DEWS2004 I-2-02

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

竹原　幹人† 　中島　伸介†† 　角谷　和俊†† 　田中　克己††

† 京都大学工学部情報学科　〒606-8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科 社会情報学専攻　〒606-8501 京都府京都市左京区吉田本町
E-mail: †{takehara,nakajima,sumiya,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　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評価手法を提案する．また，このトラスト値を利用した検索質問と検索結果の修正手法により，検索エンジンの改善を行う．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，生成されるコンテンツやこれらを結ぶ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．

キーワード　Blog, Web, 情報検索, 質問修正

A Trust Value Calculation Method for Web Searching based on Blogs

Mikihito TAKEHARA†, Shinsuke NAKAJIMA††, Kazutoshi SUMIYA††, and Katsumi TANAKA††

† Infomatics of the faculty of Engineering, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Infomatics, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{takehara,nakajima,sumiya,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　In this paper, we presume the field which Blog contributors know well by analysis of the Blog sites, and propose a trust evaluation technique for Web contents based on this presumption. Moreover, we improve a search engine by the modification technique of search queries and search results by useing the trust value. Although the rank of famous sites is usually high by way of ranking methods based on link structural analysis, it is not fit for searching minor but useful sites. Furthermore, in the situation in which real time information dissemination is increasing, contents and their links are not be supported, so conventional technique is not efficient. Therefore, we propose a technique of ability to recommend useful sites in analyzing Blog sites which have given evaluations for Web contents.

Key words　Blog, Web, information retrieval, query modification

1. はじめに

近年 Web 上における双方向の情報のやりとりが活発化してきており，電子掲示板や Wiki [6] といた Web を訪れるユーザ自身がコメントを書き込み参加するタイプのシステムが広く利用されてきている．これらのシステムの上では，コミュニケーションにとどまらず，例えば商品の評判情報に関する意見交換をすることも少なくない．また最近では，他の Web ページを参照し書き手の独自の視点により積極的にそのページを批評する内容の記事を記すというタイプの Blog(Weblog) が広まりつつある．

一方，Web 上から広く情報を探し出すために使われている検索エンジンには問題点が挙げられる．一つ目には，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が挙げられる．Web 上には日々新しいページ群が追加されるため，その膨大な Web 空間すべてを回り尽くすのは困難である．また，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 [5]．このため，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追うことができない．

二つ目として，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，有名なページのみが上位に提示されやすいという問題点がある．リンク解析

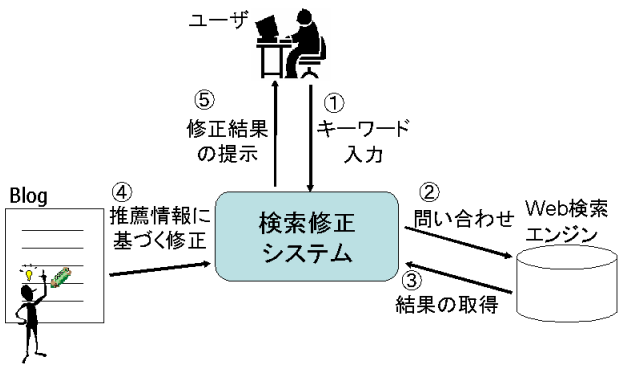

図 1　Blog 情報を用いた検索結果修正の概要図

型手法では，Web 上で他の多くのページからリンクされているページであるほど価値が高いと判断され，結果的に多くのページ・多くの人に知られている有名なページが上位にランク付けされる．しかし，検索エンジンを利用するユーザは有名なページを欲しているとは限らず，自分の欲している情報の載った有用なページを求めている．このため，すべてのリンクが推薦であると見なすリンク解析型の手法ではなく，他の人が推薦するページを重要視するという手法の方が直感的であると思われる．

そこで本論文では，Blog 情報を用いた Web ページのランキング手法と検索システムについて提案する．Blog では，書き手が自発的に新たに発見した Web ページに対する参照とともに文章を添えて記事にするため，Blog サイトを解析することにより，Web 上に新たに追加されたコンテンツや最近の話題などを追いやすくなる．また，Blog の記事の中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と思われる．さらに，すべての Blog の記事の下す評価を均等に扱うのではなく，あらかじめ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ということを熟知度という形で推定し，評価されるページはその分野についての信頼性が高いものとする．これにより，単に Blog 記事で評価されたスポーツに関する記事という形ではなく，野球について詳しい人が良い評価を下したスポーツの記事という形で提示することが可能になる．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を行い，システムの検証を行った (図 1).

以降，第 3 章にて Blog サイトからどのように Web コンテンツに対する評価を取得するかを述べる．ここで，Blog の書き手がどのような分野の知識について詳しいのかを表す熟知度を解析により計算するための手法を述べる．また，この熟知度を利用して Web 上のページの信頼性を評価値という形で一般的に提示するための手法を述べる．

第 4 章では，第 3 章で述べたページの評価値を利用して，どのように検索システムの枠組みを構築するのかを述べる．ここでは，ユーザの入力する検索キーワードからその語がどのようなカテゴリに属するのかという情報を付加させる検索質問の拡張の手法につて述べる．また，Blog 情報を用いて現在存在する検索エンジンの出力結果をより改善するための手法を述べる．

第 5 章で，これらの手法を基に作成したプロトタイプシステムの実装と動作概要を述べる．また，このプロトタイプシステムを通じて得られた問題点や改善点についてまとめる．このためにまず，第 2 章で基本的事項と関連研究について説明する．そして，第 6 章で結論を述べる．

2.　基本的事項及び関連研究

2.1　基本的事項

2.1.1　Blog について

Blog とは，Web 上で見つけた情報を個人的にメモしておくという形を発端とし，現在は他の Web ページを参照しその内容について書き手の独自の視点から記した文章を添えるという形式となっている．また，日々記事が追加されるため時間変異が大きく，Blog 情報を用いることで最近の話題等を追いやすい．さらに，そのような Blog のサイトを構築するためのスクリプトも充実してきており，例えば MovableType [7] では，タイトルや日付といった情報だけでなく，記事の分類・過去ログの自動管理・ひな形といった機能を持っていて，ユーザがより記事を書くことに集中できるよう作られている．なお最近では，このような Blog 環境の変化に伴い，独自の視点の記事と他のページへの参照というスタイルにとらわれず，書き手が自由な形式で記事を書き上げていく Blog サイトも増えてきている．本論文で想定する Blog サイトとは，他の Web ページへの参照に書き手の独自の視点による評論的な文章を添える形式の記事を載せるサイトである．

2.1.2　Blog クローラ

我々のグループでは，幅広く Blog に関する研究を行うために，Web 上から実際の Blog のサイトや記事の情報を集めてくるための作業を行っており，このための Blog クローラを製作している．現在は，特定の Blog ポータルサイト上から新しく更新された Blog サイトの情報を RSS(RDF Site Summary: We サイトの概要を記述したメタデータファイル) の形式で取得し，RSS の記述を基に実際の Blog サイトへアクセスすることで書き手や記事の情報を取得し，RDBMS に保存するという流れになっている．本研究で制作したプロトタイプでは，我々のグループが製作した Blog クローラにより集められたデータを用いて実験を行っている．. 収集した Blog のデータは今後，新たな情報検索システム作りのために利用することを想定している (図 2).

2.1.3　茶　　筅

茶筅 [10] は，奈良先端科学技術大学院大学の自然言語処理講座で開発されている日本語の形態素解析エンジンであり，他のシステムからの利用が容易となるよう用意されている．本論文では，Blog の書き手の熟知度を推定するために Blog の記事に含まれる文章から特徴語をキーワードとして抽出することを行うが，この作業のために茶筅を利用した形態素解析を行っている．

2.2　関 連 研 究

2.2.1　オークション取引における評判管理

山本らは，オークションなどのオンライン取引において取引後互いに評価をつけ合うシステムを想定し，ポジティブな評価をつけるかどうかに基づく評価システムと，ネガティブな評価

図 2　Blog クローラを利用した Blog 検索システム

をつけるかどうかに基づく評価システムのどちらがうまく機能するかをシミュレーションにより検証している [2]．Web 上のコンテンツに対する評価情報を考え，またそれらを検索エンジン上で利用することを想定している点で，本論文とは異なる．

2.2.2　ピアの評判情報の取得

Sepandar らは，Peer to Peer ネットワークにおいて，まだ見知らぬピアが良いピアであるかどうかをあらかじめ評価づけるために，既知のピアすべてに対象とする未知のピアの評価情報を尋ね，その評価に自ピアから見た既知のピアへの評価情報を重みづけて，擬似的な評価を求めるというアイディアを述べている [1]．すなわち，あるピア i から見た未知のピア k の評価はピア i の既知のピアを j として $\sum_j (c_{ij}c_{jk})$ の形で表される (c_{ij} はピア i からみたピア j の評価値)．Web 上のコンテンツが評価され，またその評価結果を用いて検索エンジンを改善することを想定しているという点で，本論文と異なる．

2.2.3　Annotea

Annotea は，W3C の Annotea Project [4] にて行われている Web・アノテーションシステムである．Annotea は，掲示板等のないどのような Web ページであっても複数の訪問者が注釈を寄せることができ，また寄せられた注釈を閲覧することもできるシステムである．そしてさらに、寄せられた注釈をメタデータとして利用することも想定されている．コンテンツに寄せられるコメントや評価情報をインターネット上に広く存在する Blog の記事から取得するという点で，本論文とは異なっている．

2.2.4　評判情報検索

立石らは，Web 上から商品に関する評判情報を取得するための手法について述べ，通常の検索エンジンとの精度比較による評価を行っている [3]．立石らは自然言語解析によるアプローチを取っており，「好き」「お勧め」といった良い評価を与える単語と商品名に対し，て文末表現や助詞の出現から評価として適正かどうかの判定ルールを作り，評判情報収集の精度を上げている．本論文では，Blog の記事から参照されたページは書き手

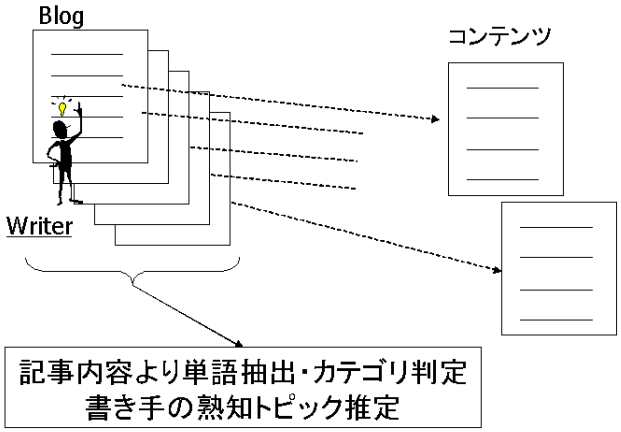

図 3　書き手の熟知度の取得

からどの程度良い評価を与えられているのかを判定する手法を提案するが，ここで立石らのアプローチを参考にできるものと考える．

3.　Blog サイトの持つ評価情報

3.1　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か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そもそも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 (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 Blog サイト同士の結びつきによりコミュニティ性が生まれる．そして，このコミュニティ内である特定のトピックに対し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 サイトを判定することができる．このような判定を統計的に行うことによりそれぞれの Blog サイトの特性を計り，これらを基に Blog サイトの信頼性を各特性の度合いを反映したプロファイルとして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。

3.2　書き手の熟知度の取得

本論文では，検索結果の改善を行うという目的のために，Blog サイトの持つ多岐にわたる特性の中から，特に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 (図 3).

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 [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，どのような書き手がどのような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．

3.3 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する．立石らの研究 [3] 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．

3.4 コンテンツの信頼性の提示

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値 (3.3 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カテゴリ毎に提示することができる (図 4)．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，その評価値から書き手の熟知度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，一つの評価値を求める．例えば，ある特定のカテゴリについての熟知度が書き手 a は 0.9, b は 0.2, c は 0.2 であり，この三人が特定のあるページについてそれぞれ 0.8, 0.3, 0.2 という評価を与えていた場合，正規化処理により求められるこのコンテンツの一意の評価値は 0.65... となる (図 5).

また，Blog の書き手 i について，特定のトピックについて書き手の熟知度 k_i，これら書き手が特定のページに p_i の評価を

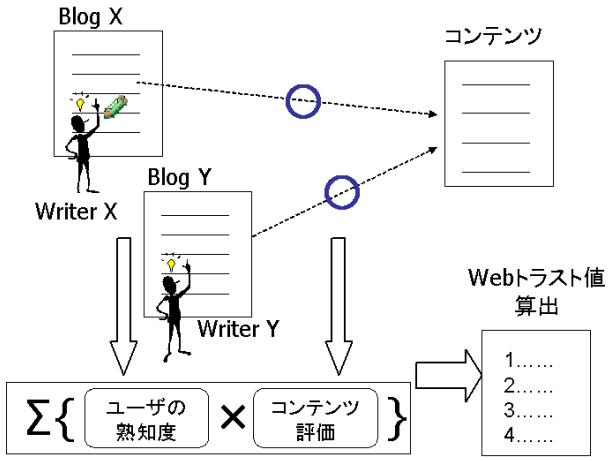

図 4　コンテンツの信頼性を表すトラスト値の算出

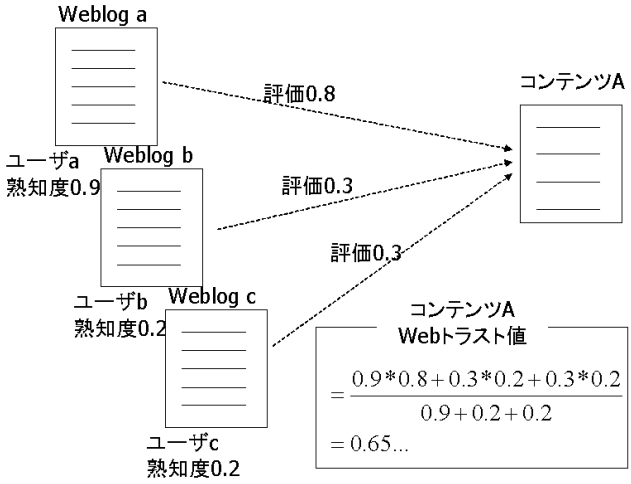

図 5　コンテンツの評価を定めるための正規化処理の例

つける場合の，このページのこのトピックについてのトラスト値 $T(p)$ とした場合、定式化すると以下のようになる

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

この操作を書き手の熟知度のカテゴリ毎に行うことにより，コンテンツの一意の評価値をカテゴリ毎に求めることができる．本論文では，このような評価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しとする，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．

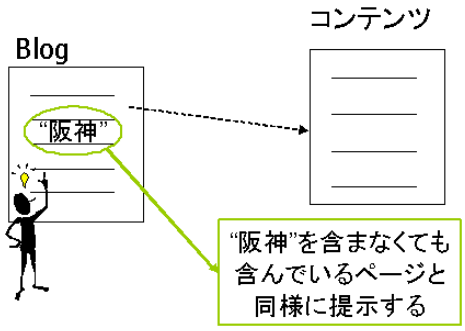

図 6　Blog 記事文章も参照先ページ検索に利用

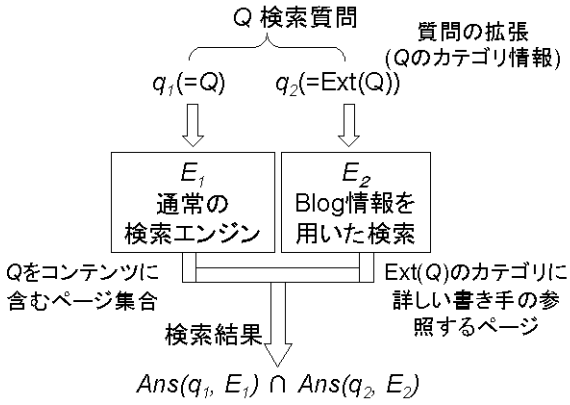

図 7　検索質問の拡張

4.　Blog 情報に基づく検索結果の改善

4.1　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．

他の Web ページを参照し良い評価を下している Blog 記事は，このままではその Web ページにとってのメタデータと見ることはできない．しかし，記事の内容や Blog サイトの信頼性が吟味されることにより，Blog の記事は参照先の 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してい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ことで，ユーザの入力する質問 Q のキーワードがコンテンツ内に直接含まれるかどうかだけで判定を行う従来型の検索エンジンをより拡張することができる．その具体的な利用方法として，4.2 節で参照先キーワードの補完を，4.3 節で検索質問の拡張を説明する．

4.2　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，提案手法では，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 (図 6)．これは，Blog 情報を用いて Web ページを検索するための仕組みとして用いることができる．

4.3　検索質問の拡張

二つ目として，通常の検索エンジンの出力結果をより吟味したものに改善するための手法が考えられる．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 (図 7).

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，3.2 節での手法と同様に，OpenDirectory [8] 等のカテゴリ検索サービスを利用して取得する．今，3.4 節の手法により各 Web ページにはカテゴリ毎の評価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，その評価値の高い Web ページを優先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 8).

そして，通常の検索エンジンの結果と Blog 記事に対する検索の結果の両方をユーザ側に提示するため，両者を統合して出力する必要がある．ここでは，Blog 記事により与えられる Web ページの評価値を用い，通常の検索エンジンの出力結果のランキングを行い提示することを考える．具体的には，カテゴリ毎の Web ページの評価値に基づき，

- 評価値の高い順に Web ページをランキングする．
- 検索キーワードの単語の属するカテゴリについて，そのカテゴリでの評価値の得られていないページをフィルタリングして表示させない．
- 評価を与えているような Blog 記事そのものを参照先ページとともにユーザ側に提示する．

といった手法が考えられる．これにより，単にコンテンツ内にキーワードを含むかどうかだけの判断ではなく，カテゴリ的に詳しい人からの支持があるかどうかという判断を用いて，提示コンテンツの信頼性を高めることができる．

5.　プロトタイプシステム

3 節と 4 節で述べた提案手法を基にプロトタイプシステムを作成し評価実験を行った．このプロトタイプの処理の流れと，評価実験を通じた考察を行う．

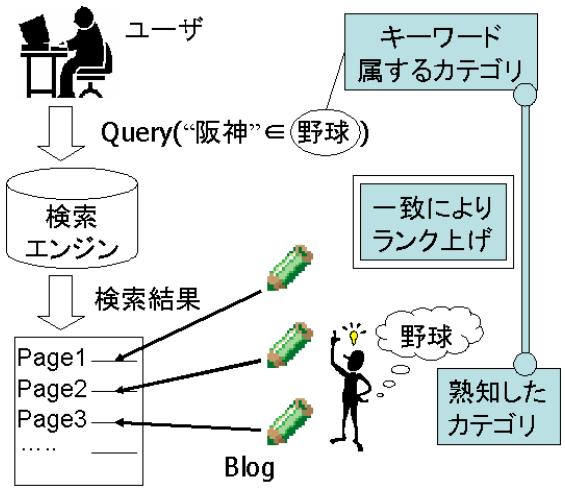

図 8　キーワードのカテゴリ一致による検索結果の改善

5.1　実 装 環 境

プロトタイプで用いた Blog の情報は，2.1.2 節で述べた Blog クローラを用いて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(同一ページ含む時別々の記事なら重複許す) の有効なデータを基に，システムを作成した．システムでは，書き手ごとの熟知度算出の部分に Perl を用い，Blog 情報を基にした検索システム部には Visual Sutido .NET の C#を用いた．また，Blog の記事等のデータは MySQL を用いて RDBMS により管理している．Blog の記事からの単語抽出には茶筅 [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 [9] を利用した．また，実験に用いたマシンは OS: Windows2000/CPU: Pentium4 1.5GHz/メモリ: RDRAM 1GB である．

5.2　インターフェース

ユーザの側から見たシステムの利用法について述べる．

Blog 情報を基にした検索を行うシステムの画面は，図 9 のようなものである．ユーザはまず，検索したい事項についてキーワードにてシステムに入力する．システムは，キーワードを基にした通常型の検索エンジン・キーワードのカテゴリ推定・Blog 情報からの評価値に基づき，該当するページ・そのページを参照している Blog 記事・どのようなカテゴリについて評価を得られているのかをまとめたものを一覧として表示する．

5.3　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） Blog 記事の書き手ごとに，ある特定の人により書かれたすべての記事を取得する．

（2） 記事中のすべての本文とタイトルについて茶筅 [10] による形態素解析を行い，名詞 (固有名詞含む) と未知語 (主にカタカナ語・アルファベット語) と判定されたものを集計する．

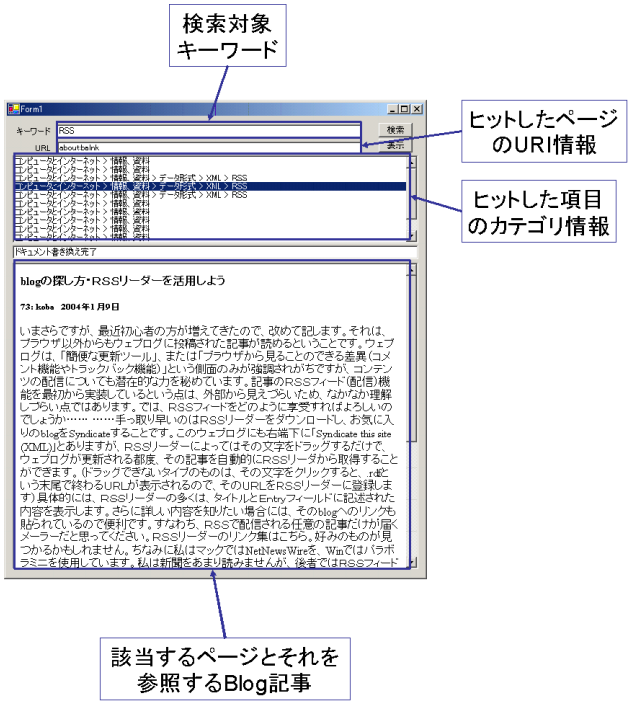

図 9　プロトタイプシステムのインターフェース

（3） 集計された単語について出現頻度の高いものから順に 40 個を取得し，これをその書き手の特徴キーワードとする．

（4） 個々の特徴キーワードごとに，Yahoo!Japan [9] 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階層的に取得する．複数のカテゴリが該当する場合は上位 10 個までを取得し，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．

（5） 書き手ごとに最大 400 個あるカテゴリ情報を保存しておく．この操作をすべての Blog 記事の書き手ごとに行う．これらはプロトタイプシステム上での動作に用いる．

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キーワードの形として入力する．

（2） 入力されたキーワードを基に，Yahoo!Japan [9] 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階層的に取得する．複数のカテゴリが該当する場合は上位 10 個までを取得し，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） ユーザの入力したキーワードを基に Google による検索を行い，その結果上位 500 件までを取得する．

（4） 上記の結果一ページごとに，そのページを参照しているような Blog 記事を探し出し，同時にその Blog 記事を書いている書き手も取得する．

（5） 該当する Blog 記事の書き手の熟知度より，詳しいとされるカテゴリ情報すべてを取得する．

（6） 詳しいとされるカテゴリすべてについて，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

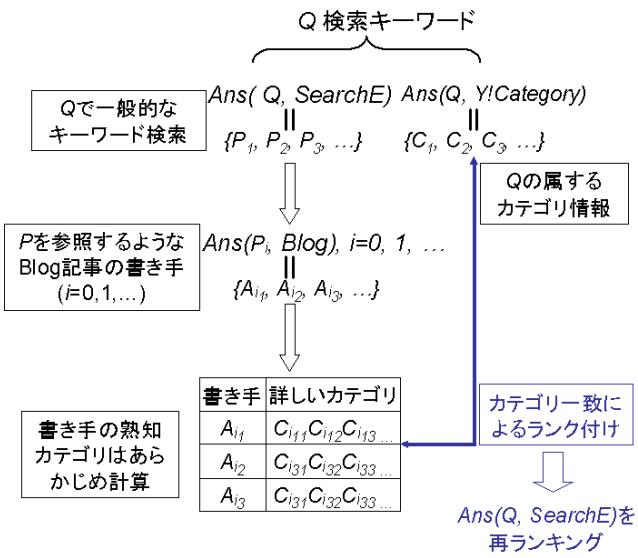

図 10　プロトタイプシステムでの処理の流れ図

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．

（7） 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，をセットとして保持する．

（8） 一致した事項についてそれらを一覧としてユーザ側に提示する．

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 10 に示す．

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，Google 等の一般的な検索エンジンを使うほかに Blog 情報を用いた緩和も考えられる．これは 4.2 節で述べた考えを用い，処理を以下のように置き換えたものである．

- ユーザの入力したキーワードを記事の本文中に含むような Blog 記事を探し出し，それら該当する Blog 記事すべてについて記事中から参照されているページを取得する．

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．このため，各ページごとの評価値計算は省いており最終的な評価値によるランキングに基づいた並び替えは行わず，該当するものを順に提示するものとしている．

5.4 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．Google の返す結果上位 100 件により先に実験を行っていたところを，この問題を解決するために上位 500 件までに拡張して行うようにしたが，それでも結果が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へ参照をしているような Blog 記事があまりなく，Blog 記事中は常時追加削除されうるニュースサイト上の特定の記事に参照していることが多く，Google のような検索エンジンではそのようなページが結果として返されてくることが少ないことが原因ではないかと思われる．
- 上記のような Google を介したアプローチの問題点を補うものとして，Blog 情報を用いた緩和アプローチを用いることができた．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．
- 単に該当するような Blog 記事からの参照先ページのみを提示するのではなく，ページと該当する Blog 記事・カテゴリの情報も同時に提示することにより，ユーザに参照先ページに対する一つのとらえ方を文章の形で与えることができていると思われる．Blog 記事の内容が書き手の独自の視点からの感想という形になっているため，ユーザに参照先ページの内容理解をより深めることができた．
- 逆に，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，参照先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，参照先が一般的なニュースサイトの特定の記事である場合にのみ見られた．
- 現在はプロトタイプシステムのため，ページと Blog 記事とタイトルと書き手の名前をそのまま表示する形を取っているが，これらはユーザにより効果的に見せる必要があるため，そのための手法を検討する必要がある．

これらを踏まえ，今後改善していくべき点として以下のものが挙げられる．

- 今回はおよそ 1200 記事ほどを対象とした解析を行ったが，Blog クローラを通じてより多くの Blog 記事を取得し，それらを対象とした解析を行う必要がある．これにより，Gogle 的アプローチでも該当する結果を多く提示できるものと思われる．
- 熟知度の重みや Blog 記事からのページへの評価の度合いを考慮して，結果として提示するものをランキング表示していくという本論文中での提案手法を，実際に実装し検証する必要がある．
- 現在 Google を介したアプローチで数十秒，Blog 情報を用いた緩和アプローチで 10 秒ほどの処理時間を要するため，最終的に Web アプリケーションとしてユーザに利用してもらうためには，これらの処理速度を改善する必要がある．
- ページと Blog 記事情報を効果的にユーザに提示するために，表示方法を工夫する必要がある．例えば，似たような意

見を持つ Blog 記事同士は一つのまとまりとしてクラスタリングした結果をユーザに提示することなどが考えられる．

- 現在は，システムを実際に利用するユーザの視点での主観的な意見・考察にとどまっているが，より客観的に提案手法の有効性を示すために何らかの数値的尺度を取り入れる必要がある．検索システムの制度を評価するための指標である適合率・再現率を基にしたものが考えられる．

6. 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，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．これは，今後も増え続けるであろう Blog サイトをシステムが有効に活用するための利用法の大きな一つであると思われる．

提案手法ではキーワードの属するカテゴリ情報を用いたが，複数のキーワードを基にする場合や複数の異なるカテゴリが候補となった場合に，どのようにすべきかという問題も残されている．それらも含めて，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必要がある．

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] 山本 仁志:消費者間オンライン取引における評判管理システムの分析, 経営情報学会誌,vol.12,No.3,2003,pp.55-69
[3] 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
[4] Annotea Project, http://www.w3.org/2001/Annotea/
Marja-Riitta Koivunen, Annotea: An Open RDF Infrastructure for Shared Web Annotations, WWW2001
http://www10.org/cdrom/papers/488/index.html
[5] Sepandar D. Kamvar: Extrapolation Methods for Accelerating PageRank Computations, WWW2003
http://www2003.org/cdrom/papers/refereed/p270/
kamvar-270-xhtml/
[6] WikiWikiWeb
http://c2.com/cgi/wiki
[7] Movable Type
http://www.movabletype.org/
[8] OpenDirectory
http://dmoz.org/
[9] Yahoo!Japan
http://www.yahoo.co.jp/
[10] 形態素解析システム 茶筌
http://chasen.aist-nara.ac.jp/