解説

WWW上の検索サービスの技術動向

林　良彦　　小橋喜嗣
NTT情報通信研究所

「検索の凡人」

今やWWW上の検索サービスは，仕事や娯楽のための情報収集にとって不可欠な道具となっているといってもよいだろう．Webブラウザといくつかの定番といわれる検索サービスを使いこなすノウハウがあれば，たとえば「フランスワールドカップでの日本代表の成績」を調べることができる．さらに多少の語学力があればもっとよい．現地の言語で書かれた細々した情報だって探し出すことができるだろう．実際，インターネット関連の雑誌や書籍では，さまざまな検索サービスの紹介や使いこなし方の記事が頻繁に取り上げられている（たとえば文献1)）．また，「検索の鉄人」なるイベントがWWW上で開催され，もっと複雑な情報をいかに効率よく検索するかのテクニックが競われたりしている．しかしながら，このようなコンテストが開催されること自体，現状の検索サービスの道具としての未成熟さを示しているともいえないであろうか☆．たとえば，多くの「検索の凡人」達（失礼!）は以下のような経験を持っているに違いない．

- （凡人の不満1）いくつかの検索サービスを渡り歩く，あるいは，使い分けないと適切な情報が得られない．
- （凡人の不満2）一度の質問で的確な情報を得ることが難しい．質問によっては非常に多くの検索結果が得られるが，すべてを見て回れるほど暇じゃない．
- （凡人の不満3）せっかく世界中にさまざまな情報があるのに，言葉の壁が邪魔をして，欲しい情報を得ることができない．あるいは，それらしい情報を見つけても読みこなすことができない．

さて，WWW上で利用できる検索サービスは，大きく2つのタイプに分類されることはすでに多くの方がご存じであろう．1つは，Yahoo!などに代表されるタイプで，ディレクトリサービスなどとも呼ばれる．もう1つは，ロボット型サーチエンジン（以下，サーチエンジンと略する）などと呼ばれるタイプである．代表例としては，AltaVistaやgooなどがある．以下では，ロボット型サーチエンジンを中心に，上記のような凡人の不満を解決するためにどのような試みがなされているのかを技術的な観点から紹介する．

連携プレーで多くの情報をカバーするサーチエンジン

凡人の不満1は，まさにWWW上の情報検索に特有の問題であるといえる．通常の方法でアクセスできるページの数は1997年12月の時点で3億2千万にものぼると推定されている[2]．この中のどこかに「私の欲しいあの情報」が埋もれているかもしれないのだ．サーチエンジンの価値を決める観点はいろいろあろうが，まずは，どれだけ多くのページが検索可能かが1つのファクタであることは間違いない．実はこの問題は，サーチエンジンをどのように構成するかというアーキテクチャの問題と深く関わっている．

●ロボット型サーチエンジンの原理

まずサーチエンジンの原理を示すために，**図-1**に最も基本的な構成を示す．ロボットは，WWW上の情報リソース（以下，適宜ページとも呼ぶ）に埋め込まれたハイパーリンクを再帰的に辿り，ひたすら情報リソースを収集する．出発点のページさえ適切に選べば，膨大な数のページがいずれ収集できるはずである．ひとたび，検索対象たるページ群が収集できれば，あとは通常のテキスト情報検索システムの構築法（たとえば文献3)）と変わるところはない．すなわち，インデクサを用いてインデックスファイルを構築しておけば，実際の検索処理は，インデクサと対になったテキスト検索エンジンが処理してくれる．検索サービスゲートウェイは，（後述するような知的な処理を実現しないのであれば）ユーザとテキスト情報検索エンジンの間のデータ形式などの変換を行うことで事足りる．実に単純な仕組みではないか….

残念ながら，なるべく多くのページを検索可能とするという目標は，図-1に示したような基本的な構成を単純に実装したのではまったく達成できない．たとえば，高速にリソース収集が可能なロボットを実装するには，数々のプログラミングの工夫やテクニックが必要である．これをクリアしてもディスク容量やインデクサの限界が次には待ちかまえているかもしれない．どうやら仕組みは単純でも，本当に使い物になるものを作るのは大変そうである．たとえば，定番サーチエンジンの1つといわれているAltaVistaは，基本的には

☆　「検索の鉄人」の企画(http://tetsujin.arena.ne.jp/)自体を非難しているのではないことは十分ご理解いただきたい．

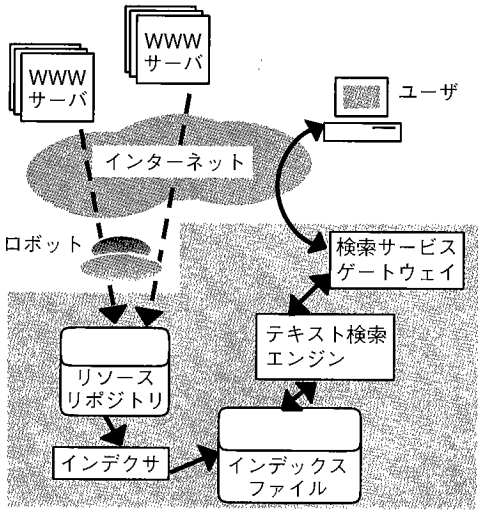

図-1　ロボット型サーチエンジンの基本構成

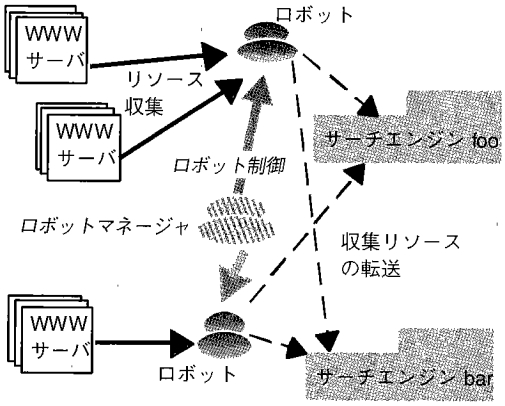

図-2　山名らによる分散型WWWロボットの構成

図-1のようなアーキテクチャを採用しながらも，強力なハードウェア群による並列／並行処理や高速ネットワークへの接続などを駆使することで，膨大なページの収集とインデックス化を可能としている[4]が，3億2千万ページをカバーしきれてはいない[2]．

●「他力本願型」メタサーチエンジン

自ら作るのが大変であれば，すでに運用されているいくつかのサーチエンジンを利用してしまうサーチエンジンを作ってはどうだろう．もっと上品にいえば，ユーザがさまざまなサーチエンジンを渡り歩くのを代行するものだ．このようなタイプをメタサーチエンジンと呼ぶ．代表例としては，MetaCrawler[5]がある．メタサーチエンジンは，利用する各サーチエンジンに適合するようにユーザからの情報要求を変換し，各サーチエンジンからの検索結果を重複を排除するなどして「適当に」マージする．このような処理をするプログラム（ラッパーなどと呼ばれる）を利用するサーチエンジンごとに作る必要がある．Lawrenceの考察[2]によれば，大手のサーチエンジンであっても収集しているリソースには片寄りが存在するため，メタサーチエンジンは有効なアーキテクチャといえる．ただし，上記のラッパーの問題のほかに，各サーチエンジンからの検索結果をランキングを含めて「適切に」マージすることが難しいといった他力本願であるがゆえの問題点がある．

●分散ロボットの協調動作による効率的なリソース収集

上記でサーチエンジンが収集しているリソースには片寄りがあると述べた．こうなると，いくつかのサイトのロボットが収集したリソース自体を共有しようというのは，自然な考え方であろう．ただし，重複して収集されているリソースも相当の量になるから，もしこれらのロボットが互いに協調することにより，重複したリソースの収集を極力避けることができれば，効率はさらに高まる．

山名らのグループでは，まさにこのような考え方に基づいた検討を進めている[6]．彼らは，図-2に示すような，ロボットとそれが収集したデータを用いるサーチエンジンとを分離したアーキテクチャを想定している．すなわち，各サーチエンジンは互いに独立な検索サービスを行うものとして考えられている．各ロボットが各WWWサーバからリソースを取得し，それを実際に検索サービスを提供するサーチエンジンへ配布するというモデルにおいて，ネットワーク負荷，分散収集時間を最小化するように各ロボットに対してそれが担当すべきWWWサーバの集合を割り当てるアルゴリズムを提案している．

これに対して能登らは，1つの検索サービスを最初からメタサーチエンジンとして実現するアプローチを提案しており，各サーチエンジンが有する専用のロボット間の効率的な協調方式を検討している[7]．各サーチエンジンは，専用のロボット群が収集したリソースのみを対象にインデックスを作成する．よって，山名らの場合のように各サーチエンジンへ収集データを配布する必要はない．また，あるサーチエンジンにおけるロボットのリソース収集ポリシーが明確に定義される場合（例：jpドメインのみ収集），それを単独で1つの検索サービスとして利用することができる．

●相互協調型メタサーチエンジン

メタサーチエンジンが利用する各サーチエンジンが共通の検索プロトコルに基づいて検索処理を受け付け，結果を返却するようにすれば，従来の単純なメタサーチエンジンの欠点を克服することができる．また，個々のサーチエンジンをビルディングブロックとして組み合わせて独自のメタサーチエンジンを構築する可能性も生まれる．このようなアプローチをここでは相互協調型メタサーチエンジン（図-3）と呼ぶことにする．STARTS[8]は，このようなアーキテクチャを可能とするための検索プロトコルの代表的な例である．このようなプロトコルが普及すれば，先に述べたような単純なメタサーチエンジンの欠点が解消される．

• 各サーチエンジンは，その特性をメタデータとして公開するので，メタサーチエンジンは，このメタデータを評価することにより，ユーザの質問に応じて適切なサーチエンジンを選択できる可能性が生まれる．

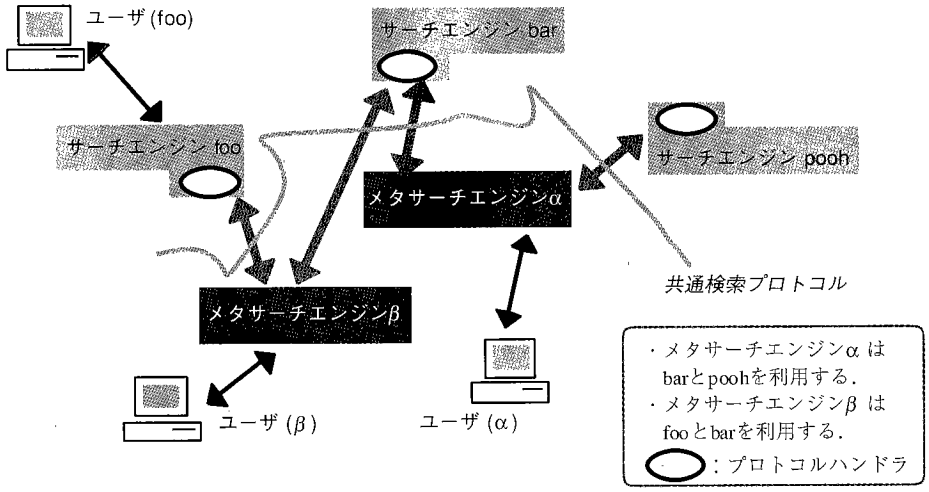

図-3　相互協調型メタサーチエンジンアーキテクチャ

• 各サーチエンジンで用いる検索結果のランキングに関係する情報（タームの頻度情報など）を公開することにより，メタサーチエンジン側で各サーチエンジンからの検索結果をランキングを含めて適切にマージできるようになる．
• 共通の検索プロトコルを遵守する限り，新しいサーチエンジンを容易に追加することができる．

インタラクティブで使いやすいサーチエンジン

凡人の不満2は，実はサーチエンジンに限った問題ではなく，テキスト情報検索システム全般にも当てはまる問題であるが，ここでは，サーチエンジンはWWWという情報空間において，ユーザが望む情報へ到達する過程（情報ナビゲーション）を支援すべき道具であるという観点にたって，凡人の不満2に関係する技術的トピックを紹介する．「検索の凡人」にとっては，1回の質問で望ましい情報を的確にゲットすることは難しい．現在のWWWのように雑多な情報が大量に地球規模で存在すればなおさらである．したがって，サーチエンジンは，ユーザの試行錯誤の過程をインタラクティブに支援すべきである☆2．つまり，(1) ユーザのアクション（情報要求）に対してどのような検索結果が得られたのかを何らかの意味で整理して提示すること，(2) 次のアクションを決定するために有益な情報を提供すること，が重要となる．

●適合フィードバック

インタラクティブな情報検索の古典的な手法として，適合フィードバック（relevance feedback）と呼ばれる手法がある[3]．この手法においては，初期の検索結果からユーザの情報要求に比較的適合したと思われるページをユーザ自身が選択する．システムは，選択されたページ中の単語分布の情報などをもとに新たな質問を生成し，2回目の検索を行う．現在インターネット上でサービスされているサーチエンジンでは，Exciteなどがこのような機能を提供している．Exciteにおいては，検索結果の各リソースについて，“More like this”と記されたリンクが用意される．これをクリックすることで容易にフィードバックを行うことができる．フィードバックというアクションをユーザに決断させるためには，それに足る情報をサーチエンジンが提供する必要がある．つまり，検索結果内のあるページが適合文書であるかをユーザがある程度の「あたり」をつけられるような情報をサーチエンジンが付与することが望まれる．ページの内容が的確に要約表示されればユーザは大いに助かるだろう．テキストの要約技術[10]の進展，特に，情報要求に対する動的なテキスト要約技術の進展が望まれる．

●検索結果の分類整理

質問として，“アンカー”というようなタームを想定してみよう．多数の検索結果には少なくとも，HTML文書に関する（ハイパーリンクのタグ）ようなページと，構造工学（建物の土台となる構造物）に関係するようなページが混在するに違いない．このような場合，検索結果であるページの集合を適切に分類整理してくれる機能があるとよいだろう．分類整理の方法論としては，あらかじめ定められたカテゴリ分類体系におけるいずれかのカテゴリへ各ページを割り振る「カテゴリ分類（categorization）」と，互いに類似した内容を持つページを動的にグループ化する「クラスタリング（clustering）」という2つがある．両者にはそれぞれ長所・短所があるが，Xerox PARCにおける研究[11]では，一貫してインタラクティブな情報検索の観点からクラスタリングの適用を提案している．しかしながら，検索実行の度に検索結果をクラスタリングするにはある程度の計算パワーを必要とするためか，大規模なサーチエンジンにおける適用例はない．筆者らの開発したTITANにおいては，検索結果のクラスタリングを実験的にサポートしている．**図-4**にTITANの検索結果の画面例を示す．

☆2 ユーザの行動履歴やブックマークなどからユーザモデルを構築し，これを情報フィルタリングなどに適応しようという研究が活発化している[9]が，サーチエンジンへの適用については具体例が少ないこと，紙面が限られていることから，本稿では取り上げない．

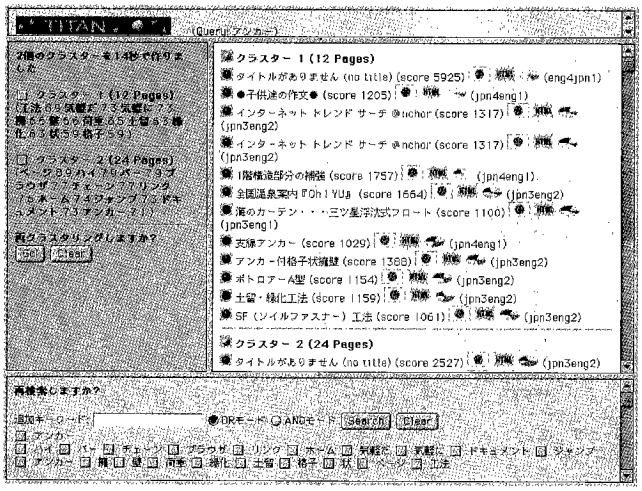

図-4　TITAN(http://titan.mcnet.ne.jp/)の検索結果画面例

●関連タームの提案

サーチエンジンが有している語彙に関する知識やインデックスを作成する際のタームの頻度分布や共起☆3情報などを利用することにより，初期の質問に関連する（したがって，追加しうる）語のリストを提示することがいくつかのサーチエンジンで行われている．たとえばAltaVistaにおいては，“Refine”というボタンを選択することによりこの機能を利用することができる．なお，TITANにおいては，検索結果をクラスタリングする際に抽出した特徴的な単語のリストを提示することにより，ユーザによる質問の変換を支援している．この機能を利用すれば，絞り込み的な検索や連想的な2次検索が可能となる．

●GUIによる検索条件の直接操作

鷲崎らは，GUIを用いた直接操作により検索条件を視覚的に調整するようなインタフェースを検討している[12]．このインタフェースでは，1つのタームに対応するページ集合がVenn図のように円として表示される．これらの円はマウスドラッグにより自由に重ね合わせることが可能であり，たとえば，（アンカー AND 構造工学）のような検索条件に適合するページ件数をVenn図上で即座に知ることができる．つまり，適当な件数の結果が得られるように検索条件を視覚的に調整することができる．

言語の壁を越えるサーチエンジン

今後のインターネットのさらなる普及を考えると，さまざまな言語で記述された情報がWWWで利用可能となると考えられる．自分の好む言語を用いて世界中のさまざまな情報を検索したい（これをクロスランゲージ情報検索と呼ぶ）というニーズは今後さらに増大し，凡人の不満3は万人の切望となるかもしれない．自分の好む言語（X言語）によって他の言語（Y言語）によって記述された情報を検索し，さらに，その検索結果が適切であるかを判定できるようにするためには，処理のいずれかの段階でX言語とY言語の間での変換／対応付けが必要となる．さらに検索結果のページを読みこなすには翻訳などの支援が必要となる．

●クロスランゲージ情報検索の方法論

クロスランゲージ検索の方法としては，3つ考えられる．(1) 検索対象を質問の言語に合わせてあらかじめ翻訳しておく（検索対象翻訳型），(2) 情報要求を受け付けるたびに検索対象の言語へ翻訳する（情報要求翻訳型），(3) 特定の言語とは独立したニュートラルな表現へ双方を翻訳する（中間言語型），である．中間言語型は，機械翻訳の歴史と現状から見ても困難そうに見える．検索対象翻訳型では，時に膨大な量の検索対象（しかもその中でどれだけが実際に検索されるかは不明）をあらかじめ翻訳し保存しておくコストが必要である．やはり，現状では情報要求翻訳型が現実的といえる．

●情報要求の翻訳

Oard[13] は，情報要求翻訳の方法論を大きく，辞書に基づく方法とコーパス☆4に基づく方法とに分類している．前者は，二言語間の対応辞書を用いる方法である．TITANなどのWWW上で公開されているクロスランゲージサーチエンジンは，この方法に基づいている[14]．この方法においては，質問中のタームに対して複数の訳語が存在する場合にそれをどう扱うか（たとえば，“buck”というタームを「ドル」とするのか，「鹿」とするのか，あるいは（ドル OR 鹿）として扱うのか），辞書にないターム（未知語）をどのように扱うか，複合語やフレーズなどの制約をいかに利用するか，などの課題がある．

一方，コーパスに基づく方法とは，あらかじめ文／段落／ファイルといった単位で対応がつけられた文書のペア（バイリンガルコーパス）を大量に用意し，これを統計的手法により処理することにより，タームの対訳関係を自動的に学習するというものである．この方法においては，大量のバイリンガルコーパスをどのように用意するか（作成するのには当然相応のコストがかかる），ある分野に関するコーパスで学習した対訳関係が他の分野にもうまく適用できるとは限らない（分野移行性の問題），という問題があり，WWW上のサーチエンジンに今すぐ適用するのは困難であろう．

現在，この分野では活発な研究が行われるようになってきている．WWW上の情報検索というような適用先においては，辞書に基づく方法をベースとせざるをえないが，適切な訳語をいかに選択するかについては，たとえば翻訳先言語におけるモノリンガルコーパスにおけるタームの共起統計☆5を利用するというような方向性[15] が有望であると考えられる．

☆3 ある範囲（文書や段落など）でターム x, y が共に出現していること．

☆4 (計算機処理可能な)言語データのこと．

☆5 たとえば，(“bank” AND “transfer”)という英語の情報要求の翻訳として，(“堤防” AND “転送”)よりも(“銀行” AND “振替”)の方が適切であろうことを日本語コーパスにおけるターム間の共起情報を用いて判定する．

●検索結果の閲覧支援

検索結果を分かりやすく提示することの重要性についてはすでに述べたが，検索結果が他国語で表示される場合はなおさらである．TITANでは，ページタイトルは名詞句であるという仮定のもとで，単純化された方法によりタイトルの日本語への翻訳を提供している[14]．しかし，これが実際のユーザの意志決定にどれほど有効であるかは未評価である．一方，CLINKSというサーチエンジンを開発した鈴木らは，英語ページからキーワード抽出を行い，これを日本語キーワードに翻訳して提示することが有効であることを実験によって示している[16]．究極的には，たとえばテキスト要約と機械翻訳を適切に組み合わせることにより，クロスランゲージなテキスト要約を目指す必要があろう．

●多言語化へ向けて

ここまではある言語ペアを前提として議論してきたが，機械翻訳の場合と同様に，当然のごとくさらに多くの言語への拡張が期待されてくる．すでにヨーロッパでは，英語に対する語彙知識データベースであるWordNet[17]をヨーロッパの諸言語へ拡張しようというプロジェクト（EuroWordNet）が活動している[18]．EuroWordNetでは，トップレベルオントロジーと呼ばれる言語共通な高位レベルの概念体系のもとに，ILI（Inter-Lingual-Index）と呼ばれるノードを介して各国語の概念ノードがリンクされることにより，全体として多言語語彙知識データベースが構築される．また，クロスランゲージ情報検索への適用を意識して，ドメインラベルの付与などの拡張も試みられている[19]．英語を対象としたWordNetをベースとすることがどの程度可能かはまだ明らかではないが，さらにほかの言語へ拡張していくような方向性も当然考えられる．

なお，筆者らのグループでは，相互協調型メタサーチエンジンアーキテクチャによってさまざまな言語を扱うサーチエンジンを統合することによる多言語情報検索の枠組みを検討している．特に，情報要求や検索結果の翻訳を取り扱うことのできる検索プロトコルについて提案を行っており[20]，複数の海外のサイトと協力して多言語情報検索の実験サービスをWWW上で行う予定である．

サーチエンジンのこれから??

現在，実験的色彩の強いものも含め多くのサーチエンジンが公開されているが，数種類のサーチエンジン（膨大なインデックスを持ち高速に動作するもの）が定番として利用されるようになってきている．また，この状況はしばらくは変わらないだろう．しかしながら，このような「巨大百貨店」のようなサーチエンジンの限界を感じておられる「検索の凡人」の方も多いのではないだろうか．今後は，何らかのポリシーで特化された「専門店」のようなサーチエンジンが多く構築され，これらを知的に使い分ける「検索の達人」たる知的な情報エージェントの出現が望まれる．そのためには，各々の専門的エンジンの特性や能力をどのように表現し流通させるか（メタデータの問題），知的情報エージェントがいかにうまくユーザに合わせて（ユーザ適応，意図理解の問題）これらの専門店サーチエンジンを使いこなすか（プランニングの問題）などを解決していく必要がある．また，これらの大前提として，XMLのようなコンテンツ記述の新しい枠組み[21]のもとで，的確な情報抽出を可能とするテキスト解析技術の高度化や，オントロジーの構築[22]も必要となってくるだろう．

参考文献

1) 原田昌紀: サーチエンジン徹底活用術, オーム社（1997）.
2) Lawrence, S. and Giles, C.L.: Searching the World Wide Web, Science, Vol.280, 3, pp.98-100（Apr. 1998）.
3) Witten, I.H., Mofat, A. and Bell, T.C.: Managing Gigabytes, Van Nostrand Reinold, New York（1994）.
4) DIGITAL Equipment Corporation: About AltaVista Search（1997）. http://altavista.digital.com/av/content/about_our_technology.htm
5) Selberg, E. and Etzioni, O.: Multi-Service Search and Comparison Using the MetaCrawler, In Proc. 4th World Wide Web Conference, pp.195-208（1995）.
6) 山名, 田村, 河野, 亀井, 原田, 西村, 浅井, 楠本, 篠田, 村岡: 分散型WWWロボットによるWWW情報収集, 第9回データ工学ワークショップ（DEWS'98）論文集, DEWS98-（28）（1998）.
7) 能登, 竹野, 小橋: インターネット検索サービスのための分散型情報収集, 情報処理学会マルチメディア，分散，協調とモバイルシンポジウム（DICOMO'98）論文集, pp.391-398（1998）.
8) Gravano, L., Chang, K., Garcia-Molina, H., Lagoze, C. and Paecke, A.: STARTS: Stanford Protocol Proposal for Internet Retrieval and Search（1997）. http://www-db.stanford.edu/~gravano/starts.html
9) Resnik, P. and Varian, H.R.（Eds.）: Recommender Systems, Comm. ACM, Vol.40, No.3, pp.56-89（1997）.
10) 奥村, 難波: テキスト自動要約技術の現状と課題, 北陸先端科学技術大学院大学情報科学研究科, Research Report IS-RR-98-0010I（1998）.
11) Hearst, M.: Interfaces for Searching the Web, Scientific American, pp.68-72（Mar. 1997）.
12) 鷲崎, 林, 菊井: WWW上の情報探索システムにおけるインタラクティブインタフェース, インタラクティブシステムとソフトウェアIV 日本ソフトウェア科学会WISS'96, pp.1-10, 近代科学社（1996）.
13) Oard, D.: Alternative Approaches for Cross-Language Text Retrieval, In Cross-Language Text and Speech Retrieval, Papers from the AAAI-97 Spring Symposium, Technical Report SS-97-05, pp.154-162, The AAAI Press（1997）.
14) Hayashi, Y., KiKui, G. and Susaki, S.: TITAN: A Cross-Linguistic Search Engine for the WWW, ibid, pp.56-62（1997）.
15) Kikui, G.: Term-List Translation using Mono-Lingual Word Co-occurrence Vectors, In Proc. of COLING-98 and the 36th ACL（to appear）（1998）.
16) 鈴木, 井ノ上, 橋本: クロスルンガル情報検索結果の閲覧支援のための主要キーワード対訳表示の効果, 情報処理学会自然言語処理研究会, NL126-14（1998）.
17) Miller, G.A.: A Lexical Database for English, Comm. ACM, Vol.38, No.11, pp.39-41（1995）.
18) Vossen, P., Diea-Orzas, P. and Peters, W.: The Multilingual Design of the EuroWordNet Database, In Proc.of IJCAI-97 Workshop on Ontologies and Multilingual NLP, pp.41-47（1997）.
19) Vossen, P.: EuroWordNet: A Multilingual Database for Information Retrieval, Paper Presented at DELOS Workshop on Cross-Language Information Retrieval（1997）. http://www.let.uva.nl/~ewn/docs/P011.ps
20) 巌寺, 林, 菊井, 小橋, Leong, M., Choi, K.: 多言語分散情報検索アーキテクチャに関する検討, 情報処理学会自然言語処理研究会, NL127-9（1998）.
21) 浦本, 武田: インターネット上での情報の記述と交換方式の最近の動向, 人工知能学会誌, Vol.13, No.4, pp.519-527（1998）.
22) 溝口理一郎ほか: 大規模知識ベースに関する調査研究ーオントロジー工学に関する調査研究－報告書, 日本情報処理開発協会（1998）.

（平成10年7月30日受付）