

求める情報へユーザを導く ナビゲーション型の情報検索技術

「検索」はポータルサービスを利用するユーザ行動の起点として欠かせない機能になっています。本稿では、今の場面に必要な情報をユーザへ適切に提供できるポータル検索サービスの実現に向けた次世代の情報検索技術について紹介します。

かたおか りょうじ とだ ひろゆき
片岡 良治 / 戸田 浩之

うえまつ ゆきお ふじむら こう
植松 幸生 / 藤村 考

べっしょ かつじ にしおか しゅういち
別所 克人 / 西岡 秀一

NTTサイバーソリューション研究所

ナビゲーション型検索サービスの時代へ

ブロードバンドの普及に伴い、日常活動のさまざまな場面でインターネットを利用する機会が増えていますが、その中でも特に中心的な役割を果たす検索サービスの領域が急激に広がろうとしています。Web上の画像・動画・音楽コンテンツを対象としたマルチメディア専門の検索サービスや、時々刻々発信されるニュース記事やブログ記事を専門に扱う検索サービスなど、膨大な情報を扱うさまざまなポータルサービスにおいて「検索」は行動の起点となる欠かせない機能となっています。

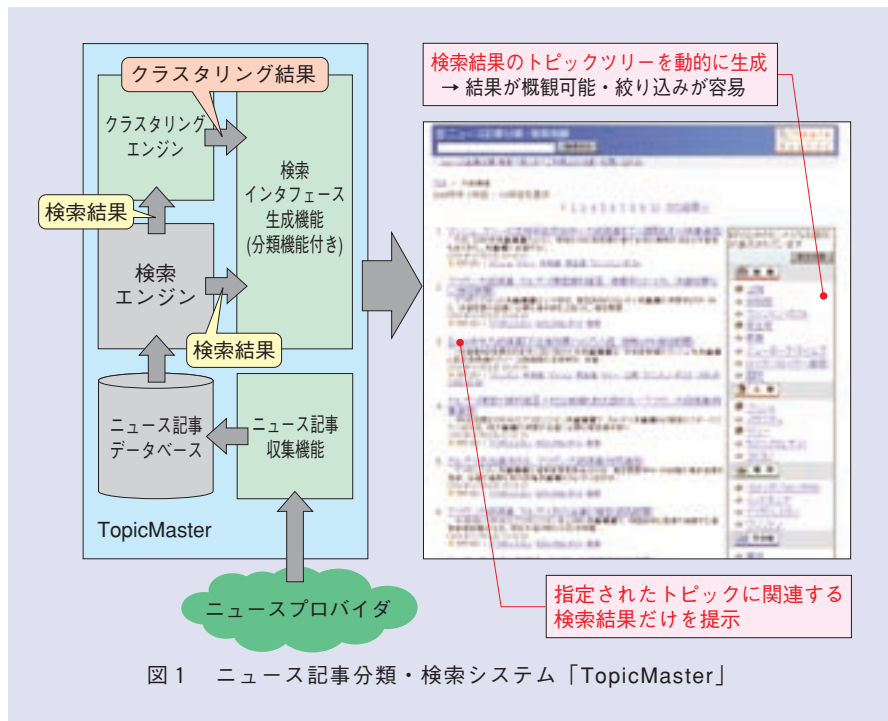
これまで検索サービスは、インターネットの普及段階と呼応して、主要情報を案内する第1世代のディレクトリ型、情報の大量化に伴いキーワードによる直接検索を可能とする第2世代の検索エンジン型と進化してきました。そしてこれからのインターネット利用が多様化・広範化される第3世代では、ユーザが必要な情報にさまざまなスタイルで出会うことを支援するナビゲーション型検索サービスの時代になると

NTTサイバーソリューション研究所では考えています。本稿では、ナビゲーション型の次世代検索サービスを実現する技術を、いくつかの研究開発事例を通して紹介します。

ニュース記事分類・検索システム「TopicMaster」

インターネット上の情報量は日々増加の一途をたどっており、検索サービ

スを利用する際の検索結果の量も膨大になりがちです。現行の検索サービスでは、検索結果をある基準でランク付けしてリスト表示するのが一般的ですが、検索結果が膨大になるにつれ、欲しい情報をリストをたどって探し出すのはとても困難になります。また検索キーワードを追加することで検索結果の量を絞りたいと思っても、適切なキーワードが思い浮かばないというも



どかしさを感じる場面も少なくありません。

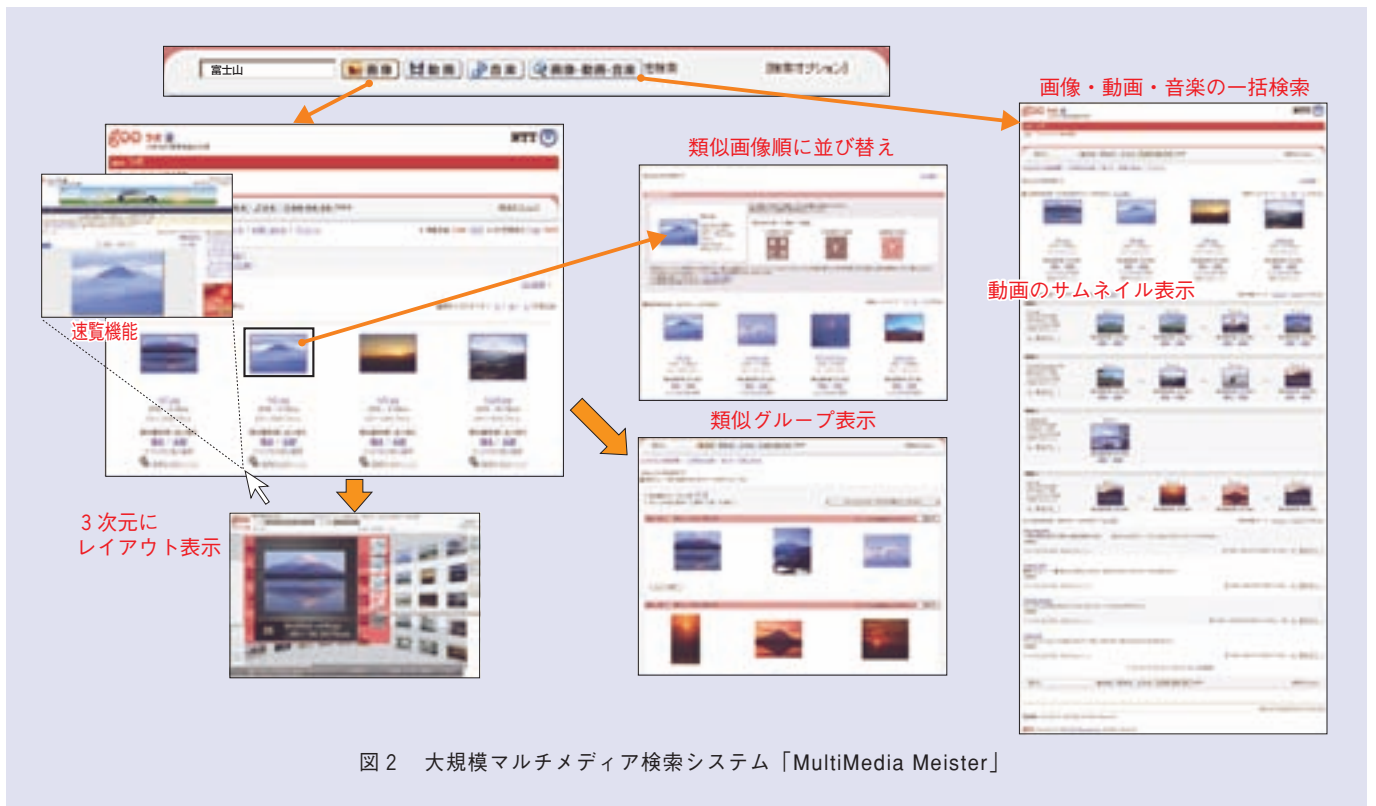
このような現行の検索サービスの不便さを解消する目的で、検索結果に含まれる固有名詞（人物、組織、場所など）をトピックとして動的に抽出し、抽出したトピックを基に検索結果を適切に分類するカテゴリ構造（トピックツリー）を生成・提示する技術を確立しました⁽¹⁾（図1）。またこの技術をニュースサービスに適用したニュース記事分類・検索システム「TopicMaster」を開発しました。トピックツリーを通して膨大な検索結果の概観が可能になるとともに、ツリー中のトピックをクリックするだけで簡単に絞り込みが行えるようになります。

大規模マルチメディア検索システム「MultiMedia Meister」

膨大な検索結果により、欲しいものが探しづらいという不便さは静止画や動画の検索においても同様ですが、特にマルチメディアコンテンツの場合は、マルチメディアならではの曰く言い難いイメージによる検索機能が重要となります。従来のキーワード入力による検索手法だけでは、例えば“富士山”というキーワードで画像を検索したとき、検索結果の中から「こんな景色の画像だけ選びたい」と思っても、「こんな景色」を的確にキーワード表現できないと求める画像をうまく選別できません。NTTサイバーソリューション研

究所では、これまでに開発した世界最速クラスの文書検索エンジン⁽²⁾の性能をさらに向上させることで、マルチメディア向けのWeb検索機能を提供できる検索システム「MultiMedia Meister」を開発しました。

MultiMedia Meisterは、従来のキーワード入力による検索に加え、静止画や動画を構成する画像の特徴（色および形状）に基づき、検索結果を分類する機能や視覚的な類似性で絞り込む機能を提供できます⁽³⁾（図2）。ユーザは、どのような画像が検索されたかを容易に把握できるようになるとともに、検索された画像そのものを指定し「こんな感じ」という条件で直感的に画像を探せるようになります。画



像特徴に基づく画像どうしの類似性解析は一般に時間を要する処理ですが、MultiMedia Meisterでは文書情報と画像特徴データの高速検索を同時に行えるため、数億もの画像データの検索を実用的な速度で実行できます。

マルチファセット・ブログ検索システム「BLOGRANGER」

ブログによって配信される情報には、マスメディアでは取り上げられない情報や最新の話題、商品・サービスの評判といった消費者の生の声が多く含まれており、マーケティング情報としての活用が期待され、ブログ情報をターゲットにしたブログ検索サービスも多数出現しています。しかし、現行サービスの多くは、日付順やアクセス頻度順等で単純にランキングされた検索結果を表示するものであり、ブログのように多種多様な情報が頻繁かつ大量に更新されるものを対象とした場合、ユーザが求めている情報にたどり着くことが難しい場合もありました。

このような現行サービスの課題を解決するために開発したブログ検索システム「BLOGRANGER」は、キーワード入力による検索結果を、有用な4種類のフィルタを通して多角的（マルチファセット）に分類・整理して利用者に提示することで、入力したキーワード以外の視点からも情報を探し出せるようにしたものです。トピック別（TopicMasterと同様の固有名詞による分類）、ブロガー別（ブログ記事の著者単位での分類）、リンク先別（ブログ記事の参照先Webページによ

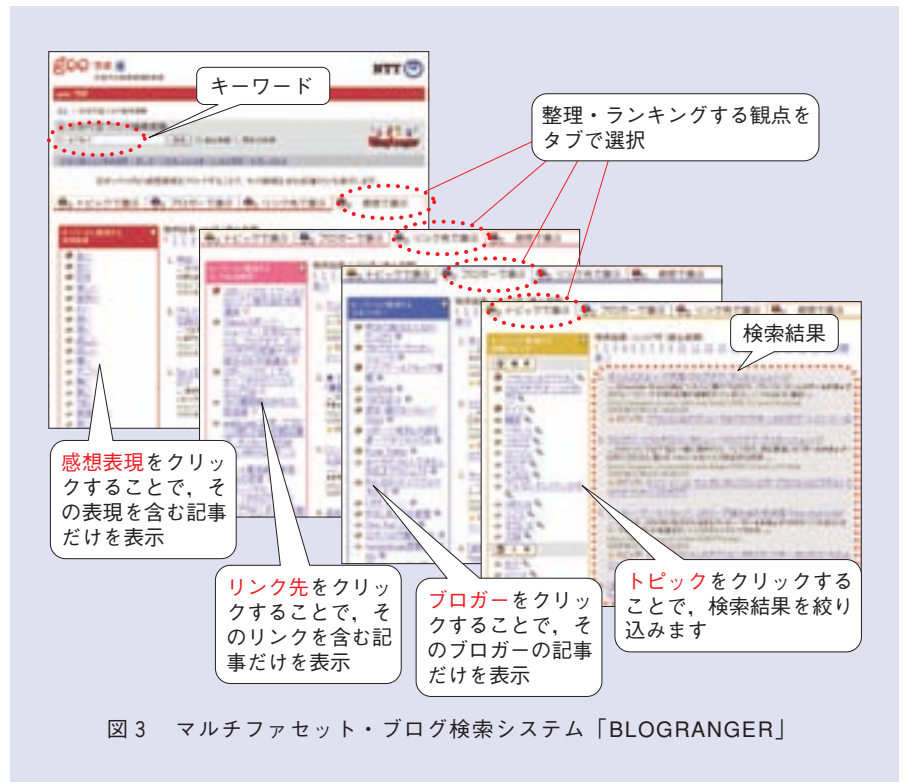


図3 マルチファセット・ブログ検索システム「BLOGRANGER」

る分類)、感想表現別（「面白い」「素晴らしい」といった評価や感想表現による分類）の4つの視点から検索結果を分類・整理して表示することができます（図3）。

4つのフィルタのうち、ブロガー別フィルタには、ブログ単位で同一の著者（ブロガー）が記事を編集しているというブログの特徴が生かされています。ブロガーが多くの人からリンクされるような良い記事を書いているかどうか、またそのブロガーが良い記事をクリックしているかどうかを、ブログ記事の参照関係からそれぞれ数値化し、それを基にブログ記事を注目度順にランキングします⁽⁴⁾。このランキングに基づき、検索結果に含まれる記事を書いたブロガー一覧を求めることで、膨大な

検索結果の中から信頼度の高いブロガーが書いた記事を容易に検索できるようにしました。

日本語概念検索システム「関心事アンテナ」

これまで紹介した事例は、検索サービスによる検索結果が膨大になってしまうことを問題視し、検索結果の絞り込みを支援するタイプのナビゲーションを指向した技術でした。一方、現行の検索サービスにはこれと相反する問題もあります。検索結果は、あくまでも入力したキーワードを含む情報であるため、初めに入力するキーワードが不適切であると、思いのほか検索結果が少なくなってしまうという場面も実は珍しくありません。つまり、ユーザ

の入力したキーワードからユーザの検索意図を汲んで検索対象を広げるようなナビゲーション技術も、次世代の検索サービスには不可欠となります。

日本語概念検索技術は、このような課題に対処するもので、ユーザの入力したキーワードが含まれるかではなく、キーワードや文章が指し示す概念を判別し、ユーザが入力したキーワードと意味内容が類似した文書を網羅的に収集します。概念検索では単語間の意味内容の類似関係を求めるための辞書が必要であるため、インターネット上の情報のように辞書に登録されていない未知語が次々と書き込まれる情報源の検索では、未知語が検索精度を低下させてしまう問題がありました。今回開発した技術では、収集した文書中から辞書に登録されていない未知語が検出されると、その文書に含まれるすでに辞書に登録されている単語間の意味内容の類似関係から、検出された未知語に関する意味内容の類似関係を自動的に求めて辞書に追加できるようにしました⁽⁵⁾。これにより、これまで概念検索が苦手としていたインターネットの検索サービスにも応用が可能となりました。

NTTサイバーソリューション研究所では、日本語概念検索技術を最新ニュース記事やブログ記事の検索に応用したシステム「関心事アンテナ」を開発しました(図4)。自分の求める情報に関するキーワードや文章を思いつくままアンテナとして登録しておけば、自分の関心事に関する情報をもれなく探し出せます。



図4 日本語概念検索システム「関心事アンテナ」

プッシュ型トピック通知システム「TopicAlert」

ニュースやブログのように頻繁に新鮮な情報が発信されるサービスでは、鮮度の高い情報をタイムリにユーザへ提供できる即時性も情報検索の重要な要件です。このような背景から、自分の知りたい情報が発信されるとそれをメール等によりユーザに通知するアラートサービスが普及しつつあります。

現行のアラートサービスでは、単純な検索条件の設定しかできないため、ブログのように多種多様な情報が頻繁かつ大量に更新されるものを対象とした場合、自分の好みに適合した情報だけを受け取ることが困難です。また設

定した検索条件と新着情報を逐一照合する現在の手法では、即時性の高いサービスを提供するために大規模なシステム環境が必要になるという運用上の課題もあります。

プッシュ型トピック通知システム「TopicAlert」は、効率よい情報照合処理を実現した高速XML^{*1}フィルタリング技術⁽⁶⁾を活用することで、従来のシステム環境のまま、詳細な検索条件設定による情報の選別を即時に実行可能としました(図5)。高速XMLフィルタリング技術では、ユーザが通

*1 XML: 文書やデータの意味や構造を記述するためのマークアップ言語の1つで、「タグ」と呼ばれる特定の文字列で元の文章に構造を埋め込んでいく言語の規格。

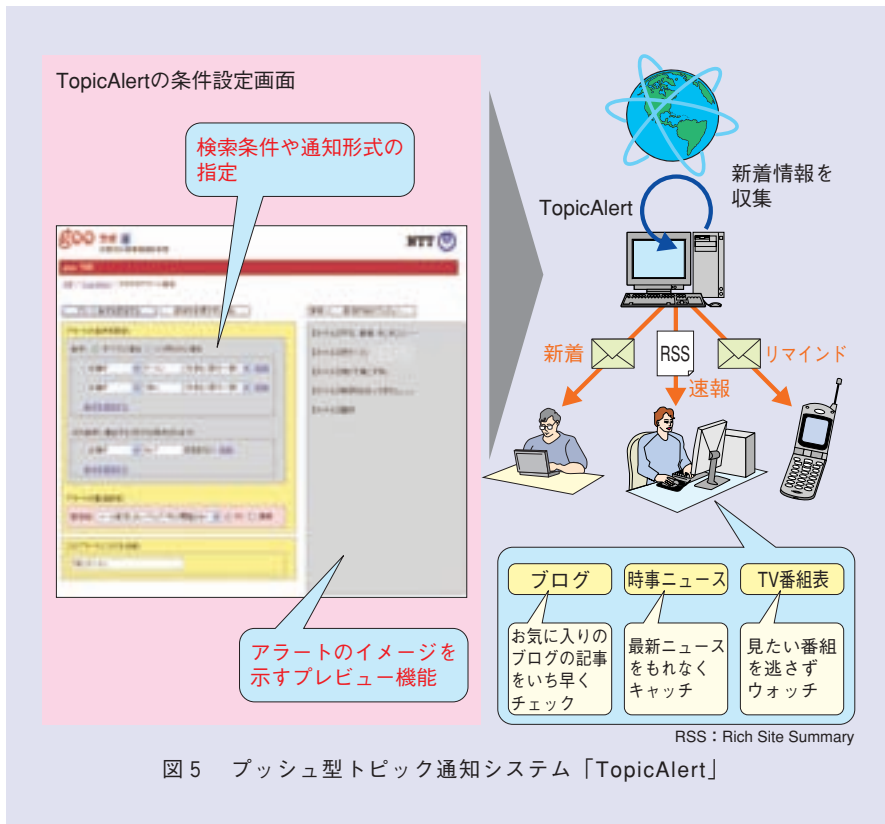


図5 プッシュ型トピック通知システム「TopicAlert」

知を希望するトピックの検索条件をXPath*²形式に変換し、構造化します。その際、別のユーザが入力したキーワードやジャンルなどの検索条件と比較し、共通部分を統合します。これにより、共通部分については処理を繰り返して実行することなく一括して行えることになるため、効率的な照合処理が可能となります。一方、インターネットから取得した新着情報は、XML形式のデータに変換されます。このXML形式のデータを先頭から順に読み取り、統合した検索条件と順次照合します。これにより、新着情報と関連

のある検索条件の絞り込みができるため、必要最小限の照合を一度実行するだけで、その情報が必要なユーザをすべて特定することが可能です。

今後の展開

ここで紹介した技術は、NTTグループ総合ポータル「goo」上に開設された実験サイト「gooラボ」⁽⁷⁾上での実証実験を通して、技術面・運用面での実用性検証を実施しています。MultiMedia Meisterは、実験フェーズを完了し、「goo画像・動画・音楽検索」サービスの検索エンジンとして商用化されました。今後も安全・安心・便利な検索サービスの提供に向け、次世代の情報検索技術に関

する研究開発を進めていきます。

参考文献

- (1) 戸田・中渡瀬・片岡：“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, データベース, Vol.46, No.SIG13 (TOD27), 2005.
- (2) 竹野・井上：“分散型高速情報収集/全文検索システムInfoBee/Evangelist,” NTT R&D, Vol.52, No.2, pp.78-84, 2004.
- (3) 植松・竹野・片岡：“親ページ分類を用いたWEB画像の実用的な検索方法,” FIT2004, 2004.
- (4) K. Fujimura, T. Inoue, and M. Sugizaki：“The EigenRumor Algorithm for Ranking Blogs,” In Proceedings of the WWW 2005 Workshop on the Weblogging Ecosystem, Aggregation, Analysis and Dynamics, 2005.
- (5) 別所・奥：“未知語の概念ベクトル推定手法,” 第164回自然言語処理研究会, 情報処理学会, 2004.
- (6) 鬼塚・兵藤・内山・西岡・山室：“XMLストリーム処理: XPath 評価から SVG レンダリングまで,” 第51回デジタル・ドキュメント研究会, 情報処理学会, 2005.
- (7) <http://labs.goo.ne.jp/>



(後列左から) 西岡 秀一/ 別所 克人/
植松 幸生
(前列左から) 片岡 良治/ 藤村 考/
戸田 浩之

安心・安全・便利な次世代のポータル検索サービスの実現に向け、引き続き研究開発を進めていきます。

◆問い合わせ先

NTTサイバーソリューション研究所
メディアコンピューティングプロジェクト
TEL 046-859-3394
FAX 046-855-1730
E-mail kataoka.ryoji@lab.ntt.co.jp

*2 XPath: XML文書の中の特定の要素を指し示す記述方法を定めた規格。