共に，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 3 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事が分かる．同様に，図 4 からも，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．これより，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わかる．メジャーな流行語になりうると考えられる．これが拡張型流行語の典型的な特徴であり，一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．

ここで，図 5 に,「AKB48」に関する記事を投稿したブロガー数と，年代別発言割合の時間推移を時系列的に比較したものを示す．同様に，図 6 に,「女子会」に関する記事を投稿したブロガー数と，年代別発言割合の時間推移を時系列的に比較したものを示す．

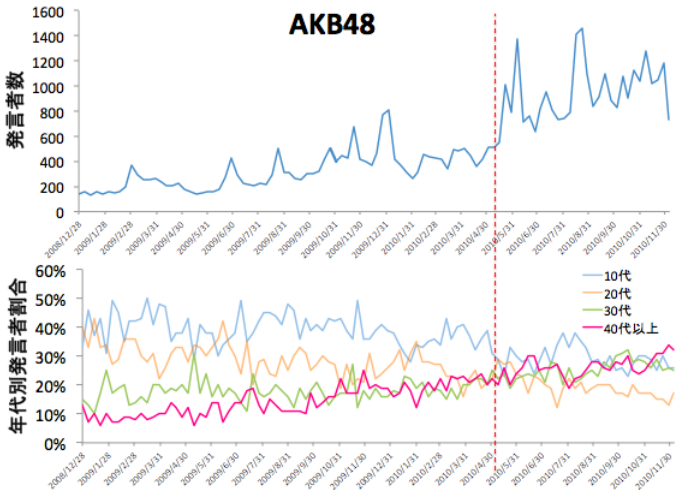

図 5　発言者数と年代別発言数割合の時間推移の関連性（「AKB48」の場合）

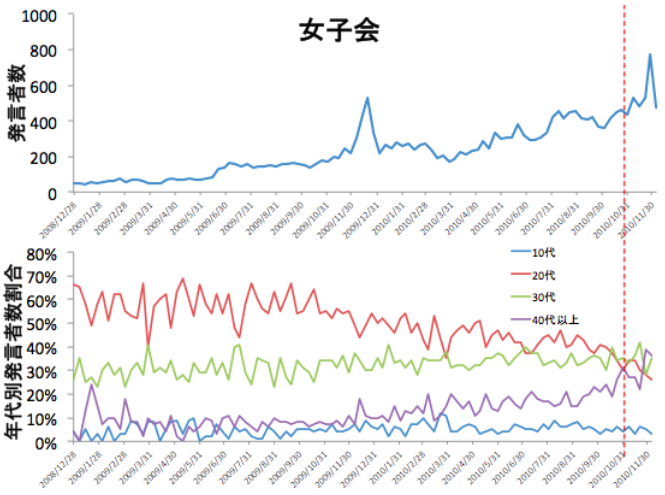

図 6　発言者数と年代別発言数割合の時間推移の関連性（「女子会」の場合）

図 5，図 6 より，拡張型流行語の特徴１（記事数の増加），特徴２（コミュニティ間での発言者の拡大）を踏まえて，AKB48 では 2010 年 5 月下旬，女子会は 2010 年 10 月下旬が，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，これを発見することが拡張型流行語の兆しを発見する鍵となると考えている．

3.2 拡張型流行語の早期発見手法

本節では，拡張型流行語の発見手法について説明する．ここで，コミュニティ別発言割合の時間推移において，最も割合の大きなコミュニティの発言割合を C_{top} とし，その他のコミュニティ i での発言割合を C_i とする．図 3 を例にとると，2009 年頃までは 10 代が C_{top} に該当し，その他が C_i となる．このとき，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を検出するための判別式を以下に示す．

$$\frac{C_i(t)}{C_{top}(t)} > \theta \qquad (1)$$

$C_{top}(t)$ は，最も話題になっているコミュニティの時刻 t での発言割合，$C_i(t)$ は，その他のコミュニティ $_i$ での時刻 t での発言割合，θ は閾値である．

すなわち，最も話題になっているコミュニティに対する他のコミュニティの比の割合がある閾値を超えた際には，その他のコミュニティが，最も話題となっているコミュニティに近づいているのかがわかる．予備実験を通じて，適切な閾値を設定する．これに基づいて，対象とする候補語が閾値を超えた場合，拡張型流行の兆しとして判別することができる．

「AKB48」では，2010 年 5 月下旬がターニングポイントであるとすると，閾値は 0.7 前後に設定することができる (図 7 参照)．「女子会」では，2010 年 10 月下旬がターニングポイントであるとすると，閾値もやはり 0.7 前後と設定することが可能である (図 8 参照)．

実際の閾値の設定は，今後行う予備実験などを通じて，適切な値を設定する．また，男女別，地域別，グループ別等のコミュニティの広がりの検出方法の改良を行う．

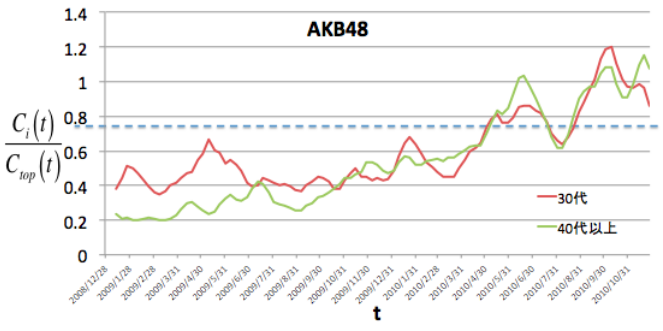

図 7　コミュニティ間の発言数割合の時間推移（「AKB48」の場合）

3.3 想定するアプリケーション

アプリケーションとしては，ユーザが興味を持つトピックに関する流行語の兆しをそのユーザに提示するシステムを想定している．利用者としては 2 つのケースを検討している．1 つ目は，ユーザが興味も持っているトピックを入力し，このトピックに関する流行語の兆しを提示するケース．2 つ目は，ブロガーに対してそのブロガーが興味をもつトピックに関する流行語の兆しを提示するケースである．ブロガーが興味を持つトピックは，そのブロガーの過去の投稿記事を解析することで推定する．

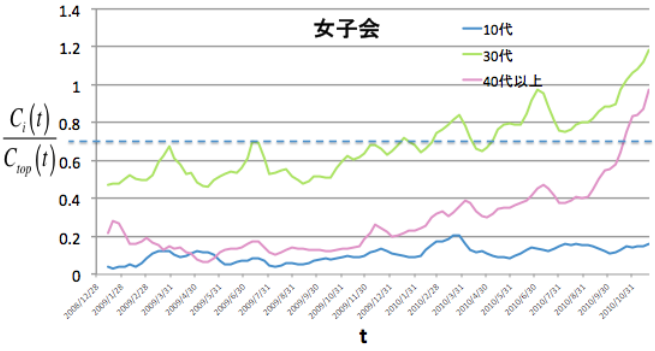

図 8　コミュニティ間の発言数割合の時間推移（「女子会」の場合）

流行語の兆しの発見はオンデマンドで行うのではなく，収集したブログ記事の中で発言数が，ある閾値以上のキーワードに対してその発言数の増加状況やコミュニティ間での拡大状況を監視することで，流行語の兆しの発見を随時行っていく．このコミュニティ間での広がりは，年代別，男女別およびコミュニティ別での分類に基づいて行う．コミュニティの分類に関しては，先行研究 [1] にて検出している 12000 領域のコミュニティに対して，分類されてるブロガーの集合を 1 つのコミュニティとみなすことを考えている．

本アプリケーションの効果について説明する．ユーザが興味を持っているトピックに関する流行語の兆しを提示するケース（上述 1 つ目のケース）では，自分があまり詳しくない領域においても流行語の兆しを容易に検索可能となる．ブロガーに対してそのブロガーが興味をもつトピックに関する流行語の兆しを提示するケース（上述 2 つ目のケース）では，他のコミュニティのみで流行していたキーワードが自分が属するコミュニティに拡大し始めた際にそのキーワードを知らせてもらうことが可能となる．また，ある領域における流行語の兆しをいち早く把握することは，マーケティングの観点からも重要であり，本アプリケーションの有効性は高いと考えている．

4. ま と め

単にブログの発言数の増加に着目するのではなく，コミュニティ間での流行語の広がりに着目し，ブログ記事の時系列解析を行うことにより，将来有望な拡張型流行語の「兆し」を検知し流行語の早期発見手法の提案した．

今後の課題としては，システムの実装及び評価実験に基づく発見手法の改良を行う予定である．

謝　　辞

本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」，および文部科学省科学研究費補助金若手研究 (B)（課題番号：20700089）による．ここに記して謝意を表します．

文　　献

[1] 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 年 9 月.

[2] 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 年.

[3] 福原知宏、中川裕志、西田豊明 ：感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出, 第 20 回人工知能学会大会 2E1-02, 2006 年 5 月.

[4] 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 年 12 月.

[5] 灘本裕紀，堀内　匡：ブロガーの注目情報を用いた株価変動予測の試み，第 6 回情報科学技術フォーラム講演論文集, Vol.2, pp.369-370，2007 年 9 月.

[6] 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 年 11 月.

[7] 内海和夫, 乾孝司, 村上浩司, 橋本泰一, 石川正道. 大規模テキストマイニングによる医療分野の社会課題・技術トレンド抽出. 研究・技術計画学会第 22 回年次学術大会, pp.684–687, 2007 年.