3 日本人のための検索技術を目指して
― goo における日本語検索の取り組み―

笹島　繁（NTT レゾナント（株）ポータル事業本部）
sasajima_shigeru@goo.jp

浜野 輝夫（NTT レゾナント（株）ポータル事業本部）
ht@nttr.co.jp

日本人による Web 検索エンジンの利用形態

1997 年から検索を中心とした国産ポータルサイトとしてサービスを提供している goo[☆1] では，現在も「行動支援メディア」として，インターネットユーザの生活行動を支援すべく，常にユーザ満足を優先してサービス開発を行っている．ユーザ行動が顕著に現れる Web 検索において入力されるキーワードやクリックログなどを分析してみると，大多数の日本人が検索で求めているのはやはり「日本語」の情報であり，かつ「日本という地域」に関連する情報であることが分かる．これは，日本のユーザの生活行動が，日本語という言語を使い，日本という地域に密着したものであるからと考えられる．

ここでは，特に，日本語という言語的側面と，日本という地域的な側面から，「日本人向け」を強く意識して対応している goo のさまざまな取り組みについて紹介する．

Web 検索における日本語使用，および日本語ページ閲覧の現状（言語的な特徴）

まず，goo の Web 検索で，実際にはどの程度の日本語依存度があるのかについて，現状のユーザ利用動向に基づいて以下の 3 つの観点から調査した．

■入力される検索キーワードにおける日本語の比率

goo の Web 検索において入力されるキーワードのうち，「英数字，記号のみ」から構成されるキーワード入力は 11.8%である．これ以外のキーワード入力は，日本語のキーワード入力であると仮定すると，全入力キーワード中，日本語のキーワード入力が占める割合は 88.2%になる．すなわち，全入力キーワード中の約 90% が日本語ページを探すものであると考えられる．

■検索対象ページの言語指定比率（日本語ページのみ／すべての言語のページ）

goo における Web 検索では，「日本語ページのみ」を検索の対象とするのか，日本語を含む「すべての言語のページ」を対象とするのかを，ユーザが指定できるようになっている．ただし初期設定では日本語ページのみを検索対象とするようになっている．ユーザが，あえて初期設定を変更して「すべての言語ページ」を対象とした検索を行う比率は，全検索回数中わずか 0.65% でしかない．すなわち，99.35%が日本語ページのみを検索対象としていることになる．

■検索結果の中での日本語ページのクリック率

上記に関連するが，検索結果のどの URL をユーザがクリックしたか，というログ情報からユーザが求めていると思われるページの言語を推定することができる．このログ情報からの日本語サイトのクリック率は 99.5% であり，検索ユーザのほとんどは日本語ページを閲覧していることが分かる．

このように，goo ユーザの Web 検索利用動向を分析すると，ほとんどのユーザが，日本語のキーワードを使用して，日本語のページを探していると結論付けることができる．

地理的な特徴

前節で述べたような言語的特徴以外に大多数の日本のユーザの自明な特徴として，日本という地域に生活しているという点が挙げられる．日本という地域に生活して

☆1 http://www.goo.ne.jp/

いるが故に，必然的に日本という地域に密着した情報を求める頻度が非常に高くなる．gooが約1,500人のユーザを対象として実施したアンケート調査では，ニュース，天気予報，地図，路線検索などの日常生活の役に立つ地域情報収集の頻度がきわめて高いとする人が80%近くを占めた．gooに限らず，ほとんどの日本の総合ポータルサイトでは，これらの地域情報は，日本という地域に関するものだけを提供している．このような日本のユーザによる検索サービスに対する地理的依存性の大きさを考慮すると，いかにして日本の地域に密着した地域情報に関する検索サービスを，量と質の両面から充実させ，日本人の生活パターンに適合したかたちで提供するかが重要な課題となる．

日本語検索キーワード入力における取り組み

ユーザにとって，検索するということは，キーワードを選択することから始まる．「検索キーワード」は事実上検索結果を大きく左右するため，非常に重要である．それにもかかわらず，その選定は現在でも勘や経験に頼ることが多く，特に初級ユーザはあまり意識していないことが多い．このような初級者ユーザが効率よく簡単に所望のWebページを検索できるようにするためには，ユーザが入力する検索キーワードに対して別の適切な候補を提示したり別の表現に変換したりするなどの検索キーワード入力支援が有効である．以下では，gooにおける検索キーワード入力支援の取り組みについて紹介する．

ユーザがキーワードを入力する際の課題

検索キーワード入力において初級者ユーザが直面する具体的な課題として，日本語の表記ゆれ問題，および絞り込みキーワード選定の問題がある．

■日本語の表記ゆれ

通常，検索サービスでは入力されたキーワードと完全に同一のワードを含むページを検索結果として提示するが，日本語には英語とは違った「表記ゆれ」に起因する問題が存在する．「表記ゆれ」には，たとえば同音異義語による人名の誤表記などの「表記の誤り」に起因するものと，送り仮名などのさまざまなパターン等による「表記の違い」に起因するものがあり，検索結果としてユーザが求めるものを必ずしも提示できているとはいえない場合もある．

gooにおける統計データでは，これらは検索回数の上位1,000件のうちの約12%，平均して8回に1回の割合で発生している．

■絞り込み検索のための追加キーワード選定

gooにおいては，Web検索クエリーの約7割は1語のみから構成されている．このように1語で検索して大量の検索結果が表示された場合，絞り込みのためにキーワードを追加して，AND検索を実行しながら絞り込むことも多い．しかし，この絞り込みのために新規に追加キーワードを考えなければならないことが，初級者ユーザにとって非常に難しい作業となっている．複数キーワードによる絞り込みが容易にできれば，ユーザはさらに素早く望む検索結果に到達することができる．

キーワード入力支援の考え方

gooにおいては，前節で述べた「日本語の表記ゆれ」や「絞り込み検索のための追加キーワード選定」といった問題を解決するために，ユーザが入力した検索キーワードに対して，適切な修正処理や追加キーワードの自動推薦提示などのキーワード入力支援を行っている．

「日本語の表記ゆれ」については，ユーザが入力した検索キーワードに対して，まず知識ベースに基づいて，「表記誤り」や「表記違い」の可能性がないかをチェックし，「表記ゆれ」を吸収している．「絞り込み検索のための追加キーワード選択」については，入力されたキーワードに対して，時事性なども考慮して追加入力すべきキーワードを関連語として提示する．具体的には，以下のようになる．

■検索キーワードの「表記の違い」を吸収

送り仮名の違いや長音（のばして発音）の表記の違い，ひらがな表記とカタカナ表記の違いなどの表記ゆれについて，言葉を追加・修正してもほぼ問題のない範囲でダイレクトに自動補正する．

※表記のゆれの例：

「宝クジ」→「宝くじ」，「宝クジ」

「年賀ハガキ」→「年賀ハガキ」，「年賀葉書」，「年賀はがき」

■「表記の誤り」に対して正しい検索キーワードを推薦提示

同音異字で誤った検索キーワードが入力された場合や，正式名称と略称，固有名詞で表記間違いが比較的多いものについて，これら検索キーワードが検索された際に正しい表記を推薦候補として表示する．

※推薦の例：

（誤）「バクダット」→（正）「バグダッド」

■絞り込みのための関連する追加キーワードを推薦提示

入力された検索キーワードについて，「goo」で検索されるキーワードの動向をもとに，関連性が高いキーワ

図-1　gooキーワードアシストβ版による関連語の表示例

ードを検索結果画面で表示する．たとえば映画のタイトルで検索した場合，その映画の続編のタイトル名や，企業名で検索した場合に，その企業の代表的な商品名を表示することで，検索結果の絞り込みを容易にする．

日本語キーワード知識ベースの構築

前節で述べた表記ゆれの吸収や追加キーワードの推薦提示を実現するためには，ユーザが入力した検索キーワードを自動的に修正，追加するための知識ベースを構築する必要がある．そのために，主にユーザが日常入力する検索キーワードから，特に時事性が高くユーザニーズのあるキーワードを即時に選定することが重要である．

ただし，キーワードランキングに表れる検索キーワードの出現頻度は，たとえ上位のものであっても全体の出現頻度に占める割合はそれほど大きくはないので，単純なキーワードランキング上位以外の情報にも目を向ける必要もある．具体的には以下のような手法から適切なものを選定することで構築している．

■キーワードランキングからの抽出

日ごとのキーワードランキングを用いて，(a) 検索回数が多く話題になったキーワード，(b) 検索回数自体は比較的少ないが，一定期間毎日検索されているキーワード（これは一種の定番キーワードとも言える），(c) 一定期間における検索回数の推移から，検索回数が上昇傾向にあるキーワード（すなわち，これから話題になりそうなキーワード）を抽出する．

■新規ワードでの読み仮名の一致（同音異義語の吸収）

新たに出現した検索キーワードに対して読みがなをふり，同一の読みがなのキーワードを推薦語の登録候補として抽出する（「週刊誌」（正）⇔「週間誌」（誤）のような候補を抽出するため）．

■ログ分析による関連語グループの抽出

goo ではユーザの検索履歴を分析し，入力されたキーワードの時間的相関関係などに基づいて，絞り込み検索に有効と考えられる追加キーワードを関連語グループとして抽出している．たとえば，検索に使われる異なる複数のキーワードでも，(a) 1 人の利用者が入力した複数のキーワードのうち，時間間隔が短いもの，(b) 複数の利用者によって入力された異なるキーワードのうち時系列の相関関係があるもの，などについては，異なる視点から同一の Web ページを検索しようとしていると考えて，関連語グループとして抽出している．

図 - 1 に，goo の「キーワードアシストβ版」[☆2] による関連語グループ表示の例を示す．これは，ユーザがキーワードを入力し，スペース（空白）を押すことで，追加単語や，置き換え候補単語が表示される（関連語は 1 日単位で自動的に変動する）．

Web 検索結果の編集

検索サービス自体は万人にとって重要なものであるが，ユーザ層はさまざまに分化しつつある．アンケートをとると，検索結果の量が少ないという意見と，多すぎるという意見が必ず混在しており，上級ユーザは，よりシンプルに速く多くの結果が出ることを望んで，初級ユーザは逆に検索結果が多すぎて選択しにくい，という感覚を持っている．上級ユーザのためにはひたすら検索の基本機能の向上を行うことになるが，初級ユーザのためには，Web 検索だけでは，ユーザの要望を必ずしも確実に満たせないという観点から，検索キーワードに対して，ユーザニーズの高そうな情報を同時に表示することで，Web 検索結果を補完しようと考えている（**図 - 2**）．

ただし，この情報の選定は非常に困難な面もあり，たとえば「旅行」というキーワードにおいても，「旅行に行く場所を決めたい」「旅行先は決まっているが周辺の観光・ホテル情報を知りたい」「旅行後の日記などの情

☆2 http://search.goo.ne.jp/gka/

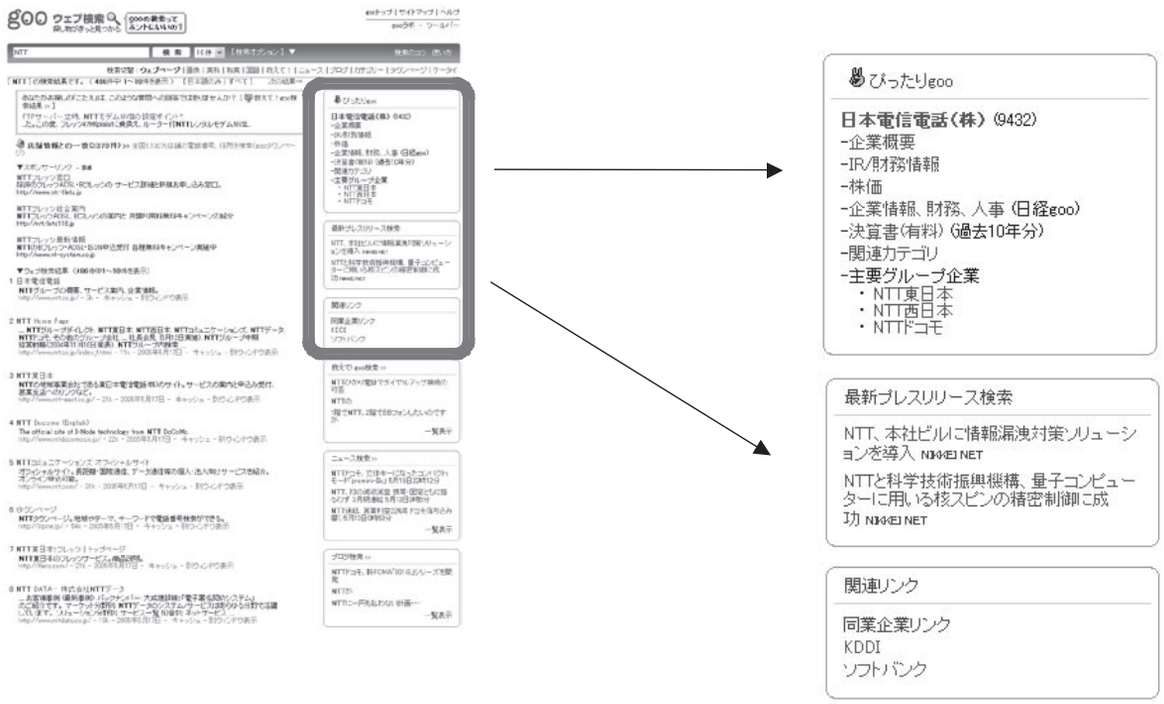

図-2　キーワードに連動したWeb検索結果の編集

報を知りたい」等々，本来知りたい情報は多岐に渡る．ユーザ層の違い，目的の違いを分析して，よりユーザニーズにあった検索結果を追求していく必要がある．

地域検索の取り組み

日本人の生活に密着した“役に立つ”地域検索サービスを実現するためには，日本に関連する地域情報を，(a) インターネット上から抜けがないように網羅的，かつ広範囲に収集し，(b) これらの情報を容易に利用できるように構造化し，(c) 日本のユーザがこれらの情報を最終的に利用する状況に適合した形態で，使いやすく分かりやすく提示することが必要である．ここでは，一例として，レストランなどの地域の店舗情報を検索サービスとして提供するケースを取り上げて，gooにおける地域情報検索サービスへの取り組みを紹介する（**図-3** 参照）．

地域店舗情報の収集

gooでは，複数の大型店舗情報サイト数社と提携して情報の提供を受けている．これらの大型店舗情報サイトでは，一般的に1サイトあたり数千店の店舗情報が掲載されている．しかし，これらの店舗情報サイトだけではインターネット上に提供されているすべての店舗情報を網羅することはできない．特に，大型店舗情報サイトに掲載されていない地方の中小の店舗の被覆率は30～50%程度と想定され，日本人の生活に役立つ地域情報検索サービスを実現するためには，とても十分であるとは言いがたい．このため地方の中小店舗情報は，インターネット上からクローラなどで追加的に収集する必要がある．gooでは，スタッフが収集した地方の店舗情報サイトを基礎として，jpドメイン配下の中小店舗情報を定期的に収集している．このようなサイトは約300サイトあるが，いわゆるDeep Webに対する収集は行っていないため，実際には約150サイト，店舗数にして約3万件のデータを自動的に収集している．gooでは，このクロールした中小店舗情報と，前述した提携している大型店舗情報サイトに掲載されている店舗情報とを併せて，合計約6万店舗のレストラン情報を検索することが可能である．

地域店舗情報の構造化

クローラによって自動的に収集した店舗情報については，HTMLの構文木を解析し，店舗名とこれに対応する住所，電話番号，店舗のジャンル，店舗写真，店舗の紹介文などのデータ項目を，知識ベースに基づいて自動的に抽出している．ただし，現状では完全には自動的に抽出できない店舗情報もあるので，最終的には抽出データの正当性を標本データから人手で確認し，知識ベースの修正を行っている．このようにして店舗情報を構造化

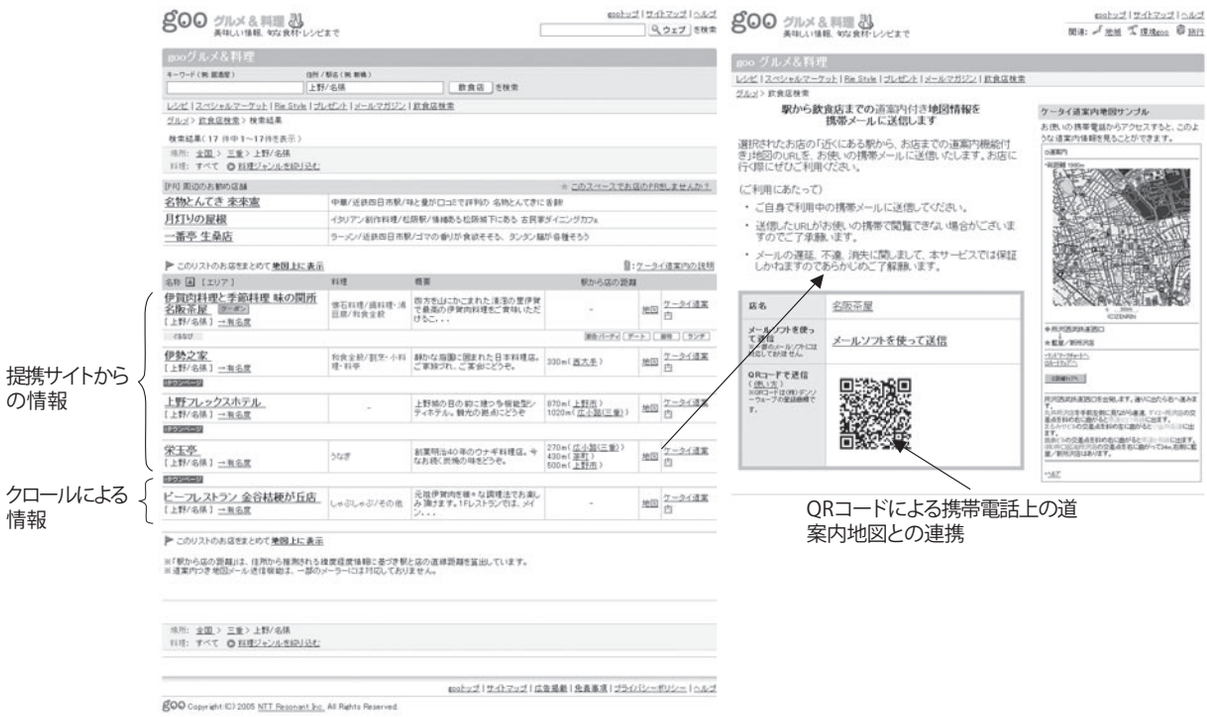

図-3　地域情報検索結果

することで，もともとはさまざまな店舗情報サイトにバラバラな形式で掲載されていた店舗情報群を，あたかも単一のデータベースを扱うように横通しに検索することが可能となる．たとえば，地域として「東京」を，ジャンルとして「ラーメン」を選択すると，複数の異なる店舗情報サイトに掲載されていた店舗情報群から該当するものを検索結果として提示することができる．また自動抽出された住所情報からその店舗の緯度経度を算出して，最寄りの駅名とその駅までの距離といった付加的な情報も検索結果に併せて表示することが可能となる．

ユーザインタフェース（スクロール地図，携帯電話連携）

レストランなどの店舗情報等を検索する場合，ユーザの最終目的は，自分がその店舗に実際に行って食事や買い物をすることである．このため，レストランの場所，すなわち地図情報の検索が多くの場合必須となる．gooでは，この地図情報の検索インタフェースとして，マウスのドラッグ操作によって，ユーザが地図上の見たい地点にスムーズに視点を移動できるスクロール地図インタフェースを試行的に提供している（goo ラボ「エリア情報検索実験」，**図-4** 参照）．ユーザはカーナビのように，通りなどに沿って視点を移動していくことが可能となるため，直感的に自分が行きたい店舗を探し出すことが可能である．このようにして，店舗情報等の地域情報と地図情報を取得すると，ユーザはこれらの情報に基づいて実際に店舗に出かけることになる．近年の日本人の行動様式における特徴として，カメラ付携帯電話の帯同が広く普及している点が挙げられる．そこで，gooでは，PC の画面上に表示された店舗情報の検索結果の近隣に，その画面の URL に該当する 2 次元バーコード（QR コード）を提示し，ユーザがこれをカメラ付携帯電話で撮影することで，当該検索結果の URL を取り込むことを可能としている．これによって，ユーザは外出先においてでも，カメラ付携帯電話から当該検索結果（すなわち，目当ての店舗情報や地図情報）を容易に閲覧することができる．

将来に向けた取り組み（goo ラボ関連）

以上，goo における日本人に適した検索技術開発の取り組みについて述べた．gooでは，さらに新しい日本人向け検索サービスを，goo ラボ☆3 上で公開している．

■日本語自然文検索実験「Web Answers」

Web 検索において，キーワードを入力する代わりに

☆3　http://labs.goo.ne.jp/
goo ラボは，NTT 研究所などで開発された次世代検索技術をいち早く公開し，その可能性を一般ユーザに体感していただくための新技術の実験場．

図-4　スクロール地図上への表示

自分が知りたい“答え”を尋ねる質問文を直接入力すると，その質問文に対応する回答を直接出力する技術．たとえば「2008年のオリンピックの開催地はどこ？」という質問文を入力すると，質問文を分析して自動的に適切なキーワード群を生成してWeb検索を実行し，取得したWeb検索結果の内容をさらに分析して，「北京」という答えを出力する．このシステムには人名・地名などを検索結果から高速に抽出する独自の日本語解析技術が使われている．（2005年5月9日終了）

■ニュース記事分類・検索実験「Topic Master」

キーワード入力でWeb上からニュース記事を検索した結果を，人物名，組織名，場所といった適切なトピック群に自動的に分類することで，ユーザが大量の検索結果から所望の記事だけを簡単に探し出すことができる．

■ Webページパーソナライズ高度化実験「パーソナルサマリ」

ユーザが，ブラウザ上でさまざまなWebコンテンツを切り張りすることで，自分好みのWebコンテンツを自由に生成する技術．gooラボ上では，Webコンテンツとして RSSに特化したバージョンを公開している．

これらの新しい検索技術については，gooラボの場などを通じてユーザの意見を取り入れながら，さらにブラッシュアップを図り，実際のgoo商用サービスとして展開していく予定である．

参考文献

1）大久保，杉崎ほか：WWW検索ログに基づく情報ニーズの抽出，情報処理学会論文誌，Vol.39, No.7, pp.2250-2258 (July 1998).
2）Toda, H. and Kataoka, R.: A Clustering Method for News Articles Retrieval System, Proceedings of WWW'05, pp.988-989 (2005).
3）Saito, K. and Nagata, M.: Multi-Language Named-Entity Recognition System based on HMM, Proceedings of ACL-2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, pp.41-48.

（平成17年6月28日受付）