トラックバックに基づく Blog コミュニティへの PageRank 適用

2002MT064 小笠原 崇人　　2002MT079 菅沼 由貴
指導教員　河野 浩之

1　はじめに

Web の発達に伴いアクセス可能な情報量が増加することにより，有用なコンテンツを効率よく取得することが重要になっている．そこで，有効性の高い情報な情報を効率的に取得する仕組みを構築することの意義は大きい．現在 Web 空間では，さまざまな検索エンジンが使用され，より有用な情報の取得をサポートしている．最も大きな検索エンジン Google[1] で使用されている PageRank アルゴリズム [2] は，リンク構造に着目しランキングを行っている．しかし現在急速に発達してる Blog 空間のみにおけるものはない．

本研究では，Blog のエントリーの中から有用なエントリーの抽出を行う際，トラックバックに着目する．エントリーに対するトラックバック数を見ることにより，そのエントリーに対する議論がどのくらい活発に行われているかを知ることができると考える．つまりトラックバックの数が多いほど議論が多く行われており，関心の高いエントリーであると推測することができる．またエントリー間でのトラックバックのリンク構造を見ることにより，エントリーが受けている支持の流れを見ることができると考える．つまり多くのトラックバックを受けているエントリーからトラックバックを受けている事は，より有用なエントリーであると推測することができる．本研究では，この考えを基に Blog 内のエントリーのトラックバックのつながりをリンク解析により抽出し，PageRank アルゴリズムを適用させてエントリーに数値を与える事により，より有用なエントリーを抽出することを目指す．

2　Blog に対する PageRank 技術

2.1　トラックバックのリンク解析に関する研究

トラックバックにおけるリンク解析に関する研究として，中島らが動的に生成される Blog のリンク構造の解析手法の提案と，解析を基に信頼性の高い Blog の判別に関する調査研究を行っている [3]．その研究において，中島らはスレッドにおける Blog 特性の規則性として，スレッドが立ち上がった初期にエントリーを提供することが多い Blog 投稿者を Topicfinder，スレッドでの議論が盛んになる直前にエントリーを提供することが多い Blog 投稿者を Agitator，他の Blog エントリーから参照されることが多い Blog 投稿者を Opinion Leader，他の多くのエントリーを参照する事が多い Blog 投稿者を Summraizer，あるトピックス関するスレッドに対してエントリーを投稿することが多い Blog 投稿者を Fan としており，それぞれを Blog 内で見つけることによりその Blog の情況を効率よく取得できると考えている．

2.2　トラックバックの定義付け

トラックバックにはエントリーを参照している参照トラックバックと参照されている被参照トラックバックがある．**図 1** に，トラックバックにおけるリンクと逆リンクの例を示す．

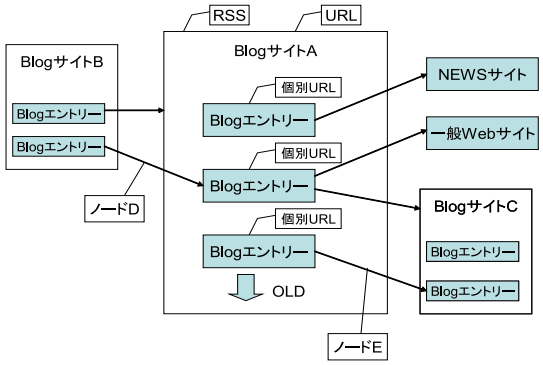

図 1　参照するエントリーと参照されるエントリー

Blog サイト A において Blog サイト B より参照されているノード D を被参照トラックバック，Blog サイト A において Blog サイト C を参照しているノード E を参照トラックバックと定義する．

2.3　PageRank[2] とは

PageRank とは，Web ページ間のリンクから Web ページのランク付けを行う手法であり，中心的なページを見つけるためのものである．その基本概念は，「有名なページは有名なページへリンクを張る」というものである． 詳しく言えば，あるページの PageRank は，そのページから発するリンクの数で割った数が，それぞれ被リンク先のページの PageRank に加算されるというアルゴリズムを繰り返して得られた物である．

もし u が Web ページとした場合に，F_u は u にリンクをされているページの集合である．また，N_u を u から出ているリンクの数（$N_u = |F_u|$）とし，c を一般化のための定数，そして u からリンクされているページ集合を B_u と定義する．この時ページ u における PageRank の値 $R(u)$ は以下の式 (1) によって計算される．

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{1}$$

3　Blog コミュニティの PageRank 適用

本研究では，トラックバックを収集・リンク解析し，PageRank を適用することによりエントリーに数値を与え，数値を基にランキングを行なう．

図 2 にシステム構成図を示す．

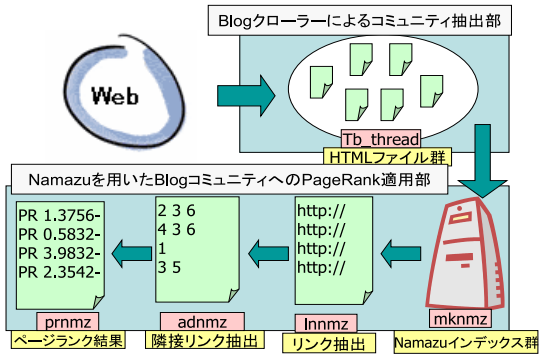

図 2　Blog コミュニティの抽出と PageRank アルゴリズムの適用

本研究における提案内容は以下の 2 つである．

- Blog クローラーによるコミュニティ抽出
- Blog コミュニティへの PageRank 適用

3.1　エントリーの収集ツール

実験の対象となるコミュニティを形成するために，我々はトラックバック抽出に特出したクローラーである Tb_thread[4] を使用する．Tb_thread とはトラックバックリンクを辿り Blog エントリーを収集し，Blog エントリーのスレッド化を視覚的に表示する Perl プログラムであり，関連記事がどのように拡散していったかを見ることが可能である． 以下のアルゴリズムを再帰的に実行する事によりトラックバックリンクの取得，表示を行う．

1. 起点となる URL を指定
2. PingURL に_mode=rss を付加して RSS を取得

Tb_thread は，ある Trackback Ping URL に対して送信された Ping のリストを、Ping URL にクエリパラメータ _mode=rss を付加することによって，RSS データをレスポンスとして取得している．

3.2　PageRank 値の算出方法

Tb_thread を用い得られた HTML 群に対して，Namazu インデックス [5] を用い PageRank の値を出すための手順を下記に示す．

1. 収集したエントリーに番号を与える (mknmz)
2. リンクを抽出 (lnnmz)
3. 隣接リンクを抽出 (adnmz)
4. PageRank 値を計算 (prnmz)
5. ランキングの表示

収集した HTML ファイルのリンクを抽出する為に Namazu の lnnmz コマンドを用い，インデックス化された検索対象の HTML ファイルに含まれるハイパーリンクを抽出する．

次に namazu の adnmz コマンドを用い，lnnmz によって抽出されたハイパーリンクより検索対象の HTML ファイル間の相互のハイパーリンク構造を抽出する．その結果作成される NMZ.field.adjacency は文書間のリンク関係を文書 ID で記したファイルで，隣接リストそのものである．得られる NMZ.field.adjacency の例を**図 3**，その遷移図を**図 4** に示す．

1	2 3 4 5 7
2	1
3	1 2
4	2 3 5
5	1 3 4 6
6	1 5
7	5

図 3　adnmz 結果例

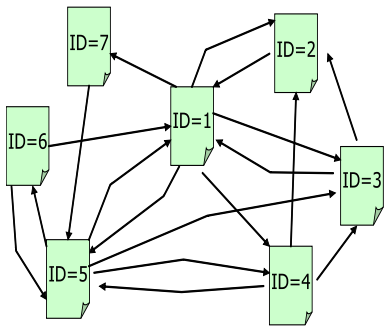

図 4　左図より得られる遷移図

抽出したリンク間の繋がりからページランクを算出するには prnmz[6] を用いる．prnmz とはページ間の繋がりを示した NMZ.field.adjacency を用い，推移状態行列の最大固有値に属する固有ベクトルを求めるプログラムである．具体例として，**図 3** にて用いた NMZ.field.adjacency の例を prnmz を実行して推移状態行列の最大固有値に属する固有ベクトルを求めている様子を**図 5** に示す．

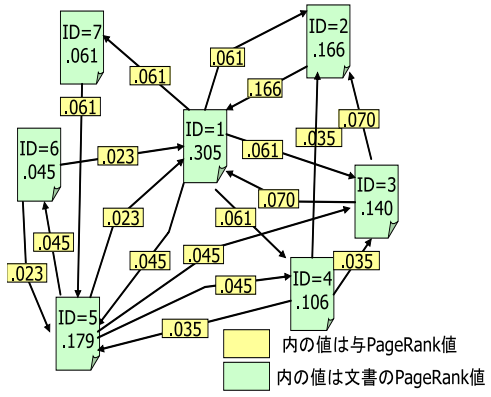

図 5　prnmz を用いた計算

以上の手順を踏まえる事により，クローラーによって収集されたトラックバックの繋がりを持つ Blog エントリー群に対して，PageRank 値を求めることが出来る．

4　実験結果

実験の対象とするエントリーの話題は，05 年 10 月に発生したパキスタン地震被災者支援，05 年 11 月に発覚

した Blog 少女母毒殺未遂事件，耐震強度偽造問題，そして日記主体の Blog の中から，横峯さくらさんの日記，眞鍋かをりさんの日記とする．

4.1 抽出コミュニティの解析

それぞれの話題を基に収集してできた集まりをコミュニティとする．収集したデータの中から Blog 少女母毒殺未遂事件について形成されたコミュニティについて見ると，コミュニティ内の総エントリー数は 97 エントリー，総トラックバック数 (総 TB 数) は 264 本，参照トラックバックを複数出しているエントリー数 (複 TB 数) は 49 エントリーあった．他の話題のコミュニティの詳細を**表 1** に示す．

表 1　抽出コミュニティ解析結果

	総エントリー数	総 TB 数	複 TB 数
パキスタン	73 個	100 本	15 個
Blog 少女	97 個	264 本	49 個
耐震偽造	148 個	204 本	30 個
横峯さくら	93 個	121 本	10 個
眞鍋かをり	69 個	81 本	6 個

4.2 解析データによる分類

解析した結果を見るとこれらのコミュニティは 2 つのグループに分ける事ができる．

1 つ目のグループは総エントリー数に比べ，総トラックバック数がかなり多く，複数参照トラックバックを出しているエントリー数の割合が高いコミュニティである．これらは収集したデータが密な繋がりを持っているといえる．1 つ目のグループに当てはまるのは 05 年 10 月に発生したパキスタン地震，05 年 11 月に発覚した Blog 少女母毒殺未遂事件，耐震強度偽造問題である．

2 つ目のグループは，総エントリー数と総トラックバック数がほとんど変わらなく複数の参照トラックバックを出しているエントリーの割合が低いコミュニティである．これらは密な繋がりを持っていない．2 つ目のグループに当てはまるのは，横峯さくらさんの日記，眞鍋かをりさんの日記である．

1 つ目のグループに分類されたエントリーの話題は，ニュースや事件など議論が比較的に盛んに行なわれる．それに対して，2 つ目のグループに分類されたエントリーの話題は，日記など議論が行なわれる事がほとんどない．本研究では 1 つ目のグループを議論型コミュニティ，2 つ目のグループを日記型コミュニティと呼ぶことにする．議論型コミュニティから Blog 少女母毒殺未遂事件のトラックバック解析と日記型コミュニティから横峯さくらさんの日記のトラックバック解析を**図 6** に示す．

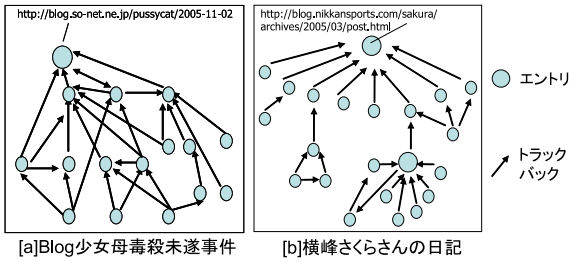

図 6　議論型コミュニティ及び日記型コミュニティのトラックバック解析図

Blog 少女母毒殺未遂事件のトラックバック解析図は，多くのエントリーが複数の繋がりを持っており，参照トラックバックの多いエントリーや，被参照トラックバックの多いエントリーなどの特徴が得られた．多くのエントリーが密な関係を持っているコミュニティは，トラックバックにより議論が活発に行なわれた結果であり，Blog ならではの特徴であると考えられる．

表 2　Blog 少女 PageRank 値上位 5 位

	エントリー名	PageRank 値	被参 TB 数 (**割合**)	支持率
1 位	「グレアムヤング毒殺日記」16 歳少女が傾倒した殺人キラーの話	0.15439	50 本 (19%)	54%
2 位	母を殺人未遂容疑の高 1 女子，ブログで動機示唆	0.11932	14 本 (5%)	15%
3 位	女子高生を逮捕 /静岡	0.07035	21 本 (8%)	22%
4 位	タリウム母親毒殺未遂、ネットで劇物情報収集か	0.06568	20 本 (8%)	21%
5 位	高 1 女子，母親に劇物？事件とグレアム・ヤング	0.06116	13 本 (5%)	14%

表 3　横峯さくら PageRank 値上位 5 位

	エントリー名	PageRank 値	被参 TB 数 (**割合**)	支持率
1 位	桜前線は今 いずこ?	0.23739	37 本 (30%)	40%
2 位	こんにちは (*^-^*) (横峯さくらの日記)	0.18720	48 本 (40%)	52%
3 位	BMW ダイナミックゴルフで横峯さくらは 13 位	0.10499	6 本 (5%)	6%
4 位	さくらの父 キャディーやめる?	0.10129	3 本 (2%)	3%
5 位	女子プロゴルファー横峯さくら	0.09013	7 本 (6%)	7%

横峯さくらさんの日記のトラックバック解析図は，起点となるエントリーにトラックバックが集中し，他のエントリー間の繋がりがほとんど見られなかった．

4.3 PageRank アルゴリズムの適用

議論型コミュニティと日記型コミュニティのエントリーに対して PageRank のアルゴリズムを適用し，PageRank 値の高い順にランキングする．得られた結果の上位 5 位のエントリー名，PageRank 値，被参照トラックバックの数，総トラックバック数に対するそのエントリーの被参照トラックバック数の割合 (割合)，総エントリー数に対するそのエントリーに参照トラックバックを送っているエントリー数 (支持率) を求めた．

まず議論型コミュニティから Blog 少女母毒殺未遂事件の結果を**表 2** に示す．上位のエントリーは被参照トラックバックを多く持ち，話題に沿ったエントリーが挙げられた．特に 1 位のエントリーは 50 本もの被参照トラックバックを持つ．これは総トラックバック数の約 20% にあたり，総エントリー数の約 50% のエントリーからトラックバックを受けている．被参照トラックバックを多く持つ事は，コミュニティ内から支持を受けているエントリーであり，有用なエントリーであると考えられる．

次に日記型コミュニティから横峯さくらさんの日記の結果を**表 3** に示す．上位のエントリーは被参照トラックバックを多く持つエントリーが挙げられたが，話題とは関連の薄いエントリーも存在した．特に 1 位のエントリーは 37 本もの被参照トラックバックを持っていたが，話題と反れたエントリーであった．また，3 位以下のエントリーが持つ被参照トラックバックは少なく，ほとんどのトラックバックが 1 位と 2 位に集中していることが分かる．

5 考察

話題によってコミュニティの構造が異なり，上位に挙がるエントリーの傾向も異なっていた．

議論型コミュニティではエントリーの繋がりが深く，複雑なリンク構造を持っていた．密なコミュニティが形成された要因は，まずその話題がとても興味深く，社会から大きな関心を示されていたこと，そして Blog の特徴の 1 つであるトラックバックにより議論が活発に行なわれたことが考えられる．このようなコミュニティの存在は Blog ならではの特徴であると考えられる．またランキング上位に挙げられたエントリーは，被参照トラックバックを多く持ち，議論の中心となっているエントリーと考えられる．議論型コミュニティではこのような議論の中心となるエントリーを抽出でき，満足な結果を得られた．

日記型コミュニティでは，起点エントリーにトラックバックが集中し，単純なリンク構造を持っていた．これは議論型コミュニティのような議論を活性化させているトラックバックとは異なり，日記エントリーの相手にメッセージを送るために出されたトラックバックが多いことが考えられる．また PageRank のランキングを見ると，1 位には横峯さくらさんとは関係のない話題のエントリーが挙がり，2 位に起点となった横峯さくらさんのエントリーが挙がった．1 位のエントリーが話題とは異なったエントリーであったように，日記型コミュニティでは議論型コミュニティとは異なり，話題に対する議論が行なわれる事がほとんどなく話題が分散しやすい特徴があると考えられる．

6 おわりに

Blog のエントリーを基にコミュニティを抽出し，コミュニティ内からより有用なエントリーの取得を目指した．議論が多く行なわれる話題のコミュニティでは，密なコミュニティ形成を発見できた．そして PageRank アルゴリズムを適用した結果，多くトラックバックを受けており，コミュニティ内の中心と思われるエントリーが高い PageRank 値を得るというよい結果を得ることができた．しかし議論があまり行なわれない話題のエントリーでは，話題とは関係の薄いエントリーが高い値を得るなどよい結果を得ることはできなかった．

今後の課題としては，有用なエントリーの抽出精度の向上が挙げられる．トラックバックのリンク構造以外に，コンテンツに注目し，コンテンツマイニングなどの技術を取り入れることで抽出精度が上がると考えられる．

謝辞

本研究を進めるにあたり，有益なアドバイスをいただいた指導教員の河野浩之先生や研究室の皆さんに深く感謝いたします．

参考文献

[1] Google,Inc: Google，http://www.google.com
[2] Lawrence Page, et al: “ The PageRank Citiation Ranking: Bringing Order to the Web ”, Stanford Digital Libraries Working Paper，(1998)
[3] 中島伸介， 舘村純一， 日野洋一郎，原 良憲， 田中克己: “ Weblog 解析に基づくコンテンツの信頼性評価の検討 ”，DBSJ Letters， Vol.3，No.1
[4] Tatsuhito Miyagawa: Trackback スレッド化，http://blog.bulknews.net/cookbook/blosxom/trackback/tb_thread.html，(accessed 2005.10)
[5] 馬場 肇:“ Namazu システムの構築と活用 ”，ソフトバンク パブリッシング株式会社 (2003.7)
[6] 馬場 肇: Google の秘密 - PageRank 徹底解説，http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html，(accessed 2005.10)