# ユーザインタラクション分析に基づくブログ記事の注目度推定

宮田　章裕† 　川島　晴美† 　藤村　考†

† 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　〒 239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †{miyata.akihiro,kawashima.harumi,fujimura.ko}@lab.ntt.co.jp

**あらまし**　本論文では，コメント・トラックバックといったユーザ同士のインタラクションの分析を行い，その分析結果に基づいてブログ記事がどの程度注目されているか推定する手法を提案する．ソーシャルブックマーク数はブログ記事の注目度を精度良く表す指標であるが，サービス利用者が少ないためソーシャルブックマークされているブログ記事が少ないという問題がある．そこで提案手法では，ソーシャルブックマークされているブログ記事のコメント・トラックバックを分析して統計モデルを構築し，ソーシャルブックマークがない記事に対しても注目度を推定することを目指す．評価実験では，提案手法が高い精度でブログ記事の注目度順位を推定できることを確認した．

**キーワード**　ブログ，コメント，トラックバック，ソーシャルブックマーク，注目度

# Popularity Estimation of A Blog Entry based on User Interaction Analysis

Akihiro MIYATA†, Harumi KAWASHIMA†, and Ko FUJIMURA†

† NTT Cyber Solutions Laboratories, NTT Corporation　1–1, Hikarinooka Yokosuka–Shi Kanagawa 239–0847 Japan
E-mail: †{miyata.akihiro,kawashima.harumi,fujimura.ko}@lab.ntt.co.jp

**Abstract**　We present a method of estimating the popularity of a blog entry based on an analysis of user interactions such as comments and trackbacks. Although the number of social bookmarks can be an indicator of popularity, social-bookmarked blog entries are very limited. To estimate the popularity of a blog entry that is not social-bookmarked, we analyze comment/trackback attributes of a blog entry that is social-bookmarked and construct a popularity estimation model. Using this model, the popularity of a blog entry that is not social-bookmarked but received comments/trackbacks can be estimated. The experiment results showed that our model is well suited for estimating the popularity of a blog entry.

**Key words**　Blog, Comment, Trackback, Social bookmark, popularity

## 1. はじめに

ブログ開設数は増加の一途を辿っており，2008 年 3 月時点で国内だけでも約 2000 万サイトが開設されているという報告がある．ブログ記事には，時事問題に対する非マスコミの立場からの考察や，商品に対する一般消費者の率直な好評・悪評などの貴重な情報が数多く埋もれており，情報源として決して無視できない存在になっている．しかし同時に，ブログ空間は玉石混淆と評されているように，個人の日記や備忘録のようなブログ記事も大量に存在している．そのため，第三者にとって有益な記事を精度良く検索する技術が切に求められている．

本論文ではその中でも特に，読者から注目を集めているブログ記事の発見に主眼を置いている．ところが，ブログ記事の注目度推定に利用できそうな指標は限定的である．アクセス数は外部から取得することが難しいし，被リンク数，ソーシャルブックマーク数はブログ空間においては非常にスパースである．

そこで我々は，ソーシャルブックマークされており，かつ，コメントまたはトラックバックがあるブログ記事集合を利用して注目度推定の統計モデルを構築する手法を提案する．このモデルを利用すれば，ソーシャルブックマークされていない記事に対しても，コメント・トラックバックがあれば注目度を精度良く推定することができると考えられる．

## 2. 既存のブログ記事の注目度推定手法

本研究では特に，読者から注目を集めているブログ記事の発見に主眼を置いている（3.1 節参照）．本章では，ブログ記事の注目度推定に関連する既存の指標・手法について述べる．

### 2.1 アクセス数

ブログ記事の注目度を推定するためのもっともシンプルな指標はアクセス数である．アクセス数が多いブログ記事は多くの読者から注目されているといえる．

しかし，一般にアクセス数やアクセスログは公開されておらず，この情報を外部から取得することは困難である．ブログサービスによってはアクセス数がブログ記事上に表示されるものもあるが，国内だけでも数十社のブログサービス業者が乱立している現状では，アクセス数を網羅的に取得することは事実上不可能である．

### 2.2 被リンク数

被リンク数もブログ記事の注目度を推定するための指標となりえる．例えば，PageRank [1] や HITS [2] で提案されているように，価値のある Web ページ（ブログ記事を含む）から多くのリンクを受けているブログ記事ほど注目度が高いと判定することができる．

しかし，大半のブログ記事には外部からのハイパーリンクが存在しないことが報告されている [3]．つまり，従来の Web ページ検索手法のように被リンク数が多い Web ページに高スコアを与えるような分析手法では，大半のブログ記事はスコアが 0 になってしまう．

### 2.3 ソーシャルブックマーク数

アクセス数，被リンク数以外にブログ記事の注目度を示す指標として，ソーシャルブックマーク数が注目されている．ソーシャルブックマークとは，不特定多数のユーザが互いのブックマークを Web 上に共有できる仕組みのことである．大勢から注目されているブログ記事はソーシャルブックマーク数が大きくなる傾向があるため，ソーシャルブックマーク数はブログ記事の注目度と捉えることもできる．山家らは，リンク構造分析によるランキング（PageRank）とソーシャルブックマーク数を利用したランキングを併用して検索結果を向上させる手法を提案している [4]．高橋らは，ソーシャルブックマークにおける Web ページを Authority，ユーザを Hub とみなし，HITS の概念を Web ページとユーザの関係に拡張し，Web ページを評価する手法を提案している [5]．

ただし，ソーシャルブックマークサービスの利用率は低く，自宅からインターネットを行っている 13 歳以上の男女の 7 %にとどまっている [6] という問題がある．

### 2.4 コメント数・トラックバック数

コメント・トラックバックもブログ記事の注目度を測定するための指標になりえる．注目されているブログ記事や洞察力があるブログ記事にはコメント数が多いという調査結果があり [7] [8]，物議を醸すような内容の記事では読み手・書き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 [7]．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．コメント・トラックバックは前述の被リンクやソーシャルブックマークほどスパースではなく，どのジャンルにおいてもコメント数またはトラックバック数が 1 以上の記事は比較的多いことが分かる（表 1 参照）．我々が収集した 313,799 件のブログ記事（4.2 節参照）においても，43.2 %にあたる 135,627 件の記事には 1 件以上のコメントまたはトラックバックが送信されていた．前述のアクセス数，被リンク数，ソーシャルブックマーク数と違い，コメント・トラックバックはブログ記事上に表示されているので，記事をクロールするだけで容易に情報を取得できるというメリットもある．

しかし，ブログ記事の注目度とコメント数・トラックバック数は必ずしも相関関係にあるとはいえない．たとえば，少数の常連メンバが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ことが報告されている [7]．常連メンバがチャットのようにコメントを利用して世間話をしている記事も散見される．このような記事が多くの第三者にとって有益でないのは自明である．

## 3. ユーザインタラクション分析に基づくブログ記事の注目度推定

### 3.1 研 究 目 標

本研究は，第三者にとって有益なブログ記事を検索する技術の確立を目指している．一般に Web ページ（ブログ記事を含む）検索サービスは，ページそのものを分析する技術と，ページに対する外部からの情報を分析する技術から成っている．前者は BM25 などに基づいて検索クエリに対するページの文書適合度を算出する手法が代表的である．後者はアクセス数や被リンク数などを分析してページの注目度を分析する手法が主流であり，本研究では後者に主眼を置いている．

現状でも，ブログ記事の注目度測定に関連する指標はいくつかある（2.1〜2.4 節参照）．しかし，アクセス数は外部から情報を取得することが困難であるし，被リンク数，ソーシャルブックマーク数はブログ記事においてはスパースであるという問題がある．コメント数・トラックバック数は比較的多くの記事に送信されており，記事上に表示されているので外部から情報を取得できるが，必ずしもブログ記事の注目度と相関関係にはないという問題がある．

上記の問題点をふまえ，本論文では**多くのブログ記事の注目度を精度良く推定できるアルゴリズムの確立**を研究目標として設定する．

### 3.2 提 案 手 法

我々は，”記事の読者と作者がコメントで意見を交わし合う，””記事を投稿する際に参考にした記事にトラックバックを送信する”などの，ブログ記事に起因するユーザインタラクションを多面的に分析することでその記事の価値を測定する**反響特性分析**という手法を提案してきた [9]．表 2 に示すのは，反響特性分析で利用するコメント・トラックバックの属性である（以降「反響素性」とする(注1)）．

コメント・トラックバックは比較的多くのブログ記事から取得できるので，記事に対する注目度と反響素性の関係をモデル

(注1)：表 2 中の< 15 >〜< 18 >は記事本文の属性であるが，これは記事本文とコメント・トラックバックの関係 (”記事の長さが短いのにコメントを送信した人数が多い”等) も重要と考えているため利用している．

表 1　ブログ記事検索結果上位 50 件の中で各条件を満たす記事の件数
Table 1　Blog entries of top 50 search results that satisfy each condition.

| ジャンル | 検索語 | 外部からの被リンク数が 1 以上 | SBM 数が 1 以上 | コメント数またはトラックバック数が 1 以上 |
|---|---|---|---|---|
| 映画 | 容疑者 X の献身 | 1 件 | 0 件 | 8 件 |
| | グーグーだって猫である | 2 件 | 0 件 | 11 件 |
| | 不都合な真実 | 2 件 | 0 件 | 6 件 |
| IT | iPhone | 7 件 | 0 件 | 3 件 |
| | Macbook | 0 件 | 1 件 | 5 件 |
| | ネットブック | 3 件 | 0 件 | 8 件 |
| 政治・経済 | 汚染米 | 1 件 | 0 件 | 5 件 |
| | 北京オリンピック | 5 件 | 0 件 | 5 件 |
| | ノーベル賞 | 0 件 | 0 件 | 6 件 |
| | 内閣支持率 | 1 件 | 0 件 | 8 件 |
| | 金融危機 | 1 件 | 1 件 | 7 件 |
| | 解散総選挙 | 3 件 | 0 件 | 8 件 |

SBM ＝ソーシャルブックマーク
検索は 2008 年 11 月 26 日に goo ブログ検索（http://blog.goo.ne.jp）の「適合度順・goo ブログのみ」オプションを用いて行った．

化できれば，より多くのブログ記事の注目度を精度良く推定できると思われる．そこで我々が注目するのはソーシャルブックマーク数である．ソーシャルブックマークされているブログ記事はスパースであるが，2.3 節にて前述のとおり，ソーシャルブックマーク数は記事に対する注目度を精度良く表しているといえる．

上記の点をふまえ，ソーシャルブックマークされており，かつ，コメントまたはトラックバックがあるブログ記事集合を利用して，記事に対する注目度を反響素性で表現する統計モデルを構築する手法を提案する．このモデルを利用すれば，ソーシャルブックマークされていない記事に対しても，コメント・トラックバックがあれば注目度を精度良く推定することができると考えられる．

### 3.3　統計モデルの構築

本論文では，データセット（4.2 節参照）ごとに統計モデルを構築することとし，モデル構築には回帰式を用いる．目的変数は，国内の大手ソーシャルブックマークサービスはてなブックマーク(注2)におけるソーシャルブックマーク数を式 (1) のようにスコア化したものを利用する．$SBMCount_i$ はブログ記事 $i$ のソーシャルブックマーク数，$round$ は小数点以下を四捨五入する関数であり，ブログ記事 $i$ のソーシャルブックマーク数をスコア化したものを $SBMScore_i$ とする．

$$SBMScore_i = round(SBMCount_i/10) \quad (1)$$

説明変数はブログ記事 $i$ が持つ反響素性を利用すべきであるが，反響素性＜ 1 ＞～＜ 18 ＞の中には素性間に相関がみられるものがあるため(注3)，反響素性をそのまま説明変数に利用すると多重共線性が生じる可能性がある．そこで，事前に反響素性に対して主成分分析を行い，得られた主成分を説明変数に利用する．主成分分析を行うために，まず式 (2) のように反響素性のスコア化（平均 0，標準偏差 1 に標準化）を行う．$J$ は反響素性の数 18，$x_{ij}$ はブログ記事 $i$ の反響組成＜ j ＞，$\mu$，$\sigma$ はそれぞれデータセット中における $x_{ij}$ の平均値，標準偏差であり，$x_{ij}$ をスコア化したものを $y_{ij}$ とする．

$$y_{ij} = (x_{ij} - \mu)/\sigma \quad (j = 1, 2, ..., J) \quad (2)$$

次に，$y_{ij}(j = 1, 2, ..., J)$ を用いて主成分分析を行う(注4)．$pc_{ik}$

表 2　反 響 素 性
Table 2　Response attributes.

| | |
|---|---|
| ＜ 1 ＞ CM 送信者数 | CM を送信した人数 |
| ＜ 2 ＞平均 CM 送信数 | CM 送信者 1 人あたりの平均 CM 送信数 |
| ＜ 3 ＞ CM リンク率 | CM 総数における，CM 送信者の URL が書いてある CM 数の割合 |
| ＜ 4 ＞ CM タイトル空欄率 | CM 総数における，CM タイトルが空欄の CM 数の割合 |
| ＜ 5 ＞ CM 送信者空欄率 | CM 総数における，CM 送信者名が空欄の CM 数の割合 |
| ＜ 6 ＞平均 CM 文字数 | CM に含まれる平均文字数 (絵文字は除く) |
| ＜ 7 ＞平均 CM 絵文字数 | CM に含まれる平均絵文字数 |
| ＜ 8 ＞平均 CM 内リンク数 | CM に含まれるハイパーリンク数 |
| ＜ 9 ＞記事・CM 類似度 | 記事と CM 集合の文書類似度 |
| ＜ 10 ＞初 CM 受信経過時間 | 記事が投稿されてから最初の CM を受信するまでに経過した時間 |
| ＜ 11 ＞最終 CM 受信経過時間 | 記事が投稿されてから最終の CM を受信するまでに経過した時間 |
| ＜ 12 ＞平均 CM 間隔 | CM 受信時間間隔の平均値 |
| ＜ 13 ＞ TB 送信者数 | TB を送信した人数 |
| ＜ 14 ＞最終 TB 受信経過時間 | 記事が投稿されてから最終の TB を受信するまでに経過した時間 |
| ＜ 15 ＞記事文字数 | 記事に含まれる文字数 (絵文字は除く) |
| ＜ 16 ＞記事絵文字数 | 記事に含まれる絵文字数 |
| ＜ 17 ＞記事画像数 | 記事に含まれる画像数 (絵文字は除く) |
| ＜ 18 ＞記事リンク数 | 記事に含まれるハイパーリンク数 |

CM=コメント，TB ＝トラックバック

(注2)：http://b.hatena.ne.jp
(注3)：たとえば 4.2 節で後述の D0 では，反響素性＜ 5 ＞と＜ 6 ＞の Pearson 相関係数が約 0.62 であった．
(注4)：主成分分析には MacOS 版 R2.8.0 の prcomp 関数を利用．

を第 $k$ 主成分とすると，これは式 (3) のように表すことができる．$w_{ijk}$ は結合係数，$K$ は主成分数である．$K$ は通常，累積寄与率が 80 %を超える程度に大きい値であれば十分であるが，後述の評価実験では精度検証が目的であるため，本論文では $K = J = 18$ とする．

$$pc_{ik} = \sum_{j=1}^{J} w_{ijk} y_{ij} \qquad (k = 1, 2, ..., K \leqq J) \tag{3}$$

最後に，$SBMScore_i$ を目的変数，$pc_{ik} \quad (k = 1, 2, ..., J)$ を説明変数として回帰分析を行う(注5)．回帰式によるブログ記事 $i$ の $SBMScore_i$ の推定値を $EstScore_i$ とすると，回帰式は式 (4)〜(6) のようになる．$c$ は定数項，$f(pc_i)$ は 1 次の項，$g(pc_i)$ は 2 次の項，$v_k$，$v_{kk'}$ は各項の偏回帰係数である．

$$EstScore_i = c + f(pc_i) + g(pc_i) \tag{4}$$

$$f(pc_i) = \sum_{k=1}^{J} v_k pc_{ik} \tag{5}$$

$$g(pc_i) = \sum_{k=1}^{J-1} \sum_{k'=k+1}^{J} v_{kk'} pc_{ik} pc_{ik'} \tag{6}$$

## 4. 評価実験

### 4.1 実験の目的

この実験の目的は，提案手法がどの程度の精度でブログ記事の注目度を推定できるか評価することである．

### 4.2 データセット

データセットは実際のブログ記事をクロールして作成した．その際，記事のトピックにできるだけ偏りが生じないよう配慮した．具体的には，2008 年 9 月 21 日〜11 月 18 日の期間に人気検索語(注6)，時事トピック（北京オリンピック，汚染米など），IT 関連用語（OS 名，プログラム言語名など），家電商品名，映画タイトル，ミュージシャン名，書籍名，タレント名，スポーツ選手名などでブログ記事検索(注7)した結果の上位最大 300 件をクロールした．さらに，収集したブログ記事へのトラックバック送信元の記事，およびコメント送信者の署名 URL がブログ URL である場合はそのブログに含まれる最新記事を最大 3 件，最大 3 段階で再帰クロールした．このようにして収集した 313,799 件のブログ記事からなるデータセットを D0 とする．

次に，D0 の中でソーシャルブックマーク数が 1 以上の 3,874 件からなるデータセットを D1 とする．また，D1 の中で，記事本文もしくはソーシャルブックマークタグ(注8)に特定の IT 関連用語（IT，Internet，インターネット，ネット，Web，ウェブ，Computer，コンピュータ，PC，パソコン，Mac，マック，Windows，ウィンドウズ，Linux，リナックス，Ajax，OSS，オープンソース）が 1 つ以上含まれる 1,360 件を抽出して D4 とした．D1 の中で D4 に含まれなかった記事の内容を確認すると，時事問題，ライフハック，料理レシピに関するものが目立った．

ここで，提案手法は主にコメントの属性を分析するので（反響素性 18 個中 12 個がコメントに関する属性），D1/D4 の中からコメントが 1 つ以上ある 2,736/1,052 件を抽出して D2/D5 とした．表 3 にあるとおり，D2/D5 はソーシャルブックマーク数の平均値が D1/D4 よりやや高いデータセットとなっている．

また，D2/D5 の中には未来の日時で投稿されている記事や，不正な文字列を含んでいて正しく分析できない記事などが含まれていた．そこで，データセットの整合性を保つために D2/D5 からこれらを取り除いた 1,700/642 件を D3/D6 とした．なお，D2 と D3 の間，および，D5 と D6 の間には標本としての有意差は見られなかった(注9)．以降，評価実験には D3 と D6 を用いる．

表 3　データセット情報
Table 3　Dataset properties.

| データセット名 | D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|---|
| | （ジャンル：指定なし） | | | （ジャンル：IT 関連） | | |
| 記事数 | 3874 | 2736 | 1700 | 1360 | 1052 | 642 |
| SBM 数の最小値 | 1 | 1 | 1 | 1 | 1 | 1 |
| SBM 数の最大値 | 457 | 457 | 457 | 457 | 457 | 457 |
| SBM 数の平均値 | 14.85 | 19.86 | 20.2 | 26.04 | 32.66 | 34.34 |
| SBM 数の標準偏差 | 32.81 | 37.73 | 39.95 | 45.36 | 49.51 | 54.05 |

SBM=ソーシャルブックマーク

### 4.3 実験手順

実験は，ブログ記事のジャンルを指定しない場合（D3 を使用）と，ジャンルを IT 関連に限定する場合（D6 を使用）の 2 通り行う．

まず，式 (1) に基づいてデータセット中の全ブログ記事の $SBMScore$ を算出する．

次に，提案手法の統計モデルを利用して $EstScore$ を算出する 10-fold 交差検定を行う．具体的には，データセットをランダムに 10 組に等分し，9 組を統合して式 (2)〜(6) に基づいて統計モデルを構築する．そして，残りの 1 組の記事群の反響素性をそのモデルに当てはめて，各記事の $EstScore$ を算出するという作業を行う．10 組の中でモデル構築に利用する 9 組を入れ替えながら，この作業を ${}_{10}C_9 = 10$ 回繰り返す．

### 4.4 実験結果

提案手法を注目度順に検索できるブログ記事検索サービスに適用することを視野に入れ，性能評価は注目度の順位推定精度という観点から行った．具体的には，Kendall の順位相関係数

(注5)：回帰分析には同じく R の stepAIC 関数を利用．
(注6)：http://ranking.goo.ne.jp/keyword
(注7)：http://blog.goo.ne.jp,「適合度順・goo ブログのみ」オプション利用．
(注8)：はてなブックマークではユーザが Web ページをブックマークする際に「タグ」とよばれるキーワードを付与することができる．ここでは全ユーザが付与したタグを統合して，その中に特定の IT 関連用語が含まれているかどうか判定した．
(注9)：Kolmogorov-Smirnov 検定を行ったところ，D2，D3，D5，D6 は正規分布でないことが分かった．そこで D2 と D3，および，D5 と D6 の各ペアについて Wilcoxon の順位和検定を行ったところ，p 値はそれぞれ 0.641，0.640 となり，有意差がないことが分かった．

$\tau$ 値と Spearman の順位相関係数 $\rho$ 値(注10)を用いて，分析対象のブログ記事群内における各記事の注目度順位（$EstScore_i$ が大きいほど順位が上位）をどの程度正しく推定できているか評価した．

また，ベースライン手法として，コメントを多く集めているブログ記事ほど注目度が高いと推定する手法を採用した(注11)．この手法では，コメント数が多い記事ほど注目度順位が上位になる．

表 4 に示すのが，ジャンルを指定しない場合の各手法の $\tau$ 値および $\rho$ 値である．$\tau$ 値について全交差検定の平均値を見ると，提案手法（0.414）がベースライン手法（0.203）を上回る結果を示していることが分かる．各手法の交差検定 ID1〜10 の結果群の間で Welch の t 検定(注12)を行うと $p = 3.46E-5$ となり，1 %水準で有意差がみとめられた．また，$\rho$ 値については提案手法（0.444）がベースライン手法（-0.033）を大幅に上回っており，Student の t 検定(注13)でも 1 %水準の有意差がみとめられた（$p = 1.45E-11$）．

表 5 に示すのが，ジャンルを IT 関連に限定する場合の各手法の $\tau$ 値および $\rho$ 値である．$\tau$ 値について全交差検定の平均値を見ると，こちらの場合も提案手法（0.448）がベースライン手法（0.218）を上回る結果を示していることが分かる．各手法の結果群の間で Student の t 検定(注13)を行うと $p = 1.60E-5$ となり，1 %水準で有意差がみとめられた．$\rho$ 値についても同様に提案手法（0.529）がベースライン手法（0.333）より優位であり，Student の t 検定(注13)にて 1 %水準の有意差がみとめられた（$p = 5.54E-5$）．

表 4　各手法の $\tau$ 値・$\rho$ 値（ジャンル：指定なし）
Table 4　Evaluation results (All topics).

| 交差検定 ID | $\tau$ 値 | | $\rho$ 値 | |
|---|---|---|---|---|
| | 提案手法 | BL 手法 | 提案手法 | BL 手法 |
| 1 | 0.449 | 0.094 | 0.487 | -0.154 |
| 2 | 0.423 | 0.271 | 0.473 | -0.001 |
| 3 | 0.372 | 0.299 | 0.392 | 0.122 |
| 4 | 0.400 | 0.328 | 0.422 | -0.042 |
| 5 | 0.407 | 0.125 | 0.420 | -0.068 |
| 6 | 0.362 | 0.179 | 0.360 | -0.041 |
| 7 | 0.447 | 0.217 | 0.509 | 0.096 |
| 8 | 0.419 | 0.045 | 0.439 | -0.146 |
| 9 | 0.394 | 0.292 | 0.437 | -0.065 |
| 10 | 0.469 | 0.180 | 0.501 | -0.037 |
| 平均 | 0.414 | 0.203 | 0.444 | -0.033 |

BL=ベースライン

表 5　各手法の $\tau$ 値・$\rho$ 値（ジャンル：IT 関連）
Table 5　Evaluation results (IT topics).

| 交差検定 ID | $\tau$ 値 | | $\rho$ 値 | |
|---|---|---|---|---|
| | 提案手法 | BL 手法 | 提案手法 | BL 手法 |
| 1 | 0.453 | 0.191 | 0.556 | 0.303 |
| 2 | 0.459 | 0.234 | 0.567 | 0.318 |
| 3 | 0.407 | 0.290 | 0.482 | 0.420 |
| 4 | 0.357 | 0.216 | 0.469 | 0.359 |
| 5 | 0.411 | 0.246 | 0.457 | 0.336 |
| 6 | 0.386 | 0.062 | 0.406 | 0.188 |
| 7 | 0.430 | 0.238 | 0.560 | 0.360 |
| 8 | 0.511 | 0.069 | 0.576 | 0.252 |
| 9 | 0.597 | 0.427 | 0.682 | 0.509 |
| 10 | 0.475 | 0.210 | 0.533 | 0.286 |
| 平均 | 0.448 | 0.218 | 0.529 | 0.333 |

BL=ベースライン

## 5. 考　　察

提案手法がベースライン手法を上回る精度で注目度順位を推定できることが評価実験の結果から確認できた（4.4 節参照）．また，多くのブログ記事から取得できるコメント・トラックバックという指標を利用する提案手法が，外部からの取得が難しい指標（アクセス数）やブログ空間においてスパースな指標（被リンク数・ソーシャルブックマーク数）を利用する既存手法よりも多くのブログ記事を分析できることは自明である．よって，提案手法は多くのブログ記事の注目度を精度良く推定するためのアプローチとして適していると考えられる．現在我々は，ソーシャルブックマークされていない記事に対しても提案手法が同様の精度を発揮できるかどうか検証するための準備を進めている．

また，ジャンルを指定しない場合，IT 関連に限定する場合のどちらにおいても提案手法が有効であることも確認できた．ここから，提案手法はある程度ジャンル非依存に性能を発揮できると考えられる．ただし，これはソーシャルブックマークが 1 件以上あるブログ記事からなるデータセットを用いた場合の実験結果であることに留意しなければならない．つまり，IT 関連，時事問題，ライフハック，料理レシピなどのようにソーシャルブックマークされやすいジャンルの記事には提案手法は有用であるが，他のジャンルにおいても同様の結果が得られるかどうか判断するためにはさらなる検証実験が必要である．

ここで，スパムに関しては検討の余地が残っているといえる．ブログ記事に対して，公序良俗に反する内容のコメントが送信されたり，記事の内容とは全く関係がない営利目的サイトからトラックバックが送信されたりするスパム行為が発生することがある．ただし，統計モデル構築に利用するブログ記事に関しては，ソーシャルブックマーク数が一定以上のものであればスパムの問題は少ないと思われる．なぜならば，一定数以上のユーザからソーシャルブックマークされるような記事を書くブログオーナー（そのブログの作成者）は自身の記事をよくメンテナンスしている場合が多く，その記事にスパムコメント・ト

(注10)：推定した順位データが実際の順位データを忠実に再現できているほど，$\tau$ 値および $\rho$ 値は大きくなる．
(注11)：ブログ記事の注目度を推定する一般的な指標としては被リンク数も考えられるが，データセット中の大半の記事の被リンク数は 0 であったため，ここでは被リンク数ではなくコメント数を利用した．
(注12)：Kolmogorov-Smirnov 検定の結果 2 群とも正規分布であり，F 検定の結果 2 群は等分散でなかったため，Weltch の t 検定を採用した．
(注13)：Kolmogorov-Smirnov 検定の結果 2 群とも正規分布であり，F 検定の結果 2 群は等分散であったため，Student の t 検定を採用した．

ラックバックが送信されてもブログオーナーがすぐに削除すると思われるからである．

一方，構築された統計モデルを用いて注目度を推定する対象となる記事までもソーシャルブックマーク数が一定以上のものだけに限定することは，「多くのブログ記事の注目度を推定する」という目標（3.1 節参照）に反するので望ましくない．しかし，対象を限定しない場合，ブログオーナーが頻繁にスパムコメント・トラックバックを削除しているとも限らず，スパムが推定精度に悪影響を及ぼすおそれがある．スパム除去技術の確立は本研究の直接の目標ではないため，この問題の根本的解決は各ブログホスティング業者のスパム除去への取り組みに期待するところではあるが，いくつかアイデアはある．そのうちの 1 つとして，ブログ記事に送信されたコメント群の中からブログオーナーによって送信されたものを特定し [10]，ブログオーナーによる一番最後のコメントよりも前に送信されたコメントのみを分析に用いる方法がある．ブログオーナーが自分のブログ記事にコメントをする際には，それ以前に記事に送信されているコメントを読んだ上でコメントを書くはずであり，その記事にスパムコメントが送信されていれば気付いて削除すると思われる．そのため，ブログオーナーによる一番最後のコメントよりも前に送信されたコメントの中にはスパムコメントが含まれている可能性が低く，これらのコメントのみを用いればスパムコメントによる悪影響が軽減できると思われる．

我々がこれまでに行ってきた研究 [9] [11] との差分としては，まず第一に，モデル構築にソーシャルブックマークを利用した点である．ブログ記事の価値を推定するモデルを構築するためには記事に対する価値判断の正解データが必要で，これまで我々は人が実際に記事に目を通して価値判断を行う方法で正解データを作成してきた．この方法は質の高い正解データが作成できる反面，非常に作成コストがかかるという欠点がある．その点，(1) のように Web 上に既に存在するソーシャルブックマークという人間による価値判断結果の指標を活用すれば，低コストで正解データ作成，モデル構築が可能になる．ただし，ソーシャルブックマークされるブログ記事のジャンルは IT 関連などに偏りがちなので，ソーシャルブックマーク数を利用して構築したモデルで他のジャンルの記事の注目度を推定する際はアルゴリズムを改善する必要があるかもしれない．

第二の差分は，回帰式でモデルを表現した点である．以前，我々はモデル構築に Support Vector Machine（SVM）を利用していた [11]．しかし，SVM は 2 値判定問題などであれば比較的高速にモデルを構築できるが，今回の問題のようにスコアや順位を推定する問題ではモデル構築時間が膨大になってしまう場合が少なくない．その点，回帰分析は SVM よりもモデル構築に要する時間が大幅に短く，モデルの表現も比較的シンプルである．これは，提案手法を実サービスとして実現する際に大きなメリットとなる．今後はモデル構築に重要，あるいは不要な反響素性をつきとめ，推定精度のさらなる向上を図る方針である．

## 6. おわりに

本論文では，多くのブログ記事の注目度を精度良く推定できるアルゴリズムの確立を研究目標として設定した．提案手法は，注目度推定の統計モデルを構築する際に，ソーシャルブックマークされており，かつ，コメントまたはトラックバックがあるブログ記事集合を利用するアプローチを採った．ブログ記事の注目度順位を推定する評価実験では，提案手法がベースライン手法を上回る精度を示し，その有用性が確認できた．

提案手法は将来的に，多くの読者から注目を集めているブログ記事の検索サービスに適用可能と思われる．第三者にとって有益な記事を精度良く検索することが難しい現状をふまえると，このようなサービスには新規性および大きな需要があると考えている．

## 文　献


[1] S. Brin and L. Pagee: “The anatomy of a large-scale hypertextual web search engine”, Proceedings of the seventh international conference on World Wide Web (1998).

[2] J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, Journal of the ACM, **46**, 5, pp. 604–632 (1999).

[3] K. Fujimura, T. Inoue and M. Sugisaki: “The eigenrumor algorithm for ranking blogs”, Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem at the 14th International World Wide Web Conference (2005).

[4] 山家雄介, 中村聡史, アダムヤトフト, 田中克己：“ソーシャルブックマークの特性分析とそれに基づく web 検索結果の再ランキング手法”, 情報処理学会論文誌 データベース, **1**, 1, pp. 88–100 (2008).

[5] 高橋翼, 北川博之：“ソーシャルブックマークを利用したユーザ嗜好に基づくページの評価”, データ工学ワークショップ 2008 (2008).

[6] 財団法人インターネット協会：“インターネット白書 2008”, インプレス R & D (2008).

[7] G. Mishne and N. Glance: “Leave a reply: An analysis of weblog comments”, Proceedings of 3rd Annual Workshop on the Weblogging Ecosystem at the 15th International World Wide Web Conference (2006).

[8] K. Sandeep: “The multidimensionality of blog conversations: The virtual enactment of september 11”, Proceedings of AoIR Internet Research 3.0 (2002).

[9] 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：“反響特性分析を利用したブログ記事検索手法”, 情報処理学会論文誌, **48**, 12, pp. 4041–4050 (2007).

[10] R. Hui, A. Miyata, H. Kawashima and H. Okuda: “Blog owner detection: User reference matrix”, 情報処理学会第 70 回全国大会 (2008).

[11] A. Miyata, H. Kawashima and H. Okuda: “A communication-based approach for detecting influential blog entries”, Proceedings of IADIS International Conference WWW/Internet 2008 (2008).