ブログ空間の情報伝播の特性の定量化

Quantification of Information Diffusion Characteristics in Blogspace

風間 一洋　今田 美幸　柏木 啓一郎
Kazuhiro KAZAMA　Miyuki IMADA　Keiichiro KASHIWAGI

日本電信電話 (株) NTT 未来ねっと研究所
NTT Network Innovation Laboratories, Nippon Telegraph and Telephone Corporation

This paper propose new quantification measures of information diffusion characteristics in blogspace. These measures are calculated from the number of three basic structures, which are pairs of directed edges, in information diffusion networks. Each basic structure is related to information scattering, information gathering or information transmission. We analyze and visualize information diffusion networks extracted from six blog datasets. In the result, we show that the difference of information diffusion characteristics can be discriminated by combination of three measures and human activity in blogspace can be explained by them.

1. はじめに

あるトピックに関するブログエントリの本文中のハイパーリンクに着目すれば，複数の情報源からの情報伝播経路を抽出できる．そのネットワーク形状の違いは，ブログ空間内のアクティビティを反映しているが，これらを定量化するための適切な指標が存在しない．本稿では，情報伝播の特性を表す指標として，有向エッジ対から成る情報拡散，情報集約，情報転送の 3 種類の基本構造を用いた定量化手法を提案し，実際の情報伝播ネットワークに適用して特性を分析する．

2. 情報伝播ネットワーク

2.1 ブログ空間内の情報伝播

本稿では，ブログ空間内で発生している現象を分析するために，情報伝播に着目する．

ブログの情報は一般の Web ページと比べると書かれている内容の即時性が高く，ブログエントリを長期間書き直し続けるよりも，新たな情報は新たなブログエントリとして書く傾向がある．何らかのイベントが発生した場合には，それに関する情報をネットワーク上の有用な情報源（ブログに限らず，新聞やニュースサイトの記事や，オフィシャルサイト，まとめサイトの情報なども含む）から入手し，自分のブログに書いて他の人にも伝えるという行為が連鎖的におこなわれてブログ空間内を情報が伝播する．

この場合には，同時に有用な情報源をリンクして，その内容を紹介したり，自分の意見を付け加えることが多いので，記事中のハイパーリンクに着目すれば情報伝播経路を抽出できるはずである．

しかし，ブログは CMS（Content Management System）の一種であり，技術的な知識がなくても容易に情報発信できる反面，さまざまな情報を参照しやすく配置したり，収入源となる広告を表示したり，自社サービスを紹介したりするために，ページ中にさまざまな目的のハイパーリンクを大量に埋め込むので，内容には直接関係がないハイパーリンクも多く，ハイパーリンクネットワークから実際の情報伝播経路に使われた部分ネットワークを特定・分離することは難しい．

連絡先: 風間 一洋，NTT 未来ねっと研究所，〒 180–8585
東京都武蔵野市緑町 3–9–11，kazama@ingrid.org

2.2 情報伝播ネットワークの抽出

そこで，まず特に注目されている情報源を発見し，そこから記事の本文中に存在するハイパーリンクを逆順に辿ることで，情報伝播経路の抽出を可能にする．詳しくは別論文 [風間 08] を参照して頂きたい（ただし，無閉路有向グラフ化など若干の変更点がある）．

2.2.1 ブログエントリの検索と収集

分析対象をある特定のイベントやトピックに対するブログエントリに限定するために，そのイベントのトピックを表す検索語を用いて，Yahoo! Japan の Yahoo!ブログ検索と Technorati Japan のキーワード検索 API を用いて，検索結果に含まれるブログエントリを過去から未来に渡り長期間・間欠的に収集する．

2.2.2 本文の特定とハイパーリンクの抽出

異なるプログラムやブログサービスでも基本的な文書構造は似ている点に着目し，次のような領域を 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，div 要素や td 要素で囲まれた領域で，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テキスト部分の順に探して本文部分を特定し，本文中に含まれるハイパーリンクだけを抽出する．

2.2.3 URL の正規化と選別

抽出した URL を，URL の最後の “/” や “/index.html”，サーバー名の最後の “/” の有無，利用履歴情報収集のための URL をパラメータとして指定したリダイレクタの使用など，同一のリソースを示すと推測できる場合には，それらが一意になるように正規化する．

さらに，キーワードサイト，ソーシャルブックマーク（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・手動挿入された場合には参照しているブログエントリがまとまって抽出されるが，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．

2.2.4 注目されている情報源の発見

本手法では，ブログから注目されている情報源を起点として，ハイパーリンクを逆向きに辿って伝播経路を特定する．た

だし，ブログの本文部分から参照されているすべての情報源が必ずしも有用とは限らず，特定のブログからしか参照されない個人的な写真データや，ブログサービスによって本文中に自動挿入されてしまう関連サービスへのリンクなども含まれている．そこで，異なるサーバからの被リンク数を計算し，指定された閾値以上のリンク先を，注目されている情報源とする．

2.2.5 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時間的変化とそれに伴う特性の変化を調べたり，情報伝播とは無関係に機械生成されたために時間的な因果関係が矛盾しているようなスパムを除去するために，ノードとエッジの生成時刻を特定する．ノードに関しては，収集済のブログの場合はその生成時刻を使用し，そうでない場合は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．エッジに関しては，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，サーバの設定時刻が大幅に狂っている場合は，時間範囲を指定して分析対象から除外する．

2.2.6 無閉路有向グラフ化

ブログは通常新しいブログエントリを追加する形式でおこなわれるので，新しいブログエントリが古いブログエントリをリンクすることになる．この特徴を利用して，自己ループや双方向リンク，スパム目的の機械的ハイパーリンク生成など，この原則に従わないエッジやノードを除去することで情報伝播の因果関係を保証し，情報伝播ネットワークを無閉路有向グラフ化する．

2.2.7 情報伝播ネットワークの作成

次に，生成時刻の古いブログエントリから順番に，リンク先がすでにノードとして登録されていた場合にはリンク先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てから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．

2.3 情報伝播ネットワークの特徴

以上の処理で得られた情報伝播ネットワークは，通常は複数の情報源を起点ノードとして放射状にエッジが広がる形状を持つ．イベントやトピックによってはさらに複雑な形状になるが，これはそれらに対してブログ空間が反応する度合いによって，単に情報が拡散するだけでなく，集約されたり，転送される部分構造が生まれ，それによって放射状構造が複雑化したり，互いに接続するからである．

つまり，このようなネットワーク構造の特性を定量化できれば，ブログ空間内のアクティビティを推測できると考えられる．

3. 情報伝播ネットワークの特性の定量化

3.1 新たな定量化指標の必要性

ネットワークの特性を定量化するための指標は，次数，平均距離，クラスタ係数など，すでにいくつも存在するが，本稿で対象としている情報伝播ネットワークに適用した場合には，あまり適切とは言えなかった [風間 08].

この一つの理由は，本稿の情報伝播ネットワークは放射状構造の集合，およびその重ね合わせであり，平均経路長は短いものの，クラスタ係数は非常に小さく，スケールフリーネットワークを前提とした既存の定量化指標では適切に評価できないと考えられる．

次に，既存のネットワーク分析の定量化指標は，ノードに主眼を置き，それらの関係から特性を把握するものが多かったが，

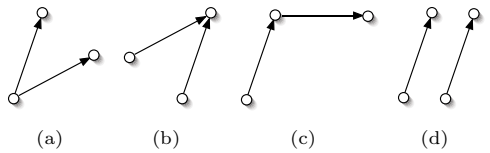


図 1: 有向エッジ対の関係

情報伝播を扱うためには，ノードではなくエッジに主眼を置き，さらにその方向性にも注目した定量化指標が必要になる．

最後に，ネットワークの成長・融合を扱うためには，連結成分に依存する手法は時間につれて分析対象が変わってしまう点であまり好ましくない．

3.2 情報伝播ネットワークの基本構造

既存のネットワークの基本構造として，クラスタ係数の三角構造 [Watts 98]，ハイダーの認知的バランス理論の三者関係 [Heider 58] などが挙げられる．しかし，これらの基本構造はノード間の関係に着目したものであり，情報伝播のような「流れ」を扱うためには，ノードではなく有向エッジ間の関係に着目する必要がある．

情報伝播ネットワークを構成する基本単位を有向エッジ対とすると，任意の 2 本の有向エッジの関係は図 1 に示すように，同一ノードを始点とする場合（図 1(a)），同一ノードを終点とする場合（図 1(b)），同一ノードをそれぞれ始点と終点とする場合（図 1(c)），ノードを共有しない場合（図 1(d)）の 4 つのいずれかに分類できる．

エッジが互いに無関係な図 1(d) を除くと，図 1(a) は情報の拡散，図 1(b) は情報の集約，図 1(c) は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している．本稿では，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．

3.3 定量化指標

情報伝播ネットワークをグラフ $G = (V, E)$（V はノード集合，E はエッジ集合），それに含まれる n 個のノードをそれぞれ $v_i (i = 0, ..., n-1)$，ノード v_i からノード v_j への有向エッジを e_{ij} とすると，グラフ G に含まれる各基本構造の数である情報拡散構造数 $N_s(G)$，情報集約構造数 $N_g(G)$，情報転送構造数 $N_t(G)$ は，有向エッジ対の接続点であるノード v_i の入次数 $indeg(v_i)$ と出次数 $outdeg(v_i)$ から，次のように求めることができる．

$$N_s(G) = \sum_{i=0}^{n-1} \frac{outdeg(v_i) \times (outdeg(v_i) - 1)}{2}, \quad (1)$$

$$N_g(G) = \sum_{i=0}^{n-1} \frac{indeg(v_i) \times (indeg(v_i) - 1)}{2}, \quad (2)$$

$$N_t(G) = \sum_{i=0}^{n-1} indeg(v_i) \times outdeg(v_i). \quad (3)$$

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

$$P_s(G) = \frac{N_s(G)}{n}, \quad (4)$$

$$P_g(G) = \frac{N_g(G)}{n}, \quad (5)$$

$$P_t(G) = \frac{N_t(G)}{n}. \quad (6)$$

これらを複数情報源を持つ情報伝播ネットワークの特性を定量化する指標として用いる．

情報拡散構造は情報を発信する場合に生成されるので，$P_s(G)$ は情報発信に関する指標である．$P_s(G)$ は情報源の注目度が高い，又は注目されている情報源が限られるほど，値が大きくなる．

情報集約構造は複数の情報源を同時に参照してブログエントリを書く場合に生成されるので，$P_g(G)$ は情報受信に関する指標である．$P_g(G)$ が大きい場合は，情報の比較や議論などが活発におこなわれているが，$P_g(G)$ が小さい場合は，単に情報を紹介しているだけの傾向が強いと考えられる．

情報伝達構造は三つのブログ間の情報伝播の連鎖を示すので，$P_t(G)$ は情報中継に関する指標である．$P_t(G)$ が大きくなる場合は，その情報をわざわざリンクしてまで他の人に知らせる価値があるとみなされていると考えられる．ただし，実際に三つのブログ間を情報が伝播していても，中間ノードが始点ノードにない有益な情報を与えない場合はハイパーリンクではショートカットされている可能性があり，情報伝達度はこの現象を反映しているはずである．つまり，$P_t(G)$ が大きくなる場合は，情報源から伝播される価値のある情報が提供されている場合に加えて，中間ノードでも有益な差分情報が提供されていてショートカットが起こらない場合だと考えられる．

4. 評価

4.1 データセット

分析には，表1に示す6個のデータセットを使用した．この表の各欄は，収集に使用した検索語，収集されたブログエントリの生成期間，収集ブログエントリ数，ブログ本文から抽出された総ハイパーリンク数，抽出された情報伝播ネットワークのノード数，エッジ数，ノード数/エッジ数，情報源数を示している．それぞれ，2008 年に起こった洞爺湖サミットの開催（7/7〜7/10），iPhone 3G 発売（7/11），Google ストリートビュー公開（8/5），毎日新聞英語版サイトの低俗記事問題に関する新事実発覚（8/12），福田首相辞職（9/1），Google Chrome 公開（9/2）という出来事を意図した検索語を用いてデータセットを収集した．2.2 の方法に従ってデータセットを抽出しており，抽出の際の情報源の被リンク数の閾値は 10 とした．なお，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っており，また Google のように異なる URL でも実体は同じような場合も別々に扱われている．

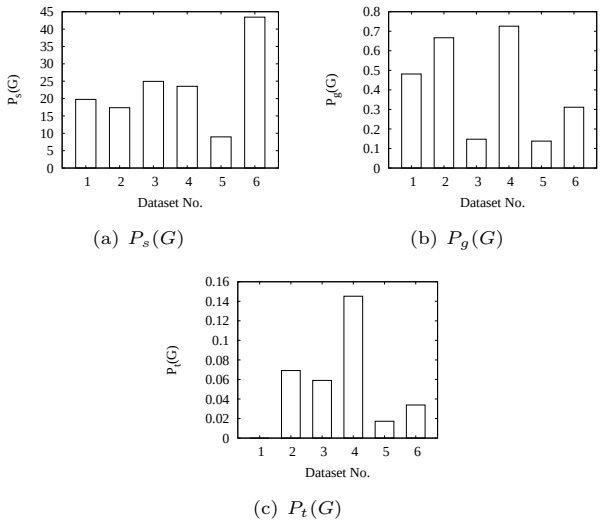
表1を分析すると，ノード数とエッジ数には 0.99 という強い相関があり，トピックの違いにあまり影響されない．しかし，検索されたブログエントリ数と抽出されたノード数の間には 0.48 と中程度の相関しかなく，単語出現頻度と情報伝播ネットワークはトピックによって異なる傾向を示すことがわかる．また，データセット 2 と 6 はノード数とエッジ数がほぼ同じでも情報源数が異なることに注意されたい．

4.2 基本構造数

各データセットの基本構造数 $N_s(G)$ ，$N_g(G)$，$N_t(G)$ を表2に示す．$N_s(G) >> N_g(G) > N_t(G)$ のような関係が成り立っている．$N_s(G)$ が一番大きい理由は，情報伝播ネットワークの基本は情報拡散だからである．$N_t(G)$ が一番少ない理由として，本手法では情報を見るだけでブログを書かない情報伝播の末端の多数の潜在的な読者を検出できないこと，実際には情報連鎖の中間ノードが新たな情報を付加しない場合にハイパーリンクがショートカットされることが考えられる．$N_s(G)$ ，$N_g(G)$，$N_t(G)$ とノード数との相関係数は，それぞれ 0.83406，0.821859，0.655873 となり，$N_t(G)$ は比較的ネットワーク規模の影響を受けにくいことがわかる．

表 2: 基本構造数

	1	2	3	4	5	6
$N_s(G)$	2629	11310	6763	8590	1041	28195
$N_g(G)$	64	434	40	265	16	202
$N_t(G)$	0	45	16	53	16	22



(a) $P_s(G)$ (b) $P_g(G)$

(c) $P_t(G)$

図 3: 各データセットの $P_s(G)$，$P_g(G)$，$P_t(G)$

4.3 可視化

次に，これらのデータセットから抽出した情報伝播ネットワークを，Pajek で 3 次元の FR（Fruchterman-Reingold）アルゴリズムを用いて可視化した結果を図2に示す．図2(a) と図2(e) は情報源を中心に一段階拡散した単純な構造であるのに対して，他はより複雑な構造を持つことがわかるが，可視化結果だけから特性の違いを識別するのは困難である．

4.3.1 情報伝播特性の評価

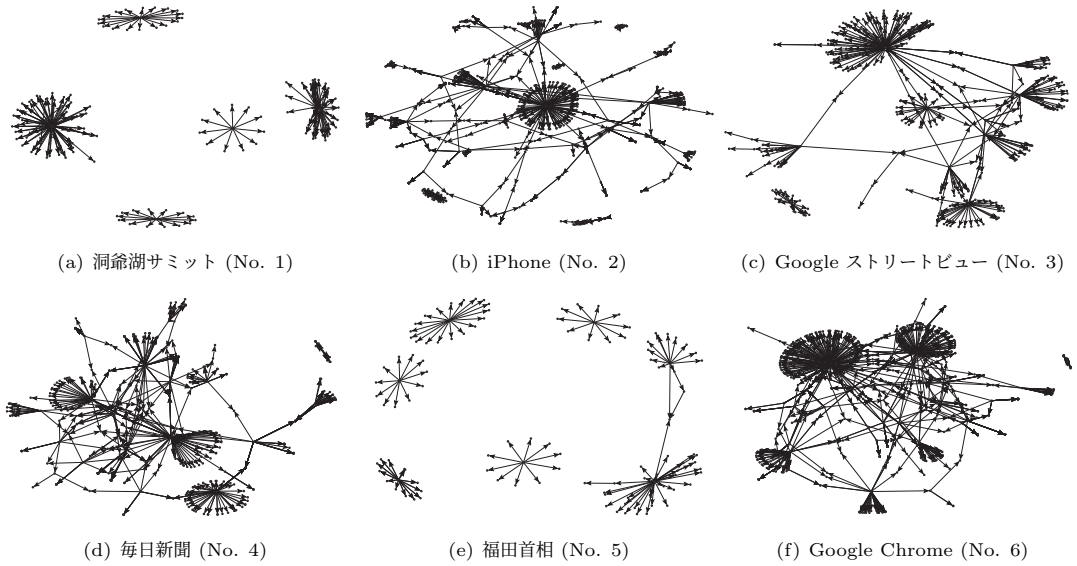
まず，各データセットの $P_s(G)$，$P_g(G)$，$P_t(G)$ の値を図3に示す．さらに，得られた各データセットの情報伝播特性を比較して，平均値の 150%より大きい場合を “+”，平均値の 50%より小さい場合を “–” として作成した表を表3に示す．この方法では，全部で 9 通りの組み合わせパターンが考えられるが，表1や図2だけでは区別がつきにくいデータセットも，別のパターンとして区別できている．

表 3: データセットの情報伝播特性

	1	2	3	4	5	6
$P_s(G)$					-	+
$P_g(G)$		+	-	+	-	
$P_t(G)$	-			+	-	

表 1: 評価に用いたデータセット

No.	検索語	期間	エントリ数	リンク数	ノード数	エッジ数	情報源数
1	洞爺湖サミット	7/8〜10	3785	7478	133	189	8
2	iPhone	7/10〜17	10801	25854	651	777	44
3	Google ストリートビュー	8/5〜25	1376	5474	271	295	10
4	毎日新聞	8/11〜25	3863	13898	365	522	21
5	福田首相	9/2〜5	2738	5435	116	123	8
6	Google Chrome	9/1〜9	1926	6764	649	772	27



(a) 洞爺湖サミット (No. 1) (b) iPhone (No. 2) (c) Google ストリートビュー (No. 3)

(d) 毎日新聞 (No. 4) (e) 福田首相 (No. 5) (f) Google Chrome (No. 6)

図 2: 情報伝播ネットワークの可視化結果

データセット 4 の場合は情報集約度 $P_g(G)$ と情報伝達度 $P_t(G)$ が高いが，これはマスコミは同業の毎日新聞の不祥事を意図的に報道しなかったために，それを補うかのようにブログ・Wiki などの CGM が新事実の報道や事実のまとめ作業をおこなったからだと思われる．

データセット 2 とデータセット 6 は，情報伝播ネットワークのノード数・エッジ数がほぼ同じでも情報源数は異なり，情報伝播特性では前者は情報集約度 $P_g(G)$ が高く，後者は情報拡散度 $P_s(G)$ が高いという違いがある．これは，前者はマスコミ，関連オフィシャルサイト（アップルの関連プロダクトサイト，ソフトバンクモバイル，Yahoo! Japan），ブログなど多くの情報源からもたらされた情報を，それぞれのブログで比較・批評している傾向が強いからであり，後者は主に Google のオフィシャルサイトから集中的に情報がもたらされているからである．

これに対して，可視化結果であきらかに未発展なネットワーク構造を持つデータセット 1 とデータセット 5 の値は全体的に小さいが，これは福田内閣が政策および政権放棄という点であまり国民に受け入れられていなかったことを示している．ただし，データセット 1 の情報拡散度 $P_s(G)$ と情報集約度 $P_g(G)$ の値がそれほど低くないのは，実は北海道洞爺湖サミット連動プロジェクトとして上演されたミュージカル「葉っぱのフレディ　—いのちの旅—」とその協賛企業，およびエコライフについてのアンケートとそのサービスにログインするリンクを同時に多数リンクしている（図 2(a) の左右）からであり，実は当初想定していた洞爺湖サミットそのものに関する情報を提供しているブログエントリは少なかった．

5. おわりに

本稿では，情報伝播ネットワークの特性を表す新しい定量化手法を提案し，実際に異なるトピックに対して抽出した 6 個の情報伝播ネットワークの可視化では判別が難しいような内部構造の違いを特性値の組み合わせパターンで識別できること，さらに特性値の大きさで情報伝播ネットワーク上のアクティビティを説明できることを示した．

今後は，現在の特性値の計算では値域が決まらないなど扱いづらい点の改良や，さらなる事例分析を行う必要がある．

参考文献

[Heider 58] Heider, F.: *The Psychology of Interpersonal Relations*, John Wiley & Sons, Inc., New York (1958)

[風間 08] 風間 一洋, 今田 美幸, 柏木 啓一郎：ブログ空間における情報伝播ネットワークの抽出と分析, in *WebDB Forum 2008* (2008)

[Watts 98] Watts, D. J. and Strogatz, S. H.: Collective dynamics of 'small-world' networks, *Nature*, Vol. 393, No. 4, pp. 440–442 (1998)