特集：Webアプリケーションを背景としたWeb利用

UDC 002.5(03)：025.4：621.396.74(100)：681.3.02

Web情報を用いた事典検索サイトの構築

藤井　敦*

インターネットを通じて様々な情報が配信されるようになり，知らない言葉や事柄に日常的に遭遇するようになった。しかし，既存の事典では新語や専門用語を調べられないことが多い。近年は，World Wide Webを事典の代用とする機会も増えてきた。しかし，既存の検索エンジンでは不要な情報が検索されたり，情報が整理されていないといった問題がある。本稿は，Webを事典的に活用するための新しい検索サイトについて紹介する。本サイトは，Webページ群を用いて事典を自動的に編纂・更新し，さらにユーザを飽きさせない様々な検索機能を備えている。

キーワード：World Wide Web，事典，検索サイト，情報検索，情報抽出，情報の組織化

1．はじめに

インターネットを利用して誰もが手軽に情報を発信できるようになったことが主な要因となり，「情報洪水」と呼ばれるほど大量の情報が氾濫するようになった。このように日々増え続ける情報に囲まれた生活環境の中で，我々は未知の言葉に日常的に遭遇する。

知らない言葉や事柄について調べるための情報源として，昔から国語辞典や百科事典がある。しかし，既存の辞典や事典は頻繁に改定されるわけではないため，日々生み出される新しい事柄や専門技術に関する言葉は収録されていないことが多い。また，既存の言葉に対する新しい定義は収録されておらず，既存の定義ですら全て収録されているわけではない。そこで，冊子体・電子版といった媒体の形式によらず「量的問題」が発生する。

それに対して，近年はWorld Wide Web上の検索エンジンを辞書のように使って調べものをすることが日常的になっている。Webには専門性が高い最新情報が存在するためである。Webが流行りはじめた当初に比べれば検索エンジンの性能は向上し，目的の情報が簡単に見つかることも多くなった。しかし，検索要求によっては，依然として何を入力すればいいのか分からないことや，膨大な検索結果から欲しい情報をどうやって選択すればよいか分からないことがある。またWebには統制がないため，誤字，誤解，嘘などの低品質の情報も存在する。そこで「質的問題」が発生する。

筆者は，Webを事典的に利用することを目的として，Webページ群から量質ともに優れた事典情報を構築する手法を提案した[1]。本手法は，Webに含まれる良質な説明情報を選択的に抽出し，さらに用語の意味や分野に応じて情報を整理する。本手法を用いて約60万語の見出し語を含む大規模な事典を自動的に編纂し，Web上のブラウザを使って利用するための検索サイトを構築した。本検索サイトは研究用の試験的なプロトタイプではなく，一般ユーザでも「使える」サイトを目指している。本稿は，事典の編纂手法と検索サイトの機能について説明する。

2．検索サイトの主旨

検索サイトを使えるものにするためには，多数のユーザがアクセスしてもサーバがダウンしないように耐久性を向上させるといったハードウェアに関する側面から，コンテンツ（事典）の品質向上や検索インタフェースの利便性向上といったソフトウェアに関する側面まで幅広い工夫の余地がある。本稿ではソフトウェアに関する側面からの工夫に焦点を当てる。

本サイトの主旨は，とにかくユーザを飽きさせないことである。ネットサーフィンが流行る主な理由は，マウスのクリックによってハイパーリンクをたどるだけで様々な情報を簡単に取得できる点にある。言い換えれば，それ以上先に進めないような行き止まりに陥ってしまうとユーザの不満は大きくなる。これは本検索サイトのユーザにも当てはまる。すなわち，

- どんな見出し語を入力すればよいか分からない。
- 入力した語が見出し語として登録されていないために何も検索されない。
- 検索された説明が分かりにくい，もしくは説明に

* ふじい あつし　筑波大学，科学技術振興事業団CREST
〒305-8550 茨城県つくば市春日1-2
Tel. 029-859-1401
（原稿受領　2003.3.10）

なっていない。

・情報が古くて役に立たない。

などの理由で検索行為の中断を余儀なくされると，ユーザは検索サイトの利用をやめる。約1,000人の被験者を対象にした調査の結果，見出し語のヒット率がユーザの満足度と強く関連することが分かった。

以上をまとめると，使える検索サイトを構築するためには，ユーザの入力に対して常に何らかの意味のある応答をし，またユーザが目的の説明を見つけた場合でも，次の検索へ自然に誘導するような仕組みが必要になる。

3．検索サイトの機能

3.1 概要

事典的Web検索サイトの概要を図1に示す。本稿執筆現在，見出し語数は約60万で既存の事典を凌駕する規模である。さらに，新語検出機能によって見出し語数は今後も増加する。以下，図1に基づいて，事典を編纂する過程（オフライン処理）とユーザが事典を検索する過程（オンライン処理）について説明する。

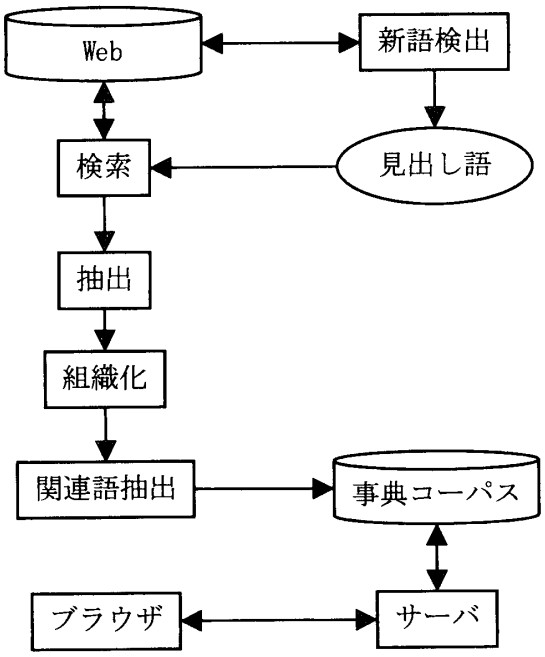

図1　事典的Web検索サイトの概要

3.2 事典の編纂手法

オフライン処理では，まず「新語検出」によって見出し語の候補をWebから自動的に収集する。ここでは，新聞サイトなどの更新頻度が高い「旬な」サイトから新しい言葉を取得し，事典の見出し語を拡張する。すなわち，従来の事典は一旦作ると何年も更新されないのに対して，本サイトの事典は継続的に自己進化を続けていくことが可能である。具体的には，新聞サイトから取得した記事に対して「形態素解析」と呼ばれる計算機処理を施して単語を抽出し，新しい単語の組合せ（「残留性有機汚染物質」など）を見出し語の候補として検出する。

見出し語の候補が検出されると，それぞれの候補に対して，「検索」「抽出」「組織化」を順番に実行する。

「検索」では，原理的には既存のWeb検索エンジンと同様の処理を行う。すなわち，見出し語を含むページを世界中から網羅的に集める。しかし，用語の説明はWebページ全体ではなく，一部（断片）に過ぎないことが多い。そこで，「抽出」によって見出し語を説明している段落を検出する。ここでは，用語の説明に固有の言語表現や文書構造を用いて説明段落を抽出する。言語表現は「○○とは△△である」などのパターンである。文書構造とは，Webページを記述しているHTMLタグの構造である。具体的には，段落を記述する「P」や，見出し語を記述する「H」などのタグを手掛かりにして説明段落を抽出する。抽出が終了すると，見出し語に関する説明が多数取得される。

しかし，人間が編纂する事典は，単一の用語に対して分野や意味に応じて説明を分類するなどして注意深く組織化されている。このような操作を計算機処理で代替するのが「組織化」の役割である。ここでは，テキストを一定のカテゴリ（分野やジャンル）に分類する手法を用いて，用語説明を専門分野に分類する。例えば，「パイプライン」という言葉は，コンピュータ分野では「処理の方式」，建築分野では「油送管」の意味で使われる。説明を分野に応じて整理することで，このような意味の違いを区別でき，ユーザは特定の意味に関する説明だけを調べることができる。

詳細は割愛するものの，組織化ではさらに，説明が信頼できるか，言葉として正しく書かれているか，説明らしいレイアウト（HTMLタグの構造）であるか，などの種々の基準によって説明の品質を評価し，良質の説明から順番にユーザに提示する。

最後に，「関連語抽出」によって見出し語に関連する言葉を説明段落から検出する。具体的には，良質の説明に頻出する語を優先的に抽出する。これらの語は，オンライン検索時にユーザの情報要求を絞り込んだり，一つの検索から次の検索に誘導するために用いる。

3.3 事典の検索方法

オンライン処理では，ユーザはWebブラウザを使って事典を検索する。図2は，見出し語として「ウォークスルー」を入力した場合の検索結果である。この用語には「ゲームの攻略法」「3次元仮想空間内の移動」「ソフトウェア開発工程の検証」「車の種類」「ゲート型の金属探知機」など多数の意味があり，図2では

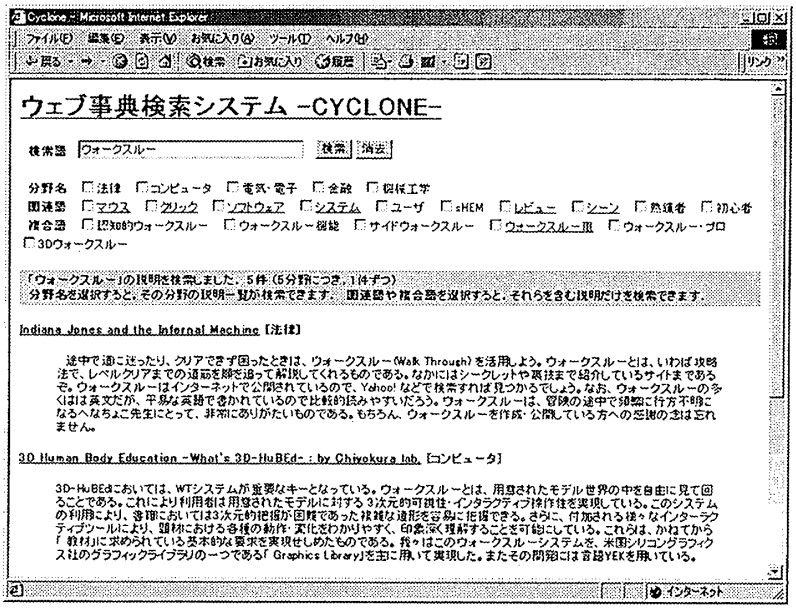

図2　入力語「ウォークスルー」に対する検索結果

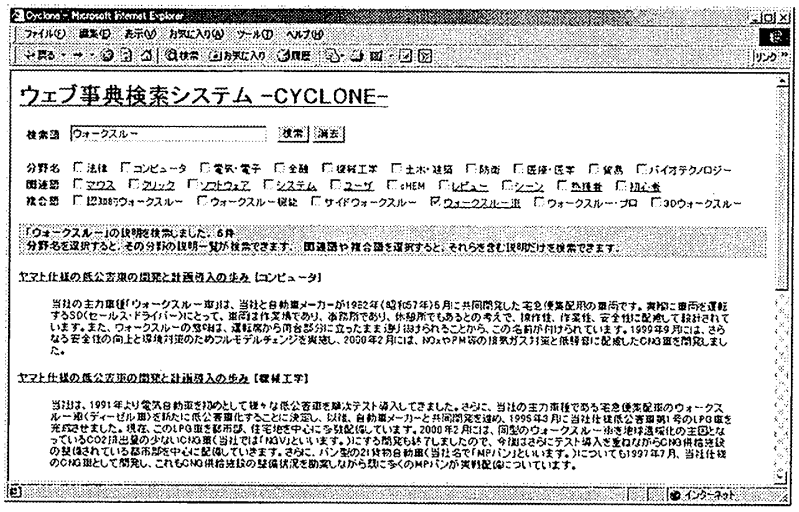

図3　「ウォークスルー車」で絞り込んだ結果

最初の2つの意味に関する説明が表示されている。

ここで，「分野名」「関連語」「複合語」から適当な項目を選択して（マウスでクリックしてチェックを入れて）再検索すると，選択された項目に該当する説明だけに絞ることができる。例えば「分野名」で「コンピュータ」を選択すれば，「ウォークスルー」の意味のうち，「3次元仮想空間内の移動」や「ソフトウェア開発工程の検証」に関する説明だけを見ることができる。

「関連語」は見出し語とともによく現れる言葉であり，ユーザの情報要求を絞り込んだり，説明の観点を変更するために有効である。例えば「マウス」を選択すれば，コンピュータ分野の説明のうち「3次元仮想空間内の移動」に絞り込むことができる。

「複合語」とは，見出し語を含む語である。複数の意味を持つ多義語は，前後に他の語を連結させることで意味が特定されることが多いため，複合語は多義性の解消に有効である。図3は「ウォークスルー車」を選択して再検索した例である。この例では，「ウォークスルー」と呼ばれる車種に関する説明だけが提示されている。

関連語や複合語が見出し語として登録されている場合はリンクが張られる（図2や図3で下線が付いている語）。そこで，ユーザはネットサーフィンと同じ要領でリンクをたどるだけで関連する言葉の意味を次々に調べることができる。

漠然とした要求はあるものの，何を見出し語とすればよいか分からないユーザ（場面）も想定している。例えば「電子メールに感染するもの」と入力すれば，図4の「概念検索で得られた用語」の欄に示されるように「マクロウイルス」のような見出し語を提示する。具体的には，説明情報に対する全文検索を行うことで，事典の逆引き機能を実装している。図5は，図4から「マクロウイルス」を選択し，クリックした場合の検索結果である。

また，入力した見出し語が事典にな

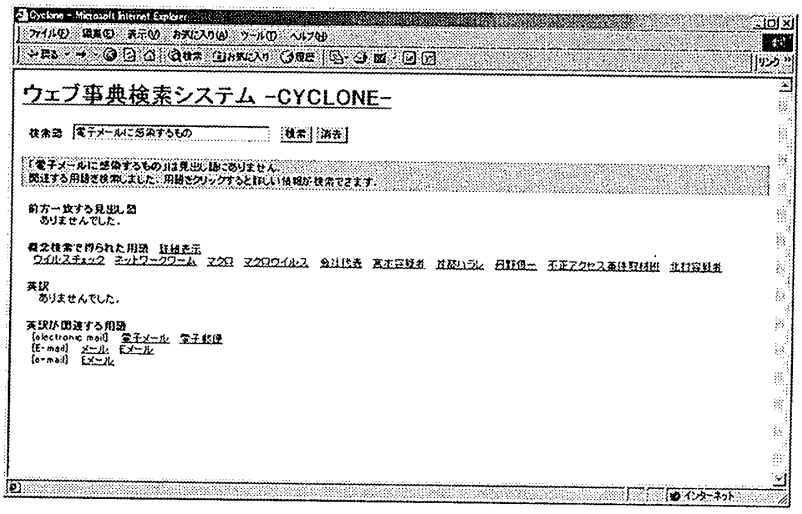

図4　「電子メールに感染するもの」を入力した場合の検索結果

い場合は，同義語を提示する機能を持っている。ここでは，和英辞書を用いて，同じ英訳に対応する日本語を同義語として検出する。図6は「レイテンシー(latency)」を入力した例である。一番下の行には「待ち時間」などの同義語が提示されており，これらの語を選択することで必要な情報を取得することができる。

4．おわりに

World Wide Webを事典のように活用するための検索サイトについて紹介した。本サイトの特長は，

- 見出し語が充実していてユーザの入力に対するヒット率が高い
- 新語検出機能によって新しい見出し語を自動的に登録する
- 入力した語が見出し語にない場合でも種々の補完機能によって意味のある情報を提示する
- 見出し語にならない漠然とした入力に対して具体的な見出し語を提示する

といった点にあった。今後は，試験運用を行った上で本サイトをWeb上で公開する予定である。

図5 「マクロウイルス」を選択した場合の検索結果

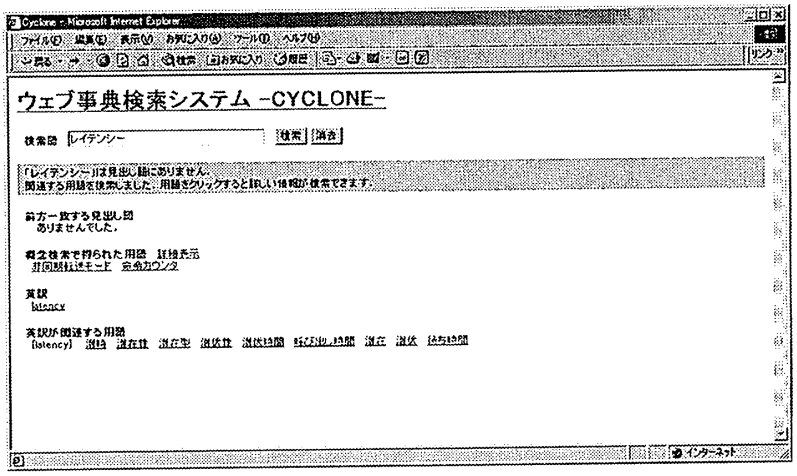

図6 「レイテンシー」を入力した場合の検索結果

謝　辞

本研究の一部は，情報処理振興事業協会の未踏ソフトウェア創造事業によって行われました。プロジェクトマネージャーの喜連川優先生（東京大学）からは多数の有益なコメントを頂きました。この場を借りて深謝致します。

参考文献

1）藤井敦，石川徹也．World Wide Webを用いた事典知識情報の抽出と組織化．電子情報通信学会論文誌，Vol. J85-D-II，No. 2，pp. 300-307，2002.

Special feature: Web Usage in the Context of Web Application. Using the Web to build an encyclopedic searching site, Atsushi FUJII (University of Tsukuba, CREST, Japan Science and Technology Corporation (1-2 kasuga, Tsukuba-shi, Ibaraki 305-8550))

Abstract: Given the growing number of contents available via the Internet, we usually encounter unknown words. However, it is often difficult to obtain descriptions for new words and technical terms in existing dictionaries and encyclopedias. The World Wide Web, which contains an enormous volume of up-to-date information, is a promising source to obtain new term descriptions. However, existing Web search engines often retrieve a number of extraneous pages and retrieved information is not organized as in hand-crafted encyclopedias. This paper describes a new searching site in which users can utilize the Web as an encyclopedia. Our site automatically compiles and updates an encyclopedia using Web pages, and has a number of useful retrieval functions.

Keywords: The World Wide Web / Encyclopedias / Searching sites / Information retrieval / Information extraction / Information organization