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

中島　伸介† 舘村　純一†† 日野洋一郎† 原　良憲†† 田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻
†† NEC Laboratories America, Inc.
E-mail: †{nakajima,hino,tanaka}@dl.kuis.kyoto-u.ac.jp, ††{tatemura,hara}@sv.nec-labs.com

あらまし ユビキタス・ブロードバンド基盤は人々が常にオンラインであるという環境をもたらしつつある．このような中で，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．

キーワード 情報信頼性, トラスト, Web 情報検索, Weblog

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

Shinsuke NAKAJIMA†, Junichi TATEMURA††, Yoichiro HINO†, Yoshinori HARA††, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
†† NEC Laboratories America, Inc.
E-mail: †{nakajima,hino,tanaka}@dl.kuis.kyoto-u.ac.jp, ††{tatemura,hara}@sv.nec-labs.com

Abstract Broadband infrastructure and ubiquitous computing have brought an environment that people are always online to World Wid Web. In this circumstance, 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.

Key words Information trust, Web Information Retrieval, Weblog

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獲得することが困難となっている．そこで，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．従来技術においては Google [1] 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 [2] 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．Weblog はある言い方をすれば「積極的に他のコンテンツに対してリンクを貼った，興味に基づいて記述した Web 上のコメント集」である．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．つまり，Web コンテンツそのものの内容やリンク構造から信頼性を見積もるのではなく，これらの Web コンテンツに対してコメント与えている Weblog を通じて信頼性の評価を行うものである．すなわち，有用な Web コンテンツの推薦手法の構築を目標としたうえで，信頼性の高い Weblog サイトの判別手法について検討し，さらに信頼性の高い Weblog サイトを通して，有用な Web コンテンツを判別する手法について検討することを本研究の目的とする．

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．

以下，本論文の構成を示す．２節では Weblog の概要および関連研究について述べる．３節では Web コンテンツ信頼性の推定を目的とした Weblog 解析について述べる．４節では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について述べる．５節では Weblog スレッドに関する調査実験および考察を述べる．６節ではまとめと今後の方向性について述べる．

2. Weblog の概要および関連研究

Weblog は，アメリカにおいては 1999 年以降，急速に発達し標準化が進みつつある．単に「blog」と呼ぶことも多く，書き手は「blogger」と呼ばれる．日本では元々「Web 日記」と呼ばれるサイトが数多く存在しており，広義での Weblog と定義できる．つまり，現在の Weblog に近いサイトは元々は日本において発達し，1999 年以降のアメリカでの発達と共にその標準化が進んでいるといえる．最近では,「MovableType [3]」などの Weblog サイト構築ツールなども公開されている．また,「はてなダイアリー [4]」等のようにホスティングサービスを行っているサイトも増えており，誰でも簡単に Weblog サイトを立ち上げるための環境が整っているといえる．

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最近（例えば一週間）に書かれた「エントリ」と呼ばれる個別書き込み記事を複数表示している．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，この点が Web 掲示板とは異なる．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（Rich Site Summary もしくは RDF Site Summary）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．

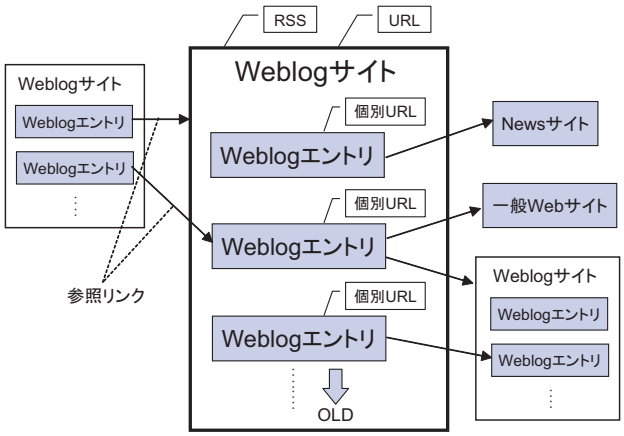

図 1　典型的な Weblog サイトの例

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，対象としているエントリに対して自分の意見をコメントとして直接書き込む方法と，対象としているエントリの個別 URL と共に自分の Weblog サイトにエントリを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 [5] である．通常のリンクでは，参照元はリンクを貼られたことを知ることはできないが，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．

Weblog サイトであることの条件は明確なものはないが，本研究では RSS を保持するものを Weblog と扱うことにしている．ただし，ニュースサイトの中には RSS を公開しているものもある．したがって，RSS が存在しても，明らかにニュースサイトであり Weblog とは考えにくいと認められる場合には，これを除外して考える．

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 [6] が挙げられる．彼らは，25000 の Weblog サイトとその中の 750000 本のリンクについて解析している．また，Blogspace と名づけたハイパーリンクによる Weblog 群のつながりに注目し，この Blogspace における 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には，まず各 Weblog の特性について評価を行うべきと考えている．そこで本節では，Weblog 解析に関する方針および手法について述べる．まずは，RSS 等を利用して Weblog クローラにより，Weblog データをクローリングするが，その後の解析

手順としては，以下のように考えている．

（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，お互いに参照し合うことで，ある話題に関するエントリの集合を形成することがある．この関連性の高いエントリ同士のつながりを，スレッドと呼ぶ．

本研究では，Weblog スレッドを「あるイベント（ニュース，トピック）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（図 2 参照）．

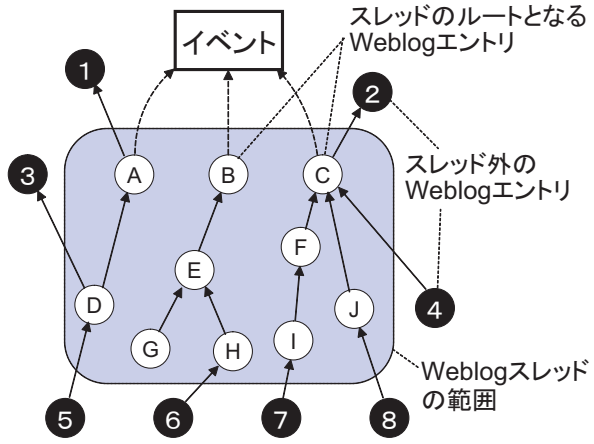

図 2 Weblog スレッド

図 2 中の白丸が Weblog スレッド内のエントリであり，黒丸が Weblog 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．

Weblog 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．この時，スレッドをシステムが判別して抽出する場合には，Weblog 同士の意味的関連をどのように判定するかが問題となる．

Weblog スレッドの判別方法を以下に示す．

- **スレッド内のルートとなる Weblog エントリの判定**

以下の条件のいずれかを満たす Weblog エントリもしくはニュースサイトの記事をスレッドのルートとする．

– 対象となるイベントとの関連が強く，それより上位の Weblog エントリおよびニュースサイトの記事がない

– それらが存在していたとしても対象となるイベントとの関連性が低い

図 2 の例によると，「C」の Weblog エントリは，リンク上位に黒丸 2 の Weblog エントリが存在するが，黒丸 2 が対象となるイベントとの関連が低いためにスレッドには含まれないので，このスレッドのルートとなる．

- **スレッド内のルート以下の Weblog エントリの判定**

以下の条件のいすれかを満たす Weblog エントリは，スレッドに含まれるものとする．

– ルートとなる Weblog エントリの下位に存在し，対象となるイベントとの関連性が強い．

– 1 段上位の Weblog エントリのとの関連性が強い．

図 2 の例によると，D〜J の白丸により表示されている Weblog エントリは，イベントとの関連性が高いか，1 段上位の Weblog エントリとの関連性が高いことにより，スレッドに含められている．逆に 4 〜 8 の黒丸は，イベントとの関連性が低く，1 段上位のエントリとの関連性が低いことにより，スレッドには含められない．

なお，イベントと各エントリとの関連性や，エントリ同士の関連性の判定方法については，キーワードマッチングや，出現単語に基づく特徴ベクトル同士の類似度などに基づいて行うことを検討している．

3.2 各 Weblog サイトの特性の判別

前節では，各エントリの特性からスレッドの特定について議論した．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 の特性の判別を行うことを検討する．

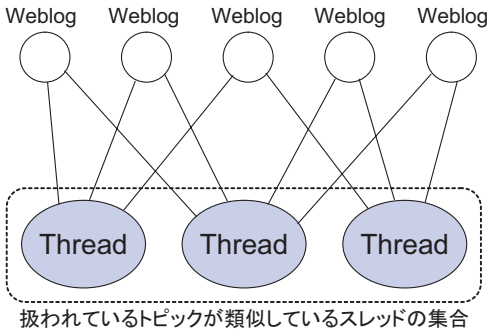

図 3 スレッドと Weblog との関係

図 3 にスレッドと Weblog との関係を示す．Weblog 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 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 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 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 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，スレッドが立ち上がった初期段階において，エントリを提供することが多い Weblog 投稿者である．（図 4 参照）

図 4 のグラフの横軸は，Weblog スレッドの立ち上がりからの経過時間であり，縦軸は Weblog 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 Weblog

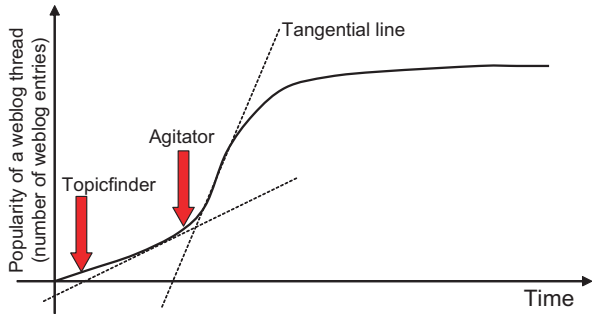

図 4　Topicfinder および Agitator

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，Weblog 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．

判別方法は，あるトピックに関する成長前のスレッドに対して Weblog エントリを投稿した後に，そのスレッドが最終的に成長する割合が高いと統計的に判断できれば，このエントリを提供した Weblog サイトをそのトピックに関する Topicfinder と判定する．

（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 4 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．

判別方法は，あるトピックに関する成長前のスレッドに対して Weblog エントリを投稿した直後に，そのスレッドが成長する割合が高いと統計的に判断できれば，このエントリを提供した Weblog サイトをそのトピックに関する Agitator と判定する．

（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 5 参照)

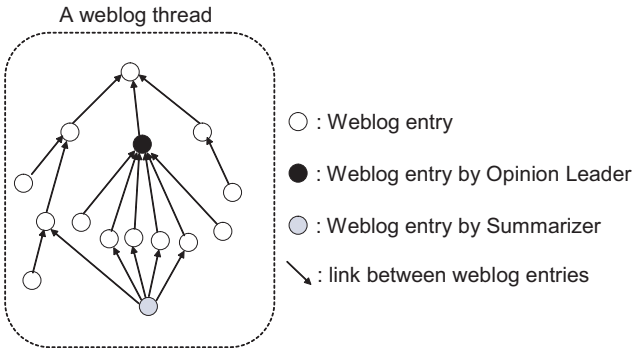

図 5　Opinion Leader および Summerizer の例

図 5 において，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．

判別方法は，あるトピックに関する Weblog スレッド内の総リンク数に対する被参照リンク数の割合に基づいて Opinion Leader 候補のエントリを決定する．そして，この候補となったエントリを，同一のトピックにおいて，ある個数以上保持する Weblog サイトを，そのトピックに関する Opinion Leader と判定する．

（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 5 参照) 図 5 において，灰色のノードが Summarizer によるエントリを示す．Summerizer 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．

判別方法は，あるトピックに関する Weblog スレッド内の総リンク数に対する参照リンク数の割合に基づいて Summarizer 候補のエントリを決定する．そして，この候補となったエントリを，同一のトピックにおいて，ある個数以上保持する Weblog サイトを，そのトピックに関する Summarizer と判定する．

（5） **Fan**

Fan とは，あるトピックに関する Weblog スレッドに対して，エントリを投稿ことが多い Weblog 投稿者である．

あるトピックに関する Fan を特定することができれば，Weblog スレッドに対する Fan によるエントリの状況によって，そのスレッドのトピックを特定することが容易になる．

判別方法は，あるトピックに関する Weblog スレッドへのエントリをある個数以上保持する Weblog サイトを，そのトピックに関する Fan と判定する．

ただし，各 Weblog は全てのトピックに対して，一様の役割を担うわけではなく，あるトピックに関しては，「Topicfinder」である Weblog が，別のトピックでは「Agitator」であることも有り得る．また，それぞれの特性（役割）については，ランク値を設定して，それぞれの役割に関するレベルを評価する．

3.3　目的の特性の Weblog サイトの検索

本節では，ある特性の Weblog のうちランク値が高い Weblog の検索手法の概要について述べる．

過去に判別および抽出されたスレッドに関しては，特徴キーワード抽出等によりそのスレッドを対象としているイベントの内容（トピック）は表現されているものとする．また，既に判別および抽出されたスレッドに関しては，これに対する Weblog の特性は明らかになっている．

ここで Weblog 検索のクエリが，キーワード，関連する Weblog，関連するスレッド，のそれぞれの場合に対する Weblog 検索手法の概要を述べる．

- キーワードに基づく検索

検索キーワードに関する幾つかの過去のスレッドを特定し，

この中で目的の役割に関してランク値の高い Weblog を提示する．

- 関連する幾つかの Weblog に基づく検索

提示された幾つかの Weblog が属する共通の過去のスレッドを特定し，この中で目的の役割に関してランク値の高い Weblog を提示する．

- 関連するスレッドに基づく検索

提示された幾つかのスレッドに関して，これらが共通で含んでいる複数の Weblog を特定し，さらにこれらが属する共通の過去のスレッド（クエリとして提示されたもの以外）を抽出する．そこで，クエリとして提示されたものも含め，これらのスレッドの中で目的の役割に関してランク値の高い Weblog を提示する．

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

3 節において，Weblog 解析方法およびこれに基づく，Weblog サイトの特性の判別方法について述べた．本節ではこれらの結果を利用し，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議論する．

想定するアプリケーションのとして，以下の 2 つの例について検討する．

- Web 情報検索時の信頼性および適時性の高いコンテンツの提供
- 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

想定するこれらのアプリケーションの詳細について以下に述べる．

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 が上位でランクしているサイトは，他の数多くのランク上位のサイトからリンクを貼られているサイトである．つまり Google のランク上位のサイトは，他のサイトが何らかの理由でリンクを貼る重要性を認めたサイトであるが，どのような観点でその重要性を感じているのかは不明である．また，1 節で述べたように 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，Fan の存在等により Weblog スレッドにて扱われているトピック（Weblog コミュニティのカテゴリ）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，単にランクが高いというだけではなく，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 Weblog 投稿者の役割として，Topicfinder や Agitator を判別し，これらの Weblog エントリを監視することで，発達する直前および発達しそうな Weblog スレッドの推測を試みる．また，彼らの Topicfinder や Agitator としてのレベルや，Weblog スレッド内での Opinion Leader の存在に基づいて，Weblog スレッド自体の重要性の推定を試みる．以上より，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場合がある．また，Weblog などに書き込まれるニュース記事の補足情報は，事前にクローリングすることが不可能であるため（ニュースが発表されるまでは存在しないため），従来の検索エンジンで取得することは難しい．

そこでこれらの問題を解決するような，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 エントリを特定することを試みる．これを提示することで，ユーザは公式な立場では発表し難いような情報についても，ニュース記事掲載後の早い段階から取得することが可能になる．

5. Weblog スレッドに関する調査実験および考察

3 節で，スレッドモデルおよび Weblog サイトの特性に関する仮説を，4 節で，Weblog サイトの特性値の利用に関する仮説を提案した．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．

- 3.1 節で定義した理想的な Weblog スレッドは，Weblog エントリ同士の意味的な関連性を考慮すべきであるが，トラックバックの情報のみを用いて構造を抽出しており，理想的な Weblog スレッドを抽出することはできない．
- Weblog サイトの特性は，複数のスレッドに対する統計的解析により判定するものであるが，今回は一つのスレッドのみを観察している．

5.1 調査実験の方法

本論文においては，TrackBack Voyager [7] 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．本論文にて行った，トラックバックに基づく Weblog スレッドの抽出手順を以下に示す．

（1） トラックバックが存在し，かつ，トラックバック付きリンクによって他の Weblog エントリを参照していない Weblog エントリを探し，これをルートエントリとする．

（2） ルートエントリから順番にトラックバックを辿り，トラックバックリンクによって接続されている Weblog エントリをスレッドに加えていく．

（3） 最終的にトラックバックリンクが無くなるまで，もしくは明らかに扱っているトピックがルートエントリと関連が無くなるまで，続ける．

（4） (3) までに取得された Web エントリの集合を Weblog スレッドとする．

この手法により取得されたトラックバックリンクに基づく Weblog スレッドのイメージを図 6 に示す．

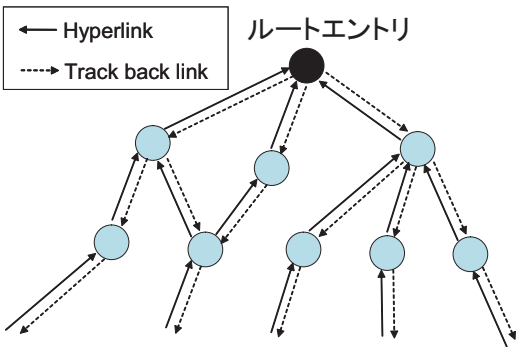

図 6　トラックバックリンクに基づく Weblog スレッドのイメージ

以上のように取得した Weblog スレッドに対して，投稿されたエントリ数の時間的変化を示した Weblog スレッドにおける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．

5.2　Weblog スレッドのモデルに関する考察

5.1 節で説明した調査実験の方法に基づいて Weblog スレッドの抽出を行い，スレッドモデルに関する考察を行う．以下に取得した Weblog スレッドの例を示す．

図 7 に“ トラックバックの使い方 ”に関する Weblog スレッド，図 8 に“ Bulkfeeds RSS ”に関する Weblog スレッド，図 9 に“ Google からの特定サイトの削除 ”に関する 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している．

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（トラックバック付きの）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．トラックバックリンクそのものに関しては省略している．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．

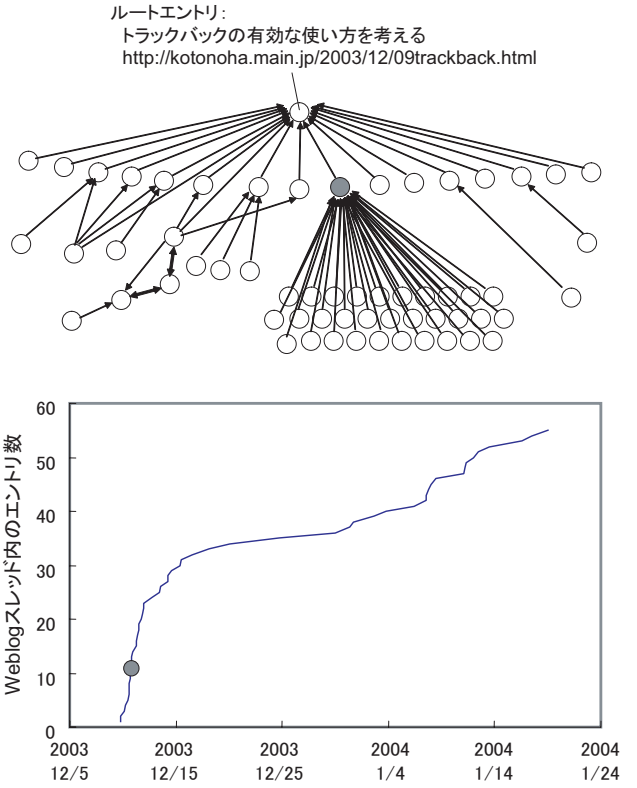

図 7 “ トラックバックの使い方 ”に関する Weblog スレッドの調査実験

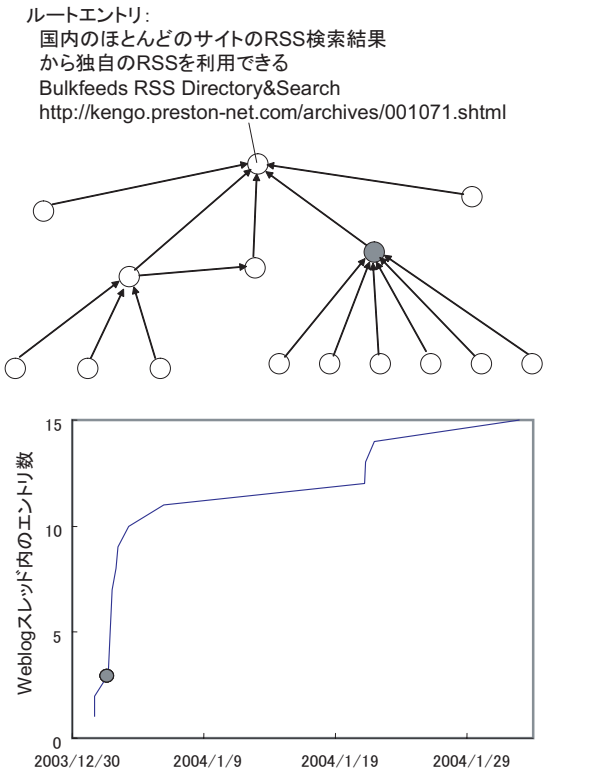

図 8 “ Bulkfeeds RSS ”に関する Weblog スレッドの調査実験

5.2.1　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 7，図 8，図 9）から共通していえることは，各スレッドの成長過程は急激にエントリ数が増加する

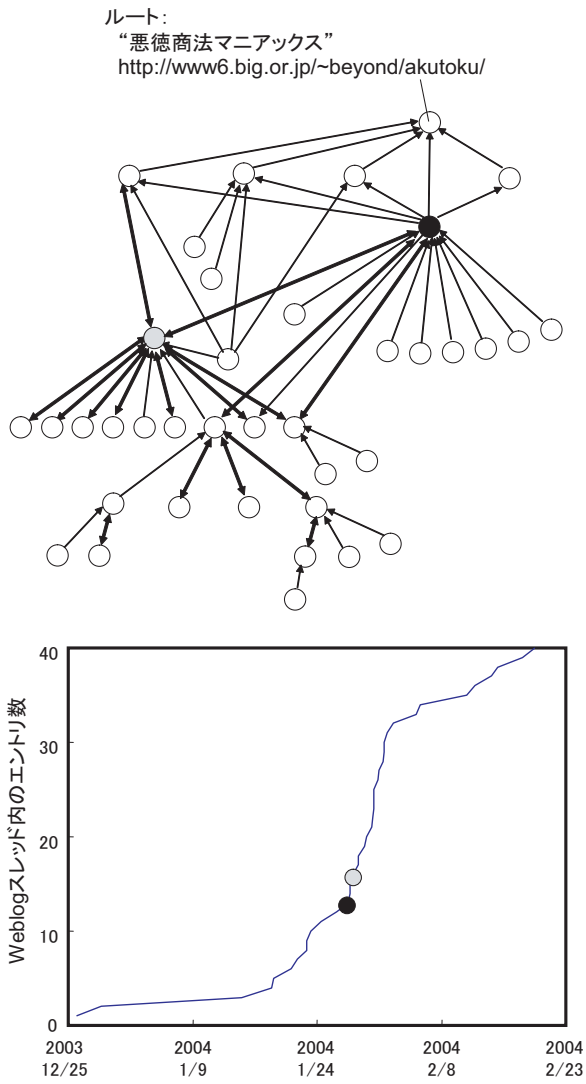

図 9 “ Google からの特定サイトの削除 ”に関する Weblog スレッドの調査実験

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多くのユーザに認知されるまでに最初の停滞期が存在し，その後に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のではないかと考えている．

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 7 および図 8 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．

5.2.2 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（図 7，図 8，図 9）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 7 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 8 中の灰色のノードに対しては 6 本（同 40%）のリンクが貼られている．図 9 では，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．

5.3 Weblog サイトの特性に関する考察

本節では，3.2 節で説明した各 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，他の多くのエントリから参照されることが多い Weblog エントリを提供する Opinion Leader について考察する．5.2.2 節でも述べたとおり，図 7，図 8，図 9 の各々において被参照リンクの多いエントリが存在しているが，これを提供する Weblog サイトが Opinion Leader 候補である．もし，他の幾つかのスレッドにおいても，同様に被参照リンクが多い Weblog 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 7，図 8，図 9 からも分かるように，スレッド内の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．

次に参照リンクを多数保持するエントリを提供する 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せず，比較的多い参照リンクを保持するエントリにおいても，スレッドの議論の内容をまとめたようなエントリではなかった．Weblog スレッドには，Summarizer 的なエントリがそもそも存在しないということも考えられるが，統計的な解析に基づいて判断すべきと考える．

Topicfinder および Agitator についてであるが，これらの判別のためには，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，5.2.1 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，Weblog スレッドの自動抽出プログラムを構築することで，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．

5.4 自動スレッド抽出における問題点

調査実験を通して明らかになったソフトウェアによる自動スレッド抽出における問題点について述べる．

- Weblog 投稿者のミスによる問題
  - 参照先 URL を本来はエントリの個別 URL とするところを Weblog サイトトップページの URL を記述している場合，参照しているエントリを特定できない．
  - 反対に，トラックバック元の URL が個別 URL ではなく，Weblog サイトのトップページの場合，トラックバック元のエントリを取得できない．
- Weblog 投稿者の追加編集による問題

– Weblog エントリを書き込んだ後日に，同じエントリに対して編集（追加や削除）することが可能であるが，エントリの書き込み日時としては初めに書き込み日時が記録されるため，後日追記した内容に関しては時系列的な整合性が取れない．

– エントリ書き込み日時と，トラックバックリンクを付けた日時が同一とは限らない．

- その他の問題

– いわゆるトラックバックのプロトコルに則っていない被リンクをトラックバックとして処理するサイトも存在する．（標準化されていない）

解決策としては，Weblog に関する仕様の標準化を進めることや，エントリのバージョン管理を行うことが必要であると考える．

6. おわりに

本論文のまとめを以下に示す．

- 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た．
- 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した．
- 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] Google, http://www.google.com
[2] Lawrence Page, et al: ”The PageRank Citation Ranking: Bringing Order to the Web”, Stanford Digital Libraries Working Paper, (1998).
[3] movabletype.org, http://www.movabletype.org/
[4] はてなダイアリー, http://d.hatena.ne.jp/
[5] ３分でわかるトラックバック, http://kotonoha.main.jp/weblog/000255_trackback.html
[6] Ravi Kumar, et al: ”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
[7] TrackBack Voyager, http://holic.org/b2uvoyager.php