

信用度に基づく blog 情報フィルタリング

Web Information Filtering based on blog Trust

中島 伸介[▼] 田中 克己[▲]

Shinsuke NAKAJIMA Katsumi TANAKA

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。

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

Web を介したユーザ間の即時的情報流通の一つとして blog が広まりつつあり、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、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、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。

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8月20日時点で PING.BLOGGERS.JP [1] に登録されている数だけでも 21 万件を超えており、平均的にみると質が高いとはいえない。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。

2. blog の概要および関連研究

2.1 blog の概要

図 1 に典型的な blog サイトの例を示す。

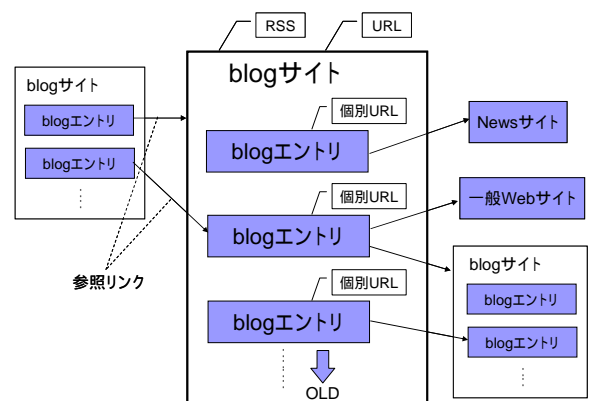


図 1 典型的な blog サイト

Fig.1 A Typical blog Site

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 を迎れば、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 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

[▼] 正会員 独立行政法人情報通信研究機構
snakajima@nict.go.jp

[▲] 正会員 京都大学大学院情報学研究科
独立行政法人情報通信研究機構
tanaka@dl.kuis.kyoto-u.ac.jp

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 を配信しているものを blog と扱うことにしている。

2.2 関連研究および技術

2.2.1 blogによる情報の広がり

blog 解析に関する関連研究としては、Kumar らや Gruhl らが、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 [2]。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 [3]。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。

2.2.2 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 [4]。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。

2.2.3 blogによる情報の広がり

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 コンテンツの信頼値の算出方式を提案している [5]。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 コンテンツの信頼値を定義している。

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。

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)。

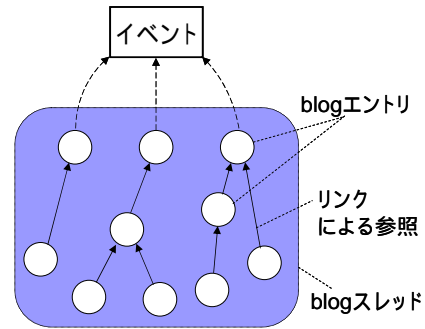


図 2 blog スレッド

Fig.2 A blog Thread

3.1 blog 情報の信用度

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)

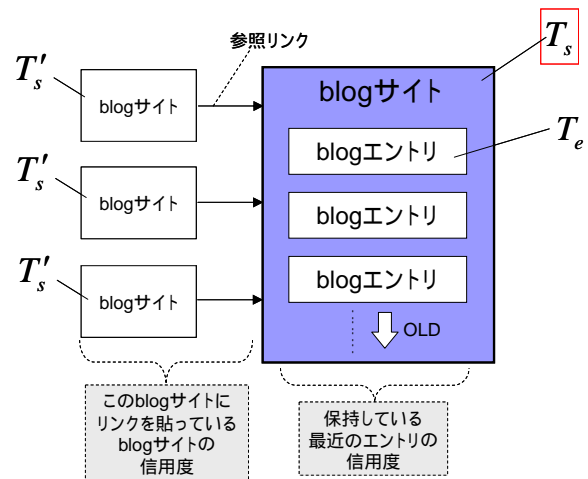


図 3 blog サイトの信用度の要素

Fig.3 Factor of Trust of a blog Site

$$T_s = \alpha_1 \cdot \sum_{n_e} (x_i \cdot T_e) + \beta_1 \cdot \sum_{l_s} T'_s \quad (1)$$

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_2 は 0 から 1 の間の値をとる係数である。

n_e および l_e は、それぞれ保持するエントリー数と、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。

3.1.2 blogエントリーの信用度

blog エントリーの信用度に関する仮説を以下に示す。

- 信用度の高いblog エントリーからの被リンクが多ければ、そのblog エントリーの信用度は高い。

上記仮説に基づいて提案する blog エントリーの信用度算出式を示す。(図4参照)

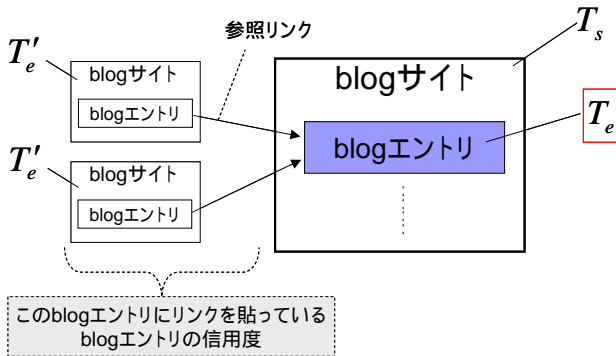


図4 blog エントリーの信用度の要素

Fig.4 Factor of Trust of a blog Entry

$$T_e = \alpha_2 \cdot \sum_{l_e} T'_e \quad - (2)$$

ただし、 T'_e はこのblogエントリーに対してリンクを貼っているblogエントリーのトピックXに関する信用度を示す。 l_e は、このblogエントリーにリンクを貼っているその他のblogエントリー数である。

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 エントリー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)。

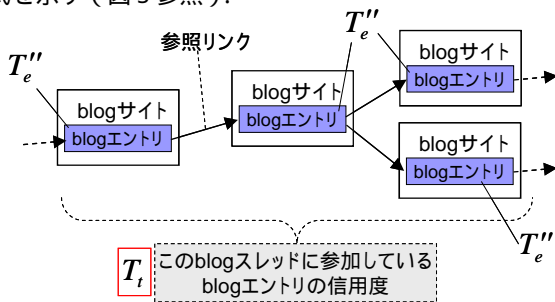


図5 blog スレッドの信用度の要素

Fig.5 Factor of Trust of a blog Thread

$$T_t = \alpha_3 \cdot \sum_{n'_e} T''_e \quad - (3)$$

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エントリーのトピ

ックXに関する信用度を示す。

3.2 blog 情報の見込み信用度

前節で blog エントリーの信用度についてその算出式を提案したが、blog エントリーが投稿された時点では、このエントリー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、エントリー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 エントリー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エントリーの見込み信用度

3.1.2 節で述べたように、blogエントリーの信用度は、他のエントリーからの被リンク数とそのエントリーの信用度を元に算出される。ただし、投稿直後のblogエントリーへのリンクは存在しないので、投稿直後のblogエントリーの見込み信用度の算出式を、属するblogサイトが保持する過去のエントリー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ー投稿から時間 t_s を経過するまでとし、その後は3.1.2節のエントリー信用度の算出式を採用する。

$$expected T_e = \alpha_4 \cdot \left\{ \sum_{l_e} T'_e + \left(\frac{t_s - t}{t_s} \right) \cdot \bar{T}'_e \cdot \bar{l}_e \right\} \quad - (4)$$

ただし、 $expected T_e$ は、投稿直後のblogエントリーの見込み信用度を示す。 T'_e は、このblogエントリーに対してリンクを貼っている他のblogエントリーの信用度を示す。 \bar{T}'_e および \bar{l}_e は、このエントリーが属するblogサイトの過去のエントリーに対し、リンクを貼っていたエントリーの平均信用度と平均個数を示す。

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エントリーが数多く参加しているかどうか大きな要素となる。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エントリー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

$$expected T_t = \alpha_5 \cdot \left\{ \sum_{n'_e} (expected T''_e) + \left(\frac{t_g - t}{t_g} \right) \cdot \bar{T}''_e \cdot \bar{n}'_e \right\} \quad - (5)$$

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エントリー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、 \bar{T}''_e および \bar{n}'_e は、このスレッドにエントリーを供給しているblogサイトの過去のエントリーの平均信用度と平均個数を示す。 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。

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)。

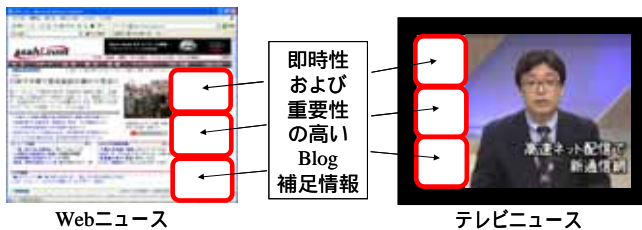


図6 ニュースコンテンツへのblog情報の提示

Fig.6 Supplementary blog Information to News Contents

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。

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)。

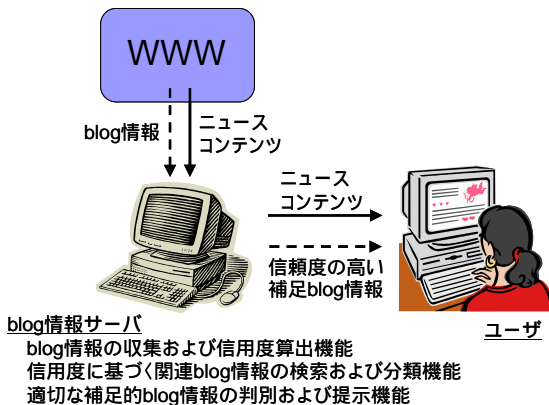


図7 システムの概要

Fig.7 System Setup

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。

1. 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。

2. 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。
3. 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。
4. 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20万件以上登録しており、blogエントリを300万件以上取得している。

5. まとめ

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の検討および提案を行った。この中で、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。また、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。さらに、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。

[文献]

- [1] PING.BLOGGERS.JP, <http://ping.bloggers.jp/>
- [2] 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>
- [3] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: "Information Diffusion Through blogspace", The Thirtieth International World Wide Web Conference (2004). <http://www2004.org/proceedings/docs/1p491.pdf>
- [4] 中島伸介, 館村純一, 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, pp.109-112, 2004年6月
- [5] 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, pp.101-104, 2004年6月

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。

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。