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

A Trust Value Calculation Method for Web Searching based on Blogs

竹原 幹人♥ 中島 伸介♦
角谷 和俊◆ 田中 克己♠

Mikihito TAKEHARA Shinsuke NAKAJIMA
Kazutoshi SUMIYA Katsumi TANAKA

本論文では，Blog サイトの解析により Blog 投稿者が詳しい知識を持つ分野を推定するとともに，この推定に基づいた Web コンテンツのトラスト値の算出手法を提案する．また，このトラスト値の利用と検索質問の拡張に基づき試作した検索システムについて述べる．

リンク構造解析に基づく Web ページのランキングは，一般に有名なサイトのランクが高くなるが，マイナーではあるが有用なサイトを検索することが困難である．また，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リンク群は未整備であることが多く，従来手法は適用できない．そこで，Web コンテンツの評価を与えている Blog サイトを解析することで，ユーザに有用なサイトの推薦が可能な手法を提案する．

In this paper, we propose a way to guess the field which Blog contributors know very well by analysis of the Blog sites, and a way for evaluating trust degree of Web contents. Moreover, we describe our prototype search engine which uses the proposed trust-degree computation method together with a query expansion technique.

Usually, the ranking score of famous Web sites is computed high by way of ranking methods based on link structural analysis. Although those ranking methods are useful, it is not suitable for finding minor but useful Web sites. Furthermore, in the situation in which real time information dissemination is increasing, links are not be supported, so conventional techniques are not effectively used. In this paper, we propose a way to recommend useful Web sites by analyzing Blog sites which have given evaluations for the corresponding Web sites.

1. はじめに

Web検索エンジンは，多くのユーザに利用されているが，幾つかの技術的課題を抱えている．一つ目は，検索エンジンで使うためのインデックス構築のために手間と時間がかかることである．Googleで用いられているランキング手法であるPageRankはその計算に数日を要する[3]ため，最近更新されたWebページや今話題となっているトピックを完全に網羅することができない．二つ目は，現在の検索エンジン構築手法で主流となっている解析に基づくランキングアルゴリズムでは，有名なページが上位に提示されやすく，マイナーではあるがあるコミュニティに対しては重要なページが上位にランキングされないという問題がある．

そこで本論文では，Blog情報を用いたWebページの信頼性を表すトラスト値の算出方式と，トラスト値による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のかの推定もでき，さらにBlogの文章が参照しているページに対する評価と取れるため，書き手が参照先のページについてどの程度の評価を下しているのかも推定できる．これにより，単にBlog記事で評価されたページというものではなく，どのようなバックグランドを持った人が良い評価を下したページであるのかという形でユーザに推薦することができる．本論文では，これらの考えに基づき検索システムのプロトタイプの製作と検証を行った(図1)．

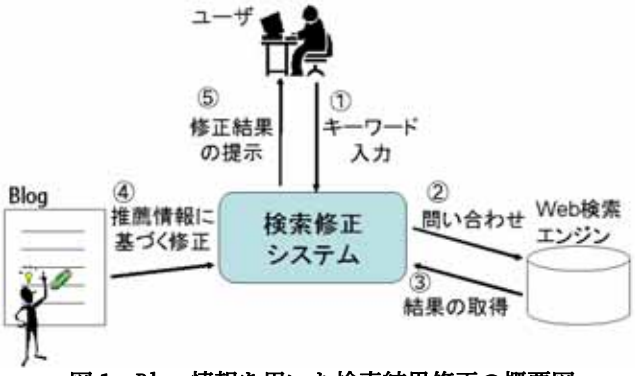

図 1 Blog 情報を用いた検索結果修正の概要図
Fig.1 Retrieval System Image based on Blogs

2. Blog サイトの持つ評価情報

2.1 Blog サイトの信頼性の推定

本論文では 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[1]．

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(すばやさ)，記事中で参照するコンテンツを他の信頼できる Blog サイトも紹介しているのかどうか(正確さ)，他のサイトからより多く支持されているか(参照)，などが要素として挙げられる．

また Blog では，記事中から他の Blog サイトへのリンクを張ることもよく見られ，ここからコミュニティの結びつきが生まれ，コミュニティ内で特定のトピックに対する一連の議論の流れが起こる．それらの議論の中から，議論の最初の起

♥ 学生会員 京都大学大学院情報学研究科修士課程
takehara@dl.kuis.kyoto-u.ac.jp
♦ 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
◆ 正会員 兵庫県立大学環境人間学部環境人間学科
sumiya@shse.u-hyogo.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

点となる記事を書いた，一連の議論をより大きく盛り上げる文章を書いた，などの議論に影響を与えた Blog サイトの特性を計り，Blog サイトの各特性を反映したプロファイルを構築することが考えられる．これを利用し，「最近盛んに議論されているトピックが欲しい」などの側面を利用した検索というものも考えられる。

2.2 書き手の熟知度の取得

本論文では，Blogサイトの持つ多岐にわたる特性の中から，どのようなカテゴリの知識についてBlogの書き手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という手法を取る．あるBlogサイト上の一人の書き手による記事すべてについて，記事の中から複数のキーワードを抽出して，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という情報を基にして，元のBlog記事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定めこれを熟知度とする．

具体的には，まず，各Blog記事の文章を形態素解析等にかけて名詞と判定された語句を抽出しこれを記事についてのキーワードとする．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[4]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．

2.3 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[2]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．

2.4 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(2.3 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 2)．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(2.2 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．

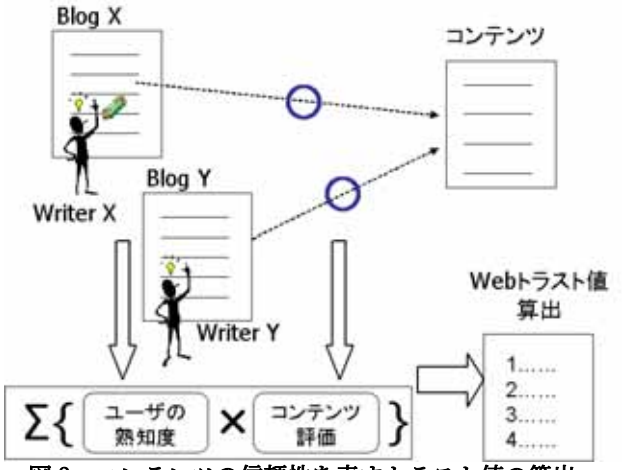

図 2 コンテンツの信頼性を表すトラスト値の算出
Fig.2 Trust Value Calculation of Contents

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

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．

3. Blog 情報に基づく検索システムの構築

3.1 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，3.2 節で参照先キーワードの補完を，3.3 節で検索質問の拡張を説明する．

3.2 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．

3.3 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，2.2 節での手法と同様に，OpenDirectory[4]等のカテゴリ検索サービスを利用して取得する．今，2.4 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 3).

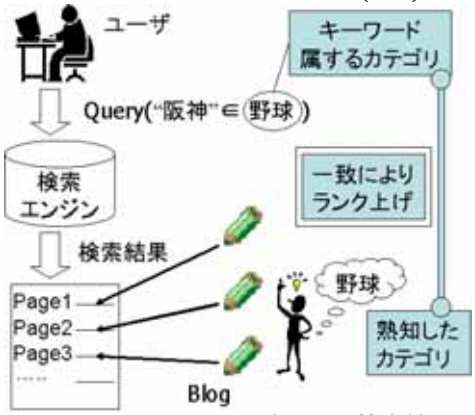

図 3 キーワードのカテゴリ一致による検索結果の改善
Fig.3 Modification of Search Result on Keyword Category

4. プロトタイプシステム

4.1 実装環境

図 4 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(別の Blog 記事が同一 URI 参照時重複許す)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[6]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[5]を利用した．

4.2 システムの処理の流れ

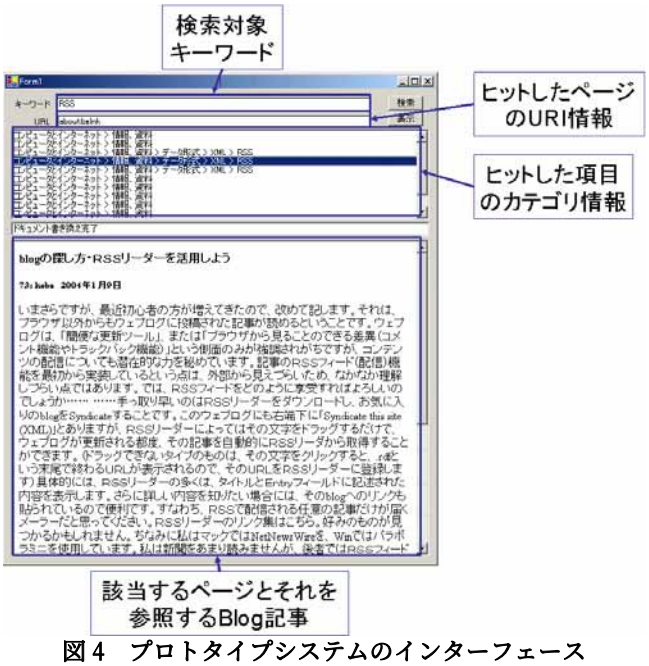

図 4 プロトタイプシステムのインターフェース
Fig.4 System User Interface

システムの動作に前もって行っている，Blog 記事の書き手ごとの熟知度計算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[6]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[5]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．

次に，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．また流れ図を図 5 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[5]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．

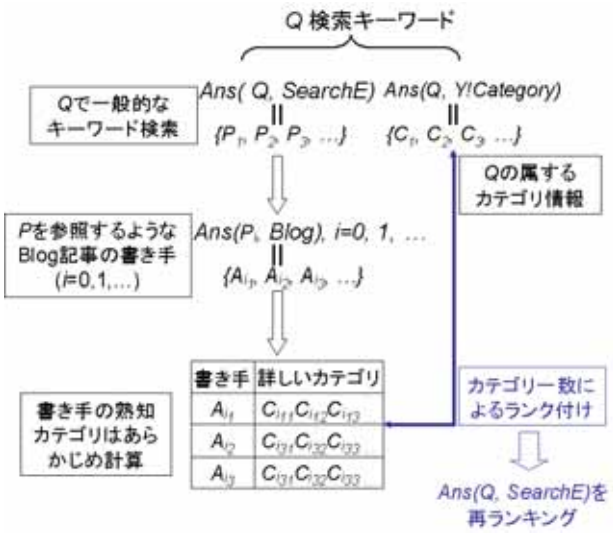

図 5 プロトタイプシステムでの処理の流れ図
Fig.5 Prototype System Flow Figure

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，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 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. おわりに

Blog サイトのカテゴライズや新着記事のあるサイトの提示・どれだけ多くの他の Blog サイトに紹介されているかなどを一元管理して提示するための Blog ポータルサイトが現在いくつか立ち上がってきているが[7][8]，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はまだ提案されていない．また，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ような試みもまだなされていない．本論文ではそのための手法について提案を行い，また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．今後も，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 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」によります．ここに記して謝意を表します.

[文献]

[1] Sepandar D. Kamvar: The EigenTrust Algorithm for Reputation Management in P2P Networks, WWW2003 http://www2003.org/cdrom/papers/refereed/p446/p446-kamvar/
[2] 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
[3] Sepandar D. Kamvar: Extrapolation Methods for Accelerating PageRank Computations, WWW2003 http://www2003.org/cdrom/papers/refereed/p270/kamvar-270-xhtml/
[4] OpenDirectory http://dmoz.org/
[5] Yahoo!Japan http://www.yahoo.co.jp/
[6] 形態素解析システム茶筌 http://chasen.aist-nara.ac.jp/
[7] Bulkfeeds http://bulkfeeds.net/
[8] BlogPeople http://www.blogpeople.net/

竹原 幹人 Mikihito TAKEHARA
京都大学大学院情報学研究科修士課程在学中．2004 年京都大学工学部情報学科卒業．日本データベース学会学生会員.

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.

角谷 和俊 Kazutoshi SUMIYA
兵庫県立大学環境人間学部環境人間学科教授．1998 年神戸大学大学院自然科学研究科博士後期課程修了，工学博士．マルチメディアデータベース，データ放送の研究開発に従事．IEEE Computer Society, ACM, 映像情報メディア学会，情報処理学会，日本データベース学会等各会員.

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学教授．1976 年京都大学大学院博士前期課程修了, 工学博士. 主にデータベース，マルチメディアコンテンツの処理の研究に従事．IEEE Computer Society , ACM , 人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員.