6W-8　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—

赤峯享　河合英紀　喜田弘司　松田勝志　福島俊一

NEC　インターネットシステム研究所

1. はじめに

近年、携帯電話からインターネットに接続するための環境が整備され、i モードユーザに代表されるモバイルインターネットユーザが急増している。これに伴い、携帯電話からアクセス可能なWEBページが日々増加し、数年後には、数百万ページになると予想される。

現状のモバイルコンテンツの検索サービスは、検索対象ページをページ作成者／編集者が登録する登録型が主流であり、規模が数百万ページになると、鮮度を保ち網羅的に収集することは困難である。今後は大規模ページを検索対象とした高精度のクローラ型検索エンジンが必須である。また、携帯電話は、文字入力が不便で、画面が小さいため、簡単に検索するためには、ユーザの目的に合致した優良サイトへ誘導するナビゲーション機能や、位置依存のモバイルページに簡単にアクセスできる機能も重要である。

本稿では、インターネット上の豊富なページを携帯電話から簡単に検索することを目的に開発したモバイル指向WWWサーチエンジン WihAir の概要と、その基盤となるクローラ型の全文検索機能について報告する。なお、本エンジンの他の特徴である先読みナビ機能は参考文献[1]で、位置依存検索機能は参考文献[2]で詳細を述べる。

2. 特長

本サーチエンジンは、以下の機能を特長とする。

(1)クローラ型全文検索機能
　インターネット上の豊富な情報をカバーし、高精度に検索する全文検索機能。

(2)先読みナビ機能
　ユーザの検索目的を先読みして提示することで、検索操作の手間を大きく軽減するナビゲーション機能。

(3)位置依存検索機能
　モバイル利用に合わせて、モバイルコンテンツを場所別に自動分類し、検索する機能。

3. システム構成

モバイルサーチエンジンのシステム構成を図1に示す。本システムは、モバイルページを自動収集する収集部と、収集したページを解析し、情報を抽出し、インデックスを作成する解析・登録部と、検索を行う検索部から成る。

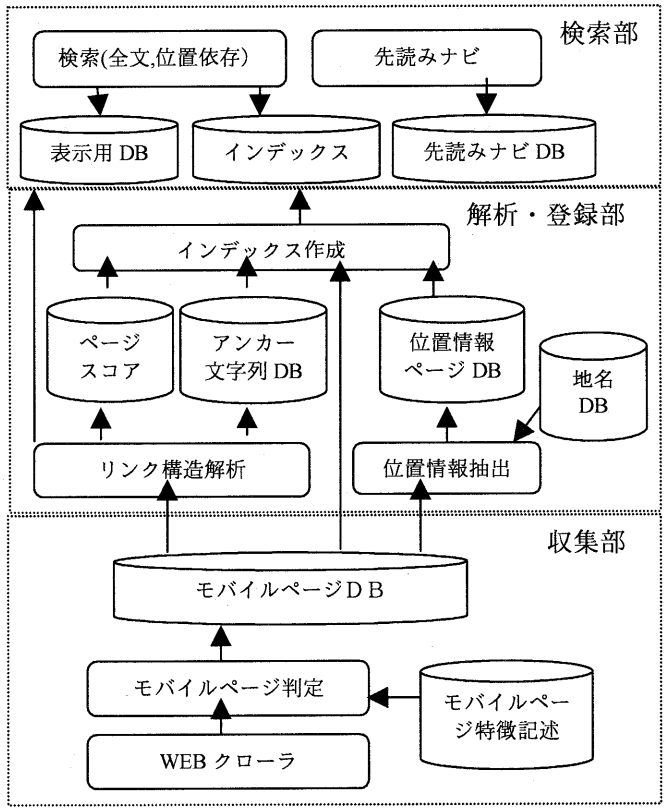

図1　システム構成

3.1　収集部

収集部は、WEBクローラによりページを自動収集し、携帯電話からアクセス可能なモバイルページを選別する。モバイルページの判別は、「ページタイプサーチ」[4]を用いて、ページサイズ、使用タグ、キーワード等の様々な情報から総合的に判断する[3]。モバイルページDBには、モバイルページのみを格納する。

3.2 解析・登録部

解析・登録部は、モバイルページから位置情報を抽出する位置情報抽出モジュール、ページ間のリンク構造を解析し、アンカー文字列・ページスコア・結果表示用の要約を作成するリンク構造解析モジュール、検索用インデックスを作成するインデックス作成モジュールから成る。位置情報抽出モジュールは、モバイルページのテキストを走査し、そのページがどの場所に関連しているのかを自動判別してインデックス付けする[2]。

Development of a mobile-oriented WWW search engine WithAir
(1) -system outline-
Susumu Akamine, Hideki Kawai, Koji Kida, Katsushi Matsuda and Toshikazu Fukushima
Internet Systems Research Laboratories,　NEC Corporation

3.3 検索部

検索部は、解析・登録部で作成した情報を用いて、キーワードに適合するページを全文検索／位置依存検索する検索モジュールと、ユーザの検索目的を先読みして優良サイトへナビゲートする先読みナビモジュール[1]から成る。

4. クローラ型全文検索機能

全文検索機能の特徴であるリンク構造を利用したランキング方式と検索結果の要約作成方式について説明する。

4.1 ランキング方式

携帯電話からの検索では、一度に表示できる件数が最大で5件程度と少ないため、実用的なサーチエンジンを構築するためには、高精度の検索が必須である。本サーチエンジンは、論文等を対象とした伝統的な検索エンジンで良く用いられるtf*idfを元にしたランキングではなく、他サイトからのリンク情報を元にランキングを行う。全文検索の対象として、WEBページ内の文字列だけでなく、リンク元ページのアンカー文字列も検索対象にし、アンカー文字列でヒットしたページを優先するように、以下の順でランキングする。

(1) リンク元ページのアンカー文字列がヒットしたページ。複数のページがヒットした場合は、ヒットしたリンク元サイトが多いページを優先する。なお、自ページのタイトルもアンカー文字列とする。

(2) WEBページ中の文字列でヒットしたページ。 複数のページがヒットした場合は、サイテーションエンジン[3]で計算したページスコアの高いページ(=多くの重要なページからリンクされたページ)を優先する。

例えば、図2に示すように、「Jリーグ」というキーワードで検索する場合、(a)「Jリーグ」をアンカー文字列とする2サイトからリンクされ、ページ内に「Jリーグ」という文字列を含まないページは、(b)「Jリーグ」をアンカー文字列とするページからリンクされず、ページ内に「Jリーグ」という文字列を多数含むページよりも上位にランキングされる。

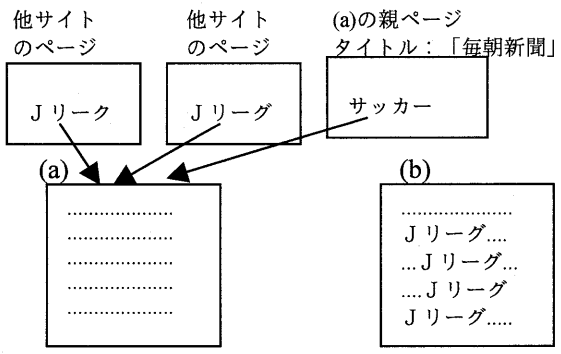

図2 アンカー文字列によるランキング／要約

4.2 検索結果の要約作成方式

画面の小さい携帯電話の場合、ユーザが複数の検索結果を一画面で比較して、アクセスするページを決定しようとすると、検索結果としては、タイトル程度しか表示することができない。ページ作成者が記述するページのタイトルは、通常のブラウザには表示されないため気にかけられないことが多く、タイトルを記述していないページや、検索結果として表示しても意味をなさないタイトル(例えば、「新しいページ」というタイトル)のページが多数存在する。また、そのページのタイトルとしては適していても、他のサイトに同一のタイトルのページが存在するため、検索結果のタイトルとしては不適切なものも多い。例えば、「Jリーグ」というキーワードで検索した場合に、上位の検索結果のタイトルが全て「Jリーグ」として表示されてしまうと、検索ユーザはどのページにアクセスすべきか判断することができない。

本サーチエンジンでは、リンク元ページのアンカー文字列(自ページのタイトルを含む)から表示用のページ要約を以下の方式で自動的に作成する。

(1)被リンクページのアンカー文字列の中で最も多くのページに共通に出現する文字列を選択し、それをそのページのタイトルとする。

(2)自ページのタイトルと、親ページ(もしくはサイトのトップページ)のタイトルを合わせて、そのページの要約とする。

例えば、図2の(a)のページ場合、ページのタイトルが「新しいページ」という記述であっても、「Jリーグ」というアンカー文字列とする2ページからリンクされているため、このページのタイトルは「Jリーグ」になる。さらに、親ページのタイトルが「毎朝新聞」であるので、このページの検索結果としての表示は、「Jリーグ←毎朝新聞」となる。

5. おわりに

インターネット上の豊富な情報を携帯電話から簡単に検索するためのサーチエンジン WithAir のシステム構成について報告した。また、全文検索機能のランキング方式と検索結果の要約作成方式について説明した。今後、ランキング方式と要約作成方式の評価・改良を行い、実用化を進めていく。

参考文献

[1]河合ほか、モバイル指向WWWサーチエンジンの開発(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

[2]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

[3]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検索ランキングシステム、情報処理学会研究報告 DBS-120-2、2000年

[4]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