Vol.2013-DBS-158 No.1
2013/11/26

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

朝永 聖也[1,a)]　中島 伸介[1,b)]　稲垣 陽一[3,c)]　中本 レン[3,d)]　小倉 僚[1,e)]　張 建偉[2,f)]

概要：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の発見を行うために，過去の流行語に対してどの程度早くから言及していたのかを分析することによる，ブロガー先読み度判定手法を提案する．具体的には，その流行語が語り始められた時点を推測し，その時点から流行のピークを迎えるまでの期間において，対象となる流行語に関してどの程度早期に言及していたのかを評価する．本稿では，提案する先読み度判定手法について説明すると共に，本手法で必要となる流行語候補のカテゴリ分類について評価を行ったので，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

Measurement of Bloggers' Buzzword Prediction Ability Based on Analyzing Frequency of Early Mentions of Past Buzzwords

Abstract: The purpose of this study is to discover good predictors in blogosphere, as one of methods to detect promising buzzwords. In order to find good predictors, we propose a method for evaluating bloggers' buzzword prediction ability by analyzing how early bloggers mentioned past buzzwords. Concretely, we predict the time when a buzzword began to be mentioned, and evaluate how early the buzzword was mentioned in the period from the beginning time to the peak. In this paper, we describe the analysis method of bloggers' buzzword prediction ability, and report the evaluation on buzzword classification.

Keywords: 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である．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”が考えられる．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 [1][2]．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”に着目する．この手法は，“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”という仮定に基づいている．

[1] 京都産業大学
Kyoto Sangyo University
[2] 筑波技術大学
Tsukuba University of Technology
[3] 株式会社きざしカンパニー
Kizasi Company
[a)] i1358068@cse.kyoto-su.ac.jp
[b)] nakajima@cse.kyoto-su.ac.jp
[c)] inagaki@kizasi.jp
[d)] reyn@kizasi.jp
[e)] g1044228@cse.kyoto-su.ac.jp
[f)] zhangjw@a.tsukuba-tech.ac.jp

既に，我々はこれまでの研究において，「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 [3][4]」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 [5]」について報告している．しかしながら，過去の流行語を自動検出や，流行語の出現から流行語のピークまでの期間の推定を実現していなかったため，流行語に対する先読み度を精度良く算出することが困難であった．そこで本稿では，過去の流行語の自動検出や，流行語の出現からピークまでの期間の推定方法を提案する．また，流行語のみを事前に引用するだけではなく，流行のピーク時における共起語（ニュアンス）を含めて言及しているかどうかを判定することで，より高精度に先読みブロガーの判定を行うことを目指している．これは，例えば「iPhone5」の発売前に，「iPhone5 欲しいなぁ」と投稿するだけでは先読みしているとは言えないため，「iPhone5」に新たに搭載された機能やスペックについても併せて言及していることを高く評価しようとする方法である．

すなわち，本研究では，流行語の成長期間を推測し，この期間中どの程度早くから対象の流行語について（ピーク時のニュアンスを含めて）言及していたかを評価することにより，ブロガー先読み度を判定する手法の提案を目的としている．

以降，2 節にて，関連研究について述べる．3 節にて，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，4 節にて，流行語候補のカテゴリ分類評価実験について述べる．最後に 5 節にて，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 [7]．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 [8]．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 [9]．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 [10]．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 [11]．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 [12]．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 [13]．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 [14]．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 [15]．

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

本提案手法は，過去のブログ記事分析によりシード語（過去の流行語）を抽出し，このシード語について早期から言及しているブロガーを先読みブロガーとして判定することを目指している．なお，解析用データとしては，kizasi.jp[16] 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）を対象としている．

ここで提案手法の処理の流れを以下に示すと共に，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の推定（3.3 節）
(4) シード語の先読みポイントの算出（3.4 節）
(5) ブロガーの先読み度の算出（3.5 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 [6] 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．

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

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 ($l \leq n$)．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

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[16] 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．

なお，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．共起語集合間の類似度算出手法としては，jaccard 係数，simpson 係数，コサイン類似度等の各種手法に対する評価実験結果を踏まえて検討する．

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．

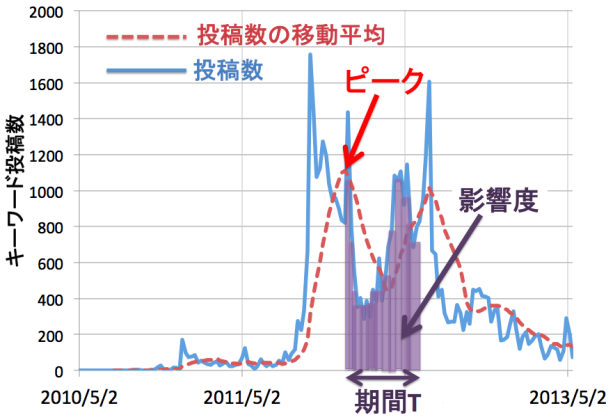

図 1 影響度の概念図

3.3 シード語の成長期間の推定

本節では，過去の流行語として抽出したシード語に対して，いつ頃から語られ始め，ピークを迎える迄にどの程度の期間を必要としたのか，すなわちシード語の成長期間を推定する手法について述べる．

3.3.1 ピーク時の話題とそれ以前の話題との類似度計算

シード語の成長期間を推定するために，まずはシード語が表す流行語について，いつから語られ始めたのかを推定する．このとき，流行したとき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 5」がシード語である場合，「iOS 5 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 5」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 5 には Siri という名の音声アシスタント機能が搭載されるらしい」という内容を，「iOS 5」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．そこで，「iOS 5」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，それより以前の期間におけるシード語「iOS 5」との共起語集合との類似度計算を行うことにより，流行のピーク時のニュアンスを含めて，そのシード語が示す話題がいつ頃から語られ始めているのかを推定する（図 2 参照）．

なお，シード語の共起語集合は，シード語に対し共起度の高いキーワードを集め，一般語を削除し，対象となるシード語の流行時の特有キーワードのみを残した上位 400 語程度のキーワード群である．シード語の流行時の特有キーワードとは，流行する以前までは，あまり投稿されていなかったが，流行と共に投稿数が増加したキーワードである．

シード語の流行時の特有キーワードを発見する手法としては，ピーク以前の期間において，ピーク時の各共起語を含むブログ記事数が時間経過と共に増加傾向にあるかど

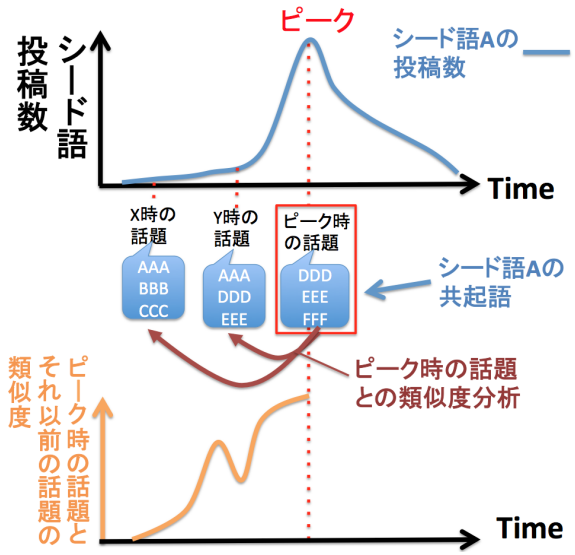

図 2 ピーク時の話題とそれ以前の話題との類似度計算

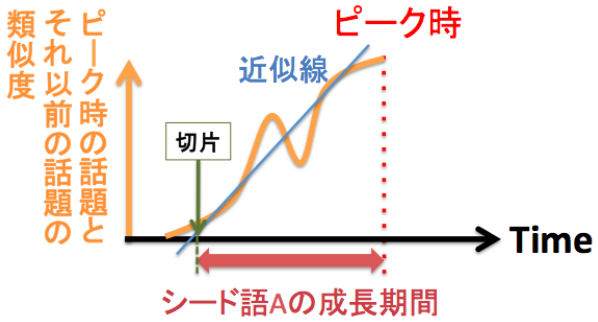

図 3 シード語の成長期間の判定 (1)

うかをスピアマンの順位相関係数を用いて計算する．この時，かなり強い正の相関（係数 ρ が 0.6 以上）があれば，この共起語は，投稿数が増加傾向にあるキーワードであり，シード語の流行時の特有キーワードであるとみなす．

3.3.2 シード語の成長期間の判定

シード語のピーク時の話題とそ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，そ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 3）．

(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 4）．

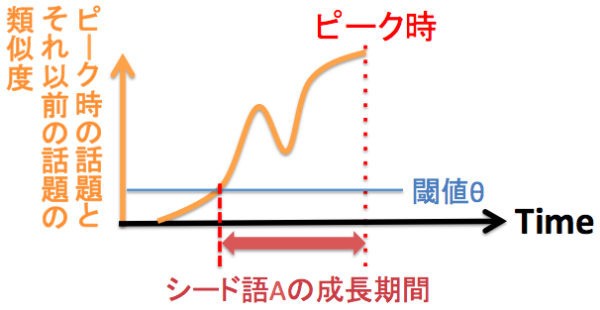

図 4 シード語の成長期間の判定 (2)

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．

3.4 シード語の先読みポイントの算出

本節では，シード語の先読みポイントの算出方法について説明する．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るものであり，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．

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

$entry_{all}$ 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．

3.5 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 5 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．

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

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．

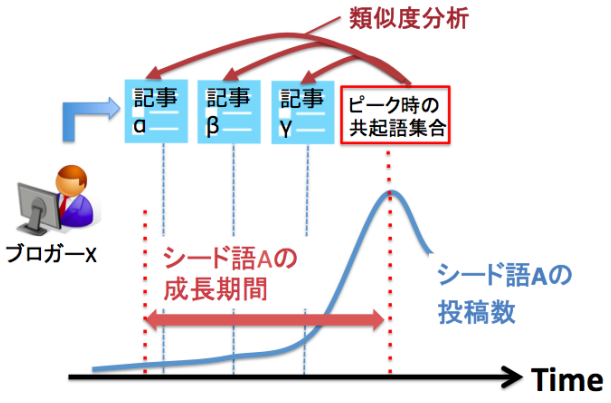

図 5 ブロガー先読み度判定のためのブログ記事の類似度分析

4. 流行語候補のカテゴリ分類評価実験

本節では，3.2.2 節で述べた「シード語候補のカテゴリ分類」に対して行った評価実験について説明する．3.2.2 節でも述べたが，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．さらに，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．

分類先のカテゴリは，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 [6] により，約 13,000 件の熟知グループに分類されているが，カテゴリの粒度が細か過ぎるため，熟知グループを約 120 件の上位カテゴリにグループ化し，約 120 件の熟知グループ単位で，シード語候補の分類を行う．

4.1 シード語のカテゴリ分類方法

本評価実験では，与えられたシード語が分類されるべきカテゴリを精度良く判定できるかどうか，すなわち，“シード語と関連が深い熟知グループを適切に判定できるかどうか”を評価することが目的である．

与えられるシード語候補と熟知グループとの意味的な近さを表す「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”の類似度で表現し，各シード語候補ごとに類似度の高い熟知グループを抽出する．

なお，本評価実験で用いた類似度は，jaccard 係数 (6)，simpson 係数 (7)，コサイン類似度 (8)，共起度順位点重み付きコサイン類似度 (9) である．

$$jaccard(S,C) = \frac{|S \cap C|}{|S \cup C|} \tag{6}$$

$$simpson(S,C) = \frac{|S \cap C|}{min(|S|,|C|)} \tag{7}$$

$$cosSim(S,C) = \frac{|S \cap C|}{\sqrt{|S|^2}\sqrt{|C|^2}} \tag{8}$$

$$cosSim_{Rank}(S,C) = \frac{\sum_{i=1}^{N}(r_{Si} \cdot r_{Ci})}{\sqrt{\sum_{i=1}^{N} {r_{Si}}^2}\sqrt{\sum_{i=1}^{N} {r_{Ci}}^2}} \tag{9}$$

ここで，S は，クローリングしたブログ全体で，シード語候補と共起度の高い上位 400 語のキーワードを集めた共起語集合である．C は，対象熟知グループ内で，「対象熟知グループを表すキーワード」と共起語度の高い上位 400 語のキーワードを集め共起語集合である．また，r_S は，S が持つ各キーワードに対し，共起度の高い順に，400 から 1 まで割り振った順位点であり，r_C は，C が持つ各キーワードに対し，共起度の順位点の合計点が高い順に，400 から 1 まで割り振った順位点である．N は比較するキーワード数であり，今回は 400 である．

したがって，これら 4 つの類似度算出方法を用いて，シード語と関連が深い熟知グループを判定し，どの手法が最も優れているかを調べる．

なお，比較対象として，先行研究 [1][2] で用いた「各熟知グループ内における，対象シード語について言及しているブロガー割合による手法」を採用した．この手法は，“熟知グループ P のメンバーの大多数が，シード語 Q について言及していれば，シード語 Q は熟知度グループ P と関連が深い”という考え方に基づいたものである．

4.2 実験・評価手法

本評価実験で用いたシード語は，過去に流行したキーワードである「AKB48」「Android」「Facebook」「女子会」「K-POP」「スマートフォン」の 6 語である．これらのシード語に対して，関連が深い熟知グループを正しく自動判定できるかどうかを，4 種類の提案手法と先行研究 [1][2] による手法について調べることが本実験の目的となる．なお，6 個の各シード語が分類されるにふさわしい上位 10 件のカテゴリを，人手により事前に作成している．これらを“熟知グループの理想ランキング”と呼ぶ．この理想ランキングに近い，ランキング結果を算出する手法が最も優れているということになる．

実験に用いるブログ記事は，「2012.07.01〜2012.09.30 (期間 1)」「2012.09.30〜2012.12.31 (期間 2)」「2012.12.31〜2013.04.01(期間 3)」「2013.04.01〜2013.07.02 (期間 4)」の約 3 ヶ月毎の各期間に投稿された記事を対象とし，熟知グ

ループの共起語集合を各期間ごとに作成する．複数の期間を扱う理由としては，時代の流れに応じて，シード語と関連するカテゴリも少しずつ変化することが考えられるため，各期間において関連カテゴリの変化が見られるかを確認するためである．

なお，比較評価尺度としては，DCG 及び $nDCG$ を用いる．DCG は減損累積利得と呼ばれ，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．すなわち，単純に理想ランキングに含まれるだけではなく，順位も合致している程高く評価できる指標である．なお，$nDCG$ とは，正規化減損累積利得であり，理想ランキングの DCG の値を 1 として正規化したときの値である．DCG および $nDCG$ の算出式を以下に示す．

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \tag{10}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{11}$$

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．なお，今回は，事前に人手で作成した熟知グループの理想ランキングに対して，上位から順に 10,9,8...,1 のような利得スコアを割り振り，DCG を計算した．この対象シード語候補ごとの DCG の結果を，各 $IDCG_p$ として扱うものとする．

4.3 関連度に基づく熟知グループランキングの評価結果

提案手法に基づくランキング結果の内，類似度として simpson 係数を用いた結果を A，jaccard 係数を用いた結果を B，コサイン類似度を用いた結果を C，共起度順位点重み付きコサイン類似度を用いた結果を D とし，前述の先行研究による手法併せて，各 nDCG の結果を表 1〜6 に示す．

表 1〜6 の結果より，提案手法 A,B,C による各スコアは異なるものの，ランキング結果が全く同じ結果となり，そのため $nDCG$ の値も同一であった．

これに対し，提案手法 D による $nDCG$ の値は，提案手法 A,B,C とは異なる結果となった．表 6 の「スマートフォン」や表 4 の「女子会」に対するランキング結果では，提案手法 A,B,C に比べて比較的優れた結果となったが，その他のシード語に対するランキング結果では，逆に提案手法 A,B,C の結果を下回ることが多かった．当初は，共起語集合の共起順位を考慮している手法 D が最も良い結果となることを期待していたが，必ずしもそのような結果を得ることはできなかった．

表 1 理想ランキングに対する各手法の nDCG（AKB48）

期間	提案手法 A	B	C	D	先行研究
期間 1	**0.617**	**0.617**	**0.617**	0.549	0.091
期間 2	0.605	0.605	0.605	**0.627**	0.144
期間 3	**0.549**	**0.549**	**0.549**	0.300	0.104
期間 4	**0.572**	**0.572**	**0.572**	0.547	0.102
平均	**0.586**	**0.586**	**0.586**	0.506	0.110

表 2 理想ランキングに対する各手法の nDCG（Android）

期間	提案手法 A	B	C	D	先行研究
期間 1	0.834	0.834	0.834	**0.838**	0.584
期間 2	**0.880**	**0.880**	**0.880**	0.819	0.584
期間 3	**0.806**	**0.806**	**0.806**	**0.806**	0.579
期間 4	**0.813**	**0.813**	**0.813**	0.765	0.531
平均	**0.833**	**0.833**	**0.833**	0.807	0.569

表 3 理想ランキングに対する各手法の nDCG（Facebook）

期間	提案手法 A	B	C	D	先行研究
期間 1	**0.476**	**0.476**	**0.476**	0.403	0.327
期間 2	0.384	0.384	0.384	**0.402**	0.327
期間 3	0.384	0.384	0.384	**0.401**	0.327
期間 4	**0.491**	**0.491**	**0.491**	0.455	0.327
平均	**0.433**	**0.433**	**0.433**	0.415	0.327

表 4 理想ランキングに対する各手法の nDCG（女子会）

期間	提案手法 A	B	C	D	先行研究
期間 1	**0.391**	**0.391**	**0.391**	0.373	0.033
期間 2	0.302	0.302	0.302	**0.397**	0.034
期間 3	0.443	0.443	0.443	**0.469**	0.033
期間 4	0.333	0.333	0.333	**0.385**	0.033
平均	0.367	0.367	0.367	**0.406**	0.033

表 5 理想ランキングに対する各手法の nDCG（K-POP）

期間	提案手法 A	B	C	D	先行研究
期間 1	**0.829**	**0.829**	**0.829**	0.683	0.486
期間 2	0.676	0.676	0.676	**0.708**	0.506
期間 3	**0.658**	**0.658**	**0.658**	0.618	0.503
期間 4	**0.689**	**0.689**	**0.689**	0.645	0.550
平均	**0.713**	**0.713**	**0.713**	0.663	0.511

表 6 理想ランキングに対する各手法の nDCG（スマートフォン）

期間	提案手法 A	B	C	D	先行研究
期間 1	0.775	0.775	0.775	**0.909**	0.050
期間 2	0.678	0.678	0.678	**0.869**	0.071
期間 3	0.553	0.553	0.553	**0.754**	0.071
期間 4	0.676	0.676	0.676	**0.700**	0.071
平均	0.670	0.670	0.670	**0.808**	0.066

今後は提案手法 D のチューニングを検討すると共に，他のシード語についても詳細に分析することで，どのような場面でどの手法を採用するべきかを検討する予定である．いずれにせよ，提案手法の結果は，先行研究の結果を全てのシード語において上回っており，提案手法の基本的な考え方については有効性は高いと考えられる．

３ヶ月毎の４つの期間に対して実験を行ったが，*nDCG* の値に違いが出ている通り，各期間にて算出したランキングは異なる結果となった．すなわち，期間毎にブログ記事の内容も異なるため，これらの記事内容の違いを反映して，シード語に関連の深いカテゴリ（熟知グループ）をその時期に応じて算出できる可能性を示した．

また，各手法において，上位 10 件の熟知グループをランキングさせたが，特に上位 1,2 位については適切にランキングしているケースが数多く見られた．したがって，提案手法にはまだまだ改良の余地はあるものの，利用範囲をランキング上位に絞れば，十分有用性はあると考えている．

5. まとめ

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，過去の流行語に対する言及時期と内容を分析することによる「ブロガー先読み度」の判定手法を提案した．また，流行語候補のカテゴリ分類について実験を行い，良好な結果を得た．

今後は，未実装部分である，“シード語の成長期間の推定手法（3.3 節）”，“シード語の先読みポイントの算出手法（3.4 節）”，“ブロガーの先読み度の算出手法（3.5 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．

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] 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.

[2] 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.

[3] 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.

[4] 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．

[5] 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．

[6] 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．

[7] 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．

[8] 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．

[9] 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．

[10] 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．

[11] 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．

[12] 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．

[13] 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．

[14] 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)．

[15] 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)．

[16] kizasi.jp:ブログから、話題を知る、きざしを見つける, http://kizasi.jp/