推薦論文

反響特性分析を利用したブログ記事検索手法

宮田章裕[†] 松岡寿延[†,☆] 岡野真一[†,☆]
山田節夫[†] 石打智美[††]
荒川則泰[†,☆☆] 加藤泰久[†,☆☆☆]

本研究では，ブログ記事を取り巻く人々のインタラクションの様子「反響特性」を利用したブログ記事検索手法を提案する．ブログ記事は被リンク数が非常に少ない等，従来の Web ページとは異なった性質を持つため，既存の検索手法が提示する記事を人間が読んでも有用と感じられない場合がある．また，既存検索手法は大衆から注目を集めている情報の提示に重点を置いているため，個々の閲覧目的に応じた記事を提示することができない．そこで本研究では，ブログ記事の書き手・読み手の行動は記事に対する人間の判断を反映しているという仮説に基づき，彼らのインタラクション情報を多面的に分析して「広さ，強さ，速さ，長さ」からなる多次元指標「反響特性」を定義する．この反響特性を用いることによって,「幅広い人から関心を集めている記事」や「長期間関心を集め続けている記事」というように，人間が実際に読んで有用と感じられる記事の多面的な観点から検索することが可能になる．提案手法が有用な記事を提示できるか検証実験を行ったところ，既存手法よりもユーザの満足度が高くなるという結果が得られ，提案手法に一定の有効性が確認できた．

Blog Search Method Based on Analysis of Response to a Blog Entry

Akihiro Miyata,[†] Hisanobu Matsuoka,[†,☆] Shinichi Okano,[†,☆]
Setsuo Yamada,[†] Satomi Ishiuchi,[††] Noriyasu Arakawa[†,☆☆]
and Yasuhisa Kato[†,☆☆☆]

We present a method of searching blog entries that takes into account the response derived from the interaction of the blogger and readers. Here, response is composed of width, strength, speed, and length. These indicators enable users to find entries in which many readers are interested, in which a small group has an enthusiastic interest, in which a few members are chatting at intervals of a few seconds, and in which many readers have shown interest over a long period. An experiment conducted with a prototype system is presented. For a particular situation (searching movie reviews), our method has shown better results than an existing blog search service.

1. はじめに

2006 年 3 月の時点でブログの国内開設数は 860 万件を超え[1)]，今やブログはメディアとしての地位を確立しつつある．2006 年 2〜3 月に我々が実施したアンケート調査（対象：ブログ・SNS 開設者 1,015 名）においても，約 4 割（393 名）がブログ記事検索サービスを利用して自ら能動的にブログ記事を閲覧したことがあると回答した[2)]．

読み手の目的も実に多種多様で，前述の被験者にどのようなブログ記事を見つけようとしたか問うと (N=393，複数回答),「自分が探している情報・答え（53.69%）」,「クチコミ情報（20.61%）」や「楽しい読み物（30.53%）」,「情報交換できそうな人の記事

† 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
†† 日本電信電話株式会社 NTT 知的財産センタ
NTT Intellectual Property Center, NTT Corporation
☆ 現在，日本電信電話株式会社 NTT サービスインテグレーション基盤研究所
Presently with NTT Service Integration Laboratories, NTT Corporation
☆☆ 現在，日本電信電話株式会社 NTT ネットワークサービスシステム研究所
Presently with NTT Network Service Systems Laboratories, NTT Corporation
☆☆☆ 現在，日本電信電話株式会社
Presently with NTT Corporation

本論文の内容は平成 18 年 11 月のグループウェアとネットワークサービスワークショップ 2006 にて報告され，GN 研究会主査により情報処理学会論文誌への掲載が推薦された論文である．

(12.72%)」という回答が得られた．また，米国で行われた意識調査でも「エンタテインメント」,「ニュース」,「他者との交流」と様々な閲覧目的が確認されている[3]．つまり，国内外問わずブログ記事の閲覧目的は非常に多様であり，目的に応じた記事に読み手をナビゲートする技術が今後いっそう求められることは明らかである．

しかし，ブログ記事は被リンク数が非常に少ない等，従来の Web ページとは異なった性質を持つため，既存の Web ページ検索技術をそのままブログ記事検索に適用しても，人間が実際に目を通したときに有用と感じられない記事が提示される場合が少なくない．また，既存検索手法は大衆から注目を集めている（アクセス数が多い，被リンク数が多い等）情報の提示に重点を置いているが，注目を「どのように」集めているかということまで考慮しているものは少ない．それゆえ，前述のとおりユーザの閲覧目的は様々であるのに，個々の閲覧目的に応じた記事を提示することができない．

そこで本研究では，人間が実際に読んだときに有用と感じられるようなブログ記事を検索でき，その結果を多面的に選択できる技術の確立を目指す．まず我々は，ブログ記事を実際に書いた，あるいは読んだ人間の記事に対するインタラクション情報を利用して記事の分析を行う．なぜならば，ブログ記事の書き手・読み手の行動は，その記事に対する人間の判断を反映していると思われるからである．そして，このインタラクション情報を多面的に分析して,「広さ・強さ・速さ・長さ」からなる反響特性という指標を導出する．このように，実際に記事を読んだ人間が起こした行動を，多面的（広さ，強さ，速さ，長さ）に分析する手法「反響特性分析」を利用したブログ記事検索手法を提案する．提案方式を用いれば,「幅広い人から関心を集めている記事」や「長期間関心を集め続けている記事」というように，人間が実際に読んで有用と感じられる記事を多面的な観点から検索することが可能になる．

以降，2 章で現状のブログ記事分析に利用されている属性について述べ，3 章で反響特性を利用したブログ記事検索手法を提案する．4 章では利用属性の調査・検討について述べ，5 章では反響特性スコアの導出アルゴリズムについて述べる．そして，6 章で提案手法の評価実験について報告し，7 章で本稿の結論と今後の方針を述べる．

2. 現状のブログ記事分析に利用されている属性

現在，ブログ関連サービス・研究事例では，記事を評価・分析する際に様々な属性が利用されている．ここでは，各属性の利用形態，利用の実施例について言及する．

2.1 アクセス数

大半のブログサービス事業者では，自社が提供しているブログのアクセス数をランキング形式で公開している．これにより，多くの読み手の関心を集めているブログ記事を発見することができる．

しかし，アクセスランキングには，上位がほぼ固定的であるという問題がある．たとえば，2006 年 4〜8 月の 100 日間で goo ブログ[4] のアクセスランキング上位 100 位以内に登場した 10,000 件のブログのうち，上位 10 位以内に登場するユニークなブログはわずか 76 件である．これは，数百万件と存在しているブログの中でごく一部のものしか目立っていないことを示唆している．

また，ブログのサイト構造に起因する問題もある．一般的なブログではトップページに複数の記事（最新 10 件程度）が表示されるケースが多い．そのため，トップページにアクセスした閲覧者がどのブログ記事を読んでいるのか，アクセス履歴から識別することはできない．

2.2 ブックマーク数・FOAF 数

ブログ記事がブックマークされたとき，その記事は読み手から関心を持たれたと判断できるだろう．有益な記事は多くの読み手からブックマークされることが各種ソーシャルブックマークサービスにおいても確認できる．しかし，記事単位でブックマークされるような記事は資料的性質が強いものに偏りがちである（プログラミング Tips，面接のコツ，冠婚葬祭のマナー等）．

対して，クチコミ情報記事や交流が起きている記事といったブログ記事に目立つ内容は長期にわたって保存すべき性質のものではない場合が多く，ブックマークされる機会はさほど多くないと思われる．ブログサイト単位でブックマークされている例は珍しくないが（著名人や友人のブログである場合が多い），この場合は読み手の関心を記事単位で評価することが難しい．最後の点においては FOAF（人のつながりをメタデータで表現したもの，友人のブログサイトリストとしてブックマークのように使われることが多い）の場合も同様である．

2.3 リンク構造

現在，主要な Web 検索サービスでは PageRank[5] や HITS[6] 等のリンク構造解析に基づくページスコア判定手法が用いられている．この手法は，価値のあるページからリンクを受けているページほど価値が高いという理念に基づいており，Web 検索においてはかなりの効果をあげている．しかし，従来の Web ページと比較するとブログ記事が外部からリンクを受けることは少なく，他のブログ記事からリンクされているブログ記事の数は約 1%程度にすぎないという報告もある[7]．

この点を考慮して考案された手法が EigenRumor[7] である．この手法では，ブログの編集主体がほとんどの場合一個人である点に着目し，各ブログ記事のスコアをブロガの属性として集約している．このため，1 つ 1 つは被リンクがない・少ないブログ記事であっても，1 人のブロガが執筆したブログ記事集合全体で見れば被リンクに基づいた評価が可能となり，検証実験では約 9.3%（PageRank の場合は 1.15%のみ）のブログ記事に非ゼロのスコアを付与することができたという．

2.4 コミュニケーション

ブログ記事上で行われているコミュニケーションを利用して記事の特徴を分析する手段も検討されている．Mishne らは，記事上で行われるコミュニケーションの 1 つであるコメントの数と，記事の人気が関係していると指摘している[8]．彼らはアクセス数と被リンク数をブログの人気の指標としており，これらの数値とコメント数には正の相関が見られたという．

また，分析対象は Web 掲示板のスレッドであるが，CMINER というコミュニケーション分析システムも考案されている[9]．このシステムでは Web 掲示板のスレッド上における各ユーザの発言数や発言間隔，ユーザ間の関わり合いの深さ等を計測することができる．

3. 反響特性分析を利用したブログ記事検索手法の提案

3.1 研究目標

2 章で述べたとおり，ブログ記事分析に利用可能な属性はいくつかあるが，人間が実際に読んで有用と感じられる記事を多面的な観点から検索できるサービスの実現を考えた場合，これらの属性はそれぞれに問題点を含んでいる．

アクセス数は上位が固定的であり一部の記事しか目立たないため，検索という行為にそのまま適用するのは難しい．また，ブログは 1 つのページに複数の記事が表示される場合が多く，記事単位でアクセス数を把握できない場合が多々ある．被ブックマーク数も記事の人気を示す属性であるが，ブックマークされるような記事は資料的性質が強いものに偏りがちである．リンク構造分析は従来の Web ページ検索では大きな効果をあげているが，ブログ記事は被リンク数が少ないため，ブログサイト単位で被リンク数を集計するという工夫[7] を行っても，非ゼロのスコアを付与できる記事は全体の 10%弱である．

逆に，検索という行為は検索キーワードを用いて行われる場合が大半であるため，検索キーワードへの適合度のみを考慮して記事検索を行うという解もあるだろう．実際，2006 年 12 月時点で大半のブログ記事検索サービスが提供する「適合度順検索」とは，検索キーワードへの文書適合度のみが考慮されている場合がほとんどである．しかし，ブログは文中で固有名詞の愛称・略称が多用される等，非常に表記揺れが大きいメディアである．内容も玉石混淆と称されることが多く，個人の備忘録を目的としたもの，クチコミ情報の発信を目的としたもの，コミュニケーションを目的としたもの，それらが混ざったもの等，様々なタイプの記事が混在している．しかも，簡単な操作で記事を作成することができるため，このような玉石混淆の記事が日々大量に発生し続けている．さらに，コメントやトラックバックといった機能を利用して，書き手や読み手が動的に文章を書き加えることもしばしばである．よって，文書適合度という情報は必要ではあるが，これだけで我々が目指しているような記事検索が行えるとは考えにくい．

上記のような問題があるため，既存の検索手法だけでは不十分であり，(1) 検索結果の記事を人間が実際に読んでみても有用と感じられない場合が多い．また，既存手法では大衆から注目を集めている（アクセス数・被ブックマーク数・被リンク数が多い等）情報の提示に重点を置いてきたが，ブログ記事検索に対するニーズを調査すると，注目を「どのように」集めているか知りたいという声や，注目されている記事以外にも共感できる記事，交流できる記事を発見したいといった要望が多く見受けられているため，(2) 記事が注目を集めているかどうか測定するだけでは個々の閲覧目的に応じることができない．

そこで本研究では，(1) 人間が実際に読んだときに有用と感じられるようなブログ記事を検索でき，その結果を，(2) 多面的に選択できる技術の確立を目指す．

3.2 反響特性

我々は，ブログ記事を実際に書いた，あるいは読ん

だ人間の記事に対するインタラクション情報を利用して記事の分析を行う．なぜならば，ブログ記事の書き手・読み手の行動は，その記事に対する人間の判断を反映していると思われるからである．たとえば，人気がある記事や洞察力がある記事にはコメント数が多いという調査結果があり[8),10)]，物議を醸すような内容の記事では書き手・読み手がコメントを利用して議論を始める傾向が見られることも指摘されている[8)]．また，資料的価値が高い記事は，他の記事からの参照行為であるトラックバックを多く受ける傾向も見受けられる．読み手からコメントを受けると書き手の執筆意欲が向上するという現象[11),12)]も，書き手と読み手のインタラクションの 1 つととらえてよいだろう．ブログの機能・構成に鑑みても，記事単位でコメントやトラックバックといったコミュニケーション機能を備えており，記事の編集主体がブロガという形で見えやすく，書き手や読み手のインタラクションが発生しやすい．よって，このように行動情報を利用するアプローチは，ブログ記事を分析する場合には特に有効であると思われる．

そこで，あるブログ記事を取り巻く人々が互いに，あるいは記事そのものとインタラクションを起こしている状態を，その記事が「反響を呼んでいる」というイメージでとらえ，そのインタラクションの様子を「反響特性」と定義する．すなわち，記事を取り巻く書き手・読み手の行動や，彼らの行動の発端である記事を分析することで導出される指標が反響特性である．また，多面的な検索を実現するためには反響特性は一次元指標ではなく，下記のような複数の要素からなる多次元要素であるべきと考えた．

- **広さ**：インタラクションしている人の幅の広さ
- **強さ**：インタラクションの行為の強さ
- **速さ**：インタラクションが発生する速さ
- **長さ**：インタラクションが継続する長さ

上記のように，反響特性を多次元要素の指標とすることでブログ記事の様子をより詳細・多面的に分析することが可能になる．たとえば，「コメント数」という一次元情報だけでも記事が多くの人の関心を集めているかどうか推測できるが，少数の常連メンバが活発な議論を行っているような記事はコメント数が多いわりに世間的には人気がないといった例も報告されている[8)]．これに対して，反響特性では「幅広い人がある程度の関心を示している記事（広さ：大，強さ：普通）」や「少数の人が熱烈に関心を示している記事（広さ：小，強さ：大）」，「少数の人がチャットのように数十秒間隔でコメントを交わし合っている記事（広さ：小，強さ：大，速さ：大）」，「幅広い人から長きにわたって関心を集め続ける記事（広さ：大，長さ：大）」等を区別することができる．

このように，実際に記事を読んだ人間が起こした行動を，多面的（広さ，強さ，速さ，長さ）に分析する手法「反響特性分析」を利用したブログ記事検索手法を提案する．

4. 利用属性の調査・検討

ここでは，提案手法で利用可能な属性の調査・検討を行う．3.2 節で述べたとおり，反響特性を導出するためには読み手・書き手の行動を収集して分析する必要がある．また，彼らの行動の発端となった記事そのものの属性を考慮する必要もあるだろう．しかし，アクセス数のように容易に取得できない指標も多々ある．そこで，比較的容易に取得可能であり，かつ，ブログ記事を取り巻く人々のインタラクションとして重要な役割を果たしているコメントとトラックバックを中心に利用することにした．**表 1** に示すのは，利用する属性と反響特性の関係である．

なお，トラックバックもコメント同様，送信者のユニーク数や 1 人あたりの送信数等を分析すべきである

表 1 利用属性
Table 1 Attributes.

属性	説明・反響特性との関係
コメント元数	コメントを送信している人のユニーク数.「広さ」に寄与.
1 人あたりのコメント数	コメント送信者 1 人あたりの平均コメント送信数.「強さ」に寄与.
初コメントまでの時間	記事が投稿されてから初めてコメントが送信されるまでの時間.「速さ」に寄与.
最新コメントまでの時間	記事が投稿されてから分析時点での最新コメントが送信されるまでの時間.「長さ」に寄与.
平均コメント時間間隔	各コメントが送信された時間間隔の平均値.「速さ」に寄与.
トラックバック数	記事に送信されたトラックバックの数.「広さ」に寄与.
初トラックバックまでの時間	記事が投稿されてから初めてトラックバックが送信されるまでの時間.「速さ」に寄与.
最新トラックバックまでの時間	記事が投稿されてから分析時点での最新トラックバックが送信されるまでの時間.「長さ」に寄与.
平均トラックバック時間間隔	各トラックバックが送信された時間間隔の平均値.「速さ」に寄与.
その他	コメントリンク率，記事の文字数，記事中のリンク数，記事中の画像数．コメントリンク率とはコメント総数に対する，コメント送信者の URL が書いてあるコメント数の割合.

表 2 調査結果概要（CM=コメント，TB=トラックバック）
Table 2 Surveyed attributes (CM=Comment, TB=Trackback).

(1) 分析対象：全ブログ記事 981,197 件

	平均値	最小値	最大値
CM 数	3.9	0.0	2,000.0
TB 数	0.8	0.0	544.0
記事の文字数	483.9	1.0	11,163.0
記事中のリンク数 *1	0.6	0.0	255.0
記事中の画像数 *2	1.1	0.0	66.0

*1 投票サイト等へのリンクは含まれていない.
*2 絵文字アイコンは含まれていない.

(2) 分析対象：CM が 1 つ以上あるブログ記事 339,291 件

	平均値	最小値	最大値
CM 数	4.8	1.0	2000.0
CM 元数	3.4	1.0	308.0
1 人あたりの CM 数	1.3	1.0	105.5
初 CM までの時間	35.7 hour	2.0 sec	2,191.1 day
最新 CM までの時間	92.6 hour	2.0 sec	2,191.4 day
平均 CM 時間間隔	30.1 hour	2.0 sec	1,095.7 day
CM リンク率	0.6	0.0	1.0

(3) 分析対象：TB が 1 つ以上あるブログ記事 140,479 件
（ただし「TB 数」以外の項目は TB 付与日時が取得できた 14,613 件）

	平均値	最小値	最大値
TB 数	2.5	1.0	544.0
初 TB までの時間	18.2 day	6.0 sec	11,309.8 day
最新 TB までの時間	20.9 day	17.0 sec	11,309.8 day
平均 TB 時間間隔	19.3 day	17.0 sec	11,309.8 day

が，1 つの記事に 1 人が複数回トラックバックを送信するケースはあまり見受けられないので，今回これらの指標の利用は見送った．

4.1 分析に利用したデータ

反響特性分析のアルゴリズム決定のためには，表 1 であげた属性の統計的特徴を把握しておく必要があると考え，各属性を実際のブログ記事から抽出して分析を行った．

分析のために収集したブログ記事は 1,130,002 件である．これらは 1997/2/24〜2006/11/27 に投稿されたものであり，投稿されてから少なくとも 1 カ月以上経過した後に収集した．これは，記事が投稿された直後に収集してしまうと，今後付与されるかもしれないコメントやトラックバックを数え損ねてしまう恐れがあるからである．なお，コメントやトラックバックが付与された日時が記事投稿日時よりも過去になっている記事（まれに未来の日時で記事を投稿するケースがあるため）や，URL としては存在しているが中身がない記事は分析対象から外したため，最終的には 981,197 件のブログ記事に対して分析を行った．その調査結果概要を**表 2** に示す．

表 3 分布の詳細（CM=コメント，TB=トラックバック）
Table 3 Distribution of attributes (CM=Comment, TB=Trackback).

(1) 分析対象：CM が 1 つ以上あるブログ記事 339,291 件

パーセンタイル	0.1%	75.0%	99.9%
CM 数	-	6.0	93.0
CM 元数	-	4.0	44.0
1 人あたりの CM 数	-	1.5	8.9
初 CM までの時間	62.0 sec	28.8 hour	127.9 day
最新 CM までの時間	-	82.0 hour	203.8 day
平均 CM 時間間隔	97.5 sec	24.2 hour	85.3 day
CM リンク率	-	1.0	-

(2) 分析対象：TB が 1 つ以上あるブログ記事 140,479 件

パーセンタイル	0.1%	75.0%	99.9%
TB 数	-	2.0	117.0
初 TB までの時間	28 sec	10.1 day	1,445.6 day
最新 TB までの時間	-	14.0 day	1,543.1 day
平均 TB 時間間隔	34 sec	13.1 day	1,445.6 day

表中の「-」は未測定，もしくは測定不能を示す.

4.2 コメント/トラックバックの利用状況

表 2 (1) は今回分析対象とした全ブログ記事 981,197 件に関するデータである．

表 1 で示した属性を利用するためにはコメントかトラックバックが付与されていることが必要であるが，表 2 (1) によると 1 つのブログ記事にコメントは平均 3.9 個，トラックバックは平均 0.8 個付与されていることが分かる．ただし，ネットワーク上のデータはべき乗分布に従う場合が多いため，コメント・トラックバックを大量に集めているブログ記事がごく少数存在し，大半のブログ記事はコメントもトラックバックも付与されていないという可能性もある．

そこで，コメント/トラックバックが 1 つ以上あるブログ記事の数を調査すると，それぞれ 339,291 件/140,479 件であった．重複を考慮すると，計 420,671 件（分析対象の 42.9%）が 1 つ以上のコメントもしくはトラックバックを持っていることになり，これらの記事に対しては反響特性分析を行うことができる．

4.3 コメントに関する詳細検討

表 2 (2) は今回収集したブログ記事の中でコメントが 1 つ以上あった 339,291 件に関するデータである．また，**表 3** (1) に示すのが分布の詳細である．

これらの表を見ると，たとえば「コメント元数」は 75.0 パーセンタイルで 4 個☆であるが最大値は 308 個であり，やはりべき乗分布の傾向が見受けられる．「コメントリンク率」以外の指標に関しても同様の傾向であった．ブログ記事検索のアプローチとして，各指標

☆ 分析対象エントリの 75.0%がコメント元数 4 個以下であるということ.

(「コメント元数」等）が極端に大きい/小さい記事だけに注目するという解もあるが，3.1 節で述べたように多面的な検索を実現するためにはごく一部の記事のみが評価される方法は適切ではない．逆に，値が極端な場合はスパムコメント/トラックバック等によるノイズの可能性もあり，このような値は解析対象から外すという解もある．しかし，実際に「コメント元数」が最大（308 人）のブログ記事を閲覧してみると政治問題について大勢が意見を交わし合っている記事であり，これらのコメントはスパムとはいえない．

そこで，各指標の値に「有効上限値/有効下限値」を設定し，それ以上/以下の値は「有効上限値/有効下限値」と同等と見なすアプローチを採ることにした．この方法であれば，各指標で極端な値を持つブログ記事を異常値として切り捨てることなく，かつ，このような記事ばかりが極端に目立つことがなくなり検索の多面性を維持できると思われる．今回は 99.9 パーセンタイルの値を有効上限値，0.1 パーセンタイルの値を有効下限値とした．

4.4 トラックバックに関する詳細検討

表 2 (3) は今回収集したブログ記事の中でトラックバックが 1 つ以上あった 140,479 件に関するデータである．また，表 3 (2) に示すのが分布の詳細である．ただし,「初トラックバックまでの時間」,「最新トラックバックまでの時間」,「平均トラックバック時間間隔」に関してはトラックバック元ブログ記事の生成日時が取得できた 14,613 件に関するデータである．

トラックバックに関しても，75.0 パーセンタイルで 2 個に対して最大値は 544 個とべき乗分布の傾向が確認できたので，コメントと同様に 99.9 パーセンタイル値を有効上限値，0.1 パーセンタイル値を有効下限値とした．

4.5 ブログ記事の収集方法に関する検討

4.1 節でも述べたが，記事が投稿されてから収集するまで一定期間をおかなければ，今後付与されるかもしれないコメントやトラックバックを数え損ねてしまう恐れがある．しかし，対象ブログ記事すべてにコメントやトラックバックが付き終わるのを待ってからでないと分析できないのでは，実際の検索サービスとしては成り立ちにくい．

実サービスとして運用する際は，すべての記事ではなく，ある一定量の記事にコメント/トラックバックが付き終わるまで待って収集する方が現実的と思われる．ここで，今回の分析対象が少なくとも 1 カ月以上前に投稿されたブログ記事であるので，最新コメント/トラックバックを最終コメント/トラックバックとほぼ同義と見なせば表 3 の値が参考になる．例えば全体の 75%の記事にコメント/トラックバックが付き終わっていれば十分という判断であれば，表 3 (1)，(2) の 75.0 パーセンタイル値の期間だけ待てば済む．つまり，コメントに関していえば投稿されてから 82.0 時間後，トラックバックも考慮するのであれば投稿されてから 14.0 日後に収集すればよい．

5. 反響特性スコアの導出アルゴリズム

反響特性（*hankyo*）のスコアは，式 (1) のように「広さ（*width*），強さ（*strength*），速さ（*speed*），長さ（*length*），その他（*etc*）」の各スコアの線形和で表現することとした．W_1〜W_5 は各スコアの重み変数であり，可変である．

$$\begin{aligned} Score(hankyo) = {} & W_1 \times Score(width) \\ & + W_2 \times Score(strength) \\ & + W_3 \times Score(speed) \\ & + W_4 \times Score(length) \\ & + W_5 \times Score(etc) \end{aligned} \quad (1)$$

式 (2)-1〜5 のように,「広さ」のスコアは「コメント元数（*uniqCmSenderNum*），トラックバック数（*tbNum*）」,「強さ」のスコアは「1 人あたりのコメント数（*cmNumPerSender*）」,「速さ」のスコアは「初コメントまでの時間（*firstCmElapsedTime*），平均コメント時間間隔（*avgCmInterval*），初トラックバックまでの時間（*firstTbElapsedTime*），平均トラックバック時間間隔（*avgTbInterval*）」,「長さ」のスコアは「最新コメントまでの時間（*latestCmElapsedTime*），最新トラックバックまでの時間（*latestTbElapsedTime*）」,「その他」のスコアは「コメントリンク率（*linkCmRate*），記事の文字数（*entryLength*），記事中のリンク数（*linkNum*），記事中の画像数（*imageNum*）」の各スコアの線形結合から導出している．C_{w1}〜C_{e4} は各スコアに係る定数である．

$$\begin{aligned} & Score(width) \\ & \quad = C_{w1} \times Score(uniqCmSenderNum) \\ & \qquad + C_{w2} \times Score(tbNum) \end{aligned} \quad (2)\text{-}1$$

$$\begin{aligned} & Score(strength) \\ & \quad = C_{st} \times Score(cmNumPerSender) \end{aligned} \quad (2)\text{-}2$$

$$\begin{aligned} & Score(speed) \\ & \quad = C_{sp1} \times Score(1/firstCmElapsedTime) \\ & \qquad + C_{sp2} \times Score(1/avgCmInterval) \\ & \qquad + C_{sp3} \times Score(1/firstTbElapsedTime) \\ & \qquad + C_{sp4} \times Score(1/avgTbInterval) \end{aligned} \quad (2)\text{-}3$$

$$Score(length) = C_{l1} \times Score(latestCmElapsedTime) + C_{l2} \times Score(latestTbElapsedTime) \quad (2)\text{-}4$$

$$Score(etc) = C_{e1} \times Score(linkCmRate) + C_{e2} \times Score(entryLength) + C_{e3} \times Score(linkNum) + C_{e4} \times Score(imageNum) \quad (2)\text{-}5$$

Score (*uniqCmSenderNum*) 等の各ブログ記事属性のスコアは式 (3) のように求めている．HT，LT はそれぞれ「有効上限値」,「有効下限値」である（4.3 節，4.4 節参照）．x_i はブログ記事（id=i）の各指標（コメント元数等）の値，$Max(x)$ は全ブログ記事中の各指標の最大値である．また，有効上限値/有効下限値を設けてもコメントリンク率以外の各指標は依然べき分布となっているので，$Log(x)$ で自然対数をとり正規化を行っている．

表 2，3 においてべき乗分布の傾向が見られた「コメント元数，トラックバック数，1 人あたりのコメント数，最新コメントまでの時間，最新トラックバックまでの時間，記事の文字数，記事中のリンク数，記事中の画像数」の各スコアは式 (3)-1 で導出する．x_i はブログ記事（id=i）の各指標（コメント元数等）の値，HT は「有効上限値」である．また，有効上限値を設けても依然べき分布となっているので，$Log(x)$ で自然対数をとり正規化を行っている．

$$Score(x_i) = \begin{cases} Log(x_i)/Log(HT) & : x_i \leq \mathrm{HT} \\ 1 & : x_i > \mathrm{HT} \end{cases} \quad (3)\text{-}1$$

同様に表 2・3 においてべき乗分布の傾向が見られたが，値の逆数を利用する「初コメントまでの時間，平均コメント時間間隔，初トラックバックまでの時間，平均トラックバック時間間隔」の各スコアは式 (3)-2 で導出する．LT は「有効下限値」であり，式 (3)-1 と同じく対数正規化を行っている．

$$Score(1/x_i) = \begin{cases} Log(LT)/Log(x_i) & : x_i \geq \mathrm{LT} \\ 0 & : x_i < \mathrm{LT} \end{cases} \quad (3)\text{-}2$$

表 2，3 で比較的一様な分布を見せた「コメントリンク率」のスコアは式 (3)-3 で導出する．$Max(x)$ は対象指標の全ブログ記事中における最大値である．

$$Score(x_i) = x_i/Max(x) \quad (3)\text{-}3$$

6. 評価実験

6.1 実験計画

3.1 節で述べたとおり，本研究の目標は，(1) 人間が実際に読んだときに有用と感じられるようなブログ記事を検索でき，その結果を，(2) 多面的に選択できる技術の確立である．つまり，提案手法の有効性を評価するためには (1)，(2) が実現できているかどうかという観点から実験を行う必要がある．本稿では (1) の観点から実施した検証[13] について，以降その詳細を述べる．

6.2 実験目的とシーン設定

この実験の目的は，ユーザが実際に読んだときに有用と感じられるようなブログ記事を提案手法が提示できるかどうか検証することである．ただし,「有用」という曖昧な表現だけでは被験者によって価値判断が異なり，適切な検証が行えないと考えられる．

そこで，本実験では「多くの情報が第三者にとって分かりやすく書かれている映画のレビュー記事を探す」というシーンを設定し，多くの情報が分かりやすく書かれているほど「有用」と判断することにする．以降 6 章内においては，この観点で「有用」という表現を用いる．

6.3 実験準備

実験を開始する前に，有用な記事の性質を調査するため映画に関する記事の収集を行う．人間が実際に目を通して有用と感じた記事の各属性（「コメント元数」等）の分布を調べ，それを教師データとすることで，有用な記事の提示に適した式 (1)〜(2) の各パラメータ（W，C）を決定するためである．

ブログ記事収集手段としては，goo ブログ[4] で提供されている「goo ブログ検索」を利用する．具体的には，複数の映画タイトルでそれぞれ記事検索を行い，検索結果として得られた全記事☆の中から無作為に計 100 件を抽出する．そして，今回は 1 つずつ著者が実際に閲覧して有用かどうか 0〜3 点の 4 段階で採点した（以降，この採点結果を「ユーザ評価」とする）．さらに，検索時に利用した映画タイトルと記事本文の文書適合度も求める☆☆．

表 4 に示すのは，各記事の属性とユーザ評価との間のスピアマンの相関係数 ρ である．表中の※はコメントが 1 個以上付いている記事 64 個に関するデータで

☆ この実験を実施した 2006 年 6 月時点の goo ブログ検索結果として得られる全記事数は 500 件である．

☆☆ 今回は，goo ブログ検索における文書適合度順検索結果順位の逆数を文書適合度とした．

表 4 各パラメータとユーザ評価の相関係数
Table 4 Coefficiients between attributes and the user rating.

	ρ
文書適合度	0.03
コメント元数	0.62
1 人あたりのコメント数	0.49
初コメントまでの時間※	0.40
最新コメントまでの時間※	0.69
平均コメント時間間隔※	0.61
トラックバック数	0.85
コメントリンク率※	0.73
記事の文字数	0.47
記事中のリンク数	0.10
記事中の画像数	0.16

ある．これらの相関係数に基づいて式 (1)〜(2) の各パラメータを決定する．すなわち，$W_1 \times C_{w1} = 0.62$，$W_1 \times C_{w2} = 0.85$，$W_2 \times C_{st} = 0.49$，$W_3 \times C_{sp1} = 0.40$，$W_3 \times C_{sp2} = 0.61$，$W_4 \times C_{l1} = 0.69$，$W_5 \times C_{e1} = 0.73$，$W_5 \times C_{e2} = 0.47$，$W_5 \times C_{e3} = 0.10$，$W_5 \times C_{e4} = 0.16$ となる．なお，今回利用したブログ記事上にトラックバックの受信日時が記載されているものはわずかであったため,「初トラックバックまでの時間，最新トラックバックまでの時間，平均トラックバック時間間隔」に関する情報は取得・利用できなかった．そのため，式 (2) において C_{sp3}，C_{sp4}，$C_{l2} = 0$ となる．

6.4 実験内容

まず，6.3 節で用いたものとは異なる映画タイトル（映画 A，B の 2 つ）をキーワードにして 6.3 節と同様にブログ記事検索を行い，検索結果として得られるブログ記事それぞれ 500 個のすべてを収集する．次に，収集した記事すべてに対して 6.3 節で求めたパラメータを適用した式 (1)〜(3) を用いて反響特性スコアを導出する.「広さ」に寄与している「コメント元数，トラックバック数」,「長さ」に寄与している「最新コメントまでの時間」がユーザ評価と高相関なので,「広さ：大，長さ：大」という反響特性を持つ記事ほどスコアが高くなる．

そして，20〜50 代の会社員 15 名に対し，既存手法（goo ブログ検索）による検索結果上位 5 件（文書適合度順）に含まれるブログ記事のセットと，反響特性スコア上位 5 件に含まれる個々の記事に対して，有用な記事としての満足度を 0〜3 点の 4 段階採点してもらう.「有用」の定義は 6.2 節，採点基準は 6.3 節におけるユーザ評価と同一である．この際，順序効果の影響を相殺するため，被験者を 2 組に分けセットを閲覧する順番が異なるようにする．

表 5 全被験者の満足度（N=15）
Table 5 Ratings by subjects (N=15).

	映画 A	映画 B
既存手法	1.09	0.98
提案手法	1.71	2.20

6.5 実験結果と考察

表 5 に実験結果を示す．表中の数値は，各手法が提示したブログ記事に対する満足度を，セット単位・全被験者で平均したものであり，どちらの映画においても提案手法が既存手法よりも高い満足度を被験者に与えていたことが分かる．

実際に既存手法が提示した記事を見ると,「今度この映画を観よう」程度の内容しかなくクチコミ情報としては有用でないものや，文章が乱雑で第三者には難解なものが多々見受けられた．これに対して，提案手法が提示した記事は大半が映画のストーリや感想，関連情報を丁寧に説明していた．このような記事では，多くの人がコメントを寄せる場合が多く，最大で 37 名が記事への共感や感想（「内容が役立った」,「私も面白いと思った」等）を語り合っているものがあった．また，長期間にわたってトラックバックを受けている場合も多く，3 カ月にわたって 84 件のトラックバック（スパムはなかった）を集め続け，他の記事から参照・引用されているものもあった．なお，この記事の作者は過去に 300 件以上の映画クチコミ記事を執筆しており，手馴れた文章で映画の内容や感想を上手に紹介しているという印象を受けた．

このように，提案手法では実際に多くの人が目を通し，コメントやトラックバックという関心を示す行為を行った記事を優先的に提示する．そのため，文書適合度が高いと機械的に判定されただけの記事が表示される既存手法よりも被験者が読んだときの満足度が高く，表 5 の結果に至ったと思われる．ここから，提案手法は「多くの情報が第三者にとって分かりやすく書かれている映画のレビュー記事を探す」というシーンにおいては，被験者が読んだときに既存手法よりも有用と感じられる記事を提示できたといえる．

7. 結　　論

本研究では，人間が実際に読んだときに有用と感じられるようなブログ記事を検索でき，その結果を多面的に選択できる技術の確立を目指して，反響特性を利用したブログ記事検索手法を提案した．

反響特性とは，記事を取り巻く書き手・読み手の行動や，彼らの行動の発端である記事を分析することで

導出される指標であり,「広さ，強さ，速さ，長さ」の多次元要素からなる．この指標を用いれば,「幅広い人から関心を集めている記事」や「長期間関心を集め続けている記事」というように，人間が実際に読んで有用と感じられる記事を多面的な観点から検索することが可能になる．

本稿では反響特性を導出する際にコメント・トラックバックに関する属性を中心に利用した．実際にブログ記事約 100 万件を収集してこれらの属性の実態調査を実施し，調査結果に基づいて反響特性スコアの導出アルゴリズムを決定した．評価実験では,「幅広い人から長期間関心を集めている」記事を提示するとユーザの満足度が高くなるという結果が得られた．ここから，特定シーンにおいてではあるが，提案手法はユーザが実際に読んだときに有用と感じられる記事を提示できることが確認できた．

今後の課題としては，反響特性導出の精度向上，適用可能シーンのさらなる検討があげられる．精度向上に関しては，今回利用したコメント・トラックバック属性の妥当性の検証，新たに利用する属性の検討が必要である．適用シーンの検討に関しては，今回評価実験を行った「多くの情報が第三者にとって分かりやすく書かれている映画のレビュー記事を探す」以外のシーンにおける提案手法の有効性評価，ユーザインタビュー等の実施によるニーズの把握が必須である．また，我々が現在開発しているブログコミュニティプロファイリング技術「XappaLinks」[14),15)] では頻繁に交流しあっているブロガ群をブログコミュニティとして抽出しているが，その交流場所である記事自体の性質を分析する際にも本提案の技術を適用する方針である．

謝辞 ブログサービス運営者の立場から，本研究に対し適切かつ貴重なアドバイスをくださった株式会社 NTT データの藤村剛氏に感謝の意を表します．

参 考 文 献

1) 総務省：ブログ及び SNS の登録者数 (2006). http://www.soumu.go.jp/s-news/2006/060413_2.html（2007/9/19 現在）.
2) 松岡寿延，岡野真一，宮田章裕，石打智美，荒川則泰，加藤泰久：ブログコミュニケーションにおけるユーザ意識調査報告，電子情報通信学会第 7 回 Web インテリジェンスとインタラクション研究会 (2006).
3) AOL, Inc.: Blog Trends Survey (2005). AOL Press Releases (2005/09/16).
4) goo ブログ. http://blog.goo.ne.jp（2006/8/10 現在）.
5) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proc. 7th International World Wide Web Conference* (1998).
6) Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment, *J. ACM*, Vol.46, No.5 (1999).
7) Fujimura, K., Inoue, T. and Sugisaki, M.: The EigenRumor Algorithm for Ranking Blogs, *WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem* (2005).
8) Mishne, G. and Glance, N.: Leave a Reply: An Analysis of Weblog Comments, *WWW 2006 Workshop on the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2006).
9) 井上智雄，小林哲郎，池田謙一，重野 寛，岡田謙一：ウェブ掲示板を対象としたネットワークコミュニティ分析システム：CMINER, 情報処理学会論文誌, Vol.45, No.1, pp.131–141 (2004).
10) Krishnamurthy, S.: The multidimensionality of blog conversations: The virtual enactment of september 11, *Internet Research 3.0* (2002).
11) Gumbrecht, M.: Blogs as protected space, *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2004).
12) Trevino, E.M.: Blogger motivations: Power, pull, and positive feedback, *Internet Research 6.0* (2005).
13) 宮田章裕，松岡寿延，岡野真一，石打智美，荒川則泰，加藤泰久：反響特性分析に基づいた Blog 記事マイニング，情報処理学会グループウェアとネットワークサービスワークショップ 2006 (2006).
14) 宮田章裕，松岡寿延，岡野真一，石打智美，荒川則泰，加藤泰久：XappaLinks: Blog コミュニティ参加支援システム，電子情報通信学会 2006 総合大会 (2006).
15) 岡野真一，松岡寿延，宮田章裕，石打智美，荒川則泰，加藤泰久：ブログコミュニケーション支援システム XappaLinks, 電子情報通信学会第 7 回 Web インテリジェンスとインタラクション研究会 (2006).

(平成 19 年 5 月 17 日受付)
(平成 19 年 9 月 3 日採録)

推 薦 文

本論文では，閲覧者行動に基づき Blog 記事の反響を広さ，強さ，速さ，長さという多側面から分析・提示している．行動情報の利用に新規性があり，記事の多面的マイニングはユーザニーズも高く実用的である．評価実験において有用性を実証している点も高く評価できる．よって，新規性，有用性のいずれにも優れて

いる論文と考えられ，推薦論文に値すると判断した．

（グループウェアとネットワークサービス研究会
主査　宗森　純）

宮田　章裕（正会員）

2005 年慶應義塾大学大学院理工学研究科修士課程修了．2006 年同大学院理工学研究科博士課程入学，現在に至る．また 2005 年日本電信電話株式会社入社，現在 NTT サイバーソリューション研究所．Web マイニング，グループウェア，ヒューマンインタラクション関連の研究に従事．DICOMO2004 最優秀プレゼンテーション賞，GN ワークショップ 2006 ベストペーパー賞，CollabTech2007 Best Paper Award，平成 19 年山下記念研究賞．

松岡　寿延

1994 年上智大学大学院物理学科修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスインテグレーション基盤研究所．

岡野　真一（正会員）

1998 年慶應義塾大学大学院理工学研究科修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスインテグレーション基盤研究所．

山田　節夫

1992 年東京電機大学大学院情報科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT サイバーソリューション研究所．

石打　智美

1990 年筑波大学大学院理工学研究科修了．同年日本電信電話株式会社入社．現在，NTT 知的財産センタ．

荒川　則泰

日本電信電話株式会社，NTT ネットワークサービスシステム研究所．

加藤　泰久（正会員）

1990 年京都大学大学院工学研究科応用システム科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，研究企画部門．