DEWS2006 3B-i7

利用者のリンク選択を考慮したWebページ検索手法

内山　紀明† 　鈴木　優†† 　川越　恭二†† 　西村　俊和††

† 立命館大学 理工学部 情報学科　〒 525-8577 滋賀県草津市野路東 1-1-1
†† 立命館大学 情報理工学部　〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: †{uchiyama,suzuki,kawagoe}@coms.ics.ritsumei.ac.jp, ††tnt@is.ritsumei.ac.jp

あらまし　本稿では，利用者のリンク選択を Web 検索システムへの問合せ手法の一つとして考えることにより，利用者にとって有用な Web ページを半自動的に検索し，利用者に提示する方法について述べる．現在，多くの Web 検索システムでは，問合せとして利用者が入力したキーワードを想定している．ところが，利用者の検索要求は一般に曖昧であり，正確にキーワードとして表現することが困難である場面が多い．一方，利用者は Web ページの閲覧を行う際，利用者の検索意図に応じてリンクを選択することが多いと考えられる．つまり，リンクの選択は利用者の検索行動の一つであると考えることができる．そこで本稿では，利用者がリンク選択を行った際，リンク元周辺のテキスト，リンク先の Web ページの文書内容の両方から利用者の検索意図を抽出し，検索システムへの問合せとして利用する方法について提案を行う．また，提案手法を用いた検索システムを作成し，キーワード検索システムとの比較のために 11 点平均適合率を用いた評価実験を行った．その結果，提案手法を用いることにより，利用者のリンク選択から検索意図を抽出し，有用な Web ページを提示できることが確認できた．

キーワード　情報検索，Web とインターネット，問合せ処理

A Web Page Retrieval Method Based on the Users' Link Selection

Noriaki UCHIYAMA†, Yu SUZUKI††, Kyoji KAWAGOE††, and Toshikazu NISHIMURA††

† Department of Infomation, College of Science and Engineering, Ritsumeikan University.
Nojihigashi 1-1-1, Kusastu, Shiga, 525-8577 Japan
†† College of Information Science and Engineering, Ritsumeikan University.
Nojihigashi 1-1-1, Kusastu, Shiga, 525-8577 Japan
E-mail: †{uchiyama,suzuki,kawagoe}@coms.ics.ritsumei.ac.jp, ††tnt@is.ritsumei.ac.jp

Abstract　In this paper, we propose a novel method of suggesting useful Web pages, based on the links which users selected. Recently, search engines commonly used in the Web search focus on keywords for specifying characteristic of users' queries. However, if users' interest are vague, the users cannot express the users' interests as keywords, then the users cannot get useful Web pages. On the other hand, when users search Web pages without using Web search engines, the users click the links depends on the users' interests. Therefore, we assume that the achievement of which links users select should also be considered as the queries of users. In this paper, we propose a method of generating users' queries by users' behavior of clicking links.

Key words　Infomation Retrieval, Query Processing, WWW

1. はじめに

インターネットを用いて有用な情報を得るために，利用者はあるページから別のページへのリンクを選ぶ行動と，リンク選択により得られたページの内容を精査する行動を繰り返す．ここで，ページとは利用者が情報を得るために閲覧する文書であり，HTML 文書に代表される Web ページである．次に，利用者は選択したリンク先のページを精査し，有用な情報が得られなければ，一つ前のページに戻ってリンクを選択し直したり，または，新たにキーワード検索による絞込みを行うことによって有用な情報を得ようとする．ここで，リンクを選択する行動によって新しいページを表示し，リンク先のページの内容を精査する行動により，ページに存在する様々な情報から必要な情報の取捨選択を行うことができる．よって，これらの行動は利用者が情報を得るために有効であると考えられる．しかし，一つ前のページに戻る行動では，一度精査しているページが再度

表示されるため，新たな情報を得ることができない．また，新たなキーワードによる検索行動においては，検索結果ページには適合候補ページへのリンクリストがページ内容の中心として表示されるため，検索結果ページを精査することで得られる情報は，キーワード検索システムによっては該当キーワード周辺の文章や，作成日などであり，必要な情報の全てを得られることは少ない．よって，これらの行動は利用者が有用な情報を得るために直接必要な行動とはいえず，補助的な行動と考えられる．全体の行動に対する補助的な行動の割合が増えると，効率的な情報収集が困難となる．

また，利用者がページを閲覧する際，リンクによりつながったページを移動することで，利用者にとって有用な情報を取得しているという点に着目すると，リンクを選択することでページの移動が発生するが，利用者は移動した先のページに自身の必要な情報があると予測してリンクを選択する．それはすなわち，キーワード検索の際に，必要な情報の存在するページの内容を予測してキーワードを選択する行為と同等であると考えられる．つまり，利用者が検索意図をキーワードとして表現することとリンクを選択することは，同等の検索意図を持った検索行動であると見なすことができる．

そこで本稿では，利用者の検索意図を利用者自身の作成するキーワードではなく，ページ閲覧時の基本的な行動であるリンクの選択から抽出し，そこからキーワードを自動的に選択し問合せとして利用する検索手法の提案を行う．これにより，従来の検索手法では，ページ閲覧により生じた検索要求を利用者が検索意図として表現する必要があるところを，提案手法では，そのページに移動するためのリンク選択からキーワードを抽出し，それらを利用して検索意図を表現することが可能になる．その結果，利用者はページを閲覧した後に生じた検索要求を検索意図として表現して検索を行うことなく，ページの閲覧と同時に利用者にとって有用なページの候補を取得することで，効率的な情報収集が可能になる．

2. 従 来 手 法

従来，利用者がインターネットにおける情報収集に用いる一般的な方法として，キーワードによるページ検索システムと，ページを入力とした類似ページ検索がある．本章では，これらキーワード検索と類似ページ検索の特徴を挙げ，その問題点について述べる．

2.1 キーワード検索

キーワード検索では，利用者が入力したキーワードに対して，事前に収集した全世界のページの中から，キーワードを含むページへのリンクを順位付リストとして表示する．そのため，検索結果へのリンクが大量に表示される場合が多く，利用者は検索結果ページを一つ一つ精査し，有用な情報が書かれたページを探索する必要がある．また，利用者はキーワード検索の際，自分の必要とする情報が記述されたページ内容を予測し，そのページに存在する的確なキーワードの組合せを作成する必要がある．なぜなら，インターネットには単純な単一のキーワードを含むページが大量に存在するためである．そのため，利用者の必要な情報から連想するキーワードが曖昧である場合，そのキーワードが存在するページ群には必要な情報内容と異なったページも含め，様々なページが同一の検索結果として表示される．

2.2 類似ページ検索

類似ページ検索とは，検索時にキーワードを入力するのではなく，単一のページ内容を基に，含まれる単語やその意味などを入力として内容の類似したページを検索する手法である．しかし，この手法では，検索目的となる情報に類似したページをあらかじめ用意しなければならない．つまり，類似したページを検索するために別途キーワード検索を行う必要が生じるため，2.1 節のキーワード検索における問題点も同時に存在する．

また，ブログと呼ばれる出来事・趣味などに関し自分の意見を日記形式のコンテンツとして書き込むことのできるページや，世の中の出来事をまとめたニュースサイトと呼ばれるページなど，1 ページに記載される話題が多岐にわたるページが増加している．このようなページを入力として検索を行った場合，ページ内容から抽出した大量のキーワードを用いて他のページとの類似度を算出するため，入力ページ内の利用者が重要と考える部分に類似する検索結果ページが出力されないことがある．また，出力されても入力ページのどの部分に類似したかによって検索目的のページの内容とかけ離れたページが出力されることがある．さらに，出力された場合，検索結果のページ数は大量になることが多く，検索精度が低下する．また，検索結果をさらに絞り込むための検索作業が必要になる点も問題である．

2.3 リンクに関連する研究

以下に，インターネットにおけるリンクに関連する研究を示し，提案手法との差異について述べる．

2.3.1 リンク情報を考慮した Web 検索システム

インターネットにおけるリンクの情報から Web 検索を行う手法として，大森ら [1] は，Web ページに使用されている HTML タグの統計から，リンクに対して前方と後方のページに含まれるキーワードの重みを考慮した検索データベースを提案している．大森らは，評価実験を通して検索対象ページ群からリンクしている前方ページ情報を考慮した検索データベースと，検索対象ページ群へリンクしている後方ページ情報を考慮した検索データベースを比較している．この研究は，リンクを中心とした前後のページを利用している部分で提案手法と類似しているが，この研究では，あるページに存在する全てのリンク先と，そのページにリンクする全てのページを考慮して重み付けを行っているのに対し，提案手法では，利用者が実際に選択したリンクの前後ページのみを考慮している部分が異なる点である．

2.3.2 利用者選好の半順序性に着目した Web 探索とナビゲーションの個別化

山口ら [2] は，利用者の嗜好を反映したページ検索を支援するために，複数のリンクから一つを選択するといった順序による評価を用いて利用者の選好を半順序グラフとして表現した上で，閲覧中ページに含まれるリンク集合のランキング手法を提案している．この研究では，利用者の明示的な選好入力によらず，ページ閲覧行為から半順序グラフをリアルタイムに作成し，

これによって利用者の選好に合致したページを推薦する手法について述べている．この中で，ページのリンクを辿ることで利用者の検索意図を表すページの提示が可能であるという部分で提案手法と類似しているが，この研究では，リンク選択履歴を利用者の選好として保存し，そのつながりを用いて利用者の嗜好を判断している．よって，利用者の嗜好に合致するページを推薦するためには，ある程度利用者のリンク選択を学習させる必要がある．これに対し，提案手法では，リンク選択の履歴に頼らず，一度のリンク選択から利用者の検索意図を抽出している部分が異なる点である．

2.4 提案手法における既存の問題の解決

従来手法の問題点として，利用者の検索意図をキーワード作成や類似したページといった単一の情報により表現している点が挙げられる．利用者の検索意図は一般に曖昧であり，それをキーワードや類似したページによって表現するためには慣れを必要とする．そのため，インターネット利用における利用者の習熟度によって検索精度が変化する．

そこで，本稿ではリンク選択から検索意図を抽出し，そこから問合せを自動的に作成するために，選択されるリンクの周辺情報と，リンクの周辺情報に対応するリンク先ページ内のキーワードを利用する．ここで，リンクの周辺情報とは，リンクを形成する文字列そのものや，リンクを含む文章の前後一定範囲の文章に含まれるキーワードを指し，利用者がそのリンクを選択するための判断材料となる情報である．リンク周辺情報と，それに対応するリンク先ページ内のキーワードを利用することにより，ページ閲覧とリンク選択というインターネット利用者の自然な行動から検索意図を抽出し，利用者によるキーワード作成や類似ページ選択という検索意図の明確な表現を必要とせずに検索システムへの問合せを作成することが可能になる．また，問合せ作成にリンク周辺情報のみを用いる場合に比べ，リンク先ページ内のキーワードを考慮することによって，リンク周辺情報に含まれる不要なキーワードを除外することが可能になり，検索精度を向上させることが可能になる．さらに，リンク先ページ内のキーワードのみを用いた類似ページ検索に比べ，リンク周辺情報を考慮することによって，リンク先ページ内の複数の話題の中から，利用者がそのページに期待する部分を重視した検索を行うことが可能になる．

3. 提 案 手 法

本章では，まず 3.1 節においてインターネットにおけるページ間のリンク接続構造について考察し，リンクによってつながったページ同士の関連性について述べる．次に，3.2 節において利用者の検索意図を定義し，リンク選択によるインターネット利用が利用者のどのような意図をあらわしているかについて述べる．さらに，3.3 節で本稿における問合せ作成手法について述べる．また，3.4 節で 3.3 節の手法の拡張方法について述べる．

3.1 インターネットにおけるリンク構造

インターネットに代表される World Wide Web では，様々な情報を持つ HTML ファイルなどのページと，そのページ同士を繋ぐリンクによって形成されており，リンク先のページは同一サイト内のページの場合や，別サイトのページの場合がある．

ここで，ページ作成者がページにリンクを設置する場合，作成者はリンク先のページ内容を把握していると考えられる．そのため，ページ作成者はリンク先ページの内容について，リンクの前後の文章やリンクを形成する文字列自体を用いてリンク先のページ内容について記述する．反対に，リンク先のページではリンク元のページ内容を把握していない場合が多い．なぜなら，一般にリンクされた側がリンクされたことを知る手段がなく，リンク元のページ内容を知ることができないためである．また，ページ作成段階で他のサイトからリンクされると分かっていることは少なく，それを考慮に入れたページ作成を行うことは少ない．そのため，リンク先のページからリンク元のページに対しては関連性は低いが，リンク元ページのリンク周辺の文章からリンク先のページに対しての関連性が高いと考えることができる．

3.2 利用者の検索意図

利用者はインターネットを用いて自分の必要とする情報を取得する．この際，何について知りたいのか，どのような情報を得たいと考えているのかという検索意図が存在する．

利用者はインターネットを利用する際，一つのページの内容を精査し，次にそのページに存在するリンクを一つ選択する．さらにリンクされているページでも同様の内容精査とリンク選択を行う．これらの行動の繰り返しがインターネットにおける利用者の基本的なページ閲覧行動と考えられる．ここで本稿では，利用者が内容精査とリンク選択を行ったページを移動元ページ，リンク選択により移動した先のページを移動先ページと定義する．ページ閲覧行動時，利用者は移動先ページに自分の目的とする情報があると期待して移動元ページ内のリンクを選択する．さらに，3.1 節で述べたようにリンク周辺の文章内容とリンク先ページの内容には関連性があると考えられるので，利用者はリンク先ページが必要な情報の存在するページか否かの判断基準として，選択するリンク周辺の文章内容を用いていると考えることができる．つまり，リンク選択は，利用者がリンク周辺の文章内容と自身の検索意図との適合評価を行い，適合する時そのリンクを選ぶという情報検索を行っていることと同等であると考えられる．

3.3 リンク選択とリンク先ページ内容を用いた問合せ作成手法

利用者は移動先のページ内容に対して，移動元のページのリンク周辺情報に沿った内容を期待している．そこで，リンク周辺の文章に含まれるキーワードと，移動先ページの文章に含まれるキーワードの中から移動元ページのキーワードが存在する位置周辺のキーワードを用いて利用者の検索意図を抽出する．なお，提案手法の概要を図 1 に示す．また，抽出したキーワード集合の包含関係ベン図を図 2 に示す．

まず，利用者がページ内の一つのリンクを選択したとき，そのリンクを含む文章とその前後の文章から，内容を表すキーワードを抽出する．このとき抽出したキーワードの集合を A とし，図 2 では実線 A で囲まれた部分を指す．次に，選択され

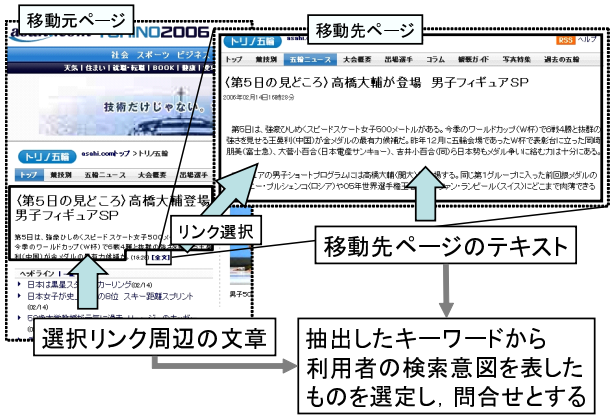

図 1　提案手法の概略

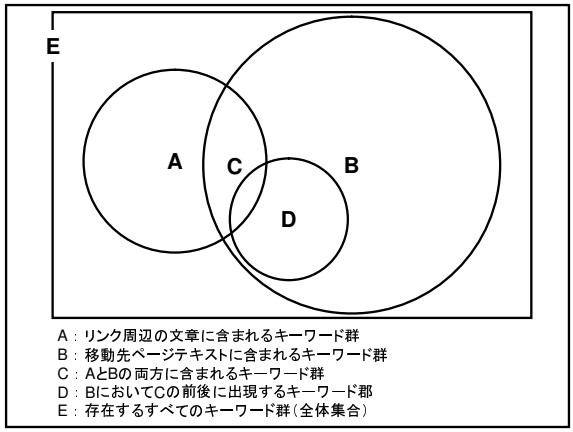

図 2　抽出キーワード集合の包含関係ベン図

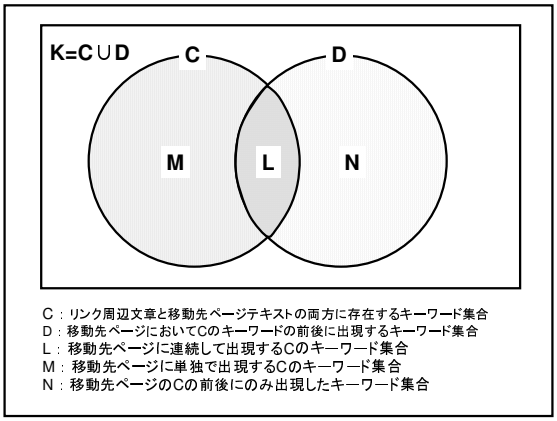

図 3　キーワード集合 K におけるキーワード関係ベン図

たリンクによる移動先ページからページ本文から HTML タグなどページ内容に関係ないものを取り除き，テキストのみを抜き出す．このテキストから同様に内容を表すキーワードを抽出し，キーワードの集合を B とする．図 2 では，このキーワード集合は実線 B で囲まれた部分を指す．ここで，キーワード集合 A と B に共通して存在するキーワード集合を C とすると，

$$C = A \cap B$$

と表すことができる．さらに，キーワード集合 B において C のキーワードの前後に出現するキーワード集合を D とすると，D のキーワードは対応する C のキーワードと同一の文章内にある可能性が高く，その文章の特徴を表すキーワードであるといえる．そのため，C と D を合わせたキーワード集合を K とすると，K は

$$K = C \cup D$$

と表すことができ，キーワード集合 K は移動先ページ内容の中でも利用者がそのページに期待した内容を表すキーワードで構成されていると考えられる．よって，このキーワード群が利用者のリンク選択とページ移動から抽出された検索意図を表していると考えられる．

次に，キーワード集合 K 内のキーワードを利用者の検索意図を表している順になるよう順位付けを行う．まず，それぞれのキーワードの出現位置の特性に沿ってキーワード集合 K を 3 個に分割する．分割した集合を L, M, N とし，それぞれの範囲を C, D で表したものを以下に示す．また，そのベン図を図 3 に示す．

$$L = C \cap D$$
$$M = C \cap \overline{D}$$
$$N = D \cap \overline{C}$$

キーワード集合 L は選択されたリンク周辺の文章と移動先ページテキストの両方に存在するキーワードであり，かつ移動先ページで任意の 2 個以上が一定の範囲内に共通して出現しているキーワードの集合である．つまり，利用者にとって移動先ページに必要な情報が存在するか否か判断の材料となる複数のキーワードが，移動先ページで連続する位置に出現しており，選択されたリンクの周辺文章が現す内容と，移動先ページ内の L のキーワードが複数含まれる文章の内容が同一である可能性が高い．そのため，キーワード集合 L に含まれるキーワードが利用者の検索意図を強く反映していると考えられる．

次に，キーワード集合 M は選択されたリンク周辺の文章と移動先ページテキストの両方に存在するキーワードであるが，移動先ページでは，前後に同様のキーワードが存在せず，単独で出現しているキーワードの集合である．よって，キーワード集合 M に含まれるキーワードは，利用者の検索意図を反映している可能性が高いが，移動先ページの内容と関連の薄い場所に存在している可能性も考えられる．

最後に，キーワード集合 N は選択されたリンク周辺の文章には出現しないが，移動先ページでその前後に出現したキーワードの集合である．よって，キーワード集合 N に含まれるキーワードは利用者がリンク選択時に移動先ページ内容を予測する際の材料としておらず，リンク選択における検索意図を表しているとはいえない．

これらのキーワード集合 L, M, N から問合せを作成するために，本提案手法ではキーワード集合 L を用いる．ここで，以降このキーワード集合 L に含まれるキーワードを意図表現キーワードと呼ぶことにする．意図表現キーワードの内，出現回数の多いものから順に適切な個数を問合せキーワードとして用い，問合せキーワードをすべて含むページを検索結果として利

用者に提示する．ここで出現回数とは，L に含まれるキーワードの移動元ページのリンク周辺文章と移動先ページテキストにおける出現回数の内，意図表現キーワードの前後一定距離に他の意図表現キーワードが出現した回数とする．この出現回数の定義により，抽出されたキーワードの文書全体の出現回数によらず，キーワードの前後関係により判断できる文書内容に基づいたキーワードの順位付けが可能になる．これは，ページに大量に出現するが，文章内容を表すことの少ない，ものを数える単位や月日を表す単位といった一般名詞などが，1 度でも連続して出現することで意図表現キーワード内の上位に順位付けされないための定義方法である．

3.4 問合せキーワード数による動的なキーワード抽出

3.3 節で，ページにリンクが存在する場合，移動元ページのリンク周辺の文章，および移動先ページの文章から，キーワードを抽出し，移動元ページのリンクと移動先ページに共通して出現し，かつ移動先ページで一定距離内に複数出現するものを問合せとして利用する手法を提案した．しかし，この手法のリンク周辺の文章の範囲のとり方，および移動先ページにおける意図表現キーワード同士の距離のとり方によって，作成された問合せによる検索結果は変わり，それに伴って検索精度にも変化が生じると考えられる．

そこで，問合せキーワード数を先に決定し，リンク周辺情報を含む文章の範囲，および意図表現キーワード同士の距離を，利用者がリンクを選択した時点で動的に決定する手法を提案する．これにより，リンク周辺の文章量や移動先ページの文章量の多寡によるキーワード数の変化を防ぎ，結果として検索精度のばらつきを抑えることができる．ここで，問合せキーワードに n 個の意思表現キーワードを用いるとすると，移動先ページ内の全キーワードの出現場所の中で，ある意思表現キーワードの前後何キーワードまでの意思表現キーワードをキーワード集合 L に含めるかというキーワード数を s，選択されたリンクを含む文章と，その前後いくつの文章をリンク周辺情報の抽出対象とするかという値を r とすると，リンク周辺情報を抽出する範囲と移動先ページにおける意思表現キーワード同士の距離を動的に決定する手順を以下に示す．

[**手順 1**] $s = 1$，$r = 1$ として意図表現キーワードの抽出を行う．

[**手順 2**] 出現回数により順位付けを行い，結果として得られた問合せキーワード数が n 以上の場合は終了する．n 未満ならば**手順 3** へ進む．

[**手順 3**] r を固定した上で，s に 1 加算して順位付けを計算し直す．

[**手順 4**] 問合せキーワード数が n 以上の場合は終了する．n 未満かつ $s \leqq 3$ ならば**手順 3** へ戻る．それ以外ならば**手順 5** へ進む．

[**手順 5**] $s = 1$ に設定し，r に 1 加算して意図表現キーワードを抽出し直して**手順 2** へ戻る．

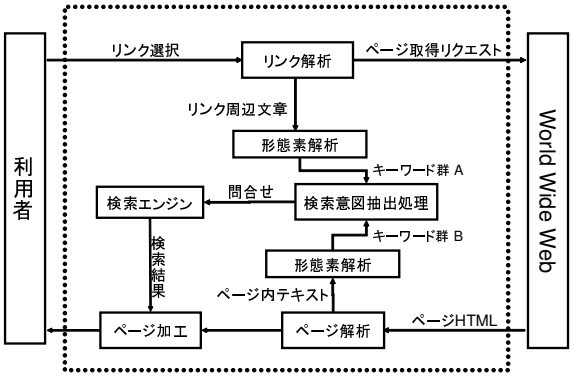

図 4 システム概要

4. リンク選択を問合せとする Web ページ検索システム

3.3 節で述べた利用者のリンク選択と移動先ページ内容から抽出した検索意図を用いた Web ページ検索を実現するために，利用者のページ閲覧リクエストとリンク選択を中継する Web システムを作成した．このシステムでは，中継過程においてリンク周辺のキーワードと移動先ページのキーワードを抽出し，重み付けを行った上で問合せを作成する．さらに，この問合せを用いて Web ページ検索を行った結果を移動先ページと共に提示する．以下に，4.1 節においてシステムの概要を述べる．次に，4.2 節でシステムの動作手順について述べる．

4.1 システム概要

本システムの概要を図 4 に示す．また，本システムの外観を図 5 に示す．本システムは，利用者と World Wide Web の間で動作し，利用者のリンク選択によるページ取得リクエストを受け取り，リクエストをオリジナル Web ページに転送すると同時に，選択されたリンク周辺の文章からキーワードを抽出する．次に，オリジナル Web ページから移動先ページの HTML を受け取り，ページテキストのみ抜き出す．さらに，抜き出されたページテキストからキーワードを抽出する．そして，リンク周辺文章から得られたキーワードと，移動先ページテキストより得られたキーワードから，キーワードの出現頻度と 3.3 節で述べた手法により意図表現キーワードの抽出を行い，その中から問合せキーワードを作成してキーワード検索を行う．さらに，その検索結果のリンク一覧を移動先ページと共に利用者に送信する．

利用者は，閲覧にブラウザを用い，システムによって縦方向に 2 分割表示された画面を見て操作する．縦方向に分割された画面の内，上側には閲覧ページへ直接移動する際の URL をシステムに入力する欄を設ける．下側には URL 入力やリンク選択による移動先のページと共に，本システムにより提示された Web ページ検索結果をリンク一覧として表示する．なお，この Web ページ検索結果はマウスでドラッグすることでブラウザ内の自由な位置に移動させることが可能であり，ページの閲覧に支障をきたすことはない．

図 5　システム外観

4.2　システム内部処理手順

図 4 に示した本システムの内部処理手順を以下に示す．

- **Step1 リンク周辺情報抽出**

利用者のリンク選択を受け取り，ページ取得リクエストを移動先ページを持つ Web サーバに対して送信すると同時に，リンク周辺情報を抽出するため，移動元ページの選択されるリンク周辺の文章を 3.4 節の手法を用いて抜き出し，解析する．なお，本システムでは，Web ページは HTML を用いて書かれていると想定しているので，HTML のブロック要素 1 個をページにおける文章 1 文とする．これは，一般的にブロック要素がページの見た目上の改行を伴い，ブロック要素内の文章内容が一つの意味を表していると考えられるためである. [3]

本システムでは，リンク周辺文字列からキーワードを抽出するために形態素解析を用い，形態素解析器として MeCab [4] を用いる．

形態素解析によって抽出するキーワードは，一般名詞・固有名詞・未知語とする．未知語とは，茶筌では解析できない半角英数字や辞書にない単語である．しかし，未知語にはページの特徴を表すキーワードが含まれているため採用した. [5] リンク周辺の文章から，1 文字で構成される語，数字・記号のみで構成される語，あらかじめ設定しておいた不要語を除いたキーワードを抽出する．

- **Step2 移動先ページ内容抽出**

Web サーバから移動先ページの HTML を受け取り，そこから移動先ページの内容を表すキーワードを抽出するために，ページ本文から HTML タグや JavaScript，コメントなどを取り除き，テキストのみを抜き出す．なぜなら，一般に HTML タグはページを形成する論理的な意味合いを表しており，JavaScript はページ内でのコンテンツの動作を定義する．さらに，コメントは利用者に対して表示されることはなく，これらはページ内容を表しているとはいえないためである．抜き出したテキストから，**Step1** と同様に形態素解析を用いて一般名詞・固有名詞・未知語からなるページ内容を表すキーワードをページ全体の文章から抽出する．

- **Step3 問合せの作成**

Step1，**Step2** により抽出したキーワード集合をそれぞれ A, B として，3.3 節で述べた手法により利用者の検索意図を表す 3 個の問合せキーワード $keyword_1$, $keyword_2$, $keyword_3$ を得る．なお，この問合せキーワードの個数は 5.1 節の実験により決定した．次に，このキーワードをすべて含むページを検索するために，問合せを以下のように作成する．

$$keyword_1 \text{ AND } keyword_2 \text{ AND } keyword_3$$

この問合せをキーワード検索エンジンに送信する．本システムではキーワード検索エンジンに Yahoo! JAPAN の Web ページ検索システムを用いた．

- **Step4 システムによるページ HTML 加工**

本システムでは，利用者に表示するページの HTML に対して行う加工内容を以下に示し，それぞれの方法を述べる．

（1） リンクを形成する URL 要素を，Web サーバへ直接アクセスするための URL ではなく，リンク URL を入力変数としたシステムへアクセスするための URL に変更する．その結果，閲覧したどのページのリンクを選択しても常にシステムが利用者とページの中継を行うことにより，全てのリンク選択を利用者の検索行動とすることができる．

まず，**Step1** により送信されたページ取得リクエストによって，Web サーバから受け取ったページの HTML からリンクを構成する HTML 要素を抜き出す．リンクを構成する HTML 要素には<a>要素を用い，移動先ページの URL には href 属性の値を用いる．次に，そのリンクが選択されたときに href 属性値をシステムの入力変数とし，システムにアクセスするようにリンク要素を置き換える．本システムでは，ページに HTML ファイルを想定しているので，href 属性値の取りうる値の中でも拡張子が.htm または.html である要素と，URL が/（スラッシュ）で終わる要素に対してのみ加工を行う．また，実際には移動先ページ URL を表す方式には，http://から始まりページのファイル名まで完全に記述する絶対パス方式と，現在表示しているページを起点として，移動先ページまでの道筋を記述する相対パス方式がある．絶対パス方式の記述の場合は href 属性値をそのままシステムの入力変数とし，相対パス方式の場合は表示ページの URL を用いて絶対パス方式に変換した後システムの入力変数とする．

（2） 表示ページと共に，検索結果を結果ページへのリンクとして提示するための JavaScript をページ HTML に追加する．移動先ページと同時に移動元ページのリンク選択による検索結果を提示することで，利用者のページ内容精査 → 検索キーワード作成 → 検索結果表示というプロセスをなくすことができる．なお，提示する検索結果ページへのリンクに対しても **Step4-(1)** におけるリンク加工を行う．

5.　実験と評価

本章では，4. 章で作成したシステムを用いて提案手法における有効性を測定する実験と，実験結果の評価について述べる．実験は予備実験と評価実験の二つを行い，予備実験によってシステムが作成する問合せキーワード数の妥当性を示す．評価実

表 1　各問合せキーワード数に対する 11 点平均適合率

キーワード数	11 点平均適合率
1	0.350
2	0.621
3	0.660
4	0.367
5	0.238
6	0.093
7	0.0
8	0.0
9	0.0
10	0.0

験では，予備実験で得られた問合せキーワード数を用いて，リンク周辺情報と移動先ページ内容を考慮した提案手法を用いたキーワード検索と，従来手法であるキーワードのみを利用した検索を行い，それぞれの実験結果に対して 11 点平均適合率を用いて互いの検索精度を比較することで，提案手法の有効性を示す. [6]

5.1　問合キーワード数決定のための予備実験

3.3 節で述べた手法を用いた Web ページ検索システムとして 4. 章のシステムを作成した際，検索意図抽出処理により作成される問合せキーワードの数を順位付け上位のキーワード 3 個とした．しかし，提案手法では問合せキーワード数によってリンク周辺情報を含む文章の範囲と，意図表現キーワード同士の距離を定義しているため，問合せキーワード数の変化に伴って抽出されるキーワードそのものが変化し，結果として問合せキーワードが変化する可能性がある．そのため，問合せキーワード数の変化により検索精度が変化すると考えられる．そこで，適切な問合せキーワード数を求めるための予備実験を行う．

5.1.1　実験方法と結果

提案手法におけるリンク周辺の定義について，問合せとして利用するキーワード数による動的な範囲決定をせず，リンク周辺の定義をリンクを含むブロック要素内の文章と，その前後のブロック要素内の文章を併せた文章に固定し，意図表現キーワード間の距離を 1 とした上で以下の方法で実験を行う．

まず，ページ X と，そのページに存在するリンク Y をランダムに設定する．次に，これら X, Y と，リンク Y による移動先ページ Z をシステムの入力として Web ページ検索を行う．X, Y, Z を固定した上で，問合せに用いるキーワード数を 1 個，2 個，$\cdots$，10 個と変化させ，それぞれの検索結果の上位 100 件を取得する．

予備実験に対する評価は，検索結果のページ内容をもとに，適合ページと不適合ページに筆者が分類し，問合せに用いるキーワード数毎の検索結果に対して 11 点平均適合率を求めた．その結果を表 1 に示す．この結果，キーワード数を 3 個と決定した．なお，表 1 のキーワード数 7〜10 に対する 11 点平均適合率が 0.0 を示しているのは，問合せキーワードを全て含むページが 1 件も存在しなかったためである．

5.2　検索精度を用いた有効性評価実験

提案手法を実現するために，4. 章で述べたシステムを作成した．本節では，提案手法を用いたシステムと，キーワード検索システムとして Yahoo! JAPAN の Web ページ検索を利用した有効性評価実験を行い，提案手法の有効性を示す．

5.2.1　実 験 方 法

実験方法として，提案手法を用いたシステムと，キーワード検索システムに対し，同等の検索意図を持つ要素を与え，それぞれの検索結果を評価する．まず，検索意図を持つ要素として 5.1.1 節と同様にランダムに選択したページとリンク X, Y，リンク Y の移動先ページ Z を用い，それぞれのシステムに対して以下の入力を与える．

- **提案手法を用いたシステム**

ページ X のリンク Y 周辺の文章と，移動先ページ Z を入力として与える．これは，提案手法を用いたシステムにおいて，ページ X のリンク Y を選択したことと同意である．

- **キーワード検索システム**

4.2 節 **Step1** と同様の方法に対してリンク周辺の文章のみを与えることで，ページ X におけるリンク Y 周辺の文章からキーワードを抽出する．次に，それらのキーワードの中から出現頻度の高い順に 3 個を選択し，それら全て含む Web ページを検索する形式で入力として与える．

以上の入力より得られた 3 つのシステムの検索結果は Web ページの順位付きリストであり，それらから上位 100 件を取得する．リンク Y を選択する際の移動先ページに期待した内容を基準として適合ページと不適合ページを判断し，システム毎に 11 点平均適合率を求めた．この際，それぞれの実験の再現率における全文書中の適合文書の数は，検索結果上位 100 件における適合ページの数とした．

5.2.2　結果と考察

実験結果を表 2 に示す．表 2 は提案手法・キーワード検索・類似ページ検索を用いた各システムの，0.0 から 1.0 までの 0.1 刻みの再現率における補完適合率と，それらの平均である 11 点平均適合率を表すものである．また，最下段には各システムにおける上位 100 件の検索結果内にいくつ適合ページが存在したかを示した．さらに，縦軸に補完適合率，横軸に再現率をとった実験結果の再現率・適合率曲線を図 6 に示す．

結果から，まず各システムの 11 点平均適合率を比較すると提案手法を用いたシステムのものが最も高い値を示していることが見て取れる．さらに，図 6 に示すように，提案手法を用いたシステムの各再現率における補完適合率は，常にキーワード検索システムの値を上回っていることが分かる．また，上位 100 件中の適合ページ数も提案手法を用いたシステムの値が最も高い．この結果は，利用者が検索意図をキーワードとして表現する従来のキーワード検索に比べて，提案手法を用いたシステムが，検索結果の上位に検索意図に適合したページを多く提示することが可能であることを示している．よって，利用者が検索意図をキーワードとして表現することで問合せとする手法に比べ，インターネットにおけるリンク選択時のリンク周辺情報と，移動先ページ内の文章から動的なキーワード抽出を行い，問合せを作成する提案手法が検索精度が高いことがいえる．

表 2　各システムの検索精度実験結果

再現率	補完適合率	
	提案手法	キーワード検索
0.0	0.649	0.462
0.1	0.649	0.462
0.2	0.649	0.462
0.3	0.649	0.462
0.4	0.649	0.462
0.5	0.649	0.462
0.6	0.649	0.259
0.7	0.649	0.258
0.8	0.538	0.233
0.9	0.471	0.233
1.0	0.425	0.216
11 点平均適合率	0.602	0.361
適合ページ数	34	11

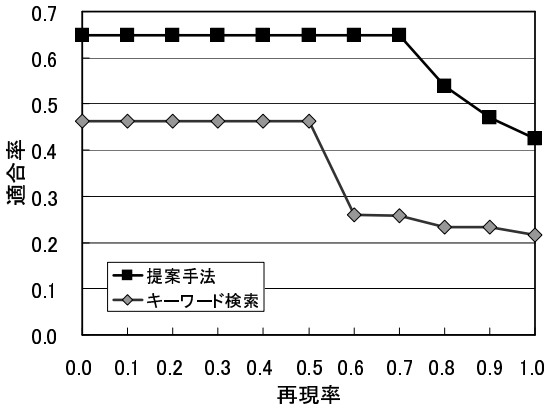

図 6　再現率・適合率曲線

6. おわりに

本稿では，インターネット利用におけるリンク選択時のリンク周辺のキーワードと，移動先ページ本文中の目的とする情報付近にあるキーワードを抜き出した．さらに，そのキーワードの中から，リンク周辺の文章と移動先ページの両方に出現し，互いに近い単語間距離に存在するものを利用者の検索意図を表すキーワードとして選択し，これらを問合せとして利用する手法を提案した．また，利用者と Web サーバの通信を中継するシステムを作成することで，利用者のリンク選択と同時に移動先ページと共に利用者の検索意図を表すページを提示し，提案手法を利用しないキーワード検索システムとの比較を行うことで提案手法の有効性を示した．その結果，利用者の作成する検索キーワードからではなく，リンク選択という直感的なインターネット閲覧操作からページの検索と提示が可能になった．

今後の課題は以下のとおりである．

- **テキスト以外の表現によるリンクからの周辺情報抽出方法の検討**

本稿では，利用者の検索意図を抽出するために，ページのリンク周辺情報と，移動先ページ全体の文章から意図表現キーワードを抽出し，それらから問合せを作成する手法を提案した．この際，利用者によって選択されたリンクに比較的近い位置に存在する文章が，そのリンクから移動することのできるページの内容を説明し，利用者がリンク選択時の判断基準になると考え，リンク周辺情報を抽出する対象とした．しかし，提案手法では，リンクが画像として埋め込まれている場合や，文章以外の形で表現されている場合は考慮されていない．近年では，テキスト以外の方法で表現されたページも増加しており，これらのページへの移動や，ページ内のリンク選択時における提案手法の適用では，利用者の検索意図を表す問合せを作成することは難しい．そこで，これらのページへの移動の際にも，適切に利用者の検索意図を表す問合せを作成するためのリンク周辺情報の抽出方法を検討を行う．

- **連続したページ移動からの検索意図抽出方法の検討**

本稿では，あるページのリンクを利用者が選択する行動を，リンクによる移動先ページを検索結果とする一種の検索行動と捉え，移動元ページのリンク周辺情報と，移動先ページ全体の文章から意図表現キーワードを抽出することで，問合せを作成する手法を提案した．ここで，提案手法では，利用者がリンクを選択した際，そのリンクを中心とした移動元ページと移動先ページのみを利用者の検索意図を示す要素として利用している．しかし，リンク選択とページ閲覧を利用者の基本的な行動の一つと考えると，この行動が連続して行われた際には，その一連の行動を新たな検索行動と捉えることも可能であると考えられる．そこで，連続したリンク選択とページ閲覧から，利用者の検索意図を抽出する手法を検討を行う．

文　献

[1] 大森貴博, 笹塚清二, 水谷正大：“リンク情報を考慮した web 検索システム”, 情報処理学会自然言語処理研究報告, **Vol.99**, No.2, pp. 49–56 (1999).

[2] 山口雅史, 大島裕明, 小山聡, 田中克己：“利用者選好の半順序性に着目した web 探索とナビゲーションの個別化”, 電子情報通信学会 第 16 回データ工学ワークショップ（DEWS2005）論文集 (2005).

[3] “HTML 4.01 specification”, http://www.w3.org/TR/1999/REC-html401-19991224/ (1999).

[4] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.jp/ (2006).

[5] 中谷圭吾, 鈴木優, 川越恭二：“利用者の要求に応じた web リンク自動生成手法”, 電子情報通信学会 第 16 回データ工学ワークショップ（DEWS2005）論文集 (2005).

[6] 北研二, 津田和彦, 獅々堀正幹：“情報検索アルゴリズム”, 共立出版 (2001).