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

朝永　聖也† 　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: †{g0946920,nakajima}@cse.kyoto-su.ac.jp, ††zhangjw@a.tsukuba-tech.ac.jp,
†††{inagaki,reyn}@kizasi.jp

あらまし　有望な流行語候補を世間に広まる前に発見することを目標とし，流行に対して鋭敏な反応をしている先読みブロガーの発見を目指す．そこで我々は，メジャーな流行語に対して，事前に言及した頻度等を分析することで，ブロガーの先読み度を算出する手法を提案する．本稿では，話題を先読みしている期間に応じたスコア算出方式を求め，スコア上位ブロガーを先読みブロガーと判定する方法について検討を行ったので報告する．

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

Method for Measuring Bloggers' Buzzword Prediction Ability Based on Analysis of Past Buzzword Mention Frequency

Seiya TOMONAGA†, Shinsuke NAKAJIMA†, Jianwei ZHANG††, Yoichi INAGAKI†††, and Reyn NAKAMOTO†††

† Faculty of Computer Science and Engineering, Kyoto Sangyo University　Motoyama, Kamigamo, Kita-ku, Kyoto-City, Kyoto, 603–8555 Japan
†† Faculty of Industrial Technology, Tsukuba University of Technology　Amakubo 4–3–15, Tsukuba-City, Ibaraki, 305–8520 Japan
††† Kizasi Company, Inc.　6th floor Tomoe Nihonbashi Building, 20–14 Hakozaki-Cho, Nihonbashi, Chuo-ku, Tokyo, 103–0015 Japan
E-mail: †{g0946920,nakajima}@cse.kyoto-su.ac.jp, ††zhangjw@a.tsukuba-tech.ac.jp,
†††{inagaki,reyn}@kizasi.jp

Abstract　Our goal is to discover good predictors in blogsphere in order to realize early detection of buzzwords. Thus, we propose a method for measuring bloggers' buzzword prediction ability based on analysis of past buzzword-mention frequency. This paper discribes how to calculate the score of buzzword prediction ability according to a period of time predicting buzzword candidates.

Key words　blog mining, blogger's buzzword prediction ability

1. はじめに

流行語は世間に知れ渡ってから始めて知ることが多く，流行する前に知ることは大変困難である．しかし，マーケティングの観点において，流行を先取りすることは重要である．一方，多くの人達が利用しているブログでは，世間に知れ渡っていない流行語候補が，限られた人々の中で語られている可能性がある．そこで，流行に鋭敏な反応をし，話題を先取りして記事にしている「先読みブロガー」を発見することができれば，流行語候補の発見に繋がると考え，著者らは「先読みブロガー」の発見に着目した研究を行なっている．これまでの研究で，ブログ記事の履歴と世間の話題の推移を時系列的に分析することに基づいた手法 [1] を，既に報告している．本稿では，流行語が世間に広まる前に投稿している頻度を分析することに基づく手法について検討する．なお，流行語を発見するアプローチとして，世代間やコミュニティ間の流行語の広がり方を分析した先行研究 [2] [3] があるが，本稿では，世代間やコミュニティ間で流行語が広がる過程ではなく，コミュニティ内で流行語が広が

る過程に着目するため，本稿で扱う流行語とはコミュニティ内で話題となったキーワードである．

以降，流行を先取りしている「先読みブロガー」は，過去の流行語についても，世間で話題になる前に投稿していると仮定し，より早く流行語を話題となる前に言及している頻度が高いブロガー程スコアを高くつける先読み度を算出する．これにより,「先読みブロガー」を発見する手法について検討検討する．2 章で，関連研究について述べる．3 章で，提案手法で使用する世間の話題の推移を表現した辞書の作成手法について述べる．4 章で提案手法である「流行語の事前言及頻度分析に基づく先読み度分析」について述べる．5 章で話題を先読みしている期間に応じたスコア算出方式について考察を述べる．最後に，6 章でまとめと，今後について述べる．

2. 関 連 研 究

ブログ等の分析に基づくトレンド分析に関する関連研究として以下が挙げられる．

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 [5]．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 [6]．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 [7]．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 [8]．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 [9]．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 [10]．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ごを重要語として判別する手法を提案している [11]．

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，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. 世間の話題推移を表現する辞書の作成

本研究では，どのようなコミュニティが存在し，過去にどのようなキーワードが話題となったかを分析する必要がある．そこで，ブロガーの興味・関心ごとにコミュニティを分類し，ブログコミュニティごとの話題推移をキーワードの集合で表現する．具体的には，月ごとに対象ブログコミュニティで最もよく語られた共起語等を関連語として抽出し，400 語の関連語辞書を作成する．この関連語辞書を 24ヶ月分用意することで，世間の話題の推移を表現していると考える．なお，ブログコミュニティは，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 [4]」において作成された熟知グループであり，2013 年 3 月 22 日時点で，11,494,400 ブロガーの 153,554,298 エントリーを対象としている．以下で，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 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，この「熟知グループ」に基いてコミュニティを分類する．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 12,000 程度の分類を作成する．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．

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

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．

3.2 ブロガー熟知度スコアに基づく熟知グループの特定

ブロガーがどの熟知度グループに所属しているかを特定するため，熟知度スコアを算出する．基本的なアイディアとしては，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，各ブロガーは熟知度グループごと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．

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

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．

また，広告収入や特定サイトへの誘導を目的として自動的に生成されたスパムブログを排除するため，投稿エントリ数と投稿時間数に基いたルールを設定して，スパムと疑わしいブログを検知する半自動スパムフィルタと，スパムブログを学習したベイズ分類器による全自動スパムフィルタを作成し，熟知グループでノイズとなるブロガーの排除の精度を向上させている．

4. 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

本章では，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について述べる．基本的なアイデアとしては，ブログコミュニティ内で話題となる前に流行語 (話題となったキーワード) について言及していた頻度に基づき，「先読みブロガー」を判定する．つまり，「先読みブロガー」であるなら，話題となった時点より過去において，流行語 (話題となったキーワード) を含む記事を投稿していると仮定している．具体的には，対象コミュニティの話題の変化を表現した 24ヶ月分の関連語辞書における関連語に着目し，月ごとに関連語の共起度順位スコアを割り振る．次に，時間の経過と共に共起度の順位が上昇していく関連語は，対象コミュニティにおいて話題となった語句と捉え，後に共起度の順位が上昇していく関連語ほど高い先読みポテンシャルを付与する．その後，先読みポテンシャルが高い関連語を含む記事を多く投稿しているブロガーに，高い先読み度を算出する．これにより，先読み度に基づいたブロガーランキングにおいて，上位のブロガーを「先読みブロガー」と考えることができる．

4.1 先読みポテンシャルに基づく関連語の重み付け

本節では，関連語に対し，月ごとに先読みポテンシャルを付与する手法について述べる．まず，24ヶ月間の関連語辞書における各関連語に対し，月ごとの共起度順位スコア R を求める．共起度順位スコア R は，以下の式 (4) で求める．なお，N は 1ヶ月分の関連語辞書に含まれる語彙数である．

$$R(\text{共起度順位スコア}) = (N + 1) - \text{共起度の順位} \quad (4)$$

関連語ごとに R を集計することにより，図 1 のように関連語の共起度順位スコアの時系列変化を分析できるようになる．

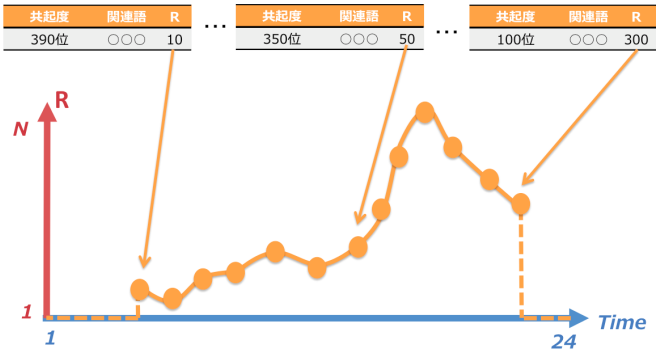

図 1　R 算出に基づく関連語「○○○」の時系列変化グラフ

次に，対象となる月 T_i の関連語から先読み期間 s までの期間において，共起度順位スコア R が最大となる月 T_{peak} を求め，月 T_{peak} から月 T_i までの差と，月 T_{peak} 時の共起度順位スコア R_{peak} から月 T_i 時の共起度順位スコア R_i までの差の面積を求める．この面積が先読みポテンシャルである．先読みポテンシャルは，対象ブログコミュニティにおいて話題となるより前ほど高くなる．つまり，先読みポテンシャルは，共起度の順位差と共起度順位が最大となる時点までの差の面積に基いて算出されるため，共起度順位スコア R が大きく上昇する時点より過去になるほど高くなる．先読みポテンシャルの算出式を式 (5) に，算出イメージを図 2 に示す．

$$P(c, w, t, s) = (T_{peak} - T_i) \times (R_{peak} - R_i) \quad (5)$$

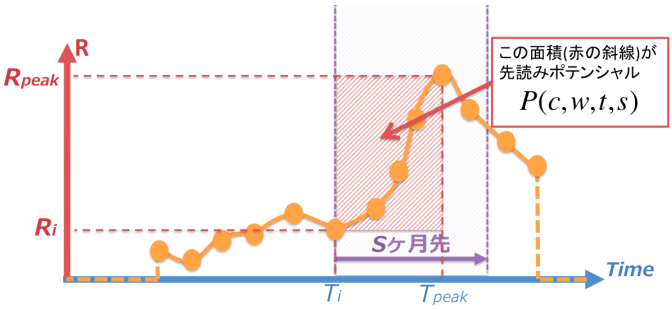

図 2　c トピックにおける関連語 w の T_i 時点における先読み期間 s の先読みポテンシャル P

なお，c はトピック，w は関連語，t は対象月 T_i，先読み期間 s は月 T_i から共起度順位が最大となる月 T_{peak} を求めるときに考慮する期間，つまりブロガーが先読みしている期間を想

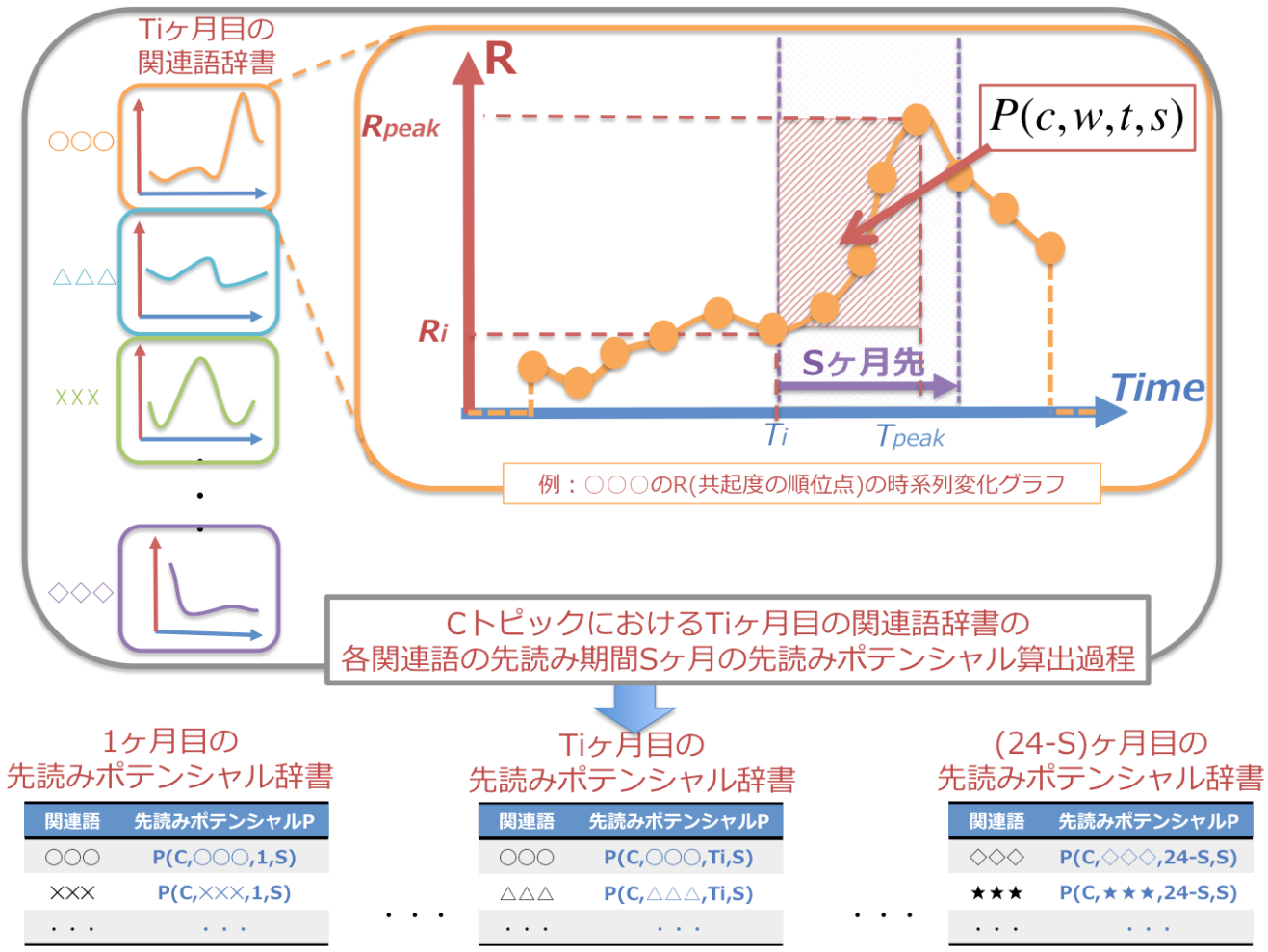

図 3　先読みポテンシャル算出過程

定した期間である．先読みポテンシャルの算出は，図 3 のように，月ごとに各関連語に対して行い，最終的には，対象トピックごとの 1 から (24 – 先読み期間 s)ヶ月目までの先読みポテンシャル辞書が得られる．

4.2　先読みポテンシャルに基づくブロガー先読み度算出手法

本節では，ブロガーの先読み度を算出する手法について述べる．対象コミュニティの各ブロガーに対し，対象トピックの 24ヶ月分の関連語辞書における関連語を，どの月に，どの程度の頻度で言及しているか調べる．これにより，話題を先取りしている傾向の強い「先読みブロガー」を発見する．ただし，ブロガーごとに投稿している記事数が違うため，対象トピックに関する記事数を考慮し，1 記事当たりの先読みポテンシャルの合計値を求める．また，月ごとに先読みポテンシャルは変わるため，ブロガーがどの程度先読みしているかを示す先読み度は，月ごとに算出する．以上より，先読み度 $pScore$ は，月ごとに，関連語 w の先読みポテンシャル $P(c,w,t,s)$ と関連語 w を含む記事数 E を掛け合わせたスコアを基に算出する．ただし，多数の記事を投稿したブロガーと少数の記事を投稿したブロガーで不公平な差が生じるのを防ぐため，c トピックに関する記事を投稿した頻度で割っている．これより，先読み度は式 (6) で算出される

$$pScore(c,t,s) = \frac{1}{E(c,\forall_w \in U,t)} \sum_{w=1}^{N} [P(c,w,t,s) \times E(c,w,t)] \tag{6}$$

ただし，N は 1 ヶ月分関連語辞書に含まれる語彙数である．また，$E(c,\forall_w \in U,t)$ は tヶ月目の c トピックに関する記事数，U は c トピックの関連語辞書に含む全関連語，$E(c,w,t)$ は tヶ月目の w を含む c トピックに関する記事数である．

4.3　先読みポテンシャル算出の問題点

本節では，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における，先読みポテンシャルの算出に対する問題点について述べる．本手法では，各月の関連語に対して，より以前からブログ内で言及していれば高い先読みポテンシャルを付与することにしている．しかし，実際にどの程度の期間を先読みしたのかを知ることは困難であり，先取り期間 s を適切に設定することは容易ではない．ただし，算出される先読みポテンシャルの大きさは，この先取り期間 s の値に大きく影響されるため，s を適切に設定することは重要であるといえる．また，現状の算出手法では，流行語として適切な語の先読みポテンシャルを高く算出し，一般語等の先読みポテンシャルを低く算出できているか不明である．そこで，現手法の先読みポテン

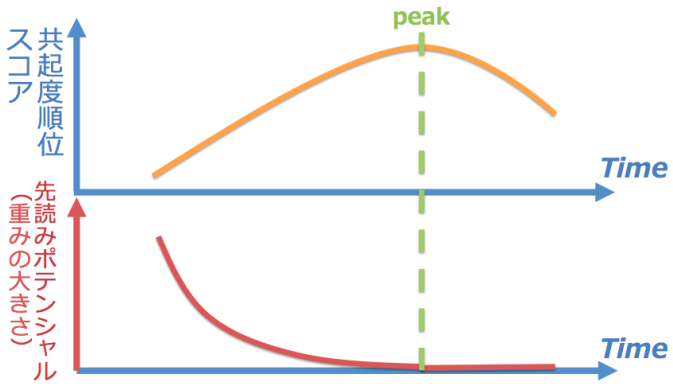

図 4　世間に広がっていくキーワード先読みポテンシャルの変化

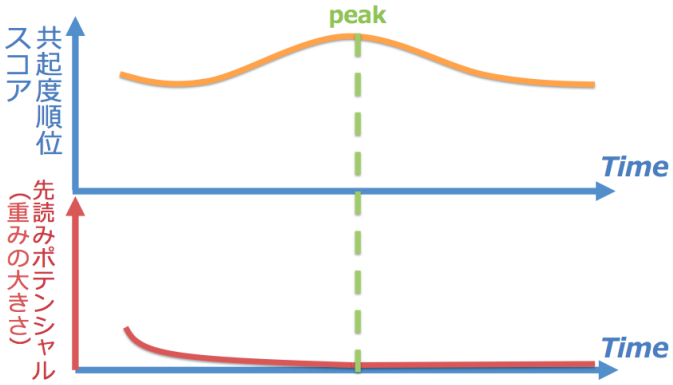

図 5　既に話題となったキーワードの先読みポテンシャルの変化

シャル算出における問題点について考察する．

まず，共起度順位スコアの時系列変化に対する先読みポテンシャル算出のイメージを，図 4，図 5 に示す．コミュニティ内で時が経過するごとに話題となった関連語は図 4 のように，話題となる時点より過去であるほど，先読みポテンシャルは高くなる．一方，コミュニティでよく語られる関連語は共起度順位スコアは高い，既に話題となっているため，共起度順位スコアの上昇率が低く，図 5 のように，先読みポテンシャルは低くなる．しかし，共起度順位が周期的に変化するような関連語は，共起度順位の上昇率が高くなり，先読みポテンシャルが高くなってしまう可能性がため，何らかの対策が必要がある．

ここで，実際の対象トピックに対して先読みポテンシャルを算出し，先読みポテンシャルの高い上位の関連語の例を示す．レシピトピックにおいて，共起度順位スコアの上位 1000 までを対象とした関連語辞書を 2010/11〜2012/10 まで作成し，先読み期間を 5ヶ月と設定した先読みポテンシャルの辞書を 2010/11〜2012/05 まで算出した例が表 1 である．表 1 より，先読みポテンシャルの上位に「レタス」,「白菜」,「茄子」といった一般的な語が含まれており，流行語として不適切な語で占められている．これは，共起度順位スコアが下位の方で変動しており，ピークとなる最大の共起度順スコアは流行語ほど高くないが，上昇率が高いため，先読みポテンシャルが高くなっているからだと考えられる．そこで，ピークとなる最大の共起度順位スコアがある一定以上ない語を除外するフィルタリングを行う．具体的には，対象月 T_i から先読み期間 s までにおいて，共起度順位スコアの最大値 R_{peak} が 300 以上の関連語のみを抽出する．つまり，ピークとなる共起度順位が 100 位以上となる関連語のみを対象とする．このフィルタリングによる結果が表 2 である．表 2 より,「タニタ」や「塩麹」などのレシピトピックにおける実際の流行語を，先読みポテンシャルの上位に含むことができた．よって，フィルタリングを行うことによって，一般的な語をある程度除外することは可能であるとかんがえられる．

なお,「タニタ」や「塩麹」の共起度順位に対する，先読みポテンシャルは，図 6，図 7 のようになる．図 6，図 7 より，共起度順位が最大となる 5ヶ月前が先読みポテンシャルが最も高くなり，共起度順位が最大となった後では先読みポテンシャルが低くなることがわかる．このように，流行語において先読みポテンシャルが高くなり，一般的な語を排除した妥当な先読みポテンシャルの辞書作成を今後検討する必要がある．

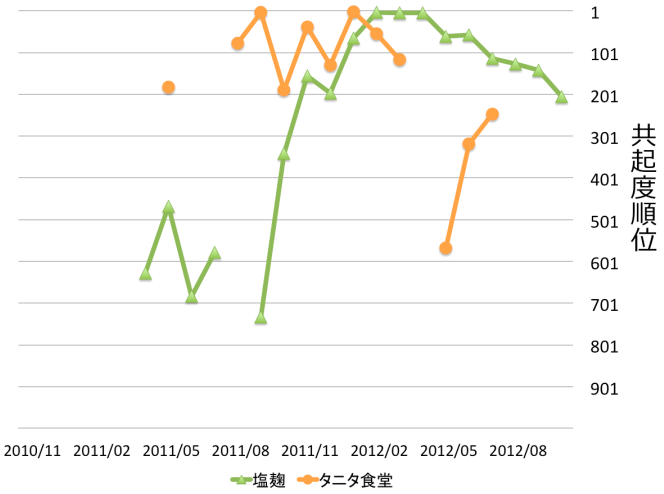

図 6　共起度順位の時系列変化（レシピコミュニティにおける「塩麹」と「タニタ食堂」の例）

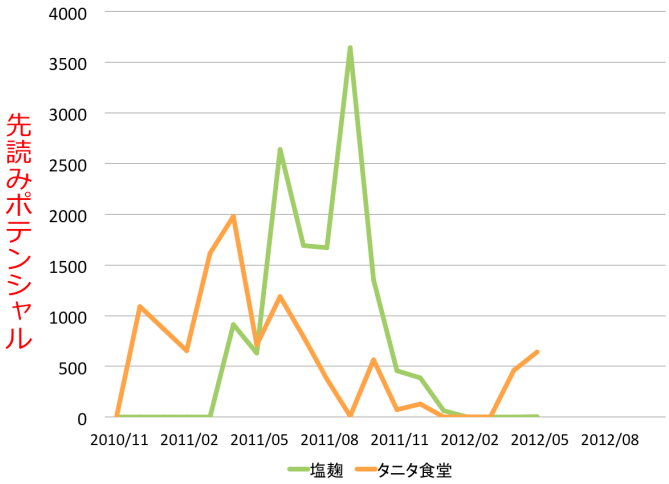

図 7　先読み期間 s を 5 に設定したときの先読みポテンシャル（レシピコミュニティにおける「塩麹」と「タニタ食堂」の例）

5. おわりに

我々は，流行語を世間に広まる前に言及している頻度を分析

表 1　レシピトピックにおける関連語辞書の先読み期間 5ヶ月の先読みポテンシャル算出結果 (一部抜粋)

	先読みポテンシャル $P(Recipe, w, T_i, 5) = (T_{peak} - T_i) \times (R_{peak} - R_i)$				
順位	2010/11	2011/03	2011/07	2011/11	2012/03
1	レタス	佃煮	白菜	チョコレート	茄子
2	なかし	バジル	チキン	海老	なす
3	こしょう	トッピング	習った	手順	バジル
4	載せて	簡単です	炊飯器	大豆	ピーマン
5	作れば	主婦	美味しさ	麹	佃煮

表 2　レシピトピックにおける関連語辞書の先読み期間 5ヶ月の先読みポテンシャル算出結果 (一部抜粋)

	先読みポテンシャル $P(Recipe, w, T_i, 5) = (T_{peak} - T_i) \times (R_{peak} - R_i) \quad when \quad R_{peak} \geqq 300$				
順位	2010/11	2011/03	2011/07	2011/11	2012/03
1	なかし	基本レシピ	新レシピ	麹	料理レシピ集
2	ケンタロウさん	シリコンスチーマー	体脂肪計タニタ	参考レシピ	クックパット
3	四十九日	ケンタロウさん	レシピサイト	塩麹	レシピ募集
4	楽天レシピ	レシピ紹介	ホットケーキミックス	体脂肪計タニタ	節約レシピ
5	参考レシピ	お弁当レシピ	タニタ	基本レシピ	ジュレ梅

することによって，話題を先取りしている先読みブロガーを発見する手法について検討した．今後は，先読みポテンシャルの妥当性の検証，および実際のブロガーに対して提案手法に基づいた先読み度を算出し,「先読みブロガー」と判定されたブロガーが今後も先読みし続けるかを検証する．また，周期的な話題は先読みする価値が低いと考えられるため，周期性を持つ話題を排除する手法について検討する．

6. 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] 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．

[2] 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.

[3] 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．

[4] 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．

[5] 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．

[6] 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．

[7] 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．

[8] 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．

[9] 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．

[10] 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．

[11] 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．