

# タッチ操作によるモバイル機器向け情報検索支援

Information Retrieval Support with Touch Operation for Mobile Terminals

岡本 昌之

渡辺 奈夕子

大村 寿美

■OKAMOTO Masayuki

■WATANABE Nayuko

■OMURA Sumi

近年、スマートフォンやスレート端末<sup>(注1)</sup>など、キーボードレスのモバイル機器を用いてWebの情報検索を行う機会が増えている。しかし、タッチパネルなどを用いた検索文の入力が煩わしいという課題がある。

東芝は、閲覧中のWebページから検索される可能性の高い単語を自動抽出し、関連情報検索をタッチ操作だけで行えるようにする情報検索支援機能“NX!™サーチ”を開発した。NX!サーチは、Webページの本文領域を推定して本文テキストを抽出する本文抽出技術、本文テキストから検索される可能性の高いキーワードを推測するキーワード抽出技術、及びタッチ操作による検索操作を実現する関連情報検索技術によって実現される。キーワード抽出についての性能評価とともに、実際の利用による被験者調査を行った結果、この機能が情報検索支援に有効であることがわかった。

Opportunities to search for information on websites using mobile terminals without a keyboard, such as smartphones and slate terminals, have been increasing in recent years.

With this trend as a background, Toshiba has developed the NX!™ Search information retrieval support technology that can automatically extract query keywords with a high probability of retrieval from the webpage currently being browsed, allowing users to search for related information using only touch operation. The features of NX!Search include a body-text extraction technology, a keyword extraction technology that estimates query terms with a high probability of retrieval from the body text, and a related-information search technology requiring only touch operation. The results of performance evaluations and experiments by trial subjects have confirmed the effectiveness of NX!Search for the support of information retrieval activities.

## 1 まえがき

近年、パソコン(PC)だけでなくスマートフォンやスレート端末などのモバイル機器を用いたWebページ閲覧の機会が増えるとともに、情報検索を行う機会も増えてきている。しかし、キーボードによって文字入力ができるPCと異なり、モバイル機器の多くは数少ないボタンやタッチパネルを用いるため、検索文を入力する煩わしさは大きな課題である。実際、携帯電話を用いた検索はPCの場合に比べて少ないことが報告されている<sup>(1)</sup>。

東芝は、モバイル機器での検索を簡単にすることを目的に、現在閲覧中のWebページからユーザーが検索する可能性の高い単語を推測し提示する機能“NX!サーチ”を開発した。この機能によって、閲覧中にアイコンを押すだけで、そのとき検索される可能性の高い単語の一覧を表示し、検索したい単語をタッチするだけで関連情報を検索できる。

任意の状況でユーザーが検索したい単語を推測するのは困難だが、現在閲覧中のページについて関連情報を調べる際には、検索文はそのページについての単語であることが多い。したがって、閲覧中のWebページの文章から検索文候補とな

(注1) 本体前面にタッチパネル式ディスプレイを搭載したスマートフォンよりもディスプレイの大きなモバイル機器。

る単語を抽出することで、関連情報検索時に操作に掛かるコストを下げる可以考虑。NX!サーチの利用により、数回のタッチ操作だけで関連情報検索を行うことができるようになる。この機能は、2010年6月にスマートフォン向けβ版がリリースされた<sup>(2)</sup>。

NX!サーチには、当社のコア技術である本文抽出技術、本文テキストから検索される可能性の高いキーワードを推測するキーワード抽出技術、及びタッチ操作による検索操作を実現する関連情報検索技術が利用されている。ここでは、機能の概要と利用される技術、そして被験者評価の結果から得られた知見について述べる。

## 2 NX!サーチの概要

NX!サーチは、モバイル環境での検索文入力の手間を削減しつつWeb閲覧からの関連情報検索ニーズを満足する情報検索支援機能である。利用方法の概要は、以下のとおりである。

まず、ユーザーはWebブラウザを起動してWebページを閲覧する。閲覧中に関連情報を知りたくなった場合、画面上部に表示される通知アイコンをタッチする。このアイコンは、NX!サーチが利用可能な状態の場合、青色で表示されている



(図 1(a))。すると、現在の Web ページ内の単語のうち、検索に利用されやすいと NX!サーチが判断したキーワードの一覧が表示される (図 1(b))。

キーワードは検索されやすさの順に並べられている。下位の候補を表示する場合は、画面全体を横にスライドすることで表示される。NX!サーチでは、最大 16 個のキーワードが選択できる。また、キーワードのボタンは、押す領域によって通常のキーワード検索、ニュース検索、及びトリビア (Wikipedia) での雑学的知識) 検索が選択できる。例えば、キーワード“田中久重”について詳細を知りたい場合、“田中久重”と書かれたボタンの“トリビア”部分を押しさえよ。もし、候補の中に適当なキーワードを見つけられなかった場合や、関連情報以外の情報を検索する場合は、画面下部の検索ボックスから直接キーワードを入力する方式を利用することもできる。

いずれかのキーワードを押すことで、検索が行われ検索結果が表示される (図 1(c))。検索結果は画面全体を縦方向にスクロールすることで下位の結果も表示される。見たい検索結果があれば、そのスニペット (Web ページの説明又は引用) をタッチすることで次の Web ページが表示される。

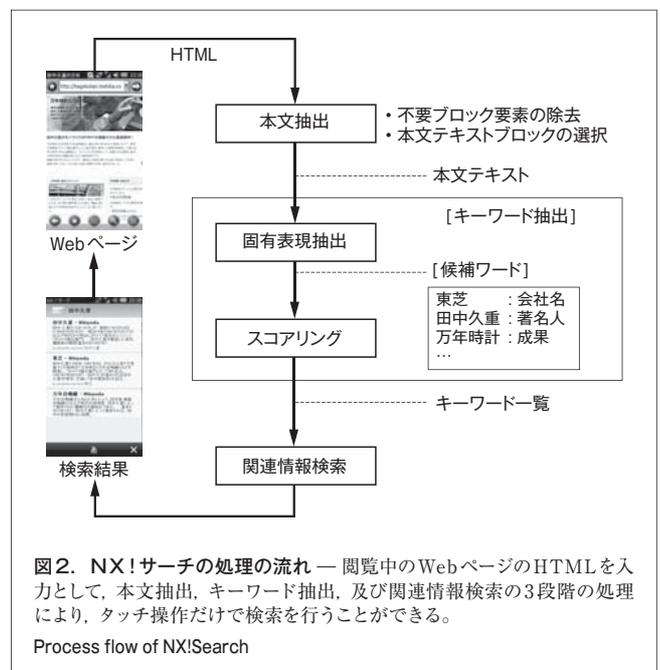
このように、ユーザーは数回のタッチ操作だけで関連情報を検索できる。

### 3 NX!サーチの処理の流れ

NX!サーチにおける処理の概要を図 2 に示す。NX!サーチでは、閲覧中の Web ページの HTML (Hyper Text Markup Language) を入力として用い、本文抽出、キーワード抽出、及び関連情報検索の 3 段階の処理が行われる。詳細を以下に述べる。

#### 3.1 本文の抽出

本文抽出では、HTML 文書を解析して本文領域を推定し、



内部のテキストを抽出する。ここで本文領域とは、ニュース記事などの Web ページのうち、広告部分や関連記事リンクなどを除いた記事コンテンツ部分である。本文抽出処理は、次の手順で行われる。

まず、HTML 文書を、大見出しを表す <h1>、表を表す <table> などのブロック要素ごとに分割し、ブロックごとにテキストの文字数、ハイパーリンク<sup>(注2)</sup>を表す <a> タグで囲まれたアンカーテキスト長<sup>(注3)</sup>、句読点数、及びストップワード<sup>(注4)</sup>に基づくスコア (得点) を算出する。次に、テキストの文字数がしきい値より少ない場合や、あらかじめ指定されたストップワードを含むなど、

(注2) 文章中に埋め込まれた文書や画像の位置情報。  
(注3) HTML 文書の中で、ハイパーリンクが設定された文書列。  
(注4) 一般的なことばなので、処理対象外とする単語。

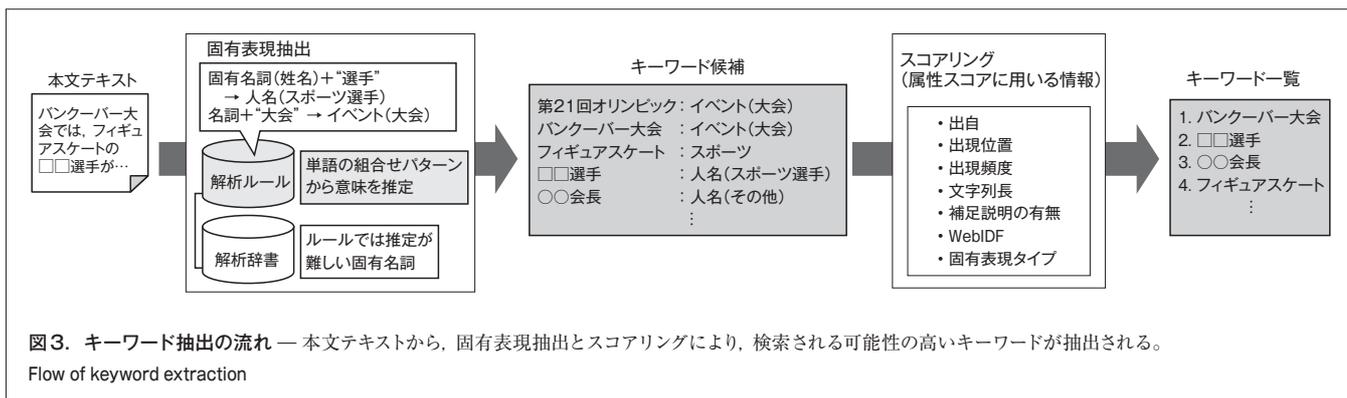


図3. キーワード抽出の流れ — 本文テキストから、固有表現抽出とスコアリングにより、検索される可能性の高いキーワードが抽出される。

Flow of keyword extraction

本文らしくない特徴を持つブロックを除去する。その後、残ったブロックを連結し、スコアの高いブロックを残して本文テキストとみなす。なお、長文ページなどで処理時間が掛かるのを防ぐため、一定の長さ以上のテキストでは途中で処理を打ち切る。

### 3.2 キーワードの抽出

3.1節で得られた本文テキストに対して、キーワード抽出を行う(図3)。

まず、固有表現抽出を適用し、固有表現クラスを付与する。固有表現とは、人物名、企業名などの固有名詞や、日時、数量などの表層表現であり、それぞれの意味を表す属性のラベルが固有表現クラスである。これらは辞書と解析ルールによって付与される。例えば、辞書に“□□(人名)”, “選手(スポーツ選手)”が登録されている場合、登録されていない語“□□選手”に対して、ルール“人名+スポーツ選手→スポーツ選手”を適用することで、固有表現クラス“スポーツ選手”を付与する。辞書には、解析ルールへの入力となる基本的な単語や、ルールによる解析が困難な単語が登録される。

このように、固有表現クラスが付与された単語の一覧がキーワード候補となる。なお、この処理は、テレビの視聴ができるPCで用いられている視聴シーン関連情報検索アプリケーションソフトウェア<sup>(3)</sup>と同じものである。

次に、スコアリング処理が行われる。得られたキーワード候補は数が多い場合もあり、そのまま出力すると選択に手間が掛かる。この処理では、それぞれのキーワード候補に対してスコアを付与する。スコアは、見出しに含まれるか本文に含まれるかなどの出自、出現位置、文書内の出現頻度、文字列長、単語直後の補足説明の有無、WebIDF (Web Inverse Document Frequency)、及び意味属性ごとのスコアの線形和として算出される。ここで、WebIDFとは、Web上でどの程度頻出する単語かをあらかじめ計算したスコアである。線形和の重みは、ユーザーがどのようなキーワードを検索するかを事前調査した結果に基づき決定した。

最終的に、スコア上位のものがユーザーに提示されるキーワードとなる。キーワード抽出が完了すると、画面に通知アイコンが表示される。NX!サーチでは、ユーザーが通知アイコン

をタッチすると、1画面当たり8個、最大で16個までのキーワードが表示される。

### 3.3 関連情報の検索

提示されたキーワードのうち1個をユーザーがタッチすると、そのキーワードを使って検索が行われる。ユーザーがタッチする場所に応じて検索ジャンルが指定される。指定された検索ジャンルに応じて、タッチされたキーワードに加え、特定の単語とのAND検索<sup>(注5)</sup>、あるいは検索対象のWebサイト指定に関する条件が追加され検索が実行される。

## 4 キーワード抽出方式の評価

この機能の有効性を検証するため、性能評価及び被験者の自由利用による精度評価を行った。なお、この章で述べる評価では、当社のスマートフォン T-01A に搭載されたWebブラウザ (Internet Explorer<sup>®(注6)</sup> Mobile) のアドオン<sup>(注7)</sup>として実装したものを用いた。T-01Aは、OS (基本ソフトウェア) としてWindows Mobile<sup>®(注8)</sup> 6.5を搭載し、CPUの動作周波数は1 GHz、SDRAM (Synchronous DRAM) の容量は256 Mバイト、表示部は4.1型シングルタッチ ディスプレイである。

### 4.1 キーワード抽出の処理性能

ニュースやブログなど、複数分野の八つのWebサイトに対して、キーワード抽出処理に掛かった時間を4回ずつ測定した結果の平均を表1に示す。

表1より、ユーザーがWebページを読む時間に比べてキーワード抽出処理が十分速く動作することがわかる。この処理はWebページを読み終わると同時にバックグラウンドで行われるため、ユーザーがページを読み終わり関連情報を検索したいと思った時点では、既にキーワードの抽出が終了している。

### 4.2 被験者実験を通じたキーワード抽出の精度評価

実際のユーザーが自由利用した際の印象を調査するため

(注5) キーワードで検索する手法の一つで、二つ以上の単語を指定し、それら全てを含む検索方法。

(注6)、(注8) Internet Explorer, Windows Mobileは、米国Microsoft Corporationの米国及びその他の国における商標又は登録商標。

(注7) ソフトウェアに追加される拡張機能。

表1. クエリワード抽出時間

Term extraction times on various webpages

ページID	HTMLサイズ (kバイト)	候補ワード数 (個)	抽出時間 (s)
1	50.9	44	1.580
2	184.0	100	4.759
3	218.1	85	2.581
4	137.7	13	2.021
5	39.0	49	2.131
6	62.0	46	2.630
7	40.0	48	1.093
8	49.0	47	1.736

ID : Identification

に、14名の被験者による利用評価を実施した。評価の進め方は、以下のとおりである。

各被験者は約1時間、自由にWebブラウジング (Webの閲覧)を行う。検索したいタイミングで通知アイコンを押し、NX!検索機能呼び出す。検索に利用したいキーワードが表示されている場合は選択して検索し、表示されていない場合は画面下部の検索用テキストボックスを用いて検索する。

自由利用の後、NX!検索を利用したときに提示されたキーワード一覧を確認し、それぞれの単語について3: 検索したい、2: キーワード一覧に出てよい、1: キーワード一覧に不要、の3段階で評価値を付ける。この評価では、評価指標として検索に利用したいキーワードが1個以上提示される割合である、提示ワード利用率を用いた。

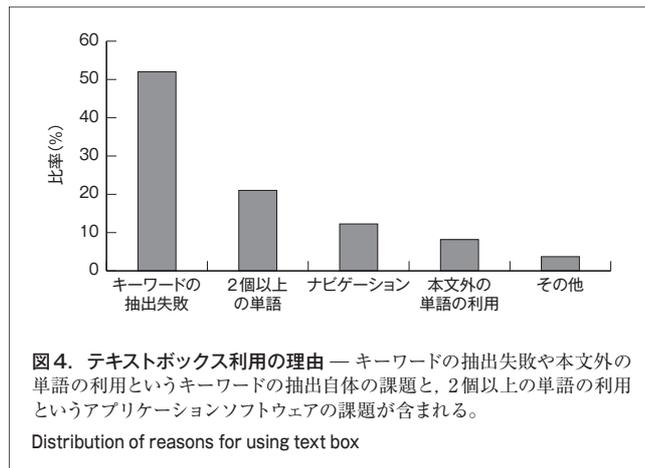
評価結果を表2に示す。実際に検索を行ったうち、NX!検索機能で提示されたキーワードを利用した割合 (A/(A+A'))は85.7%であり、その分だけ検索操作に掛かる手間を減らすことができたと言える。

評価後、被験者に検索用テキストボックスを用いて検索した理由を確認した。図4に示す分布から、キーワードの抽出失敗や、本文外の単語の利用のようなキーワード抽出自体の課題と言えるものが60%存在している。これらはキーワード抽出方式の改良によって解決すべきものである。また、2個以上の単語利用は現在のUI (User Interface)仕様ではサポートされていないため、アプリケーションソフトウェアの課題として今後の対応が求められる。

表2. 被験者評価の結果

Results of evaluation tests by trial subjects

項目	平均
閲覧ページ数	40回
提示キーワード選択回数 (A)	10.2回
キーワード入力回数 (A')	1.7回
提示ワード利用率 (A/(A+A'))	85.7%



## 5 あとがき

ここでは、数回のタッチ操作だけで閲覧中のWebページの関連情報検索を行うNX!検索機能の概要と、利用される技術について述べた。また、性能評価及び検索支援の効果検証に関する被験者評価を実施した結果、NX!検索が情報検索支援に有用であることを確認した。

今後は、キーワード抽出方式の精度向上に取り組むとともに、この技術の他の端末への適用や様々なユーザーに適した方式について検討を進めていく。

## 文献

- (1) Kamvar, M. et al. Computers and iPhones and mobile phones, oh my!: a logs-based comparison of search users on different devices. Proc. SIGIR-08. 2008, p.801 - 810.
- (2) 東芝 研究開発センター. "当社スマートフォン向け情報検索アプリケーション (NX!検索 (β版))を公開". 東芝ホームページ. <http://www.toshiba.co.jp/rdc/rd/detail\_j/1007\_02.htm>. (参照2011-03-18).
- (3) 山崎智弘 他. 番組視聴中に気になったシーンを簡単に検索できるキーワード抽出技術. 東芝レビュー. 63, 4, 2008, p.26 - 29.



岡本 昌之 OKAMOTO Masayuki, Ph.D.

研究開発センター ヒューマンセントリックラボラトリー 研究主務、博士 (情報学)。コンテキストウェア技術及び情報抽出技術の研究・開発に従事。情報処理学会, 人工知能学会, ACM会員。Humancentric Lab.



渡辺 奈夕子 WATANABE Nayuko

研究開発センター ヒューマンセントリックラボラトリー。ユーザーインタフェース技術及び情報抽出技術の研究・開発に従事。情報処理学会会員。Humancentric Lab.



大村 寿美 OMURA Sumi

デジタルプロダクツ&サービス社 設計開発センター デジタルプロダクツ&サービス設計第3部グループ長。モバイル分野におけるユーザー支援アプリケーションの開発に従事。情報処理学会会員。Design & Development Center