流行先読みブロガー発見のための流行語分析手法

朝永　聖也† 　中島　伸介†† 　稲垣　陽一††† 　中本　レン††† 　張　建偉††††

† 京都産業大学大学院　先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学 コンピュータ理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†††† 筑波技術大学 産業技術学部　〒 305–8520 茨城県つくば市天久保 4-3-15
††† 株式会社きざしカンパニー　〒 103–0015　東京都中央区日本橋箱崎町 20-14 日本橋巴ビル 6F
E-mail: †i1358068@cc.kyoto-su.ac.jp, ††nakajima@cse.kyoto-su.ac.jp, †††{inagaki,reyn}@kizasi.jp,
††††zhangjw@a.tsukuba-tech.ac.jp

あらまし　有望な流行語候補を早期に発見する手法の一つとして，流行語先読みブロガーの発見を目指している．この流行語先読みブロガーは，過去の流行語を早期に投稿しているブロガーであるため，過去の流行語をどの時点で投稿しているかに着目することが重要だと考えている．具体的には，対象流行語の流行時の話題が語り始められた時期（成長期間）を推測することで，流行語先読みブロガーの発見が可能となると考えている．そこで，本稿では，対象流行語の成長期間の推測に有効な流行語の分析手法について検討したので，報告する．

キーワード　ブログマイニング, ブロガー先読み度, 流行語発見

1. はじめに

近年，流行語は様々なメディアで注目されている．流行語は世間で話題のキーワードやヒット商品・コンテンツを最も的確に捉えたキーワードであり，2013 年では,「今でしょ！」「お・も・て・な・し」「じぇじぇじぇ」「倍返し」等のキーワード [1] や,「コンビニコーヒー」「パズル＆ドラゴン」等の商品・コンテンツ [2] が注目された．このような流行語は世間に知れ渡ってから初めて知ることが多く，流行する前にこれを早期に発見することは大変困難である．しかしながら，マーケティングの観点において，流行語をいち早く検知することは大変重要だと考えられる．そこで我々はブログ分析に基づいた流行語の早期発見手法の開発に向けた研究を行っている．ブログや SNS 等の CGM は，一般ユーザによって情報発信されるものであり，これを分析することで世間にはまだ知れ渡っていないような未来の流行語候補を検出できる可能性がある．

流行語の早期発見手法として，“将来世間に広まりそうな流行語候補を推測する方法” と，“流行に敏感な先読みブロガーを発見し，このブロガーが発信する情報から流行語候補を検出しようとする方法” の 2 通りを考えている．これまでの研究で，前者の “将来世間に広まりそうな流行語候補を推測する方法” に関する研究については，既に実施しており，ある程度の成果を得ている [3] [4]．したがって，後者の “流行に敏感な先読みブロガーを発見し，このブロガーが発信する情報から流行語候補を検出しようとする方法” に取り組んでおり,「各ブロガーの投稿記事の履歴が，未来の話題に近いのか，過去の話題に近いのかを分析する手法 [5] [6]」や,「過去のメジャーな流行語を特定した上で，この過去の流行語に対して事前に言及した頻度を分析する手法 [7] [8]」について報告している．前者の「各ブロガーの投稿記事の履歴が，未来の話題に近いのか，過去の話題に近いのかを分析する手法 [5] [6]」では，あるトピックレベルでのコミュニティ内において，他のブロガーよりもコミュニティ内話題を早期に投稿している傾向を分析するため，相対的に話題を先取りしている事で先読みブロガーとなり，実際に大きな流行を先取りしているブロガーの発見が困難であった．一方後者の「過去のメジャーな流行語を特定した上で，この過去の流行語に対して事前に言及した頻度を分析する手法 [7] [8]」では，過去のメジャーな流行語を特定した上で，この過去の流行語を早期に投稿している傾向を分析する手法であり，実際に大きな流行を先取りしているブロガーの発見が可能だと考えられる．しかし，単純に，過去の流行語を早期に投稿している傾向を分析するだけではなく，過去の流行語および過去の流行語の流行時の話題を含む記事を早期に投稿している傾向を分析する必要があると考えられる．例えば,「iOS7」が公開されるまでに,「iOS7 を早くダウンロードしたい」といった内容を早期に投稿しているブロガーは先読みブロガーではない.「iOS7 では UI がフラットデザインになる」といったように,「フラットデザイン」といったような流行時特有のキーワードも含めて，早期に記事を投稿しているブロガーである．

そこで先行研究 [8] では，対象流行語語の流行時の話題を抽出し，その話題が投稿された時期 (流行時の話題が語られ始めてから，ピークを迎えるまでの期間) を成長期間として推定する手法について提案してきたが，具体的な推定手法，および推定に必要な流行語の分析については未検討であった．

よって，本稿では，対象流行語の流行時の内容を早期かつ，的確に捉えることができる先読みブロガーを発見するために，成長期間の推定に必要となる流行語の分析に関する実験および評価を行う．

以降，2 章にて，関連研究について述べる．3 章にて，これまで先行研究 [8] において提案を行った，先読みブロガー発見までの全体の流れである，流行語の事前言及頻度分析に基づくブロガー先読み度分析手法について述べる．4 章にて，今回新

たに検討を行った，具体的なシード語の成長期間の推定手法について述べる．5 章にて，成長期間算出に関する評価として，シード語に対する流行のピーク時の話題抽出実験に対する評価について述べる．最後に 6 章にて，まとめと今後の課題について述べる．

2. 関 連 研 究

ブログ等の分析により流行語やトレンドを発見もしくは抽出しようとする関連研究を以下に挙げる．

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 [10]．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 [11]．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 [12]．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 [13]．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 [14]．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 [15]．古川らは，ブログにおける話題の伝搬が語とブロガーの影響力によって起こるという仮説の下で，伝搬の情報から議論の連なりやすい語を重要語として判別する手法を提案している [16]．横山らは，潜在的ディリクレ配分法を用いてブログ記事のトピックを推定することで，情報伝播のネットワークを抽出する枠組みを提案している [17]．小阪らは，注目話題を早期に発見するために，話題頻度の推移を学習データとして用い，話題が全体に波及するかどうかを判別する分類器の作成を行っている [18]．片上らは，電子掲示板のリンク情報に基づき，伝播パターンと流行度合いの測定を行い，サポートベクターマシンにより流行的話題の予測を行っている [19]

以上の通り，既に広まったキーワードの検出，可視化を目的とした研究や，話題の伝播に関する研究は行われているが，過去に流行を先取りしていたブロガーを発見し，そこから流行語やトレンドを効率的に取得することを目指した研究はなされていない．

3. 流行語の事前言及頻度分析に基づくブロガー先読み度分析手法

本章では，提案する先読みブロガー発見手法の処理の流れについて説明する．なお，解析用データとして，kizasi.jp [20] にて保持しているブログデータ（2013 年 9 月 6 日時点で，12,103,387 ブロガー，172,018,786 エントリー）を対象としており，本手法では，先読みブロガー発見のための学習データとして用いる過去の流行語を，シード語と呼称している．

まず，各ブロガーがどの分野について専門性を持っているかを，熟知度という指標でカテゴリ毎（ブロガーコミュニティ毎）に分類する．次に，先読みブロガーを発見するための学習データとなる，シード語（過去の流行語）を抽出し，カテゴリ（トピック毎）に分類する．次に，シード語（過去の流行語）の成長期間の推定を行い，各時点の先読みの価値を先読みポイントとして算出する．最後に，先読みポイントを用いて，先読み傾向の強さをブロガーの先読み度として算出する．

以上の流れを具体的に，括弧内に示した各節において詳細を説明する．

（1） ブロガーグループの分類と熟知度判定（3.1 節）
（2） ブログアーカイブ解析によるシード語の抽出（3.2 節）
（3） シード語の成長期間推定に基づく先読みポイントの算出（3.3 節）
（4） ブロガーの先読み度の算出（3.4 節）

3.1 ブロガーグループの分類と熟知度判定

提案手法では，流行語に対していち早く反応する傾向を有する先読みブロガーの判定を目的としているが，各先読みブロガーがどの分野における先読み能力が高いかを示す必要がある．なぜなら，ある先読みブロガーが「インターネット」に関する話題において先読み能力が高いとしても，「経済」に関する話題において先読み能力が必ずしも高い訳ではないためである．また，我々が発見しようとしている先読みブロガーは，分類された該当分野に対してある程度熟知していることを想定している．したがって，各ブロガーを話題別のブロガーグループに分類し，その各カテゴリ内におけるブロガーの熟知度に基づいたランキングを行う．

なお，ブロガーグループは，著者らが過去に開発したブロガーの潜在的なコミュニティの分類とその熟知度レベルによるランキングシステム [9] において作成された熟知グループを採用する．以下，ブロガーが過去に投稿したエントリに含まれる「あるトピックを表すキーワードおよびこれに関連する特徴語」の頻度から，そのキーワードが表すブロガーの熟知度を算出し，熟知グループを特定する過程について説明する．

3.1.1 熟知グループおよび共起語辞書の作成

あるトピックに関して熟知するブロガーの集合を「熟知グループ」と呼び，これに基いてブロガーを分類する．まず，「熟知グループ名」として，ブログでよく言及されるトピックを自動抽出したキーワード群と，独自に開発した生活体験シソーラス LETS を用いて，約 13,000 程度の分類を作成する．（ただし，本研究においては，13,000 の分類カテゴリでは，話題がやや細かすぎるため，これらをグルーピングした 120 件程度のカテゴリを採用する．）次に，直近 2 年分のブログエントリを対象とし，「熟知グループ名」との共起度が高い 400 語のキーワードを抽出し，共起語辞書を作成する．

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．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．ブログ記事の総語数を 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．

$$LogLog\ Score = log\frac{N_{xy} \cdot N}{N_x \cdot N_y} \cdot logN_{xy} \quad (1)$$

なお，共起語の選定には自らの生活体験を表すような語句を優先的に採用し，不適切な語句を排除することにより，実体験に即したブログエントリを記述するブロガーを分類できる精度を上げている．また，新しいトピックに対応するため，熟知グループは 1 週間間隔で更新している．

3.1.2 ブロガー熟知度スコアに基づく熟知グループ判定

ブロガーがどの熟知グループに所属しているかを判定するため，熟知度スコアを算出する．基本的なアイデアとしては，対象熟知グループに関連するトピックを含んだエントリの投稿数に基づき算出する．なお，各ブロガーは熟知グループ毎に異なる複数の熟知度スコアを有する．つまり，あるブロガーが「経済」と「政治」に関する熟知グループに属する場合，このブロガーは「経済」に対する熟知度スコアと「政治」に関する熟知度スコアを別々に有することになる．

ここで，対象熟知グループ g_i に対する，あるブログ記事 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．

$$relevance_{g_i}(e_k) = \sum_{j=1}^{n} \alpha_{ij} \cdot \beta_{ji} \cdot \gamma_{ij} \quad (2)$$

ただし，n はこの熟知グループ g_i の共起語数であり，今回は $n = 400$ である．α_{ij} は熟知グループ g_i の共起度順位 j 番目の共起語 ω_{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，共起度順位の高い語句の重みを大きくするための工夫であり，共起度順位 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} は熟知グループ g_i の j 番目の共起語 ω_{ij} の共起度である．そして，γ_{ij} は順位 j 番目の共起語 ω_{ij} が該当記事 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．

次に，対象熟知グループ g_i に対するブロガー b の熟知度スコアを $knowledge_{g_i}(b)$ 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 はブロガー b が投稿した記事である．m はブロガー b が対象期間内に投稿した記事数である．l はブロガー b が対象期間内に投稿した記事に出現した共起語数である ($l \leqq n$)．したがって，l/n はブロガー 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，関連性の低い記事を大量に投稿した場合に，そのブロガーの熟知度が高くなってしまう問題に対して，記事数の増加の影響を緩和させている．最終的に対象となるブログコミュニティに対する熟知度スコアが，設定した閾値を超えれば，そのブロガーが属するものと判定する．

3.2 ブログアーカイブ解析によるシード語の抽出

本節では，ブログアーカイブより，シード語候補を抽出し，カテゴリ分類後，シード語候補のブログ投稿数に基づく影響度算出により，シード語の認定を行う過程について説明する．

3.2.1 ブログ分析によるシード語候補の抽出

提案手法ではシード語（過去の流行語）を使って，ブロガーの先読み分析を行うため，ブログ分析によりシード語候補の抽出を行う必要がある．

シード語候補を抽出するにあたり，ブログで話題になったキーワードを取り上げている kizasi.jp [20] の話題ランキング (アーカイブ 2 年分) を利用する．手順としては，まず，kizasi.jp より上位 100 までに入ったキーワードを抽出する．その後，抽出したキーワードから重複語，一般語，総出現数が少ないキーワード，周期性のあるキーワードを除外し，残ったキーワードをシード語候補とする．周期性のあるキーワードとは，特定の周期で出現するキーワードである．例えば，1 年周期であれば「夏祭り」や「正月」等のキーワードが挙げられる．4 年周期の「オリンピック」「FIFA ワールドカップ」「WBC」，1 ヶ月周期の「給料日」なども該当する．

3.2.2 シード語候補のカテゴリ分類

各シード語が，どの分野に関連する流行語であるのかを判別するため，シード語候補をカテゴリ毎に分類する必要がある．また，最終的に，先読みブロガーを効率よく発見するためには，シード語候補について内容を熟知している熟知ブロガーを中心に分析することを考えている．その意味からも，各シード語候補がどのカテゴリと意味的に近いのかを判定する必要がある．なお，分類カテゴリとしては，3.1 節にて説明した熟知グループを利用する．

分類方法としては，シード語候補と熟知グループの意味的な近さを表す関連度を算出することによって行う．この関連度は，“シード語候補の共起語集合”と“熟知グループの共起語集合”の類似度により表現する．

シード語候補の共起語集合は，全ブログ記事中における共起度の高いキーワード上位 400 個としている．熟知グループの共起語集合は，各熟知グループに属するブロガーが投稿した該当カテゴリに関連するブログ記事中における，共起度の高いキーワード上位 400 語としている．

なお，先行研究 [8] でのカテゴリ分類評価実験の結果を踏まえ，共起語集合間の類似度算出手法は，共起度順に重みを付与したコサイン類似度を用い，最も類似度の高い上位 1,2 を関連していると判定する手法を用いることを検討している．

3.2.3 影響度に基づくシード語の認定

提案手法では，シード語が示す過去の流行語を，世間に広まる以前から言及していたブロガーを先読みブロガーと認定しようとしているため，認定されるシード語はある程度重要なキーワードに絞る必要がある．シード語候補の重要性の評価は，ブログにおける該当キーワードの投稿数のピーク以降の期間 T に

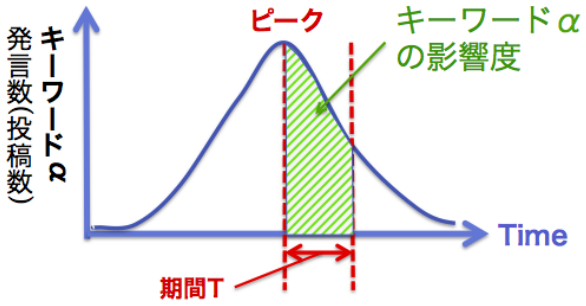

図 1　影響度の概念図

おけるブログ投稿数の累計を，シード語候補の影響度として算出することで行う．

具体的な算出方法としては，シード語候補毎にブログ投稿数を調べ，該当シード語候補の過去 2 年間の投稿数に対し，必要に応じて移動平均を算出し，投稿数のピークを確認する．このピークを迎えた時点が社会的認知が最も高くなった時点であるといえる．このピーク以降の期間 T における投稿数の累計が非常に少なくなっている場合には，このシード語候補はピーク後に世間から忘れ去られるようなキーワードであると考えられるため，ピーク以降も投稿数があまり減少しないようなシード語候補を社会的な影響度が高いキーワードであると判断し，シード語として認定する（図 1 参照）．

なお，シード語候補毎に，関連度の高い熟知グループを幾つか求める．さらに各熟知グループ毎に，影響度の高い上位数個のシード語候補を，その熟知グループのシード語とする．

3.3　シード語の成長期間推定に基づく先読みポイントの算出

本節では，シード語の成長期間推定に基づく，各記事のある時点の先読みの価値を表現した，先読みポイントの算出手法について述べる．なお，今回新たに検討を行った，具体的なシード語の成長期間の推定手法については，4 章で述べる．

先読みポイントとは，シード語の成長期間内において，対象時点の先読みの価値を表したスコアである．この先読みポイントは，シード語の成長期間内に投稿された（このシード語に関する）ブログ記事に対して付与され，この期間の開始時点が最も高く，終了時点（ピーク時）が最も低い値となる．ここで，ブログ記事 $entry_i$ に付与される先読みポイント $PredictionPoint_i$ の算出式を式 (4) に示す．

$$PredictionPoint_i = \frac{(entry_{all} + 1) - order_i}{entry_{all}} \qquad (4)$$

$entry_{all}$ は，シード語の成長期間内にて対象シード語について投稿しているエントリ数である．$order_i$ は，シード語の成長期間内において，シード語を早期に投稿した順序である．すなわち，エントリ数 $entry_{all}$ が 100 の場合には，$order_i$ が 1 から 100 のエントリに対する先読みポイントは，順番に $1, 0.99, \cdots, 0.02, 0.01$ という値が付与される．

3.4　ブロガー先読み度の算出

本節では，各シード語に対するブロガー先読み度の算出方式について説明する．各ブロガーのブロガー先読み度が高く算出されるための条件を以下に示す．

- 対象シード語に関するブログ記事を，シード語の成長期間内で早期に投稿している．
- 上記ブログ記事の投稿数が多く，その内容がシード語のピーク時の話題と類似している（図 2 参照）．

そこで，あるシード語 A に対する，あるブロガー x の先読み度 $PredictionScore_{(A,x)}$ の算出式を以下に示す．

$$PredictionScore_{(A,x)} = \sum_{k=1}^{N} Sim(D_A, entry_k) \times PredictionPoint_k \qquad (5)$$

N は，ブロガー x が成長期間内に投稿したブログ記事数である．D_A は，シード語 A のピーク時の共起語集合であり，$Sim(D_A, entry_k)$ は，シード語 A のピーク時の話題とブログ記事 $entry_k$ との類似度である．$PredictionPoint_k$ は，3.4 節にて説明したブログ記事 $entry_k$ に対する先読みポイントである．

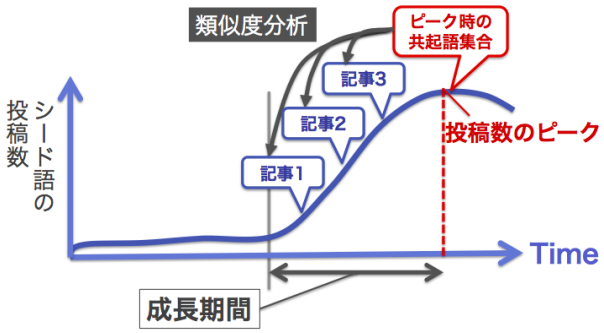

図 2　ブロガー先読み度判定のためのブログ記事の類似度分析

このブロガー先読み度を，対象カテゴリにおける複数のシード語において算出することにより，ある特定分野における「先読みブロガー」は，その後もこの分野に関しては「先読みブロガー」であり続ける可能性が高いかどうかについて評価を行うことができると考えている．

4.　シード語の成長期間の推定手法

本章では，新たに検討したシード語の成長期間の推定手法について述べる．

シード語の成長期間を推定するためには，まずはシード語が表す流行語について，どの時点から語られ始めたのかを推定する．このとき，流行のピーク時の話題の内容とかけ離れていないことを確認する必要がある．

例えば，「iOS 7」がシード語である場合，「iOS 7 はいつリリースされるんだろう？」といった記述内容は，先読みブロガーでなくとも投稿することが可能である．つまり，「iOS 7」というシード語のみを記述するだけでは，流行を先読みしているとはいえない．一方，「iOS 7 では UI がフラットデザインが採用されるらしい」という内容を，「iOS 7」リリース前からブログに投稿していれば，この関連カテゴリに関するある程度の先読み能力があると考えられる．

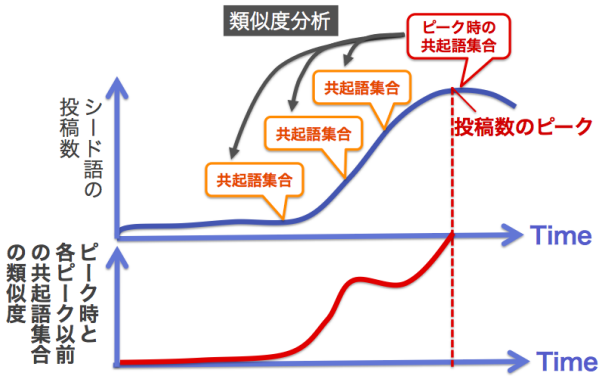

図 3　ピーク時の話題とこれ以前の話題との類似度計算

そこで,「iOS 7」というシード語が流行のピークを迎えた際に，どのような共起語と共に語られているかを，その共起語集合により表現し，これ以前の期間におけるシード語「iOS 7」の共起語集合との類似度計算を行うことにより，流行のピーク時に話題となったキーワード（「iOS 7」の場合，フラットデザイン等）を含めて，そのシード語が示す話題がどの時点から語られ始めているのかを推定する方法を考案している（図 3 参照）．しかし，この推定を精度良く行うためには，流行のピーク時に話題となるような「流行時特有キーワード」を抽出する必要がある．よって，4.1 節にて，シード語に対する流行時の話題の抽出手法について述べ，4.2 節にて，流行時の話題とこれ以前の話題との類似度計算による成長期間算出手法について述べる．

4.1　シード語に対する流行のピーク時の話題抽出手法

本節では，シード語の流行時の話題の抽出手法について述べる．シード語と共に，流行時周辺で投稿されたキーワードを抽出するには，シード語に対する共起語を抽出することが考えられる．しかし，単純に，共起回数を基に抽出した共起語集合だけでは，一般語，頻繁にシード語と共に投稿されているキーワードや，ブログでよく話題になるようなキーワード等も含まれ,「流行時特有キーワード」をより多く含むようなキーワード集合を作成するのは困難であると考えられる．したがって，以下の手順を踏まえ,「流行時特有キーワード」をより多く抽出することを目指す．

手順 1　シード語の出現数を計算する
手順 2　シード語の流行時 (最大出現数となる時点の前後 1 週間) における共起語集合を抽出する
手順 3　手順 2 の各共起語の出現数を計算する
手順 4　シード語の出現数 (手順 1) と各共起語の出現数 (手順 3) のスピアマンの順位相関係数の高い上位キーワードを抽出する

スピアマンの順位相関係数を用いることで，シード語と出現数の変化と類似している共起語の抽出が可能である．つまり，シード語に寄り添うように出現数が増減してきたキーワードの抽出が可能であると考えている．

4.2　流行のピーク時の話題とこれ以前の話題との類似度計算

本節では，シード語の流行時の話題とこれ以前の話題との類似度計算により，成長期間を推定する手法について述べる．シード語のピーク時の話題とこれ以前の話題との類似度により，シード語が出現した時期から投稿数が最大となる時点までの変化を調べることができる．この類似度曲線の立ち上がり付近において，シード語と共起する「流行時特有キーワード」について話題がブログ上で語られ始めたと考えられる．

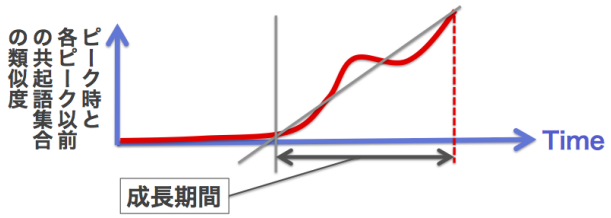

図 4　シード語の成長期間の判定 (1)

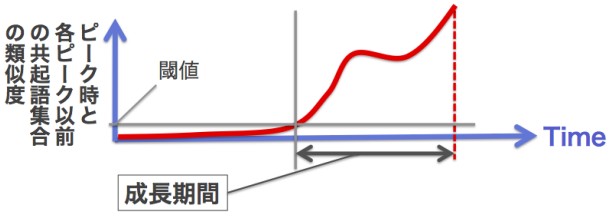

図 5　シード語の成長期間の判定 (2)

シード語のピーク時の話題とこれ以前の話題との類似度により，シード語が出現した時期から投稿数が最大となる時点までの変化を調べることができる．この類似度曲線の立ち上がり付近において，シード語か示す話題がブログ上で語られ始めたと考えられる．しかしながら，これ以前の話題とピーク時の話題との類似度は，完全にゼロとなる保証はなく，バックグランドノイズのような形で，それ程高くない類似度となることが考えられる．したがって，シード語の成長期間の開始時点を判定する手法としては，以下の 2 つの手法を検討している．

（1）“ピーク時の話題と成長期間の話題との類似度”の一次近似線の切片を，シード語の成長期間の開始時点とする（図 4）．

（2）“ピーク時の話題と成長期間の話題との類似度”に，適当な閾値 θ を設定し，閾値 θ 最初に超えた時点を，シード語の成長期間の開始時点とする（図 5）．

以上により，シード語の成長期間の開始時点を求め，ここからシード語の投稿数が最大となる時点（ピーク）までを，このシード語の成長期間として判定する．

5.　シード語の流行時の話題抽出実験および評価

本章では，シード語の流行時の話題抽出実験および評価について述べる．

流行語先読みブロガーを発見するために，対象シード語の成長期間において，対象シード語を含む記事を早期に投稿している傾向を分析する手法を検討している．この手法は,「流行語先読みブロガーは，流行時の話題を，流行以前から早期に投稿している」という仮定に基づき，流行時の話題と，流行以前の各

話題を比較することで，流行時の話題が出現し始め流行するまでの成長期間の推定し，期間に応じた先読みの価値を設定しようとする手法である．この成長期間を適切に推定するためには，流行以前において，流行時の話題が，どの辺りから出現し始めたかを適切に判定する必要がある．そこで，流行時に話題になるような「流行時特有キーワード」をより多く含むような，流行時の話題を抽出することにより，精度の良い推定を行うことを検討している．

5.1 実験手法

本節では，シード語の流行時の話題抽出において，「流行時特有キーワード」をより多く含む手法を調査する．具体的には，以下の 3 手法において，対象シード語のピーク時前後の共起語集合を素データとしたとき，「流行時特有キーワード」をより多く抽出可能な手法の判定である．

（1） 共起回数の高いキーワード (ベースライン)

（2） loglog スコアの高いキーワード (LogLog スコア順)

（3） 対象シード語の出現数とその共起語の出現数のスピアマンの順位相関係数の高いキーワード (順位相関係数順)

LogLog スコアは，相互情報量とも呼ばれている MI スコアに対し，共起度頻度を積極的に評価する共起度の尺度であり，シード語と共起キーワードの中で，特徴的な共起語が抽出できる．一方，スピアマンの順位相関係数は，シード語の出現数とその共起語の出現数において，時系列的な分布に相関があるかどうかを評価するノンパラメトリックな指標であり，時間の経過と共に，シード語の出現数の増加傾向の変動と類似している共起語が抽出できる．

本実験で用いたシード語は，2012 年〜2013 年において DVD の売上が上位であったアニメの 13 タイトルである．また，解析データとして，kizasi.jp で保持しているブログデータ（2011.9.28 から 2013.12.04）を用い，対象シード語が最も投稿された時点の前後 1 周間（合計 3 週間）を基に，対象シード語毎に共起語集合を抽出した．なお，対象となる共起語は，対象シード語と共起するキーワードではなく，対象シード語および “アニメ” と共起するキーワードである．さらに，スピアマンの順位相関係数を計算するために，2011.9.28 より 1 週間毎のキーワードの出現数も抽出した．

5.2 評価手法

「ベースライン」「LogLog スコア順」「順位相関係数順」の内，正解データとなる「流行時特有キーワード」および「関連語」をどの程度含んでいるかを適合率，再現率で評価する．

なお，「流行時特有キーワード」および「関連語」は，対象シード語が最も投稿された時点の前後 1 周間（合計 3 週間）における対象シード語に対する各共起語について，以下の項目のアンケートを実施し，　項目 3 と回答されたキーワードを「流行時特有キーワード」とした．

項目 1　対象シード語と関連するキーワードである（関連語）
項目 2　対象シード語と関連しないキーワードである（ノイズ）
項目 3　対象シード語と関連するキーワードかつ，流行時に話題となるキーワードである（流行時特有キーワード）

アンケートの対象者は，対象シード語（対象アニメタイトル）について，実際に全て閲覧した人であり，得られた回答数は，表 1 のようになっている．

表 1　実験に用いた対象流行語，アンケート回答数，およびその共起語数

対象シード語	アンケート回答数	共起語数
Free!	3	369
PSYCHO-PASS	3	542
ジョジョの奇妙な冒険	4	982
ソードアート・オンライン	3	521
とある科学の超電磁砲 S	2	204
ニャル子さん	3	603
はたらく魔王さま	3	263
ラブライブ	4	483
リトルバスターズ	3	692
進撃の巨人	2	991
翠星のガルガンティア	2	350
中二病でも恋がしたい	3	1589
氷菓	3	1252

5.3 評価結果

対象シード語毎に，ベースラインに基づく手法の共起語集合，loglog スコア順に基づく手法の共起語集合，順位相関係数順に基づく手法の共起語集合の適合率を図 6 に，再現率を図 7 に示す．さらに，対象シード語毎に，アンケートにおいて「関連語」と回答された結果についても，同様に適合率を図 8 に，再現率を図 9 に示す．

この図 6，7 の結果より，「流行時特有キーワード」を最も多く含むのは，対象シード語と共に増加するようなキーワードを抽出した順位相関係数順に基づく手法であった．また，単純に共起回数を考慮したベースラインよりも，対象シード語に対して特徴的なキーワードを抽出できる loglog スコア順の方が「流行時特有キーワード」をより多く含む傾向があることが分かった．また，順位相関係数順の再現率が 50 % を越えるのに対し，適合率が約 15 % 程度だった事から，全体の約半数の「流行時特有キーワード」の抽出に成功した事を示していると考えられる．また，図 8，9 の結果より「関連語」をより多く含むかどうかについては，どの手法も大きな差はない事が分かった．

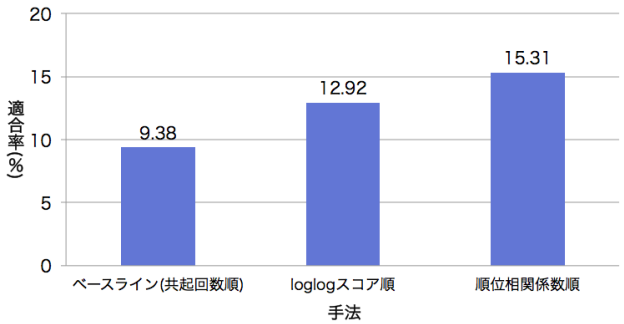

図 6　各手法に基づく対象シード語の共起語上位 100 語に対する「流行時特有のキーワード」の平均適合率

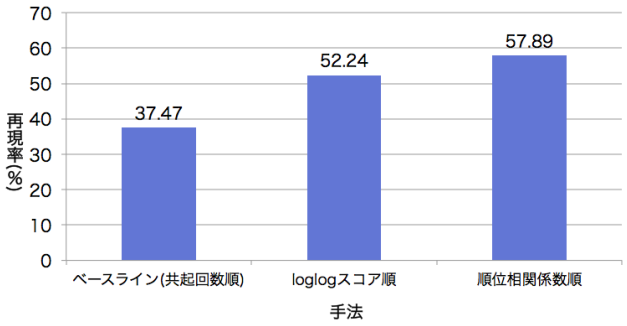

図 7　各手法に基づく対象シード語の共起語上位 100 語に対する「流行時特有のキーワード」の平均再現率

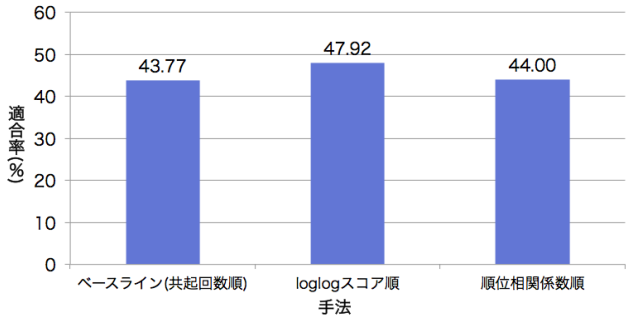

図 8　各手法に基づく対象シード語の共起語上位 100 語に対する「関連語」の平均適合率

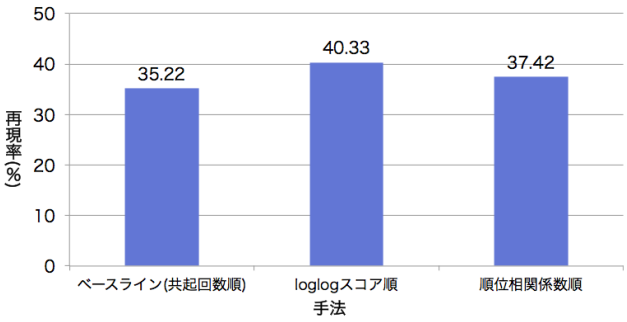

図 9　各手法に基づく対象シード語の共起語上位 100 語に対する「関連語」の平均適合率

以上より，現在検討している手法のなかで,「流行時特有キーワード」をより多く抽出するためには，シード語と共に成長してきたキーワードを取得できる，順位相関係数順を用いる事が有効であると考えられる．

6.　まとめと今後の課題

我々は，流行に鋭敏に反応するブロガー（先読みブロガー）群を発見し，彼らの発信情報から流行語候補を早期発見する手法の開発を目指している．本論文では，先行研究で未検証であった，シード語の成長期間算出が先読みブロガーの発見に繋がるかどうかを評価するために，成長期間算出において重要な「流行時特有キーワード」の抽出について検討を行った．実験の結果，シード語と共に成長してきたキーワードを抽出することが,「流行時特有キーワード」の発見に繋がることが確認できた．

今後は，提案手法の有効性を確かめるため，さらなる実験を行い，成長期間算出の詳細な評価を行うと共に，未実装部分である，“シード語の成長期間推定に基づく先読みポイントの算出（3.3 節）”，“ブロガーの先読み度の算出手法（3.4 節）” の実装と評価を行うことで，実用化に向けた取り組むを進める予定である．

7.　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．

文　　献

[1] 「現代用語の基礎知識」選，2013 ユーキャン新語・流行語大賞発表，http://singo.jiyu.co.jp/
[2] 「2013 年ヒット商品ベスト 30」が発表！ 1 位は「コンビニコーヒー」，日経トレンディネット，http://trendy.nikkeibp.co.jp/article/pickup/20131028/1053200/
[3] Shinsuke Nakajima, Jianwei Zhang, Yoichi Inagaki and Reyn Nakamoto. Early Detection of Buzzwords Based on Large-scale Time-Series Analysis of Blog Entries, 23rd ACM Conference on Hypertext and Social Media (ACM Hypertext 2012), pp.275-284, June 2012.
[4] 中島伸介，張建偉，稲垣陽一，中本レン，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，情報処理学会論文誌:データベース (TOD56)，2013 年．
[5] Shinsuke NAKAJIMA, Adam JATOWT, Yoichi INAGAKI, Reyn NAKAMOTO, Jianwei ZHANG, Katsumi TANAKA: “Finding Good Predictors in Blogsphere Based on Temporal Analysis of Posting Patterns”, DBSJ Journal, Vol.10, No.1, pp.13-18, June 2011.
[6] 朝永聖也，中島伸介，Adam JATOWT，稲垣陽一，Reyn NAKAMOTO，張 建偉，田中克己． ブログ記事の時系列分析に基づくブロガー先読み度分析手法の提案． 第 3 回ソーシャルコンピューティングシンポジウム（SoC2012），SoC2012 講演論文集 pp.79-84，2012 年 6 月．
[7] 朝永聖也, 中島伸介, 張建偉, 稲垣陽一, 中本レン， 流行語の事前言及頻度分析に基づくブロガー先読み度判定手法の提案， 第 5 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 2013) C1-2，2013 年 3 月．
[8] 朝永聖也，中島伸介，稲垣陽一，中本レン，小倉 僚，張 建偉，流行語に対する早期言及頻度分析に基づくブロガー先読み度判定手法の提案，情報処理学会研究報告 データベース・システム研究会報告，2013-DBS-158(1)，1-8，2013-11-19．
[9] 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．
[10] 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．
[11] 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．
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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（日本知能情報ファジィ学会誌）, Vol.21, No.4, pp.557–566，2009 年.
[17] 横山 正太朗，江口 浩二，大川 剛直，“潜在トピックを用いたブログ空間からの情報伝搬ネットワーク抽出”，電子情報通信学会論文誌，Vol.J93-D，No.3，pp.180-188 (2010).
[18] 小阪有平，安村禎明，上原邦昭，“ブログのカテゴリ分類に基づく注目話題の早期検出”，人工知能学会全国大会 (第 23 回) 論文集，3B2-1 (2009).
[19] 片上大輔, 大久保亮介, 新田克己, 電子掲示板のリンク情報に基づく流行的話題の予測, 人工知能学会論文誌, Vol. 21 (2006) No. 6 P 459-472
[20] kizasi.jp:ブログから、話題を知る、きざしを見つける, http://kizasi.jp/