CYCLONE: 事典的 Web 検索サイト

CYCLONE: An Encyclopedic Web Searching Site

藤井　敦 [1)]　伊藤　克亘 [2)]　秋葉　友良 [3)]
Atsushi FUJII　Katunobu ITOU　Tomoyosi AKIBA

1) 筑波大学 図書館情報学系（〒305-8550 つくば市春日 1-2　E-mail: fujii@slis.tsukuba.ac.jp）

2) 名古屋大学大学院情報科学研究科 メディア科学専攻（〒464-8603　名古屋市千種区不老町 1
E-mail: itou@is.nagoya-u.ac.jp）

3) 産業技術総合研究所 情報処理研究部門（〒305-8568 つくば市梅園 1-1-1　E-mail: t-akiba@aist.go.jp）

ABSTRACT. The World Wide Web, which contains an enormous volume of up-to-date information, is a promising source to obtain encyclopedic knowledge. We built a searching site, in which users can utilize an encyclopedic corpus consisting of descriptions for 600,000 entry words. For this purpose, we propose a method to automatically produce encyclopedic corpora from the Web. First, we analyze Web pages to extract new words, which are used as entry words in our corpus. Second, we search the Web for pages containing new entry words. Third, we analyze the HTML layout of retrieved pages to extract paragraph-style fragments that potentially describe the term. Finally, we quantify the extent to which each fragment correctly describes the term, based on several criteria, including reliability and language properties. Page fragments associated with high scores are selected as descriptions in a resultant corpus. In addition, descriptions are organized on the basis of domains and word senses, so that users can selectively obtain descriptions in specific domains.

1　背景と目的

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を主な要因として，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，既存の言葉に対して新しく作られた意味や用法は収録されておらず，既存の意味ですら全て網羅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態によらず「量的問題」が発生する．

最近では，World Wide Web を一種の辞書のように使って，知らない言葉や事柄を調べることが一般的になってきた．Web には既存の情報源にはない多種多様な情報が存在するからである．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．しかし依然として，検索要求によっては，どのようなキーワードを入力すればいいのか分からない場合や，膨大な検索結果から欲しい情報だけをどうやって選択すればよいか分からない場合がある．また，検索された個別のページは互いに関連性がなく，人間が編纂する事典のように情報が整理されていない．さらに，Web には統制がないため，誤字，誤解，虚偽といった低品質の情報も存在する．すなわち「質的問題」が発生する．

本稿は，上記の「量的・質的問題」を解決して Web を事典的に利用することを目的とした検索サイト「CYCLONE」について紹介する [3, 4, 5, 15, 16, 17] ．CYCLONE の基盤技術は，Web から大規模かつ高品質の事典（コンテンツ）を自動構築する編纂機能である．具体的には，言葉や事柄に関する説明情報を Web ページから抽出し，良質な情報だけを選択して専門分野（コンピュータ，金融，医療など）ごとに分類する[*1]．その結果，ユーザは入力したキーワードに関する説明を分野ごとに閲覧することが可能である．それと同時に，検索サイトを継続的に運用し，さらに「使える」サイトとして位置付けるためには，構築したコンテンツを活用するための検索機能の開発やインタフェースの設計などにも取り組まなければならない．

2　CYCLONE の主旨

検索サイトを「使える」ものにするためには，サーバの耐久性向上，ページのリンク切れに対応するキャッシュの充実といったハードウェアに関する側面から，事典コンテンツの品質向上，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．本章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．

CYCLONE の検索サイトとしての主旨は「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辿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ると，ユーザの不満は大きくなる．これは，本サイトのユーザにも当てはまる．すなわち，

[*1] 辞典と事典の違いは，辞典が言葉を中心に収録しているのに対して，事典は事柄（「帝銀事件」など）も収録している点にある．CYCLONE は言葉と事柄の両方を見出し語の対象としているので「事典」的検索サイトと呼んでいる．

- 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動の中断を余儀なくされた場合，ユーザは検索サイトの利用をやめる．事実，約 1000 人の被験者を対象にした調査の結果，入力した語に対する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ている [15].

以上をまとめると，ユーザの入力に対して常に何らかの意味のある応答を返すことが必要である．また，ユーザが一つの用語に関する説明を見つけて当初の目的を達成した場合でも，別の用語の検索へ自然に誘導するような仕組みが必要である．

3 CYCLONE の概要

CYCLONE は，見出し語の対象となる語を集めて事典コンテンツを自動的に構築する機能と，構築されたコンテンツをユーザインタフェースによって利用に供する検索機能で構成されている．

CYCLONE でどのような検索が可能であるかを直感的に理解することは，後で説明する技術的な詳細を理解する上でも重要であろう．原理的には，Web ブラウザ上で動作するインタフェースに見出し語を入力すると，その語に対する説明が分野ごとに整理されて表示される．図 1 は「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 1 には最初の 2 つの意味に関する説明が表示されている．これらの説明は，Web ページから段落単位で抽出（抜粋）された情報である．各説明の上にある下線付きの文字列は，抽出元のページタイトルであり，この部分をクリックすることでページ本体にジャンプできる．すなわち，一般の Web 検索エンジンにおいて，検索されたページタイトルとサマリ（要約）を表示する手法と同じである．

検索語を入力するボックスの下にある「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り込むことができる．例えば「分野名」として「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使われる言葉であり，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」は見出し語を含む関連語である．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．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（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．

4 事典コンテンツの構築

(1) 概要

CYCLONE のシステム構成を図 2 に示す．この図に基づいて事典コンテンツの構築手法について説明する．まず「新語検出」によって見出し語の候補を Web から収集する．次に，それぞれの見出し語候補に対して「検索」「抽出」「組織化」を順番に実行することで説明を取得し，専門分野ごとに分類する．そこで「パイプライン（処理方式/油送管）」のように分野によって意味が異なる多義語の説明を区別することができる．ただし，全ての見出し語候補に対して必ず説明が得られる訳ではない．

検索処理は，通常の Web 検索エンジンと同様に，見出し語を含むページを Web から網羅的に集める．ただし，集めたページの全てが見出し語に関する説明を含んでいる訳ではない．また，説明を含んでいる場合でも，ページ全体ではなく，特定の範囲に限定されていることが多い．そこで，抽出処理は，HTML タグを用いて見出し語に関する説明を段落の単位でページから抽出する．その結果，見出し語に関する説明の候補が多数集められる．さらに，組織化処理によって，分野分類と順位付け（ソート）を行い，分野ごとに質の高い説明を選択する．最後に「関連語抽出」によって見出し語を特徴付ける語を取得する．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むために利用する（3 章参照）．

(2) 新語検出

CYCLONE で利用されている事典コンテンツの見出し語数は，本稿執筆現在，60 万語に達している．そこで，現状でも十分に有用性が高い資源である．しかし，更新頻度が高いという Web の特徴を利用すれば，Web から新しい見出し語を継続的に取得して事典コーパスを更新し，有用性をさらに高めることができる．言い換えれば，既存の事典が一度作ると長期間改訂されないのに対して，CYCLONE では事典コンテンツの内容を短いサイクルで自律的に更新することが可能である．

CYCLONE における「新語」とは，事典コンテンツの見出し語として登録されていない語に関する便宜上の総称である．実際には，新語は以下のように分類できる．

a) 既存の語であるにも拘らず，事典の見出し語として登録されていない語
b) 新しく作られた言葉（「ムネオハウス」など）
c) 異表記（「ら致」と「拉致」など）
d) 既存の語に対する新しい用法（後部座席から前方の座席に車内で移動できるようにする機能やその機能を持った乗用車を指す「ウォークスルー」など）

また，不特定多数の著者が執筆した不均質な Web ページ群を対象にするため，文書の不備（誤字，脱字など）に起因して，未知語問題が生じることもある．

このように様々な側面のある新語のうち，CYCLONE では表記上の「新語」だけを検出の対象とする．そこで，上記のうち，a，b，c だけを対象とする．d は，既収録語に対して，図 2 における「検索」以降の処理を定期的に行うことで，説明情報を更新することで対処している．

日本語では語の検出（認定）自体が困難であり，様々な手法が提案されている [8]．特に，事典の見出し語に特有の語構成を考えた場合は，語は単独の形態素で構成される場合だけでなく，複合名詞に代表される複合語や複雑な構造を持つフレーズの場合もある（人工知能分野における「説明に基づく学習」など）．また，近年は b に分類される新語として「ちょぼら（ちょっとしたボランティア）」「モーニング娘。（句点が語の一部として含まれている）」「１０９（デパート名）」のような特殊な語が増えており，新語検出を一層困難にしている．

CYCLONE では，語の説明を取得するために，1 日ごと，1 週間ごと，1ヶ月ごとなど期間を定めて，その間に新たに収集した文書集合から新語を検出する．継続的に語彙を拡張しながら検出することが前提となるため，既知語のリストを参照しながら，未登録の新語を検出す

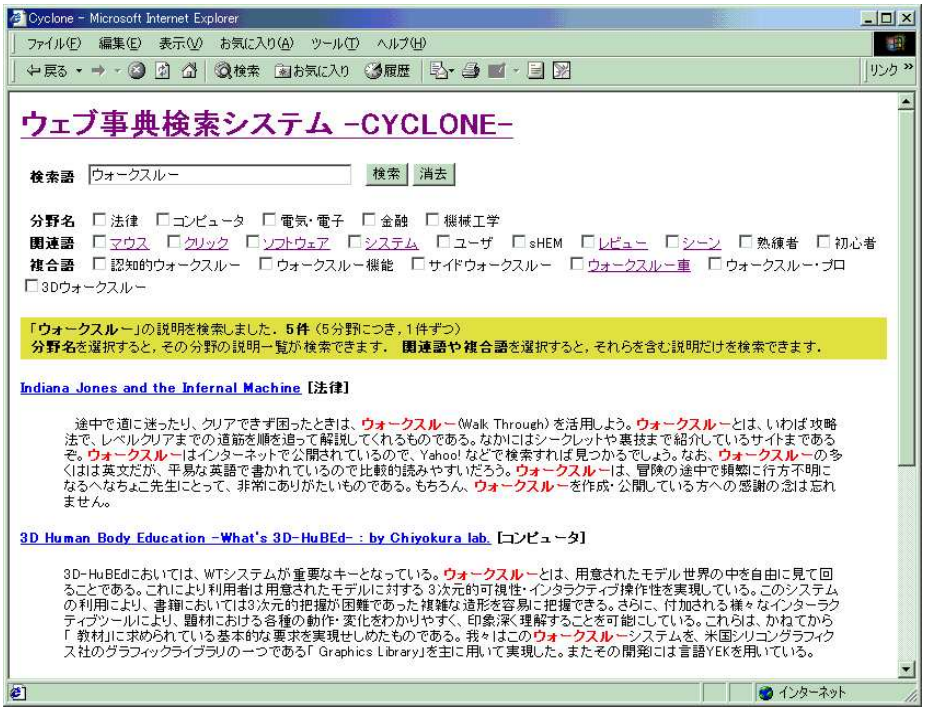

図 1: 入力語「ウォークスルー」に対する検索結果

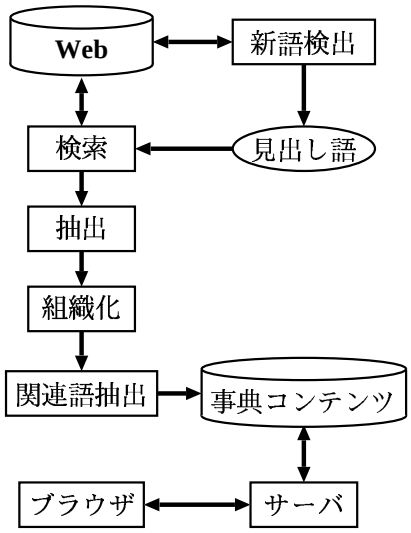

図 2: CYCLONE のシステム構成

る．そこで，以下に示す手順によって，既知語リストを用いて新語を検出する．

1) 新語検出の対象となるページを形態素解析する．前処理として，HTML タグを削除し，さらに半角文字を全角文字に変換して統一する．
2) 形態素列から，名詞（数字を含む），未知語，記号（「・」など）の連続を新語候補として抽出する．この処理では「説明に基づく学習」のようなフレーズは抽出されない．他方で「03-3333-3333」のような電話番号や単なる数字（列）が抽出されるため，新語候補数は膨大になる．
3) 対象とする文書全体の抽出結果から，新語候補の頻度リストを作成する．
4) 頻度リストから，既知語リストに登録されている語を削除する．この段階で大半の語は削除される．
5) 4 で作成した頻度リストから，低頻度語を削除する．ただし，低頻度語も既知語リストには追加する．そこで，ある期間内に低頻度で出現した形態素列は，当該期間に関しては，低頻度であるために削除され，当該期間以降は既知語リストに登録されているために新語候補として検出されない．その結果，連絡先の電話番号や数字を削除でき，さらに，誤字，脱字，形態素解析誤り（の一部）も削除できる．

意味のない新語候補は，後続の処理で説明情報が得られない可能性が高いため，その段階で淘汰される．そこで，新語検出の段階では精度よりも網羅性を重視している．また，多様な見出し語を検出するため，字種や長さに関する制限は加えない．他方において，不要な候補を増加させると，後段の処理効率を低下させてしまうため，運用におけるコストを考慮する必要がある．

過去 10 年間の新聞記事（CD-ROM 版）を用いたシミュレーションの結果，月平均で 800 語近い新語を約 84%の精度で検出することに成功している [16].

(3) 検索

検索処理では，既存の Web 検索エンジンと同じように，対象用語を含むページを網羅的に収集する．原理的には，既存の検索エンジンを利用することが可能である．しかし，多数の用語を用いた大規模な実験を行うためには，通信などのコストが膨大になる．そこで，独自にページ収集を行い，検索エンジンを実装した．収集し

たページは「茶筌」[*2]を用いて形態素解析して単語に分割し，単語単位で索引付けを行い，キーワードによる検索を可能にしている．

(4) 抽出

一般に，用語の説明はページ全体ではなく特定の一部分であることが多いので，抽出規則によって特定の領域を抽出する．説明の範囲を正確に特定することは容易ではないため，段落を説明の単位と見なす．具体的には，HTML のタグ構造を利用してページのレイアウトを解析し，対象用語を含む段落を抽出する．

HTML の規格では，段落を示すタグとして <P> がある．しかし，著者によって記述スタイルに揺れがあり，実際には様々なタグが段落の表現に使用される．そこで，現在は以下に挙げるようなタグを段落の特定に使用している．

> <BLOCKQUOTE>,
, <DD>, <DFN>, <DIV> <LI>, <P>, <TD>, <TT>

対象用語を含んでいなくても説明としてふさわしい段落もある．例えば，用語解説ページ特有のレイアウトとして，用語を強調して見出し語化し，後続の段落でその意味を説明する記述形式がある．このような形式に対処するために，見出し語化に使われやすいタグを定義し，対象用語にそれらのタグが付いている場合には後続の段落を説明の候補として抽出する．現在，見出し語化のタグとして，以下に挙げるタグを使用している．

> <B>, <DT> <EM>, <FONT>, <Hx>, <I>, <STRONG>, <TT>, <U>

ここで，<Hx> における x は文字サイズを制御する数値である．

なお，見出し語表現としては,「ウォークスルー」のように対象用語だけが書かれている場合だけでなく「ウォークスルーとは?」のような典型的な表現も考慮する．

(5) 組織化

抽出処理が終了した段階では「対象用語が説明されている可能性が高そうな段落」が収集されて，雑然と並んでいるだけである．そこで，組織化処理によって，対象用語が適切に説明されている段落を選択し，さらに既存の事典と同じように語義や分野に応じて分類を行う．実際には，語義と分野は関連していることが多いので [14]，分野を分類することで間接的に語義を区別する．そこで，個別の用語ごとに語義を定義する必要はない．

まず，説明の都合上，対象用語が関連する 1 つ以上の分野があらかじめ分かっていると仮定する．組織化処理の目的は，ある関連分野 c に対して最適な用語説明 d を選択することである．確率的手法の観点から捉えれば，c に対して $P(d|c)$ を最大化する d を選択することに相当する．

実際の処理では，まず全ての分野に対して $P(d|c)$ を計算し，$P(d|c)$ の値がある閾値以上の d だけを選択する．その結果，対象用語が関連する分野と，適切な用語説明を同時に特定することができる．そこで，はじめに仮定したように，対象用語が関連する分野をあらかじめ知っておく必要はない．

ここで，ベイズの定理によって式 (1) が成り立つ．

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \tag{1}$$

式 (1) の右辺において，分母 $P(c)$ は定数として扱うため，分子だけが組織化処理の中核である．$P(c|d)$ は用語説明 d が分野 c に関連する度合を定量化し，$P(d)$ は d が用語説明（事典情報）として妥当である度合を定量化する．両者をそれぞれ「分野モデル」「事典モデル」と呼ぶ．すなわち，特定の分野との関連度が高く，かつそれ自身が（分野とは無関係に）用語説明として妥当であるような情報が優先的に選択される．

用語説明を一般の文書と見なせば，既存の文書分類（categorization）手法によって分野モデルを実現することができる．本研究では，確率的な分類手法 [7] を用いて，式 (2) によって $P(c|d)$ を推定する．

$$P(c|d) = P(c) \cdot \sum_i \frac{P(w_i|c) \cdot P(w_i|d)}{P(w_i)} \tag{2}$$

ここで，$P(w_i|d)$，$P(w_i|c)$，$P(w_i)$ はそれぞれ，d，c，全分野から無作為に選んだ単語が w_i である確率を表す．単語として，形態素解析によって得られる名詞を利用する．$P(c)$ は定数として扱い，$P(w_i|d)$ は用語説明における語の統計情報を用いて最尤推定法で計算する．

$P(w_i|c)$，$P(w_i)$ の計算には，複数の分野に関する語の統計情報が必要である．しかし，分野が付与された大量の文書は入手や作成が高価なので，クロスランゲージ社（旧ノヴァ社）が機械翻訳用に作成した専門語辞書を利用している[*3]．この辞書は，合計約 100 万件の日英対訳を定義しており，以下に示す 19 の専門分野を分野モデルの推定に利用する．

> 航空・宇宙，バイオテクノロジー，ビジネス，化学，コンピュータ，土木・建築，防衛，地球環境，電気・電子，原子力・エネルギー，金融，法律，数学・物理，機械工学，医療・医学，金属，海洋・船舶，プラント，貿易

本辞書の日本語項目から「茶筌」を用いて単語を抽出した．日本語の用語説明は英単語を含むこともあるので，英語項目も併用して $P(w_i|c)$，$P(w_i)$ を計算した．

辞書は，項目に関する頻度情報を含んでいないため，統計情報の抽出には適さないことが多い．しかし，専門語辞書の項目は複合語が多いため，複合語を構成する単語の頻度分布を抽出することは可能である．また，分野特有の単語は多くの複合語に含まれやすいので，本手法は妥当な近似である．今後は，技術分野だけでなく，スポーツや芸能などの分野も導入する必要がある．

事典モデルを実装するためには，用語説明としての妥当性について検討する必要がある．まず，言語的な妥当性がある．対象の用語について説明していない抽出結果は排除する必要がある．次に，情報の信頼性に関する妥当性がある．一般の出版物に比べると，Web ページは誤りや虚偽を含むことが多い．そこで，言語的に妥当であっても，信頼性が低い用語説明は排除しなければならない（尤もらしい嘘をつくことは可能であるため）．さらに，抽出処理で利用した HTML のレイアウト構造に基づいて，用語説明らしい段落を特定することも可能である．

以上の検討から，式 (3) に示すように，事典モデルを言語モデル $P_L(d)$，信頼性モデル $P_R(d)$，レイアウト（構造）モデル $P_S(d)$ に分解する．

$$P(d) = P_L(d) \cdot P_R(d) \cdot P_S(d) \tag{3}$$

言語モデルを作成するために「茶筌」を用いて日立デジタル平凡社「CD-ROM 世界大百科事典」（約 8 万語収録）を単語に分割し，CMU-Cambridge ツールキット [2] を用いて単語のトライグラムモデルを学習した．トライ

[*2] http://chasen.aist-nara.ac.jp/

[*3] http://www.nova.co.jp/

グラムモデルとは単語の生起が直前の 2 単語にのみ依存すると考えた言語モデルであり，統計的な機械翻訳や音声認識に応用されて成果を上げている．ここで，説明対象用語（見出し語）の表層的な違いは重要ではないので，見出し語を事前に共通の特殊記号に置換した．また，d に含まれる対象用語も同様の特殊記号に置換する．そこで，事典によく現れる表現（「X とは」や「と呼ばれる」など）を含む段落ほど高いスコアが与えられる．

信頼性モデルを作成するために，Google[*4] でページの順位付けに使用されている PageRank [1] を計算するモジュールを実装して利用している．すなわち，Web 上における参照（リンク）関係を利用して，あるユーザが各ページに到達する確率（PageRank 値）を計算する．直感的には，権威のある（評判の良い）ページから参照されているページほど PageRank 値は高くなる．権威のあるページとは，別の権威あるページから参照されているようなページである．ただし，PageRank 値はページに対して計算される値である．そこで，$P_R(d)$ の値として，説明 d が抽出されたページの PageRank 値を与える．

レイアウトモデルを統計的に実現することは困難であるため，現状では経験的な値によってレイアウトモデルを実現している．具体的には，対象用語が HTML タグによって見出し語として扱われている段落には 1 を与え，それ以外の段落には 0.5 を与える．

以上まとめると，組織化処理は，分野モデル，言語モデル，信頼性モデル，レイアウトモデルの合計 4 つの確率モデルを利用する．約 2000 語の情報処理用語を用いた評価実験の結果，4 つモデルを全て用いた場合に，用語説明の検索精度が最も良くなった [15]．

(6) 関連語抽出

関連語抽出の基本原理は，各用語の説明情報（段落）に頻出する語を検出することである．そこで，適切な語の単位を検出する処理と，検出した語と見出し語の関連度を評価する尺度が必要になる．そこで，まず段落を「茶筌」で形態素解析して，品詞情報に基づいて（複合）語を構成し，関連語の候補とする．具体的には，名詞，動詞連用形，未知語，記号の連続を語として抽出する．さらに，段落における出現頻度と抽出元の段落に対する組織化のスコア（式 (1) 参照）を統合して関連語をソートし，上位の関連語から優先的に提示する．すなわち，良質の説明でよく使われる語が優先される．

実際は，見出し語を含むかどうかによって関連語を 2 種類に分類する．以降では，見出し語を含む語を「複合語」，含まない語を「関連語」と呼ぶ．図 1 では，見出し語「ウォークスルー」に対する関連語と複合語が評価値の高い順に左から表示されている．

複合語は，例えば「ウォークスルー」に対する「ウォークスルー車」である．多義語は，前後に連接する語によって語義を特定できる場合があるため，複合語による情報要求の絞り込みは有効である．

関連語とは，例えば「ウォークスルー」と共出現する「マウス」や「レビュー」などの語である．これらの語は複合語と同様に多義性を解消する目的のために有効である．さらに，同じ語義に関する説明であっても，説明の観点が変わる場合がある．関連語はこのような観点の違いを区別するためにも有効である．例えば「特許法」という見出し語に対して「特許料」「存続期間」のような関連語が提示される．これらの関連語は特許法の説明を異なる観点から調べるために有効である．また「RISC」に対する関連語として「CISC」を選択すると，両者の相違点に関する説明を調べることができる．

[*4] http://www.google.com/

5 事典コンテンツの検索

事典コンテンツ検索の基本原理は，3 章で説明したように，見出し語を入力して分野ごとに説明を取得することである．

しかし，ユーザが入力した語が見出し語として収録されていなかったり，何を入力すればよいか分からない場合もある．そういった場合でも意味のある応答をするために，種々の補完機能を実装している．

まず，文字列の部分一致検索によって，ユーザの入力語と事典の見出し語を柔軟に照合する．その結果,「コンピュータ」と「コンピューター」のような表記の揺れに対処することが可能になる．

また，ユーザの入力語に対する同義語を提示する機能を持っている．ここでは，組織化処理における分野モデルの作成に利用した「専門用語辞書」((5) 章参照）を用いて，同じ英訳に対応する日本語を同義語として提示する．例えば「レイテンシー（latency）」を入力すると「待ち時間」のような同義語が提示されるので，これらの語を選択して説明を閲覧することで，必要な情報を間接的に取得することができる．

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,「マクロウイルス」のような見出し語が提示されるので，それらの語を選択してクリックすることで説明を閲覧することができる．ここでは，事典の「逆引き」を行っている．具体的には，用語説明を個別の文書と見なして単語単位で索引付けし，確率型の検索手法 [13] によって，入力された検索質問（単語，句，文など）に適合する用語を確率スコアの高い順番に提示する．図 3 は，質問として「電子メールに感染するもの」を実際に入力した場合の検索結果である．ここで「概念検索で得られた用語」の欄に，関連する見出し語が複数提示されている．図 4 は，図 3 から「マクロウイルス」を選択した結果である．

このように，様々な検索機能を併用することで，常にユーザに意味のある応答をし，さらに関連語どうしにリンクを張ることによって，一つの検索から別の検索への自然な誘導が可能になった．

6 本研究の位置付け

CYCLONE の構築を研究と捉えた場合，複数の観点からそれぞれ異なる位置付けが可能である．

自然言語処理の観点からは，言語知識の獲得と見なすことができる．既存の研究では，文法，概念体系，対訳などの獲得がある．Web を対象とした研究では，対訳関係にあるページの抽出がある [12]．

情報検索の観点からは，目的指向型検索 [11] の一種と見なすことができる．CYCLONE のユーザは，ある用語の意味が知りたいという目的を持っている．

人工知能の観点から解釈すれば，ユーザにとって有用な情報を Web から集める知的エージェントや Web マイニングとも関連する．Web マイニングにおける処理として，コンテンツ解析，リンク解析，ログ解析がある [9]．CYCLONE ではコンテンツ解析を中心に行っている．しかし，信頼性モデルの計算においてページ間のリンク関係を利用しているという点で，リンク解析も行っている．CYCLONE のユーザが増えれば，ログ解析も可能になるだろう．

また，CYCLONE で構築した事典コンテンツは，人間が利用するだけでなく，事典情報を用いたシソーラスの構築 [6] や質問応答システム [10] のような計算機用辞書としての応用も可能である．通常これらの計算機処理

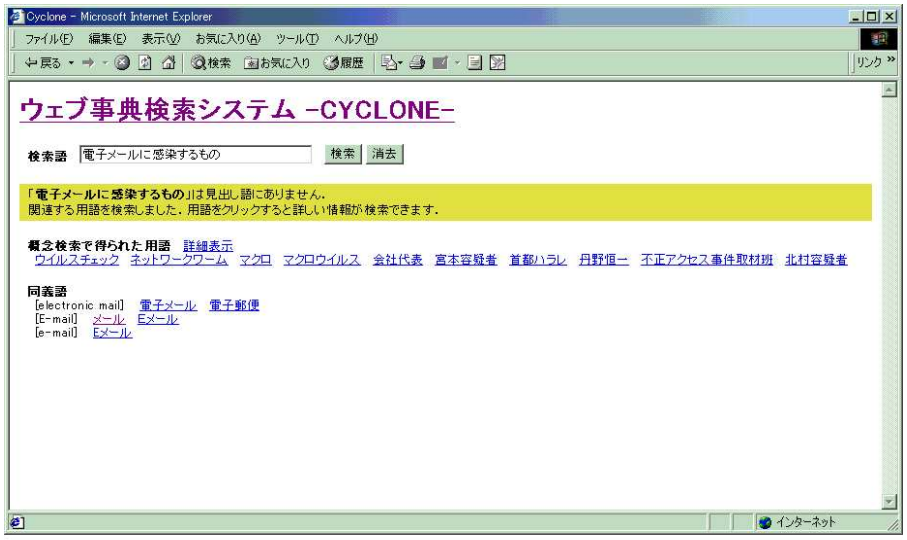

図 3: 「電子メールに感染するもの」に対する検索結果

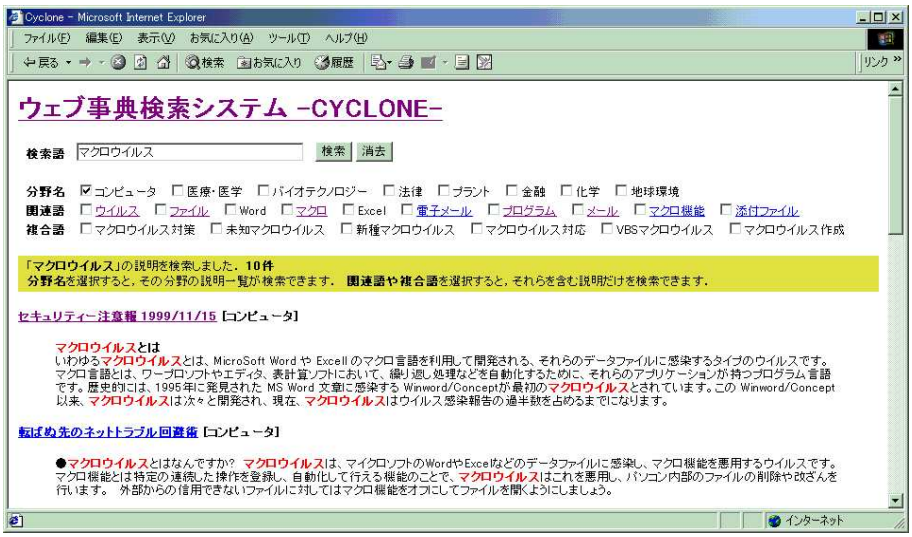

図 4: 「マクロウイルス」を選択した場合の検索結果

は事典の規模が小さいことがが障害になるものの，その障害は今後解消されることが期待できる．事実，情報処理技術者試験を対象にした自動回答システムにおいて，事典コンテンツを知識源として利用し，その有効性を実験によって評価した [4].

7 本事業のまとめと今後の課題

World Wide Web を事典のように活用するための検索サイト「CYCLONE」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．

CYCLONE は，Web 上の検索エンジンとして一般ユーザが使うだけでなく，6 章で説明したような分野の研究を促進するためのプラットフォームとしての有用性も高い．今後は，ユーザのログ解析などを通して，より利便性の高い検索サイトへ改善することが課題である．

8 参加企業及び機関

財団法人京都高度技術研究所
（プロジェクト実施管理組織）
契約名: CYCLONE: 最強事典サイトの構築

9 謝辞

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．

10 参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[2] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech'97*, pp. 2707–2710, 1997.

[3] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 488–495, 2000.

[4] Atsushi Fujii and Tetsuya Ishikawa. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 196–203, 2001.

[5] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Producing a large-scale encyclopedic corpus over the Web. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1737–1740, 2002.

[6] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539–545, 1992.

[7] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 162–167, 1994.

[8] Kyo Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology*, Vol. 3, No. 2, pp. 259–289, 1996.

[9] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *SIGKDD Explorations*, Vol. 2, No. 1, pp. 1–15, 2002.

[10] Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 181–189, 1993.

[11] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A machine learning approach to building domain-specific search engines. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 662–667, 1999.

[12] Philip Resnik. Mining the Web for bilingual texts. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527–534, 1999.

[13] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[14] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[15] 藤井敦, 伊藤克亘, 石川徹也. WWW は百科事典として使えるか? –大規模コーパスの構築–. 情報処理学会研究報告, 2002-NL-149, pp. 7–14, 2002.

[16] 藤井敦, 伊藤克亘, 秋葉友良. 事典的 Web 検索サイトの構築. 言語処理学会第 9 回年次大会発表論文集, pp. 129–132, 2003.

[17] 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.