人工知能学会研究会資料
SIG-SWO-A401-05

blog 解析に基づく Web 情報検索の信頼性向上技術

中島 伸介* 竹原 幹人** 舘村 純一***
日野 洋一郎** 原 良憲*** 田中 克己**,*

* 独立行政法人 情報通信研究機構 ** 京都大学大学院 情報学研究科
*** NEC Laboratories America, Inc.

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．blog 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．blog 記事は情報の即時性の観点からも情報源としても重要となりつつあり，ある意味で世論を反映した知識の宝庫であると考えている．我々は，これら blog 情報を解析に基づき Web 情報検索の信頼性を向上させることを目的とした手法を提案する．１）ニュースコンテンツに対して信頼性および適時性の高い補足情報を付加することを目的とした blog スレッドの抽出および解析，および，２）Web 検索エンジンの検索精度の向上を目的とした blog 情報に基づくトラスト値の算出方式，である．また，各々の手法に対して実験を通じて考察を行ったので報告する．

1. はじめに

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．blog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．

Web掲示板では多くの場合,書き手が不明であるため，どのようなバックグランドを持つ書き手が書いたのかが分からず，書き込み内容の信憑性を判断するための情報が十分とはいえない．一方，blogの場合は，書き手（以下，blogger）が過去にどのような記事を書いているのかを容易に把握できるので，例えば“このbloggerはUNIXに関して詳しそうなので，彼が書いたUNIX関連のエントリは信用できる”等のように，blog記事に対する評価が行いやすいといえる．つまり，閲覧するユーザは安心してblog記事を参照することができると考えている．

blogサイトの中には，単に個人の日記を綴ったものもあるが，社会問題に関して真面目に議論しているものも数多く存在する．また，多くのblog記事の更新頻度は非常に早く，対象となるニュースやイベントが起きたその日にblogエントリの書き込みが行われることも少なくない．したがって，blog記事は情報の即時性の観点からも，情報源としても重要となりつつあり，ある意味で世論を反映した知識の宝庫であると考えている．

そこで我々は，これらblog情報を解析することにより,Web情報検索の信頼性を向上させることを目的とした手法を提案する．１つ目は，１）例えばニュースコンテンツ等に対し，信頼性および適時性の高い補足情報を付加することを目的としたblogスレッドの抽出および解析[1]，２つ目は，２）Web検索エンジンの検索精度の向上を目的としたblog情報に基づくトラスト値の算出方式[2]，である．

以下，本論文の構成を示す．２節ではblogの概要および関連研究について述べる．３節ではニュースコンテンツに対し，信頼性および適時性の高い補足情報を付加することを目的としたblogスレッドの抽出および解析について述べる．４節ではWeb検索エンジンの検索精度の向上を目的としたblog情報に基づくトラスト値の算出方式について述べる．５節ではまとめと今後の方向性について述べる

2. blog の概要および関連研究

図１に典型的な blog サイトの例を示す．

blog 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 blog 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．

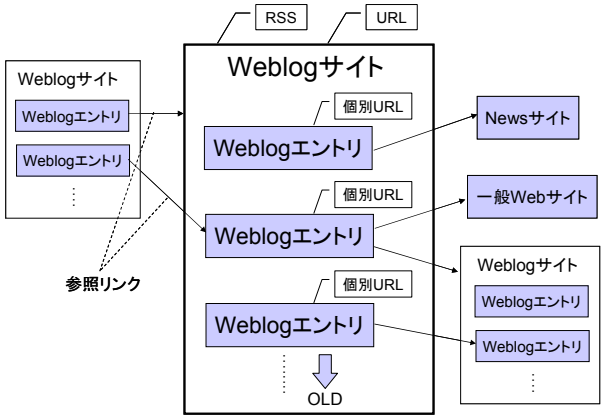

図１ 典型的な blog サイト

また，blog 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 blog 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 blog 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 blog 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 blog 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である．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 blog 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，blog サイトの定義は明確なものはないが，本研究では blog 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 blog 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 を公開していても明らかに blog サイトではないニュースサイトに関しては，これらを除外して考える．

blog に関する関連研究としては，Kumar ら[3]および，Gruhl ら[4]は，blog 空間の進化や広がりに関する調査研究を行っている．

Kumar らは，25,000 の blog サイトとその中の 750,000 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 blog 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．

Gruhl らは，11,000 以上の blog サイトにおける 400,000 以上の blog エントリについて解析している．この中で，blogspace におけるマクロな視点によるトピックの伝播の特徴付けと，ミクロな視点による個々の blog 同士のトピックの伝播の特徴付けを試みている．この中で，blogspace において内部的に発生する議論である Chatter と，外的要因により発生する Spikes という尺度を用いて，トピック伝播のモデル化を行っている．

これらの研究は，あくまでも blog による情報の広がりに注目したものであり，適時性および重要性の高い blog 記事の取得および提示方法について検討するものではない．

関連技術としては，Bulkfeeds[5] や MyblogJapan[6]等の blog 検索サービスがある．ただし，提供する blog 情報のランキングに関しては，特徴ベクトルをベースにした類似度に基づいたものであったり，単にアクセス数や被リンク数を利用したものであたったりする．つまり，blog 情報の信用度を評価した上でのランキングは行われていない．

3. blog スレッドの抽出および解析手法

本節では，blog スレッドの抽出および解析手法について述べる．blog スレッド内における blogger の役割について注目し，ある特定のトピックに関するスレッドに対して影響力のある blogger を特定する．これを元に，ニュースコンテンツに関する補足情報をユーザに提供する手法を提案する．

以下，3.1 節で blog スレッドの抽出および blog 解析について，3.2 節で blog スレッドに関する調査実験について，3.3 節で blog 解析に基づくニュースコンテンツへの補足情報の提示について述べる．

3.1 blog スレッドの抽出および blog 解析

blog スレッドの特定

blog 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，blog スレッドを「あるイベントについて意味的関連性の高い blog エントリのつながり」として扱う（図 2 参照）．図 2 の白丸が Weblog 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 blog 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．

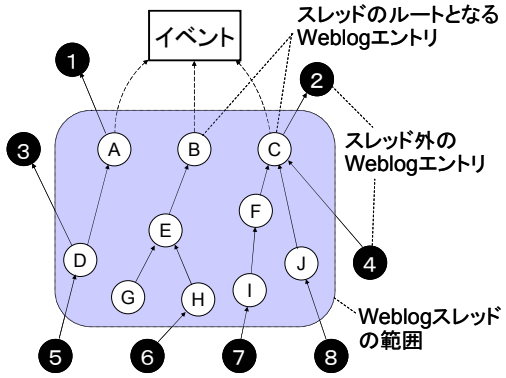

図 2　blog スレッド

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．

各 blog サイトの特性の判別

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 blog サイトの特性の判別を行うことを検討する．

blog 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 blog 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 blog サイトの特性の判別を行うことが可能と考えた．ここでは，トピック毎のスレッドの集合において，各 blog 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 blog 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 blog 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 blog 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 blog 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 blog 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 blog 投稿者である（図 3 参照）．Agitator は，自らのエントリによって，blog 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 blog 投稿者である．Agitator のエントリを監視することで，blog 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．

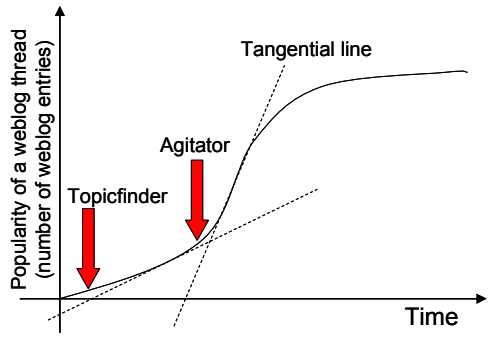

図 3 Topicfinder および Agitator

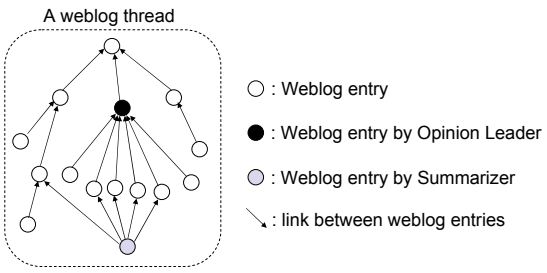

図 4 Opinion Leader および Summarizer

（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 blog エントリから参照されることが多い blog 投稿者である（図 4 参照）．図 4 では，各ノードが blog 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 blog 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 blog エントリを参照することが多い blog 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 blog 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．

3.2 blogスレッドに関する調査実験

本節では，このうち，スレッドモデルおよびblogサイトの特性について，事例に基づいた議論を行う．blog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blog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．

- blogエントリ同士の意味的な関連を考慮しない．
- データ数が十分ではなく統計的解析していない．

なお，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blogエントリの集合を抽出し，これをblogスレッドとした．取得したblog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，blogスレッドに関する考察を行った．

3.2.1 blog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 blog 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 blog 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．

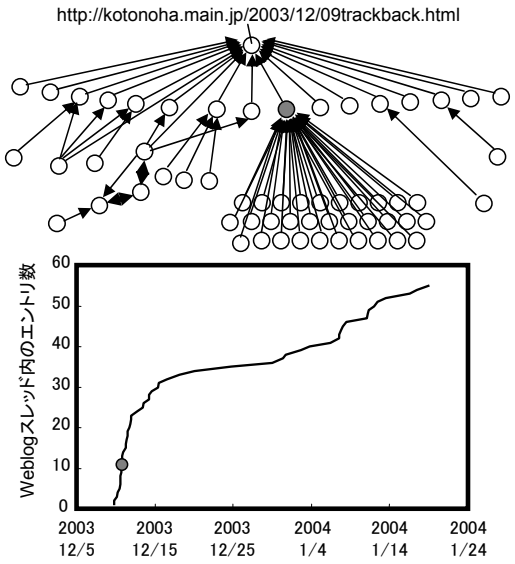

図 5 blog スレッドの調査実験結果１

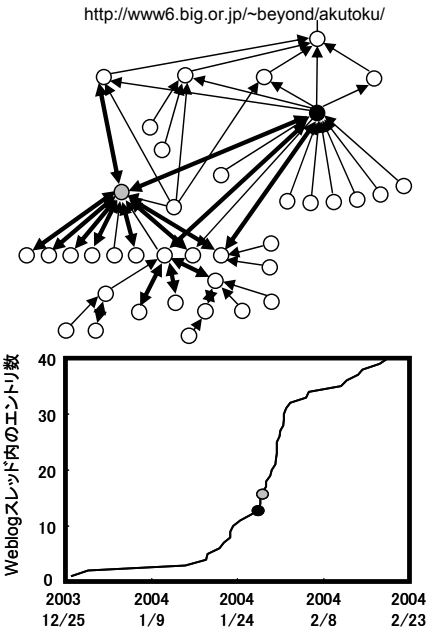

図 6　blog スレッドの調査実験結果 2

また，各図下部のblog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．

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．

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．

3.2.2　blogサイトの特性に関する考察

本節では，各blog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．3.2.1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blog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．blog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，3.2.1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，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．

3.3　blog解析に基づくニュースコンテンツへの補足情報の提示

信用度に基づく blog 情報フィルタリングを利用したアプリケーションとしては，幾つか考えられるが，本論文ではニュースコンテンツへの補足情報の提示システムへの応用を検討する（図 7 参照）．

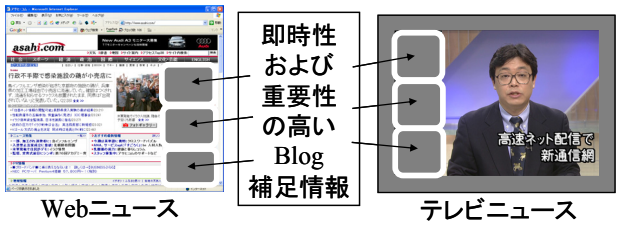

図 7　ニュースコンテンツへの補足 blog 情報の提示

ニュースコンテンツを提供するメディア媒体としては，テレビや新聞の Web サイトなどがある．これらのニュース提供者は有名であれば有名であるほど，

ユーザからの信頼度は高いといえるが，その社会的立場から発表できない内容の情報も存在することが考えられる．

これに対して，blog は基本的には個人によって執筆されるものであり，社会に対するしがらみは大きくないことに加えて，個人の独自の視点に基づく意見が書かれていることが多い．したがって，いろいろな立場の人のいろいろな見解を知るためには，blog 情報は有用であると考えている．

ただし，blog は個人が簡単に開設することができ，必ずしも質の高いものばかりではないが，本論文で提案する信用度によるフィルタリングを利用することで，即時性および重要性の高い blog 情報を取得して提示することが可能になる．

4. blog 情報に基づくトラスト値の算出方式

本節では，blog情報を用いてWebページの信頼性を表すトラスト値の算出方式と，トラスト値を用いたランキングに基づく検索システムを提案する．blog サイトを解析することにより，blog記事の書き手がどのような分野の知識について詳しいかを推定し，さらにblog記事内でbloggerが参照先のページについてどのように評価しているのかを推定する．Web ページのトラスト値を算出するために，bloggerがリンク参照しているWebページに対してどのようなコメントをつけているのかということを解析するが，このときのblogger自身の信頼性も考慮すべきである．したがって，各トピックに対して熟知しているbloggerを特定することも併せて試みる．

そこで，上記アイデアを採用した検索システムのプロトタイプの製作を行い，システムの検証を行った(図8)．

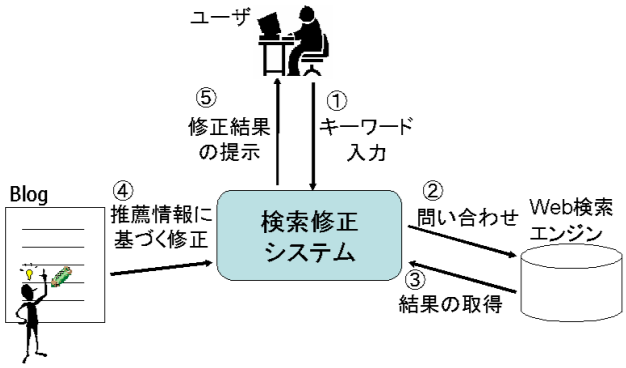

図 8　blog 情報を用いた検索結果修正の概要図

4.1　blog サイトの持つ評価情報

4.1.1　blog サイトの信頼性の推定

本テーマではblogの記事中から他のWebページへ良い評価を下しているのを取得することを目的としているが，その前にそのような blog のサイト，つまり blog の記事の書き手自身が信頼できるのかどうかを推定する必要がある．

blog サイトには，タイトル・日付・書き手の名前・記事の属するカテゴリといった blog の記事そのものに付随する情報以外にも，blog サイトの信頼性を決定するための要素が挙げられる．例えば，どれだけ多くのユーザに読まれているか(人気)，最近の注目のトピックやニュースを早く記事として載せているか(更新の早さ)，他の信頼できる blog サイトを記事中で参照し，肯定的に紹介しているか(正確さ)，他のサイトからより多く支持されているか(支持)，などが要素として挙げられる．

4.1.2　書き手の熟知度の取得

本テーマでは，blogサイトの持つ多岐にわたる特性の中から，どのようなトピックについてbloggerが詳しい知識を持っているのかという指標を熟知度として求める．あるblogサイト上のエントリのすべてについて，エントリの中からキーワードを抽出する．それらのキーワードがどのようなカテゴリに属する言葉なのかを基にして，元のblogエントリの書き手がこのカテゴリごとにどの程度詳しい知識を持っているのかを特定し，これを熟知度とする．

具体的には，各blog記事の文章を形態素解析にかけて名詞と判定された語句を抽出する．これをエントリに関するキーワードとする．次に，ある一人の書き手により書かれた記事すべてについてこのキーワードを集計しその出現頻度を取り，頻度の高い上位の語いくつかをこの書き手の特徴キーワードと定める．そして，個別の特徴キーワードごとにそれがどのようなカテゴリに属する言葉なのかを，OpenDirectory[8]等のカテゴリ検索サービスを用いて階層的情報として取得する．例えば「野球」という単語の場合，OpenDirectoryを用いた検索では「Top: World: Japanese: スポーツ: 野球」という階層的位置にあるカテゴリに属する単語であると取得できる．このようなカテゴリ情報に，元の特徴キーワードの出現頻度に応じた数値を添え，これをカテゴリ毎の詳しさの指標とする．この解析をblogの書き手ごとに行うことにより，どの書き手がどの分野についてどの程度詳しいのかというデータとして利用することができる．

4.1.3　記事からの良評価の取得

blog の記事の中では他のページへの参照が含まれるが，それらのページすべてが良い評価を与えられた上で参照されているとは限らない．そこで，各 blog 記事が参照先のページに対し肯定的な評価を下しているのかどうかを，簡易な言語解析により判断し，評価度を求める．立石らの研究[9]を基に，記事中の他ページへの参照箇所周辺で「好き」「最高」といった単語の単純なマッチング処理と否定表現の有無により，参照先のページに良い評価を与えているのかどうかを判断する．ここでは，他ページを参照している箇所からどの程度離れた出現箇所かと肯

定的表現の単語の種類により，評価の度合いを値として判断することを想定している．他の近似的手法としては，現在のリンク解析的手法と同じようにすべての参照を同じ一定の評価を下しているものと見なす場合や，blog 記事の書き手に具体的に数値として投稿してもらうなどの場合が考えられる．

4.1.4　コンテンツの信頼性の算出

複数の blog の書き手について，他ページに対しての良い評価の度合いである評価度(4.1.3 節)を合わせることにより，参照されたページのコンテンツそのものの信頼性を提示することができる(図 9)．ある一つの特定のページについて複数の書き手が評価を下している場合，その評価度から書き手の熟知度(4.1.2 節)における詳しさの度合いに応じて重み付けした正規化処理により，一位の値を求める．ある特定のカテゴリについて，blog の書き手 i の熟知度を k_i，この書き手がある特定のページに p_i の評価をつける場合，このページのこのカテゴリについてのトラスト値 $T(p)$ を定式化すると以下のようになる．

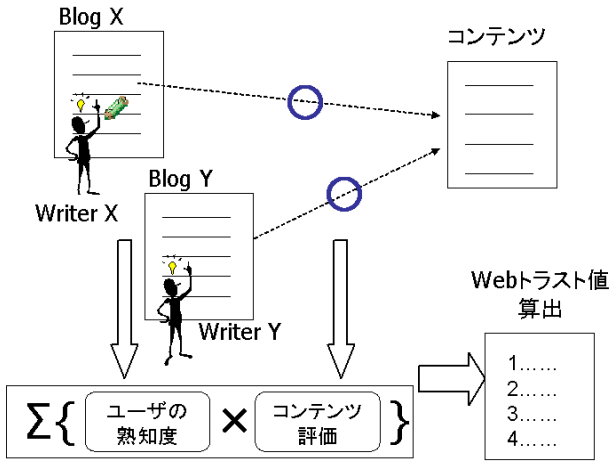

図 9　コンテンツの信頼性を表すトラスト値の算出

$$T(p)=\frac{k_1 * p_1 + k_2 * p_2 + \ldots}{k_1 + k_2 + \ldots}=\frac{\sum_i k_i * p_i}{\sum_i k_i}$$

この操作を存在するすべてのカテゴリ毎に行うことにより，コンテンツの一意の値をカテゴリ毎に求めることができる．本論文では，このような値をコンテンツに対する信頼性の一種ととらえ，トラスト値と呼ぶことにする．つまりトラスト値とは，blog サイト自体の信頼性を推定し，信頼できる blog サイトから良い評価を持って参照されたページを良いとする，コンテンツの信頼性を表す指標である．これにより例えば，野球というトピックに含まれるキーワードを記事の中で多く記す blog の書き手がいる場合，その書き手が記事中で肯定するページは野球というトピックに対してのトラスト値が高く，野球という内容についてそのページのコンテンツは信頼性が高いとなる．そして，そのようにして求めるトラスト値の具体的な利用法として，検索結果の改善に用いるという手法を提案する．これについては次の節で述べる．

4.2　blog 情報に基づく検索システムの構築

4.2.1　blog 情報を用いた検索

通常の検索エンジンでは，ユーザの入力する質問キーワード Q(*Query*)と Web ページのコンテンツの内容 C(*Content*)から，Q のキーワードが本文の中に含まれているような C を探しだし，それを各々の持つランキング手法に基づき並び替え提示している．本論文の提案する手法は，この C と Q に blog の記事情報 B を加えた中で，通常の検索エンジンの出力結果を改善することでユーザにとって有用に情報を提示するものである．blog の記事の内容や blog サイトの信頼性を吟味されることにより，blog の記事を参照先の Web ページのコンテンツ内には直接は書かれていないがコンテンツの内容をより詳しく説明するメタデータの一種であると見なせる．このように blog 情報をメタデータとして用いるための具体的な利用方法として，4.2.2 節で参照先キーワードの補完を，4.2.3 節で検索質問の拡張を説明する．

4.2.2　参照先キーワードとしての利用

blog の記事から参照している他の Web ページについての説明文章であると見なすという手法が考えられる．これは，参照先のページに直接は書かれていないが，その内容に意味的に近い用語が参照元の blog の記事中には含まれていることが多いことを利用する．例えば，ユーザが Q という検索キーワードを入力した場合，通常の検索エンジンではその Q という単語そのものが本文に含まれるページしか提示できないのに対し，この手法では，Q を含むような内容の文章である blog 記事を見つけ出し，その記事から参照されているページをユーザに提示することが可能になる．

4.2.3　検索質問の拡張

blog 情報をユーザの入力する検索質問を拡張するために利用できる．これは，一方でユーザの入力する検索キーワードを通常の検索エンジンに入力して結果を受けとり，他方で検索キーワードを基にした他の情報を付け加えて検索質問の拡張を行って，その拡張情報を基に blog 情報による検索を行い，この blog 情報による検索を利用して先の通常の検索エンジンの出力結果のページ集合から適切なページを優先し，最終的にユーザに提示しようというものである．

検索キーワードを基にした拡張情報として，具体

的には検索キーワードの単語がどのようなカテゴリに属するのかという情報を利用する．これは，4.1.2 節での手法と同様に，OpenDirectory 等のカテゴリ検索サービスを利用して取得する．今，4.1.4 節の手法により各 Web ページにはカテゴリ毎のトラスト値がつけられていると想定すると，検索キーワードから推定したある特定のカテゴリについてトラスト値の高い Web ページを優先して表示するという流れになる．これにより，最終的にユーザに提示するページの適合性を，カテゴリ的な一致によるものと blog 情報からの評価値によるものの双方から判断していることになる(図 10)．

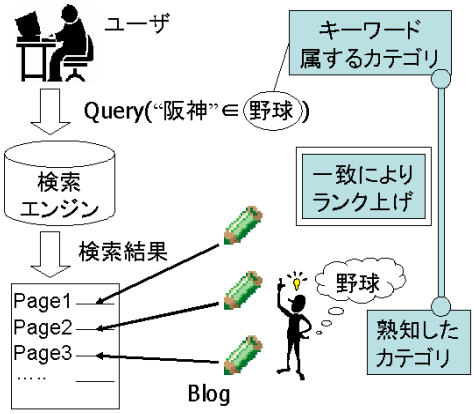

図 10　キーワードのカテゴリ一致による検索結果の改善

4.3 プロトタイプシステム

4.3.1　実装環境

図 11 にプロトタイプ画面を示す．プロトタイプで用いた blog の情報は，我々の開発している blog クローラを用いて事前に取得してきた実際の blog 記事のデータである．今回，170 の blog サイトと記事の書き手の情報，それらの人の書く 1185 個の blog 記事，それらの記事から参照されている 2061 個の Web ページの URI(同一ページを参照する時別々の blog 記事なら重複許す)の有効なデータを基に，システムを作成した． blog の記事からの単語抽出には茶筌[10]による形態素解析を用い，単語からの属するカテゴリ情報の取得には Yahoo!Japan のディレクトリ型検索システム[11]を利用した．

4.3.2　システムの処理の流れ

システムが動作するにあたって，blog 記事の書き手ごとの熟知度計算は前もって処理している．この処理の流れを以下に示す．

1. すべての blog の記事を書き手ごとに集計し取得する．
2. 記事中のすべての本文とタイトルについて茶筌による形態素解析を行い，名詞と未知語(主にカタカナ・アルファベット語)を集計する．そして出現頻度の高いものから 40 個を取得し，書き手の特徴キーワードとする．
3. すべてのキーワードについて Yahoo!Japan のディレクトリ検索を用い，キーワードの属するカテゴリ情報を最大 10 まで取得する．該当するカテゴリがない場合はその特徴キーワードは使わないこととした．
4. 上の操作により得られた書き手ごとに最大 400 個のカテゴリを書き手の熟知したカテゴリとする．

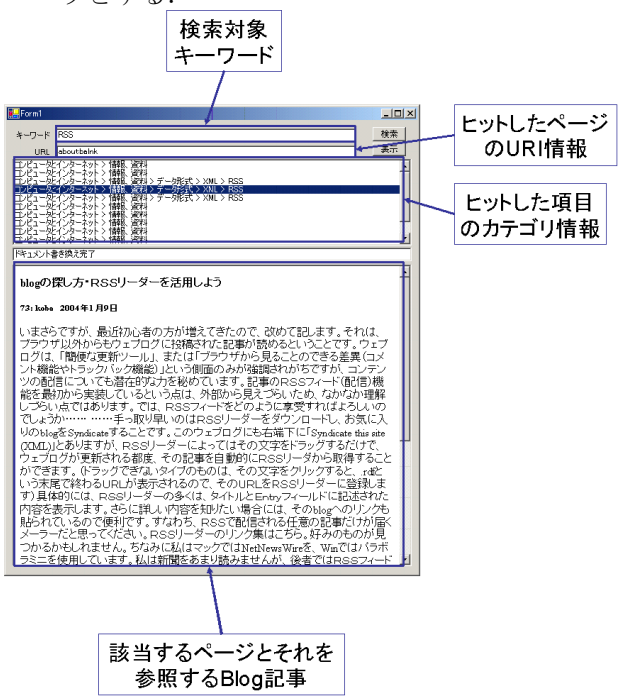

図 11　プロトタイプシステムのインターフェース

また，これらの前処理に基づくデータを利用して，システムがどのように動作しているのかを以下に示す．

1. ユーザがシステムに検索したい事項を単語で入力する．
2. 入力された単語を Yahoo!Japan のディレクトリ検索にかけ単語の属するカテゴリ情報を最大 10 個取得する．該当するカテゴリがない場合はここで処理を終了する．
3. ユーザの入力した単語で Google による検索を行い，その結果上位 500 件までを取得する．結果の各ページごとに，ページを参照するような blog 記事を探し，同時にその blog 記事の書き手も取得する．
4. 該当する blog 記事の書き手が詳しいとするカテゴリ情報すべてについて，先にユーザの入力キーワードより推定されたカテゴリ一つずつと比較を行う．このとき，カテゴリの階層構造を利用し，書き手の詳しいカテゴリ

がユーザ入力キーワードのカテゴリよりも上位に位置するものも，比較により一致したものと見なす．

5. 一致したカテゴリについて，カテゴリ情報・blog 記事のタイトルとその内容・blog の書き手の名前・参照先ページ，をセットとしてユーザに提示する．

これらの前処理とプロトタイプシステムの処理の流れを表したものを図 12 に示す．

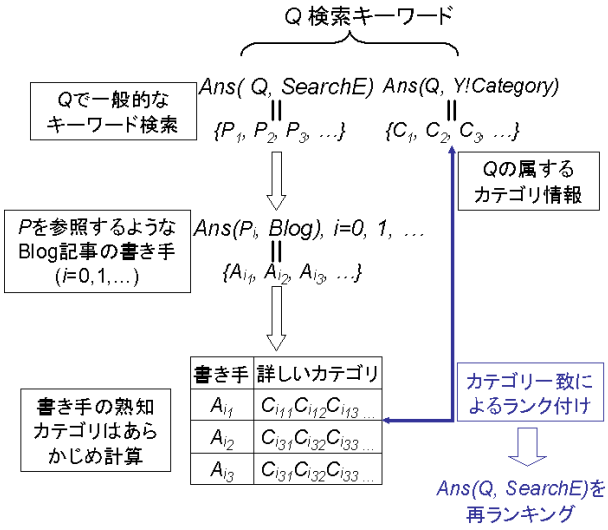

図 12　プロトタイプシステムでの処理の流れ図

ここで，プロトタイプシステム上での処理の流れの 3 番目の処理では，4.3.2 節で述べた考えを用い，以下のような blog 情報を用いた緩和も考えられる．

3. ユーザの入力したキーワードを文中に含むような blog 記事を探し，その記事から参照されたページを取得する．

プロトタイプシステムではこのような処理も比較対照として実装している．ここでは，前者を Google を介したアプローチ，後者を blog 情報を用いた緩和アプローチと呼ぶことにする．

なお今回は，各カテゴリごとに書き手がどの程度詳しいのかの処理は行わずにすべてのカテゴリについて等しく詳しいとし，記事中での参照先についてどの程度良いと評価を下しているかについても参照リンクが存在するならすべて一様に良いと評価しているものとした．

4.3.3　考察

いくつかのキーワードを基にプロトタイプシステムを通じて行った実験結果に対する考察を以下に述べる．

- Google を介したアプローチでは，カテゴリ一致まで含めると最終的に該当する結果がほとんど得られないことが多かった．これは，そもそも Google の検索結果として返すページ群と， blog の記事中で参照されるようなページ群とで，ページの数や種類が異なることが原因ではないかと思われる．
- blog 情報を用いた緩和アプローチにより，該当する検索結果の件数を大きく増やすことができた．またそれらの多くは検索キーワードと内容の深い blog 記事と参照先ページであることが多く，適した結果を返していることを確認できた．
- blog の記事の内容が，本論文で想定するような書き手の独自の視点による文章と特定の他のページへの参照という形式ではなく，例えばニュースサイトなどのページをそのまま引用しただけのものがいくつか見られた．これは，書き手が評価しているとは見なせず適しないと思われる．

5.　まとめと今後の課題

本論文では，blog 情報を解析することにより，Web 情報検索の信頼性を向上させることを目的とした手法として，1）ニュースコンテンツに対して信頼性および適時性の高い補足情報を付加することを目的とした blog スレッドの抽出および解析，2）Web 検索エンジンの検索精度の向上を目的とした blog 情報に基づくトラスト値の算出方式，を提案した．

1）信頼性および適時性の高い補足情報を付加することを目的とした blog スレッドの抽出および解析では，blog コンテンツの信頼性の推定目的とした blog 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， blog スレッドに関する調査実験および考察を行った．今後は，blog 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．

2）Web 検索エンジンの検索精度の向上を目的とした blog 情報に基づくトラスト値の算出方式では，blog 記事そのものを利用して書き手の熟知度を計り，またそれを利用して参照された Web ページの信頼性を推定する手法，およびこれらの手法を取り入れて検索エンジンをより改善する手法について提案を行うと共に，この手法を実践するプロトタイプを通じて考察を行った．今後は，Web コンテンツの信頼性の提示手法やユーザの入力するキーワードの拡張手法についてもより検討を重ね，さらなる改善に取り組む予定である．

[謝辞]
本研究の一部は，平成 16 年度科研費特定領域研究(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 : 16016247　代表 : 田中克己), および 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. ここに記し謝意を表します.

[文献]

[1] 中島伸介, 舘村純一, 日野洋一郎, 原 良憲, 田中克己, 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, DBSJ Letters, Vol.3, No.1. (掲載決定).
[2] 竹原幹人, 中島伸介, 角谷和俊, 田中 克己, Web 情報検索のための Blog 情報に基づくトラスト値の算出方式, DBSJ Letters, Vol.3, No.1. (掲載決定).
[3] 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
[4] D. Gruhl, et al:Information Diffusion Through Blogspace, The Thirteenth International World Wide Web Conference (2004). http://www2004.org/proceedings/docs/1p491.pdf
[5] bulkfeeds, http://bulkfeeds.net/
[6] MyBlogJapan, http://www.myblog.jp/
[7] TrackBack Voyager, http://holic.org/b2uvoyager.php
[8] OpenDirectory, http://dmoz.org/
[9] 立石健二: インターネットからの評判情報検索, 情報処理学会研究報告, 2001-NL-144-11, pp.75-82 (2001)
[10] 形態素解析システム茶筌, http://chasen.aist-nara.ac.jp/
[11] Yahoo!Japan, http://www.yahoo.co.jp/