E-027

WWW を利用した広域検索型辞書システム
A Wide Area Coverage Dictionary Using World Wide Web

三鍋 洋司† Yoji Minabe　吉田 卓哉† Takuya Yoshida　加藤 貴司† Takashi Katoh　児玉 英一郎† Eiichiro Kodama　ベッド B. ビスタ† Bhed Bahadur Bista　高田 豊雄† Toyoo Takata

1．はじめに

現在，我々の周りで利用されている用語の数は増え続けている．広辞苑の収録語数は毎年約 5 千語の割合で増えており，広辞苑に掲載されていない用語を含めると増加した用語の数は計り知れない．我々が未知の用語の意味を調べるためにいくつかの検索ツールを利用するが，どの検索ツールにも欠点があり利便性が高いとは言い難い．そのため，より利便性の高い検索ツールを開発することは重要である．そこで本稿では，利便性の高い検索ツールに求められる要求分析を行い，その結果に基づき，メタ Web 辞書と Web ページからの語義抽出を組み合わせ，日英/英日対訳機能を付加した広域検索型辞書システムの提案を行う．また，提案システムの実装，評価について述べる．

2 検索ツールの現状と分析

2．1 既存の検索ツールの分析

本節では， Web 辞書，検索エンジンによる WWW 上の検索について，それぞれの特徴を分析する．紙数の都合上短所のみ列挙する．

(1) Web 辞書

書籍辞書を電子化し，WWW を通して利用できるようにした辞書サービスである．

・Web 辞書で扱われる専門分野は狭く，そのほとんどは，一般的な用語か IT 分野における専門用語に限られている．

(2) 検索エンジン

検索エンジンを用いて，用語の説明文が掲載されている Web ページを探すサービスである．

・検索可能語数が多い反面，検索の手間が大きい．

検索エンジンはキーワードに対して，大量の URL を表示するため，利用者はその中から用語の解説文を手作業で探しださなければならない．検索エンジンで用語の説明文を検索する多くの場合，キーワードを工夫する必要があり，検索に要する手間は大きい．

・検索に要する時間も長く，最終的に用語の解説文が見出せない場合もある．

・Web ページ上の解説文にはしばしば誤字や本質的な間違いが含まれており，解説文の精度は必ずしも高くない．

2．2 利便性の高い検索ツールの要求

前節の分析結果より，利便性の高い検索ツールの要件として，以下のような要素が考えられる．

・検索の容易さ: 検索が容易にできること．検索結果の出力の仕方などにも左右される．

・検索時間: 用語の検索を始めてから解説文が得られるまでに要する時間が短いこと．

・収録語数: 多くの分野の専門用語や造語などが収録されており，収録語数が多いこと．

・更新速度: 新しい用語の説明文が得られるようになるまでに要する日数が短いこと．

・説明文精度: 用語に対する説明文に間違いがないこと．

・携帯性: 時間や場所を問わず利用できること．

前節の検索ツールと従来からある書籍辞書と電子辞書を併せた分析結果を表 1 に示し，各検索ツールで検索できる用語の包含関係を図 1 に示す．

表 1 各検索ツールの分析評価

	書籍辞書	電子辞書	Web 辞書	検索エンジン
検索の容易さ	○	◎	◎	×
検索時間	○	◎	◎	×
収録語数	○	×	×	◎
更新速度	×	×	○	◎
説明文の精度	◎	◎	◎	△
携帯性	×	○	○	△

(◎非常に良い，○良い，△悪い，×非常に悪い)

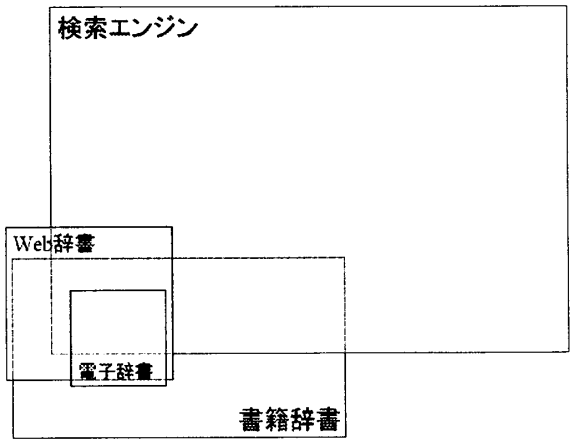

図 1 各検索ツールで検索できる用語の包含関係

2．3 利便性向上のためのアプローチと現状

本節では Web 辞書，Web 検索の欠点を改良するためのアプローチを考察する．

・Web 辞書

Web 辞書は収録語数が非常に少ないという欠点を持っている．すなわち現在，国語辞典，英和辞典，和英辞典などの一般的な辞典と IT 分野における用語辞典は Web 辞書化されているが，他の多くの分野においては進んでいない．

†岩手県立大学ソフトウェア情報学部

また，Web 辞書を統合するメタ Web 辞書型システムの開発により収録語数を増やす研究[1] が行われている．現段階では，検索したい用語についてシステムに登録されている複数の Web 辞書で検索を行い，用語の解説文が得られた Web 辞書のリンクを表示する．これにより，利用者は用語の解説文が得られない Web 辞書に対して検索する手間が省ける．また，利用者がその存在を知らない Web 辞書に対しても検索を行うため，利用者が直接 Web 辞書を回って検索するより用語の説明文を得られる可能性が高くなる．

・検索エンジン

検索エンジンにおいては，検索の難易度が高いことと，検索に時間を要するという欠点を持っている．現在，検索エンジンから得られる大量の URL から Web ページを収集し，用語の解説文を自動で抽出を行う Web ページ抽出型システムの研究がいくつか知られている．その中で，桜井らの研究[2] では，辞書などで用いられる用語の説明文を 12 種類のパターンに分類し，WWW 上より収集した Web ページがいずれかのパターンと一致するかを判定することで説明文の抽出を行っている．更に，抽出した説明文から用語の語義ごとに上位概念を設定し，複数の語義を持つ用語に対して複数の説明文を用意できるように開発されている．また，吉田の研究[3] では説明文だけではなく，日英対訳も Web ページより抽出できるシステムの開発を行っている．

以上，Web 辞書と検索エンジンは相補的な関係が多々あるにもかかわらず Web 辞書の統合を行ったメタ Web 辞書型と，WWW 上の Web ページから説明文を抽出する Web ページ抽出型を同時に扱った研究はなされていないのが現状である．

3．WWW広域検索型辞書システムの提案

3．1 システムの概要

本節では前述したメタ Web 辞書型及び Web ページ抽出型の二つを統合したシステム(WWW 広域検索型)の構築について述べる．利用者は複数の Web 辞書や Web ページの解説文を一つの Web 辞書のように扱うことができ，検索効率も更に向上すると考える．また，本論文ではメタ Web 辞書型の先行研究 [1]で実装されていなかった説明文の整形，統合を行い，さらに，Web ページ抽出型の先行研究 [3]で行われていなかった動的な検索や検索時間の短縮も実現する．

3．2 システムの設計

本システムは，用語を入力として受け付け，その用語に対する日英説明文及び日英対訳を，WWW 上の Web ページ及び Web 辞書から抽出し出力する．本システムはインタフェース部のリトリーバ，メタ Web 辞書部，Web ページ抽出部，用語データベース（以下，用語 DB）から構成される．本システムのモデルを図 2 に示す．

以下，本システムの中心部であるメタ Web 辞書部と Web ページ抽出部について述べる．

3．2．1 メタ Web 辞書部

上述したメタ Web 辞書型の先行研究 [1]の検索結果は，各 Web 辞書の検索結果を順次表示するにとどまっており，最終的にはそれぞれの利用者が各 Web 辞書の検索結果を閲覧し，解説文を比較しなければならない．これに対し，本システムは各 Web 辞書の検索結果を収集し，意味が重複している説明文の統合処理を行い，利用者に表示する．また，複数の Web 辞書で重複して説明されている意味を一般的な意味であるとし，そのような説明文を先に表示する．重複説明文の検出方法は，各説明文内から，カタカナまたは漢字の連続を抽出し，抽出した文字列の重複数が，抽出した文字列数の 2 割を超えていれば同じ意味の説明文と解釈する．

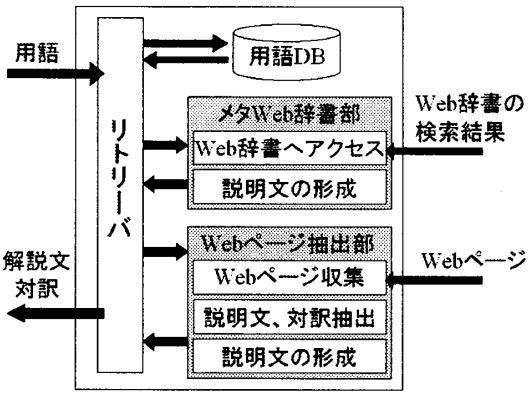

図 2 システムモデル

3．2．2 Web ページ抽出部

前述した Web ページ抽出型の先行研究 [3]では用語の説明文及び対訳の生成には 15 分から 60 分程度の時間が必要であった．また，あらかじめシステムの管理者が解説文生成システムに検索を要求しておき，用語 DB に登録を行っておかなければ利用できなかった．これに対し，本システムでは徹底したスレッド化による効率的な Web ページ収集により，説明文生成時間の短縮を図った．また，用語 DB に登録されていない用語が入力された場合には，即座に解説文の生成を行い，出力するように設計した．

説明文の抽出方法は HTML タグと説明文特有の文体などを利用し，特定のパターンに当てはまるものを説明文候補として抽出した．そして，抽出した各説明文候補に点数付けを行い，点数の高いものから 3 つまでを説明文として出力する．以下，説明文抽出方法と点数付けの方法を述べる．なお，本稿の点数付け方式の関連研究として，藤井らの研究による説明文要約方法 [4]があるが紙数の都合上，比較等は省略する．

（1）説明文抽出方法

・抽出する説明文候補と抽出時の点数付け

先行研究で用意されていた説明文を抽出する文章のパターンは日本語説明文で 18 パターン，英語説明文で 20 パターンののべ 38 パターンがあった．本システムでは従来のパターンに 8 パターンを追加し，46 パターンとしている．本システムの点数付けは，様々な専門分野から採録された 40 の用語(例: 平和維持活動，アポトーシス，など)を用いて最適化を図った．表 2, 3 にそれぞれ日本語説明文の抽出に用いる抽出パターンの例，英語説明文の抽出に用いる抽出パターンの例を示す．表中の〔キーワード〕は検索する用語とスペースから成る文字列を表し，〔キーワード＋α〕は検索する用語と，☆や※など，見出しなどに使われる特殊な記号を含んだ文字列から形成される．また，〔説明文〕は本システムが説明文として抽出する任意の文字列を表す．

表 2　日本語説明文の抽出に用いる抽出パターン

見出しがあるパターン	
説明文として抽出するパターン	抽出時の点数
〔キーワード＋α〕とは〔説明文〕	３０点
〔キーワード＋α〕とは〔説明文〕	１０点
〔キーワード〕 〔説明文〕	３０点
〔キーワード〕〔説明文〕	１０点

表 3　英語説明文の抽出に用いる抽出パターン

見出しがあるパターン	
説明文として抽出するパターン	抽出時の点数
〔キーワード＋α〕is〔説明文〕	３０点
〔キーワード＋α〕is〔説明文〕	１０点
What is〔キーワード〕 〔説明文〕	４０点
〔キーワード〕〔説明文〕	１０点

（２）説明文の点数付けの方法

本システムの説明文への点数付けは前述した抽出時の点数のほか，見出しの有無とその内容，抽出した説明文の内容，文字数の 4 項目により行われる．

・見出しによる点数付け

WWW 上の用語集などには説明文の他に見出しも記載されている場合が多い．本システムでは説明文抽出時に見出しと解釈される部分がれば見出しの内容により点数を加点する．表 4, 5 にそれぞれ日本語説明文用，英語説明文用の見出しによる加点例を示す．

表 4 見出しに点数を付加するパターン（日本語説明文用）

点数を加点するパターン	抽出したときの点数
〔キーワード〕とは	３０点
〔キーワード〕	２０点
〔キーワード〕＋５文字以内	５点

表 5 見出しに点数を付加するパターン（英語説明文用）

点数を加点するパターン	抽出したときの点数
What is〔キーワード〕	２５点
〔キーワード〕	１５点
〔キーワード〕 is	１０点

・抽出した説明文候補の内容による点数付け

抽出した説明文候補中に「[キーワード]とは，」などのような説明文特有のパターンが含まれている場合は説明文である可能性が高い．よって，このような表現が含まれている説明文にはその内容により加点を行った．また，「意味」「定義」「こと．」「ことである．」のような表現を含む文章には加点を行い，「私は」「わが社」「俺」などを含む文章は一般的な意味の説明文であることが少ないため減点を行っている．このような点数付けのうち日本語説明文，英語説明文に用いたものの例をそれぞれ表 6,7 に示す．

表 6 説明文の内容から点数を付加するパターン（日本語説明文用）

従来のシステムの例	
点数を加点するパターン	抽出時の点数
「〔キーワード〕とは，」で始まる文章	２５点
「〔キーワード〕は」で始まる文章	２０点
「〔キーワード〕が」で始まる文章	５点
本システムで付加したパターン	
「意味」「定義」を含む説明文	１５点
「我々」「わが社」「俺」「筆者」「私は」などを含む説明文	－２０点

表 7 説明文の内容から点数を付加するパターン（英語説明文用）

従来のシステムの例	
点数を加点するパターン	抽出時の点数
「〔キーワード〕 is～」で始まる文章	３０点
「〔キーワード〕 are～」で始まる文章	１５点
本システムで付加したパターン	
「definition」「meaning」を含む説明文	１５点

・抽出した文字数による点数付け

上述した方法で用語を抽出した場合，説明文の長いものがより適切な説明文であることが多かった．このことより，本システムでは抽出した文字数の多いものにボーナスを付加している．文字数による点数の加点の例を表 8 に示す．

（３）日英対訳の抽出方法

本システムでは説明文候補の抽出と同じように特定のパターンにあてはまる文字列を対訳候補と抽出している．本システムでは日本語から英語，英語から日本語それぞれの対訳を抽出するために各 25 パターンの抽出パターンを用いた．抽出に用いたパターンの例を表 9 に示す．

表 8 文字数による加点

見出しがないパターン	
	抽出時の点数
２０文字以下	-１０点
３０文字以上	１０点
（省略）	（省略）
２００文字以上	４５点

表 9 対訳候補として抽出するパターン

記号を利用したもの	
説明文として抽出するパターン	抽出時の点数
〔キーワード〕：〔対訳候補〕	７点
〔キーワード〕 /〔対訳候補〕	５点
〔キーワード〕,〔対訳候補〕	３点
括弧を利用したもの	
〔キーワード〕 【〔対訳候補〕】	２０点
〔キーワード〕 （〔対訳候補〕）	１５点
日本語を利用したもの	
〔キーワード〕とは英語で〔対訳候補〕	３５点
〔キーワード〕 英名：〔対訳候補〕	５０点

（４）日英対訳の点数付けの方法

上記の方法で抽出した対訳候補が既に対訳候補と認識されていれば，持ち点を統合する．また，キーワードまたは

対訳候補がカタカナで始まる場合，対訳の頭文字を予想することができる．予想される頭文字が対訳候補と一致する場合70点を加算している．表10に対応表の一部を示す．

表 10 頭文字に対して加算するパターン

検索対象語がカタカナの場合	
ア	a e i o u h A E I O H
イ	i e y j I E Y J
ウ	u w v U W V
検索対象語が英語の場合	
a A	ア
b B	バビブベボ
c C	チ

4．実装と評価

4．1 実装

本システムでは，リトリーバにおける検索用語の入力部分，及び検索結果の出力部分のインタフェース部はJSPを用いて実装した．また，リトリーバ内の用語DBへのアクセス部，メタWeb辞書部，Webページ抽出部はJAVAを用いて実装した．DBMSにはMySQLを用い，Webページの収集にはInfoSeekのgoogle検索を用いた．メタWeb辞書部で利用するWeb辞書にはYahoo辞書，EXCITE辞書，Goo辞書，アスキーデジタル用語辞典，IT用語辞典e-word，ケンブリッジ英英辞典を用いた．本システムではWeb辞書は，(1) 解説文の内容にほぼ間違いがないと想定されること，(2) ある程度の用語数が掲載されていること，(3) サービスの継続性を考慮し，管理者が個人ではなく会社規模の組織で運営されていること，の3項目を満たすものとした．

4．2 評価

評価は，本システムに用語を入力し得られた説明文及び対訳の正誤（検索精度）と説明文生成に要した時間（検索速度）の2つの項目で行った．評価には検索精度，検索時間に差が生じやすいWebページ抽出部のみを用いて行った．また，回線速度: ADSL 12 Mbps，CPU: Pentium4 2.4GHz，メモリ: 1GbyteのPC上で行い，入力する用語には初級システムアドコンパクト用語辞典 [5]及びカタカナ語辞典 [6]からランダムに選択した160語を用いた．上記の環境によるWebページ抽出型先行研究と本システムの実験結果を表11に示す．

表 11 実験結果

	検索精度(%)				検索速度(秒)
	説明文		対訳		
	日	英	日	英	
本システム	77	72	85	91	180
先行研究	75	72	85	92	1200

また，本システムでは用語を入力するだけで用語の意味を検索することができるため，検索難易度は低い．メタWeb辞書部の検索時間は短いが，抽出部の検索時間は従来のものより大幅に短縮できたとはいえ，短いとは言い難い．収録語数に関しては，メタWeb辞書部においては大幅には改善されていないが，Webページ抽出部ではWWW上のWebページに対して検索を行っており，全体としては十分多いと考えられる．更新速度は検索エンジンを利用した場合と同程度の速度を実現している．説明文精度は抽出部においては7割程度であり改善の余地があるが，メタWeb辞書部では精度の高い説明文が得られる．Web辞書，本システムのメタWeb辞書部，本システム全体の比較を表12に示す．

表 12 本システムの評価

	Web辞書	本システムのメタWeb辞書部	本システム全体
検索の容易さ	◎	◎	◎
検索時間	◎	◎	○
収録語数	×	△	○
更新速度	○	○	◎
説明文の精度	◎	◎	△～○
携帯性	○	○	○

（◎非常に良い，○良い，△悪い，×非常に悪い）

5．まとめ

本稿ではWWW広域検索型辞書システムの提案及び実装を行った．実装した環境に対して検索精度，検索速度の2つの観点から評価実験を行い，これらに加え，検索の容易さ，収録語数，更新速度，携帯性の4つの観点から従来の検索ツールとの比較を行った．その結果，本システムのWebページ抽出部は既存の検索ツールと同等の検索精度を維持し，大幅に検索時間を短縮できた．また，メタWeb辞書部においては，複数のWeb辞書の統合により収録語数を増やすことができた．更に，Webページ抽出部とメタWeb辞書部の統合により多くの用語の意味を調べることが可能とし，更に英語の専門用語の日本語説明文を取り出すといった対訳機能を実現している．結果として本システムは既存の検索ツールより高い利便性を実現している．

今後の課題として，Webページ抽出部の時間短縮，説明文の精度の向上が考えられる．時間短縮については，自動用語収集エンジンによりバックグラウンドで常時用語の説明文及び対訳を収集しデータベースに登録しておく手法が考えられる．説明文の精度については，形態素解析を行い説明文特有の文型パターンによる説明文の解析を行うことや，WWW上から用語集特有のタグパターンなどを分析し，用語集の可能性の高いWebページを抽出する手法の適用が考えられる．

参考文献

1) 南野朋之，奥村学: Web上の辞書を利用したメタ辞書の構築，情報処理学会第65回全国大会，3T6-1 (2003).

2) 桜井裕，佐藤理史: ワールドワイドウェブを利用した用語検索の実現，情報処理学会研究報告，2000-NL-137, pp.23-29 (2000).

3) 吉田卓哉: World Wide Webを利用した日英相互変換辞典の構築，岩手県立大学ソフトウェア情報学部卒業論文，(2003).

4) 藤井敦，渡邉まり子，石川徹也: 複数Webページの要約による用語説明の自動生成，情報処理学会研究報告，2004-NL-159，pp.31-38 (2004).

5) 高度情報化利用技術研究会（編）: 2000/2001年度初級シスアドコンパクト用語辞典，ピアソン (2000).

6) 徳川宗賢(編): ポケットカタカナ語辞典，集英社 (1999).