情報処理学会論文誌　データベース　Vol. 3　No. 2　102–110 (June 2010)

推薦論文

ブログ空間の情報伝播特性を用いた情報源の多面的ランキング

風間一洋[†1]　今田美幸[†1]　柏木啓一郎[†1]

本論文では，3 種類の情報伝播特性を用いた，ブロガが注目している情報源の多面的ランキング手法を提案する．ブログ検索結果から抽出した情報伝播ネットワークを各情報源から到達可能な複数の部分ネットワークに分割し，部分ネットワーク中の 3 種類の有向 2 エッジ連結部分グラフの数から定義した情報拡散度，情報集約度および情報転送度を情報源のランキングに用いる．さらに，実際の情報伝播ネットワークに適用し，ランキング結果の順位相関係数を調べるとともに，オフィシャルサイト，ニュース，CGM という 3 種類の情報源の MRR と MAP を用いてランキング手法の特性を明らかにし，情報伝播特性を用いた多面的ランキングの有効性を示す．

Multi-faceted Ranking of Information Sources Using Information Diffusion Properties in Blogspace

Kazuhiro Kazama,[†1] Miyuki Imada[†1] and Keiichiro Kashiwagi[†1]

This paper proposes a multi-faceted ranking method of information sources attracting bloggers by using three information diffusion properties. The information diffusion network extracted from searched blog entries is divided into subnetworks, each of which is reachable from an information source. We define the information scatter degree, the information gather degree, and information transmit degree as information diffusion properties by using the number of three types of directed 2-edge connected subgraphs in a subnetwork and use them for the ranking of information sources. We apply this ranking method to real information diffusion networks and analyze rank-order correlation coefficient of ranking results. Furthermore, we characterize ranking methods by using MRR and MAP of three-types of information sources: official sites, news sites and CGMs. In the result, we demonstrate the effectiveness of a multi-faceted ranking method by using information diffusion properties.

1. はじめに

ブログの登場により Web 上の情報公開コストが大幅に低下し，評判情報を含めて多種多様な情報が得られるようになった．反面，その情報量は膨大で，重複している情報や無益・有害な情報なども多く含まれることから，ユーザが有効に活用できているとはいいがたい．

本論文では，ユーザの情報探索の目的に合わせた有用な情報源の推薦を実現するために，情報伝播ネットワーク構造の情報源の影響範囲である部分ネットワークの 3 種類の情報伝播特性の使い分けによる多面的なランキング手法を提案する．まず，ある検索語でブログ検索した結果から，ブログエントリとそこから参照されている Web ページをノードとし，ハイパーリンクと逆向きの有向エッジを持つ情報伝播ネットワークを抽出する．各情報源からエッジをたどって到達可能な部分ネットワークに対して 3 種類の情報伝播特性，情報拡散度，情報集約度および情報転送度を定義し，これらの特性値は情報源がブログ空間に対して与えた影響度と見なすことで，ブロガが注目している複数の情報源を多面的にランキングする．

2. 関連研究

リンク解析ベースのランキング手法としては，被リンク数，HITS のオーソリティ度・ハブ度[1] および PageRank[2] がある．被リンク数は隣接ノードとのリンクだけを考慮するのに対し，HITS と PageRank は同様に被リンク数をベースとしながらもネットワーク全体を考慮するので，巨大なネットワークの全体情報が必要になる．本手法は局所的な情報伝播経路上のノードだけを考慮するので，データを検索 API 経由で利用する場合など，Web 空間のデータをローカルに持たずに全体情報にアクセスできない場合でも使用できる[⋆1]．

また，これまでネットワーク分析に使われてきたネットワークの基本構造は，2 者関係のダイアドと 3 者関係のトライアド[3]，クラスタ係数の三角構造[4]，ハイダーの認知的バランス理論の三者関係[5] などがあげられる．また，Milo らは複雑ネットワーク中に高い頻度で現れる構成要素のパターンであるネットワークモチーフ[6] を分析した．既存手法は n ノー

†1 日本電信電話株式会社 NTT 未来ねっと研究所
NTT Network Innovation Laboratories, Nippon Telegraph and Telephone Corporation

⋆1 HITS では検索結果をルート集合とする近傍グラフを分析するが，Link Popularity が高いオーソリティやハブが検索結果の上位に集中していたり，ある Web ページにリンクしている Web ページ一覧が取得できるなど，ネットワーク全体が既知であることを前提としている．

ド部分グラフを対象とするが，本手法は 2 エッジ部分グラフを解析する点が異なる．2 ノード部分グラフはエッジそのものであり，3 ノード部分グラフには本論文の 3 種類の基本構造が含まれる．しかし，$A \to B$，$B \to C$，$A \to C$ という 3 本のエッジからなる 3 ノード部分グラフでは，本論文で述べる情報拡散構造 $A \to B$，C と情報転送構造 $A \to B \to C$ のような情報伝播の基本的な現象を独立に扱えないことから，情報伝播のような「流れ」の基本構造の分析には，ノード間の関係ではなくエッジ間の関係を用いる必要がある．

ブログにおける情報伝播の研究としては，Adar らのブログ間の埋め込まれた URL の情報伝播に関する研究[7]，Kim らによる文章の再利用から情報伝播を探る研究[8]，Gruhl らのブログ空間における個人間のトピックの伝播を分析した研究[9] などがある．本手法は，ブログ空間に閉じずに外部の影響も考慮し，テキスト情報を使用しないために大量に存在するテキスト複製による自動生成ブログの影響を受けない．さらに，ハイパーリンク経由の情報伝播しか扱わないので，発見できる情報伝播経路は限られても，形式が整った良質な内容のブログエントリだけを対象にできるので，精度的に有利である．

3. ブログの情報源の多面的ランキング

3.1 ブログ情報利用の問題点

ブログ上では多種多様な人々による大量の情報が公開されているが，必ずしも十分に活用されていない．

まず，ブログ空間はつねに外部の影響を受けているにもかかわらず，従来のブログ検索や既存の研究では閉じた世界として扱われることが多かった．しかし，ブログエントリの情報源の多くはオフィシャルサイトやニュースであり，それらに対する感想や意見，追加情報がブログ記事に書かれるなど，互いに相補的な関係にあることが多く，外部の影響を無視できない．

また，ブログの膨大な情報の中には有益な情報が存在するが重複も多く，無益な情報や有害な情報も多い．たとえば，IICP の調査結果[10] では，2008 年 1 月の時点の日本国内のブログ総数は約 1,690 万（1 月に 1 回以上更新されているブログは約 300 万），記事総数は約 13 億 5,000 万，データ量は 42 テラバイト，スパムブログはアクティブブログの 12%を占めていると推測されている．ブログエントリの文章は比較的簡潔なので，文章だけを手がかりに有益な情報を選び出したり，他の人間が書いた文章を組み合わせて自動生成されるスパムブログを除去したりするのは難しい．

さらに，ブログの情報を処理する場合には，Web 検索で有効な PageRank のような Web 空間のハイパーリンクネットワークの全体情報が必要な手法を使うことができないので，適切なランキングが難しい．そのために，ブログ検索のランキング方法はいくつも提案されていても，結局は情報作成時間順で扱うことが多い．

3.2 ブログ情報の二次的な利用

ブログは気軽に書かれるために情報量が膨大でも，情報が不完全で自己完結的でないブログエントリは多いことから，ブログエントリは情報探索の対象ではなく，注目されている情報源を探したり，その情報源に対する評判や感想を知ったりするための二次的な手段と考える．実社会で発生したイベントを，ブログに限らず新聞やニュース，オフィシャルサイト，まとめサイトなどの一次情報源から知った多くのブロガは，それに関するブログエントリを書く．そこで，あるトピックを表す検索語で収集したブログエントリの本文中のハイパーリンクを分析すれば，そのトピックに関して特に注目されている情報源を見つけ出すことができる．この情報源は必ずしもブログであるとは限らないので，ブログ空間外部の影響を考慮した，より現実に近い分析ができる．

ただし，有用な情報源を見つけ出すのは比較的簡単でも，実際にそれらをどのように提示するかは難しい．それは，ユーザによって情報探索の目的は様々で一律に扱えないからであり，ある方針に基づいて情報源をランキングしたとしても，情報探索の目的によっては適さない場合もある．

3.3 ユーザの情報探索指向に応じた多面的ランキング

多くのブロガが情報源から入手した情報をブログエントリを書いてより多くの人たちに連鎖的に情報を伝えた結果として情報伝播ネットワークが生成される．情報源とそこから到達可能なブログエントリ集合を対で扱うことで，その情報源がブログ空間に与えた影響を，ブログエントリ集合のリンク構造やテキストの内容から推測できる．

本論文では，特に情報伝播ネットワークの構造に着目する．情報源からたどれる部分ネットワークは，情報源がブロガに与えた影響に応じて情報拡散，情報集約および情報転送という情報の基本操作を行った結果として生成され，情報源の価値や性質を反映している．これらの 3 種類の基本操作に関する情報伝播特性を定量化すれば，複数の観点から情報源の重要度を知ることができ，さらに関連している 3 種類の情報伝播特性を使い分ければ，注目されている情報，資料性の高い情報，口コミで伝わりやすい情報など，ユーザの情報探索の目的に合わせた情報源の推薦を実現できる．

この手法は，自然言語処理は用いないためにブログエントリのテキスト長の問題に制限されず，また情報源から到達可能な近傍を考慮するために，Web 空間全体のネットワーク構

造情報が得られない場合でも比較的妥当なランキングが可能である．

3.4　ソーシャルブックマークとの比較

ソーシャルブックマーク（SBM）でも，複数のユーザが情報源をリンクした結果であるブックマークを統合することで，同様のネットワーク構造が得られる．ただし，SBM は単一システム内の情報に制限されるが，本手法は多くのブログサービスを同じ枠組みでカバーできる．さらに，SBM は情報源とユーザの二部グラフ構造だが，情報伝播ネットワークは本論文で述べるようにより複雑なネットワーク構造を持つ．

4. 情報伝播ネットワークの抽出

以下に，注目されている情報源を発見し，記事の本文中に存在するハイパーリンクを逆順にたどることで情報伝播ネットワークを抽出する手法を示す．

4.1　ブログエントリの検索と収集

ブログ空間には多種多様な情報が伝播しているが，特定の情報の流れを詳しく解析するためには，特定のトピックに関するブログエントリに限定する必要がある．そこで，Yahoo! Japan または Technorati Japan のブログ検索を用いて，そのトピックを表す検索語が含まれるブログエントリを取得する．

通常は何らかの現実のイベントが発生したことを知ってからブログエントリの収集を開始するが，収集開始前のイベント発生時から，何が起こったのかを長い時間にわたり継続的に調べる必要がある．ブログ検索エンジンは時間の降順にソートした検索結果を返すので，まず最初に収集開始時刻より遡って古いブログエントリを取得するために最大 1,000 エントリ位まで検索を繰り返し，次に継続的に新しいブログエントリを取得するために更新状況に応じて適度な時間間隔を空けながら数日〜数週間程度検索を検索を繰り返し，検索結果に含まれるブログエントリを Web ロボットで収集する．

4.2　本文の特定とハイパーリンクの抽出

ブログのエントリには内容とは直接の関係がない多量のハイパーリンクが含まれるために，本文部分を特定して，その中に含まれるハイパーリンクだけを抽出する．

本文部分は，異なるプログラムやブログサービスでも基本的な HTML 文書構造が似ている点に着目して，Google AdSense 用のコメントで囲まれた領域，指定された class 属性を持つ div 要素で囲まれた領域（ドメイン名ごとに判定），div 要素や td 要素で囲まれた領域で，コメント部分に記述されている RDF の dc:description 属性に指定されたテキストと類似した領域，テキストが指定されたアンカーテキスト部分の比率，テキスト長，句読点数の条件に一番合致する div 要素や td 要素で囲まれた領域，という 4 種類の領域を順に探して特定する．

4.3　ノードとエッジの生成時刻の特定

抽出したネットワーク構造の分析に加えて，時間的変化とそれにともなう特性の変化も分析できるように．ノードとエッジの生成時刻を特定する．

ノードの場合には，収集済みのブログの場合はその生成時刻を使用し，そうでなければそのノードを一番最初にリンクした時刻を生成時刻とする．エッジの場合には，リンク元のノード生成時刻をエッジの生成時刻とする．ただし，サーバの時刻設定の狂いにより明らかに誤った時刻のブログエントリが存在する場合は，時間範囲を指定して分析対象から除外する．

4.4　情報伝播ネットワークの作成

抽出された URL は，ブログエントリを記述する際に使われた情報源である．しかし，少数の知人だけにしか意味がない情報や個人的な写真データを広く注目されている情報と同一に扱うのは情報伝播を分析するうえで適切とはいえず，前者はノイズとなることが多い．また，ブログサービスでは，サービス内の他のブログやサービスへのナビゲーションリンクを自動的に大量生成する傾向がある．

そこで，抽出したすべての URL に対して異なるサーバからの被リンク数を計算し，指定された閾値 T 以上の URL だけを特に注目されている第 1 次情報源と見なし，そこからハイパーリンクを逆向きにたどって伝播経路を特定した．T は，ノイズと見なされるような情報源がほとんど抽出されない程度の小さい値に設定する．通常は 5 から 10 の間の値を用いる．

実際には，生成時刻の古いブログエントリから順番に，リンク先 URL がすでにノードとして登録されていた場合には，その URL からブログエントリへの有向エッジを作成し，未登録のリンク先 URL が閾値以上の被リンク数を持つ場合には，ノードとして登録し，リンク先からブログエントリへの有向エッジを作成する．この結果，情報源とブログエントリをノードとし，その間のハイパーリンクとは逆方向の有向エッジを持つ情報伝播ネットワークが得られる．

実際に「毎日新聞」という検索語で抽出した情報伝播ネットワークを，**図 1** に示す．

4.5　ノイズの除去

上記の方法で残る多量のノイズは，次のように除去した．

4.5.1　URL の正規化

異なる URL であっても同一のリソースを指していたり，ユーザの閲覧履歴をサーバ側に

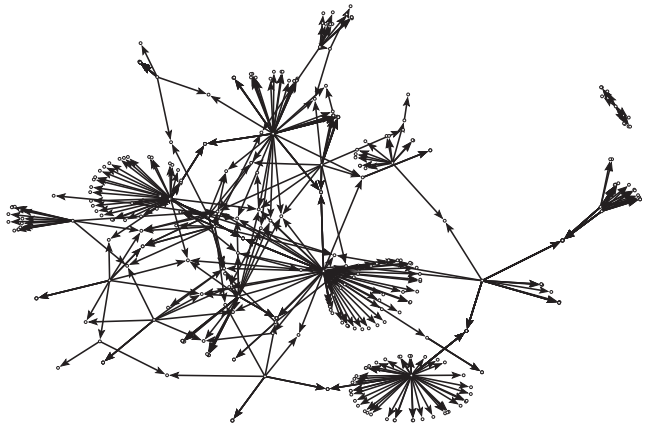


図 1　情報伝播ネットワークの例
Fig. 1　Example of information diffusion network.

記録するために特殊な符号化を施した URL を使用したりすることも多い．

前者の場合は，URL の大文字・小文字の違い，最後の “/” や “/index.html” などの違い，サーバ名の最後の “/” の有無などを正規化する．後者の場合は，URL のパラメータ部分に本来の URL を符号化していったん閲覧記録用のサーバを経由させてから，本来の URL にリダイレクトしなおしているので，閲覧記録用の URL かどうかをサーバ名で判定し，そのパラメータ部分の URL を復号化して使用する．

4.5.2　ブラックリストによる URL のフィルタリング

本文中に内容と直接の関係がないキーワードサイト，SBM（登録用），ブログランキング，メールマガジン，アフィリエイトサービス，SEO サービスなどが本文中に自動・手動で挿入されていると，それを参照しているブログエントリがまとまって抽出されるという問題がある．これは自動識別が困難なので，ブラックリストを用いて除外する．

4.5.3　因果関係に矛盾するエッジ・ノードの除去

情報伝播には方向性があり，生成時間が古いブログエントリから新しいブログエントリに情報が伝播する，すなわち古いブログエントリを新しいブログエントリがリンクするという因果関係が存在する．そこで自己リンクの削除とともに，この因果関係に従わない双方向リンクや古いブログエントリから新しいブログエントリへの逆方向リンクが存在する場合には，エッジとノードの両方を除去する．

表 1　評価に用いたデータセット
Table 1　Datasets used for our evaluation.

No.	検索語	期間	エントリ	リンク	ノード	エッジ	情報源
1	iPhone	08/7/10〜17	10,801	25,836	646	772	45
2	毎日新聞	08/8/11〜25	3,863	13,857	350	508	20
3	Google Chrome	08/9/1〜9	1,926	6,764	649	772	27
4	Doblog	08/9/28〜09/5/1	1,019	4,288	211	250	13
5	地デジカ	09/4/28〜5/12	1,179	6,025	418	583	22

ブログは日記的要素が強く，通常は新しいブログエントリを追加する形式で内容が更新されるが，既存のエントリを変更した場合には，この因果関係を必ずしも満たさなくなる．**表 1** のデータセットについて調べると，双方向リンク・逆方向リンクを合わせた存在確率が 0.089〜0.35%であり，スパムを目的とした機械的ハイパーリンク生成の場合が多いことを考慮すると，無視しても特に問題がない．この因果関係を受け入れることで，ブログの情報伝播ネットワークは無閉路有向グラフとなり，処理が簡単になる．

5.　情報伝播特性の定量化とランキング

あるトピックに関する各情報源のノードの影響範囲を特定し，その部分ネットワーク内の情報伝播に関する 3 種類の基本構造に基づいて情報源の特性を定量化し，それぞれ異なる種類の情報を優先するランキングを実現する．

5.1　情報源の影響範囲の特定と部分ネットワークへの分割

情報伝播ネットワークは均一ではなく，各情報源が周囲に与える影響の違いに応じて各部が異なる構造を持つ．そこで，情報伝播特性を詳細に調べる場合には，各情報源の影響範囲ごとに分割して，個々の部分ネットワークを分析する必要がある．

一般に，ネットワークを分割する方法として，連結成分ごとに分解する方法，つながりが弱いエッジ部分を切断する方法，密な部分を抽出する方法が使われる．しかし，本論文では，情報源から情報伝播が始まり，情報伝播ネットワークが無閉路有向グラフであるという 2 つの特徴を利用して，各情報源ノードを始点とし，有向エッジをたどって到達できる範囲までを，その情報源の影響範囲の部分ネットワークとして分割する．

ここで，情報伝播ネットワークを $G = (V, E)$（V はノード集合，E はエッジ集合），V に含まれる n（$= |V|$）個のノードを v_i（$i = 1, ..., n$），E に含まれるノード v_i からノード v_j への有向エッジを e_{ij}（$j = 1, ..., n$）とする．情報源であるノード v_k（$1 \leq k \leq n$, $d'_{out}(v_k) \geq T$）からたどることができる部分ネットワーク $G_k = (V_k, E_k)$ のノード集合 V_k

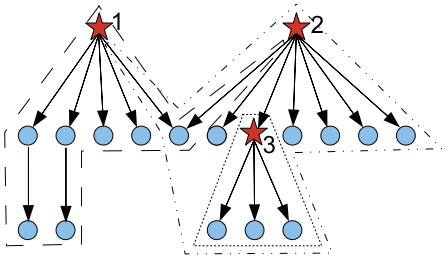


図 2　情報源と部分ネットワークの関係
Fig. 2　Relationship between information sources and their subnetworks.

とエッジ集合 E_k を次のように定義する．

$$V_k = \{v_i | v_i = v_k \vee directed\_path(v_k, v_i)\}, \quad (1)$$

$$E_k = \{e_{ij} | v_i \in V_k \vee v_j \in V_k\}. \quad (2)$$

ここで，$d'_{out}(v_k)$ は異なるサーバへの出次数（ハイパーリンクネットワークでは入次数）であり，$directed\_path(v_k, v_i)$ は，ノード v_k からノード v_i の間に有向パスが存在することを示す．E_k では，エッジ e_{ij} の始点 v_i が $v_i \in V_k$ であるならば終点 v_j に対してつねに $v_j \in V_k$ が成り立つが，v_j において他の情報源からも同時に情報が伝播する影響も考慮するために $v_i \notin V_k$ でも $v_j \in V_k$ である場合も含むことに注意する．

この情報源と部分ネットワークの関係を**図 2** に示す．星印の 3 つの情報源は，それぞれ点線で囲まれた対応する部分ネットワークを持つ．ただし，この部分ネットワークには，情報源からたどれるエッジだけでなく，到達可能なノードに入るエッジも含む．また，図 2 の情報源 1 と情報源 2 のように，2 つの部分ネットワークに重複部分が存在したり，情報源 2 と情報源 3 のように，ある情報源の部分ネットワークが別の情報源の部分ネットワークに含まれたりすることもある．

5.2　情報伝播ネットワークの基本構造

情報伝播ネットワークおよびその部分ネットワークを構成する基本単位として，2 エッジ部分グラフに着目する．2 エッジ部分グラフとは，グラフを構成する任意の 2 エッジで構成されるグラフであり，特に 2 エッジが連結している場合を 2 エッジ連結部分グラフと呼ぶ．

無閉路有向グラフの場合には，**図 3** に示すように，それが含む任意の有向 2 エッジ連結部分グラフは同一ノードを始点とする場合（図 3 (a)），同一ノードを終点とする場合（図 3 (b)），同一ノードをそれぞれ始点と終点とする場合（図 3 (c)）の 3 種類に分類でき，それぞれ情報の拡散，情報の集約，情報の転送という情報伝播ネットワークにおける基本構造を表す．本論文では，それぞれ情報拡散構造，情報集約構造，情報転送構造と呼ぶ．

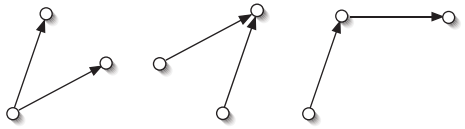


図 3　3 種類の 2 エッジ連結部分グラフ
Fig. 3　Three types of directed 2-edge connected subgraphs.

5.3　情報伝播特性の定量化

情報源であるノード v_k（$1 \leq k \leq n$）からたどれる部分ネットワーク $G_k = (V_k, E_k)$ に含まれる各基本構造の数である情報拡散構造数 $N_s(G_k)$，情報集約構造数 $N_g(G_k)$，情報転送構造数 $N_t(G_k)$ は，2 エッジ部分グラフの接続点であるノード v_i の入次数 $d_{in}(v_i)$ と出次数 $d_{out}(v_i)$ から，次のように求められる．

$$N_s(G_k) = \sum^{v_i \in V_k} \frac{d_{out}(v_i) \times (d_{out}(v_i) - 1)}{2}, \quad (3)$$

$$N_g(G_k) = \sum^{v_i \in V_k} \frac{d_{in}(v_i) \times (d_{in}(v_i) - 1)}{2}, \quad (4)$$

$$N_t(G_k) = \sum^{v_i \in V_k} d_{in}(v_i) \times d_{out}(v_i). \quad (5)$$

すなわち，あるノード v_i について考えた場合に，情報拡散構造数は v_i を始点とする 2 本の出エッジの組合せ数，情報集約構造数は v_i を終点とする 2 本の入エッジの組合せ数，情報転送構造数は v_i を中間点とする入エッジと出エッジの組合せ数である．

異なる部分ネットワークを互いに比較できるように，部分ネットワーク G_k のノード数 $|V_k|$ で正規化して，以下のように情報拡散度 $P_s(G_k)$，情報集約度 $P_g(G_k)$，情報転送度 $P_t(G_k)$ を定義する．

$$P_s(G_k) = \frac{N_s(G_k)}{|V_k|}, \quad (6)$$

$$P_g(G_k) = \frac{N_g(G_k)}{|V_k|}, \quad (7)$$

$$P_t(G_k) = \frac{N_t(G_k)}{|V_k|}. \tag{8}$$

5.4　多面的ランキング

本論文では，情報伝播ネットワークの各情報源の特性の定量化指標である情報拡散度 $P_s(G_k)$，情報集約度 $P_g(G_k)$，情報転送度 $P_t(G_k)$ をランキングに使用する．

情報拡散度は，情報源から発信された情報がどの程度注目されているかを示す．被リンク数と比較的性質が近いが，情報拡散度は情報伝播経路の下流の注目度も考慮する点が異なる．

情報集約度は，情報源から発信された情報が，複数の情報を引用して比較・議論するようなブログエントリからどの程度参照されたかを示す．

情報転送度は，情報源から発信された情報がどの程度広く多段に伝播したかを表す．ただし，ある情報源 A の存在を別のブログエントリ B を読んで知ることは頻繁にあるが，ブログエントリ B が情報源 A にない有益な差分情報を与えない場合は，ブログエントリ C では情報源 A に直接リンクするショートカット現象が生じる．つまり，単に情報が伝播するだけでなく，途中で差分情報が付け加わるほど値が大きくなる．

本論文では，ユーザが情報探索の目的に合わせてランキングに用いる定量化指標を切り替えることで，情報源の多面的ランキングを実現する．今回作成した試作システムでは，ランキング表示ウィンドウを開くと，まず最初に情報源を情報拡散度順で表示する．さらに，ユーザが情報拡散度を選択すれば権威があり注目度が高いような情報が，情報集約度を選択すれば文章を書くときに役立つような客観的な資料性の高い情報が，情報転送度を選択すれば主観的で議論を活発にしたり口コミで伝わりやすい情報がランキング上位に来るので，個々の情報の個別に表示したり，ランキング上位 9 件を一括表示したりすることができる．

ただし，情報集約度と情報転送度は，あまり話題が盛り上がらないようなトピックでは値が 0 になることがあり，これは特にショートカット現象が生じる情報転送度の場合には顕著である．そこで情報集約度または情報転送度の場合には，まず値で比較し，値が同じ場合にさらに情報拡散度で比較して順位を決定する，2 段階のマルチレベルランキングを行う．

6. 評　　価

6.1　データセット

評価には，表 1 に示すように，iPhone 3G 発売，毎日新聞の低俗記事問題の新事実発覚，Google Chrome 公開，Doblog 障害・閉鎖，地デジカ著作権問題というイベントを意図した検索語を用いて収集した 5 個のデータセットを用いた．情報源の異なるサーバからの被リンク数の閾値は，スパムブログの影響を最小限とするために，各データセット共通で $T = 10$ とした．表 1 の各欄は，収集に使用した検索語，収集したブログエントリの生成期間，ブログエントリ数，ブログ本文中の総ハイパーリンク数（ブラックリストにより自動挿入リンクは除去済み），情報伝播ネットワークのノード数，エッジ数，情報源数を示す．ただし，スパムブログや目的のトピックに合致しないブログエントリも若干残り，また異なる URL でも実体は同じサービスは別々に扱われている．

なお，2 個の 2.8 GHz Quad-Core Xenon を持つ Mac Pro を用いて，一番大きいデータセット 1 の処理にかかる時間は，ディスク上のリンク情報ファイルを読み込んで情報伝播ネットワークを抽出するまでにかかった時間は 46 秒であり，さらにすべての情報源の情報拡散度・情報拡散度・情報転送度を計算してから情報拡散度で情報源をランキングするためにかかった時間は 87 ミリ秒であった．なお前者の処理はマルチスレッド化されていない部分も残っており，さらなる高速化も可能である．

表 2　順位相関係数の平均値と分散
Table 2　Average and variance of rank correlation coefficient.

	出次数		情報拡散度		情報集約度	
情報拡散度	0.966	(0.00129)	—		—	
情報集約度	−0.00672	(0.196)	−0.0669	(0.166)	—	
情報転送度	0.619	(0.0581)	0.503	(0.0641)	0.281	(0.104)

6.2　ランキングの相関の分析

各ランキング結果の類似度合いを，次のスピアマンの順位相関係数 ρ を使って調べる．

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i}{N(N^2-1)} \tag{9}$$

N は要素数，d_i は順位差である．なお，3 種類の情報伝播特性と比較するために，既存のリンク解析によるランキングによく用いられる出次数も用意した．

各ランキング手法の組のすべてのデータセットに対する順位相関係数の平均値を**表 2** に示す．括弧内の値は分散である．これから，出次数と情報拡散度の間には強い相関があるが，出次数と情報転送度，情報拡散度と情報転送度の間には中程度の相関があることが分かる．この結果は，情報転送度，情報拡散度，情報転送度のランキング結果が異なっていることを示している．

6.3　情報源の抽出精度と分類結果

データセットからの情報源の抽出精度と抽出された情報源の種類によるランキング結果の違

表 3　情報源の抽出精度と分類結果
Table 3　Classification and precision of information sources.

	1	2	3	4	5
オフィシャルサイト	11	3	11	1	4
ニュース	22	5	14	7	10
CGM	4	8	2	3	5
その他（適合）	0	1	0	0	0
その他（不適合）	8	3	0	2	3
合計	45	20	27	13	22
抽出精度	0.82	0.85	1.0	0.85	0.86

いを調べるために，オフィシャルサイト，ニュース，CGM（Consumer Generated Media），その他の適合情報源および不適合情報源の 5 種類に分類した．オフィシャルサイトとは，トピックに該当する企業や組織などの Web サイトであり，特にサーバの最上位ディレクトリまたは該当コンテンツの格納ディレクトリのインデックスページの URL とした．ただし，データセット中に 1 つだけとは限らない．ニュースは，新聞社とオンラインニュースが公開しているニュース記事や，オフィシャルサイトのプレスリリースである．CGM は，ブログ，Wiki，掲示板に代表されるような，ユーザによって作成された情報源である．その他の情報源は，適合する場合は検索サービスへのリンクであり，適合しない場合は検索語と無関係なページやスパムなどである．

この情報源の抽出精度と分類結果を，**表 3** に示す．情報源抽出精度の平均は 0.88 である．精度低下の要因は，本文抽出誤りや機械的な本文中へのリンク挿入が原因による情報源の誤抽出と，スパムブログだった．たとえば，データセット 1 の 8 個はスパムであり，情報源とブログエントリによる二部グラフ構造を持っていた．また，分類内容を調べると，ブログ空間から注目されている情報源はオフィシャルサイトとニュースの割合が多い．これらと CGM の比率はトピックによって大きく変化し，ブログ空間でそのトピックに関する議論がどの程度盛り上がったかを反映していると推測される．

6.4　ランキング手法の特性分析

各ランキング手法の特性を分析するために，Web 検索や QA システムの結果の質の評価に用いられる MRR（Mean Reciprocal Rank）[11] と MAP（Mean Average Precision）を用いた．通常は MRR と MAP では質問と正解の評価用データに対して性能を評価するが，本論文では質問（検索語）に対するランキング結果に対して 3 種類の情報源をそれぞれ正解と見なした場合の 3 つの評価用データを作成し，それぞれの性能を比較することでランキング手法の特性の違いを分析した．

表 4　MRR の結果
Table 4　Results of MRR.

	出次数	情報拡散度	情報集約度	情報転送度
オフィシャルサイト	1	1	0.384	0.5
ニュース	0.9	0.9	0.653	1
CGM	0.356	0.356	0.814	0.86

MRR は，各課題の適合文書が最初に見つかった順位の逆数を全課題に対して平均した値で，次の式で表される．

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{r_i} \tag{10}$$

Q は課題数，r_i は課題 i で最初に見つかった適合文書の順位である．

MAP は，課題ごとに各適合文書が見つかった順位における精度の平均（平均精度）[12] を求めて，さらに全課題に対して平均した値で，次の式で表される．

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\sum_{j=1}^{N}x_j}\sum_{k=1}^{N}x_k\sum_{l=1}^{k}\frac{x_l}{k} \tag{11}$$

Q は課題数，N は調べるランキング結果数，x_i は第 i 位の文書の適合・不適合を表す変数で，適合ならば $x_i = 1$，不適合ならば $x_i = 0$ とする．$\sum_{j=1}^{N} x_j$ は適合文書数であり，さらに検索結果の 10 位までを評価対象（$N = 10$）とした．

一般的に MRR は一番高い順位だけを評価するが，MAP は正解の網羅性を考慮することから，両者を組み合わせて両面から評価する．

表 1 に示した 5 つのデータセットに対して，情報源の種類がオフィシャルサイト，ニュース，CGM である場合に正解と見なした場合の出次数，情報拡散度，情報集約度，情報転送度の 4 種類の MRR と MAP の結果を**表 4** と**表 5** に示す．

MRR は，出次数と情報拡散度はほぼ同じ傾向を示すが，情報集約度，情報転送度ともにかなり異なる傾向を示した．上位になりやすい情報源の種類は，出次数と情報拡散度ではオフィシャルサイト，情報集約度では CGM，情報転送度ではニュースと，顕著に分かれている．これは，情報集約度が高いとは被リンク数を多く集める情報であること，情報集約度が高いとは同時に参照される議論になりやすい情報であること，情報転送度が高いとは

表 5　MAP の結果
Table 5　Results of MAP.

	出次数	情報拡散度	情報集約度	情報転送度
オフィシャルサイト	0.654	0.654	0.0894	0.314
ニュース	0.473	0.473	0.234	0.669
CGM	0.144	0.144	0.350	0.417

より広範囲に広がる情報であることを考えると，違いが生じる理由は直感的に理解できる．ただし，情報伝播特性の値に関係するのは，あくまで情報自身の性質であり，オフィシャルサイト，ニュース，CGM という種類の違いは，ある性質の情報はどの種類であることが多いという二次的なものにすぎない．

MAP も，MRR と似た傾向を示すことから，出次数と情報拡散度ではオフィシャルサイト，情報集約度では CGM，情報転送度ではニュースが，最上位に限らず全体的に上位にランキングされる傾向があることが分かる．

なお，情報集約度は他と比べると全体的に低い値を示すが，これはデータセット 1 の情報集約度を用いてランキングした場合の結果の上位をスパムが占めていたからであった．つまり，機械的に作られたスパムブログが持つ同じ Web サイト群を複数のブログエントリがリンクする構造は，まさに情報集約構造であり，ずばぬけて高い情報集約度を示すという問題がある．適切なランキングのために，さらなるスパムブログの除去が必要である．

6.5　ランキングの順位差の分析

出次数と情報拡散度は MRR，MAP ともにほぼ同じ結果であるが，詳細に調べると，ランキング上位はほぼ同じだが，MRR や MAP の対象外となる下位では順位に違いが見られた．これは，被リンク数が非常に多い情報源では，情報源が直接持っている情報拡散構造の影響がほとんどだが，被リンク数がそれほどではない情報源では，その情報伝播経路の下流に存在する情報拡散構造の影響が相対的に大きくなるからである．

そこで，どの種類の情報の場合に情報伝播先でも情報拡散構造が生まれやすいかを調べるために，各データセットでオフィシャルサイト，ニュース，CGM のそれぞれで，順位差がある文書で，どの程度の順位差があるかを調べた．以下のように，情報源の種類ごとに，出次数と情報拡散度の順位の差が生じた場合の順位差の平均をとった．これを ARD（Average Rank Difference）と呼ぶことにし，さらに全データセットの平均もとった．

表 6　ARD の結果
Table 6　Results of ARD.

	1	2	3	4	5	平均
オフィシャルサイト	−6.6	−3	0.3	0	0	−1.85
ニュース	1.46	0	−0.142	−0.2	0	−0.224
CGM	2	0.333	0	−0.5	0	0.367

$$ARD = \frac{\sum_{i=1}^{N} d_i}{\sum_{i=1}^{N} \theta(|d_i|)} \tag{12}$$

N は調べるランキング結果数，d_i は順位差，$\theta(x)$ は x が 0 以下の場合 0，x が 0 より大きい場合 1 を返すステップ関数を表し，$\theta(|d_i|)$ は順位差がない場合は 0，順位差がある場合は 1 である．ただし，順位がまったく同じ場合は 0 とする．

各データセットに対するオフィシャルサイト，ニュース，CGM の ARD の値と平均値を，**表 6** に示す．この結果から，ランキング上位の結果が同じでも，ランキング下位においては，情報拡散度は出次数よりもオフィシャルサイトの順位が低くなる傾向があり，CGM の順位が高くなる傾向があることが分かる．これは出次数の代わりに情報拡散度を用いた場合には，あまり注目されていないオフィシャルサイトと，被リンク数は多くなくても広く影響を与える CGM の間に順位の逆転が起こっていることを示している．

7. おわりに

本論文では，ある検索語でブログを検索した結果から，ブロガが注目している情報源を起点とする情報伝播ネットワークを抽出し，各情報源からたどれる部分ネットワークの 3 種類の情報伝播特性を使い分けることで，ブログの情報源を多面的にランキングする手法を提案した．さらに，ランキングに使用する情報伝播特性によって，オフィシャルサイト，ニュース，CGM のような種類が異なる情報源が高く評価されることを示した．

これは，注目されている情報，資料性の高い情報，口コミで伝わりやすい情報など，ユーザの情報探索の指向に合わせてランキング結果を変化させるための基礎となる技術である．ただし，本論文では，情報伝播特性の性質を調べる目的もあり，情報集約度と情報転送度についてマルチレベルランキング化するだけでランキングに使用したが，実際のランキングにおける他の手法との組合せについては，さらなる検討の余地がある．

現在は情報源の他のサーバからの被リンク数を用いてスパムブログを除去しているが，そ

れだけではまだ十分ではなく，特に抽出時の閾値 T を下げたときにスパムブログが多く抽出され，特に情報集約度に無視できない影響を与える．そこで機械生成されるスパムブログが持つリンク構造の特徴を用いたスパムブログ排除機能の実装が必要である．

参 考 文 献

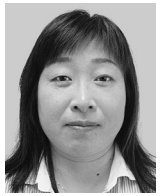
1) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
2) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. 7th International Conference on World Wide Web*, Brisbane, Australia, pp.107–117 (1998).
3) 安田　雪：ネットワーク分析—何が行為を決定するか，新曜社 (1997).
4) Watts, D.J. and Strogatz, S.H.: Collective dynamics of 'small-world' networks, *Nature*, Vol.393, No.4, pp.440–442 (1998).
5) Heider, F.: *The Psychology of Interpersonal Relations*, John Wiley & Sons, Inc., New York (1958).
6) Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U.: Network Motifs: Simple Building Blocks of Complex Networks, *Science*, Vol.298, No.5594, pp.824–827 (2002).
7) Adar, E. and Adamic, L.A.: Tracking Information Epidemics in Blogspace, *WI '05: Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pp.207–214 (2005).
8) Kim, J.W., Candan, K.S. and Tatemura, J.: Efficient Overlap and Content Reuse Detection in Blogs and Online News Articles, *WWW '09: Proc. 18th International Conference on World Wide Web*, pp.81–90 (2009).
9) Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A.: Information Diffusion Through Blogspace, *Proc. 13th International Conference on World Wide Web*, ACM, pp.491–501 (2004).
10) 総務省情報通信政策研究所調査研究部：ブログの実態に関する調査研究の結果 (2008). http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html
11) Voorhees, E.M.: The TREC-8 Question Answering Track Report, *Proc. 8th Text Retrieval Conference*, pp.77–82 (1999).
12) Buckley, C. and Voorhees, E.M.: Evaluating Evaluation Measure Stability, *SIGIR '00: Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.33–40 (2000).

(平成 21 年 12 月 20 日受付)
(平成 22 年　2 月 23 日採録)

(担当編集委員　渡辺 知恵美)

風間　一洋（正会員）

1988 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．2005 年京都大学大学院情報学研究科システム科学専攻博士課程修了．博士（情報学）．現在，NTT 未来ねっと研究所主任研究員．Web 情報検索，Web マイニングの研究に従事．人工知能学会，日本ソフトウェア科学会，ACM 各会員．



今田　美幸

1990 年 3 月横浜国立大学大学院教育学研究科修士課程修了．同年日本電信電話株式会社入社．2005 年 3 月早稲田大学大学院情報生産システム研究科博士課程修了．博士（工学）．現在，NTT 未来ねっと研究所勤務．プライバシー保護データマイニング，ユビキタスネットワーキングサービス，ネットワークセキュリティ，フォールトトレラントシステムの研究に従事．電子情報通信学会会員．

柏木啓一郎

2006 年 3 月早稲田大学理工学部卒業．2008 年 3 月早稲田大学大学院理工学研究科修士課程修了．2008 年 4 月日本電信電話（株）入社．現在，NTT 未来ねっと研究所社員．分散システムにおけるアクセス制御の研究と精度保証付き数値計算の研究に従事．電子情報通信学会会員，日本応用数理学会各会員．