リンク構造の時間特性に着目したWeblog 解析に基づくコンテンツの信頼性評価の検討

Evaluating Content Trust Based on Weblog Analysis Adjusted to Time Current Characteristics of Its Link Structure

中島 伸介♥　舘村 純一♦
日野 洋一郎♣　原 良憲♦
田中 克己♠

Shinsuke NAKAJIMA　Junichi TATEMURA
Yoichiro HINO　Yoshinori HARA
Katsumi TANAKA

近年，Web を介したユーザ間の即時的情報流通が広まりつつある．Weblog はその一例であり，互いに関連しあうコンテンツが常時生成され続けている．従来，Google などの検索エンジンでは，蓄積されたコンテンツから信頼性の高いものを選択するのに静的なリンク構造を利用してきたが，Weblog のような動的特性を持つコンテンツには対応し切れていない．本研究では，常時生成されるリンク構造の動的特性に着目した Weblog の解析手法を提案し，信頼性と適時性の高いコンテンツの抽出・評価の可能性について議論する．

Recently, circulation of information through WWW between users is spreading. Weblog, which is one of information circulation environment, usually creates web contents that have relevance each other. Though conventional Web search engines like Google use static link structure in order to provide Web page rankings, they cannot effectively use weblogs that have a link structure in growth process. Thus, we propose evaluating content trust based on weblog analysis adjusted to weblog content analyzing method adjusted to time current characteristics of its link structure, and discuss how to extract and evaluate trustworthy and timely contents.

1. はじめに

信頼性の高い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といえるが，従来技術においてはGoogleなどの検索エンジンが与えるWebページのランキングを基に，そのWebコンテンツの有用性を推測しているのが現状である．しかしながら，Page Rank[1]等のリンク構造解析によるランキングは，十分発達した静的なリンク構造をもつWebコンテンツに対して有効な手法であり，ユーザによるリアルタイムの情報発信が増加している状況においては，生成されるコンテンツやこれらを結ぶリンク群は未発達であり，必ずしも有効ではない．

このWebを介した即時的情報流通方式の１つとして，Weblogが挙げられる．これらWeblogサイトがWeb上で提供されている情報に対する考えを記述しているケースが多いことから，これを解析することでWeblogが評価しているWebコンテンツの信頼性を見積もることができるのではないかと考えた．そこで，まずは生成されるリンク構造の動的特性に着目したWeblogの解析手法を提案する．この中で信頼性の高いWeblogの判別手法や，発生直後のイベントに関するWeblogスレッドの成長予測手法について検討する．そしてこれらを利用することで，信頼性と適時性の高いWebコンテンツの抽出・評価の可能性について議論する．なお，本研究にて指す“Webコンテンツの信頼性”とは，通信の保障やセキュリティに関するものではなく，情報の内容そのものに関する信頼性である．

2. Weblog の概要および関連研究

図１に典型的な Weblog サイトの例を示す．

Weblog サイトは，そのトップページに「エントリ」と呼ばれる個別書き込み記事を新しいものから数件表示している．通常は Weblog サイトの管理者のみがエントリを追加することができる．新しいエントリが追加されれば，古いエントリはトップページからは削除されるが，各エントリが保持している個別 URL を辿れば，トップページから削除された後でも閲覧することが可能である．

また，Weblog サイトトップページについては，RSS と呼ばれる XML で記述されたサイトの要約を公開していることが多く，RSS のみを巡回することで Weblog サイトの更新情報等を取得することが可能となっている．他人の Weblog エントリに対して，何らかの意見を述べる手段としては，コメントとして直接書き込む方法と，自分の Weblog サイトのエントリの中に対象の URL と共に書き込む方法がある．また，自分の Weblog サイトのエントリから貼るリンクにも２種類存在する．通常のリンクおよびトラックバックリンク[2]である．

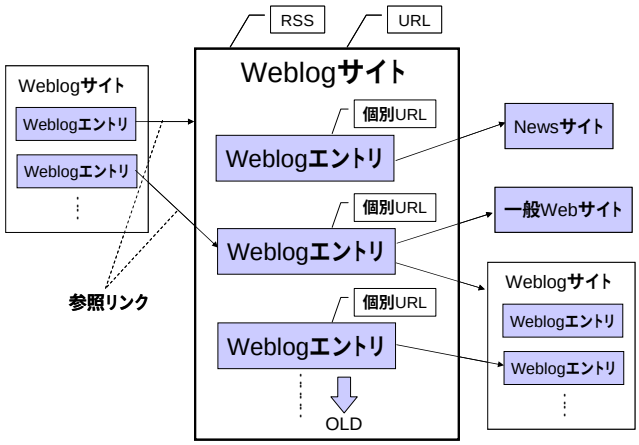

図１　典型的な Weblog サイト

Fig.1　A Typical Weblog Site

♥ 正会員　独立行政法人情報通信研究機構　けいはんな情報通信融合研究センター　snakajima@nict.go.jp
♦ 非会員　NEC Laboratories America, Inc.
{tatemura, hara}@sv.nec-labs.com
♣ 学生会員　京都大学大学院情報学研究科修士課程
hino@dl.kuis.kyoto-u.ac.jp
♠ 正会員　京都大学大学院情報学研究科
tanaka@dl.kuis.kyoto-u.ac.jp

トラックバックリンクはリンクを貼ったことをリンク参照元に知らせる機能があり，参照された Weblog エントリの投稿者がリンクを貼られたことを知ることができる．なお，Weblog サイトの定義は明確なものはないが，本研究では Weblog とは考えがたいニュースサイトを除き RSS を保持するものを Weblog と扱うことにしている．

Weblog 解析に関する関連研究としては，Kumar らの Weblog 空間の爆発的進化に関する調査研究[3]が挙げられる．彼らは，ハイパーリンクによる Weblog 群のつながりに注目し，blog コミュニティの抽出とこの blog コミュニティの進化に関する調査研究を行っている．ただし，Weblog および参照している Web コンテンツの信頼性評価を目的としているものではない．

3. Web コンテンツ信頼性の推定を目的とした Weblog 解析

Weblog が参照する Web コンテンツの信頼性を議論するため，まず各 Weblog の特性について評価を行うべきと考えた．そこで RSS 等に基づいて，Weblog データをクローリングし，以下に示す手順で解析を行う．

（1） Weblog スレッドの特定（3.1 節）
（2） 各 Weblog サイトの特性の判別（3.2 節）
（3） 目的の特性の Weblog サイトの検索（3.3 節）

3.1 Weblog スレッドの特定

Weblog エントリは，共通の話題について触れたり，お互いに参照し合ったりすることで，スレッドと呼ばれるエントリの集合を形成する．本研究では，Weblog スレッドを「あるイベントについて意味的関連性の高い Weblog エントリのつながり」として扱うスレッド内のエントリであり，黒丸がスレッド外のエントリである．白丸のうち A,B,C と書かれたものがスレッド内のルートとなるエントリである．スレッド内のエントリのうち，ルートとなる Weblog のみ，ニュースサイトであることも認める．なお，この「イベント」については，URI の有無は問わない．

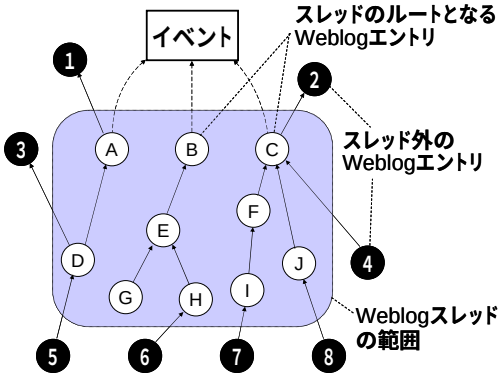

図 2 Weblog スレッド

Fig.2 A Weblog Thread

スレッドの特定方法としては，リンクによる接続が無い場合においても，同じイベントに関して言及しているエントリが存在すれば，同じスレッドに属するとみなす．

3.2 各 Weblog サイトの特性の判別

本節では，スレッド内における各エントリの位置付けを評価することで，そのエントリが記述されている Weblog サイトの特性の判別を行うことを検討する．

Weblog サイトはスレッドにエントリを提供している．逆に言えば，各スレッドは，何らかのアイデンティティを持った Weblog サイトからエントリの提供を受けている．したがって，扱われているトピックが類似しているスレッドの集合において，エントリの位置付けを統計的に解析することで，エントリを提供している Weblog サイトの特性の判別を行うことが可能と考えた．本研究では，トピック毎のスレッドの集合において，各 Weblog サイトは何らかの役割を担っているものという仮説を立てた．以下に，スレッドにおける Weblog サイトの特性（役割）に関する仮説を示し，それぞれについて説明する．

（1） Topicfinder

Topicfinder とは，議論が盛んに行われた Weblog スレッドにおいて，スレッドの初期段階に，エントリを提供することが多い Weblog 投稿者である（図 3 参照）．図 3 のグラフの横軸は，スレッドの立ち上がりからの経過時間であり，縦軸はスレッドに対するエントリ数である．つまり，Topicfinder は，成長前の段階からスレッドにて議論するための良いトピックを見つけることが多い Weblog 投稿者であるといえる．Topicfinder のエントリを監視することで，スレッドが将来成長するかどうかの判断材料にすることができる．

（2） Agitator

Agitator とは，議論が盛んに行われた Weblog スレッドにおいて，スレッドでの議論が盛んになる直前にエントリを提供することが多い Weblog 投稿者である（図 3 参照）．Agitator は，自らのエントリによって，Weblog スレッドの議論が盛んになるきっかけを作っている可能性が高い Weblog 投稿者である．Agitator のエントリを監視することで，Weblog スレッドが成長する時期を予測するための判断材料にすることができる．

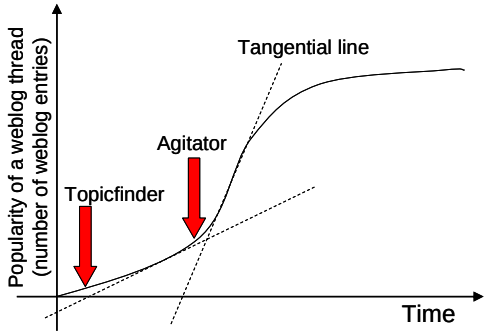

図 3 Topicfinder および Agitator

Fig.3 Topicfinder and Agitator

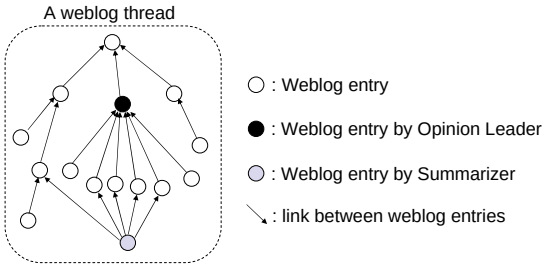

図 4 Opinion Leader および Summarizer

Fig.4 Opinion Leader and Summarizer

（3） Opinion Leader

Opinion Leader とは，あるトピックに関するスレッド内において，他の Weblog エントリから参照されることが多い Weblog 投稿者である（図４参照）．図４では，各ノードが Weblog エントリを示し，黒いノードが Opinion Leader によるエントリを示す．Opinion Leader のエントリを監視することで，あるトピックに関する Weblog コミュニティにおける重要な見解を効率よく取得することができる．

（4） Summarizer

Summarizer とは，あるトピックに関するスレッド内において，他の多くの Weblog エントリを参照することが多い Weblog 投稿者である（図４参照）．図中の灰色のノードが Summarizer を示す．Summarizer のエントリを監視することで，あるトピックに関する Weblog スレッドをまとめたような書き込みを効率よく取得できる可能性がある．

4. 信頼性と適時性の高いコンテンツの抽出・評価の可能性の検討

本節では，信頼性と適時性の高いWebコンテンツの抽出および評価方法の可能性について述べる．

4.1 Web情報検索時の信頼性・適時性の高いコンテンツの提供

従来の検索エンジン，例えばGoogleのページランキングは，Webページの重要性を判断する際の尺度に成り得るが，そのページの重要性がなぜ高いのかということについてはユーザは判断できない．また， Googleのランキングのためのリンク構造解析は，十分発達したリンク構造を想定しており，動的にリンク構造が変化するようなコンテンツに対しては必ずしも有効ではない．そこでこれらの問題を解決するような，Web情報検索時の信頼性・適時性の高いコンテンツの提供手法について以下の方針に基づいて検討する．

・Weblog解析において，Weblogスレッドにて扱われているトピックを判別しておくことで，Webコンテンツがどのようなコミュニティから，どのような観点で評価されているのかを把握することを試みる．これにより，どのような観点で評価されているコンテンツであるのかを含めて，検索結果をユーザに提示する．
・Weblog解析により，TopicfinderやAgitatorを判別し，これらのWeblogエントリを監視することで，議論が活発になる直前および議論が活発になりそうなWeblogスレッドの推測を行う．これにより，将来重要性が高いWeblogスレッドに発達しそうなものを早期に発見し，信頼性・適時性の高いWeblogコンテンツを提供する．

4.2 信頼性・適時性の高いニュース記事の補足コメントの提示

有名なニュース配信サイトは，信頼性および適時性の高い情報（ニュース）を配信しているといえるが，有名であるため発表したくてもできない情報が存在していることもあり得る．そこでこれらの問題を解決するような，補足コメントの提示手法について以下の方針に基づいて検討する．

・対象としているニュース記事を参照しているWeblogエントリのクローリングを行い，Topicfinder，Agitator，Opinion Leader，Summarizerの存在等に基づいて，重要性が高そうなWeblogエントリの特定を試みる．これを提示することで，公式な立場では発表し難い情報も，ニュース記事掲載後の早い段階から提供が可能になる．

5. Weblogスレッドに関する調査実験および考察

本節では，このうち，スレッドモデルおよびWeblogサイトの特性について，事例に基づいた議論を行う．Weblogサイトに関して統計的な解析を行うためには，大規模なデータ収集が必要であるが，本論文ではWeblogエントリのトラックバックを手作業で辿ることで，幾つかのスレッドに関する事例を収集した．この調査実験の制限を以下に示す．

・Weblogエントリ同士の意味的な関連を考慮ない．
・データ数が十分ではなく統計的解析はできていない．

なお，本論文においては，TrackBack Voyager[4]という，トラックバック情報検出サイトを利用して，トラックバックリンクによりつながりを持つWeblogエントリの集合を抽出し，これをWeblogスレッドとした．取得したWeblogスレッドに対して，エントリ数の時系列変化グラフと，トラックバックリンクに基づくリンク構造グラフを生成して，Weblogスレッドに関する考察を行った．

5.1 Weblogスレッドのモデルに関する考察

本節ではスレッドモデルに関する考察を行う．図５および図６に Weblog スレッドのリンクグラフおよびエントリ数の時系列変化を示す．各図上部のリンクグラフ中の○印は Weblog エントリを示し，これらを結ぶ矢印はリンクの参照関係を示している．太線の両端矢印は，相互リンクを示す．

また，各図下部のWeblogスレッドのエントリ数の時系列変化を示すグラフでは，縦軸がエントリ数で横軸が日付となっている．グラフ中にプロットされた●印は，同色のリンクグラフのエントリに対応する．

5.1.1 スレッドの成長過程

ここでは，スレッド内のエントリ数の増加をそのスレッドの成長とみなす．各図（図5，図6）からいえることは，各スレッドの成長過程は急激にエントリ数が増加する成長期と，エントリの増加量がほとんどない停滞期が見られることである．恐らく，最初のエントリが投稿されてから，スレッドの存在が認知されるまでに最初の停滞期が存在し，その後に

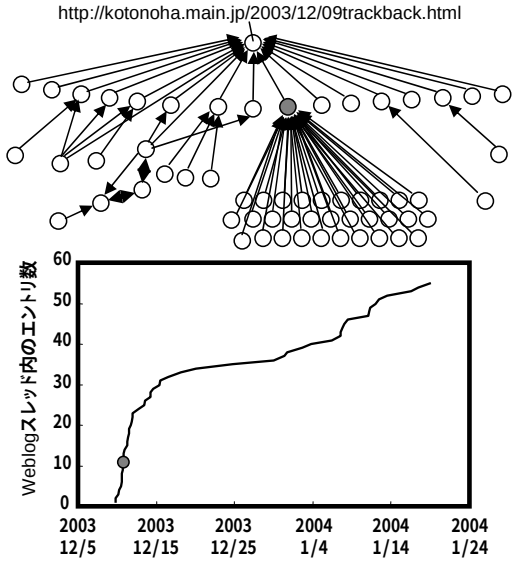

図５ Weblog スレッドの調査実験結果１

Fig.5 Experimental Result for a Weblog Thread 1

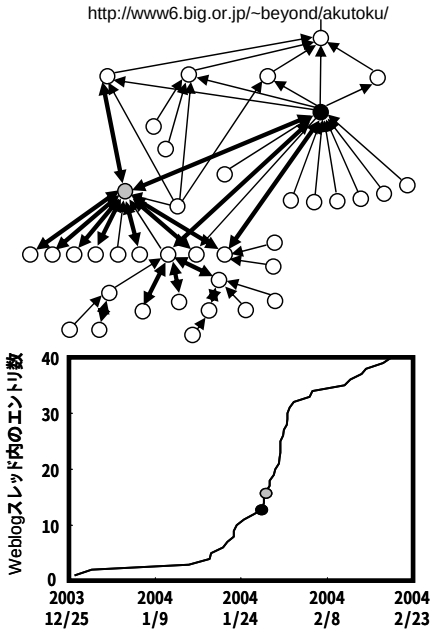

図 6 Weblog スレッドの調査実験結果 2

Fig.6 Experimental Result for a Weblog Thread 2

多くのユーザに認知されると共に議論が盛んになる成長期となる．さらにその後，ある程度議論が収束するもしくはユーザの関心が薄れることで停滞期となると考えている．

ただし，スレッドが対象とするイベントが，ニュースにて大きく取り上げられた場合においては，図5のように初期の停滞期が存在せずに，初めから成長期に入る場合もある．

5.1.2 スレッド内のリンク構造

スレッド内のリンク構造に関する各図の共通点は，リンクの参照関係には偏りがあり，灰色および黒色で示されたノードのように，これを参照しているエントリが特に多いノードが存在していることである．図5中の灰色のノードに対しては31本（スレッド内の全てのリンクの46%）のリンクが貼られており，図6中の灰色のノードに対しては12本（同19%），黒色のノードに対しては10本（同16%）のリンクが貼られている．各図のリンクグラフを見れば容易に予測できるが，これらの参照しているエントリが多いノード（エントリ）は，各々のスレッドにおいて重要な役割を担っているといえる．

5.2 Weblogサイトの特性に関する考察

本節では，各Weblogの特性に関して，調査実験結果に基づいて考察する．まず，Opinion Leaderについて考察する．5.1.2節でも述べたとおり，図5，図6の各々において被参照リンクの多いエントリが存在するが，これを提供するWeblogサイトがOpinion Leader候補となる．そして，他の多くののスレッドにおいても，同様に被参照リンクが多いエントリを提供していればOpinion Leaderと判定される．これらOpinion Leader候補のエントリは，図5，図6からも分かるように，エントリ数の時系列変化を示したグラフにおいて，スレッドの急激な成長の前に提供されたエントリであるといえる．したがって，Opinion Leader候補であるエントリは，Agitator的な存在である可能性がある．データ量を増やして統計的な解析を行う必要があると考える．

次にSummarizerについてであるが，参照リンクを顕著に数多く保持するエントリは存在しなかった．Weblogサイトには，Summarizerがそもそも存在しないということも考えられるが，今後の統計的な解析に基づいて判断すべきである．

TopicfinderおよびAgitatorの判別のためには，取得したスレッドにおける時系列解析を統計的に行う必要があり，本論文にて行った実験データでは不十分である．ただし，5.1.1節でも述べたように，スレッドの成長過程においては，成長期と停滞期が見られることが確認できており，TopicfinderおよびAgitatorの定義に利用する条件である急激な成長以前という時期を特定することは可能であると考える．今後，統計的解析に必要なデータ収集を行い，TopicfinderおよびAgitatorに関する解析を行う．

6. まとめと今後の課題

本論文ではWeblogコンテンツの信頼性の推定目的としたWeblog の解析手法について検討し，信頼性と適時性の高いWeb コンテンツの抽出･評価の方法について検討すると共に，Weblog スレッドに関する調査実験および考察を行った．

今後は，Weblog スレッド抽出ソフトを実装し，統計的な実験を通じて仮説の検証やアプリケーションの実現に向けた検討を行う予定である．

[謝辞]

本研究の一部は，平成 15 年度文部科学省科学研究費特定領域研究(2)「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」(課題番号：15017249)，および京都大学２１世紀ＣＯＥプログラム「知識社会基盤構築のための情報学拠点形成」による．ここに記して謝意を表します．

[文献]

[1] Page, L., Brin, S., Motwani, R., Winograd, T.: "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Libraries Working Paper, (1998).

[2] ３分でわかるトラックバック，http://kotonoha.main.jp/weblog/000255_trackback.html

[3] Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: "On the Bursty Evolution of Blogspace", The Twelfth International World Wide Web Conference (2003). http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477-kumar.htm

[4] TrackBack Voyager, http://holic.org/b2uvoyager.php

中島 伸介 Shinsuke NAKAJIMA
独立行政法人情報通信研究機構勤務. 2004 京都大学大学院情報学研究科博士後期課程修了，博士(情報学). 日本データベース学会，情報処理学会，人工知能学会，環境システム計測制御学会各会員.

舘村 純一 Junichi TATEMURA
NEC Laboratories America 勤務, 1994 東京大学大学院工学系研究科情報工学専攻博士課程修了，工学博士．情報処理学会，ACM, IEEE Computer Society 各会員.

日野 洋一郎 Yoichiro HINO
京都大学大学院情報学研究科修士課程在学中．2004 京都大学工学部情報工学科卒業．日本データベース学会学生会員.

原 良憲 Yoshinori HARA
NEC Laboratories America 勤務. Department Head. 1983 年東京大学工学系研究科電気工学専攻修士課程終了．主にハイパーメディアシステム関連の研究開発に従事．情報処理学会、ACM 各会員.

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科教授. 1976 京都大学大学院修士課程修了．工学博士．主にデータベースの研究に従事．情報処理学会，日本データベース学会, 人工知能学会, ACM, IEEE Computer Society 等各会員.