情 報 学 基 礎 66−15
デジタル・ドキュメント 32−15
（2002. 3. 15）

ブラウザ履歴による検索キーワードの意味的多義性の解消に関する研究

小川 功[†1] 羽田久一[†2]
今井正和[†3] 砂原秀樹[†4]

近年、Webのリンク構造を用いた検索技術が注目を集めている。しかしこれらの検索技術でも、意味的多義性を持つ検索キーワードが入力された場合、同一キーワードの異なる意味に関連するWebページが検索結果として返される場合がある。

本稿ではブラウザ履歴をデータベース化し、検索時にデータベースから検索キーワードの関連語を抽出することで、検索キーワードの多義性を解消するシステムを提案する。一般に、名詞である同音異義語が複合語内に存在すれば、前後の単語から同音異義語の意味が特定できる場合が多い。検索キーワードとして名詞が使われブラウザ履歴に検索キーワードが含まれる場合、その語の前後の単語を調べることで多義性を解消できた。

A Research on the Dissolution of the Semantic Polysemy of the Retrieval Keyword by the Browser History

Isao Ogawa,[†1] Hisakazu Hada,[†2] Masakazu Imai[†3]
and Hideki Sunahara[†4]

Recently, many people observe the retrieving technology using link structure peculiar to the Web. In this case , when a polysemous word is given as a retrieval key, the different result from the target information may be chosen with this methodology.

In this paper, we propose the system which solves the polysemy of the retrieval keyword by accumulating the browser history into the database and extracting some words related to the retrieval keyword. Generally the meaning of a polysemous word can identified with the coocurrence words. Therefore, if the set of coocurrence words can be stored, we can utilize the data to identify the meaning of the polysemous word.

1. はじめに

インターネットの急速な普及により、多くの人々がWorld Wide Web (WWW) を利用するようになった。従来よりWWW上のWebページから情報を獲得する方法として、WWW検索システムの利用がある。WWW検索システムには、キーワードで検索するものが多い。これらの検索システムでは、ユーザは検索キーワードを入力し、システムは、入力された検索キーワードとWebページのインデックスを比較し、類似度を求め、これにより検索結果を出力する。しかし、Webページとの類似度をそれに含まれる単語のみで求めているために、ユーザは満足する検索結果が得られないことが多い。そこで近年、WWW特有のリンク構造を用いた検索システムが登場し、その検索精度の高さから注目を集めている[1)]。

しかしこれらの検索システムでは、検索キーワードのみで、ユーザの検索意図を判断している。そのため、不適切な検索キーワードを入力すると、ユーザが満足する検索結果を得られない場合がある。その一例として、意味的多義性を持つ検索キーワードが挙げられる。このような検索キーワードが入力された場合、同一キーワードの異なる意味に関連するWebページが検索結果として返される場合があり、ユーザが満足す

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†2 奈良先端科学技術大学院大学 附属図書館研究開発室
Research Division Digital Library, Nara Institute of Science and Technology
†3 鳥取環境大学 情報システム学科
Department of Information System, Tottori University of Environmental Studies
†4 奈良先端科学技術大学院大学 情報科学センター
Information Technology Center, Nara Institute of Science and Technology

る検索結果が得られない場合がある。例えば、「ATM」という固有名詞を検索キーワードとして入力すると、非同期転送モードの意味を持つ「ATM」と、現金自動預払機の意味を持つ「A TM」などが検索される。ユーザは検索キーワードを複数入力することにより、検索意図を詳細にシステムに伝達させることが可能である。しかし商用の検索システムでは、ユーザが検索に用いる語は一、二語程度であることが言われており[2]、検索キーワードが少数の場合でもシステムがユーザの検索意図を正しく推定する必要がある。

本稿では検索キーワードの多義性を解消することを目的とし、検索キーワードとして一語程度しか入力しないユーザに対して、ユーザが要求する情報を出力するシステムを提案する。そのための手法として、ブラウザ履歴をデータベース化し、検索時にデータベースから検索キーワードの関連語を抽出し、検索キーワードと関連キーワードで検索する。

また、このシステムに対し、検索キーワードの単一の意味に関するWebページを履歴として使用し、その検索キーワードを入力する実験を行い、評価する。

2. 関連研究

川前らの研究[3]では、ユーザのIPアドレス、検索キーワード、アクセス先のURLを検索ログとして保存し、入力キーワードの変更に対し検索ログから再帰的にキーワードディレクトリを作成することにより検索支援を行う。この手法では、検索時の総URL閲覧数の削減が可能となる。しかし、検索支援の効果がユーザが過去に閲覧したページに大きく依存し、検索結果であるURLの一覧から閲覧したWebページは、検索結果の出力の順序に大きく依存するため、これを改善することが課題となる。本研究では、クライアント側に存在するブラウザ履歴を用いて、ユーザ自身の検索支援を行うものであり、この研究とは異なる。

原田らの研究[4]では、不特定多数のユーザの検索ログから検索キーワードと閲覧したWebページを関連づける。ユーザは検索キーワードを入力すると、システムはURLのリストを求める。そしてそのURLに一番多く関連づけられている検索キーワードを関連語として提示する。この手法では、検索サーバに負荷をかけず平易な手法で検索支援が可能である。しかし、閲覧したWebページのうち、ファイルサイズが大きなWebページは多くの語を含んでおり、様々な検索キーワードにマッチするため、しばしば検索キーワードと無関係な語が関連語として出力される。本研究は、ブラウザ履歴から関連キーワードを抽出し、検索支援を行うものであり、不特定多数のユーザに対する検索支援であるこの研究とは異なる。

川越らの研究[5]では、図1のように、クライアントマシンにプロキシサーバをインストールすることにより、ユーザのWebページに対するアクセスの流れをデータベース化する。そして、その中から意味的にまとまりのあるWebページ群を抽出し、検索時にそれを提示する。意味的にまとまりのあるページ群は文書同士の関連度を計算することにより求める。文書同士の関連度は既存の手法であるTF・IDF法とベクトル空間モデルを用いる。意味的にまとまりのあるWebページ群のみを検索することにより、検索する際の総ページ閲覧数を減少させることが可能となる。この手法は、ユーザが過去に閲覧したWebページを構造化して、共有する。そして、それを他ユーザが検索する時に利用するものである。この手法は、クライアント側にある履歴を用いる手法であり、履歴を獲得するのにクライアントマシンにプロキシサーバをインストールしている点で本研究と似ている。しかし、検索キーワードの多義性の解消を目的とし、ユーザの履歴をそのユーザ自身が利用する点でこの研究とは異なる。

獅々堀らの研究[6]では、検索キーワードの多義性を解消することを目的としている。HTMLファイル中の表構造を解析することによりそのWebページの検索キーワードとなる固有名詞の意味抽出を試みている。その意味を抽出することにより、Webページ上に存在する検索キーワードの多義性の解消を試みる。例えば、表1、表2のように「ヤクルト」という単語には、球団名の意味を持つ「ヤクルト」と、企業名の意味を持つ「ヤクルト」が存在する。この場合、「ヤ

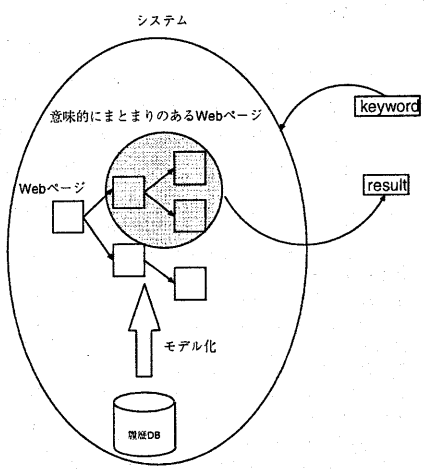

図1 アクセス履歴を用いる手法

表 1 「球団名」の意味を持つ" ヤクルト" の表

球団名	順位	氏名	位置	出身
ヤクルト	1 位	○○ ○○	投	愛工大名電高
ヤクルト	2 位	△△ △△	投	旭川実業高
横浜	1 位	×× ××	内	豊田大谷高

表 2 「企業名」の意味を持つ" ヤクルト" の表

企業名、店舗名	住所	TEL	FAX
大阪○○乳業 (株)	大阪市□□	06-...	
ヤクルト本社××支店	大阪市△△	06-...	
(株) 雪印○○○○支店	大阪市××	06-...	06-...

クルト」が持つ意味によって、表の最上位の項目が異なる。これを利用して検索キーワードの多義性の解消を試みる。しかし、この手法では、HTML ファイル中の表構造についてのタグを解析する場合、そのタグが表として使用されているかを考慮する必要がある。本研究では、ブラウザ履歴を用いており、また、個々のユーザに応じた検索支援を行う点でこの研究とは異なる。

3. 関連キーワード抽出による検索支援システム

3.1 クライアント側の履歴の活用

履歴を獲得する方法として、サーバ側で獲得する手法やクライアント側で獲得する手法が考えられる。しかし、サーバ側で獲得する手法では、数多くの Web サーバの協力が必要となるため現実的ではない。また、クライアント側で履歴を獲得する手法は、容易にユーザの履歴を獲得することが可能であるため、これを適用することとする。

3.2 過去に閲覧した Web ページの利用

クライアント側に存在する履歴を用いる手法には様々なものがある。例えば、電子メールの内容を解析する手法や、過去に閲覧した Web ページを解析する手法が考えられる。また、クライアントに存在するブックマークの内容を解析し検索支援に用いることも考えられる。電子メールの内容を解析する手法では、図 2 のように電子メールにメールマガジンなどユーザを特徴づけるメールが存在する場合、その内容を解析して、それを検索意図の判断材料とすることも考えられる。しかし、どのユーザにもユーザを特徴づけるようなメールが存在するとは限らず、また、電子メールを使用しないユーザには適用できない。また、ブックマークの内容を解析する手法では、ユーザの興味の対象を直接推定することが可能である。しかしブックマークを使用しないユーザには適用不可能であり、ブックマークにはすでに興味がない Web ページが存

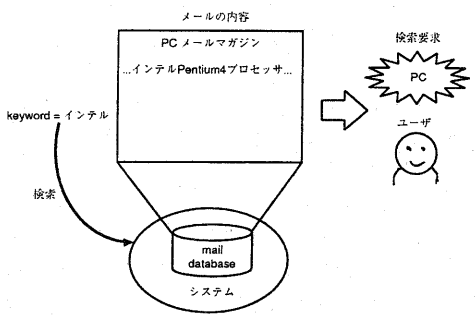

図 2 電子メール解析システム

在することがあり、その場合には有効ではない。

提案する検索支援システムでは、ユーザの検索意図をユーザ個人の興味と結びつけて考え、ユーザが興味を持った事柄に対して検索すると考える。そのため、ユーザの興味をどうやって推定するかが問題となる。小石川らの研究[7] では、ユーザの Web ページ閲覧時間とユーザ個人の関心には、ユーザの閲覧時間が長い Web ページに強い関心があることが述べられており、過去に閲覧した Web ページをユーザの興味の対象として捉えるのは妥当であると考えられる。ユーザが興味を持つには、過去に閲覧した Web ページ中に検索キーワードが現れると考え、その語の前後の単語から検索意図を推定する。これにより多義性のある検索キーワードを入力しても、ユーザが満足する検索結果が得られると考える。

3.3 システムの概要

提案システムはユーザの検索意図を推定し検索キーワードの多義性を解消することを目的とする。システムは図 3 のようにプロキシサーバ、インデックス作成部 (indexer)、検索部 (index searcher) の 3 つから構成される。

ユーザは proxy server を通して Web ページを閲覧する。その時、proxy サーバのログとして履歴 URL が蓄積される。インデックス作成部は履歴 URL のうち HTML ファイルのみを取得し、インデックス化する。検索部は、ユーザの検索時にインデックスデータベースから関連キーワードを抽出する。そして、それを検索キーワードに追加して既存の検索システムへ送信し、返ってきた検索結果をそのまま出力する。

3.3.1 プロキシサーバ

プロキシサーバではユーザがアクセスした Web ページの URL を蓄積する。蓄積された URL をブラウザ履歴として使用する。ブラウザ履歴をブラウザから直接使用しない理由は、ブラウザの種類・バージョンに

依存するためである。また、プロキシサーバのキャッシュの内容は、最新でない場合があるため使用しない。プロキシサーバはあらかじめ設定しておく。プロキシサーバに蓄積されたブラウザ履歴をユーザ個人が使用することにより、プライバシーの配慮が可能である。プロキシサーバにはフリーソフトウェアである delegate7.5.4[8] を用いた。

3.3.2 インデックス作成部

インデックス作成部の動作の概要は以下の通りである。インデックス作成部では、ページを収集する page collecter とインデックスを作成する index builder を作成した。インデックスの作成は、ユーザが Web ページを閲覧し、ログに新しい URL が書き込まれるたびに行われる。これにより、Web ページ閲覧時にのみオンラインにするユーザに対しても、インデックスが作成可能である。

(1) 蓄積された URL から html ファイルに関する URL のみを抽出
(2) 抽出された URL に対してリクエストを送信
(3) html ファイルを獲得
(4) html ファイルからタグを取り除き、形態素解析を行い、単語を抽出
(5) 単語の出現頻度と前後の共起関係を考慮して、インデックス化

以下、それぞれの項目について説明する。

- Web ページの収集
  蓄積されたブラウザ履歴のうち、音声ファイルや画像ファイルや cgi ファイルを除いた HTML ファイルのみを収集する。cgi ファイルは内容が動的に変化するため、収集しない。収集した HTML ファイルからタグや特殊フォントに関する記号を取り除く。また、"+" や"*" などの記号も正しく形態素解析を行うため取り除く。
- 形態素解析
  形態素解析プログラムには chasen2.1[9] を用いた。chasen はクライアントマシンにあらかじめインストールしておく。辞書ファイルは chasen2.1 に付属のものを用いた。形態素解析の後、表 3 の品詞について抽出する。名詞は検索キーワードとして使用されることが多いため抽出する。Web ページ上に存在する未知語は名詞であることが多かったため、未知語も抽出する。ここで、抽出した語の集合を W とする。
- インデックス化
  インデックス化の手順は以下の通りである。
  (1) 抽出された単語のうち前後 3 単語以内に存在する表 4 で示される品詞について調べ、抽出された単語の前後、それぞれ一つずつを関連語の候補とする。助詞などの非自立語は共起しやすい単語であるため、関連語の候補として取り除く。また、計算量を削減するため、3 単語以内を調べることとする。ここで候補となる関連語の集合を

表 3 抽出する品詞

抽出する品詞
名詞-一般
名詞-形容動詞語幹
名詞-サ変接続
名詞-ナイ形容詞語幹
名詞-固有名詞
未知語

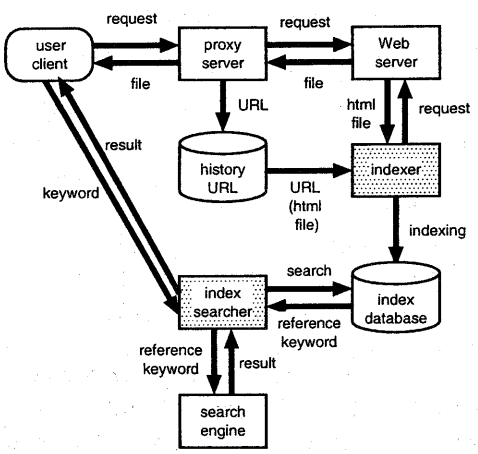

図 3 提案システムの構成図

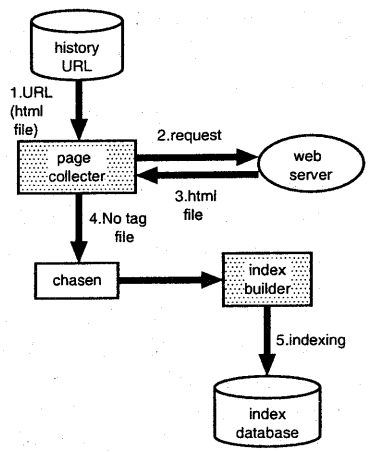

図 4 インデックス作成部

W_-, W_+ とする。

(2) W_-, W_+ についてその単語の前後 3 単語 (その単語自身を含む) 内に存在する語を調べ、関連度を調べる。W_- については後の語、W_+ については前の語について調べる。これは、候補となる関連語が様々な単語と共に出現する場合、元の語の意味を特定する単語としてふさわしくないと考えられるためである。関連度は次の式で定義する。

$ref(g) = \sum_g \frac{n(h \cap g)}{n(g)} (g \subset W_- or W_+, h \subset W)$

$n(h \cap g)$ は単語 h と単語 g が 3 単語以内に出現する回数、$n(g)$ は単語 g の出現回数である。関連度は、一つの単語についてページ毎に計算し、和をとることにより求める。関連度を用いることにより、検索キーワードと関連が深い語のみを抽出できる。

3.3.3 検索部

検索部の動作は以下の通りである。

(1) ユーザが検索キーワードを入力

(2) あらかじめ作成されているインデックスファイルに対し、検索キーワードを探索。検索キーワードが存在しなければ、既存の検索システムへ検索キーワードを送信し、検索結果をそのまま出力

(3) 検索キーワードが存在する場合には、関連キーワードを抽出

(4) 検索キーワードと関連キーワードを合わせて検索システムに送信

(5) 検索結果をそのまま出力

以下、それぞれの項目について説明する。

- 既存の検索システムによる検索
  インデックスファイルに検索キーワードが存在しない場合、既存の検索システムにそのまま検索キーワードを送信し、返ってきた検索結果をそのまま出力する。既存の検索システムとしては、一般に登録されている Web ページ数が大手検索システムの中で最大と言われている Google[1] を用い、日本語のページについて検索した。

表 4 関連語として抽出する品詞

抽出する品詞
名詞-一般
名詞-形容動詞語幹
名詞-サ変接続
名詞-ナイ形容詞語幹
名詞-固有名詞
形容詞-自立
動詞-自立
未知語

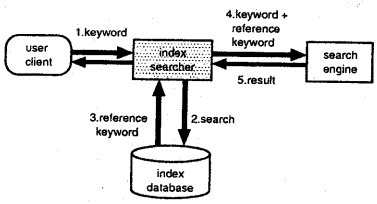

図 5 検索部

- 関連キーワードの抽出
  インデックスファイルに検索キーワードが存在する場合、その検索キーワードの関連キーワードを抽出する。関連キーワードは、インデックス作成部で算出した関連度の最も高い語とする。検索システムへ検索キーワードおよび関連キーワードを新しく検索キーワードとして送信し、検索結果をそのまま出力する。関連キーワードを追加して検索することにより、検索キーワードが一語の場合でも、ユーザの検索要求を推定することが可能である。

4. 評価実験

4.1 実験データ

検索キーワードとして表 5 のような多義性を持つキーワード 25 種類について実験した。それぞれの検索キーワードに対し、ブラウザ履歴として Yahoo[10]、Lycos[11]、Excite[12] のカテゴリ及びサイト登録されている日本語の Web ページを 1 つの意味に関して 20 件ずつ収集し使用した。分類されているカテゴリ及び説明文を元に単語の意味を判別した。Yahoo、Lycos、Excite のカテゴリ及びサイトに登録されている登録件数が合計 20 件以下である検索キーワードについては、そのページ数だけ Web ページを収集し、一つの検索キーワードについて履歴のページ数をそれぞれ等しくした。

4.2 実験内容

ブラウザ履歴に対して本システムを適用し、インデックスを作成する。作成されたインデックスに対し、関連キーワード抽出プログラムを起動し、検索キーワードを入力して関連キーワードを抽出し、既存の検索システムへ送信し、検索結果を評価した。検索結果の Web ページの評価は、Web ページの内容が判別できないページについては適合文書に含めなかった。なお、検索キーワードの中には、ここに表記した以外の意味も存在するが、簡単のため割愛した。

表 5 実験に使用した検索キーワードの一部

検索キーワード	意味
ATM	1. 非同期転送モード 2. 現金自動支払機
DV	1. デジタルビデオ 2. ドメスティック・バイオレンス
インテル	1. 半導体メーカー 2. イタリアのサッカーチーム
ボーリング	1. 球を転がす 2. 穴を掘る
HP	1. ヒューレット・パッカード (企業名) 2. ホームページ
呉	1. 日本の地名 2. 中国の三国時代の国名
オウム	1. オウム真理教 2. 鳥
AV	1. オーディオ・ビジュアル 2. アダルト・ビデオ
IC	1. 集積回路 2. インターチェンジ
中国	1. 日本の地方名 2. 世界の国名
デフォルト	1. 債務不履行 2. 初期設定
AMD	1.Advanced Micro Devices(企業名) 2.auto mount daemon
LD	1. レーザー・ディスク 2. 学習障害
FA	1. フリー・エージェント 2. ファクトリー・オートメーション
JAS	1. 日本農林規格 2. 日本エアシステム
DDR	1. ダンスダンスレボリューション 2.Double bit Data Rate

4.3 評価手法

評価手法については、提案システムと既存の検索システムの検索結果の上位 20 件について比較して評価する。ユーザは、検索結果のうち上位 20 件までしか閲覧しないという調査結果がある[13]。そのため上位 20 件を比較することは妥当であると考えられる。一般に情報検索においては、評価項目として再現率 (Recall) と適合率 (Precision) が用いられる。再現率と適合率は以下のように定義される。

$$Recall = \frac{\text{検索された適合文書数}}{\text{検索対象となる文書集合中の適合文書数}}$$

$$Precision = \frac{\text{検索された適合文書数}}{\text{検索された文書数}}$$

しかし、Web の検索においては、検索対象となる文書集合が膨大であり、検索対象となる文書集合中の適合文書数を求めることが不可能であるため、再現率を求めることが不可能である。また、検索された文書数も膨大になる場合がある。よって、本論文では、検索結果の上位 20 件について適合率を求めることにより、疑似的な適合率を求める。

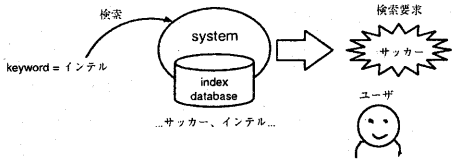

図 6 ユーザの検索要求

本研究では、図 6 のように、過去に閲覧した Web ページに検索キーワードが含まれる場合、ユーザがその検索キーワードで検索する時の検索要求は、過去に閲覧した Web ページのその語の意味に関する Web ページであると考える。それに基づき検索結果を評価する。例えば、サッカーに関する Web ページを過去に閲覧した場合、検索キーワード「インテル」で検索する時の検索要求はサッカーチーム「インテル」に関するものであると考える。この考え方は、過去に閲覧した Web ページ中に意味の分からない語が存在し、その単語の意味を調べるためにそれを入力して検索する場合に成り立つ。この考え方に基づき、検索キーワードの一つの意味に関する Web ページを履歴として持った場合に、検索結果のうちその意味に関する Web ページを適合文書として提案システムと既存の検索システムで比較する。

4.4 実験結果

実験結果を、表 6 に示す。ブラウザ履歴の項目の数字は表 5 の検索キーワードの意味の項目の数字に対応している。例えば、ブラウザ履歴が 1.ATM の場合は、ブラウザ履歴として非同期転送モードに関する Web ページが 20 件存在する。これに対し、検索キーワー

表 6 関連キーワードと既存システム、提案システムの適合率 (一部)

検索キーワード	関連キーワード	検索キーワード	関連キーワード
1.ATM	フレームリレー	2.ATM	CD
1.DV	編集	2.DV	防止
1. インテル	R	2. インテル	歴史
1. ボーリング	行う	2. ボーリング	(なし)
1.HP	hp	2.HP	公式
1. 呉	私立	2. 呉	東
1. オウム	アレフ	2. オウム	インコ
1.AV	機器	2.AV	女優
1.IC	レイアウト	2.IC	車
1. 中国	地域	2. 中国	各地
1. デフォルト	債務	2. デフォルト	ルート
1.AMD	Athlon	2.AMD	Auto
1.LD	DVD	2.LD	学会
1.FA	宣言	2.FA	機器
1.JAS	マーク	2.JAS	マイレッジサービス
1.DDR	ページ	2.DDR	Double

ド「ATM」を入力する。提案システムでは、関連キーワード「フレームリレー」が提示される。図7は既存の検索システムと提案システムの適合率をグラフ化したものである。

4.5 考　　察

多くの場合で、提案システムの適合率が既存の検索システムの適合率を上回った。しかし、関連キーワードの中には不自然なものが存在した。例えば、検索キーワード「HP」の時の関連キーワード「hp」がそれに当たる。これは、「hp の HP」(ヒューレット・パッカートのホームページ) のように一つの検索キーワードが異なる意味で共に出現することが考えられるため、不自然な関連キーワードが抽出されたと考えられる。他にも、検索キーワード「呉」の「東」のように、理解しにくい関連キーワードも存在した。これは、履歴とする Web ページが検索キーワード「呉」の場合 11 件と少ないためであると考えられる。関連度の上位 5 番以内に「三国志」が登場しており履歴とする Web ページが増えれば改善される可能性はある。

本手法では形態素解析を行い、関連度を算出して関連キーワードを抽出している。その方式上形態素解析の結果に大きく依存する部分が存在する。それは複合語及び未知語の扱い方である。複合語については、形態素解析の辞書の内容に大きく依存する。例えば、「情報処理」を「情報処理」として一語とするか、「情報」「処理」として二語とするかである。これについては、どちらが正しいとは言えない場合があり、その点は課題である。また、未知語は、その内容が日々刻々と変化するインターネット上には存在する可能性が高く、これをどう処理するかが問題となる。本研究では、未知語について品詞を調べた結果名詞であることが多かったため、名詞と同様に扱った。

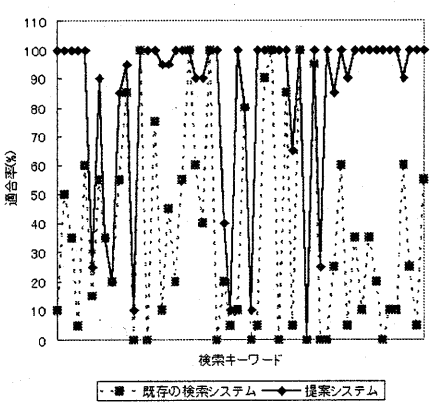

図 7　既存の検索システムと提案システムの適合率

また、本手法は関連キーワードを元の検索キーワードに追加して既存の検索システムを用いて検索する。そのため、検索結果の内容が既存の検索システムのデータベースの内容に依存する。したがって、既存の検索システムのデータベースにユーザが要求する情報がひとつも存在しない場合、満足する結果が得られない。したがって、既存の検索システムを用いる手法では、データベースの規模ができるだけ大きな検索システムを選択する必要がある。

また、日本語形態素解析プログラム chasen を使用しているため、英語のみで表記されている Web ページに対しては、単語の切れ目が誤って解析されるため、適用できない。しかし、検索キーワードが英語である場合には、日本語で表記されたページの一部にその語が独立して登場する場合には、その前後の語を調べることにより意味の判別が可能である。

5. 課　　題

実験結果から、検索キーワード「インテル」や検索キーワード「呉」等、一部で不自然な関連キーワードが抽出された。不自然な関連キーワードが抽出されず、適合率を評価しても既存の検索システムを上回るのに必要な履歴はどれくらいであるか調べる必要がある。また、過去の全ての履歴を保存しておくことは不可能であるため、履歴はどれくらい残すべきかを考慮する必要がある。

今回の評価実験は、単一の意味に関する Web ページを履歴とした場合についてのみ行った。しかし、本方式は意味が異なる Web ページが存在する場合にも適用が可能である。例えば、検索キーワード「DV」に関して実験データである「デジタル・ビデオ」の意味に関する Web ページと「ドメスティック・バイオレンス」の意味に関する Web ページそれぞれ 20 ページずつを履歴とした場合には、関連キーワードとして「編集」が抽出される。しかし、これが適当であるか判断が困難であり、評価が困難である。それは、関連キーワードを一語のみ抽出して検索するためである。関連キーワードを一つだけでなく複数抽出して、ユーザに選択させるようにすれば、本方式を用いても有効であると考えられる。複数の意味に関する Web ページが履歴として存在する場合に、有効性について検証する必要がある。

また、複数の意味に関する Web ページが履歴として存在する場合、関連キーワードの抽出に誤ることが考えられる。その場合、本システムで検索する場合が

既存の検索システムで検索する場合に比べ、閲覧ページ数が増え検索効率が低下することが考えられる。このため、検索結果だけでなく関連キーワードを提示し、ユーザに関連キーワードを選ばせてその結果をフィードバックさせるといったことが必要となる。

本手法は、興味の対象が変化した場合には適用が不可能である。例えば、同じ検索キーワード「AV」でも、昼はオーディオ・ビジュアルの意味で「AV」を入力し、夜はアダルト・ビデオの意味で「AV」を入力することも考えられる。このように興味の対象が時間的に変化する場合には、Web ページを閲覧した時刻を考慮してインデックスを作成する。そして、検索時にその時の時刻も考慮して関連キーワードを抽出すれば、興味の対象が変化した直後には対応不可能であるが、すぐにそれに追随することが可能となる。本手法を改良することにより、興味の対象の変化に対応できる。

6. おわりに

ユーザの意図を推定する手法として、ブラウザ履歴から関連キーワードを抽出することによる検索支援方式を提案した。これを用いて、検索キーワードの多義性の解消について評価した。

評価実験として、検索キーワードの単一の意味に関する Web ページを履歴として集め、関連キーワードを抽出し検索した。実験の結果、多くの検索キーワードにおいて既存の検索システムに比べ高い適合率を示した。実験結果から、単一の意味に関する Web ページが履歴として存在した場合には、検索意図が反映され有効性があると言える。

今後は、自然な関連キーワードの抽出に必要な履歴の量の検証、複数の意味に関する Web ページが履歴として存在する場合の本手法の有効性の評価、時刻を考慮したインデックスの作成をしたいと考えている。

参考文献

1) Google
http://www.google.com
2) 原田昌紀: WWW サーチエンジンの技術動向, 電子情報通信学会技術報告, SSE2000-228, pp17-22, 2001
3) 川前徳章, 青木照勝, 安田浩: ユーザ履歴を活用した検索システム, 電子情報通信学会技術研究報告, DE2000-37, pp113-120, 2001
4) 原田昌紀, 清水奨: WWW 検索システムにおける不特定多数の操作履歴の活用, 情報処理学会研究報告, DPS81-11, pp61-66, 1997
5) 川越恭二他: ネットワークアクセス行動の DB 化と WWW 検索への応用, 電子情報通信学会技術研究報告, DE2001-40, pp25-32, 2001
6) 獅々堀正幹, 岩口義広ほか: HTML 形式の表構造に対する一索引化手法, 電子情報通信学会技術研究報告, DE2001-54, pp137-144, 2001
7) 小石川将, 呉勇, 岸本陽次郎:個人の関心を利用した情報検索システム, 電子情報通信学会技術研究報告,OFS2000-47,pp11-16,2001
8) プロキシサーバ delegate
http://www.delegate.org/delegate/index.html
9) 松本裕治、北内啓、山下達雄、平野善隆、松田寛、高岡一馬、浅原正幸:日本語形態素解析システム「茶筌」version2.2.1 使用説明書、奈良先端科学技術大学松本研究室
10) Yahoo JAPAN
http://www.yahoo.co.jp
11) Lycos JAPAN
http://www.lycos.co.jp
12) Excite JAPAN
http://www.excite.co.jp
13) 風間一洋, 原田昌紀, 佐藤進也: ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法, 情報処理学会研究報告, DD24, 2000