# トラックバックネットワークに基づくSEOコミュニティの分析

風間一洋[†] 佐藤進也[†]
斉藤和巳[††] 木村昌弘[†††]

本稿では，ブログのトラックバックネットワークからサーチエンジン最適化（SEO）コンテストに参加しているSEOコミュニティを抽出し，SEOの観点から次数，近傍平均次数，クラスタ係数などのネットワークの基本統計量について分析する．さらに，これらの基本統計量に加えて，隣接行列の主成分ベクトルと，ランダムウォークによって求められる定常確率分布を用いてブログエントリをランキングし，F値と精度を求めて性能を評価する．その結果，トラックバックネットワークの構造の特徴に着目すればSEOコミュニティを分離できることと，特に主成分ベクトルを用いてブログエントリをランキングする方法が，SEOコミュニティの検出性能に優れていることを示す．

# Analyzing SEO Communities Using Trackback Networks

Kazuhiro Kazama,[†] Shin-ya Sato,[†] Kazumi Saito[††]
and Masahiro Kimura[†††]

In this paper, we analyze fundamental metrics on trackback network structures of SEO (search engine optimization) communities, which were made by participants of a SEO content. Furthermore, we rank blog entries using the fundamental metrics, the principal eigenvector of an adjacency matrix and the stationary probability distribution of the resulting random walk, and we qualify the performance of those methods in terms of F-measure and precision. The result shows that SEO communities can be extracted using the characteristic of trackback network structures and the method of ranking blog entries based on its principal eigenvector can be a very promising approach for extracting SEO communities.

## 1. はじめに

最近，Web上で閲覧できる個人的な日誌（log）の一種である**ブログ**（blog，weblogの省略形）が急速に普及しつつある．ブログを書く人間をブロガ（blogger），互いにリンクでつながれたブログの集合をブログ空間（blogspace）と呼ぶ．ブログが普通のWebページと異なるのは，技術的な知識をほとんど持っていない人でも簡単に更新・維持できるサービス（例，Blogger）やプログラム（例，Movable Type）を用いることと，**トラックバック**と呼ばれる別のブログにリンクを張ったことを相手に通知すると同時に通知元への逆リンクが自動的に掲載される機構が提供されることである．つまり，情報の流れに対して通常のリンクは逆方向だが，トラックバックは順方向であり，情報や議論の流れを効率的に追うことができる．また，通常のリンクでは，一般にWebページ単位でリンクし，リンク先の内容が更新されるのに対して，ブログのトラックバックでは，更新されても変化しない**パーマリンク**（Permalink）を用いてエントリ単位にリンクし，情報更新時には新しいエントリが追加されるだけで，トラックバックされている既存のエントリの内容は変化しない．つまり，通常のリンクの場合は時間の経過につれて単純に被リンク数が増加していくのに対して，ブログのトラックバックはたいていは新しいエントリに行われることから，ネットワークの各部が時間を反映する傾向がある．これらの特徴から，インターネット上の情報の公開や情報交換がさらに加速され，たとえば既存のメディアを介さずに重要な情報が広がったり，時期や話題ごとにコミュニティを形成したりする現象が起こっている．

ただし，リンク解析の観点から見ると，一般的なWeb情報と比較するとブログ情報は扱いが難しく，特にGoogleなどのリンク解析技術を用いたサーチエン

† NTT未来ねっと研究所
NTT Network Innovation Laboratories
†† NTTコミュニケーション科学基礎研究所
NTT Communication Science Laboratories
††† 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University

ジンのランキングに問題が発生している．たとえば，ブログは互いにリンクを張り合うためにサーチエンジンの検索結果で高い順位になる傾向があり，普通の Web ページで公開された情報との相対的なバランスが崩れやすい．また，最近のサーチエンジンの有用性は，Web ページの入次数が重要度に関係があるという点に着目した Link Popularity [1]，アンカーテキスト[2]，PageRank [3] などの評価指標によって確保されている．しかし，トラックバックを使えば他のブログから簡単に多くのリンクを張ることができる点を悪用して，意図的にリンクが密なブログ群を構築することでサーチエンジンの検索結果の順位を上げて自分のブログの閲覧数を増やそうとする SEO スパムが頻繁に行われ，その結果サーチエンジンの有用性が損なわれる問題が深刻化している．

しかし，現時点の対策は，ブログサービス提供者や開発者にコメント中のリンクやトラックバックに rel=“nofollow” 属性を指定するようにプログラムを変更してもらうことで，Web ロボットの収集やリンク解析の対象から除外する消極的な提案が行われている程度である[4]．

本稿では，SEO スパムを検出し，さらに SEO スパムに対して頑健なリンク解析手法を確立する目的で，SEO コンテストの参加者コミュニティに着目し，ブログ空間でトラックバックを用いた SEO の現状を分析する．まず，ブログのトラックバックネットワークからサーチエンジン最適化（SEO）コンテストに参加している SEO コミュニティを特定のキーワードやバナーに着目して抽出し，複雑ネットワーク理論の研究でよく行われている次数，近傍平均次数，クラスタ係数などのネットワークの基本統計量について，特に SEO の観点から分析する．さらに，これらの基本統計量に加えて主成分ベクトルと定常確率をトラックバックネットワーク構造を反映する特徴量として用いて再度 SEO コミュニティを分離して，情報検索の性能評価などで広く使われている F 値と精度を求めて，性能を評価する．その結果，トラックバックネットワークの構造に着目すれば SEO コミュニティを分離できることと，特に主成分ベクトル法を用いた場合の検出性能が優れていることを示し，その理由を考察する．

## 2. SEO の観点からのトラックバックネットワークの解析

### 2.1 SEO と SEO スパム

**SEO**（Search Engine Optimization）は，自分の Web サイトを閲覧する人を増やすために，サーチエンジンの検索結果の順位をより上位にする方法である[1]．SEO の本来の目的は，ユーザが見やすく，Web ロボットが収集しやすい Web ページを作ることであった．

しかし，Web の普及やアフィリエイト広告の登場にともない，不正な手段で順位を操作することが行われはじめた．これを **SEO スパム**と呼び，単語を操作する単語スパム（term spamming）と，リンク関係を操作するリンクスパム（link spamming）の 2 種類に大きく分類される[5]．前者は，比較的初期から行われていたスパムで，Web ページの内容とは無関係な単語をユーザに見えない色，大きさなどで埋め込むことで，無関係な単語でも検索できるようにしたり，単語の出現頻度から求める TF-IDF（Term Frequency-Inverse Document Frequency）のようなスコアリング手法で高い値を得たりする方法である．ただし，これらの不正手段の判別は比較的容易であり，現在ではサーチエンジン側でも対策が施されており，あまり効果的とはいえない．

後者は，Google のような Web ページの被参照関係を解析して重要度を判定するサーチエンジンの検索結果の順位を上げるために，ある Web ページに対して大量の内容や重要度とは無関係なリンクを張る方法である．ただし，一般に多くの他人にリンクを張ってもらうのは困難なので，自分で大量の Web サイトを管理したり，不特定多数の Web サイト管理者を集団化したりして，多くの Web ページの間で相互リンクすることになる．このような特殊なリンク構造を持つ Web ページ群を**リンクファーム**（link farm）と呼ぶ．最近はわざわざ自分でサーバを用意しなくても，既存のブログホスティングサービスを利用してブログを複数開設し，相互にトラックバックを張ることが容易にできる．

なお，SEO スパムがサーチエンジン運営側に発見された場合には，検索結果から除外されるなどの大きな危険がともなうので，Web ロボットと閲覧者に異なるページを見せるクローキング（cloaking）や，実際に SEO スパムを行っているページから目的のページに転送するリダイレクト（redirect）などの隠蔽工作も行われる[1),5]．特に，該当 Web ページを見ればすぐ分かる単語スパムと異なり，リンクスパムの証拠はブログ空間に広く分散しているために，ある Web ページが本当に SEO スパムを行っているかどうかを知るのは困難である．

### 2.2 SEO コミュニティ

そこで，本稿では**SEO コンテスト**に着目した．SEO コンテストとは，世の中でまったく使われていない，

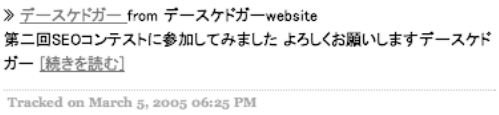


**図 1** トラックバックによる SEO スパムの例
Fig. 1 Example of SEO spam using trackbacks.

サーチエンジンでも検索されない特殊な造語に対して Web ページを作成し，Google などのサーチエンジンの検索結果の順位を競い合うコンテストである．つまり，その単語に注目すれば，SEO コンテストに参加しているブログを比較的容易に抽出できることになる．

本稿では，日本で 2004 年 10 月 17 日から 2004 年 12 月 20 日に「ゴッゴル」というキーワードで開催された第 1 回 SEO コンテストと，2005 年 1 月 21 日から 2005 年 4 月 9 日までの期間に「デースケドガー」というキーワードで開催された第 2 回 SEO コンテストを対象とした．このコンテストは特にリンクスパム技術を競うのが目的であり，使用される技術は SEO スパムの場合と変わらない．たとえば，コンテストに参加する，またはリンクファームになってよい Web ページ・ブログは,「トラックバック OK　ゴッゴル」または「トラックバック OK　デースケドガー」というバナーを張る必要があるが，これを目印にコンテスト参加者は他の参加者のブログエントリに対してトラックバックを張り，相互リンクを作成する．本稿では，これらのコンテスト参加者のブログエントリ群を，**SEO コミュニティ**と呼ぶ．

SEO コンテストのトラックバックを使った SEO の例を**図 1** に示す．このようなトラックバックを，数多くのブログエントリに対して行っている．コンテストの主な評価対象になっている Google[6] では，このようなトラックバックを数多く張られると PageRank の値に影響する．さらに，リンクのアンカーテキストも Web ページの内容と同様に索引に含まれることから，アンカーテキストになる可能性のあるテキスト（ただし，プログラムによって異なる）にキーワードを含めることも順位に影響する．

### 2.3 関 連 研 究

ブログ情報の分類やコミュニティの抽出に関して，いくつかの研究が行われている．

Kumar らは，ブログ間のリンク関係を解析して時間グラフを生成し，コミュニティとバースト性を調べた[7]．Adar らは，ブログ間のリンク関係と時刻を用いてブログ空間の情報拡散を分析するとともに，引用関係に基づいてブログエントリをランキングする手法として iRank を提案した[8]．谷口らは，収集したブログエントリに対し本文中から張られているリンクを抽出し，PageRank を用いて authority Blog を抽出し，また Betweenness Clustering によりブログコミュニティを抽出した[9]．Marlow は，他ブログへのリンクであるブログロール（blogroll）のリンクと内容を更新しても変化しないパーマリンクへのリンクを解析し，さらにそれらを用いたランキング方法を提案した[10]．

トラックバックを取り上げた研究が見あたらないのは，すでに述べた理由から，ランキングなどの応用が考えられにくいからであると思われる．しかし，本稿では，逆にトラックバックのネットワーク構造がブロガの活動を表すことから，特に SEO スパムの観点から，SEO という特殊な活動をしているコミュニティを発見・分離するための特殊なランキング手法を確立しようとする点で，既存の研究とアプローチが異なる．

## 3. 実験データ

### 3.1 ブログエントリの収集

トラックバックネットワークを解析するためには，トラックバックでつながった比較的類似した内容のブログエントリ集合を短期間に収集する必要がある．たとえば，ブログ検索では，weblogUpdates ping を使ってブログの更新情報を管理する Ping サーバからブログの一覧を入手して収集することが多い．しかし，国内のブロガの約 6 割がトラックバックをしたことがないという調査報告[11] があるように，実際のブログでは必ずしもトラックバックが頻繁に行われているとは限らないので，このような方法では収集したブログエントリのごく一部しかトラックバックネットワークに関係していないので非効率的である．

そこで，収集起点として goo ブログの「テーマサロン」（**図 2**，http://blog.goo.ne.jp/userstheme/）を利用し，そこからトラックバックをたどりながら幅優先で収集した.「テーマサロン」は，goo ブログがトラックバックを体験するために提供しているサービスであり，テーマごとに分類されているために，ある程度類似した内容を収集しやすい．さらに,「テーマサロン」に参加するブロガは，トラックバックを使ったコミュニケーションが目的であることからトラックバック数も期待でき，比較的効率良くトラックバックネットワークを収集できると考えられる．

具体的には，次の手順でブログエントリを収集した．

(1) goo のテーマサロンのあるテーマの HTML ページに掲載されているトラックバック一覧を解析して，各ブログのトラックバック元の URL リストを抽出する．


**図 2** goo ブログのテーマサロン
Fig. 2 Theme salon in goo blog.

**表 1** 対応ブログサービスと収集量
Table 1 Supported blog services and amount of their blog entries.

| サービス名 | エントリ数 | % |
|---|---|---|
| livedoor Blog | 11,266 | 36.4 |
| goo ブログ | 3,223 | 10.4 |
| Seesaa BLOG | 2,707 | 8.8 |
| Ameba Blog | 1,701 | 5.5 |
| ココログ | 1,492 | 4.8 |
| 楽天広場 | 1,398 | 4.5 |
| excite ブログ | 1,351 | 4.4 |
| ジュゲム | 680 | 2.2 |
| yaplog! | 639 | 2.1 |
| LOVELOG | 269 | 0.9 |
| ブログ人 | 357 | 1.1 |
| はてなダイアリー | 171 | 0.6 |
| その他 | 5,644 | 18.3 |

(2) URL リストに含まれる各ブログエントリの XHTML ファイルを収集する.
(3) 収集した XHTML ファイルの中に含まれる他ブログエントリからのトラックバックを表示している部分を解析し，新たなトラックバック元の URL リストを作成する.
(4) 収集開始点からの距離が $n$ 段を超えれば終了する．そうでなければ，再び (2) に戻る.

今回は $n = 10$ で収集した.

ただし，抽出対象となるトラックバック表示部分は単なる XHTML として記述されている．そこで，各ブログサービスごとに XHTML のテンプレートが決まっていることに着目し，**表 1** に示す主要なブログサービスに対して，それぞれトラックバック元の URL，ブログ名，題名を抽出するための解析ルーチンを用意した．さらに，ブログサービスのドメインとテンプレートの対応を定義しておき，収集時にはブログエントリの URL のドメイン部を抽出して，適切な解析ルーチンを適用した．この方法ではすべてのブログサービスに対応するのは難しいが，ブログサービスのユーザ数には大きな偏りがあることから，トラックバック表示部分を解析できなかった URL は全体の 18.3%にとどまり，対応したブログサービス数に対するサポートブログサービスの割合は高いと考えられる.

なお，数カ月の間同じ手法で収集を繰り返したが，サーチエンジン運営側に SEO スパムと判断される危険を恐れてか，第 1 回目の SEO コンテスト後に内容を変更したり，第 2 回目の終了に比較的近くなってから SEO を始めたりする現象も観察された．本稿では，第 2 回の SEO コンテスト終了直前が一番 SEO のネットワーク構造がはっきりしていると考え，2005 年 4 月 4 日に「テーマサロン」から「ライブドアのこと」というテーマで収集したブログエントリデータを用いる.

### 3.2 SEO エントリの判定

収集したブログエントリのうち，次のいずれかの条件を満たすものを，SEO エントリと判定した.

(1) トラックバック自動発見用に各ブログエントリの XHTML 中に埋め込まれている RDF [12] の rdf:Description 要素の情報に,「ゴッゴル」または「デースケドガー」という単語が現れるブログエントリ．たとえば，Dublin Core で定義されているメタ情報である dc:creator 属性（ブロガ名），dc:title 属性（題名），dc:description 属性（概要）が相当する [13].
(2) トラックバック先のトラックバック表示部分のブログ名または題名に,「ゴッゴル」または「デースケドガー」という単語が出現するブログエントリ.
(3) 「トラックバック OK　ゴッゴル」または「トラックバック OK　デースケドガー」というバナーが含まれるブログエントリ.

(1) と (2) は，トラックバックを使った SEO の証拠であり，(3) は SEO コンテストへの参加を示す．(1) と (2) という比較的似た条件を併用する理由は，必ずしもすべてのブログエントリが RDF 定義を持っているとは限らず☆，さらに「はてなダイアリー」のようにトラックバック表示部分に URL しか表示しないブログサービスが存在するからである.

なお，判定に本文の概要は用いても全文は使わない．これは，SEO 行為をともなわない，単に SEO コンテストに言及しているだけのブログエントリを除外する

☆ 使用したデータでは 91%が RDF を含んでいた.

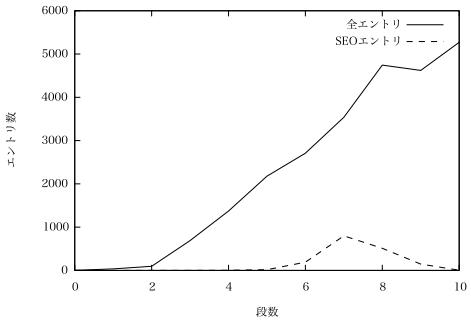


**図 3** 起点からの段数とブログエントリ数
Fig. 3 The number of steps and the number of blog entries.

ためである．逆に，SEO を行っているブログエントリでは，これらのキーワードのテキスト中の出現回数やトラックバック先のアンカーテキスト数を増やそうとするために，上記の部分にキーワードが含まれる確率が非常に高く，抽出漏れは少ないと思われる．

ただし，この判定方法では，判定漏れは非常に少なくなる半面，SEO コンテストに興味を持っているだけだったり，参加しているが熱心ではなく SEO 技術を駆使していたりするという点では適切でないブログエントリが若干存在する可能性がある．

### 3.3 収集データの分析

本方法で収集できた全ブログエントリ数は 25,254 件，総トラックバック数は 190,107 本であった．平均トラックバック数は 7.5 本であることから，効率良くトラックバックネットワークを収集できたといえる．

収集したデータの内訳を 表 1 に示す．ブロガ数が特に多いといわれている livedoor Blog と goo ブログの割合が大きいなど，一般のブログ調査と同様な傾向が得られていることから，収集起点が 1 つであってもトラックバックをたどってある程度の段数を収集することで，広範囲に収集できているといえる．ただし，同様にブロガ数が多いといわれているはてなダイアリーの比率が低い．これは Web 日記を元にした独自のシステムであることと，トラックバック元の記事にトラックバック先へのリンクが存在しなければいけないという制約があるからだと考えられる．

さらに，SEO と判定されたブログエントリ数は 1,674 件で，全体の 6.6%であった．起点からの段数と，各段で収集された非既出ブログエントリ数または非既出 SEO ブログエントリ数の関係を，**図 3** に示す．

この結果から，SEO では特に多くのトラックバックを張るために，SEO コミュニティには少ない段数でたどり着いてしまうことが分かる．そして起点と SEO コミュニティのトピックの相違から明かなように，また収集段数が増えるにつれてトピックがかなり変遷していると思われる．さらに，トラックバックをたどって収集しても，エントリ数は爆発的には増えない．この理由は，段数に応じてたどるトラックバック数は大幅に増えるが，既収集エントリに対するものが大部分を占めているからだと推測している．

次に，SEO コミュニティが非 SEO ブログエントリ群とトラックバックで直接つながっている部分を境界部分と定義し，ここだけに着目する．ここで，非 SEO ブログエントリにつながっている SEO ブログエントリは 631，SEO ブログエントリにつながっている非 SEO ブログエントリは 657 である．この間のトラックバック数を求めると，SEO ブログエントリから非 SEO ブログエントリへのトラックバックが 1,177 本，非 SEO ブログエントリから SEO ブログエントリへのトラックバックが 900 本であった．さらに，この境界部分には，SEO とは判断されないが比較的遊び的要素の強いブログが多く，著名ブログらしいものはほとんど見あたらなかった．ただし，この点に関しては，第 1 回目でコンテスト参加者以外のブログエントリに多くのトラックバックが張られる問題を引き起こしたために，開始 1 週間後からは参加者以外へのトラックバックは強く禁止されたことが影響している可能性がある．

## 4. トラックバックネットワークの基本統計量の解析

社会ネットワーク解析や Web のリンク解析に代表される複雑ネットワーク理論の研究でよく行われている基本的な統計量を解析する．次数相関やクラスタ係数などの基本統計量の標準定義が無向グラフに対して行われているが，本稿ではこのようなネットワークの基本性質を探求することを目的としていることから，一律に無向グラフとして扱う．

### 4.1 次　　数

ネットワークの中のノード $i$ の次数 $k_i$ は，ノード $i$ に張られているリンク数として定義される[14]．SEO コミュニティと非 SEO ブログエントリ群の次数分布を **図 4** に示す．

x 軸は次数 $k$，y 軸はその存在確率 $P(k)$ であり，次のように求めた．

$$P(k) = \frac{|i : k_i = k|}{|N|} \tag{1}$$

ここで，$N$ は全要素の集合を表し，$|N|$ はその要素数を表す．

全体・SEO コミュニティ・非 SEO ブログエントリ

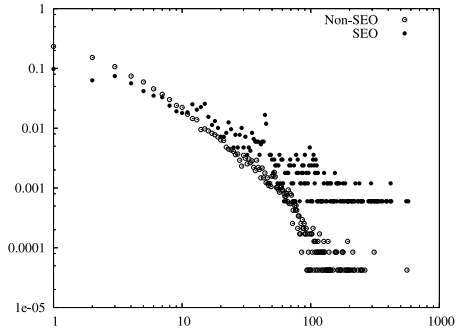


**図 4** 次数の分布
Fig. 4 Distribution of degrees.

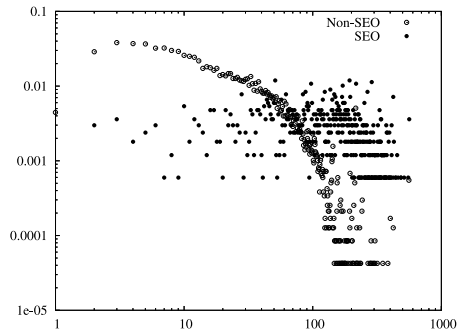


**図 5** 近傍平均次数の分布
Fig. 5 Distribution of average degrees of nearest neighbors.

**表 2** トラックバックネットワークの基本統計量
Table 2 Basic statistical parameters of blog trackback network.

| | 全体 | SEO | 非 SEO |
|---|---|---|---|
| 次数 | 10.8(23.1) | 33.6(23.1) | 9.18(16.8) |
| 近傍平均次数 | 41.5(54.0) | 156.7(102.4) | 33.3(37.0) |
| クラスタ係数 | 0.239(0.329) | 0.296(0.285) | 0.235(0.332) |

の次数の平均と標準偏差を **表 2** に示す．ただし，括弧の中が標準偏差である．SEO コミュニティと非 SEO ブログエントリ群のどちらもベキ分布を示すが，SEO コミュニティの方が指数が小さいことから，SEO コミュニティは同様に高い次数を持ったブログエントリの集団であることが分かる．

### 4.2 近傍平均次数

ネットワークの中のノード $i$ に隣接するノードの次数の平均値である近傍平均次数 $\bar{k_i}$[15] は，次のように定義される．

$$\bar{k_i} = \frac{1}{|N_i|} \sum_{j \in N_i} k_j \tag{2}$$

$N_i$ はノード $i$ に隣接するノードの集合であり，$|N_i|$ は $N_i$ の要素数である．

SEO コミュニティと非 SEO ブログエントリ群の近傍平均次数分布を **図 5** に示す．x 軸は近傍平均次数 $\bar{k}$，y 軸はその存在確率 $P(\bar{k})$ であり，次のように求めた．

$$P(\bar{k}) = \frac{|i : \bar{k_i} = \bar{k}|}{|N|} \tag{3}$$

全体・SEO コミュニティ・非 SEO ブログエントリの近傍平均次数の平均と標準偏差を 表 2 に示す．非 SEO ブログエントリ群は，近傍平均次数が高い部分ではベキ分布に従う傾向があるが，SEO コミュニティではそのような傾向が見られない．また，SEO コミュニティでは，近傍平均次数がかなり高い部分の存在確率が高く，特にその部分に集中しているのが観察できることから，隣接するブログエントリ群も次数が高い傾向があることが分かる．

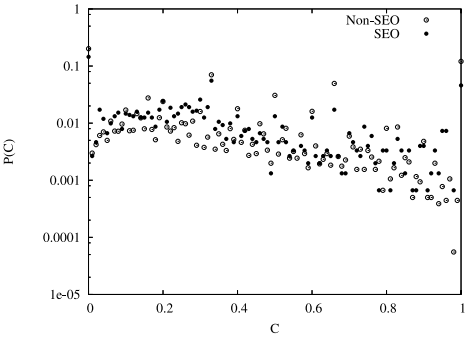


**図 6** クラスタ係数の分布
Fig. 6 Distribution of clustering coefficients.

### 4.3 クラスタ係数

クラスタ係数は，あるノードに隣接する 2 つのノードどうしにリンクが存在する割合を示す．ネットワークの中のノード $i$ のクラスタ係数 $C_i$ は，次のように定義される[16]．

$$C_i = \frac{2b_i}{k_i(k_i - 1)} \tag{4}$$

ここで，$b_i$ は，ノード $i$ とそれに隣接するノード群 $N_i$ の間のリンク数である．

SEO コミュニティと非 SEO ブログエントリ群のクラスタ係数の分布を **図 6** に示す．x 軸はクラスタ係数，y 軸はその存在確率であり，次のように求めた．

$$P(C) = \frac{|i : C - \delta < C_i <= C|}{|\tilde{N}|} \tag{5}$$

今回は $\delta = 0.01$ に設定し，$\tilde{N}$ は次数が 2 以上の要素の集合を表す．つまり，次数が 1 の要素にはクラスタ係数が定義できないので無視する．

全体・SEO コミュニティ・非 SEO ブログエントリのクラスタ係数の平均と標準偏差を 表 2 に示す．クラスタ係数の分布は，Vázquez の研究で行われたイン

ターネットの AS やルータ，Gnutella，Web，タンパク質，共著関係と比較すると，やはり Web の分布に近い[17]．また，SEO コミュニティの方がクラスタ係数が大きいが，どちらもピークがなく，特筆すべき違いは見られない．

## 5. SEO コミュニティの検出

### 5.1 ネットワーク構造の特徴量を用いたランキング

前章では，全ブログエントリと SEO コミュニティのブログエントリに対して，ネットワークの基本統計量を分析したが，SEO コミュニティの次数，近傍平均次数，およびクラスタ係数のそれぞれにおいて，全体よりも平均値が大きい傾向や，分布が異なる様子が観察された．この事実は，ネットワーク構造から導き出される特徴量を用いてブログエントリを評価すれば，ブログ空間の SEO コミュニティを発見できる可能性を示している．

これを確認するために，収集した全ブログエントリを，いくつかのネットワーク構造に関連する特徴量を用いてランキングする．SEO コミュニティに属すると判断されたブログエントリ群が高い順位を占めるとしたら，SEO コミュニティが検出できたことになる．

そこで，このような特徴量として，まず前章で分析したネットワーク基本統計量である次数，近傍平均次数，およびクラスタ係数を用いる．本稿では，これらの基本統計量を用いた手法を，それぞれ**次数法**，**近傍平均次数法**，および**クラスタ係数法**と呼ぶことにする．

さらに，SEO コミュニティのメンバはサーチエンジンの検索結果で高い順位を得ることを目指していることから，HITS（Hypertext Induced Topic Selection）[18] や PageRank[3] などの検索結果のランキングで用いられているリンク解析手法についても評価するのが望ましい．ただし，本稿では無向グラフを扱っているために，そのまま用いることはできないので，HITS と PageRank を無向化したネットワーク向けに変更したアルゴリズムを使用し，それぞれ**主成分ベクトル法**と**定常確率法**と呼ぶことにする．

### 5.2 主成分ベクトル法

主成分ベクトル法では，ネットワークの隣接行列を $\mathbf{A}$ とするとき，この行列の第 1 固有ベクトル $\mathbf{u}$ を求め，その第 $i$ 成分 $u_i$ をノード $i$ の評価値として大きい順にランキングする．

固有ベクトル $\mathbf{u}$ は，以下のパワー法を土台とした反復で求める．

**1.** $t=1$，$\mathbf{u}_i^{(0)}=1$ $(i \in N)$ と初期化する；
**2.** $\widetilde{\mathbf{u}}=\mathbf{A}\mathbf{u}^{(t-1)}$，$\mathbf{u}^{(t)}=\widetilde{\mathbf{u}}/\max_i \widetilde{u}_i$ を求める；
**3.** $\max_i |u_i^{(t)}-u_i^{(t-1)}|<\eta$ なら反復を終了する；
**4.** $t=t+1$ として **2** に戻る．

ここで，$\eta$ は終了条件を制御する正の実数であり，反復終了後に $\mathbf{u}=\mathbf{u}^{(t)}$ として結果が求まる．

### 5.3 定常確率法

定常確率法では，各要素がノード次数 $k_i$ の対角行列を $\mathbf{K}$，全要素の値が $1/|N|$ の $|N|\times|N|$ 行列を $\mathbf{U}$ とするとき，行列 $\mathbf{B}$ の第 1 固有ベクトル $\mathbf{v}$ を求め，その第 $i$ 成分 $v_i$ をノード $i$ の評価値として大きい順にランキングする．

$$\mathbf{B}=(1-\epsilon)\mathbf{A}\mathbf{K}^{-1}+\epsilon\mathbf{U} \tag{6}$$

ここで，一様ジャンプ確率 $\epsilon$ は，Ng らの論文[19] に従って $\epsilon=0.15$ に設定する．固有ベクトル $\mathbf{v}$ は，主成分ベクトル法と同様に，パワー法を土台とした反復で，$\mathbf{A}$ を $\mathbf{B}$ に，$\mathbf{u}$ を $\mathbf{v}$ に置き換えて求める．

### 5.4 F 値と精度

ランキング手法による検出性能を定量的に評価するために，情報検索の性能評価などで広く使われている F 値（F-measure）と精度（precision）を用いた．

F 値 $F(r)$ と精度 $P(r)$ は，SEO コミュニティのブログエントリ集合を $S$，各手法で上位 $r$ 番以内のエントリ集合を $M_r$ とすれば，以下のように定義される．

$$F(r)=\frac{2|M_r\cap S|}{|M_r|+|S|} \tag{7}$$

$$P(r)=\frac{|M_r\cap S|}{|M_r|} \tag{8}$$

ここで，$|A|$ は集合 $A$ の要素数を表す．

### 5.5 SEO コミュニティ検出性能の評価

**図 7** と **図 8** に，**次数法**，**近傍平均次数法**，**クラスタ係数法**，**主成分ベクトル法**および**定常確率法**を用いて得られた順位 $r$ と，F 値 $F(r)$ と精度 $P(r)$ の関係を示す．ただし，複数のブログエントリが同じ順位の場合は，その数だけプロットをスキップしていることに注意されたい．たとえば，クラスタ係数法では，係数 1 でトップとなるエントリが 1,100 程度存在したので，実際のプロットは $r=1{,}100$ 程度から開始している．また，次数が 1 の要素はクラスタ係数が 0 の場合と等しいようにランキングしている．

図 7 と 図 8 から，主成分ベクトル法が最も高い性能を示すことが分かる．$r=1{,}100$ 付近で F 値が 90%を超え，精度も 100%に近い点は特筆できる．この結果は，主成分ベクトル法は，リンクファーム特有のネットワーク構造の検出能力が高いことを示唆している．実際に，主成分ベクトル法は，目的関数として以下の Rayleigh 商を最大化する解を求めることと等

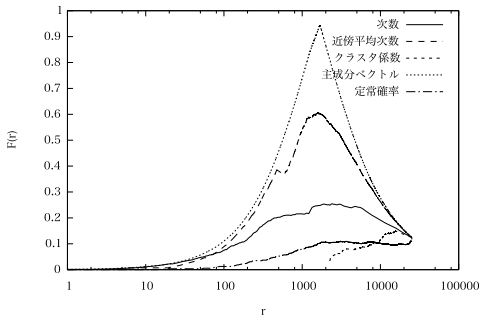


**図 7** F 値の評価
Fig. 7 Evaluation of F-measure.

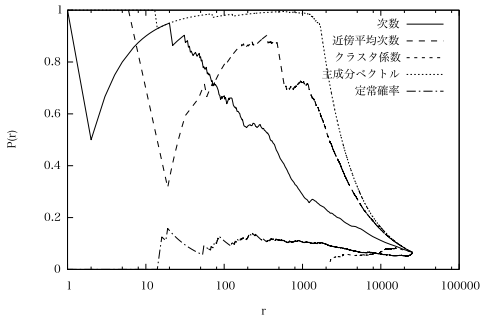


**図 8** 精度の評価
Fig. 8 Evaluation of precision.

価である．

$$G(\mathbf{u}) = \frac{1}{2}\frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}. \tag{9}$$

ここで，$\mathbf{u}^T$ はベクトル $\mathbf{u}$ の転置を表す．$\mathbf{u}$ の各要素を 0 または 1 のバイナリ値に限定すれば，$G(\mathbf{u})$ の値は，ノード集合 $\{i:\ u_i = 1\}$ からなる部分ネットワークの平均リンク数となり，主成分ベクトル法での $G(\mathbf{u})$ の最大化は，平均リンク数が大きく凝集した部分ネットワークを求めるための緩和問題を解くことにほかならない．すでに述べたように，SEO スパムで作られるリンクファームでは，特定のノード群の間でリンクが密に張られる特殊なネットワーク構造を持つが，主成分ベクトル法は，このような平均リンク数が大きく，凝集した部分ネットワークを検出していると考えられる．

同時に，この実験結果は，HITS は PageRank のような検索結果のランキングには向いていないことを示唆している．たとえば，Bharat ら[20] が，HITS はホスト間で相互に強め合うような関係がある場合やリンクを自動生成している場合には，トピックに適合しないノードが含まれて相互に結合している場合などには望む結果が得られないと述べているように，トピックを汎化する性質が強いことが知られている．すなわち，HITS や主成分ベクトル法は，ネットワーク構造に左右されすぎて検索結果のランキングには向かないが，ある種の特殊なネットワーク構造を持つ Web サイトの集合に敏感なのではないかと思われる．

これに対して，ネットワーク基本統計量に基づくランキング手法では，近傍平均次数法，次数法，クラスタ係数法の順で高い性能を示した．この検出性能の差は，各統計量での全体と SEO コミュニティの両分布の相違の程度に符合する．近傍平均次数法に関しては，主成分ベクトル法に次ぐ結果が得られたのは，リンクファーム特有のネットワーク構造に含まれる各ノードの値は当然高くなるが，同時に次数の高いノードと 1 本のリンクだけで接続しているようなノードの値も高くなってしまう問題もあるからだと推測できる．次数法に関しては，SEO 以外に次数が高い理由として，そのブログエントリが有用だと評価されてトラックバックが張られているような正当な理由が考えられるが，単に次数だけしか見ない場合には，これらは区別できないのが影響していると推測される．クラスタ係数法に関しては，図 6 から分かるように顕著な差がなく，実際の検出性能も低いことから，適していないといえる．

さらに，この実験結果は，定常確率法は単純な次数法よりも劣ることを示している．そこで，定常確率法の処理結果の上位 20 件を観察すると，SEO とアダルト以外のトピックのまともな内容のブログエントリが 17 件も占め，これらのブログエントリ群の順位は，次数法では 44 位から 245 位に落ちることから，次数はそれほど高くないことが分かった．現時点では，定常確率法は，ブログのトラックバックネットワーク全体を行動する利用者の閲覧確率を求めることと同等なことから，局所的に密なだけのリンクファームの場合には，次数は高くてもたどり着く確率はそれほど高くならないと推測しているが，有向グラフとして扱うことを含めて，さらなる検討が必要である．

ここで，定常確率法の性能が劣ることから，リンクファームは SEO スパムに有効でないと思われるかもしれないが，Google では，PageRank に加えて，ある Web ページの本文だけでなく，それに対するリンクのソースアンカー部分のテキスト（アンカーテキスト）を同時に索引付けしており[6]，オフィシャルサイトのようにある単語が示唆する最も代表的で一般的な Web ページを検索結果の上位に持っていくためには，PageRank はあまり有効ではなく[21]，アンカーテキストの効果の方が大きいことが知られている[2]．そこで，

SEO スパムに必ずしも PageRank が寄与しなくても，アンカーテキストにおける数の多さが大きく寄与してしていると推測される．

一方，ネットワークを分離するためには，多様なコミュニティ抽出法[22)~24)]がすでに多く存在する．これらの手法では，コミュニティは基本的に排他的なものとして定義され，たとえば密結合する 2 つの部分は両者の隘路で分離されることになる．ところが，このような考え方では，複数の SEO スパムを行っているエントリが，そうでないエントリに多数のトラックバックを張る場合などは，既存手法の単純な適用では不都合が起こることが容易に想定できる．そこで，本稿では，ネットワークの特徴量に基づいてノードを順位付けすることで，同じ特徴を共有するコミュニティを抽出するという，逆のアプローチをとっている．これでは，コミュニティが必ずしも排他的であることは要求しない．SEO コミュニティ検出の観点で，本手法と既存のコミュニティ抽出法を比較評価し，各手法の長所や短所を明確にすることは，今後の重要な研究課題の 1 つと考えている．

## 6. おわりに

本稿では，SEO という観点からブログのトラックバックネットワークを解析し，SEO コミュニティが持つネットワーク構造に関する特徴を明らかにした．さらに，ネットワーク基本統計量とともに，隣接行列の主成分ベクトルと定常確率を利用して，SEO コミュニティを再分離して，その F 値と精度を求め，トラックバックネットワークの構造から SEO コミュニティが分離できることと，特に主成分ベクトル法が検出法として優れていることを示した．本稿では，検索結果のランキングにおける HITS（本稿では主成分ベクトル法）の問題点を，逆に利用しているといえる．

今回は SEO コンテストのデータを分析や評価に用いたが，SEO コンテストは一種のお祭り的な雰囲気がある点が特異であり，通常の SEO スパムの場合とコミュニティのサイズやリンク構造に違いがある可能性がある．そこで，一般的なブログエントリデータに対しても評価し，その有効性を再確認する必要がある．さらに，今回は無向グラフとして扱ったが，トラックバックは方向性を持つことから有向グラフとして扱った場合の評価も必要である．

今回は SEO コミュニティが対象であったが，フィルタリングが望まれるような公序良俗に反するアダルト情報のブログ群や，現在ネットワーク上で流行しているトピックについて議論するブログ群，または熱狂的なマニアのブログ群も，同様に特徴的で比較的密なネットワーク構造を持ち，それがネットワーク構造の特徴量にも何らかの形で反映されると考えられるので，このようなコミュニティの発見も検討したい．

**謝辞** 本稿に有益な助言をいただいた上田修功部長，および本実験を手伝っていただいた NTT コムウェアの藤本裕文氏に感謝する．

## 参考文献

1) Thurow, S.: *Search Engine Visibility*, New Riders Publishing (2003).
2) 風間一洋，原田昌紀，佐藤進也：ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法，情報処理学会研究会報告 FI-59-3/DD-24-3，pp.17–24，情報処理学会 (2000).
3) Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project (1998).
4) Google: Preventing Comment Spam (2005). http://googleblog.blogspot.com/archives/2005_01_01_googleblog_archive.html
5) Gyöngyi, Z. and Garcia-Molina, H.: Web Spam Taxonomy, Technical report, Stanford University (2004).
6) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. 7th International Conference on World Wide Web*, Brisbane, Australia, pp.107–117 (1998).
7) Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the Bursty Evolution of Blogspace, *Proc. 12th international conference on World Wide Web*, pp.568–576 (2003).
8) Adar, E., Zhang, L., Adamic, L.A. and Lukose, R.M.: Implicit Structure and the Dynamics of Blogspace, *Workshop on the Weblogging Ecosystem*, the 13th International World Wide Web Conference (2004).
9) 谷口智哉，松尾　豊，石塚　満：Blog コミュニティの抽出と分析，第 6 回セマンティックウェブとオントロジ研究会，人工知能学会 (2004).
10) Marlow, C.: Audience, Structure and Authority in the Weblog Community, *International Communication Association Conference*, International Communication Association (2004).
11) goo リサーチ：第 14 回：Blog に関する調査 (2005). http://research.goo.ne.jp/Result/0504cl11/01.html
12) Six Apart: TrackBack Technical Specification (2004). http://www.sixapart.com/pronet/docs/trackback_spec
13) Dublin Core Metadata Initiative: Dublin Core

Metadata Element Set, Version 1.1: Reference Description (2004). http://dublincore.org/documents/dces/
14) Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks, *Science*, Vol.286, pp.509–512 (1999).
15) Pastor-Satorras, R., Vázquez, A. and Vespignani, A.: Dynamical and correlation properties of the Internet, *Physical Review Letters*, Vol.87, p.258701 (2001).
16) Watts, D.J. and Strogatz, S.H.: Collective dynamics of 'small-world' networks, *Nature*, Vol.393, No.4, pp.440–442 (1998).
17) Vázquez, A.: Growing Networks with Local Rules: Preferential Attachment, Clustering Hierarchy and Degree Correlations, *Physical Review E*, Vol.67, p.056104 (2003).
18) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
19) Ng, A.Y., Zheng, A.X. and Jordan, M.I.: Link Analysis, Eigenvectors and Stability, *the 17th International Joint Conference on Artificial Intelligence*, pp.903–910 (2001).
20) Bharat, K. and Henzinger, M.R.: Improved algorithms for topic distillation in a hyperlinked environment, *Proc. 21st ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '98*), Melbourne, AU, pp.104–111 (1998).
21) 高野 元, 久保進也：サイテーション・エンジン：リンク解析を用いた WWW 検索ランキングシステム, 情報処理学会研究会報告 DBS-120-2, pp.9–16, 情報処理学会 (2000).
22) Chung, F.R.K.: Spectral Graph Theory, *CBMS Regional Conference Series in Mathematics*, Vol.92, American Mathematical Society (1997).
23) Flake, G., Lawrence, S. and Giles, C.L.: Efficient Identification of Web Communities, *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, pp.150–160 (2000).
24) Girvan, M. and Newman, M.E.J.: Community Structure in Social and Biological Networks, *the National Academy of Sciences of the United States of America*, pp.7821–7826 (2002).

(平成 17 年 5 月 25 日受付)
(平成 18 年 1 月 6 日採録)

**風間 一洋**（正会員）

昭和 63 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在，NTT 未来ねっと研究所主任研究員．博士（情報学）．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．

**佐藤 進也**（正会員）

昭和 38 年生．昭和 63 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在，NTT 未来ねっと研究所主任研究員．電子情報通信学会，Internet Society，ACM 各会員．

**斉藤 和巳**（正会員）

昭和 38 年生．昭和 60 年慶應義塾大学理工学部数理科学科卒業．工学博士．同年日本電信電話（株）．平成 3 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習の研究に従事．現在，NTT コミュニケーション科学基礎研究所主任研究員（特別研究員）．NAIST 客員助教授．情報処理学会論文賞受賞（平成 9 年）．人工知能学会論文賞受賞（平成 11 年）．FIT 船井ベストペーパー賞受賞（平成 16 年）．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE 各会員．

**木村 昌弘**

平成元年大阪大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．現在，龍谷大学理工学部電子情報学科助教授．博士（理学）．ニューラルコンピュテーション，複雑系の数理モデリングおよび数理解析，Web マイニングの研究に興味を持つ．電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本数学会各会員．