情報学基礎　64-11
自然言語処理　145-11
（２００１．９．１１）

モバイルサーチエンジン WithAir の試作と評価

河合英紀　赤峯享　喜田弘司　松田勝志　福島俊一
{kawai, akamine, kida, mat, fuku}@hml.cl.nec.co.jp
NEC インターネットシステム研究所

モバイル端末での利用に特化して効率のよい検索を可能にするサーチエンジン WithAir を試作し、性能を評価した。本サーチエンジンは、(1)モバイルページを選択的に自動収集する機能、(2)モバイルページの位置情報を自動分類し、タウン情報や観光情報を検索できる機能、(3)リンク構造解析とページ要約文抽出により、全文検索結果を高精度に表示する機能、(4)人気キーワードを手軽に入力できるキーワードナビ機能、(5)検索目的を先読みして優良サイトを提示する的中ナビ機能、を備えることを特長とする。

基本性能の評価の結果、モバイルページ 130 万件を 99%の精度で自動収集でき、そのうち位置情報に関係するモバイルページ 37 万件を 93%の精度で抽出できた。さらに利便性の評価の結果、従来のモバイルサーチエンジンと比較して操作コストが最小であることを確認した。

Development and Evaluation of a Mobile Search Engine WithAir

Hideki KAWAI, Susumu AKAMINE, Koji KIDA, Katsushi MATSUDA, Toshikazu FUKUSHIMA
Internet Systems Research Laboratories, NEC Corp.

This paper describes a novel search engine WithAir, which can retrieve mobile contents at low operation cost and with high precision. WithAir has the following five features. (1) Mobile-contents-focused crawler, that can collect Web pages published for browser phones selectively. (2) Location-based retrieval method, using location-related expressions detection and classification technologies. (3) Ranking and summarization methods for mobile contents, based on link structure analysis. (4) Keyword anticipation method, that can reduce costs of query keyword input operation. (5) Purpose anticipation method, which can estimate the user's purpose from a query keyword and show the pages relevant to it.

This paper also describes the evaluation results. They show that WithAir can collect 1.3 million mobile pages with accuracy of 99%, and can select 370 thousand location-related pages with accuracy of 93%. They also show that the operation cost of WithAir is the lowest, compared with conventional mobile search engines.

1. はじめに

近年、Web に接続可能なモバイル端末が急速に普及している。特に、i モードをはじめとするインターネット接続可能な携帯電話/PHS の利用者は 2001 年 2 月時点で 1830 万人に達している[1]。それに伴い、モバイル端末で閲覧できるようにデザインされた Web ページの数も急増している。たとえば、i モードから閲覧可能なサイトは、2001 年 7 月時点で少なくとも 46000 件以上存在しており[2]、その中には質の高い人気ページも数多く含まれている。これらモバイルページの質・量両面の充実と同時に、それを探すためのモバイルサーチエンジンへの要求が高まっている。また、モバイル端末の利用場面は、出張や旅行など外出先が多いため、位置に関係した情報の検索のニーズも大きい。さらに、モバイル端末は文字入力が煩わしく画面も小さいなど、利用者の操作コストが高いため、効率のよい入力補助機能やナビゲーション機能が求められている。

従来のモバイルサーチエンジンには、(A)モバイルページのみを検索するタイプと、(B)モバイルページも Web ページも区別なく検索し、モバイル端末向けに変換して表示するタイプの 2 種類がある。OH!NEW?[2]や Yahoo!モバイル[3]、i-seek[4]など、多くの従来モバイルサーチエンジンは(A)タイプであり、モバイル用のサイトを人手で登録している。しかし、人手による登録では更新頻度に限界があり、全文検索のような網羅的な検索も不可能である。また、同じ(A)タイプでも i-Yappo[5]など一部のロボット型サーチエンジンでは、モバイルページを自動で収集している。しかし、現状の自動収集では、モバイルページのみを精度よく選択して、良質なページを上位にランキングすることができないため、検索結果が膨大になるだけで逆に良質なページにたどり着きにくくなってしまっている。一方、

(B)タイプとしては、i モード版 Google[6]があるが、Web ページを i モード端末用のフォーマットに変換するだけでは内容をつかみにくく、必要な情報にたどり着くまでに多くのページ切り替えが必要となる。

本研究では、モバイルページを選択的に自動収集し、検索結果を人気度順に表示することによって良質なサイトを素早く網羅的に検索でき、位置情報を抽出することによって外出先で地域情報を検索できるモバイルサーチエンジン WithAir を試作し、その基本性能を評価した。また、利用者の操作コストを減らすために、人気キーワードを手軽に入力できるキーワードナビや、入力キーワードから検索目的を先読みして適合サイトを提示する的中ナビを実装し、その利便性を評価した。

2. 設計方針

本研究では、前節で述べた従来の課題を踏まえて下記の 3 点を基本方針としてモバイルサーチエンジン WithAir を設計した。

(1) モバイルページの収集や、位置情報による分類を自動で行うことにより、鮮度を保った網羅的な検索を提供すること。

(2) 検索結果を人気度順で表示したり、サイトの内容を端的に表現することによって、大量の自動収集ページを高精度に検索できるようにすること。

(3) 入力補助機能や検索目的の先読み機能により、モバイル端末の制限下での利用者のコストを最小限にすること

3. WithAir の特長

上記の設計方針にしたがって開発したモバイルサーチエンジン WithAir は、次の 5 つの特長をもつ。

(1) 専用クローラによるモバイルページの自動収集
(2) 位置情報の自動抽出による地域情報検索
(3) 人気度によるランキングとページ要約
(4) キーワード入力を補助するキーワードナビ機能
(5) 検索目的を先読みして提示する的中ナビ機能

以下、図 1 に示した WithAir の画面例を参照しながら、各特長とその実装方法を説明する。

3.1 モバイルページの自動収集

大量のモバイルページを鮮度を保って網羅的に検索するためには、モバイルページを選択的に自動収集する必要がある。WithAir では、Web クローラにより収集したページから、「ページタイプ判定技術」[7]を用いてモバイルページの特徴(ページタイプ)をスコア化することにより、モバイル端末で表示可能なページのみを選別する。モバイルページの特徴としては、ページサイズ、使用タグ、キーワード等の様々な情報に着目する。今回はまず i モードページを対象として実装し、例えば下記のような特徴を記述することで、i モードページらしさをスコア化した。

図 1 WithAir の画面例

加点の条件：

i モード依存の絵文字が使用されている。

ディレクトリに/i_mode/などの表現がある。など

減点の条件：

<FRAME>など特定タグが使用されている。

テキスト部分のサイズが 2500Byte 以上ある。など

本クローラは、フォーカストクローラ[8]の一種であるが、ページタイプに着目してページを選別・収集する点を特徴としている。これにより、i モードページの収集に関して実用精度の実現が可能となった。

3.2 地域情報検索

モバイル端末の大きな長所として、「どこにでも持ち歩けて、いつでも使えること」がある。特に旅行先や外出先など、PC がその場にない環境で、その位置に関するレストラン情報や宿泊情報を検索したいというニーズは大きい。そこで、WithAir ではモバイルページの中から住所や電話番号など、位置に関係した表現を抽出して、それを元にページを地域で分類し[9]、利用者が階層構造をたどって地域を選択すると(図 1 の画面(D))、その地域のタウン情報や観光情報を優先的に検索できるようにした(図 1 の画面(E))。

図 2 に位置情報の抽出例を示す。位置情報の抽出で

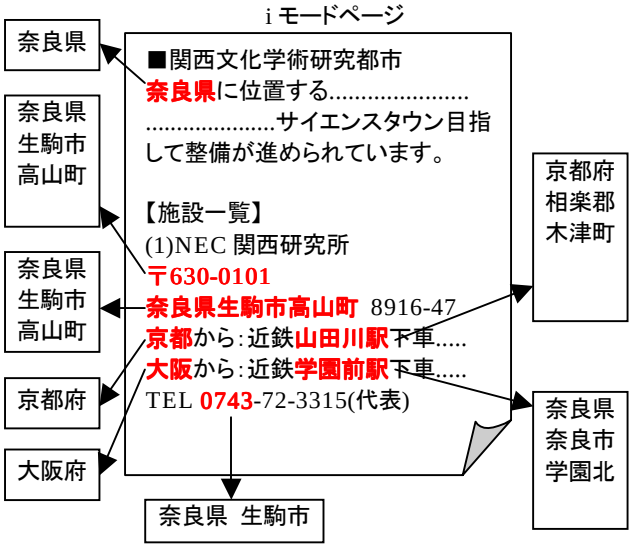

図 2 位置表現の抽出例

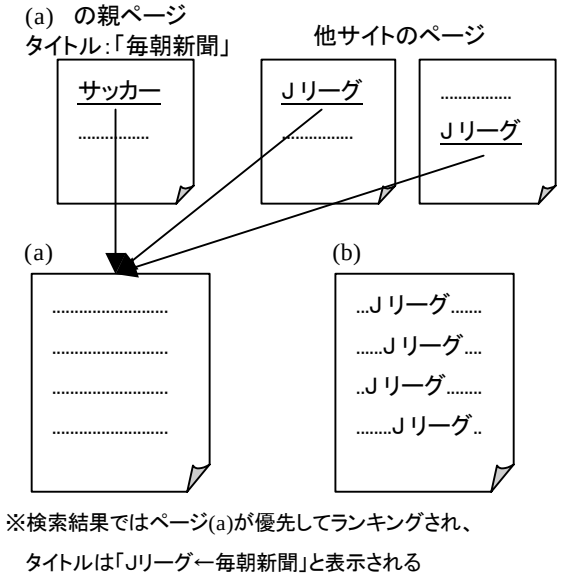

図 3 ランキングと要約文抽出の例

は、地名 DB に格納されている地名と、モバイルページに記述されている地名、市外局番、郵便番号、鉄道駅名などの位置表現とのマッチングを行う。図2では、「奈良」「京都」などの地名や、「0743」という市外局番、「630-0101」などの郵便番号、「山田川駅」などの駅名がマッチしている。次に、マッチしたモバイルページの文字列に対して、その文字列が表す位置を、都道府県→市郡→区→町村の階層表現に関連付ける。異なる地域に同一の地名がある場合は、その地名を複数の階層表現に関連付ける。図 2 では、「奈良県」という地名にマッチした場合、その地名が表す県の階層「奈良県」までを関連付けており、「0743」という市外局番がマッチした場合、その局番が表す市の階層「奈良県生駒市」までを、関連付けている。

位置情報の分類では、あるページが位置Xに関する情報であるかどうかスコア化して判定する。このスコアは、ページ内で位置 X と関連のある位置表現の出現頻度と、位置 X 以外の位置と関連のある位置表現の出現頻度の差で計算する。図2の例の場合、奈良県が5回、京都府が2回、大阪府が 1 回出現している。したがって、このページが奈良県であるスコアは 5-3=2 である。また、市の階層に注目すると、生駒市が 3 回、奈良市が 1 回、相楽郡が 1 回出現している。したがって、生駒市であるスコアは 3-2=1 であり、奈良市や相楽郡であるスコアは 1-4=-3 である。検索時には、入力された地名やキーワードに対して、スコアが閾値以上のページをその地域の情報と判定して検索する。一方、地名のみ指定されて、キーワードが指定されていない場合は、「レストラン」「グルメ」「観光」「ホテル」などのキーワードを含むページを優先して表示する。

地域情報検索の関連技術として、Web ページ中の住所表記を緯度経度に変換して検索するモバイルインフォサーチ[10]などがある。WithAir では、位置表現抽出の対象をモバイルページに限定して、ページ単位で位置情報を分類する点や、タウン情報を優先して検索できる点を特徴としている。

3.3 ランキングとページ要約文抽出

大量のページを単純に全文検索するだけでは、結果が膨大になり、かえって良質なページにたどり着きにくくなってしまう。そこで、WithAir では、「サイテーションエンジン」によるリンク構造解析[11]を用いて算出した人気度を用いてランキングを行い、人気ページを検索結果の上位に表示できるようにした。また、モバイルページでは正しくページタイトルが記述されていない場合が多いため、ページタイトルの代わりにアンカー文字列からその内容を端的に表現可能なページ要約文[12]を抽出して表示することにした。(図 1 の画面(C)下部、画面(E)のリンク文字列)

ランキングと要約文抽出の実装方法の概要を図 3 に示す。ランキングでは、ページ本文だけでなく、アンカー文字列も検索し、ページ本文よりアンカー文字列でヒットしたページを優先して表示する。ページ本文のみでヒットしているページは、人気度が高い方を優先して表示する。例えば、検索キーワードが「Jリーグ」であった場合、図3では、本文中で 4 件ヒットしているページ(b)よりも、他サイトからのアンカー文字列で 2 件ヒットしているページ(a)を優先してランキングする。

要約文抽出は、複数のページからはられたリンク中のアンカー文字列のうち、最も多くのアンカー文字列で使われている表現をそのページのタイトルとし、さらに、親ページ

のタイトルと「←」でつないで併記することによって、どんなページから引用されているかを示す。図 3 では、ページ(a)へのリンクで使われているアンカー文字列は「サッカー」が 1 回、「J リーグ」が 2 回なので、ページ(a)のタイトルは「J リーグ」となる。さらに、ページ(a)の親ページのタイトルが「毎朝新聞」であるため、検索結果には、ページ(a)を表現する要約文として「J リーグ←毎朝新聞」と表示する。

関連技術としては、Google の PageRank 技術[13]などが挙げられる。WithAir では、モバイルページに対するチューニングを行っており、ページ要約を抽出する点を特徴としている。

3.4 キーワードナビ機能

i モードでは、テンキーを使ってキーワードを入力する必要があり、利用者にとって大きな操作コストとなっている。そこで WithAir では、「あ行」「か行」のようにキーワードの読みを 1 文字ずつたどることによって、その読みで始まる人気キーワードを候補として表示するキーワードナビ機能を実装した。これにより、利用頻度の高いキーワードほど手軽に入力が可能になる。例えば、「温泉」というキーワードで検索する場合、従来のテンキー入力では、「入力モードに切り替え、11111(お)、000(ん)、3333(せ)、000(ん)、文字確定(おんせん)、変換、変換確定(温泉)、文字列確定、入力モードを解除、検索」と、テンキーと決定キーの押下(ストローク)が合計 22 回必要であったのに対し、キーワードナビでは「キーワードナビに切り替え、1(あ行)」の入力だけで図 1 の画面(B)のように人気キーワードの候補として「温泉」が現れるため、3 ストロークで検索できる。

キーワードナビ用データの例を表 1 に示す。表 1 では縮退ルビをキーにして、その縮退ルビではじまるキーワードが検索ログでの利用頻度順に格納されている。縮退ルビとは、キーワードの読みを、「あ行」「か行」などの字種別に縮退したものであり、例えばキーワード「温泉」の場合、読みが「おんせん」で縮退ルビは「あわさわ」である。

表 1 の場合、最初に利用者が「あ行」を選択すると、縮退ルビが「あ」で始まる人気キーワード上位 5 語が「アイドル」「占い」「アニメ」「オークション」「温泉」の順で表示される。次に 2 文字目で「わ行」を選択すると、縮退ルビが「あわ」で始まる人気キーワード上位 5 語が「温泉」「音楽」「インターネット」「インテリア」「淡路島」の順で表示される。

テンキー入力におけるキーワード入力補助機能には、携帯電話に保持した辞書を用いる eZiText[14]などが挙げられる。WithAir では、検索ログを利用したキーワード辞書をサーバー側に保持しており、読みの最初の数文字で残りの文字も補完する点を特徴とする。

表 1 キーワードナビ用データの例

縮退ルビ	人気キーワード
あ	アイドル、占い、アニメ、オークション、温泉
あわ	温泉、音楽、インターネット、インテリア、淡路島
あわさ	温泉、淡路島、印刷、インストール、飲食店
・・・	・・・

3.5 的中ナビ機能

モバイル端末の画面は小さいため、検索結果の URL リストや説明文などの補助情報の表示が制限される。そのため、利用者は検索目的に適合するページを見つけるために、検索結果を一つずつ閲覧しなくてはならない。そこで WithAir では、入力されたキーワードに関する典型的な検索目的と優良サイトを関連づけて提示することによって、少ない情報量で素早く的確なページへ誘導できる先読み型の新しいナビゲーション機能「的中ナビ」を提案・実装した[15]。図 1 の画面(C)上部では、「温泉」というキーワードに典型的な検索目的として、露天風呂や混浴などのこだわり検索を持つサイトと、全国の温泉宿泊施設などの一般的な情報を持つサイトを提示している。

的中ナビ用データの例を図 4 に示す。図 4 では、あらかじめ利用頻度の高いキーワードについて、典型的な検索目的とそれに適合する優良サイトを関連付けて格納している。優良サイトが複数存在する場合は、複数の優良サイトを登録しておく。検索時には、入力キーワードをキーにして検索目的を先読みし、目的ごとの優良サイトのうち 1 件を提示する。さらに、検索目的のリンクがクリックされると、キーワードと検索目的をキーにデータベースを検索し、優良サイトのリストを表示する。

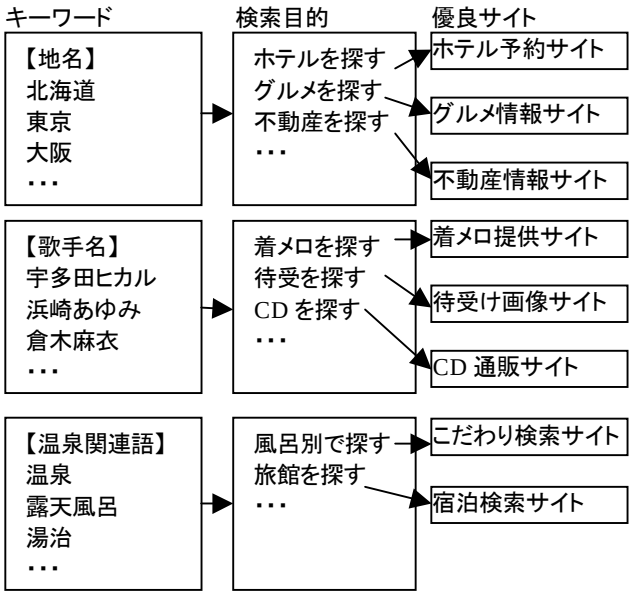

図 4 的中ナビ用データの例

4. 試作システム

図 5 に試作したモバイルサーチエンジン WithAir のシステム構成図を示す。本システムは、(1)モバイルページを自動収集する収集部と、(2)収集したページを解析して情報を抽出し、インデックスを生成する解析・登録部と、(3)検索を行う検索部、および(4)キーワード入力を補助する入力補助部の 4 つのサブシステムからなる。なお、本試作システムは、端末利用者の数を考慮して、まず i モード端末に対応した。

(1) 収集部

収集部は、Web クローラによりページを自動収集し、モバイル判定モジュールにより i モードから閲覧可能なモバイルページを選別し、モバイルページ DB に格納する。

(2) 解析・登録部

解析・登録部は、リンク構造解析モジュール、位置情報抽出モジュール、インデックス作成モジュールからなる。

リンク構造解析モジュールは、ページ間のリンク構造を解析してスコア付けし、アンカー文字列を DB に格納する。また、結果表示のためのページ要約文を抽出し、表示用 DB に格納する。

位置情報抽出モジュールは、地名 DB を参照して、モバイルページに記述されている地名、市外局番、郵便番号、駅名などの位置表現を自動抽出し、位置によるスコア付けを行って位置情報ページ DB に格納する。

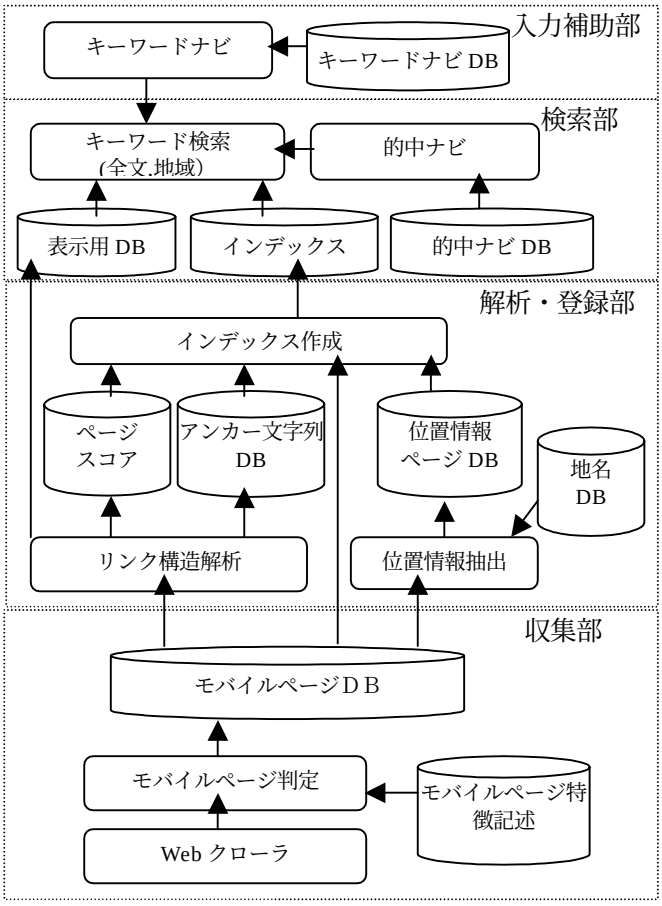

図 5 WithAir 試作システムの構成

インデックス作成モジュールは、アンカー文字列 DB、モバイルページ DB、位置情報ページ DB からそれぞれ全文検索用の文字列インデックスを作成する。

(3) 検索部

検索部は、キーワード検索モジュールと的中ナビモジュールから構成されている。キーワード検索モジュールは、キーワードによる全文検索と地域を指定した地域情報検索の 2 種類の検索方法があり、利用者の目的に応じて切り替えることができる。

(4) 入力補助部

入力補助部はキーワードナビ機能を提供する。キーワードナビ DB には、縮退ルビをキーにして、キーワードが利用頻度順に格納されている。

5. 性能評価

モバイルサーチエンジンの基本性能として、収集部のモバイル専用クローラと、解析部の位置情報抽出モジュールの評価結果を示す。

(1) モバイル専用クローラ

今回は、モバイルページとして i モード端末で閲覧可能なページを収集するようチューニングを行った。モバイル専用クローラの基本性能を表 2 に示す。適合率は、モバイル専用クローラが i モードページと識別したページ(1251 件)のうち、実際に i モード端末で表示可能であったページ(1246 件)の割合である。再現率は、人手でピックアップした i モードページ(1326 件)に対して、クローラが実際に i モードページであると正しく判定できたページ(1169 件)の割合である。

表 2 モバイル専用クローラの基本性能

収集ページ数	適合率	再現率
130 万ページ	99%	88%

(2) 位置情報抽出・検索モジュール

位置情報抽出・検索モジュールの基本性能を表 2 に示す。モバイル専用クローラで収集した i モードページ 130 万ページのうち、37 万ページが位置に関する情報を含んでいた。位置適合率は、ある特定地域の位置情報を含むと判別したページ(15 件)のうち、判別が正しかったページ(14 件)の割合である。また、位置再現率は、人手でピックアップした位置情報を含むページの集合(1957 件)のうち、位置情報を含むと正しく判別できたページ(1480 件)の割合である。さらに、地域情報適合率は、位置情報ページのうち、地域情報検索によってタウン情報や観光情報に関するページを検索できる割合である。

表 3 地域情報抽出・検索モジュールの基本性能

位置情報ページ数	位置適合率	位置再現率	地域情報適合率
37 万ページ	93%	76%	91%

6. 利便性評価

検索結果のランキング方法やキーワードナビ機能、的中ナビ機能も含めた総合的な利便性の指標として、トップページから目的のページにたどり着くまでに必要なテンキーと決定キーの最小ストローク数を測定した。図6に「旅行先のホテル」「新幹線の時刻表」「アーティストの着メロ」など、モバイルでの典型的な 10 件の検索課題に対して、WithAir とモバイルサーチエンジンとして一般によく知られた 3 つのサービス(OH!NEW?[2]、Yahoo!モバイル[3]、Infoseek[4])におけるキーワード検索とカテゴリ検索の最小ストローク数を比較した結果を示す。

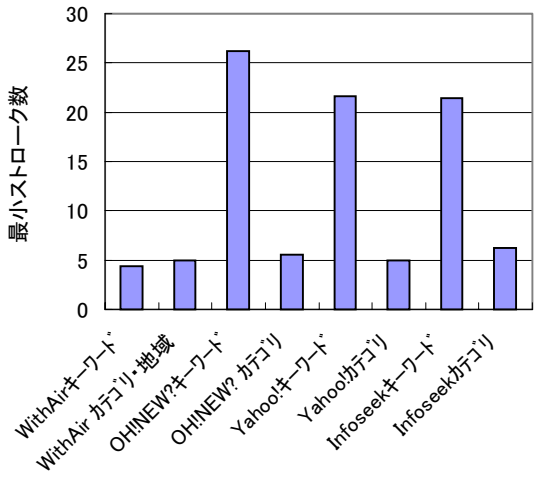

図 6 モバイル検索の最小ストローク数比較

図 6 より、従来サービスでは、カテゴリ検索の方がキーワード検索よりも最小ストローク数が少なかった。これは、キーワードを入力する必要がないからである。一方、WithAir では、キーワード検索でもカテゴリ検索と同程度の最小ストローク数であった。これは、キーワードナビ機能によって、効率よくキーワード入力が可能になったためである。カテゴリ検索では、カテゴリ体系を熟知していないと、必ずしも最小ストロークで所望のページにたどり着けないことを考慮すると、WithAir によるキーワード検索が最も操作コストが低いといえる。

7. おわりに

モバイルサーチエンジン WithAir を試作し、基本性能と利便性の評価を行った。その結果、i モードページ自動判別で適合率 99%の実用精度を達成し、130 万件のモバイルページを収集した。また、地域情報抽出で 93%の精度を達成し、37 万件の位置情報ページを抽出できた。さらに、従来サービスと比較して最も操作コストが低いことを確認した。今後は、実サービス[16]の上で、利用状況を測定しながら、評価・改良を進めていく。

参考文献

[1]日本インターネット協会監修、インターネット白書 2001、インプレス、2001

[2]OH!NEW?、http://ohnew.co.jp/i/

[3]Yahoo!モバイル、http://mobile.yahoo.co.jp/

[4]i-seek、http://iseek.infoseek.co.jp/

[5]iYappo、http://i.yappo.ne.jp/

[6]i モード 版 Google、http://www.google.com/imode

[7]松田ほか、文書タイプ分類による問題解決向き WWW 検索システムの開発と評価、情報処理学会研究報告 FI-53-2、1999 年

[8]Soumen Chakrabarti ほか、Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery、WWW8 Conference、2000 年

[9]喜田ほか、モバイル位置指向サーチエンジンの開発、第 61 回情報処理学会全国大会 1U-2、2000 年

[10]三浦ほか、モバイルインフォサーチ:移動環境下でのユーザ指向 WWW 検索、情報処理学会 MC 研究会、MC-3-7、2000 年

[11]高野ほか、サイテーションエンジン リンク解析を用いた WWW ランキングシステム、情報処理学会研究報告 DBS-120-2、2000 年

[12]赤峯ほか、モバイル指向 WWW サーチエンジン WithAir の開発(1) - システム構成 -、第 62 回情報処理学会全国大会 6W-8、2001 年

[13]Sergey Brin ほか、The Anatomy of a Large-Scale Hypertextual Web Search Engine、WWW7 Conference、1999 年

[14]zi corporation、http://www.zicorp.com/

[15]河合ほか、モバイル指向 WWW サーチエンジン WithAir の開発(2) - ナビゲーション機能 -、第 62 回情報処理学会全国大会 6W-9、2001 年

[16]BIGLOBE サーチ Attayo!ケイタイ、http://attayo.jp/i/