# 被参照インデックスページによる情報探索支援方式

4P－7

和田 義毅　新井 克也
NTT情報流通プラットフォーム研究所

## 1．はじめに

WWWでの検索では、検索式による条件指定だけでは十分な精度の検索結果が得られないことが多い。このため、検索の過程において、多くの検索者が検索結果のＷｅｂページをエントリポイントとして自力で情報探索を行い、必要な情報を収集している。

この情報探索の過程で、探索者はリンクテキストや画像、その他の文章など、Ｗｅｂページ内にある情報を用いて目的の情報の存在場所を推定している。この時に利用される情報の中で特に有効なのは、リンク集のようなインデックス情報である。

インデックス情報を持つページの多くは、同じトピックに関連する複数のサイトへのリンクリストを持っている（以下このようなページをインデックスページと呼ぶ）。このため、インデックスページを利用すれば、同じトピックに関連する情報を効率よく収集することができる。

しかし、現状では、探索者が目的に合致するインデックスページを探索することは、容易ではない。本研究では、探索者の情報探索を支援するために、探索作業の任意の時点でインデックスページを推薦できる、新しい情報探索支援方式を提案する。

## 2．既存のインデックスページ提示方法

本研究と同様に、ユーザーにインデックス的なページを提供するためのシステムとしては、図1のように、キーワード検索の際にページタイプとしてリンク集を選択することにより、キーワードを含むリンク集ページを提示するものがある[1][2]。

この方式では、リンク集など幾つかのページタイプの構造的特徴ルールを記述した特徴記述を用意し、検索者が入力した検索キーワードによる検索結果の中で、ページタイプの特徴記述に一致するページだけを提示している。

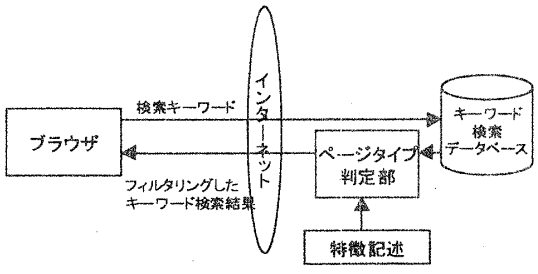


図1　従来手法のシステム構成

このため、得られる検索結果はキーワード検索結果の部分集合であり、キーワード検索以上の情報を入手できるわけではない。また、検索条件をキーワードで適切に指定できない場合には、不要な情報が多数含まれたり、必要な情報が含まれていなかったりする点もキーワード検索と同じである。

従って、このような方式でも、検索条件の指定が困難な場合には、情報探索は必要な作業である。しかし、現状では、情報探索に入った後に、インデックスページを探索するのを支援する情報がない。このため、探索者は大きな労力をかけて自らインデックスページを探し、情報を収集しなければならない。

## 3．被参照インデックスページ提示手法

本研究では、上記のような問題を解決するために、探索者が参照しているＷｅｂページにリンクを張っているインデックスページ（被参照インデックスページ）を提示する方式を提案する。

インデックスページは、同じトピックに関連する情報を持った、複数サイトへのリンクリストである。従って、参照中のページにリンクを張っているインデックスページは、参照中のページと関連した情報を持つサイトへのリンクを持っている可能性が高い。

この方式では、目的に関連する情報を持ったページを発見すると、そこから被参照インデックスページを介して、関連する情報を持った他のサイトにアクセスすることができる。

そのため、キーワード検索で十分な検索条件が指定できない場合でも、探索によって関連情報を発見できれば、被参照インデックスページを介して関連情報にアク

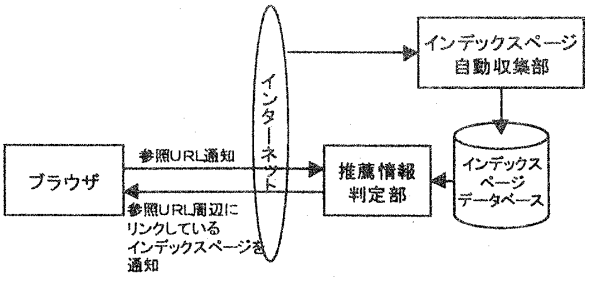


図2　提案法のシステム構成

Information Groping Support Using Refered Index Page
Yoshinori WADA and Katsuya ARAI
NTT Information Sharing Platform Laboratories

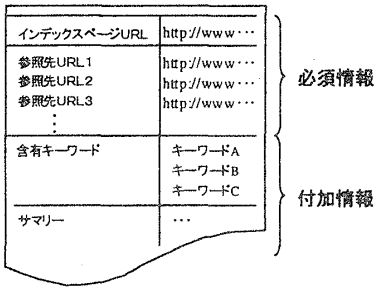


図３ インデックスページデータベースへの登録情報

セスし、効率よく情報を収集することができる。

## ４．システム構成

提案法を実現するためには、まず、探索者に提示するインデックスページを事前に収集しておく必要がある。８億ページものＷｅｂページが存在すると言われるＷＷＷの規模を考えると、この作業を人手で行うのは困難であり、自動的に収集する方法が必要となる。

自動収集の方法としては、文献１のような、文書タイプの特徴記述に基づく汎用的な方法もあるが、本方式では、インデックスページの抽出に特化したインデックスページ専用の判定関数を作成した。なお、この判定関数については次章で説明する。

上記の判定関数で集めたインデックスページは、図２のインデックスページデータベースに図３の形式で蓄積される。この時蓄積される情報としては、インデックスページのＵＲＬとインデックスページの参照先ＵＲＬが最低限必要である。実際にはスコアリングのための情報や補助的情報として、含有キーワードやサマリーなど幾つかの情報も併せて記憶する。

次に、インデックスページデータベースの情報を探索者に提示するための方法について述べる。まず、探索者は、キーワード検索やディレクトリなど任意の手段を使って情報を探索する。

探索の過程で目的に関連するＷｅｂページを発見した場合には、ブラウザに追加実装されたインデックスページリクエスト用インターフェースを用いて、参照しているページのＵＲＬを図２の推薦情報判定部に通知する。

推薦情報判定部では、インデックスページデータベースを検索し、参照先ＵＲＬの中に探索者が通知したＵＲＬを含むインデックスページを検索し、被参照インデックスページのリストを作成する。なお、多くのリンクはサイトのトップやトピックのトップページに対して張られているため、検索時には、参照ＵＲＬの上位ディレクトリを参照しているインデックスページも検索する。

推薦情報判定部では、各被参照インデックスページに対し、参照ＵＲＬ数、含有キーワード、インデックスページの参照位置などによるスコアリングを行い、スコアが上位のインデックスページを探索者に提示する。

これにより、本方式では、探索中任意の時点で、探索者が参照中のＷｅｂページにリンクを張っている被参照インデックスページを提供することができる。このため、キーワード検索で十分な検索精度が得られなかった場合でも、その後の探索行動により目的に関連する情報を発見できれば、そこを起点として関連する情報を持ったサイトに効率よくアクセスできる。

表１ インデックスページ判定関数による判定結果

| サンプル | 判定結果（%） | |
|---|---|---|
| | インデックスページ | 通常ページ |
| インデックスページ | 96．8 | 3．2 |
| 通常ページ | 0．4 | 99．6 |

## ５．インデックスページの自動判定方式

提案法のシステムを機能させるためには、探索者に提示するインデックスページが事前に収集されている必要がある。そこで、インデックスページを自動収集するためのインデックスページ判定関数を作成した。具体的には、ドメイン外へのリンク数及び比率を基本判定基準とし、バナー広告など誤判定要因を排除するための補正処理を加えた判定関数を用いた。

上記の判定関数を用いて、ＷＷＷから収集し、人手で分類したインデックスページ、通常ページ各１０００ページのサンプルに対して判定実験を行った。この結果、表１に示すように、９６．８％のインデックスページが抽出された。また、この時誤抽出される通常ページは０．４％であった。

この結果より、機械的な判定でもインデックスページは十分収集可能であり、インデックスページデータベースは自動的に作成可能であるという見通しが得られた。

## ６．まとめ

情報探索者を支援するための新しい方式として、探索者が参照中のＷｅｂページにリンクしている被参照インデックスページを推薦する方式を提案した。

また、インデックスページ判定関数を用いたインデックスページ自動収集部と、探索者から通知された URL を元に推薦インデックスページを決定する推薦情報判定部による、本方式の一実現形態のシステム構成を紹介した。

今後は、このシステムを実装し、被参照インデックスページによる情報探索支援の有効性について検証を行っていく予定である。

## 文献

[1] 松田、福島：インターネット多角的検索システムＯＴＲＯＳ－構造的特徴量によるタイプ分類と検索，情報処理学会第 57 回全国大会予稿集(3)，PP. 145-146, 1998

[2] http://netplaza.biglobe.ne.jp/keyword2.html