情報処理学会論文誌　データベース　Vol. 3　No. 3　123–134 (Sep. 2010)

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および評価

稲垣陽一[†1]　中島伸介[†2]　張建偉[†2]
中本レン[†1]　桑原雄[†1]

本研究ではブロガの体験熟知度に基づいたブログランキングシステムの開発を行った．ユーザが入力した検索キーワードに対して，関連するトピックを複数抽出し，各トピックに関するブロガの体験熟知度を算出する．これに基づいてブログエントリのランキングを行う．熟知度スコアが高いブロガ（熟知ブロガ）が書いたエントリは，熟知度スコアが低いブロガが書いたエントリよりもランキングが上位となる．ブロガの熟知度スコアは，ブロガが過去に投稿したエントリ内で，各トピックに関して共起に基づいて抽出した特徴語をどれほど使ったかを分析することで算出される．なお，開発したシステムは，視点の異なる複数のランキングを提示するとともに，エントリ投稿者（ブロガ）の特性に関する補助情報を提示している．これにより，ユーザは閲覧するブログエントリの信頼性を自分なりに判断することが可能となる．我々は開発した実証実験システムを Web 上で公開するとともに，これを用いた評価実験を行った．提案システムにより提示される熟知ブロガおよびブログエントリの妥当性が十分に高いことを確認できた．

Implementation and Evaluation of Blog Ranking System Based on Bloggers' Knowledge Level

Yoichi Inagaki,[†1] Shinsuke Nakajima,[†2]
Jianwei Zhang,[†2] Reyn Nakamoto[†1]
and Yu Kuwabara[†1]

In this paper, we propose a blog ranking system based on bloggers' knowledge level. For the query keyword that a user enters, this method extracts multiple relevant topics, calculates knowledge scores of bloggers for each topic, and ranks blog entries based on bloggers' knowledge scores. In our method, blog entries written by knowledgeable bloggers have higher rankings than those written by common bloggers. Bloggers' knowledge scores are evaluated based on their usage of topic-specific words in their past blog entries. Additionally, our system can present multiple ranking lists of blog entries from the perspectives of different bloggers' groups. This allows users to estimate the trustworthiness of blog contents from multiple aspects. We built a prototype of the proposed system and evaluated it through user testing. Our evaluation showed that our method can effectively rank bloggers and blog entries.

1. はじめに

近年，ブログや SNS などの CGM と呼ばれるコンテンツが数多く配信されるようになり，これら CGM コンテンツの中から価値の高いものを効率的に取得することが困難になりつつある．そのような背景から，ブログ検索やそのランキングに対する要求が高まっている．ブログの魅力の 1 つは，その即時性である．したがって，ブログランキングでは，エントリ投稿直後の短時間においてランキングを算出する必要がある．しかし，一般的に多様・雑多な Web ページやブログエントリを本文だけで適切にランキングするのは難しい．Web 検索エンジンでは PageRank [1] のようなリンク構造解析が効果的に併用されているが，即時性が重要なブログでは，投稿直後のほとんどリンクされていない段階でも使用できる手法が必要とされる．

我々はこの現状から，ブロガ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トピックに関するブロガ熟知度を算出し，これに基づいてブログエントリのランキングを算出しようとするブログランキングシステムの開発を行った．なお，この手法では，ランキングそのものの算出だけでなく，複数の尺度に基づいたランキングの提示が可能である．たとえば，検索キーワード「株式」で検索した際に，「株式」に対する熟知度に基づいたランキングだけでなく，「経済」，「企業買収」，「政治」などに関する熟知度に基づいた複数のランキングを提示できる．すなわち，システムが提示する 1 つのランキングだけでなく，複数の視点によるランキングの中から，ユーザ自身が選択することができる．これにより，ユーザは閲覧する情報の背景を把握することができるため，その閲覧情報の価値や信頼性を自分なりに判断す

†1 きざしカンパニー
kizasi Company, Inc.
†2 京都産業大学
Kyoto Sangyo University

ることも可能になる．開発した熟知度に基づくブログランキングシステムは，2008 年 9 月より一般公開するとともに評価実験を実施しており，本システムの有効性を確認している．

以下，実証実験システムの実装内容に関して説明するとともに，本システムにより行った評価実験について報告する．2 章では熟知度に基づくブログランキングの方式を述べる．3 章では実証実験システムの内容を説明する．4 章では開発したシステムに関する評価実験の結果について考察する．5 章では関連研究について述べ，6 章でまとめとする．

2. 熟知度に基づくブログランキング方式

本章では，熟知度に基づくブログランキング方式について説明する．前処理として，システムは熟知グループと共起語辞書を作成し，各熟知グループに関する各ブロガの熟知度を算出する．検索の際は，ユーザが検索キーワードを入力すると，システムはそのキーワードを含むブログエントリを検索し，エントリを投稿したブロガの熟知度スコアによってエントリのランキングを行い，結果をユーザに提示する．以下，熟知グループの作成，共起語辞書の作成，ブロガ熟知度の算出，スパムブログの検出ならびにブログエントリのランキング算出方法について詳細を述べる．

2.1 熟知グループの作成

本研究では，あるトピックに関して熟知するブロガの集合を「熟知グループ」と呼び，「熟知グループ名」とはそのトピックそのものである．まず，ブログでよく言及されるトピック（熟知グループ名）を抽出する．熟知グループ名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．

以下の手法により，500 程度の熟知グループ名を自動的に抽出した．

(1) 「マニア」，「ファン」，「フリーク」などの特定のトピックの専門家であることを表すキーワードでブログ検索を行い，検索結果のテキストの中から「マニア」，「ファン」，「フリーク」の直前の名詞句を取得し，その頻度を計算する（たとえば，「鉄道マニア」の場合は，語句「鉄道」を取得する）．

(2) この頻度順に整列した名詞句のうち，頻度が高いものを熟知グループ名としてリストアップする（すなわち，「鉄道」に熟知するブロガのグループが存在する場合，そのグループ名を抽出する）．

一方，我々は独自に開発した生活体験シソーラス LETS（Life Experience TheSaurus）のカテゴリから，熟知グループ名として適切なカテゴリを採用することで，熟知グループの拡充を行っている．LETS は，ブログやニュースなどに顕著に見られる，生活者の体験に強く関わるカテゴリを人手で体系的に分類，整理したものである．実体験をもとにブログエントリ内で語られているトピックのみを対象とする点が，既存のシソーラスと異なると考えている．**図 1** に示すように，LETS は感情を表す感情層と，抽象から具体までの概念を表す A，B，C 階層を持つ．感情層は「喜び」，「悲しみ」，「恐怖」，「可愛い」など，約 70 種類の感情的なものからなる．A 層のカテゴリは「芸術」，「健康」，「家族」，「音楽」など 50 種類程度の抽象的なものである．B 層は 2,000 種類程度の若干具体的なカテゴリからなる．たとえば，「音楽」の下位ジャンルである「アニメソング」や「オペラ」などがあげられる．C 層は人名，植物名，動物名，曲名など，58,000 程度のさらに具体的な固有名詞からなる．たとえば，「アニメソング」の下位ジャンルである曲名や歌手名などがあげられる．4 層の合計で LETS の全カテゴリ数は 60,000 程度である．A 層，B 層，C 層のカテゴリのうち，熟知グループ名としての提示が不適切と思われるものを除き，12,000 程度のカテゴリを用いて熟知グループを拡充している（2010 年 3 月 10 日現在）．「自殺」，「うつ病」などのカテゴリは LETS に存在するが，このような内容を書くブロガのグループとして公開することは控えるべきであるため，熟知グループの拡充には使用しない．

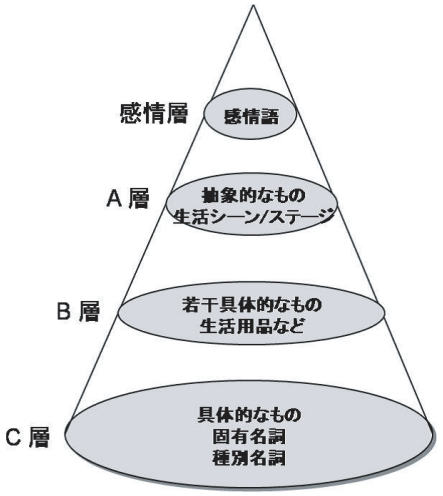

図 1　生活体験シソーラス LETS の概念図
Fig. 1　LETS's concept.

自動抽出で取得した 500 程度の熟知グループと LETS から採用した 12,000 程度の熟知

グループの重なりは 300 程度である．たとえば，「ネコ」，「サッカー」などがあげられる．自動抽出ではなく，人手で収集したカテゴリは 11,700 程度である．たとえば，猫の一種である「アメリカンカール」など，「○○マニア」のパターンではブログに出現しないが，「アメリカンカール」ファンのブロガは確実に存在する．一方，自動抽出した残りの 200 程度の熟知グループには，LETS の編集者は知らなくても，ブロガの間では著名なスポーツ選手やアーティストなどが含まれる．なお，ブログに新しいトピックが続々と登場するため，熟知グループは 1 週間間隔で更新している．

2.2　共起語辞書の作成

次に，各熟知グループに対して，共起語辞書を構築する．各熟知グループに対して，一定期間内のブログエントリを対象とし，その熟知グループ名であるキーワードとの共起度が高い n 個の語句を抽出する．共起語数 n を大きくすると処理時間が長くなり，n を小さくするとブロガの熟知度の算出精度が低くなるため，予備実験を行ったうえで，対象期間を辞書構築時点の直近 2 年分とし，共起語数を $n = 400$ とした．

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[2),3)]．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予備実験で見られた．そのため，実証実験システムでは，それらの中間の尺度 LogLog スコアを採用した．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．

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

共起語辞書の構造と例を**図 2** に示す．各行では熟知グループ（g_i）と共起語（w_{ij}）および共起度（β_{ij}）を表す．たとえば，熟知グループ「コンピュータ」に対して，「計算機」，「システム」などの語句がブログの中で同時に多く現れるため，このような語句を共起語として抽出し，共起語辞書に登録する．なお，ブログエントリは随時更新され，同一熟知グループに対しても，共起語が時間につれて変わるため，共起語辞書を定期の間隔（1 週間程度）で更新する．たとえば，「鳩山」に対して，2009 年 8 月の時点で共起度が最も高い共起語は「政権交代」であったが，2010 年 3 月の時点で共起度が最も高い共起語は「子ども手当」となった．

また，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熟知ブロガとして発見しやすくしている．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．なお，LETS（図 1）の感情層のカテゴリ（「喜び」，「悲しみ」，「恐怖」，「可愛い」など）に対しても，同様な方法で共起語辞書を構築した．

熟知グループ g_i または感情語		共起語 w_{ij}							
		$j=1$		2		...		400	
$i=1$	コンピュータ	計算機	$\beta_{1,1}$	システム	$\beta_{1,2}$	...	...	会社	$\beta_{1,400}$
$i=2$	鳩山	子ども手当	$\beta_{2,1}$	支持率	$\beta_{2,2}$	...	...	国民	$\beta_{2,400}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$i=k$	可愛い	女の子	$\beta_{k,1}$	ピンク	$\beta_{k,2}$	...	...	笑顔	$\beta_{k,400}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

図 2　共起語辞書の構造と例
Fig. 2　Structure and examples of the co-occurrence word dictionary.

2.3　ブロガ熟知度スコアの算出

各ブロガの熟知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いて算出する．なお，各ブロガ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，あるブロガが「経済」と「政治」に関する熟知ブロガである場合，このブロガは「経済」に関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．

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

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 w_{ij} の重みであり，$\alpha_{ij} = (n - j + 1)/n$ で表される．β_{ij} は熟知グループ 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．

次に，対象熟知グループ g_i に対するブロガ 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 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，そのブロガの熟知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．

なお，実証実験システムでは，ブロガの熟知度の算出に，スコア算出時点から直近 2 年間のブログエントリを使用している．解析対象エントリ数の確保という観点も含めて考慮し，対象期間を決定しているが，十分なエントリ数が確保できるのであれば，ブロガ自身の特性の変化に対応するために，対象期間をより短くすべきということも考えられる．したがって，最適な対象期間の設定については引き続き検討していく．なお，ブロガ熟知度スコアは 1 週間ごとに再計算を行い，更新する．

2.4　スパムブログの検出

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ているスパムブログが増加している．ブログ検索サービスを提供するのに，スパムブログを発見・除去することは重要な課題の 1 つである．提案システムでは，半自動スパムフィルタと全自動スパムフィルタを利用している．

半自動スパムフィルタは，スパムであることが疑わしいものを検知するためのルールを作成し，これにより検知されたものに対してのみ，人手でスパムかどうかを判定する．以下では，代表的なルールを 2 つあげる．

1 つ目は，投稿エントリ数に関するルールである．典型的なスパムブログの特徴の 1 つとして，大量のエントリ投稿を長時間にわたって行うというものがあげられる．1 日平均 10 エントリ以上 1 週間で 70 エントリ以上の投稿数のあるブロガを抽出し，スパムブログであるかどうかを，人間が最終的に判定する．**図 3** は，このフィルタで発見したスパムブログの一例である．観測期間は 2008 年 10 月 1 日から 2009 年 4 月 2 日までの約半年である．横軸は時間帯であり，縦軸はそのブロガの投稿数である．各曲線は曜日別の投稿数である．このブロガは，約半年の間に 3,000 近くのエントリ，つまり 1 週間あたり 110 件程度のエントリを投稿するとともに，時間帯もほぼ全時間帯においてエントリの投稿が行われている．人間がチェックしたところ，スパムブログであることが確認された．

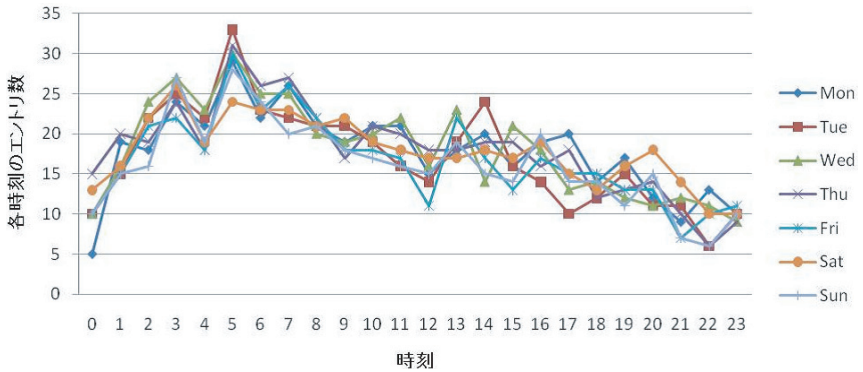

図 3　スパムブログの例：異常な投稿数
（観測期間：2008 年 10 月 1 日～2009 年 4 月 2 日）
Fig. 3　Example of spam blog: abnormal post number
(Observation period: October 1, 2008 – April 2, 2009).

2 つ目は，投稿時間の規則性に関するルールである．典型的なスパムブログのもう 1 つの特徴として，投稿に時間的な規則性が見られる．そこで，投稿時間のパターンに着目し，規則正しく決まった時刻にエントリを投稿したブロガをスパムの候補と判定する．具体的には，投稿時間が前後 30 秒以内の差で一致するエントリ数が全体の 75%以上を占めるブログをスパムの候補とした．**図 4** はこのフィルタで発見したスパムブログの一例である．エントリの投稿時間がほぼ決まっており，チェックした結果，スパムブログと確認された．

全自動スパムフィルタは，人間に判定されたスパムを学習データとし，生成したベイズ分類器である．半自動スパムフィルタと全自動スパムフィルタで検出したスパムブログ数は，1 日に数百件から数万件程度である．当然ながら，システムによりスパムブログと判定されたブログはランキング対象からは排除される．

2.5　ブログエントリのランキング算出方法

本節では，前述の熟知度スコアに基づくブログエントリのランキング算出方法について述べる．以下に，その算出手順を示す（**図 5**）．

(1)　まず，検索キーワードを含むブログエントリを検索する．たとえば，「健康」というキーワードを入力し，ブログエントリを取得する．

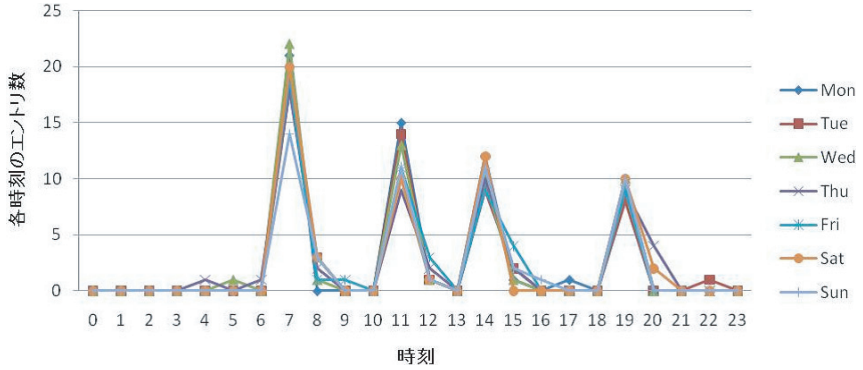

図 4　スパムブログの例：規則的な投稿
（観測期間：2008 年 10 月 1 日〜2009 年 4 月 2 日）
Fig. 4　Example of spam blog: regular post
(Observation period: October 1, 2008 – April 2, 2009).

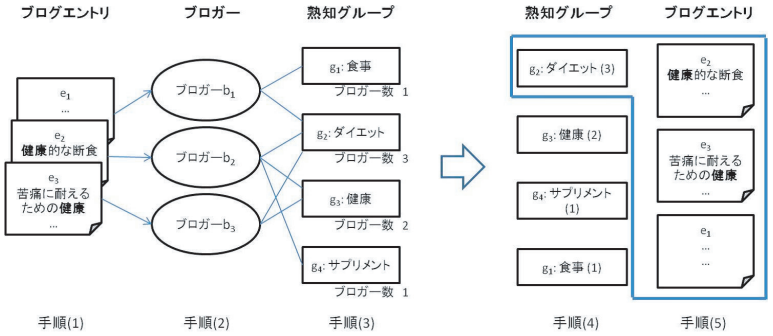

図 5　ブログエントリのランキングの例
Fig. 5　Example of blog entry ranking.

（2）　システムはこれらのエントリを投稿したブロガを特定する．図 5 の例では，「健康」のキーワードを含むブログエントリ e_1，e_2，e_3 に対して，エントリを書いたブロガ b_1，b_2，b_3 を特定する．

（3）　次に，これらのブロガが属する熟知グループを特定し，熟知グループごとにブロガ数を集計する．この際，あるブロガが複数の熟知グループに属することがある．

（4）　熟知グループをブロガ数の降順でランキングする．図 5 の例では，「g_2：ダイエット」，「g_3：健康」，「g_4：サプリメント」，「g_1：食事」の順番で熟知グループをランキングする．

（5）　各熟知グループのブログエントリを，ブロガの熟知度スコアの降順でランキングする．図 5 の例において，「g_2：ダイエット」に関するブロガの熟知度が $knowledge_{g_2}(b_2) > knowledge_{g_2}(b_3) > knowledge_{g_2}(b_1)$ の場合，そのエントリは $e_2 > e_3 > e_1$ としてランキングされる．

以上のように，検索対象であるブログエントリに対して，複数の熟知グループに関する複数のランキングを提供することができる．すなわち，1 度の検索において複数の視点（熟知グループ）からのランキングを実現できる．これにより，利用者は各視点における上位ランクのブロガが書いたエントリを選択的に閲覧することが可能になる．

3. 実証実験システム

本章では，2008 年 9 月より公開し，随時機能の改良を行っている実証実験システムについて説明する．本システムは，2010 年 3 月 10 日現在で，約 7,422,000 名のブロガにより投稿された，約 173,715,000 ブログエントリのデータを保有している．また，熟知グループの数は約 12,000 であり，熟知ブロガとされているブロガ数は，約 265,500 名となっている．公開から 2010 年 3 月 10 日現在までのシステムアクセス数は 224,794 件である．

図 6 に実証実験システム[4] のスナップショットを示す．図 6 に示すとおり，本システムではユーザから検索要求があった際の結果表示として「熟知グループ選択およびランキング表示部」と「ニュアンス比較表示部」を有する．

3.1 熟知グループ選択およびランキング表示部

図 7 に「熟知グループ選択およびランキング表示部」の拡大図を示す．A 部では，検索キーワード（たとえば，「健康」）をもとに，関連する熟知グループを関連性の高い順に表示している．この関連性は，“対象熟知グループに属する全ブロガの中で対象期間内に検索キーワードを含むエントリを投稿したブロガ数とその割合” から算出している．A 部の中のグループ名（たとえば，「ダイエット」）をクリックすれば，B 部においてそのグループのブログエントリがランキング順に表示される．これにより，「ダイエット」熟知ブロガから見た「健康」に関する書き込み，「食事」熟知ブロガから見た「健康」に関する書き込み，「仕事」熟知ブロガから見た「健康」に関する書き込み，などを選択して閲覧することができる．すなわち，ある検索キーワードにより特定される 1 つのトピックに対しても，いろい

図 6　実証実験システムのスナップショット
Fig. 6　Snapshot of the experimental system.

ろな切り口での書き込みを選択することができるのである．他の例でいえば，「沖縄米軍基地」で検索した際の，「外交」熟知ブロガ，「環境」熟知ブロガ，「沖縄」熟知ブロガらの書き込みは，各々別の視点で記述されていることが予想される．

また，スニペット部にブロガが属する熟知グループの割合情報を表示している．これにより，このエントリを記述しているブロガの特性を推定することができるため，ユーザ（閲覧者）自身がそのブログエントリの信頼度を客観的に判断することを助けている．

3.2　ニュアンス比較表示部

図 8 に「ニュアンス比較表示部」の拡大図を示す．ここでは検索語に対するブロガ全体の感情表現と特定の熟知グループの感情表現の差異が分かるように比較表示している．図 8 の左側は「健康」に関する全ブロガが書いたエントリのニュアンスであり，右側は「健康」に関して「ダイエット」に熟知するブロガのニュアンスである．検索キーワードが与えられると，全ブロガのニュアンスと選択した熟知グループのニュアンスを算出する．

全ブロガのニュアンスを求めるには，まず検索語を含む全ブログエントリを対象とし，検索語の共起語を抽出する．次に検索語の共起語リストと，LETS の感情層に属する各感情語（たとえば，図 2 の表中の「可愛い」）の共起語リスト（図 2 の表中の「女の子」，「ピンク」,...，「笑顔」）との共通部分を求め，共通語数が多い感情語上位 3 個を抽出し，全ブロガのニュアンスとして提示する．

熟知グループのニュアンスを求めるには，検索語を含む熟知ブロガが書いたエントリを分析対象とする．熟知グループのニュアンスの差異をより明瞭に示すため，対象の熟知グループにおける検索語の共起語リストとブロガ全体の共起語リストの差分をとり，その差分をもとに熟知グループのニュアンスを計測している．

図 8 の例では，全体ブロガでのニュアンスは，「恐怖」，「不快・不愉快」，「疲労」のように「健康」に対してネガティブなニュアンスが上位となっているが，「ダイエット」熟知ブロガのニュアンスには，「スッキリ」というニュアンスも含まれており，「健康」というトピックに対するとらえ方の違いが表現できているといえる．このようにニュアンスを比較することにより，対象熟知ブロガグループの特性をユーザが把握することを助けている．

4.　評 価 実 験

4.1　熟知グループ，熟知ブロガ，ブログエントリの評価

実証実験システムが提示する，熟知グループ，熟知ブロガ，ブログエントリのランキングに関する妥当性の評価を行った．トピック（検索キーワード）として，**図 9** に示す 20 トピックを使用し，システムが提示した結果に対して 5 名の被験者（共著者でない大学生）が行った評価結果を集計する形で実施した．

4.1.1　熟知グループの妥当性評価

与えられた検索キーワードに対し，実証実験システムでは，複数の熟知グループを提示する．ここで検索キーワードと無関係な熟知グループが提示されることは適切ではない．たとえば，キーワード「衆議院選挙」で検索した場合には，「政治」，「報道」，「経済」，「民主党」，「自民党」のような関連の深い熟知グループが提示されるべきである．したがって，この検索キーワードに対して提示される熟知グループの妥当性を評価した．

各被験者は，図 9 に示す検索キーワードで検索した際に，上位 5 個の熟知グループ名がこの検索キーワードに関連するかどうかを判断した．この評価では，“システムが提示した 5 個の熟知グループのうち，被験者が妥当だと判断した熟知グループ数の割合” により評価を行った．

図 7　熟知グループ選択およびランキング表示部（検索語：健康，熟知グループ：ダイエット）
Fig. 7　Part of knowledgeable group selection and ranking presentation (query keyword: health, knowledgeable group: diet).

システムが提示する熟知グループの妥当性評価結果を**図 10** に示す．k は検索キーワードであり，$k1$〜$k20$ は各検索キーワードに対する被験者の平均精度である．各検索キーワードに対する熟知グループの抽出精度は，最小 0.9 であり，最大 1 であった．20 個の検索キーワードの平均精度である $ave20$ は，0.98 と非常に高い値となった．したがって，提案システムが提示する熟知グループの妥当性が十分高いことを示した．

4.1.2　システムが認定した熟知ブロガの妥当性評価

ここでは各熟知グループでのランキング（熟知度ランキング）上位に現れる熟知ブロガが，熟知ブロガとして妥当かどうかを評価した．

図 9 に示した各検索キーワードに対して，提案システムが提示したトップの熟知グループを**図 11** に示す．各熟知グループにおける上位 5 名のブロガが，熟知ブロガとして妥当か

図 8　ニュアンス比較表示部（検索語：健康，熟知グループ：ダイエット）
Fig. 8　Part of nuance comparison presentation (query keyword: health, knowledgeable group: diet).

k_1：J リーグ	k_2：GUNDAM	k_3：スイーツ	k_4：格闘技
k_5：鉄道	k_6：ジャニーズ	k_7：株	k_8：酒
k_9：サッカー	k_{10}：野球	k_{11}：競馬	k_{12}：美術
k_{13}：サザン	k_{14}：劇団四季	k_{15}：アイドル	k_{16}：声優
k_{17}：iPod	k_{18}：ソムリエ	k_{19}：linux	k_{20}：健康

図 9　評価実験で使用された 20 トピック
Fig. 9　20 topics used for evaluation experiments.

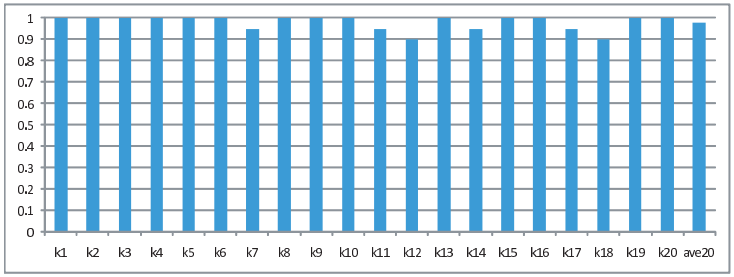

図 10　システムが提示する熟知グループの妥当性評価
Fig. 10　Appropriateness evaluation for knowledgeable groups presented by the system.

どうかを被験者により判定した．なお，各被験者は各ブロガのエントリをいくつか閲覧することで，熟知ブロガとして妥当であるかどうかを判断した．この評価では，“熟知ブロガとして提示される上位 5 名に対し，被験者が妥当だと判断した熟知ブロガ数の割合” で評価を行った．熟知ブロガの妥当性評価結果を**図 12** に示す．各熟知グループに対する熟知ブロガの抽出精度は，最小 0.65 であり，最大 1 であった．平均精度は 0.91 であり，提案システムが提示する熟知ブロガの妥当性が十分高いことを示した．精度が最小の 2 つの熟知グループは「g_{13}：サザン」と「g_{17}：アップル」であった．熟知グループ「g_{13}：サザン」に関しては，「サザン」熟知グループとシステムが判定したブロガのうち，「サザン」ではなく他のアーティストの熟知ブロガであったケースが見られたため，精度が 0.7 となった．これは熟知グループ「サザン」と他のアーティストの共起語辞書が音楽に関わる共通の単語を多く持つためである．対策としては，同種類（同階層）の熟知グループに対して，差分の共起語に高い重みを付けることが考えられる．熟知グループ「g_{17}：アップル」の共起語辞書には，アップル社に関する単語だけではなく，フルーツに関する単語も含まれていた．そのため，上位にランキングされたブロガには，リンゴを熟知するブロガも存在する．検索キーワード「k_{17}：iPod」を入力したユーザにとっては不適切なものであったため，精度が 0.65 となり，比較的低い値となった．対策としては，複数の意味を持つ熟知グループをさらに分割し，複

g_1：国内サッカー	g_2：ガンダム	g_3：洋菓子	g_4：K-1
g_5：乗り鉄	g_6：アイドル	g_7：株式	g_8：酒
g_9：高校サッカー	g_{10}：プロ野球	g_{11}：競馬	g_{12}：絵画
g_{13}：サザン	g_{14}：劇団四季	g_{15}：アイドル	g_{16}：アニメ
g_{17}：アップル	g_{18}：ソムリエ	g_{19}：linux	g_{20}：ダイエット

図 11　20 トピックのトップ熟知グループ
Fig. 11　Top knowledgeable groups for 20 topics.

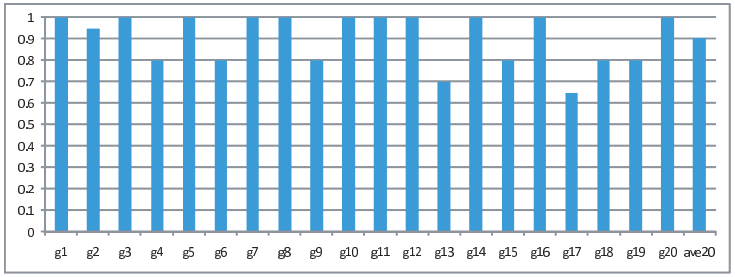

図 12　熟知ブロガの妥当性評価
Fig. 12　Appropriateness evaluation for knowledgeable bloggers.

数の熟知グループに分割することが考えられる．

4.1.3　ランク上位のエントリの妥当性評価

提案システムによる熟知度に基づくブログランキングで上位にランキングされるエントリの内容に関して，その価値や信憑性が高いかどうかを被験者により判定した．ここでは，各トピックに対して検索を行い，上位 n 個の熟知グループのランキングを表示させた際に，ランキング上位 5 件の中で信頼できるエントリの割合を調べた．上位 3 個の熟知グループを考慮した際のブログエントリの妥当性評価結果を**図 13** に，上位 5 個の熟知グループを考慮した際のブログエントリの妥当性評価結果を**図 14** に示す．

図 13 および図 14 により，上位 3 グループに限定した場合の妥当性評価結果は，平均精度が 0.75 となり，上位 5 グループの場合の妥当性評価結果は，平均精度が 0.67 となった．熟知グループの順位は，“対象熟知グループに属する全ブロガの中で対象期間内に検索キーワードを含むエントリを投稿したブロガ数とその割合” から算出されるが，上位 3 グループに限定した方が平均精度が高かったことから，この熟知グループの順位について妥当性があ

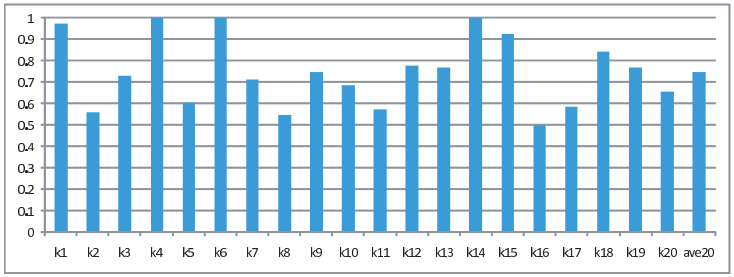

図 13　上位 3 個の熟知グループにおけるランク上位のエントリの妥当性評価
Fig. 13　Appropriateness evaluation of top-ranked entries for the top 3 knowledgeable groups.

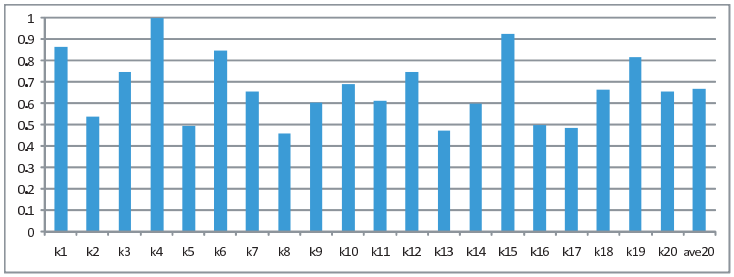

図 14　上位 5 個の熟知グループにおけるランク上位のエントリの妥当性評価
Fig. 14　Appropriateness evaluation of top-ranked entries for the top 5 knowledgeable groups.

ると考えられる．上位 3 グループに限定した場合と上位 5 グループの精度の差が大きかった検索キーワードは「k_{14}：劇団四季」であった．検索キーワード「k_{14}：劇団四季」の熟知グループを調べたところ，上位 5 個の熟知グループは「劇団四季」，「ミュージカル」，「オペラ」，「四季」，「宝塚歌劇団」であり，熟知グループ「劇団四季」，「ミュージカル」，「オペラ」と検索キーワードの関連性は「四季」，「宝塚歌劇団」より高いと考えられる．熟知グループ「四季」と「宝塚歌劇団」内のエントリには検索キーワードとの関連性が低いものが含まれたため，上位 5 グループを考慮した際の精度が低下した．

また，上位 5 グループに限定した場合の精度が 0.5 以下であった検索キーワード「k_5：鉄道」，「k_8：酒」，「k_{13}：サザン」，「k_{16}：声優」，「k_{17}：iPod」について調べた．「k_{13}：サザン」と「k_{17}：iPod」に対しては，システムが提示した熟知グループには，不適切なものが存在するため，結果としてエントリの精度が低下したものと考えられる．前項で説明したよ

うに，同種類の熟知グループに対して，差分の共起語に高い重みを付けたり，熟知グループの細分化を行ったりすることにより，「k_{13}：サザン」と「k_{17}：iPod」に関するエントリの精度が向上されることが期待できる．「k_5：鉄道」，「k_8：酒」と「k_{16}：声優」の熟知グループは比較的適切なものであったが，下位の熟知グループ内のエントリに無関係なものが見られたため，精度としては低い値となった．提案システムは，熟知度の高いブロガが書いたエントリは熟知度の低いブロガより信頼できるという考え方に基づいて，エントリをランキングしているが，エントリが検索キーワードとの関連度を深く考慮していない（キーワードを含むか含まないかのみを考慮している）．今後は，ブロガの熟知度と tf-idf をもとに算出した類似度を統合したスコアに基づくランキング手法を開発し，さらに精度を向上させる予定である．

全ブログの約 40%がスパムであるといわれているブログ空間において，エントリランキングの平均精度が 0.75 および 0.67 という値は，ブログの検索システムとしては十分高いと考える．本評価実験により，ブログ検索エンジンとして提案システムが十分有効であることを示せたと考える．

4.2　アンケート調査

前節の評価実験とは別に，公開中の実証実験システムを用いたアンケート調査を行った．以下にその結果を述べる．なお，ユーザが使用する検索キーワードは任意とした．アンケートの有効回答数は 125 件であった．

(1)　提示された熟知グループの妥当性について
システムが提示した関連熟知グループに対して，半数以上が妥当であると答えたユーザは，81%であり，提示される熟知グループの妥当性が高いことを示せた．

(2)　各熟知グループのブログの視点の特異性について
熟知グループ独特の視点で書かれたブログが抽出・ランキングされているかという質問に対して，同意と答えた人の割合は 88%であった．熟知グループごとに異なるランキングを提示することにより，異なる視点のブログを閲覧できるようにすることを目指しているが，本アンケートでもユーザ自身が熟知グループごとの視点の違いを認識することができたことを示せた．

(3)　熟知グループに所属するブロガの意見の信頼性について
熟知度順に整列したブロガ群（熟知度に基づくランキング）と直近時間順に整列したブロガ群（一般的なブログ検索のランキング方式）のどちらの意見が信頼できるかを質問した．熟知グループの上位ブロガが書いたエントリの方が信頼できると答えた人の割合は 69%であった．ブログ検索において，検索結果を直近時系列順に提示するよりも，提案手法である熟知グループに分類されたブログを熟知度に応じて提示する方式がより信頼できるという結果が得られた．

(4)　ブロガが属する熟知グループの割合の提示について
検索結果のスニペット部には，対象ブロガが所属する熟知グループの割合情報を付加した．対象ブロガがどのような熟知グループに属しているかを示すこの情報が検索に役立ったと答えたユーザは，79%であった．

5. 関 連 研 究

近年，グーグルブログ検索[5]，Technorati[6]，ヤフーブログ検索[7] などのブログ検索サイトが増加している．これらのサーチエンジンは，ブログに索引を付け，検索機能を提供する．商用サービスのほかに，ブログ検索やランキングに関する学術的な研究が進められている．たとえば，ブロガのハブ度とオーソリティ度を求めることにより，ブログエントリをランキングする EigenRumor 手法が提案されている[8),10)]．この手法では，良いブロガが書いたブログエントリにはより高いスコアを与えることで，ほかのブログにリンクを張られていないエントリでも高いスコアが得られる．Kritikopoulos ら[9] は，実リンクと虚リンクからなるリンクグラフの解析により，ブログエントリをランキングする手法を提案している．実リンクはエントリ間のハイパーリンクであり，虚リンクはエントリ間の類似度に基づいて生成されたリンクである．これらのリンク解析に基づく手法に対して，本研究はブログの内容分析に基づきブロガとブログエントリをランキングする手法を提案する．また，ブログ検索システム BLOGRANGER[10] では，検索結果から抽出された固有名詞をトピックとし，検索結果から抽出された形容詞・形容動詞を評価表現とし，ブログエントリをトピックや評価表現ごとに整理している．BLOGRANGER システムが検索結果中のブログエントリの内容分析によってトピックや評価表現などの側面からエントリを分類するのに対し，我々の提案システムは，検索結果中のエントリを書いたブロガが熟知している領域によってエントリをまとめている．

信憑性に関する研究としては，Gil ら[11] は，ウェブコンテンツが信頼できるかどうかを判断するための要素をあげている．このような要素を見つけることは実際には難しいため，最小のユーザ操作でシステムの最大の信頼度が得られる要素を抽出している．Adler ら[12] は，ウィキペディアにおいて，評判の高い作者を抽出する内容駆動の評価システムを提案している．基本的なアイデアは，ある作者による編集結果を，その後他の作者が修正すること

なくそのまま採用した場合は，元の作者はより高い評価を獲得し，逆に他の作者が修正した場合には元の作者の評価が下がるという考え方である．Andersen ら[13] は，ソーシャルネットワークの構造を分析し，信頼度に基づいた推薦システムを提案している．これらの研究が扱ったメディアと異なり，本研究はブログの信頼度に着目する．提案手法は，ブロガの過去のブログ履歴を分析することで信頼できるブロガを抽出し，ブロガの熟知度に基づいて信頼できるブログエントリを発見するものである．

6. おわりに

本研究では，熟知度に基づくブログランキングシステムを開発し，実証実験システムとして Web 上で公開した．実施した評価実験では，開発したシステムにより提示される，熟知グループ，熟知ブロガ，ブログエントリの妥当性が十分高いことを示した．

今後はさらなる精度向上を目指すとともに，信頼性の高い情報の取得を支援できるような技術を目指して，本研究を発展させていくつもりである．

謝辞　この研究は，独立行政法人情報通信研究機構の高度通信・放送研究開発委託研究「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課題ア Web コンテンツ分析技術」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．

参考文献

1) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks*, Vol.30, No.1-7, pp.107–117 (1998).
2) 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).
3) 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).
4) Kizasi Blog Search. http://kizasi.jp/labo/nict_h20/
5) Google Blog Search. http://blogsearch.google.co.jp/
6) Technorati. http://www.technorati.com/
7) Yahoo! Blog Search. http://blog-search.yahoo.co.jp/
8) Fujimura, K., Inoue, T. and Sugisaki, M.: The EigenRumor Alogorithm for Ranking Blogs, *WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2005).
9) Kritikopoulos, A., Sideri, M. and Varlamis, I.: BlogRank, Ranking Weblogs Based on Connectivity and Similarity Features, *AAA-IDEA* (2006).
10) 戸田浩之，藤村　考，井上孝史，廣嶋伸章，杉崎正之，片岡良治，奥　雅博：目的指向型ブログ検索システム BLOGRANGER の提案およびユーザ評価，情報処理学会論文誌：データベース，No.SIG 14 (TOD 35), pp.132–151 (2007).
11) Gil, Y. and Artz, D.: Towards Content Trust of Web Resources, *WWW*, pp.565–574 (2006).
12) Adler, B.T. and de Alfaro, L.: A Content-driven Reputation System for the Wikipedia, *WWW*, pp.261–270 (2007).
13) Andersen, R., et al.: Trust-based Recommendation Systems: An Axiomatic Approach, *WWW*, pp.199–208 (2008).

(平成 22 年 3 月 20 日受付)
(平成 22 年 7 月 7 日採録)

（担当編集委員　河野 浩之）

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 年）．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 Web マイニングおよび情報検索・情報推薦の研究に従事．日本データベース学会，IEEE CS 各会員．

張　　建偉（正会員）

京都産業大学コンピュータ理工学部特定研究員．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）．Web マイニング，Web 情報システム，Web 情報信憑性分析の研究に従事．日本データベース学会会員．

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．

桑原　　雄

（株）きざしカンパニー技術研究員．2003 年電気通信大学電気通信学部システム工学科卒業．2005 年電気通信大学大学院電気通信学研究科システム工学専攻博士前期課程修了．