

Blog コミュニティの抽出と分析 Detecting and Analyzing Community in BLOGs

谷口 智哉^{*1}
Tomoya Taniguchi

松尾 豊^{*2}
Yutaka Matsuo

石塚 満^{*1}
Mitsuru Ishizuka

^{*1} 東京大学大学院情報理工学系研究科

Graduate School of Information Science and Technology, The University of Tokyo

^{*2} 産業技術総合研究所 サイバーアシスト研究センター

Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology

Recently there are many blog hosting services in Japan and users are getting larger and larger. The Blog users seem to make new communities on the Internet. In this paper, we extract and analyze Blog communities based on the link structure of Blog. Blog networks can be constructed as follows: a Blog (whole entries of one Blog) is considered as a node, and a link between two Blogs is considered as an edge. We try two experiments to two areas: scoring Blog pages by PageRank and clustering Blog communities by betweenness clustering for the Blog networks on “baseball” and “Winny.”

1. はじめに

1999 年に「Blogger.com」「Pita」などの Blog が容易に作れるソフトが世に出現した。それ以来、Blog は多くのネットユーザから指示をえて、急速に広がりつつある[Blood 02]。特に、日本では 2003 年から 2004 年にかけて、「ココログ」「Livedoor Blog」「JUGEM」「エキサイトブログ」「Doblog」などのブログ・ホスティングサービスが次々とスタートし、そのユーザ数も急増している[NetRatings 04]。

Blog は、特に個人にとって、優れた情報発信の形態ではないだろうか。その理由は3つある。まず、Trackback の機能により他の Blog とのつながりを多くの人に示せることであり、議論の活発化が容易である。次に、RSS を発信することによって自らの Blog の存在・内容を第三者が簡単に取得することができることである。さらに、掲示板に書き込むのと同じような手軽さで自分が所有するサイトを作ることが可能になったことであり、これにより多くの人が情報を発信することが可能となる。

本研究では、次のような仮定をおいて、議論を進めていく。

「ひとつの Blog は1人の人間に対応し、したがって Blog 間の関係から、人間同士の何らかの関係を導くことができる。」

多くの Blog は、ひとつの Blog がひとりの人によって書かれている。そして、Blog 間でリンクや Trackback をひくのは意見の交換をしていることにあたると考えられる。Blog は、さまざまなトピックス・分野において、当人のそれまでの文脈や周りとの関係がある上で、その人の意見を読むことができるという特徴がある。したがって、Blog 間にどういう関係があり、それが全体としてどういう構成をしているのか、現実には存在する社会ネットワークと対比してどのような特徴があるのかといった Blog が織り成すコミュニティの抽出および分析は、あるトピックや分野において、どういう文脈でどういう意見が交わされているかを知る上で重要な課題である。

これまで、Web ページのリンク構造に着目し、コミュニティを抽出する研究は多く行われてきた。村田らは、Web ページのリンク

の共起性に基づいてコミュニティを抽出する研究を行っている[村田 01]。また、よい hub ページからリンクされている authority ページ、よい authority ページにリンクしている hub ページを定義し、hub と authority の2分グラフという形でコミュニティを抽出する手法もある[Kumar 99]。また、外部よりも内部にリンクが多いページ集合をもってコミュニティとする研究もある[Flake 00]。

本研究では、こういったアイデアをそのまま Blog に適用し、Blog ページの間のリンク構造から Blog の重要度を算出するとともに、コミュニティを抽出する。Web マイニングのアプローチを用いることができるか、有効かどうかを示すことを目的とする。具体的には、Blog 間のリンク構造から、PageRank[Page 98]を用いた authority Blog の抽出、および弱いリンクを切ることでコミュニティを出現させる Betweenness Clustering[Michelle 01]を行う。

Blog には、Movable Typeをはじめとするスクリプトをサーバに設置したもの、ホスティングサービスを利用したものなどがあるが、本稿では、特にインターネット上に新たに登場する人々のコミュニティを発見したいと考え、そのような人が多いと考えられるブログ・ホスティングサービスによる Blog を対象に考える。これらのブログ・ホスティングサービスは、かつサーバの用意や設定などが必要なく、文章入力と GUI による簡単な操作だけで、すぐにでも自分の Blog を作ることができるで、今まで特別な知識を持っていないといった理由により、インターネット上での情報発信のができなかった人でも情報を発信することができるからである。

2. Blog 間のリンク構造の抽出

2.1 抽出・分析の対象

Blog からは多数のリンクが張られているが、Blog を個人の情報発信の場と考えると、その核となるものは Blog の本文であろう。ここでいう本文とは、本人が Blog にエントリ(記事)を書いたときの、文書そのものを言う。この文書は HTML 形式で表現されており、ハイパーリンクにより他のページとのつながりを持っている。本稿では、この本文中のハイパーリンクを収集することにする。

Blog には、ハイパーリンクを多く張っている Blog と、少ない数しかリンクを張っていない Blog とが存在するが、同様にハイパーリンクを収集すると、Blog ごとによってのリンクを張る頻度に

よって左右されてしまう。インターネット上に新しく登場する人がリンクを多数張るとは考えにくいので、単にリンクを収集すると、その人たちのコミュニティの発見を妨げてしまうと考えられる。よって、Blog ひとつから張られリンクは、回数の多い上位3件だけとする。ただし同位が複数ある場合は、その全てのリンクが張られているとみなした。

対象としたブログ・ホスティングサービスは、「はてなダイアリー」「JUGEM」「Livedoor Blog」「ココログ」「エキサイトブログ」「goo Blog」「Doblog」「Seesaa」「Myprofile」の9つである。2004年6月29日の時点で、Bulkfeedsによると、これらの9つのブログ・ホスティングサービスに対応することにより、約80%のBlogに対応することになる[bulkfeeds]。

2.2 収集方法

収集方法は、あるBlog10個をSeed Blogとし、そのSeed Blogについてリンクを収拾し、そのリンクでつながれているBlogを新たなSeed Blogとする、という作業を5回繰り返した。

具体的には、最新の記事のアドレス(URL)をRSSから抽出し、最新の記事から過去の記事にさかのぼって、そのBlogからの全ての本文中からのリンクを収集した。アルゴリズムを示す。ただし、Source(x)はアドレスxのソースを表しExtract_next_address(y)はソースyから、一つ前に書かれた記事のアドレスを抜き出す。dは深さ、Bはクロールの対象になっているBlogの集合である。

1. [Seed Blogの設定] $B := \text{Seed Blog}$, $d:=1$.
2. [取得するBlogを決める] $b \in B$ である最新のBlogのアドレス b を取り出す。 $B:=B/b$
3. [RSSを見る] Blog b のRSSを取得し、そのうち最新のエントリのURLを p とする。
4. [エントリーの取得] Source(p)を取得する。その中に含まれている他のBlogへのリンクを取り出し、カウントをインクリメントする。
5. [前のエントリー] $p := \text{Extract_next_address}(\text{Source}(p))$. $p=\text{null}$ なら6へ。そうでなければ4へ。
6. b から張られた最も多いリンクの上位3件をエッジと見なす。 $d<5$ なら上位3件のBlogを B に加える。 $B=\{\}$ なら終了。そうでないなら、 $d:=d+1$ として、2へ。

はじめのSeed Blogを変えることにより、それぞれの特徴を持ったコミュニティに関するBlogの収集が期待できる。

Seedとしては、次の2つを選択した。ひとつめは、「Blog」と「野球」を含むWebページをGoogleで検索し、そのうち2章で示したホスティングサービスに適合するもののうち上位10件をSeed Blogとし、ふたつめは、同様に「Blog」「Winny」を含むBlogをSeed Blogとした。

2.3 収集方法の工夫点

Blogの本文中のリンクを収集するのに、大きく分けて二つの工夫しなければならない点がある。

ひとつは、Blogの本文を特定することが難しいことである。しかし、本稿では、各ブログ・ホスティングサービスごとに、ある程度HTMLのソースのテンプレートが決まっていることに注目し、そのテンプレートに対応して本文の部分だけを抜き出してきている。

もうひとつは、最新の記事から過去の記事のアドレスを発見することである。これもHTMLソースが決まった形で書かれていることを利用して過去のアドレスを抽出している。また各Blogには、過去の記事を順に閲覧できるようにリンクが張られているので、それを利用して時系列を遡るように記事を収集できる。これにより、新たな記事が追加されても、新たな記事だけを収集することが可能である。

3. Blog間のリンク構造の解析手法

Blogからリンク構造を抽出した後、PageRank[Pageら]によるBlogのランキングづけとクラスタリングによるコミュニティの同定を行う。本章では、具体的な処理についての説明を行う。

3.1 PageRankによるBlogのランキング

PageRank Citation Ranking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(オーソリティなページ)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。 n 個のページがあったときに、それぞれのページのPageRank($r_1 \sim r_n$)を持つ $1 \times n$ のベクトル R を考え、そのページ間のリンクを表す $n \times n$ の行列を考えると、次式を満たすものがページランクとなる。(cは定数)

$$R = cMR$$

表1: 「野球」のPageRank

	アドレス(URL)	PageRank	説明
1	http://info.cocolog-nifty.com/info/	4.450551	ココログ運営者からのお知らせBlog
2	http://blog.goo.ne.jp/simauma_dx/	2.158997	さまざまな一般的な話題
3	http://blog.livedoor.jp/fireflycatching/	1.805901	みんなが楽しめるFlash紹介のBlog
4	http://yosuke.exblog.jp/	1.480864	俳優の窪塚洋介さんのBlog
5	http://finalvent.cocolog-nifty.com/fareastblog/	1.369204	さまざまな問題について語るBlog
6	http://masafumi.cocolog-nifty.com/masafumis_diar	1.261015	野球を含むさまざまな話題とコンピュータ関係
7	http://forum-fsaitama.cocolog-nifty.com/info/	1.238172	埼玉フォーラムのBlog(西部ライオンズ)
8	http://siketa6.seesaa.net/	1.228047	プロ野球、メジャーから高校野球まで野球尽くし
9	http://blog.livedoor.jp/sayupong/	1.170053	西野さゆきの舞ログ
10	http://blog.goo.ne.jp/nihei-labs/	1.147428	創作サークル「二瓶幻想研究所」(阪神ファン)
11	http://marines.exblog.jp/	1.108098	千葉ロッテマリーンズのファンのBlog

また、ページランクが収束するために Random Surfer Model を考えた。これは、ある確率で PageRank はランダムにあるサイトへ加算されるというものである。本稿では PageRank が用いたように[Page98]、確率 0.15 で、全てのページに等しい確率で PageRank が加算されるものとした。

3.2 グラフ分割によるクラスタリング

Blog 間のリンク構造から、弱いリンクを取り除くことでクラスタを抽出するグラフ分割によるクラスタリング法を用いる。ここで用いる Betweenness Clustering とは、Freeman が 1979 年に提唱した Betweenness の考え方をエッジ(無方向リンク)に拡張して Clustering を行う方法である[Michelle 01]。これは、リンクの関係に着目して、リンクが多く張られているコミュニティを見つけようとするものであるといえる。

Betweenness とは、あるノードが他ノードのペアの最短パスにどのくらいの割合で含まれているかを表はかる尺度である。つまり、最短パスがそのノードを通らないといけない度合いを示す。

これをエッジに拡張すると、エッジの Betweenness は「最短パスがそのエッジを通らないといけない度合い」になる。

Betweenness Clustering とは、この Betweenness が高いエッジから順番に切っていく、Clustering を行う。これはなるべく大きなクラスタ同士に分けようとする行為である。

グラフを Tree 構造だと考えれば、一番大きい二つのクラスタに分けられるのは、一番頂上のエッジ(無方向リンク)であり、頂上のエッジは Betweenness が最も高いエッジだからである。

4. Blog 間のリンク構造の分析

3 章で述べたホスティングサービスの Blog のリンクを収集し、4 章で述べた手法により、コミュニティの抽出・分析を行う。

4.1 「野球」の Blog について

(1) 「野球」の PageRank

「野球」の Blog に対して行った PageRank の結果の上位 11 件は表 1 のようになった。

表 1 の 1 位のサイトはココログのお知らせのための Blog である。ココログは、収集した 80 個の Blog 中、46 個の Blog がココログの Blog であり、お知らせページはココログの利用者、特に初心者の方はリンクを張りやすいと考えられるから、野球と関係のない Blog の PageRank が高いのも妥当であろう。2、3、5 位は、特に分野を絞らずに語っている Blog であり、誰でも読める内容になっている。これも野球に関係なしに人気のある Blog であろう。4、9 位は俳優のページであり、Blog の書き手が有名なパターンである。6、7、8、10、11 位の Blog は野球に関連した Blog である。

Google による PageRank の上位のサイトとは、Seed となった Blog だが、それらが本稿の PageRank 上位に入っていない。これは、PageRank の上位の Blog がいずれも外に向かうリンクをたかだか 1 本しか張っていないことが原因ではないかと考えられる。1 本以下しかリンクを張っていない Blog は 80 個中 28 個しかないもので、偶発的なものではないだろう。

(2) 「野球」の Betweenness Clustering

Betweenness Clustering によってエッジ(無方向にしたリンク) 100 本を切ったグラフが図 1 である。全部でグラフ間にはエッジが 248 本あった。

まず、図 1 の中央上部のクラスタ(①)は、野球とはほとんど関係のないクラスタである。例えば <http://sasanishiki.air-nifty.com/log/> や <http://facet.cocolog-nifty.com/divers/> など、雑記に近いものが多い。しかし、このクラスタには大きな特徴がある。クラスタに属する 24 個の Blog 全てがココログを利用した

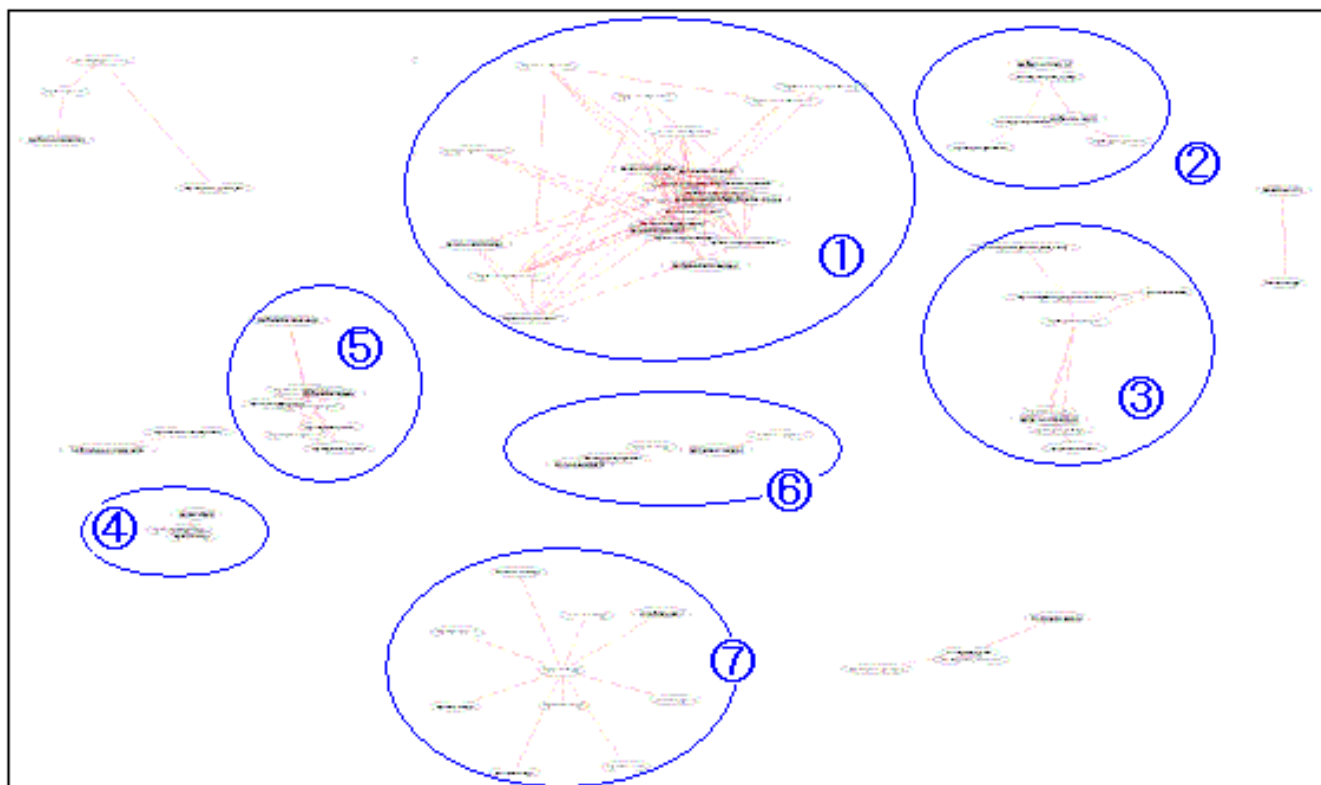


図 1 : 「野球」の Betweenness Clustering

表2: 「Winny」の PageRank

	アドレス		
1	http://kuraki.livedoor.jp/	3.082268	倉木麻衣さんのBlog
2	http://blog.livedoor.jp/weather11/	2.734883	お天気お姉さんのBlog
3	http://furukawa.cocolog-nifty.com/blog/	2.636927	ニフティ(株)代表取締役社長
4	http://copyrights.livedoor.biz/	2.343273	輸入CD規制に反対する音楽関係者
5	http://blog.livedoor.jp/memorize/	2.234231	ブロガーの血液型統計
6	http://souken.exblog.jp/	2.147456	日々雑感、日記、事象分析など徒然に
7	http://fji.jugem.cc/	2.140702	日々の出来事を写真を交えて
8	http://shinoken.tea-nifty.com/xxx/	2.022166	アメリカン★バカ情報をあなたに
9	http://www.myprofile.ne.jp/coolervinore+blog/	1.919774	日記
10	http://mobaio.cocolog-nifty.com/pickup/	1.880273	他のBlogを紹介するBlog

Blog である。この結果に限ったことではないが、同じホスティングサービス同士でのリンクは張られやすいというのが Blog の特長であろう。

同様に右上部のクラスタ(②)は 6 個ともエキサイトブログである。しかし、このクラスタでは 6 個中 4 個が野球に関連する Blog であり、しかも阪神ファンの Blog であった。

逆に右中央部の Blog(③)ではホスティングサービスは一定でないものの、クラスタに属する 8 個の Blog 全てが野球に関する Blog であった。

右下部のクラスタ(④)は 4 つともが、写真を中心とした Blog であったが、野球に関するものではなかった。写真を中心とした Blog は多数あるので、これだけでは偶然なのか、それとも Blog 作成者の趣向が近いために生じたクラスタかは判断できない。

左下部のクラスタは、全てがエキサイトブログによる Blog で、かつ(おそらく)女性によるクラスタであった。

左中央部のクラスタ(⑤)は、全てが Livedoor による Blog であった。しかし野球に関する Blog はなかった。

中央部のクラスタ(⑥)は、ホスティング、内容ともに共通の事項が少なかった。強いていえば、ゲーム・漫画関係だろうか。

中央下部のクラスタ(⑦)には全部で11つの Blog があるが、話題で大きな特徴があった。4つの Blog はサッカーについて語る Blog であるが、残り1を除く、6つの Blog は日本ハムファイターズに関する Blog であった。これは、私の人生経験上最多の日ハムコミュニティであり、大変有用なクラスタの発見となった。

以上、Betweenness Clustering によるコミュニティ抽出の分析をしてきたが、これにおける PageRank が高い上位 10 件の Blog は、ほぼ一つのクラスタに一つずつ配置されていた。これは、PageRank の高い Blog からのリンクが少ないために、Betweenness Clustering における PageRank の高い Blog につながるエッジも少なくなり、Betweenness が低くなったためと考えられる。

4.2 「Winny」の Blog について

「野球」と同様に、「Winny」「Blog」を検索後に Google 検索した Blog を出発点に収集した Blog について分析してみる。「野球」で収集した Blog は73個であった。

(1) 「Winny」の PageRank

PageRank は表2のようになった。1 位は倉木麻衣の Blog であり、2 位はお天気お姉さんの Blog である。これらは一般に人気があるサイトである。10 位は Blog を紹介する Blog である。その他 Blog は特に Winny に限っている Blog ではない。また「野球」で収集した Blog と同様に外に張るリンクがいずれも 1 本以下であった。全体でも50%の Blog が 1 本以下だったが、やは

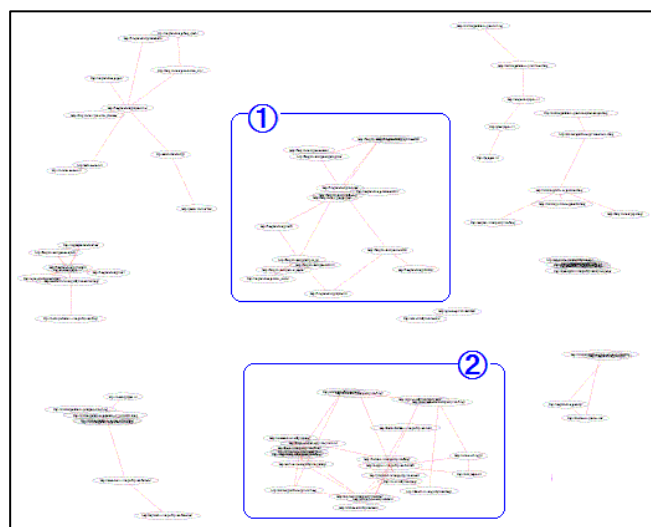


図2: 「Winny」の Betweenness Clustering

リンクを外に張っていないことも PageRank に大きく関与しているだろう。

(2) 「Winny」の Betweenness Clustering

Betweenness Clustering による方法では、全部で Blog 間には 212 本のエッジ(無方向リンク)があり、それを 100 本切ったときに、図 2 のようにクラスタが分けられた。このときも前節と同様に、中央上部のクラスタ(①)で 16 個全部が Livedoor による Blog であり、中央下部のクラスタ(②)では 24 個中 21 個がココログによる Blog であって、ホスティングサービス同士での結びつきの強さがうかがえた。

しかし、「野球」のときのように、「Winny」に関するような特定の性質の Blog はあまりなかった。Blog を全体的にみると、社会に関する話題を取り上げている Blog が多く、また Blog の記事の長さも野球に比べると長いものが多かった。これは、社会的問題に対する意見を発信する Blog が Winny について言及して、かつそれらの間ではリンクが張られている、つまり議論が行われているためだと考えられる。

4.3 「野球」と「Winny」の比較

本節では「野球」と「Winny」の分析結果を比較する。

PageRank による方法では、Blog による議論、すなわちリンクの張り合いではなく、有名 Blog が PageRank のいう「dangling page」になってしまい、有効な指標として働かない可能性がある。これはネットワークの規模を大きくするか、Random Surfer Model の確率を上げれば解決する問題かもしれない。

Betweenness Clustering によるコミュニティの抽出は、ホスティングサービスにより形成されたと考えられるコミュニティは「野球」「Winny」ともに抽出できる可能性がある。しかし話題によるコミュニティ形成は、「野球」に関する Blog では、ある程度成功したといえるだろうが、「Winny」に関する Blog では失敗している。これは「Winny」に関する Blog は、他にも様々な問題に関して言及し、様々な Blog リンクしている場合が多いので、話題によるコミュニティ形成がなされていないのか、もしくは Winny に言及しているときだけの短期的なコミュニティが形成されていたかであろう。逆に「野球」に関する話題は、特定の球団について長期に書かれる可能性が強いためコミュニティの抽出に成功したと考えられる。

5. 考察

5.1 PageRank への考察・評価

本稿では、PageRank による方法は適切な結果を得ることができなかった。それは、本稿では収集した Blog の数が少ないため外にリンクを張っていない Blog があると、その Blog の PageRank が高くなりすぎてしまうからだと考えられる。また私は Blog を個人の情報発信の場とみており、交流が盛んな Blog ほど中心的なページだと考えている。しかし、PageRank は外にリンクを張られていなくても、外からリンクを張られるだけで値が上がってしまうという性質なので、私が考えている中心的な Blog との違いは大きいだろう。

5.2 Betweenness Clustering への評価・考察

Betweenness Clustering によるコミュニティの抽出は、抽出するコミュニティによって適切かどうかが変わるだろう。例えば、野球についてしか書かれない Blog 群があれば、野球について語り合うことにより自然とその Blog 間にリンクが多くなり、Betweenness Clustering によってコミュニティとして抽出されるだろう。

だが、今回の Winny について語っている Blog のように、一過性の話題についての Blog は、意見の交換回数が相対的に少ないので、リンクも少なくなり、コミュニティとしては抽出されないだろう。

同様に、話題が多岐に渡る Blog は、リンクを色々な Blog に張ることになるので、話題によるコミュニティには関与しない可能性がある。例えば、たまに野球に語っている Blog は野球コミュニティよりも、ココログのコミュニティに入ってしまう。

これらを解決するには、ひとつの Blog が複数のコミュニティに所属することを許すことや、時系列で区切ってコミュニティを抽出することが必要である。私が考える最も有効な方法は、キーワードごとにコミュニティを抽出することである。なぜなら、今回、私はコミュニティを抽出するのに書く Blog を巡回しコミュニティが形成されているかを判断したが、そのためには Blog の記事を読まねばならず、その記事から得られた印象、すなわちキーワードが共通しているかでコミュニティの判断を行ったからである。

5.3 Blog ユーザの動向

今回の Blog のコミュニティ抽出により、「野球」「Winny」ともにホスティングサービスによるコミュニティが抽出されたので、Blog のユーザは同じホスティングサービスに対してリンクを張りやすいのではないだろうか。このことは、ブログ・ホスティングサービスを用いるユーザには、特別な知識を持たないインターネットの初心者が多いことが関係していると考えられる。なぜなら、初心者は、

検索エンジンやハブサイトを使わずに、自分のホスティングサービスのホームページから辿りやすい Blog を閲覧するだろうからである。またコメントや Trackback などは、同ホスティングサービスの方が、自分へのテストで行っているの、やり方がわかっていて、やりやすく、したがって結びつきが強くなるだろうからである。よって、同じホスティングサービスのユーザに絞って Blog コミュニティの抽出を行えば、話題に関連したコミュニティを抽出できる可能性があるだろう。逆に、同じホスティングサービスのユーザのつながりが強いのはわかっているのだから、話題に関連したコミュニティ抽出のときには同ホスティングサービスのつながりは弱く評価するのも有効であろう。

5.4 Blog 収集に対する問題点

Blog 収集についても、いくつか問題点が発見された。

ブログ・ホスティングサービスによっては、各ユーザの嗜好に対応するため様々なテンプレートを用意したり、ユーザ独自のテンプレートを使用可能にしたりしているので、本文抽出が困難な場合がある。特にホスティングサービス側がテンプレートを複数用意している場合は、特別な知識がなくても様々な形のソースを用いることが可能になってしまい、本文をだけを抜き出すことが困難となり、リンクを適切に抽出することができなくなってしまう。これに対応するには、HTML の解析をすすめるのが良いだろう。

また、最新の記事から一つ前の記事のアドレス、その一つ前の記事のアドレス…と順にとっていくのだが、途中で切れてしまうと、以降の記事が全て取れなくなってしまうという問題がある。前後の記事のアドレス関係が単なる数字の昇降順なら推測も可能だが、そうでない場合は大変困難となる。これは、月別アーカイブなどから推測する方法が考えられる。

また移転した Blog の場合、RSS を用いて最新の記事を探すと、集めようとした移転元の Blog ではなく、移転した先の Blog の記事だけしか集められない場合がある。これは、RSS で最新の記事を探すのではなく、現在わかっているアドレスから前後を順に収集する方法が考えられる。

6. おわりに

本稿では、特に日本のブログ・ホスティングサービスによる Blog におけるコミュニティの抽出と分析を行った。「野球」と「Winny」に関する Blog のリンクを自動収集するプログラムを作り自動収集させ、それらのリンクから PageRank と Betweenness Clustering によりコミュニティの抽出・分析を行った。その結果、「野球」に関しては野球に関するコミュニティが発見され、また Blog 全体として、ホスティングサービスが同じ Blog はつながりが深いということが判明した。しかし、PageRank による方法では、話題の中心の発見はうまくいかなかった。

コミュニティによっては抽出がうまくいくこともあるが、さらに多くの Blog を対象とした解析を行う必要があるだろう。また、Blog 間のコミュニティは、リンクを容易に張れるので、どういったリンクの種類があるのかといった要因を考慮しながら、コミュニティの抽出を行う必要があるのではないだろうか。リンク先・元の記事の情報を用いて、リンクの性質を表すラベル付けをすることが今後の課題である。

参考文献

- [Kumar 99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan,: Trawling the web for emerging cyber-communities, Elsevier Science B.V.,1999.
- [Flake 00] Gray William Flake, Steve Lawrence, C. Lee Giles,: Efficient Identification of Web Communities, ACM SIGKDD-2000, pp.150-160,2000
- [村田 01] 村田 剛志, 参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌 Vol.16 No.3, pp.316-323, 2001
- [Blood 02] Rebecca Blood,: We've Got Blog: How Weblogs are Changing Our Culture, Perseus Publishing, 2002.
- [NetRatings 04] Nielsen//NetRatings,: 2004 年 1 月 月間インターネット利用動向調査結果,
http://csp.netratings.co.jp/nnr/PDF/227_2004release_j_final.pdf
- [bulkfeeds] <http://bulkfeeds.net/app/stats.html>
- [Page 99] Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry:
The PageRank Citation Ranking: Bringing Order to the Web. ,
<http://dbpubs.stanford.edu:8090/pub/1999-66> ,1999.
- [Freeman 79] Freeman, L. C.: Centrality in social networks: Conceptual clarification, Social Networks, Vol.1, pp.215-239, 1979.
- [Google] http://www.google.co.jp/intl/ja/why_use.html
- [Michelle 01] Michelle Girvan; M. E. Newman, : Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821-7826 ,2002.