目的指向型ブログ検索システムBLOGRANGERの提案およびユーザ評価

戸田浩之[†] 藤村考[†] 井上孝史[†]
廣嶋伸章[††] 杉崎正之[††]
片岡良治[†] 奥雅博[†]

ブログに記載される情報は，一般の Web と比較して，エンタテイメント情報や最新の話題，商品やサービスに対する個人的な意見を多く含む．我々はこれらブログ固有の情報コンテンツとしての特徴を生かし，話題検索，評判検索等の検索目的に特化した複数のインタフェースを備えたブログ検索システムを開発した．本システムの有効性を評価するため，大規模な Web アンケート調査を行い，本システムと Web 検索および従来のブログ検索との検索結果の主観評価による比較を行った．収集した 2191 の回答を分析した結果，話題の検索および評判の検索においては，本システムの方が有効であると回答した人の割合が Web 検索の約 2 倍となる等，本システムによるブログ検索がこの分野での検索ニーズを充足する有用なツールとなりうることを示す結果となった．

BLOGRANGER: Implementation of Goal-oriented Blog Search Engine

Hiroyuki Toda,[†] Ko Fujimura,[†] Takafumi Inoue,[†]
Nobuaki Hiroshima,[†] Masayuki Sugizaki,[††] Ryoji Kataoka[†]
and Masahiro Oku[†]

Topics mentioned in blogspace are biased towards interesting/funny or entertainment-related topics compared to the generic web space and many articles contain personal opinions on goods or services. Making good use of these characteristics, we introduce a new blog search engine that provides multiple interfaces, each targeted at a different goal, e.g., topic search, blogger search, and reputation search. To evaluate the effectiveness of the system, we conducted a user survey and collected 2191 answers. For the specific search conducted, twice as many people answered that BLOGRANGER is superior to general web search.

1. はじめに

World Wide Web（以下単に Web）上での日記作成ツールとして近年急速に普及したブログは，記事の更新情報を ping サーバに通知するという push 型の要素と，記事のサマリ情報（RSS）を配信するという pull 型の要素をあわせ持つ通信手段としたことで，情報発信者と情報受信者の新しい出会いが効率的に行える等のメリットを生み，単なる日記ツールとしての枠を越え，新しい情報発信，情報共有ツールとして多くのユーザに利用されている．

2006 年 3 月末時点での国内のブログ開設数は 868 万件と膨大な数のブログが開設されており[19]，これを背景に，ブログのみにターゲットを絞ったブログ検索システムの開発が近年活発に行われている[20),21),23),24),27),30)]．

しかし，ブログ空間（ブログサイトおよびそれらに関係するリンク）は，Web 空間のサブセットであるため，検索目的によっては，従来の Web 検索をそのまま利用することも可能である．芸能人等の有名人ブロガーのブログを検索する場合はその典型例である．

一方で，ブログ空間には最新の話題に関する記事や個人の主観的な意見が多い等，従来の Web とは異なる特徴がある．これらを対象にした「今注目の話題が知りたい」や「ある製品の評判が知りたい」等の検索は，Web 検索では必ずしも有益な結果は得られないが，ブログの特徴を分析し，適切な手法を利用することで，有益な結果が得られるのではないかと考えられる．

しかし，現在のブログ検索システムの多く[20),24),27),30)]は，日付順やアクセス頻度順等で単純にランキングさ

† 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
†† NTT レゾナント株式会社
NTT Resonant Inc.

れた検索結果を表示するものであり，上記のようなユーザのニーズを必ずしも満たせていない．

我々はこの現状から，ブログの情報コンテンツとしての特徴について分析した結果に基づき，新たなブログ検索システム BLOGRANGER を提案する．このシステムでは，ブログの特徴を活かすことで，以下に示す検索を実現する．

- 話題の検索
  ブログの中に含まれる最新の話題や多くの人が興味を持つような注目度の高い話題を探す．
- 評判の検索
  ブログの中に含まれる製品やサービス等について述べた主観的な意見，感想を探す．
- ブロガーの検索
  同じ分野に興味がある人（ブログサイト）や特定の分野で注目されている人を探す．

ブログの特徴を有益に利用する手法として，BLOGRANGER では，検索結果のリストを提示するだけでなく，検索結果のブログ記事集合中に含まれる情報（話題や評判，ブロガー等）を動的に解析し，主要な情報を検索結果のリストとともに提示する．

これにより，ユーザは検索結果中に含まれる主要な話題や評判，ブロガーを効率的に発見し，関連する記事に容易にアクセスできる．

また我々は，大規模なユーザアンケートを実施し，上記で提案した検索のニーズの検証および BLOGRANGER の有効性評価を行った．

以下，本論文では，2 章に BLOGRANER 開発の背景となったブログの特徴分析と，それに基づくシステムの設計指針について示す．3 章では，2 章で述べた設計指針に基づいた BLOGRANGER のアプローチとその実現方法を述べる．4 章では実証実験で行ったアンケート調査の結果と考察について述べる．5 章で関連する技術，研究との比較を述べ，6 章でまとめる．

2. ブログ検索の目的

ブログに含まれる情報は，個人の備忘録レベルの情報も多く，内容は玉石混交である．近年の検索エンジンの性能向上により，Web 空間全体から品質の高い Web ページを見つけることが可能になっている中で，玉石混交のブログ記事のみをターゲットとする検索エンジンを開発しても役に立たないという考え方もある．しかし，我々は，ブログ記事の情報コンテンツとしての特徴を分析し，その特徴を活かした情報検索が可能になれば有用なシステムとなると考えた．以下では，我々が注目したそれぞれの特徴について示す．

ブログの情報コンテンツとしての第 1 の特徴は，「重要な」ニュースより，「面白い」あるいは「興味深い」話題やエンタテイメント系の話題が多く含まれる傾向が強いことである．これは，ブログはその書き手であるブロガー単位に作成されるメディアであるため，個人の主観に基づいて言及対象の話題が選択されるからであると考えられる．したがって，ブログから面白い話題や話題に関連する情報を効率的に収集することができれば，週刊誌等のようなエンタテイメント性の高い読み物としての利用が期待できる．

第 2 の特徴は，個人の主観による商品・サービスの評価といった「消費者の生の声」が多く含まれていることがある．これらは，商品・サービスを購入しようとする消費者が，すでに購入した人の評判を調べるため等に有用である．また，商品・サービスの提供者が市場の反応を調べマーケティングに生かすといった目的にも利用できる．

第 3 の特徴は，ブロガー単位（ブログサイト単位）に記事がまとめられるため，ブログの情報コンテンツにブロガーのパーソナリティ（嗜好や文体等）が強く反映されることである．また，ブログには，コメントやトラックバックといったコミュニティ形成を支援する機能も備えている．これらの特徴により，ブログは，同様の嗜好もしくは感性を持つ人と交流するためのソーシャルメディアととらえることもできる．この目的でブログを利用する場合には，トラックバック先の仲間（ブログサイト）やその仲間が書いた記事（ブログ記事）を探すことが重要であり，この支援によりコミュニティを活性化できる．また，特定の分野に詳しいブロガーを発見できれば，有益な情報を継続的に収集できるという利点も考えられる．

我々は，以上の分析に基づき，ブログ検索の価値を高めるため，以下を実現するブログ検索システムを提案する．

- 話題検索
  話題（面白い話題，興味深い話題）を探す．
- 評判検索
  物や事象に対する個人の意見や感想等の評判を探す．
- ブロガー検索
  同様の嗜好を持つ仲間や分野の注目ブロガーを探す．

次章では，ブログの特徴を活かして，上記で示す検索を効率的に実現する BLOGRANGER の詳細について述べる．また 4 章では，ユーザアンケートをもとに，上記で提案した検索機能のニーズ検証，およびこれら

図 1 BLOGRANGER の画面例
Fig. 1 GUI of BLOGRANGER.

の検索を行う場合に，BLOGRANGER が有効であることを示す．

3. BLOGRANGER システム

3.1 基本的なアプローチ

ブログの特徴を利用し，ユーザにできるだけ負荷を感じさせることなく，2 章で述べた検索を実現するため，我々は Scatter/Gather [3),8)] で提案された検索結果クラスタリングのアプローチを採用する．

このアプローチでは，ユーザが大きな文書集合をブラウジングする場合に，文書集合中の話題を示す構造を提示することで，文書集合の概観を可能とし，所望の文書への到達を支援する☆．

ただし，Scatter/Gather では，文書ベクトル間の類似度をもとに作成したクラスタを個々の話題の集まりとして表現し，検索結果を構造化していたのに対して，BLOGRANGER では，クラスタリングを検索結果からの主要なキーワードの抽出と見なした手法[15),17)] と同様のアプローチで，検索結果中の主要な特徴（キーワード等）をもとに検索結果の話題を表現し，構造化を行う手法を利用する．

☆ Scatter/Gather では，さらに，Gather プロセスとして，選択した複数のクラスタをマージし，再クラスタリングするプロセスも備えている．

さらに，従来手法[15),17)] が一般的な Web 検索やニュース記事検索を対象とし，検索結果の構造化を行うために名詞や名詞句を中心としたキーワードのみに注目しているのに対し，提案手法では，ブログ検索の目的を考慮し，上記のキーワードに加えて，ブログ記事中に出現する形容詞や形容動詞等の評価表現，リンク，記事の作者であるブロガーに注目し，検索結果の構造化を行う．

また，BLOGRANGER では，検索結果中の構造を，ユーザに提示するための手段をフィルタと呼び，これにより，キーワードや評価表現，リンク，ブロガー等それぞれの観点で表現した検索結果の構造を提示する．

図 1 に，このアプローチに基づくシステムの画面例を示す．ユーザが BLOGRANGER を利用する場合には，通常の検索システムと同様に興味のあるキーワードを入力し，検索ボタンを押下する．検索結果には，通常の検索結果に加えて，画面左側に検索結果の構造を示すフィルタが提示される．

フィルタは，前章で示したブログ記事に対する 3 つの検索目的を支援するために 4 つ用意されている．

- 話題検索
  - 「トピックフィルタ」
    検索結果中に含まれる話題に関連する固有名詞を抽出し，検索結果中の話題を分かりやす

く提示する．キーワードを選択することで，そのキーワードを利用した絞り込み検索が行える．
  - 「リンク先フィルタ」
    検索結果中で多くのユーザが注目するニュースやサイトを提示する．提示されているサイトを選択することで，そのサイトへリンクを張っているブログ記事のみを絞り込むことができる．
- 評判検索
  - 「評価表現フィルタ」
    検索結果中に含まれる評価表現を分析し，全体傾向の提示および詳細な表現の参照を容易にする．提示されている評価表現を選択することで，その評価表現を含むブログ記事を絞り込み，その表現がどのように利用されているか容易に閲覧できる．
- ブロガー検索
  - 「ブロガーフィルタ」
    検索結果の話題に関して注目度の高いブロガーを提示する．ブロガーを選択することで，そのブロガーが書いた記事のみを絞り込むことができる．

ユーザは検索結果上部のタブを操作することで，これらフィルタを切り替えられ，目的にあったフィルタを選択することができる．そしてフィルタを参照することで，検索結果中にどのような情報が存在するかを概観でき，所望の情報があった場合には，クリック操作 1 つで情報の絞り込みが可能になる．また，各フィルタの切替えや，複数のフィルタを利用した検索結果の分析も容易にでき，様々な観点から検索結果を分析することを可能としている．

これらフィルタに関しての詳細は次節以降に示す．

また，ブログ記事は一般に玉石混淆であるといわれ，他では手に入らない有益な情報を含む記事が存在する半面，個人の備忘録のように，そのブロガー以外が参照してもまったく意味がない記事も存在し，ある程度有益な記事を優先的に提示することが必要となる．そこで，本システムでは，ブログ記事間のリンク関係を分析することで，ブロガー，ブログ記事の注目度を分析し，それをもとにブログ記事検索結果をランキングすることとしている．これは，注目度が高い記事は，他のブロガーから注目を浴びる記事であり，そのような記事はブログを読む側にとっても有益であるとの考えに基づく．ランキングの詳細については 3.3 節に示す．

3.2 検索目的とフィルタ

前節で述べたように，BLOGRANGER では，それぞれの検索を支援するフィルタを検索結果を分析することで生成する．本節では，それぞれの検索目的に応じたフィルタのバリエーションについて示す．

まず，話題検索では，検索結果中に存在する話題を特定することが重要となる．これを実現するために考えられる方法として，ブログ記事の本文（テキスト部分）を解析する方法と，ブログ記事に多く含まれるリンクを解析する方法の 2 つが考えられる．これら 2 つは，相互に関係している場合もあるが，リンクを付与しないブログ記事や，逆にほぼリンクだけを備忘録のように記述するブログ記事も存在し，独立に存在している場合も多い．また，本文の解析で得られるのが，ブロガー自身が提供している話題なのに対し，リンクで得られるのは話題の情報源であるという違いもある．そこで，本システムでは，話題検索に関しては，ブログ記事の本文を利用する「トピックフィルタ」と，ブログ記事中に存在するリンクを利用する「リンク先フィルタ」の 2 種類のフィルタを採用した．

評判検索では，検索結果中でどのような意見や感想が存在しているか，またどのように言及されているかを見つけることが重要となる．このため，検索結果中での主要な評価表現によって検索結果を構造化する「評価表現フィルタ」を採用する．

ブロガー検索では，ユーザが興味を持つ分野で注目されるブロガーや特定のブロガーが書いた記事を見つけることが必要となるため，検索結果をブロガーごとで構造化する「ブロガーフィルタ」を採用する．

以下では，それぞれのフィルタ生成に関する課題と実現方法について示す．

3.2.1 トピックフィルタ

トピックフィルタの生成は，我々が提案している検索結果分類技術[17] をベースとする．ただし，この手法はニュース記事の検索を対象としているため，2 章で述べたブログに対する検索目的を充足すべく，より広い話題に対応する改良を行っている．

我々が提案している検索結果分類技術[17] によると，検索結果の文書集合から，ニュース記事中の話題やイベントを特定するのに有益な人物，組織，場所といった固有表現[7),☆] を自動的に抽出するとともに，抽出した固有表現の中から，記事を分類するのに適切なものを選び出すことが可能となる．しかし，最初のプロセ

☆ 実際の固有表現抽出では，これらの固有名詞に加えて，金銭表現や時間表現等の数値表現の抽出も行うが，文献 17) では，これらの数値表現は利用していない．

スで抽出する固有表現の種別は,「人物名」,「組織名」,「場所名」,「その他の固有物名」に限定されており，ブログ記事中の話題を考えると，この種別は必ずしも十分ではなく，抽出できる語彙の種別を増やす必要があると考えられる．

抽出する語彙の種別を増やす場合，上記の手法で利用されていた，固有表現抽出ツールでは，機械学習を利用する手法により，語彙の抽出を行っているため，新たな教師データを作成する必要がある．しかし，語彙の種別を増やすたびに新たな教師データを作るのは非常にコストがかかるうえ，種別が増えるにつれて，種別間の差が不明確となり，抽出精度が低下することも考えられる．

そこで，今回の提案システムでは，既存の固有表現抽出では抽出できない語彙の抽出を行うため，抽出する語彙の種別ごとに辞書を構築し，それを利用して語彙の抽出を行う方法を利用した．これによると，網羅的な辞書を用意することができれば，比較的簡単に語彙の抽出が可能となる．

しかし，辞書の構築では大きく 3 つの点が問題となる．

- 語彙の新規性
- 語彙の網羅性
- 異表記の語彙への対応

これらの問題に対応するため，今回の提案システムでは，Web 上に存在する語彙を利用することを考えた．手法の概要を**図 2** に示す．提案手法では，まず抽出したい種別の語彙が日々更新されながら存在するサイトを特定し，そのサイトから語彙を抽出する “Web ラッパ” を作成する．今回のシステムでは，映画，本，テレビ番組，CD，DVD，アニメ，ゲームのコンテンツ名の辞書構築を行っており，Web 上のポータルサイト等に存在する新着情報，ランキング等を Web ラッパの処理対象とした．例としては，goo 映画☆等があげられる．本システムでは，7 種の辞書を構築するために，40 の Web サイトを解析している．しかし，サービス提供者によっては，コンテンツが違う場合でも同じフォーマットで Web サイトを提供している場合があるため，Web ラッパの種類は 11 種類を利用している．この手法は，個々のサイトごとに人手でルールを決定するものであり，多少のコストがかかるが，これにより，つねに新しい語彙を取得することを可能とする．

次にここで取得した語彙をもとに，以下に示す「並列語獲得法」と「異表記獲得法」を利用することで，

☆ http://movie.goo.ne.jp/schedule/upcoming.html

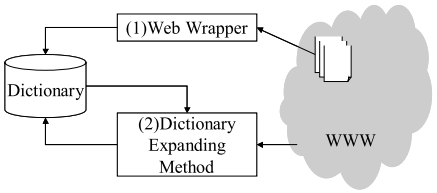

図 2 辞書構築プロセスの概要図
Fig. 2 Outline of dictionary construction process.

Web ラッパで取得した語彙を拡張し,「語彙の網羅性」と「異表記の語彙への対応」の問題を解決する.「並列語獲得法」とは，Web 上で，リストやテーブル等のように並列に並べられている構造を利用し，Web ラッパで獲得した語と同じ種別の語彙を獲得しようとする手法であり,「異表記獲得法」は，ある語彙に関する代表的な Web ページを見つけ（映画のオフィシャルサイト等），その Web ページに対するリンクのアンカテキストを解析することで，異表記の語彙を獲得しようとする手法である．それぞれの手法を以下に示す．

まず，並列語獲得法について示す．この手法では，野口ら[18] や Shinzato ら[13] が提案している手法と同様に HTML 中での繰返し構造に注目し，同列に並ぶ語彙を抽出する手法である．以下に処理のステップを示す．

(1) 以下のプロセスを複数回繰り返し，抽出された候補語と，各候補語の抽出された頻度のデータを作成．
  (a) ラッパを利用して抽出した語彙から，抽出を行う種別の語彙を少量（5〜10 程度）ランダムにサンプリングし，事例データとする．
  (b) 事例データをもとに Web 検索エンジンに問合せを行い，規定数以上の事例データを含む Web ページを特定．
  (c) 上記で特定された Web ページを取得．
  (d) HTML 文書を XML 文書と見なし，規定数以上の事例データが出現する有益なパスを特定．
  実際には，上記で特定したパスをリーフからルート方向にさかのぼったときに，最初に出現する TR もしくは，LI の出現位置を無効化したパスをもとに，有益なパスの特定を行っている．これは，テーブルの列方向やリスト構造に有益な情報が存在することが多いと考慮したためである．
  (e) 上記パスを用いて，該当するパスに存在

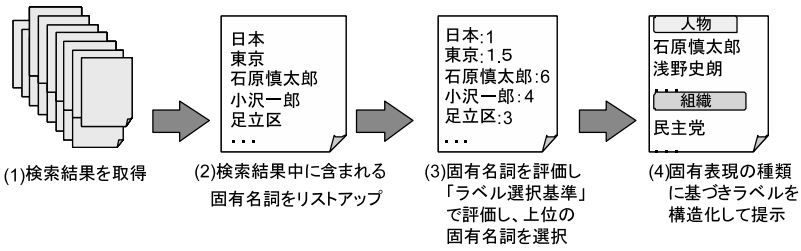

図 3 トピックフィルタの生成フロー
Fig. 3 Constructing process of Topic Filter.

する文字列を候補語として抽出.

(2) 候補語のうち，一定以下の抽出頻度の語は，ノイズの可能性があるため，候補語から除去.

(3) 個々の候補語について，別の候補語との組合せについて，Web 中での共起関係を検証し，別の候補語との共起関係が弱い語は，並列語でない可能性があるため，候補語から除去.

(4) 上記の結果，候補語として残ったデータを並列語として辞書に登録.

次に，異表記獲得法について示す．この手法では，Fujii ら[4] が提案しているように，同一のページに対するリンクのアンカテキストには，同じ内容が書かれているとの仮説をもとに，共参照関係にあるアンカテキストから同じものを指し示す語彙を獲得する．以下に処理のステップを示す.

(1) 異表記を見つけたい語を用いて，Web 検索エンジンに問い合わせる.

(2) Web 検索エンジンから得られる検索結果から，入力した語彙に関する公式ホームページもしくはそれに準ずるページ（入力語を含むアンカテキストのリンクによって，一定回数以上リンクされているページ）を特定.

(3) 上記で特定されたサイトへのリンクを収集.

(4) 上記リンクのアンカ文字列を収集し，規定数以上出現する文字列を異表記の候補として抽出.

(5) 上記の候補のうち,「〜のホームページ」等，一般的に多くのアンカ文字列に含まれる文字列，部分文字列を削除.

(6) 上記の処理で残った文字列を入力した語に対する異表記として辞書に登録.

以上の手法を利用し，2 章で示したようにブログ中にはエンタテイメント系の話題が多いことから，BLOGRANGER では，テレビ番組，映画，DVD，CD，ゲーム，本，アニメのタイトル等の辞書を構築し利用している.

トピックフィルタを実現するための処理は以下のとおりである.

図 4 トピックフィルタの例
Fig. 4 Example of Topic Filter.

- 前処理
  ブログ記事が検索システムに登録される時点で，上記で述べた手法で生成する辞書と固有表現抽出により，個々の文書にどのような固有名詞が存在するかを分析し，検索用データベースに格納する.
- 検索時の処理
  検索結果中のブログ記事に含まれる固有名詞をリストアップし，その中から，戸田らの手法[17] による，検索結果中での重要性と，検索条件との関連性に基づいた指標により，検索結果中の話題を示すのに有益な固有名詞を抽出し，トピックフィルタとして提示する．検索時の処理概要を**図 3** に示す.

図 4 には，キーワード「喰いタン」で検索した場合

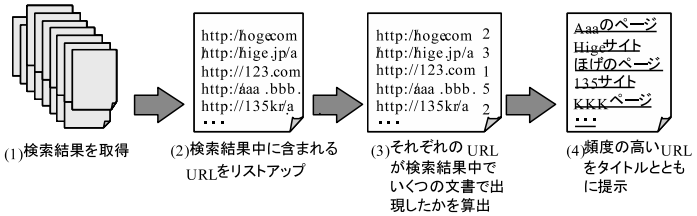

図 5 リンク先フィルタの生成フロー
Fig. 5 Constructing process of Refer Filter.

に表示されるトピックフィルタを示す．検索キーワードに該当するドラマの登場人物やドラマで舞台となった場所，同時期に放映されたドラマのタイトル等が並んでいる．たとえば，地名「大阪」を選択すると，大阪が舞台となったドラマの第 6 回の放送に関して書かれた記事が見つかり，テレビ番組で絞り込むと，同時期に放映されているドラマの視聴率に関して書かれた記事を見つけることができる．このようにトピックフィルタを利用することで，ユーザは検索結果中の話題を概観でき，また，気になる話題を示すキーワードがあればそれを選択すだけで容易に情報を絞り込み，新たな話題を知ることができる．

3.2.2 リンク先フィルタ

リンク先フィルタは，検索結果のブログ記事中で，多くのブロガーに注目されている話題の情報源を抽出提示する．しかし，ブログには，ブログ記事の内容に無関係の広告等自動的に生成されるリンクも多く含まれており，話題の情報源のリンクのみを集計するにはブログ記事に該当する領域を精度良く抽出することが必要となる．

そこで BLOGRANGER では，ブログ記事部分のみを抽出することを可能とするクローラを構築し利用している．このブログ記事部分のみを抽出するタスクに関連する研究として，様々な研究[2),10)] が行われているが，その多様性により，現実的に高い精度で抽出することはできないため，ブログプロバイダごと☆に手動でラッパを作成し，本文部分のみを抽出している．これにより，自動で生成される不要なリンクを排除し，話題の情報源となるサイトへのリンクをユーザに提示することを可能としている．

リンク先フィルタを用いることにより，たとえばキーワード「喰いタン」で検索すると，このキーワードを言及しているブログ記事中で参照している Web 上のリソースとして，「このキーワードに該当するドラマの公式サイト」や「ドラマに関係する音楽を配信するサイト」，「関係するニュース記事」等が提示され，該検索キーワードについて言及するブロガーに注目されている情報を知ることができる．さらに，これらのサイトを参照している人がどのようなブログ記事を書いているかを容易に閲覧することも可能である．

このフィルタを実現するための処理は以下のとおりである．

- 前処理
  ブログ記事が検索システムに登録される時点で，どのブログ記事中にどのリンクが存在するかを分析し，検索用データベースに格納する．また，リンク先を提示する場合に URL だけでは，何を示しているのか不明であるため，新たな URL が記事中に存在することが分かった段階で，その URL の文書を取得し，HTML の TITLE タグを利用して，URL のタイトルを取得する．
- 検索時の処理
  実際の検索時には，検索結果中に含まれる URL の中で頻度の高いものを優先的に抽出し，タイトルとともにリンク先フィルタとして提示する．検索時の処理概要を**図 5** に示す．

3.2.3 評価表現フィルタ

評価表現フィルタを生成するためには，我々はコーパスを解析して生成した評価表現辞書を利用する．この評価表現辞書を用いることで，検索結果のブログ記事から「面白い」，「素晴らしい」といった評価表現を含むセンテンスを抽出できる．この評価表現辞書は，約 7000 の形容詞，形容動詞により構成される．

BLOGRANGER では，評価表現フィルタとして，上記のようにして得たブログ記事中の評価表現を出現頻度順にランキング表示している．これにより，ブログコミュニティにおける概評を把握することができる．また，所望の評価表現を選択するだけで，たとえば，商品やサービスがどのような言い回しで評価されているかといった観点からブログ記事を容易に閲覧可能にした．

☆ ブログプロバイダによっては，複数のフォーマットが存在し，それぞれに対するラッパが必要となる場合もある．

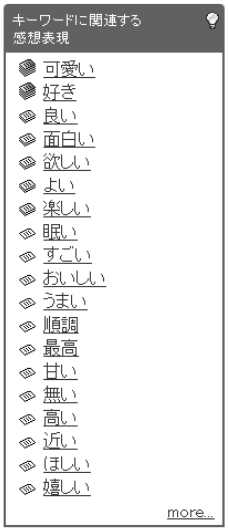

図 6 評価表現フィルタの例
Fig. 6 Example of Sentiment Filter.

図 6 には，キーワード「喰いタン」で検索された場合の評価表現フィルタを示しており，全体的な傾向を見ることで，好感が持たれていることが分かる．また，**図 7** には，評価表現フィルタで「可愛い」という表現を選択した場合の例を示している．フィルタ中の特定の表現を選択すると，その表現を含んだ文脈を簡単に表示できるとともに，その文脈中に多く出現するキーワードを表示可能としている．図中では，検索結果の概要文として「可愛い」が利用されている文脈が提示され，評価表現フィルタの「可愛い」の下には，「須賀くん」や「里奈タン」等が「可愛い」と関係するキーワードとして提示されている．この機能により，ユーザは評価表現フィルタ中に気になる表現があった場合に，その表現がどのような文脈でされているか簡単に知ることができる．これは製品の購入を考えているユーザが，事前に評判を調査するような場面にも利用可能である．

このフィルタを実現するための処理は以下のとおりである．

- 前処理
  ブログ記事が検索システムに登録される時点で，各ブログ記事中の，どの位置に，どの評価表現が存在するかを分析し，さらに評価表現が存在した場合には係り受け関係にあるキーワードを抽出する．これによって，各記事について，(評価表現，キーワード，記事中での位置）の情報を取得し，検索システム中の評価表現データベースに格納する．
- 検索時の処理
  検索結果中に含まれる評価表現の中で頻度の高いものを優先的に抽出し提示する．また，評価フィルタ中の評価表現が選択された場合には，検索結果のブログ記事中で，関連するキーワードの頻度が高いものを提示するとともに，その評価表現がどのような文脈で利用されているかを提示する．

処理概要を**図 8** に示す．

3.2.4 ブロガーフィルタ

ブロガーフィルタは，特定分野での注目のブロガーの発見を支援するともに，そのブロガーが書いた記事のみに検索結果を絞り込む機能を提供する．注目のブロガーの抽出には，我々が提案する EigenRumor と呼ぶアルゴリズム[6] を利用する．EigenRumor アルゴリズムは次節で詳しく述べるが，ブログ記事間のハイパーリンクを分析することにより，ブロガーの authority スコアと hub スコアと，ブログ記事に対する reputation スコアと呼ぶ 3 種類のスコアを算出するものである．

ブロガーフィルタの実現には，このうち reputation スコアのみを使用する．ブロガーに対するスコアである authority スコアと hub スコアを利用しないのは，ブロガーフィルタでは芸能や政治等の様々な分野のキーワードが入力された検索結果集合の中で，そのキーワードに関連する注目ブロガーを抽出することが求められるが，authority スコアと hub スコアはいずれもキーワードに依存しないグローバルなスコアだからである．reputation スコアについても後で述べるようにキーワードとは無関係に算出されるものであるが，ブロガーフィルタでは，キーワードの検索結果集合に含まれるブログ記事の reputation スコアをブロガー単位で集計することで，キーワード依存の（ブロガー）スコアとしている．なお，あるブロガーが書いた（キーワードに依存しない）全ブログ記事の reputation スコアの総和がそのブロガーの authority スコアとなっている．

ブロガーフィルタを用いることにより，たとえばキーワード「喰いタン」で検索すると，このドラマによく言及して注目度の高い記事を多く書くブロガー（ブログサイト）の一覧が表示され，さらに，ブロガーを選択すると，そのブロガーが書いた記事のみに検索結果を絞り込むことが可能になる．このような操作により，特定分野での注目ブロガーの発見を支援する．

このフィルタを実現するための処理は以下のとおりである．

- 前処理
  ブログ記事が検索システムに登録される時点で，

図 7 評価表現フィルタの例（評価表現を選択した場合）
Fig. 7 Example of Sentiment Filter (A case which a sentiment word is selected).

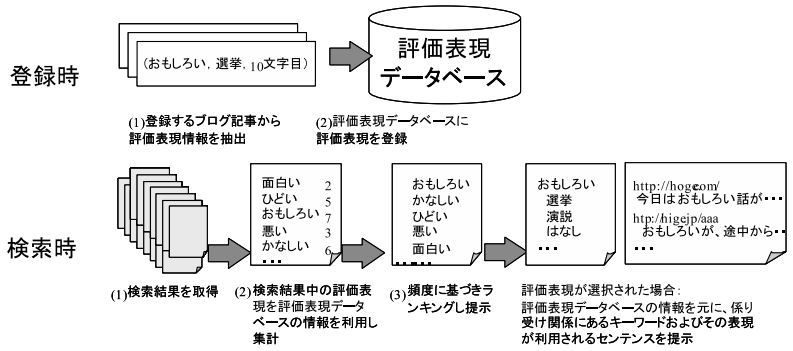

図 8 評価表現フィルタの生成フロー
Fig. 8 Constructing process of Sentiment Filter.

URL により，各ブログ記事がどのブロガーが書いた記事か判別し，検索システム中のデータベースに格納する．また，ブロガーおよびブログ記事の評価値を算出するため，ブログ記事間のリンクを収集し，リンクデータベースを作成する．この情報をもとに，次節で示す EigenRumor アルゴリズムを利用して，ブログ記事の評価値を算出し，検索用データベースに格納する．この処理は，1 日数回程度のバッチ処理として行われる．

- 検索時の処理
  検索結果中に含まれる記事を書いたブロガーを取得し，上記で示した手法によりブロガーをランキングし，検索条件に関連する話題で注目のブロガーを優先的に提示する．

処理概要を**図 9** に示す．

3.3 ランキングアルゴリズム

BLOGRANGER ではすでに述べたように，4 つのフィルタを備えることにより，検索結果の分類と概観を可能とし，所望のブログ記事への到達を支援する．しかし，ブログ記事は玉石混交であることおよびキーワードによっては膨大な検索結果集合となることから，なんらかの指標でブログ記事をランキングし，上位の記事に対してのみフィルタを適用することは，情報の取捨選択とフィルタ生成のコストの低減の両面から望ましい．そこで BLOGRANGER では，更新日時順の検索結果ランキングに加えて，EigenRumor と呼ぶブロガーおよびブログ記事の注目度に基づくランキングを提供している．注目度に基づく検索結果のランキングは，前節で述べたブロガーフィルタだけではなく，今回の検索目的のいずれを達成するにも有益であると考えられるため，BLOGRANGER が提供するすべてのフィルタについて適用可能とした．このように BLOGRANGER におけるランキングアルゴリズムは主要な技術要素の 1 つであることから以下で詳しく述べる．

ブロガーやブログ記事のランキングには，ブログサイトへのアクセス数や RSS フィードの購読数等をもとに情報をランキングする方法や，Web 検索で利用

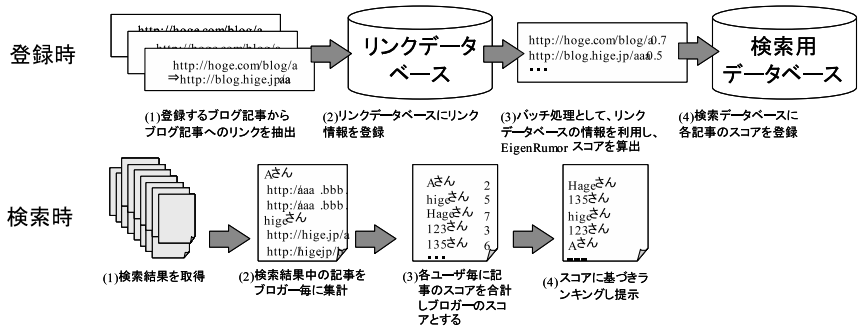

図 9 ブロガーフィルタの生成フロー
Fig. 9 Constructing process of Blogger Filter.

されるリンク解析を利用する手法[1),9)] があげられる．

しかし，アクセス数を利用するためには，個々のブログサイトへのアクセスカウンタ機能の組み込みや，ブログサービスプロバイダからのアクセス情報の提供が必要となり，網羅的に情報を収集することは困難である．RSS フィードの購読数についても幅広く普及した RSS リーダからの情報が必要となる．

一方，リンク解析を行うにはこのような問題はないが，Web で用いられている手法を単純に利用するにはいくつか問題がある．1 つはブログ空間でのリンクがスパースである点である．ブログサイト全体としての被リンクがある程度存在する場合であっても，一般にブログ記事単位ではスパースである．我々は，2004 年 10 月 16 日から 2005 年 2 月 3 日までの間に日本国内の約 10 社の主なブログプロバイダから，約 305,000 のブログサイトの約 9,280,000 のブログ記事の収集を行った[6)]．この 9,280,000 記事の中で 1 つ以上のハイパーリンクを有しているブログ記事は，1,520,000（16.3%）記事であった．しかし，このうちブログに対するリンクは，わずか 116,000（1.25%）記事にとどまっている（リンクがブログに対するものか否かは，我々が収集したデータセットに含まれているか否か判断しているため，実際にはこの値よりは若干多い）．他のブログから参照されていたブログ記事は 107,000（1.15%）であり，さらに小さい．もし，このデータセットのブログ記事に対して PageRank を計算した場合は，98.85%のブログ記事はスコアが 0 となる．これでは検索結果のランキングとして利用するには小さすぎる．もちろんブログ外からのリンクを含めると若干増えることが想定されるが，ブログからもリンクされない記事がブログ外からリンクされることは少ないと考えられる．

また，ブログ記事のスコアに関して，投稿してから記事にスコアが付与されるまでのタイムラグの問題（外部からリンクをもらわないと有益なスコアは付与されないため）もあり，新しさが重要な要素を占めるブログにおいては大きな問題となる．

我々は，これらの問題を解決するため，EigenRumor アルゴリズム[6)] を提案している．EigenRumor アルゴリズムは，ブログの編集主体が個人であるという特徴を生かし，hub スコア（情報評価能力を表すスコア）と authority スコア（情報提供能力を表すスコア）をブログ記事ではなく，ブロガーの属性としてリンクを集約して算出する．これにより，過去の実績から authority スコアが高いブロガーによって書かれた記事は，その記事自体に被リンクがない初期の段階でも，ある程度高いスコアとすることを可能にしている．

上記のデータセットの場合，ブログ記事単位ではなくブログサイト単位（ブロガー単位）でリンクを集約すると，36,200（11.9%）のブログサイトは他のブログサイトとリンクを持ち，そのうち 28,300（9.28%）のブログサイトは，他のブログサイトからの被リンクを有していた．収集したデータセット中に含まれるこの 28,300 サイトの有する記事数は 862,000 記事であり，全体の 9.3%にあたる．EigenRumor では，この 9.3%の記事について被ゼロのスコアを与えることができる．

この 9.3%という値は，決して大きな値ではないが，そもそもブログ記事の場合，本人にしか意味のない備忘録として利用されているものも少なくなく，注目すべきブログサイトは，この 9.3%に含まれている可能性が高い．

EigneRumor アルゴリズムは，HITS と同様にリンクから生成した行列の固有ベクトルを計算することにより各スコアを計算することを基本とする．ただし，HITS では Web ページ間のハイパーリンクから隣接行列とするのに対し，EigenRumor では，ブロガーとブログ記事との間のリンクから隣接行列を生成するこ

とを特徴としており，これらのリンク分析を行うことで，ブロガーの情報提供能力と情報評価能力を算出するという特徴を有する．

EigenRumor アルゴリズムは，ブログ空間だけではなく，BBS やメーリングリスト等，コミュニティ参加者の identity（アカウント ID 等）が観測できる様々なコミュニティに対しても適用できる．そこで以下では，ブロガーをエージェント，ブログ記事をオブジェクトとも呼ぶ．

EigenRumor アルゴリズムは，m 人のエージェント（ブロガー），n 個のオブジェクト（ブログ記事）から構成されるコミュニティを前提とする．

エージェント i がオブジェクト j を提供したとき，i から j への「情報提供リンク」を生成する．コミュニティにおける全提供リンクの状態を，情報提供行列 $P = [p_{i,j}]$（$i = 1 \cdots m, j = 1 \cdots n$）で表す．すなわち，エージェント i がオブジェクト j を提供した場合は，$p_{i,j} = 1$，提供していない場合は $p_{i,j} = 0$ とする．

エージェント i がオブジェクト j を $e_{i,j}$ と評価したとき，i から j への「情報評価リンク」を生成する．コミュニティにおける全評価リンクの状態を，情報評価行列 $E = [e_{i,j}]$（$i = 1 \cdots m, j = 1 \cdots n$）で表す．$e_{i,j}$ は i の j に対する支持のレベルを表すが，ここでは [0,1] の値をとることとし，1 が最大の支持レベルを表すこととする．

ブログに EigenRumor アルゴリズムを適用するためには，収集したブログ記事集合から情報提供行列 P と，情報評価行列 E を抽出する必要がある．

このためには，まずエージェント（ブロガー）の identity が識別できなくてはならないが，ブログの場合には，編集主体が個人であること，およびブログ・ホスティングサービスごとに各ブログのトップページの URL の形式がほぼ決まっているため，ブログサイトトップの URL をエージェントの identity とすればよい．また，オブジェクト（blog 記事）の identity については，基本的にブログでは，各 blog の記事エントリに永続的な URL が付与されるので，これを用いる．

この結果，情報提供リンクはブロガーのトップページの URL（エージェント）と，その配下にあるブログ記事の URL（オブジェクト）の 2 つ組の集合として表現できる．

情報評価リンクは，「あるブログ i から外部のブログある記事 j に対して，リンクが存在する」≒「ブロガー i が記事 j に対して関心がある」という仮説に基づき，リンクの有無により，情報評価リンクを $e_{i,j} = 1$ あるいは $e_{i,j} = 0$ とする．この結果，情報評価リンクはブロガーのトップページの URL（エージェント）と，その配下にある各記事エントリに含まれる外部のブログ記事に対する URL（オブジェクト）の 2 つ組の集合として表現できる．

なお，トラックバック機能により，自動的に生成されるトラックバックリンクは，情報評価リンクとは見なさない．なぜなら，トラックバックされたブロガーはトラックバックしたブログ記事に対して関心があることを意味しないからである．幸いなことに，トラックバックリンクは，その構造から通常のリンクと区別できるため，これを無視することができる．

EigenRumor アルゴリズムは，こうして取得した情報提供行列 P と情報評価行列 E の 2 つの隣接行列から，以下の authority ベクトル $\vec{a}$，hub ベクトル $\vec{h}$，reputation ベクトル $\vec{r}$ の 3 つのスコアベクトルを算出するものである．

authority スコア a_i はエージェント i（$i = 1 \cdots m$）が，過去，どの程度コミュニティから支持を受けたオブジェクトを提供してきたかを示す指標である．このスコアが高いエージェントは情報提供の面でコミュニティに貢献する能力があることを示す．ここで全エージェントの authority スコアを，$\vec{a} = [a_1 \cdots a_m]^T$ と表記し，authority ベクトルと呼ぶ．

hub スコア h_i はエージェント i（$i = 1 \cdots m$）が，過去，コミュニティに提供されるオブジェクトに対してどの程度コミュニティの方向性に沿った評価情報を投入してきたを示す指標である．このスコアが高いエージェントは情報評価の面でコミュニティに貢献する能力があることを示す．ここで全エージェントの hub スコアを，$\vec{h} = [h_1 \cdots h_m]^T$ と表記し，hub ベクトルと呼ぶ．

reputation スコア r_j はオブジェクト j（$j = 1 \cdots n$）が，どの程度エージェントからの支持を受けているかを示す指標である．このスコアが高ければ高いほど，その情報はコミュニティから支持を受けているものであることを示す．ここで全オブジェクトの reputation スコアを，$\vec{r} = [r_1 \cdots r_n]^T$ と表記し，reputation ベクトルと呼ぶ．

これらのスコアを計算するため，以下の仮説を導入する．

- authority スコアの高いエージェントが提供するオブジェクトは高い reputation スコアを持つ．
- hub スコアの高いエージェントが支持したオブジェクトは高い reputation スコアを持つ．
- reputation スコアが高いオブジェクトを提供したエージェントは高い authority スコアを持つ．

- reputation スコアが高いオブジェクトを支持したエージェントは高い hub スコアを持つ.

これらは，以下の 4 式で表現できる．

$$\vec{r} = P^T \vec{a} \quad (1)$$

$$\vec{r} = E^T \vec{h} \quad (2)$$

$$\vec{a} = P\vec{r} \quad (3)$$

$$\vec{h} = E\vec{r} \quad (4)$$

ここで，式 (1) と式 (2) を両立させるため，式 (1) と式 (2) を線形統合して，

$$\vec{r} = \alpha P^T \vec{a} + (1-\alpha) E^T \vec{h} \quad (5)$$

を利用する．ここで，α は [0,1] を定義域とする実数であり，適用先のコミュニティの特性に応じて調整されるものとする．つまり，α が 1 に近いほど，情報提供リンクを重視し，0 に近いほど，情報評価リンクを重視して reputation スコアが算出されることとなる．上記の式 (5) に，式 (3) と式 (4) を代入すると次式を得る．

$$\begin{aligned} \vec{r} &= \alpha P^T P\vec{r} + (1-\alpha) E^T E\vec{r} \\ &= (\alpha P^T P + (1-\alpha) E^T E)\vec{r} \\ &= S\vec{r} \end{aligned} \quad (6)$$

ここで，$S = (\alpha P^T P + (1-\alpha) E^T E)$ は reputation スコア推移行列と呼ぶ．上記，式 (6) を満たす $\vec{r}$ は一般的には存在しないが，S の要素が非負の実数の場合には，

$$\lambda \vec{r} = S\vec{r} \quad (7)$$

を満たす実数ベクトル $\vec{r}$ は存在する．λ は定数であって S の固有値，$\vec{r}$ は S の固有ベクトルと呼ばれる．

S は非負行列であるから，HITS と同様にべき乗法により式 (6) を繰り返し演算し，各ループで $\vec{r}$ をユークリッドノルムにより正規化することにより，$\vec{r}$ は S の固有値最大の固有ベクトル（principal eigenvector）を求めることができる．$\vec{r}$ が求められれば式 (3)，(4) により，$\vec{a}$ および $\vec{h}$ も算出できる．

以上が EigenRumor アルゴリズムの基本である．ただし，EigenRumor アルゴリズムを BLOGRANGER に適用するにあたってはいくつかの修正を行った．ランダムジャンプの導入とリンクの正規化である．

EigenRumor アルゴリズムは，HITS をベースとしているが，PageRank と HITS との大きな相違点として HITS ではランダムジャンプ[1] を導入していないことがある．ランダムジャンプは既約ではない状態推移行列を既約なものとするためには不可欠なファクタであるが，同時に固有値最大の固有ベクトルがコミュニティ全体の意見を反映したものにするという効果もある．いい換えると，ランダムジャンプを導入しなければ，固有値最大の固有ベクトルは最大勢力のサブ・コミュニティのみの意見しか反映したものとならない．HITS アルゴリズムがランダムジャンプを導入することなく良い検索結果が得られるのは，HITS では基本的にキーワードが与えられた検索結果の集合に対して，アルゴリズムを適用してスコアリングすることを前提にしているからである．そこで，BLOGRANGER では，情報評価行列 E のすべての要素に一定の割合で値を与えることによってランダムジャンプ相当を実現している．

2 つ目の修正はリンクの正規化である．EigenRumor アルゴリズムを適用するために生成した情報評価行列 E や情報提供行列 P の要素は，正規化せずに利用すると，スパマーによる大量の記事やリンクによって authority スコアや hub スコアが大きく影響を受けるという問題がある．そこで，これらの行列の正規化が不可欠である．しかし，我々の実験によれば，PageRank と同様の正規化，すなわち E や P の行ベクトルの総和が 1 となるようにした場合には，良いランキングは得られない．これは，Web ページの場合には 1 ページあたりのリンク数は平均 7〜10 といわれ，その分散も大きくないのに対し，ブログサイトの場合は，1 つのブログサイトあたりの記事数やそこから出ているリンク数の分散が大きいことによるものと考えられる．このような特徴を有する E や P を，無理に行ベクトルの総和を 1 となるように正規化した場合は，逆に，投稿記事の少ない人のリンクによる影響を強く受けてしまい，多くの記事を書く人の貢献が軽くなりすぎる結果となる．そこで，BLOGRANGER では E や P の行ベクトルの総和がリンク数や記事投稿数の平方根に比例した値に正規化する等の中間的な正規化を行った．

また，ブログコミュニティは時間的な経過にともない記事やリンクが増大していく．このため，上記のような正規化を行うことを前提としたうえで，過去のすべての投稿やリンクを同じように扱った場合には，古いブロガーが投稿した新しい記事は，新しいブロガーのものよりも相対的に低い重みとなってしまう．そこで BLOGRANGER では，たとえば 1 日経過するごとに E や P の要素に定数（たとえば 0.98）かけてリンクを減衰させる仕組みを導入した．

3.4 システム構成

BLOGRANGER のシステム構成を**図 10** に示す．BLOGRANGER では，すでに述べたように，ユーザから入力されたキーワードを含む検索結果に対して，

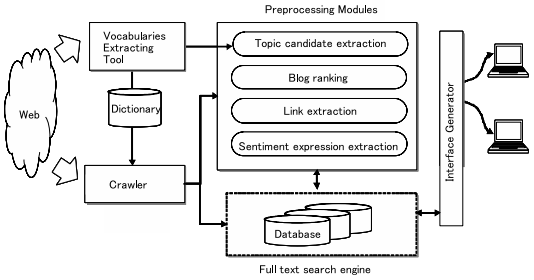

図 10 システム構成
Fig. 10 System overview.

4 つのフィルタを生成し，検索結果とともに表示する．しかし，これらのフィルタの生成を，すべて検索時に動的に実行することは実用的でない．

そこで，個々のフィルタの説明でも述べたように，事前に処理可能な部分については，ブログ記事を収集する際，もしくはバッチ処理のタイミングで事前に処理しておくことで，検索時の処理を削減している．実際には，トピックフィルタ用の語彙，リンク先フィルタおよびブロガーフィルタ用のリンク先 URL，評価表現フィルタ用の評価表現の抽出をブログ記事収集時に行い，それらの統計情報の作成，リンク先 URL のタイトル収集，およびブロガーフィルタ用のランキング計算をバッチ処理で定期的に行っている．これらの検索システムのアーキテクチャ上の工夫により，機能性と検索効率をあわせ持つようにしたことも BLOGRANGER の 1 つの技術的な特徴である．

4. 評　　価

2 章で述べたように，BLOGRANGER の目的には，特定の情報に到達するというよりは，ブログ記事中に存在する様々な情報を発見しながら本当に欲しい情報を見つけるというブラウジング的な要素も含んでおり，通常の検索システムのように正解データを作成し，その検索精度だけを評価することはかならずしも BLOGRANGER の適切な評価ではないと考えている．

また，Intelliseek がブログのコーパス[28] を提供しているが，検索課題の設定が行われていない等，少なくとも現段階では，検索システムの評価を行うためのものではない．

以上から，本システムの評価ではアンケートによる主観評価を行った．

4.1 アンケート調査方法

公募型インターネットアンケートパネルの登録者約 22 万人の中から一般男女を無作為抽出し，メールにより Web アンケートを依頼し，調査を行った．期間は，2006 年 2 月 10 日（金）〜2 月 12 日（日）であり，6,700 人に調査を依頼し，2,191 人から回答を得た．回収率は 32.7%である．

Web アンケートでは，単純に「役に立ったか」等の質問をしても有益な結果が返ってこないことが想定されるので，我々のアンケートでは，アンケート回答中に実際に，BLOGRANGER や比較対象のシステムに触れてもらいながら回答できるように工夫をした．

本アンケートの設問項目のうち本論文に関連する部分を付録 1 に示す．この調査では，40 個のキーワードを提示し，その中から 1 つを選択させる．

そして，そのキーワードによる，それぞれのシステム（一般の Web 検索[24]，一般のブログ検索[25]，BLOGRANGER[22]）の検索結果を表示し，それについて選択方式により評価を入力させた．各検索結果の表示は，アンケート回答ページに「検索ボタン」を埋め込み，その検索ボタンを押すことで，容易に表示できるようにした．

なお，アンケート回答者の負荷を考え，1 人あたり 1 つのキーワードのみで評価を行っている．

BLOGRANGER に格納しているデータは，最新 5 週間分のブログ記事記事（約 100 万件のブログサイトから収集した 1,000 万ブログ記事）である．一方の Web 検索では Google のデータベースを利用した検索エンジンであるため，数十億から数百億の Web ページを対象としていると考えられる．また，今回比較対象としたブログ検索では約 2500 万件の記事を検索対象としている．

このように今回利用した BLOGRANGER のデータベースは，Web 検索と比較して，非常に小さいものであるが，2 章で述べた，話題検索，評判検索，ブロガー検索に関していえば BLOGRANGER の方が有益であるというのが我々の 1 つの仮説である．

この調査においては，検索キーワードが何かにより，その結果が大きく異なることが予想される．そこで，公平な評価を行うため，Web 検索とブログ検索のそれぞれでよく利用されるキーワードを，以下に示す Web で公開されているキーワードランキングをもとに公平に抽出した．

- 2006 年 1 月期 goo（Web 検索）急上昇キーワードランキング（上位 20）[26]
- 2006 年 2 月 9 日テクノラティ検索語ランキング（上位 10）[30]
- 2006 年 1 月 9 日〜2 月 8 日 BLOGRANGER キーワードランキング（上位 10）

表 1 選択されたキーワードに対する検索目的
Table 1 Search goals for the selected keywords.

検索目的	割合
公式ページ	38.02%
詳細情報	36.92%
話題	48.52%
評判	21.50%
ブロガー	7.99%
その他	1.32%

表 2 Web 検索と BLOGRANGER の有益さの比較
Table 2 Comparison of the usefullness between Web search and BLOGRANGER.

回答数	Web 検索	BLOGRANGER	不明	有意差
2191	907	698	586	あり

これらをマージし重複を削除し，アンケート期間の時節語（トリノ，バレンタイン，皇室）を追加した合計 40 語をユーザに提示し，選択してもらう形とした．選択肢には，話題語が多く含まれるが，HIS，DELL，トヨタ，JTB 等の公式ページを検索する目的と考えられるキーワードも一定の割合で含まれている．また，これらのキーワードの提示順序は，アンケートにアクセスするたびにランダムに変化するようにした．付録 2 に回答者が選んだ頻度，割合とともにキーワードのリストを示す．

4.2 キーワード検索の目的

表 1 に，ユーザがアンケート中（設問 5）で選択した検索目的を示す．今回評価に利用した検索キーワードは，無作為にブログ検索と Web 検索の両方のキーワードランキングから取得したが，実際に選択されたキーワードは，付録 2 に示すように最近の話題に関連する語が比較的多く選択された．ただ，このような条件とはいえ，話題検索に興味を持つユーザが，公式ページの検索や詳細情報の検索を上回っていること，また話題語とはあまり相関のない評判やブロガーを検索するというニーズが存在することがアンケート結果から得られた．この結果から，2 章で提案した，話題検索，評判検索，ブロガー検索のニーズが存在することが確認できた．次節以降では，我々の提案するシステムがこれらの検索ニーズを満たすのに有益であるかどうかの結果を示す．

4.3 Web 検索との比較

表 2 に Web 検索と BLOGRANGER のどちらの検索結果の方が有益であったか（設問 8）を回答した結果を示す．キーワードの選択肢に最近の話題語が多いにもかかわらず，全体としては Web 検索の方が良い結果であるという回答が多かった．これは，BLOGRANGER では，検索対象が最近のブログ記事に絞られ，検索対象の記事が Web 検索の方が圧倒的に多いこと，および，検索目的として公式ページや詳細情報を探すことが，ブログ検索目的と考えられる話題や評判を探すこととほぼ同じ割合を占めたことが理由であると考える．しかし，**表 3** に示したように，設問 5 で回答した検索目的ごとに設問 8 の結果（Web 検索と BLOGRANGER のどちらが有益であったか）を見ると，2 章でブログの特徴を生かすことで有益となると仮説を立てた話題検索，評判検索，ブロガー検索については BLOGRANGER の結果の方を良いと判断した人が多いことが分かる☆．またこれらの結果について，「BLOGRANGER と Web 検索の検索結果に有益性の差はない」という帰無仮説を立て，5%の有意水準で有意差検定を行った結果，話題検索およびブロガー検索の場合には，帰無仮説は棄却され，有意差が確認できた．ただし，評判検索の場合には，帰無仮説は棄却されず，有意差は確認できなかった．

一方，**表 4** には，Web 検索と BLOGRANGER のどちらの検索結果の方が有益であったか（設問 8）の回答を，その判断理由（設問 9）で分類したデータを示す☆☆．このデータでは，2 章で述べた 3 つの検索すべてで，Web 検索より BLOGRANGER の結果の方が有効であると回答した人の割合が 2 倍以上となっている．また，上記と同様の検定を行った結果，話題検索，評判検索，ブロガー検索のすべてにおいて，Web 検索結果との間に有意な差があることが確認できた．

前者の分析と後者の分析との傾向は同様の傾向であったが，後者では，どちらの結果が良かったかを回答した直後に理由を回答した結果となっており，目的間での結果の差がより明確になっている．

両方の結果を統合すると，BLOGRANGER は，我々が 2 章で提案した検索を実現する場合には Web 検索より有益であるといえる．

4.4 従来のブログ検索との比較

また，本アンケートでは，BLOGRANGER によるブログ検索を体験することにより，ブログ検索に対する意識の変化を調査するため，通常のブログ検索と Web 検索を比較（設問 6）した後に，BLOGRANGER と Web 検索を比較する設問（設問 8）を設けている．通常のブログ検索と BLOGRANGER の結果を，ユーザの検索目的別に示した図を**図 11** に示す．縦軸は，上記の設問において，いずれかのシステムを選択したユーザがブログ検索システム（設問 6 の場合「通常の

☆ 検索目的は複数回答のため，合計値は表 2 の値とは異なる．
☆☆ 判断理由は複数回答のため，合計値は表 2 の値とは異なる．

表 3 Web 検索と BLOGRANGER の有益さの比較
Table 3 Comparison of the usefullness between Web search and BLOGRANGER.

検索目的	回答数	Web 検索	BLOGRANGER	不明	有意差
公式ページ	833	411	201	221	あり
詳細情報	809	345	257	207	あり
話題	1063	362	430	271	あり
評判	471	167	177	127	なし
ブロガー	175	51	89	35	あり

表 4 Web 検索と BLOGRANGER の有益さの比較（結果を良いと判断した理由から）
Table 4 Comparison of the usefullness between Web search and BLOGRANGER.

判断理由	回答数	Web 検索	BLOGRANGER	その他	有意差
公式ページ	754	530	62	162	あり
詳細情報	732	346	228	158	あり
話題	968	225	487	256	あり
評判	302	75	155	72	あり
ブロガー	101	7	71	23	あり

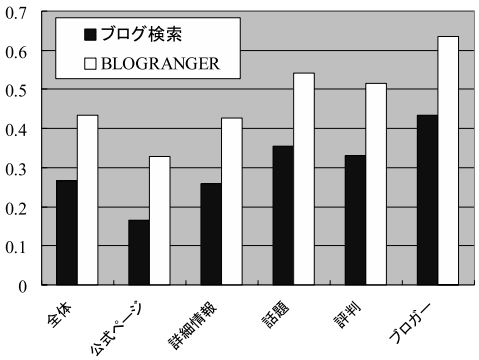

図 11 従来のブログ検索と BLOGRANGER の検索結果の有益性比較
Fig. 11 Comparison of usefullness between traditional bog search and BLOGRANGER.

ブログ検索」，設問 8 の場合「BLOGRANGER」）を選択した割合である．

まず，全体的な傾向を見ると，Web 検索の結果と比較した場合に通常のブログ検索を支持した人（通常のブログ検索の方が良い結果と答えた人）の割合が 19%だったのに対し，BLOGRANGER では 34%に大幅に向上したことが分かる．また，この結果の有意差を確認するため，「BLOGRANGER と通常のブログ検索の検索結果に有益性の差はない」という帰無仮説を立てた検定を行ったところ，5%の有意水準で帰無仮説が棄却され，有意な差があることが確認できた．しかしながら，通常のブログ検索，BLOGRANGER の両方を対象としたアンケートにおいて，ユーザの半分以上は，Web 検索を支持している．

次に，それぞれの検索目的別の結果を見ると，「公式ページ」，「詳細情報」を探す目的では，「通常のブログ検索」，「BLOGRANGER」とも半分以下のユーザにしか支持されていない．これはブログ記事がそのようなコンテンツをあまり含まないことを考えると当然の結果であるといえる．一方，ブログ記事中に目的のコンテンツが多く含まれており，我々がブログ検索をすることが有益であると考えている「話題」，「評判」，「ブロガー」を探す目的においての結果を見ると，「通常のブログ検索」では，30%から 40%程度と過半数以下のユーザにしか支持されていないことが分かる．この結果は，単に Web 検索と同様にキーワード検索のブログ記事検索を提供したとしても，必ずしも多くのユーザを満足させることができないことを示しているといえる．それに対して，BLOGRANGER の結果を見ると，ユーザの支持率は，3 つすべての検索目的で 50%を超え，「ブロガー」を検索する目的の場合には約 65%のユーザが支持していることが分かる．この結果は，今回提案するフィルタにより，これまでユーザがブログから取得できなかった情報を取得しやすくしたことが原因であると考えられる．

上記で示した「情報を取得しやすい」という部分について，通常のブログ検索と比較して，BLOGRANGER ではどのように情報を取得しやすいかについて示す．

まず，「話題」を探したいという場合，通常のブログ検索では，入力されたキーワードに対して，検索結果の一覧が得られるのみである．このため，個々の検索結果のタイトルや概要文を読むことで検索条件に関係する話題のうち所望のものが存在するか分析し，所望の情報を選択する．それに対して，BLOGRANGER の「トピックフィルタ」では，話題を直感的に理解しやすい固有名詞を利用して検索結果中の話題を図 4 に示す形で提供しており，ユーザはこれを参照することで容易に検索結果中に存在する話題を理解可能となり，所望の情報が存在した場合には，これを選択することで，特定の話題の情報だけに絞り込んだ検索を行うこ

ともできる．一方,「リンク先フィルタ」では，検索結果中で多くのブロガーによって参照されているニュースや注目のサイトを提示することで，ユーザが検索結果中で注目されている話題を容易に知ることを可能としている．両方のフィルタに共通な効果として，元々想定しなかった意外な話題の発見につながるという効果も考えられる．

また,「評判」を探したいという場合にも，通常のブログ検索では，入力されたキーワードに対して，検索結果の一覧が得られるのみであるため，検索結果中のタイトルと概要文からユーザの評判が含まれそうか判断し，それぞれの文書を閲覧してはじめて評判情報を得ることができる．それに対して，BLOGRANGER の「評判フィルタ」では，ブログ記事中に存在する評価表現を抽出し，集計した形で提示しており，これにより全体的な評判の傾向が容易に理解できる．さらに個々の評判表現を選択すると，下位構造として，どのような点についてその評価表現が利用されているかを提示したり，それら，評判表現が実際の文脈でどのように利用されているかを一覧で閲覧したりすることを可能としており，個々の詳細な表現についても，マウスで選択しながら容易に閲覧することが可能となっている．

最後に「ブロガー」に関してであるが，通常のブログ検索では，基本的に記事検索であるため，記事を経由してブロガーを探すという形となるが，BLOGRANGER の「ブロガーフィルタ」では，3 章で示したリンク関係を利用したランキングアルゴリズムを用いることで，入力されたキーワードについて言及しているブロガーのうちより注目されているブロガーを優先的に提示している．これにより，ユーザは所望の分野の人気ブロガーを容易に閲覧することが可能となっている．

以上，アンケート結果の分析から通常のブログ検索に対して BLOGRANGER が有益であるとの知見を得るとともに，BLOGRANGER の提供するフィルタが，通常のキーワード検索を補完し，話題検索，評判検索，ブロガー検索において，目的の情報の取得を支援することを示した．

4.5 各インタフェースの比較

設問 7 において，BLOGRANGER の 4 つの機能,「トピックで選ぶ」（トピックフィルタ),「ブロガーで選ぶ」（ブロガーフィルタ),「リンク先で選ぶ」（リンク先フィルタ),「感想で選ぶ」（評価表現フィルタ）の中でどれが最も有用であったかを聞いており，その結果を，**図 12** に示す．全体としてはトピックフィルタ

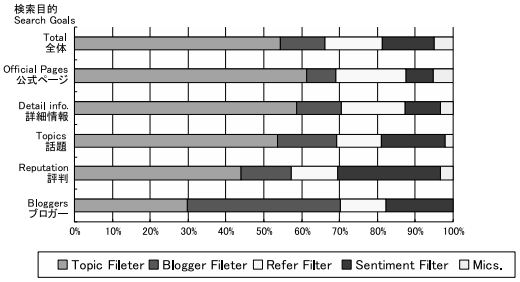

図 12 各フィルタの有用性の比較
Fig. 12 Comparison of usefullness among the proposed filters.

表 5 操作性に関するアンケート結果
Table 5 Questionnaire result for the usability.

	はい	いいえ	どちらでもない
理解しやすさ	58.97%	10.59%	30.31%
使いやすさ	56.37%	9.27%	34.14%

が最も評価が高い．これは，検索目的として話題を探すことに対するニーズが高いことによるものと考えられる．また，評判を探す目的では評価表現フィルタが，ブロガーを探す目的では，ブロガーフィルタの割合が高くなった．

これらの結果はインタフェースのデザインが設計どおり働いている結果であると考える．

4.6 操 作 性

設問 13 では，BLOGRANGER の操作方法の分かりやすさについて，設問 14 では，使いやすさについて質問している．結果を**表 5** に示す．どちらもほぼ同様の結果であり，ネガティブな回答はいずれも 10%程度であった．

これは，今回提案した BLOGRANGER がユーザにとって十分受け入れられることを示す結果といえる．

5. 関 連 手 法

近年，ブログの特徴を利用することで，通常の検索エンジンでは実現できなかった検索を可能とするシステムが登場している．

BlogPulse[21] はブログの構造を生かし，話題検索とブロガー検索を行うための様々な機能を提供している．特に “Conversation Tracker” は，リンク関係に基づきブロガー間にわたる議論の推移をトラックする機能を持っている．また，“BlogPulse Profiles” では，個人のブロガーをより詳細に知ることを可能とする機能として，特定のブロガーの参照先や，参照元，類似したブロガーを抽出する機能を持っている．我々の手法が大量のデータの中から部分的な情報を探し出すこ

とに注力しているのに対し，BlogPulse の手法はブログ空間のごく一部の情報を起点に情報を広げようというアプローチといえる．これらは相反するものではなく，相互に補完する関係にあると考えられる．

また，Nakajima らの研究[12] もブログの構造を利用した研究である．この手法では，ブログ記事およびブログ記事からリンクされる外部サイトへのリンクをもとに，複数のブロガーが特定の話題について語っているスレッドを抽出し，これを分析することで重要なブロガーを特定する手法を提案している．この手法では，リンクの張り方やスレッド内のブログ記事のポスト数，ブログ記事間の類似度をもとに，ブログのスレッドを盛り上げる Agitator や，スレッドで書かれる内容をまとめる Summarizer を抽出しており，ブロガーにタイプ付けを行う手法と見なすこともできる．一方，我々の提案する EigenRumor アルゴリズムでは，注目度という観点で一次元のランキングを行っており，Nakajima らの手法によるタイプ付けと組み合わせることによって，複数の軸でブロガーをランキングすることが可能となり，より有益なブロガーフィルタの構築が行える可能性が考えられる．

一方，blogWatcher [16),23)] では，話題検索と評判検索を行うための機能として，入力したキーワードのバースト度[5] やキーワードに関連する評判情報の推移を提示し，それをもとに特定の時期のブログ記事にアクセスすることを可能としている．これにより，時間的な記事の傾向を把握したり，特定の時期のブログ記事に容易にアクセスしたりすることを可能としている．現在の我々のシステムでは最新の記事のみを扱っていることから積極的に時系列を意識した処理は行っていないが，今後検索対象とするブログ記事を増加させる場合には，学ぶべき点が多いと考えている．特にバーストを利用した手法は，我々の技術と相互に補完する関係であると考えている．たとえば，blogWatcher で話題がバーストした時期を見つけ，それぞれの時期の記事ごとに我々の手法で関連するトピックやリンク先を見つける等，双方の機能を組み合わせることでより有益な情報の発見，目的とする情報への効率的なナビゲーションが可能であると考えられる．一方で，評判を探すインタフェースとしてポジティブな表現と，ネガティブな表現に分離し，集計結果を提示する機能については，十分な有用性を感じるまでには至っていない．この機能自体は非常に興味深く，ニーズは高いと思われるが，現状では，センテンスや文書のポジティブ，ネガティブを判別する精度が十分でなく，結局本文を参照しない限り評判は分からないと考えているからである．このため，我々は，無理にポジティブ，ネガティブに分類した結果を提示するのではなく，評価表現フィルタ（3.5 節参照）で実現しているように評価表現の出現傾向を提示し，興味がある表現が使われているセンテンスを容易に参照できる仕組みとした．

また，話題検索に関係するシステムとして，kizasi.jp [29)] の検索機能があげられる．kizasi.jp では，検索結果に関係する話題の内容を大まかに把握することを目的として，入力したキーワードとともにテキスト中に出現するキーワードを関連語として提示している．我々の手法によるトピックフィルタも同様に検索結果の内容を概観することが 1 つの目的であるが，より効果的な内容把握，絞り込み検索を行うために，直感的に内容を把握し，特定の話題に結び付きやすい固有名詞を利用する点で異なる．固有名詞を利用することで，提示するワード中に不要な語が含まれる可能性を低減させ，また，固有名詞の種類に応じたキーワードの提示を行うことで，整理された形でキーワードを提示することが可能となり，全体的な内容を把握しやすくなる．また，固有名詞を利用することで，より情報の抽出を高精度で行うことも可能になる[14]．

ブログ検索のユーザニーズを分析した研究として，Mishne [11)] らの研究があげられる．ここでは，ブログ検索システムの検索ログを分析し，検索キーワードの種別や検索システムでのユーザの振舞いについて分析している．それに対して我々は，大規模なユーザアンケートの結果をもとに，ブログ検索と Web 検索の違いについて分析している点で，従来研究とは異なる新しい知見を提供している．

6. ま と め

本論文では，2 章でブログの情報コンテンツの特徴を生かしたブログ検索システムのユーザニーズについて議論し，ブログ検索に必要と思われる検索として，話題の検索，評判の検索，ブロガーの検索をあげた．3 章では，これらの機能にターゲットを絞った新しいブログ検索システム BLOGRANGER のアプローチとその実現方法を述べた．4 章で大規模なアンケート調査により，2 章であげた検索のニーズがあること，それらの検索を実現するうえで，BLOGRANGER が，Web 検索およびこれまでのブログ検索より有益であることを確認した．5 章では，ブログの特徴を生かしたサービスおよび研究をあげ，我々の手法との関係について議論した．今後は，関連研究で述べた別の手法との融合やブログの大きな特徴であるブロガー間のつながりを意識した検索手法を検討していきたい．また，

本実験システムの運用ログをもとに，より詳細なユーザニーズの分析等もあわせて行いたいと考えている．

参 考 文 献

1) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web Search Engine, *Proc. 7th international conference on World Wide Web 7*, Brisbane, Australia, pp.107–117 (Apr. 1998).
2) Chang, C.H., Lui, S.C. and Pu. C.: IEPAD: Information Extraction Based on Pattern Discovery, *Proc. 12th International Conference of World Wide Web*, Hong Kong, China, pp.4–15 (May 2001).
3) Cutting, D., Karger, D., Pedersen, J. and Tukey, J.: Scatter/Gather: A cluster-based approach to browsing large document collections, *Proc. 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark, pp.318–329 (June 1992).
4) Fujii, A., Itoh, K., Akiba, T. and Ishikawa, T.: Exploiting Anchor Text for the Navigational Web Retrieval at NTCIR-5, *Proc. NTCIR-5 Workshop Meeting*, Tokyo, Japan (Dec. 2005).
5) Fujiki, T., Nanno, T., Suzuki, Y. and Okumura, M.: Identification of Bursts in a Document Stream, *Proc. 1st International Workshop on Knowledge Discovery in Data Streams*, Pisa, Italy, pp.55–64 (Sep. 2004).
6) Fujimura, K., Inoue, T. and Sugizaki, M.: The EigenRumor Algorithm for Ranking Blogs, *Proc. WWW 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba, Japan (May 2005).
7) Grishman, R. and Sundheim, B.: Message Understanding Conference — 6: A Brief History, *Proc. 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp.466–471 (Aug. 1996).
8) Hearst, M. and Pederson, J.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proc. 19th annual international ACM SIGIR conference on Research and development in information retrieval*, Zurich, Switzerland, pp.318–329 (Aug. 1996).
9) Kleinburg, J.: Authoritative sources in hyperlinked environment, *J. ACM*, Vol.46, No.5. pp.604–632 (1999).
10) Kushmerick, N.: Wrapper Induction: Efficiency and Expressiveness, *Artifical Intelligence*, Vol.118, pp.15–68 (2000).
11) Mishne, G. and Rijke, M.: A Study of Blog Search, *Proc. 28th European Conference on Information Retrieval*, London, UK, pp.289–301 (Apr. 2006).
12) Nakajima, S., Tatemura, J., Hino, Y., Hara, Y. and Tanaka, K.: Discovering Important Bloggers based on Analyzing Blog Threads, *Proc. WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem*, Chiba, Japan (May 2005).
13) Shinzato, K. and Torisawa, K.: A Simple WWW-based Method for Sementic Word Class Acquisition, *Proc. International Conference on Recant Advances in Natural Language Processing 2005*, pp.493–500 (Sep. 2005).
14) Suhara, Y., Toda, H. and Sakurai, A.: Event mining from the Blogosphere using topic words, *Proc. 1st International Conference on Weblogs and Social Media* (*ICWSM 2007*), Boulder, Colorado, U.S.A. (Mar. 2007).
15) Zeng, H., He, Q., Zheng, C., Ma, W. and Ma, J.: Learning to cluster web search results, *Proc. 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, pp.210–217 (Aug. 2004).
16) 奥村 学，南野智之，藤木稔明，鈴木泰裕：blog ページの自動収集と監視に基づくテキストマイニング，人工知能学会研究会資料，SIG-SW & ONT-A401-01, pp.01-01–01-08 (2004).
17) 戸田浩之，中渡瀬秀一，片岡良治：特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案，情報処理学会論文誌：データベース，Vol.46, No.SIG13 (TOD27), pp.40–52 (2005).
18) 野口正人，廣川佐千男：SoftPath を用いた同系列単語抽出方式，情報処理学会研究報告知能と複雑系，Vol.2002, No.105, pp.15–20 (2002).
19) 総務省：ブログ・SNS の現状分析及び将来予測 (2006).
20) ask.jpブログ検索. http://ask.jp/bloghome.asp
21) BlogPulse. http://www.blogpulse.com/
22) BLOGRANGER. http://ranger.labs.goo.ne.jp/
23) blogWatcher. http://blogwatcher.pi.titech.ac.jp/
24) goo ウェブ検索. http://www.goo.ne.jp/
25) goo ブログ Search. http://blog.goo.ne.jp/
26) goo ランキング. http://ranking.goo.ne.jp/
27) Google Blog Search. http://blogsearch.google.com/
28) Intelliseek, WWE-2006 Weblog Data Challenge. http://www.blogpulse.com/www2006-workshop/datashare-instructions.txt
29) kizasi.jp. http://kizasi.jp/
30) Technorati JAPAN. http://technorati.jp/

付　　録

A.1 アンケート設問

設問 4　現在あなたが検索したいキーワードを以下から 1 つお選びください.
【キーワード 40 語を提示，付録 2 参照】

設問 5　設問 4 で選んだキーワードに関してどのような情報が探したいでしょうか？
- 公式ページ
- 詳細情報（まとめサイト等）
- 話題（面白い話題，最新の話題）
- 商品やサービスの評判
- 注目ブロガー（ブログ記事を書く人）
- その他

【設問 4 で選んだキーワードの Web 検索（goo）とブログ検索（goo）の検索結果を表示】

設問 6　今回選ばれたキーワードに関して，Web 検索とブログ検索でどちらがよい情報が得られましたか？
- Web 検索
- ブログ検索
- どちらともいえない

【設問 4 で選んだキーワードの BLOGRANGER の各検索結果を表示】

設問 7　設問 4 で選ばれたキーワードに関して，BLOGRANGER の 4 つの機能,「トピックで選ぶ」,「ブロガーで選ぶ」,「リンク先で選ぶ」,「感想で選ぶ」の中でどれが役に立ちましたが？
- 「トピックで選ぶ」
- 「ブロガーで選ぶ」
- 「リンク先で選ぶ」
- 「感想で選ぶ」
- その他

設問 8　設問 4 で選ばれたキーワードに関して，Web 検索とブログ検索（BLOGRANGER）でどちらがよい情報が得られましたか？
- Web 検索
- ブログ検索
- どちらともいえない

設問 9　前問の判断ポイントは何でしょうか？
【設問 4 で選ばれたキーワードに関する BLOGRANGER 検索結果を再表示】
- 公式ページ
- 詳細情報（まとめサイト等）
- 話題（面白い話題，最新の話題）
- 商品やサービスの評判
- 注目ブロガー（ブログ記事を書く人）
- その他

設問 13　BLOGRANGER の操作方法は理解しやすかったでしょうか.
- はい
- いいえ
- どちらともいえない

設問 14　BLOGRANGER は使いやすかったでしょうか.
- はい
- いいえ
- どちらともいえない

A.2 アンケートで提示したキーワード

キーワード	選択頻度	選択割合
株	189	8.63%
確定申告	187	8.53%
トリノ	171	7.80%
ニンテンドー DS	155	7.07%
オーラの泉	125	5.71%
白夜行	111	5.07%
バレンタイン	98	4.47%
宝くじ	85	3.88%
倖田來未	79	3.61%
ライブドア	72	3.29%
ホリエモン	68	3.10%
皇室	68	3.10%
細木数子	64	2.92%
どうぶつの森	61	2.78%
トヨタ	55	2.51%
JTB	48	2.19%
ジャニーズ	45	2.05%
HIS	42	1.92%
ハローワーク	42	1.92%
綾瀬はるか	41	1.87%
功名が辻	40	1.83%
W-ZERO3	38	1.73%
中古車	38	1.73%
KAT-TUN	35	1.60%
任天堂	30	1.37%
Dell	27	1.23%
待ち組	27	1.23%
上川隆也	21	0.96%
堂本剛	19	0.87%
井上和香	18	0.82%
Web2.0	15	0.68%
青木裕子	15	0.68%
ヨンエ	14	0.64%
サイバーファーム	13	0.59%
Opera	12	0.55%
ENDLICHERI	8	0.37%
4gamer	5	0.23%
Feedpath	4	0.18%
foobar2000	4	0.18%
久石	2	0.09%

(平成 19 年 3 月 20 日受付)
(平成 19 年 7 月 4 日採録)

(担当編集委員 江口 浩二)

戸田 浩之(正会員)

1997 年名古屋大学工学部材料プロセス工学科卒業．1999 年同大学大学院工学研究科材料プロセス工学専攻博士課程前期課程修了．2007 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻博士後期課程修了．1999 年日本電信電話（株）に入社．以来，情報検索，情報抽出，Web マイニングの研究開発に従事．現在，NTT サイバーソリューション研究所所属．博士（工学）．ACM SIGIR，電子情報通信学会，日本データベース学会各会員．

藤村 考(正会員)

1984 年北海道大学工学部電気工学科卒業．1989 年同大学大学院工学研究科情報工学専攻博士課程修了．同年日本電信電話（株）に入社．以来，トランザクション処理記述言語，汎用電子チケットシステム，電子決済システム，ブログマイニングの研究開発に従事．現在，NTT サイバーソリューション研所所属．電気通信大学大学院情報システム学研究科客員教授．工学博士．電子情報通信学会，日本社会情報学会各会員．

井上 孝史

1990 年京都大学工学部電気系学科卒業．1992 年同大学大学院工学研究科電気工学第二専攻修士課程修了．同年日本電信電話（株）に入社．主にテキスト処理，情報検索の研究開発に従事．現在，NTT サイバーソリューション研究所所属．日本データベース学会会員．

廣嶋 伸章(正会員)

1998 年慶應義塾大学理工学部数理科学科卒業．2000 年同大学大学院理工学研究科計算機科学専攻博士前期課程修了．同年日本電信電話（株）に入社．自然言語処理・テキストマイニングの研究開発に従事．現在，NTT レゾナント（株）所属．言語処理学会会員．

杉崎 正之

1993 年東京理科大学理工学部情報科学科卒業．1995 年同大学大学院理工学研究科情報科学専攻博士課程前期課程修了．同年日本電信電話（株）に入社．主にテキスト情報の自動分類技術の研究開発，および，ログ解析技術の研究開発に従事．現在，NTT レゾナント（株）所属．

片岡 良治(正会員)

1985 年千葉大学工学部電子工学科卒業．1987 年同大学大学院電子工学専攻修士課程修了．同年日本電信電話（株）に入社．以来，トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．現在，NTT サイバーソリューション研究所所属．

奥 雅博(正会員)

1982 年大阪府立大学工学部電子工学科卒業．1984 年同大学大学院工学研究科電子工学専攻博士前期課程修了．同年日本電信電話公社（現 NTT）に入社し，日英機械翻訳システム ALT-J/E，日本文推敲支援システム REVISE-T，オペレータレス電話番号案内システム等における自然言語処理技術，特に日本語処理技術の研究開発に従事．新事業開発の企画，光を活用した自由で豊かな生活の実現を目指した光マーケットクリエーション活動等を経て，現在 NTT サイバーソリューション研究所において，検索をはじめとするブロードバンドインターネットサービスに関する研究開発に従事．実験サイト goo ラボ（http://labs.goo.ne.jp/）で様々な技術を公開実験中．博士（工学）．電子情報通信学会，言語処理学会各会員．