社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－DBS－134 (II) (94)
2004／7／15

信用度に基づくblog情報フィルタリング

中島伸介† 田中克己††

Webを介したユーザ間の即時的情報流通の方法としてblogが注目を浴びている．blog記事は情報の即時性の観点からも，情報源として重要となりつつあるといえる．しかしながら，爆発的に増加しているblog記事の中から信用度の高いものを効果的に検索することは容易でない．そこで本研究では，即時性および重要性の観点から信用度の高いblog記事の取得および提示手法について提案する．信用度を評価する対象として，blogサイト，blogエントリ，blogスレッドを区別しつつ，投稿直後のエントリや成長過程のスレッドの見込み信用度の算出方法を検討する．さらにこの信用度に基づくニュースコンテンツへの補足情報の提示方法についても検討する．

Web Information Filtering based on Blog Trust

SHINSUKE NAKAJIMA† and KATSUMI TANAKA††

Recently, blogs become very popular as a way to circulate information between Web users. The blog articles may be getting important from the point of view of instantaneous information circulation. However, it is difficult to retrieve high trustworthy blog articles efficiently from a set of blog articles increasing explosively in WWW. Thus, we propose a method to retrieve and provide high trustworthy blog articles from the point of view of instantaneousness and importance. We investigate how to compute not only trustworthiness of blog sites, blog entries and blog threads but also expected trustworthiness of blog entries immediately after posting and growing blog threads. Moreover, we examine a way to express the trustworthy blog information for supplementary to Web and TV news contents.

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．

blogはある言い方をすれば「ユーザが自分の興味に基づいて記述したWeb上のコメント集」である．blogの書き手（以下，bloggerという）は，それぞれ自分のblogサイトを管理し，自分の意見をそのblogサイトに書き込む．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，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る．

しかしながら，blogサイトの数は，2004年6月14日時点でPING.BLOGGERS.JP[1]に登録されている数だけでも11万件を超えており，平均的にみると質

† 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology
†† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University
および
独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

が高いとは言いがたい．したがって，決して質が高いものばかりとはいえない数多くの blog 情報の中から，信用度の高いものだけを人手で探すことは不可能である．また，検索エンジンを利用するとしても，検索エンジンのクローリングにより，blog の最新記事を即時に獲得するのは困難であることから，即時性かつ重要性の高い blog 記事を効率的に獲得する手法は確立できていないといえる．

そこで本論文では，即時性および重要性の観点から信用度の高い blog 記事の取得および提示手法について提案する．

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 blog 情報の信用度評価について述べる．4節では信用度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．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

2.1 blog の概要

blog 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 Web コンテンツであり，最新ニュースをいち早く取りあげ、独自の視点で解説をする blog や、ある特定の分野に対してコラムを展開する blog、写真を掲載しそれらにコメントを付ける blog などその内容は様々である．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 blog と定義できる．「MovableType[2]」などの blog サイト構築ツールなども公開されており，また，「はてなダイアリー[3]」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 blog サイトを立ち上げるための環境が整っているといえる．

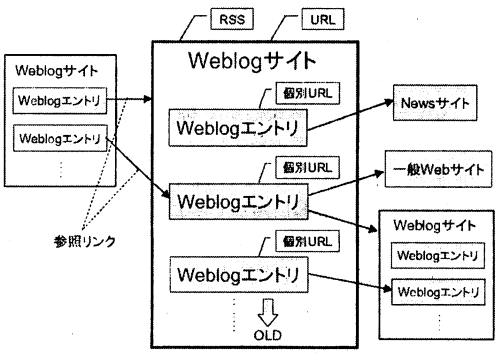

図 1　典型的な blog サイトの例

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 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，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，blog 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 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，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 blog サイトにエントリを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．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 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．

blog 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 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 が存在しても，明らかにニュースサイトであり blog 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．

2.2 関連研究および技術

- **blog による情報の広がり**

  Kumar ら[4] および，Gruhl ら[5] 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．

- **リンク構造の時間特性に着目した blog 時系列解析**

  中島らは，Web コンテンツの信頼性評価を目的とした blog 解析手法に関して提案している[6]．この中で，blog エントリが形成する blog スレッドを定義し，この blog スレッド内における blog サイトの役割の判別方法に関して議論している．blog サイトの役割としては，“Topicfinder”，“Agitator”，“Opinion Leader”，“Summarizer”などを定義し，blog スレッドのリンク構造の解析および時系列解析によってこれらの判別方法を提案している．
  この研究は，blog 解析手法としては新規性はあるものの，実質的な信頼性評価など，解析結果を踏まえた利用方法に関しては不十分な点が多い．

- **blog 情報に基づくトラスト値の算出方式**

  竹原らは，blog サイトが，参照している Web コンテンツに対して何らかの評価を示しているケースに着目し，blog 情報に基づく Web コンテンツのトラスト値の算出方式を提案している[7]．この中で blog サイトの熟知度と，blog エントリ内での評価度という指標を提案し，これに基づいて Web コンテンツのトラスト値を定義している．
  この研究は，blog エントリ内の Web コンテンツに対する評価を利用した，検索エンジン結果の修正方法を提案しているものであり，重要な blog データそのものを取得および検索しようとするものではない．

- **blog 検索サービス**

  関連技術として，Bulkfeeds[8] や MyBlogJapan[9] 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 情報の信用度評価

本節では，blog 情報の信用度評価について述べる．本論文における blog 情報の信用度とは，あるトピックに関してその blog 情報が信用するに値するかどうかという観点で評価した一つの指標である．扱うトピック毎に評価するので，ある blog 情報に対して“UNIX に関する信用度は高いが，Windows に関する信用度は高くない”という評価が可能である．信用度の評価対象となるのは，blog サイト，blog エントリ，blog スレッドである．各々の信用度は異なるため，これらを個別に扱うことで，blog 情報のより詳細な信用度評価が可能になると考えた．

blog スレッドとは，blog エントリ同士が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ものである．本研究では，blog 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 blog エントリのつながり」として扱う（図 2 参照）．

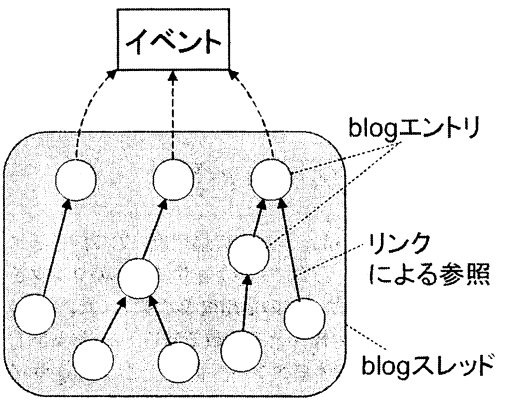

図 2 blog スレッド

図 2 中の白丸が blog スレッドを形成しているエントリである．スレッド内のエントリのうち，イベントを直接参照している blog エントリのみ，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．

3.1 blog 情報の信用度

本節では，blog 情報の算出式について説明する．

3.1.1 blog サイトの信用度

blog サイトのあるトピック X に関する信用度に関する仮説を以下に示す．これ以後議論する信用度は，あるトピック X に関するものとする．

- 信用度の高い blog エントリを数多く保持していれば，その blog サイトの信用度は高い．
- 信用度の高い blog サイトからのリンクが多ければ，その blog サイトの信用度は高い．

以上の仮説に基づいて提案する blog サイトの信用度算出式を示す．(図 3 参照)

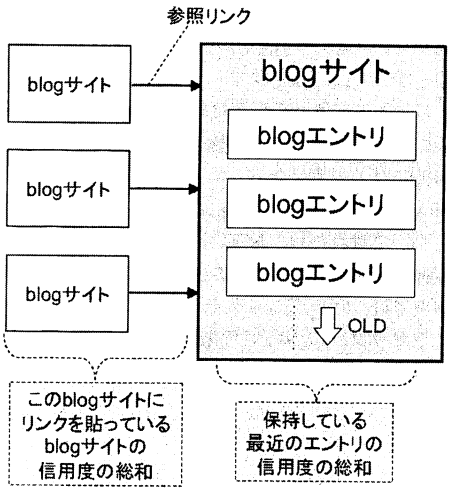

図 3 blog サイトの信用度の要素

$$T_s = \alpha \cdot \sum^{n_e}(x_t \cdot T_e) + \beta \cdot \sum^{l_s} T'_s \quad —(1)$$

ただし，T_s は blog サイトの信用度，T_e は blog エントリの信用度，T'_s はこの blog サイトへのリンクを有する他の blog サイトの信用度を示す．また，α，β は 0 から 1 の間の値をとる係数である．n_e および l_s は，それぞれ保持するエントリ数と，この blog サイトにリンクを貼っている blog サイト数であり，x_t は，時間の経過と共に減衰する係数である．

式 (1) の第一項は，保持するエントリの数と信用度により形成されている．時間経過と共に減衰する係数がかかっているため，古いエントリの重みは小さくなる．つまり，信用度の高いエントリを最近多く追加した blog サイトの信用度が高くなる．第二項は，この blog サイトへリンクを貼っている blog サイトの信用度の総和に係数 β をかけたものである．ただし，この時のリンクは blog サイトのトップページからのもののみとし，blog エントリからのリンクは含めない．

3.1.2 blog エントリの信用度

blog エントリの信用度に関する仮説を以下に示す．

- 信用度の高い blog サイトのエントリは，信用度が高い．
- 信用度の高い blog エントリからの被リンクが多ければ，その blog エントリの信用度は高い．

以上の仮説に基づいて提案する blog エントリの信用度算出式を示す．(図 4 参照)

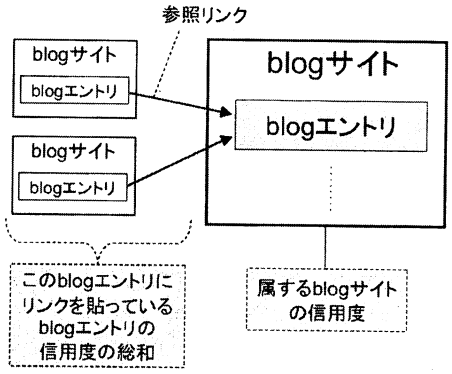

図 4 blog エントリの信用度の要素

$$T_e = \alpha \cdot T_s + \beta \cdot \sum^{l_e} T'_e \quad —(2)$$

ただし，前述の通り，T_s は blog サイトの信用度，T_e は blog エントリの信用度である．T'_e はこの blog エントリに対してリンクを貼っている blog エントリのトピック X に関する信用度を示す．また，α，β は 0 から 1 の間の値をとる係数である．l_e は，この blog エントリにリンクを貼っているその他の blog エントリ数である．

式 (2) の第一項は，1 つ目の仮説そのものであり信用度の高い blog サイトであればエントリの信用度も高くなる．第二項は，この blog エントリへリンクを貼っている blog エントリの信用度の総和に係数 β をかけたものであり，信用度の高い他のエントリからのリンクが多ければ，そのエントリの信用度は高くなる．

3.1.3 blog スレッドの信用度

blog スレッドの信用度に関する仮説を以下に示す．

- 信用度の高い blog エントリが数多く参加していれば，その blog スレッドの信用度は高い．

以上の仮説に基づいて提案する blog スレッドの信用度算出式を示す．(図 5 参照)

$$T_t = \alpha \cdot \sum^{n_e} T''_e \quad —(3)$$

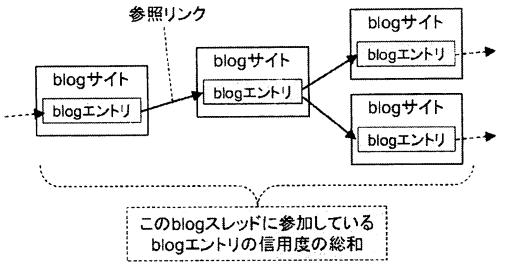

図 5 blog スレッドの信用度の要素

ただし，T_t は blog スレッドのトピック X に関する信用度，T_e'' はこの blog スレッドに参加している blog エントリのトピック X に関する信用度を示す．また，α は 0 から 1 の間の値をとる係数である．

式 (3) の右辺は，仮説そのものであり信用度の高い blog エントリが多く参加している blog スレッドのトピック X に関する信用度は高くなる．

3.2 blog 情報の見込み信用度

前節で blog エントリの信用度についてその算出式を提案したが，blog エントリが投稿された時点では，このエントリへのリンクは存在しないので，信用度の評価ができない．加えて，スレッドの成長の初期段階では，エントリの信用度が不明であるので，スレッドの信用度も評価することができない．

したがって，本節では blog エントリが投稿された時点での見込み信用度，blog スレッドの初期段階の見込み信用度を算出する方法について述べる．

3.2.1 blog エントリの見込み信用度

3.1.2 節で述べたように，blog エントリの信用度は，属する blog サイトの信用度，他のエントリからの被リンク数とそのエントリの信用度を元に算出される．ただし，投稿直後の blog エントリへのリンクは存在しないので，投稿直後の blog エントリの見込み信用度の算出式を，属する blog サイトの信用度に基づいて算出する．算出式を以下に示す．この算出式はエントリ投稿から時間 t_s を経過するまでとし，その後は 3.1.2 節のエントリ信用度の算出式を採用する．

$$expected\ T_e = \alpha \cdot T_s + \beta \cdot \left(\frac{t_s - t}{t}\right) \cdot \bar{T}_e' \cdot \bar{l}_e \quad -(4)$$

ただし，$expected\ T_e$ は，投稿直後の blog エントリの見込み信用度を示す．$\bar{T}_e'$ および $\bar{l}_e$ は，このエントリが属する blog サイトの過去のエントリに対し，リンクを貼っていたエントリの平均信用度と平均個数を示す．α は係数である．

式 (4) より，投稿直後のエントリの見込み信用度は属する blog サイトの信用度と，他のエントリからの被リンク元の平均信用度と平均エントリ個数に基づいて算出する．

3.2.2 blog スレッドの見込み信用度

3.1.3 節で述べたように，blog スレッドの信用度は，信用度の高い blog エントリが数多く参加しているかどうかが大きな要素となる．したがって，成長段階の blog スレッドの見込み信用度の算出式を，このスレッドに参加している blog エントリの信用度および見込み信用度を用いて，以下のように定義する．成長期間はスレッド生成から時間 t_g を経過するまでとし，その後は 3.1.3 節のスレッド信用度の算出式を採用する．

$$expected\ T_t = \alpha \cdot \sum^{n_e'} T_e'' + \beta \cdot \left(\frac{t_g - t}{t}\right) \sum^{n_e'} (expected\ T_e) \quad -(5)$$

ただし，$expected\ T_t$ は blog スレッドの見込み信用度，$expected\ T_e$ はこの blog スレッドに参加している blog エントリの見込み信用度を示す．また，t はスレッド形成後の経過時間，n_e' はこのスレッドに参加しているエントリ数を示す．t はスレッドの生成からの経過時間であり，t_g は blog スレッドの見込み成長期間である．

式 (5) 右辺の第一項は，スレッドに参加しているエントリの信用度の和である．第二項は，スレッドに参加しているエントリの見込み信用度の和を，経過時間 t が大きくなるにつれて重みを小さくしたものである．つまり，初期段階の blog スレッドの見込み信用度は，参加しているエントリの信用度の和に，その後の成長見込み分を加えることで算出する．

4. 信用度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 6 参照）．

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

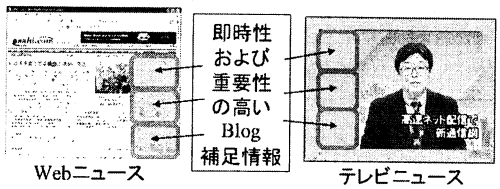

図 6　ニュースコンテンツへの blog 補足情報の提示

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．

4.1　システムの概要

提案するシステムには，即時性および重要性の観点から信用度の高い blog 記事の取得および提示を行うために，blog 情報の収集および信用度算出機能と，信用度に基づく blog 情報の検索および提示機能を保持させている（図 7 参照）．

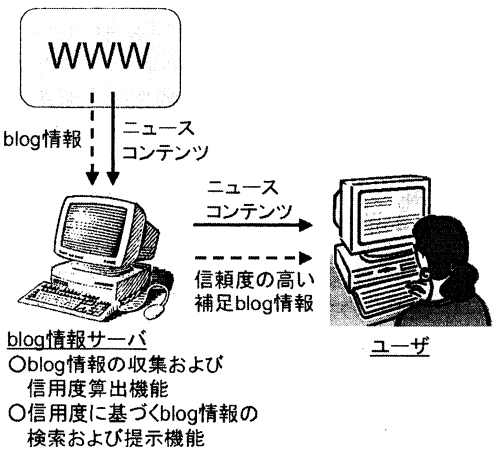

図 7　システムの概要

以下に，blog 情報の収集および信用度算出の手順を示す．

(1)　blog 情報サーバは，予め登録している RSS 集合に基づいて blog サイトを巡回して，サイト毎の blog エントリを収集する．

(2)　取得した blog エントリに対してリンク抽出を行い，blog スレッドを検出する．

(3)　取得したデータから，カテゴリ毎の blog サイト，blog エントリ，blog スレッドの信用度を算出する．

次に，ニュースコンテンツに対する信用度に基づく blog 情報の検索および提示の手順を示す．

(I)　ユーザが blog 情報サーバを介して Web ニュースサイトを閲覧する．

(II)　blog 情報サーバはユーザが閲覧している記事のトピックを抽出すると共に，他のニュースサイトでの同一のニュース記事を検索する．

(III)　抽出されたトピックに関して信用度が高い blog サイトから，対象のニュースサイトを直接参照している blog エントリ，およびこれらニュースに関して議論しているスレッドと，スレッド上の blog エントリを収集する．

(IV)　収集した blog エントリおよび blog スレッドを，ニュースコンテンツと共にブラウザの補足情報ウィンドウに提示する．

現在，システムを実装中であり，blog エントリの収集および解析と blog スレッド抽出までを実現している．現在，RSS を 10 万件以上登録しており，blog エントリを 150 万件以上取得している．

4.2　今後の課題

本節では，ニュースコンテンツへの補足情報の提示システムの実現のための今後の課題について述べる．

- **トピック抽出手法**
  各々の blog サイトは，扱うトピックに関して，得意分野および不得意分野が存在する．したがって，blog 情報の信用度を抽出する際にはどのトピックに関する信用度であるのかを明らかにする必要がある．トピック抽出に関する従来技術を押さえた上で，本システムに適した手法を検討する．
- **blog スレッド抽出**
  blog エントリのリンク参照関係に基づいた意味的つながりである，blog スレッドの信用度を算出するために，blog スレッドの抽出が必要である．blog エントリにおいて，タグ解析により，記事内のリンクのみを抽出する必要があるが，blog サイトによってそのフォーマットが異なるため，様々なパターンの blog サイトを想定した解析アルゴリズムを構築する必要がある．
- **blog 情報の信用度算出**
  blog サイト，blog エントリ，blog スレッドの信用度算出では，それぞれが影響し合っているため

に，繰り返し計算を行うことで値を収束させる必要がある．効率的に値が収束するような手法を検討する必要がある．

5. おわりに

本論文では，即時性および重要性の観点から信用度の高い blog 記事の取得および提示手法について検討した．以下に，本論文のまとめを示す．

- blog サイト，blog エントリ，blog スレッドの信用度について定義し，これらの算出方法を提案した．
- 投稿直後の blog エントリや，成長過程の blog スレッドに対する見込み信用度の算出方法を提案した．
- 信用度に基づく blog 情報フィルタリング手法を，ニュースコンテンツ補足情報提示システムに応用することを提案し，その実現方法について検討した．

今後は，プロトタイプの実装を行うと共に，これを用いた評価実験を行う予定である．

参考文献

1) PING.BLOGGERS.JP, http://ping.bloggers.jp/
2) movabletype.org, http://www.movabletype.org/
3) はてなダイアリー, http://d.hatena.ne.jp/
4) 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
5) D. Gruhl, et al: "Information Diffusion Through Blogspace", *The Thirteenth International World Wide Web Conference (2004).* http://www2004.org/proceedings/docs/1p491.pdf
6) 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, 2004 年 6 月（掲載予定）．
7) 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, 2004 年 6 月（掲載予定）．
8) bulkfeeds, http://bulkfeeds.net/
9) MyBlogJapan, http://www.myblog.jp/