DEIM Forum 2011 F6-5

閲覧 Web ページからの第 1 クエリワード抽出に基づく組み込み機器向け検索支援

岡本　昌之† 　菊池　匡晃†† 　渡辺奈夕子† 　飯田　貴之†† 　佐々木健太†
服部　正典†

† (株) 東芝 研究開発センター　〒 212–8582 川崎市幸区小向東芝町 1
†† (株) 東芝 ビジュアルプロダクツ社 コアテクノロジーセンター　〒 198–8710 東京都青梅市末広町 2-9
E-mail: †{masayuki4.okamoto,nayuko.watanabe,kenta.sasaki,masanori.hattori}@toshiba.co.jp,
††{masaaki11.kikuchi,takayuki1.iida}@toshiba.co.jp

あらまし　スマートフォンや TV のような組み込み機器における Web 検索は日常的になりつつあるが，クエリワード入力のコストは依然として課題である．クエリ予測は有望なアプローチであるが，クエリ拡張と比べ，最初のクエリワードの推薦に関する報告は少ない．本稿では，組み込み機器において閲覧中の Web ページに関連する情報をタッチ操作のような簡単な操作のみで検索可能とするための，第 1 クエリワード抽出方式を提案する．提案方式は，本文抽出，候補ワード抽出，スコアリングから構成される．スマートフォン上で動作する検索支援アプリケーションを実装し，高速に動作することを確認した．また，299URL を対象としたクローズド評価および 14 名によるオープン評価を通じ，実用的な精度でクエリワードが抽出されることを確認した．

キーワード　Web 検索，クエリワード抽出，固有表現抽出

First Query Term Extraction from Current Browsed Webpage for Embedded Device

Masayuki OKAMOTO†, Masaaki KIKUCHI††, Nayuko WATANABE†, Takayuki IIDA††, Kenta SASAKI†, and Masanori HATTORI†

† Corporate R&D Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212–8582 Japan
†† Core Technology Center, Visual Products Company, Toshiba Corporation
2–9 Suehiro-cho, Ome, Tokyo, 198–8710 Japan
E-mail: †{masayuki4.okamoto,nayuko.watanabe,kenta.sasaki,masanori.hattori}@toshiba.co.jp,
††{masaaki11.kikuchi,takayuki1.iida}@toshiba.co.jp

Abstract　Inputting query terms on a embedded device such as a smartphone or a TV is not easy though web search activity on such kind of devices is becoming popular. Query prediction is a promising approach. In the literature, however, little attention has been paid to the first query term, though there are many reports on query expansion or second query recommendation. In this report, we propose an information retrieval method which enables users to search the web with an easier operation by automatically extracts query-term candidates. The proposed method consists of body-text extraction, candidate query term extraction, and scoring processes. We also implemented a search support application that works for webpages at a practical speed on a smartphone. According to our closed evaluation with 299 webpages and open evaluation with 14 users, our method achieved practical quality.

Key words　Web search, first query term extraction, named entity recognition

1. はじめに

本稿の目的は，閲覧中の Web ページに関連する情報の検索を容易にするための，クライアントサイドで第 1 クエリワードを自動抽出する手法の提案である．

PC や携帯電話・スマートフォン，TV などの機器を用いた情報検索の機会はますます増加しているが，文字入力のコストは依然として課題である．特に，携帯電話・スマートフォン，TV などの組み込み機器においては，文字入力のコストは依然として課題である．例えば，携帯電話を用いた検索は PC と比べて少ないことが報告されている [8]．また，最近では TV などの情報機器においても Web ブラウザが搭載されたものが増えており，モバイル端末の場合と同様の課題があると考えられる．このコストを減らすためには，できるだけ少ない操作回数・操作時間で検索できる方式の実現が望まれる．

クエリ入力を減らすアプローチとしては，バックグラウンドで検索を行うクエリフリー検索 [5] や 2 番目以降の単語を提示するクエリ推薦 [6] [14] があるが，閲覧中の Web ページに応じた **1 番目の**クエリワードを提示する試みは殆ど行われていない．

あらゆるシーンでユーザが検索する可能性のある単語を推測するのは困難であるが，Web 閲覧を起点とした情報検索では，現在閲覧中のページについて関連する情報を調べる傾向が強いため，閲覧ページの関連情報検索に限定してもある程度ユーザニーズを満たせると考えられる．例えば，ニュース関連のページを読んだ直後の Web 検索クエリはそのページについての単語であることが多い [17]．これは PC 上の Web 閲覧を対象とした調査結果ではあるが，スマートフォンなどを用いた Web 閲覧においても，Web ブラウザの性能が向上にしたがい PC と同様の傾向になるものと考えられる．したがって，閲覧中の Web ページ文章から，検索クエリの候補となる単語を抽出することで，多くの場合検索操作にかかるコストを下げることが可能である．

また，検索支援におけるもう 1 つの課題は通信コストである．無線も含む通信帯域が増加しているとはいえ，ユーザが意図しないバックグラウンドでの検索は通信コストがかかる上，検索エンジン提供者にとってもユーザが意図しない検索セッションの増加は望ましくないと考えられる．

本稿では，閲覧中の Web ページに関連する情報の検索を支援するための，クライアントサイドでの第 1 クエリワード抽出方式を提案する．本方式では，HTML から本文テキストを抽出し，その後本文テキストからクエリワードの候補を抽出し，付与されたスコア順に出力する．この方式は，特に画面の解像度が小さい端末を用いてある程度まとまった量のテキストを含むニュースサイトやブログ記事を読む場合に効果があると考えられる．後で検索される可能性がある単語でも，読み終えた時点では画面上に表示されていないことがあるが，アプリケーションが提示することで，画面をスクロールして選択，といった操作の手間を減らすことが可能となる．

抽出すべき単語に関しては，様々な Web ページを対象としてユーザが検索に利用したいと考える単語を調査した．調査で得られた単語は，正解データとして登録するとともに，クローズド評価の対象とした．

また，実際に本方式をスマートフォン上に実装し，実用的な性能で動作することを確認するとともに，クローズド評価，オープン評価を通じ実用的な精度が得られることも検証する．

また，精度向上作業においては，様々な Web ページを対象としてクエリワード抽出方式を改良しつつ，副作用が発生しないことを確認する必要がある．そこで，評価システムを試作し，半自動で評価を行うことにより，後戻りの少ない改良を行った．この精度向上方式についても紹介する．

本稿の構成は以下の通りである．まず，**??** では関連研究について述べる．次に，提案するクエリワード抽出方式について **3.** で述べる．**4.** では，どのような単語が抽出対象となるかの調査結果を報告する．**5.** では，スマートフォンにおける実装例と性能評価結果について述べる．**6.** では，精度向上施策として評価ツールを利用した改良サイクルを紹介する．そして，**7.** において，最終的に得られたクローズド評価，オープン評価の結果を報告する．

2. 関連研究

これまでにクエリワードの入力コストを下げるために閲覧コンテクストを利用する手法が多数研究されてきた．主な方式として，クエリフリー検索とクエリ推薦がある．

クエリフリー検索方式では，閲覧内容に基づきクエリを自動抽出し，検索結果が直接ユーザに提示される．Letizia [12] はユーザの閲覧行動に基づき自動的にリンクを付与する．FIXIT [5] は故障時の症状に関するレポートをキーワードに対応付けることで，コピー機のマニュアルから修理情報を検索する．Query-free news search system [7] は，放送中の映像のクローズドキャプションからクエリタームを抽出する．WebTelop [13] や映像ブックマーク検索 [15] は現在の TV 映像シー ンから補完情報や関連する Web ページを検索する．閲覧 Web ページの関連情報の検索では，補完情報を検索する WeBrowSearch [22] が提案されている．

クエリ推薦は，検索エンジン上のクエリログやパーソナルなクエリログ，広告クリック等の情報を利用して 2 番目以降のクエリワードを推薦するものである．アプローチとしては，ユーザのクエリログからクエリ拡張用の単語を推薦するもの [4] [6] や，ターゲット広告用の単語を抽出するもの [21] がある．また，クエリ推薦に広告のクリックを用いるものでは [2] が，検索コンテクストの分析に意味ネットワークを用いるものとしては [10] [11] がある．

閲覧 Web ページからの第 1 クエリワード推薦に関する研究では，AQUAM [14] がある．AQUAM では，閲覧ページから意味情報を抽出し，複数のクエリワードによる検索が行われる．本稿の提案手法と近いアプローチであるが，クエリワード選択のためにバックグラウンド検索を必要とする点が異なる．

また，画面上に見えている単語を簡単に選択するための UI も提案されている．携帯電話向けクリック検索インタフェース [9] では，方向キーしか持たない携帯電話において，円による領域

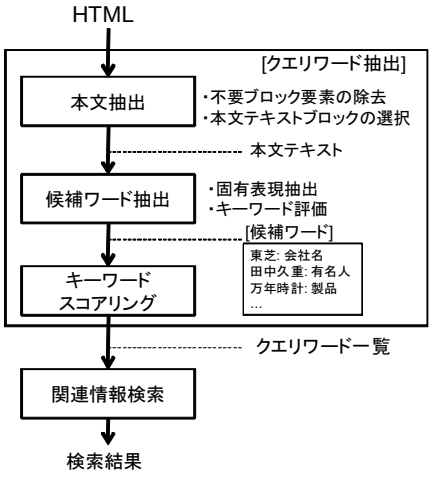

図 1　処理の流れ
Fig. 1　Process flow.

で検索語を指定することで検索を容易にしている．本稿の提案手法は，主に Web ページの内容を読み終えた後，その時点で画面上に表示されていない単語も含め関連情報の検索に利用されやすい単語を抽出する点で，利用するタイミングが異なると言える．

なお，これらの手法は本稿における提案手法と補完的に活用可能であり，今後組み合わせることも考えられる．

3. 閲覧 Web ページからのクエリワード抽出

本節では，閲覧中の Web ページに対し，検索クエリに用いる可能性の高い単語を抽出するまでの流れを説明する．

図 1 にクエリワード抽出処理の流れを示す．この処理は，本文抽出，候補ワード抽出，スコアリングの 3 つのステップから構成される．

3.1 本文抽出

本ステップでは HTML 文書を解析してユーザが注目すると考えられる本文領域を推定し，内部のテキストを抽出する．主にニュースやブログなどの Web ページから記事本文のテキストのみを抽出し，サイドバーや関連記事リンク，広告部分を除去することを目的とする．本文抽出の流れを図 2 に示す．本文抽出処理は以下の手順により行われる．

（1） HTML のブロック要素への分割

div, table, h1 などのブロック要素タグで囲まれた領域毎にテキストを分割し（図 2(a)），ブロック毎の基本スコアを算出する．i 番目のブロックの基本スコア s_i は以下の式で算出される．

$$s_i = (len - at + n \times pw) \times dump\_factor_i$$
$$dump\_factor_i = dump\_factor_{i-1} \times dump\_factor_0 \ (i > 0)$$

ここで，len はタグを除く内部テキスト長，at は内部テキストのうち a タグで囲まれたアンカーテキストの合計長，n は句読点の個数，pw は句読点の重み付け定数，$dump\_factor$ は減衰係数である．

（2） 不要ブロックの除去

内部テキスト長 len が閾値より短い場合，内部テキスト長に対するアンカーテキスト長の比率 at/len が閾値より大きい場合，あるいは予め指定されたストップワードを含む場合は不要ブロックとみなして除去する（図 2(b)）．

（3） 隣接ブロックの連結

前方から順に各ブロックを評価し，隣接ブロックを連結する（図 2(c)）．連結のためのスコア c_i は以下の式により算出される．

$$c_i = s_i \times cont\_factor_i$$
$$cont\_factor_i = cont\_factor_{i-1}/cont\_factor_0 \ (i > 0)$$

ここで，$cont\_factor$ は連続係数である．c_i が閾値以上である場合には c_{i-1} と連結し，閾値以下の場合は連結せず $cont\_factor_i$ を $cont\_factor_0$ に初期化する．

（4） 本文として採用するブロックの選択

全ブロックをスコアの降順に並び替え，スコア総和 Σc_i に対する前方からのスコアの合計の比率が閾値以上になるブロックまでを本文として採用する（図 2(d)）．本文として採用されたブロックを出現位置順に再度並び替え，内部テキストを全て連結したものを本文として抽出する．

上記手順による本文テキストの他，HTML ヘッダの meta タグに記述される title, keywords, description に記述されるテキストが合わせて抽出される．

3.2 候補ワード抽出

次に，得られた本文テキストに対して固有表現抽出を用い，候補ワード抽出が行われる．固有表現抽出手法には統計的な手法や文法ベースの手法が存在する [3] が，本稿では，固有表現抽出手法として，人物名，企業名，地名などの固有名詞や，日時，数量，金額など文字列中の表層表現の意味を表す属性のラベルである固有表現クラスに分類する辞書およびルールベースの手法を用いる．本手法は元々質問応答システム向けに開発されたもので，100 以上の固有表現クラスが定義されている [18] [19]．これらのクラスは，辞書により定義されるものと組み合わせルールにより抽出されるものもある．例えば,「松坂牛」の場合,「地名＋食品→食品」のようなルールが「松坂（地名）＋牛（食品）→松坂牛（食品）」のように適用されることで，予め「松坂牛」が辞書に登録されていない場合も固有表現クラスが付与される．

ただし，固有表現抽出の結果そのままでは，例えば「菅直人首相」という入力に対して「菅直人」と「菅直人首相」のように重なり合う複数の単語が抽出される場合がある．これらの単語を全て表示する場合，表示件数が増え，ユーザにとっては操作がかえって煩雑になる．検索クエリとしては，長い文字列の方が曖昧さが少ないと考えられるが，検索結果が得られない可能性もある．この点を考慮し，包含関係にある単語が抽出された場合は以下の条件に当てはまる場合のみ短い単語を優先し，それ以外は長い単語を優先する．

- 「菅直人」と「菅直人首相」のように姓名と役職が連続する場合
- 「覚せい剤取締法違反」と「覚せい剤取締法」のように主要な単語が変わらない修飾関係

図 2　本文抽出処理の流れ（枠で囲まれた領域が各手順での処理単位）
Fig. 2　Body-text extraction process (each text area surrounded by a square is the process unit).

この結果得られた単語の一覧が，候補ワードとなる．

3.3　スコアリング

そして，前節において抽出された各候補ワードに対しスコアが付与される．著者らは，以下の特徴がクエリに有用な手掛かりであると仮定した．

(a)　Web ページのタイトルに含まれる単語
(b)　Web ページの前半に含まれる単語
(c)　頻出する単語
(d)　文字数の多い単語
(e)　補足説明（例：丸括弧で括られた文）が後に続く単語

また，以下の特徴はクエリに用いられにくい手掛かりであると仮定した．

(f)　検索クエリとするには一般的すぎる単語
(g)　Web サイト名自体を示す単語
(h)　記事の著者や出版社を示す単語
(i)　リストで列挙された単語

これらの特徴をまとめ，以下の属性に関する各スコア（属性スコア）を各候補ワードに付与した．

- 単語の出自（タイトル，デスクリプション，キーワード，本文（(a), (g) に対応）
- 本文中の出現位置（(b), (h), (i) に対応）
- 単語の出現頻度（(c) に対応）
- 単語の文字列長（(d) に対応）
- 補足説明の有無（(e) に対応）
- WebIDF（(f) に対応）
- 固有表現タイプ（(h), (i) に対応）

各候補ワードについて，上述の属性スコアの線形和としてワードスコアが算出される．各属性スコアの係数は，**4.** にて述べる通り，実験計画法および焼き鈍し法 [1] により決定される．

最終的に，高いワードスコアが付与された単語がクエリワードとして出力される．

3.4　関連情報検索

アプリケーションは，提示されたクエリワードを利用することで検索を実行する．

例えば，スマートフォン上の検索支援 UI [16] では，スコアが上位のクエリワード一覧がユーザに提示され，そのうちの 1 つをユーザがタッチすることで，その単語をクエリとした検索処理が実行される．

また，各クエリワードには固有表現タイプが付与されているため，固有表現タイプに応じた検索を実行することも可能である．例えば，人名であればプロフィールや画像を優先して検索することができる．クエリフリー型の検索支援 UI [20] を用いる場合，検索対象のコンテンツに適した固有表現タイプの単語を用いることも可能である．

表 1　検索クエリとして利用された意味分類（上位 5 件）
Table 1　Semantic class used for search queries (top 5).

分類	頻度
人物	106
無生物（商品）	83
組織	81
無形物（タイトル，事件）	60
地域	42

4.　抽出対象ワードの調査

実装にあたり，検索クエリとして用いられる単語を選定するための調査を実施した．また，調査結果に基づき，クエリワード抽出方式の精度向上を行った．

まず，モバイル端末で 2 日に 1 回以上 Web ページを閲覧する 20 代から 50 代の 50 名（男性 35 名，女性 15 名）を対象に，クエリワードおよび検索サービスをそれぞれリストから選択する試作 UI を用いて各自約 40 分自由利用することで，検索に用いられやすい単語を調査した．

調査により行われた合計 407 回の検索において用いられた上位ワードの意味分類とその頻度を表 1 に示す．この結果を受け，**3.3** のスコアリングにおいて，これらの単語の重み付けを高くした．

正解ワードの登録に際しては，ニュース，ブログなどのジャンルからなる 151URL の Web ページについて，関連情報を調べたい単語を調査した．これらの Web ページは，エンタメ，カルチャー，スポーツ，科学技術，政治，経済，グルメ，旅行，国

際，生活の各ジャンルについて様々なドメインから収集したものを用いた．

収集されたページについて，主に IT 系技術者の 29 名により検索クエリとして使いそうな単語を抽出した結果，延べ 2,310 単語が抽出された．平均すると，1URL あたり 4.98 人が作業を行い，10.56 種類の単語が選択されている．この結果を参考に，クエリワード抽出処理の重み付けを調整するとともに，主に 3 人以上が選択した単語を後述のクローズド評価用の正解として登録した．

その後，**3.3** で挙げたどの特徴がクエリワードとして重要であるかを調査した．まず，主効果と交互作用を調査するため，実験計画法による 2^k 要因実験を実施した．1/2 部分要因実験（分解能 VI）で 32 試行の実験を行った結果，単語の出現位置と WebIDF の交互作用，単語の出自と単語の出現位置の交互作用による効果が大きく，出現頻度や補足説明の有無による効果は相対的に小さいことが分かった．そこで，出現頻度や補足説明の有無に関する係数を固定した上で，残りのパラメタを焼き鈍し法により最適化した．

5. 実装例と性能評価

本アプリケーションを，Web ブラウザのアドオンとして実装した．ブラウザが HTML を読み込むと，アドオンが呼び出され，HTML の DOM ツリー文字列がクエリワード抽出モジュールに渡される．前節で述べたクエリワード抽出処理処理が完了すると，通知される．

クエリワード抽出モジュールは Windows® XP/Vista/7, Windows Mobile® 6.5, Linux, Android 2.1 以降の各プラットフォームにおいて実装されているが，以降の節では本アプリケーションをスマートフォン T-01A（OS: Windows Mobile® 6.5（シングルタッチ），CPU: 1GHz, SDRAM 256MB，4.1 インチタッチディスプレイ，ハードウェアキーボード非搭載）において，Web ブラウザ Internet Explorer® Mobile のアドオンとして実装させたものを用いている．本モジュールは，モジュール本体，辞書，ルールあわせて約 10MB であるため，組み込み向けでも利用可能なサイズである．

5.1 検索支援 UI の実装例

本節では，モバイル環境においてテキスト入力の手間を削減しつつ Web 閲覧からの関連情報検索を行う検索支援 UI の実装例 [16] を図 3 に挙げる．

図 3(a) は Web ブラウザで Web ページを閲覧中の画面である．検索支援機能を利用可能な状態になると，図 3(a) 上部に通知アイコンが表示される．ユーザが閲覧中のページに関連する情報を知りたい場合，通知アイコンをタッチすると，ページ内で検索に利用されやすいとアプリケーションが判断したクエリワード一覧が表示される（図 3 (b)）．ユーザが，より深く知りたい場合クエリワードを見つけた場合，その単語が表示されているボタンをタッチすると，検索が行われ，検索結果が表示される．このようにして，ユーザは文字を入力せずタッチ操作だけで関連情報を検索できる．

図 3　検索支援インタフェース実装例
Fig. 3　Example use of search-support application.

表 2　様々な Web ページに対するクエリワード抽出時間．時間は 4 回試行の平均を表す．
Table 2　Keyword extraction time for various web pages. Term extraction time is the average of four runs.

ページ ID	HTML サイズ (kB)	候補ワード数	抽出時間 (msec)
1	50.9	44	1580
2	184.0	100	4759
3	218.1	85	2581
4	137.7	13	2021
5	39.0	49	2131
6	62.0	46	2630
7	40.0	48	1093
8	49.0	47	1736

5.2 クエリワード抽出性能

上記アプリケーションにおけるクエリワード抽出性能の評価を行った．HTML のサイズやクエリワード数の異なるニュース記事，ブログエントリなど 8URL の Web ページに対してクエリワード抽出処理を実行した結果を表 2 に示す．

表 2 より，ユーザが Web ページを読む時間に比べクエリワード抽出処理が十分速く動作することが分かる．したがって，スマートフォン上において実用的な速度で利用できることを確認できたと言える．

6. 精度向上

開発者の視点では，改良のために評価対象の Web ページを随時追加したいが，前節で述べた手法でのチューニングを毎回行うのは手間がかかる．

また，同様のエンジンを他の言語に適用する場合や，異なるプラットフォーム向けに適用する場合など，大部分の処理は共通だが前述の属性スコアに対するパラメタが異なる場合のチューニングを一から行うのは作業コストが大きい．

そのため，著者らは評価システムを試作し，正解データ登録から精度向上までの作業を半自動化することで効率化を図りつ

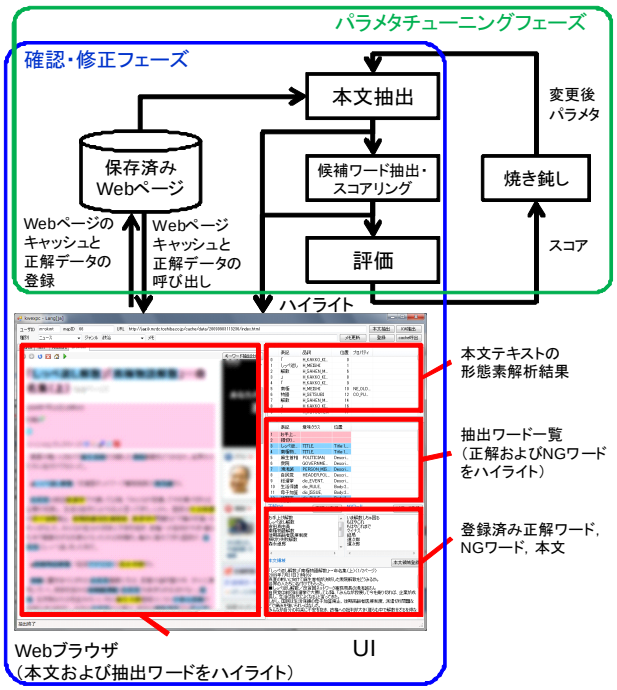

図 4　評価システムの構成

Fig. 4　Configuration of evaluation system.

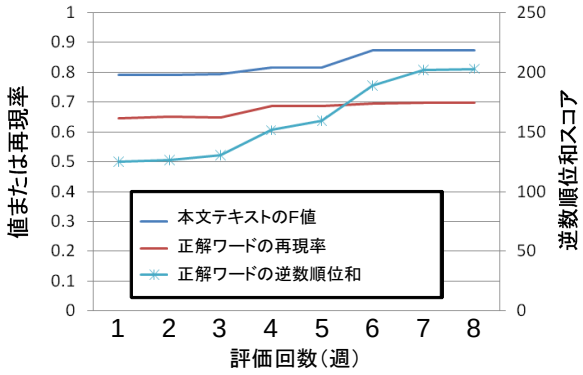

図 5　クローズド評価の推移

Fig. 5　Changes in the closed evaluation results.

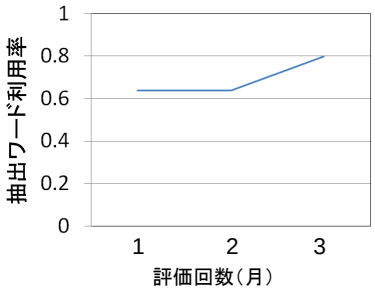

図 6　オープン評価の推移

Fig. 6　Changes in the open evaluation results.

つ精度向上を進めた．

本章では，評価システムを用いた精度向上の流れと，前章で紹介したアプリケーションの改良における作業例について述べる．

6.1　評価システム

図 4 に，評価システムの単純化した構成図を示す．

本システムの動作には 2 種類のフェーズが含まれる．1 つは確認・修正フェーズで，もう 1 つはパラメタチューニングフェーズである．確認・修正フェーズでは，抽出結果を可視化しながら開発者はクエリワード抽出モジュールや辞書・ルールの修正を行う．開発が一段落すると，パラメタチューニングフェーズにより，焼き鈍し法に基づくチューニングを行う．

まず，開発者は Web ブラウザを用いて Web ページの閲覧を行う．評価対象となるページを見つけると，そのページと正解データを登録する．正解データとして，本文テキスト，正解ワード，NG ワードが登録される．本稿では，クエリワード抽出において抽出されるべきでない単語を NG ワードとする．

その後，開発者が登録された Web ページを呼び出すと，システムにより本文およびクエリワード抽出が行われる．抽出結果はそれぞれ Web ブラウザ上にハイライト表示される．これにより，開発者は画面を見ながら抽出された本文テキストやクエリワードが適切であるかどうかを確認できる．抽出に問題があれば，開発者はモジュールや辞書・ルールを修正し，再度確認する．

修正が一段落した後，評価モジュールを実行することで，登録された全ページに対する評価指標が算出される．算出される指標として，例えば以下が利用可能である．

- 正解本文テキストに対する，本文抽出結果の文字数比率による F 値
- 正解クエリワードの再現率，適合率，F 値
- 正解クエリワードの逆数順位和，平均逆数順位

これらのスコアは開発モジュールの各リビジョンに対して算出される．スコアの変化を確認することで，開発者は副作用の有無を確認しながら後戻りの少ない修正を進めることができる．

6.2　精度向上の例

前節で述べた評価システムを用いて精度向上を行った．

5. で紹介した検索支援インタフェース向けクエリワード抽出モジュールに対して適用した．2 か月間継続的な改良を行い，週毎に 295URL の Web ページ（うち 151URL は **4.** で用いたものと同じ）クローズド評価を行った結果を図 5 に，月毎に 5 名によるオープン評価を行った結果を図 6 に示す．

クローズド評価では，主に

- 正解本文テキストに対する，本文抽出結果の文字数比率による F 値
- 正解クエリワードの再現率
- 正解クエリワードの逆数順位和

を算出することで精度の確認を行った．また，オープン評価では，5 名の被験者が 1 時間自由に Web ブラウジングを行い，検索を行いたいタイミングで **5.** の機能を呼び出し，検索に利用可能なクエリワードが抽出された比率を測定した．なお，図 5 では，比較のため最終的なチューニングを行った後のパラメタに基づき算出されたスコアを用いている．

図 5 より，正解ワードの再現率の改善幅は全期間を通じてそ

れほど大きくないものの，逆数順位和に関しては後半に大きくスコアを増やしている，つまり，より良いクエリワードが上位に出現していることが分かる．図 6 とあわせると，クエリワードが抽出されているかどうかよりも，上位にあることがオープン評価でのクエリワード利用に結び付いているものと考えられる．

7. 精度評価

5. で紹介し，**6.** で述べた改良を適用した検索支援インタフェースについて，最終的な精度を報告する．以下，クローズド評価によるクエリワード抽出精度，およびオープン評価によるクエリワード抽出精度について述べる．

7.1 クローズド評価

本節では，想定される Web ページを対象としたクローズド評価の項目および精度について述べる．

評価対象として，上記のジャンルから集められた Web ページを用い，うち 299URL の Web ページ（うち 151URL は **4.** で用いたものと同じ）に対し，前節の結果を参考に正解として 1 つ以上のクエリワードを登録した．正解は 1,148 個（URL あたり平均 3.84 個）で，作成に際し 2 人以上が登録・確認を行っている．また，本文抽出の精度評価のために，メニューや広告を含まない，ページの主要コンテンツと呼べる文章を本文の正解として登録した．

評価指標としては，以下の指標を用いた．

- 本文抽出 F 値：正解として登録した本文に対する，本文抽出結果の文字数比率による F 値
- クエリワード抽出再現率：正解として登録したクエリワードの再現率
- クエリワード抽出 8 位再現率：正解として登録したクエリワードの，抽出されたクエリワードの上位 8 件以内における再現率

本評価では，多くの場合出力可能なクエリワード個数より正解個数の方が少ないため，正解が含まれるかどうかを重視するために，適合率ではなく再現率を用いた．

正解データを登録した 299URL に対する結果は以下の通りである．

- 本文抽出 F 値: 0.874
- クエリワード抽出再現率: 0.698
- クエリワード抽出 8 位再現率: 0.539

URL あたり平均して 3.84 個正解が登録されているので，図 3(b) の UI を用いる場合，8 個のうち平均して 4，5 個の正解キーワードが提示されることになる．また，クエリワード抽出再現率と 8 位再現率の差が 0.159 と比率が小さいことから，スコアリングによるクエリワードの取りこぼしは殆ど起こっていないと考えられる．

しかしながら，候補ワード抽出において約 30%が抽出に失敗している．原因としては，スコアリング段階での失敗，固有表現抽出方式で扱える語彙の不十分さ，単語の表記自体は一般語と変わらない場合，などの原因があり，これらの解消は今後の課題である．

表 3　オープン評価の結果（各数値はユーザ毎の平均）
Table 3　Result of open evaluation (each value is the average of users).

項目	平均
閲覧ページ数	40
検索支援機能利用回数 (A)	13.9
クエリワード選択セッション数 (B)	10.2
提示ワード利用率 (B/A)	0.733

7.2 オープン評価

クローズド評価に続き，実際のユーザによる印象を調査するために提案インタフェースのクエリワード抽出に関する，モバイル端末によるオープン評価を実施した．

被験者は情報検索に慣れている 14 名である．各被験者はモバイル端末 (T-01A) を用い，約 1 時間ずつ，普段 PC や携帯端末で閲覧するのと同様に自由利用を行った．

ユーザは関連情報を調べたいタイミングで **5.1** の検索支援機能を利用し，ユーザが検索に利用したいクエリワードが表示されている場合はそれを用いて検索を行った．

自由利用の後，検索支援機能を用いた時に提示されたクエリワード一覧を確認し，それぞれの単語に対する満足度を以下の 3 段階で評価した．

3: 検索したい
2: クエリワード一覧に出ても良い
1: クエリワード一覧に不要

オープン評価では，評価指標として検索に利用したい単語が提示されている割合，つまり 1 個以上評価値が 3 である単語が含まれる割合である提示ワード利用率を用いた．

$$\text{提示ワード利用率} = \frac{\text{クエリワード選択による検索セッション数 (B)}}{\text{検索支援利用セッション数 (A)}}$$

オープン評価の結果を表 3 に示す．

表 3 より，73.3%の検索において提示ワードが利用されていることが分かる．つまり，ユーザが検索しようとした場合の入力の手間をそれだけ削減できたと言える．

8. おわりに

本稿では，携帯電話やスマートフォン，TV などの組み込み機器を対象とした，閲覧中 Web ページの内容に関係する情報の検索行動を支援するクエリワード抽出方式を提案した．抽出方式の選定にあたっては，実際の Web ページを対象に調査を行い，クエリワードとして利用されやすい単語の特徴を調査した．

また，評価システムを用いた後戻りの少ない精度向上サイクルを通じた 2 か月の改良を行い，その結果スマートフォンを対象として開発したエンジンのクローズド評価・オープン評価両方の精度測定において実用的な検索精度が得られることを確認した．

本稿で提案した手法や実装例は，2 番目以降のクエリワードの利用に関しては考慮していない．したがって 2 番目の単語の提案方式は課題である．また，提示する用語は HTML に含ま

れる単語のみ利用するため，ユーザが検索に利用したい単語とは表記が合わない場合もある．これらに関しては，シソーラスを用いた同義語や関連語の利用が考えられる．

また，提案手法の処理対象は HTML 全体であり，画面に描画されない部分も含め閲覧 Web ページ全体に対して行われている．そのため，実際には閲覧していない部分からもクエリワードが抽出される可能性がある．また，ユーザの過去の操作履歴も反映されていない．今後は，これらの点も考慮の上，抽出手法の改良および精度向上を図るとともに，他の言語への適用や他の組み込み機器への応用も進める予定である．

文　献

[1] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*, John Wiley and Sons Inc., 1988.

[2] I. Antonellis, H. Garcia-Molina, and C. Chang, "Simrank++: query rewriting through link analysis of the click-graph," In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pp.1177–1178, 2008.

[3] A. Borthwick, *A Maximum Entropy Approach to Named Entity Recognition*, Doctoral Thesis, New York University, 1999.

[4] H. Cui, J. Wen, J. Nie, and W. Ma, "Query Expansion by Mining User Logs," *IEEE Transaction on Knowledge and Data Engineering*, vol.15, no.4 , pp.829–839, 2003.

[5] P. E. Hart and J. Graham, "Query-Free Information Retrieval," *IEEE Expert*, vol.12, no.5, pp. 32–37, 1997.

[6] Q. He, D. Jiang, Z. Liao, S. Hoi, K. Chang, and E. Lim., "Web Query Recommendation via Sequential Query Prediction," In *Proceedings of the 2009 IEEE international Conference on Data Engineering (ICDE '09)*, pp.1443–1454, 2009.

[7] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin, "Query-free news search," *World Wide Web*, vol.8, no.2, pp.101–126, 2005.

[8] M. Kamvar, M. Kellar, R. Patel, and Y. Xu, "Computers and iPhones and mobile phones, oh my!: a logs-based comparison of search users on different devices," In *Proceedings of the 18th international Conference on World Wide Web (WWW '09)*, pp. 801–810, 2009.

[9] D. Komaki, K. Ohnishi, Y. Arase, G. Hattori, T. Hara, and S. Nishio, "Design and Implementation of A Click-Search Interface for Web Browsing Using Cellular Phones," *International Journal of Web and Grid Services (IJWGS)*, vol.5, no.1, pp.66–84, 2009.

[10] R. Kraft, C. Chang, F. Maghoul, and R. Kumar, "Searching with context," In *Proceedings of the 15th international Conference on World Wide Web (WWW '06)*, pp.477–486, 2006.

[11] R. Kraft, F. Maghoul, and C. C. Chang, "Y!Q: contextual search at the point of inspiration," In *Proceedings of the 14th ACM international Conference on information and Knowledge Management (CIKM '05)*, pp.816–823, 2005.

[12] H. Lieberman, "Letizia: an agent that assists web browsing," *Proceedings of the 14th international Joint Conference on Artificial intelligence (IJCAI '95)*, pp.924–929, 1995.

[13] Q. Ma and K. Tanaka, "Topic-structure-based complementary information retrieval and its application," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol.4, no.4, 475–503, 2005.

[14] F. Menemenis, S. Papadopoulos, B. Bratu, S. Waddington, and Y. Kompatsiaris, "AQUAM: automatic query formulation architecture for mobile applications," In *Proceedings of the 7th international Conference on Mobile and Ubiquitous Multimedia (MUM '08)*, pp.32–39, 2008.

[15] M. Okamoto, M. Kikuchi, and T. Yamasaki, "One-button search extracts wider interests: an empirical study with video bookmarking search," In *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '08)*, pp.779–780, 2008.

[16] M. Okamoto, N. Watanabe, M. Kikuchi, T. Iida, K. Sasaki, K. Horiuchi, T. Yamasaki, S. Omura, and M. Hattori, "First query term extraction from current webpage for mobile applications," In *Proceedings of the 9th international Conference on Mobile and Ubiquitous Multimedia (MUM '10)*, pp.19:1–19:9, 2010.

[17] M. Rahurkar and S. Cucerzan, Silviu, "Predicting when browsing context is relevant to search," In *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '08)*, pp.841–842, 2008.

[18] T. Sakai, "Advanced technologies for information access," *International Journal of Computer Processing of Oriental Languages*, vol.18, no.2, pp.95–113, 2005.

[19] T. Sakai, Y. Saito, Y. Ichimura, M. Koyama, T. Kokubu, and T. Manabe, "ASKMi: a Japanese question answering system based on semantic role analysis," In *Proceedings of Recherche d'Information Assistee par Ordinateur (RIAO '04)*, pp.215–231, 2004.

[20] 渡辺 奈夕子, 岡本 昌之, 菊池 匡晃, 飯田 貴之, 佐々木 健太, 堀内 健介, 服部 正典, "モバイル機器における閲覧 Web ページからのクエリ抽出を用いた検索支援システム," 電子情報通信学会研究報告, WI2-2010-49, 2010.

[21] W. Yih, J. Goodman, and V. R. Carvalho, "Finding advertising keywords on web pages," In *Proceedings of the 15th international Conference on World Wide Web (WWW '06)*, pp.213–222, 2006.

[22] T. Yoshida, S. Nakamura, and K. Tanaka, "WeBrowSearch: toward web browser with autonomous search," In *Proceedings of the 8th international Conference on Web Information Systems Engineering (WISE '07)*, pp.135–146, 2007.