デジタルサーチャー講座

検索を科学する

塩田 紳二

第10回 ジャストシステム ConceptSearch /R.2

近年は Google、マイクロソフト、Yahoo!など、各社が各々の検索技術を活かしたデスクトップ検索ツールを提供し、こういったツールも当たり前のように知られるようになった。だが、それ以前に日本でジャストシステムがデスクトップ検索ソフト「ConceptBase Search」を提供していた。今回はこの ConceptBase 技術を使ったデスクトップ検索ソフトの最新版（2月10日発売）である「ConceptSearch /R.2」について紹介する。

デスクトップ検索の老舗 ジャストシステムの新製品

ジャストシステムの「ConceptSearch（コンセプトサーチ）/R.2」（図1）は、Google Desktop Search やマイクロソフトの Windows Desktop Search と同じく、ローカルファイルの検索を行うツールで、2005年3月に発売された ConceptSearch の後継にあたる。

マイクロソフトや Google のものは、ウェブ検索と組み合わされているが、ジャストシステムのものは、単体アプリケーションでインターネット検索の機能は持たない。マイクロソフトや Google は、インターネットの検索エンジンを持ち、その延長上にデスクトップ検索を位置づけているが、ConceptSearch /R.2 は、ローカル検索のアプリケーションという位置づけである。インターネット検索はないものの、その分、アプリケーションとして、使い勝手を向上させるような細かい機能を搭載している。

ジャストシステムは、以前からデスクトップ検索のアプリケーションを提供してきた。1997年には、個人向けのデスクトップ検索ソフト「ConceptBase Search」を発売している。当時は、デスクトップ検索という言葉さえなく、「文章検索・要約システム」と表記されていた。同社では、こうした機能を ConceptBase 技術と呼んでいる。これは、「文書概念による類似情報抽出」技術で、簡単にいえば、文書や文章と、大量の文書の間で類似性を計算するものだ。以前紹介した GETA（連想計算エンジン）と同じような目的の技術だ。この技術は、ジャストシステムが1996年に買収した米国 Claritech 社の技術をベースにし、日本語処理技術などを組み込んだものだ。

その後、インターネットのウェブページ収集機能などを付加した「インターネットブーメラン」などを製品化。また、企業向けのデータマイニング（大量のデータの中から関連性や特徴などの情報を抽出する技術）やナレッジマネジメント分野にもこの ConceptBase 技術を応用した製品を投入している。ある意味で、同社はデ

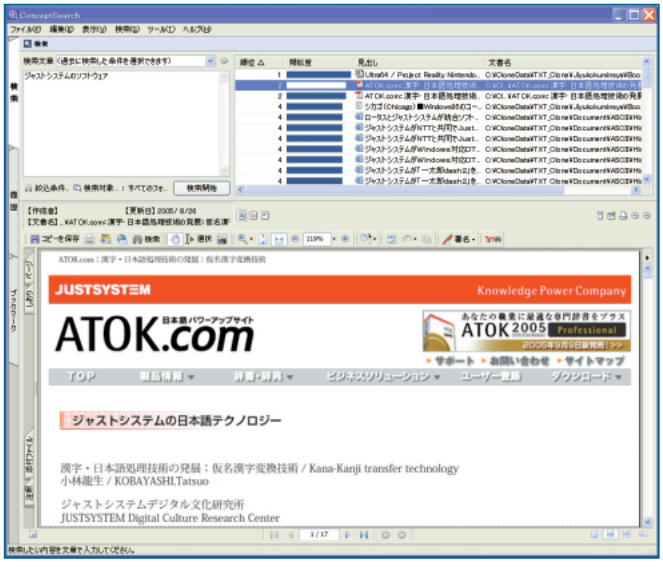

図1 ConceptSearch /R.2 は、ConceptBase 技術を使ったデスクトップ検索ソフトウェア。

インターネットマガジン／株式会社インプレスR&D
©1994-2007 Impress R&D

スクトップ検索の老舗ともいえるだろう。

自然文検索や連想検索など他社製品にはない機能を持つ

ConceptSearch /R.2には、Google Desktop SearchやWindows Desktop Searchにはない機能がある。1つは、自然文による検索機能である。ConceptBaseは、検索対象となる自然文から、手がかりとなる検索語を抽出して検索用のデータ(インデックス)を作成する。この機能を利用して、検索条件として自然文を受け付け、そこからキーワードを抽出する。また、このときすでに出来上がっているインデックスから、関連を持つ単語を同時に検索対象とする。このようにすることで、そのものズバリの検索語が対象文書になくとも、関連していそうな文書を探すことができる。

ただし、完全に自然文を理解しているわけではないので、抽象的な表現を使うと検索がうまくいかないことがある。たとえば、「マイクロソフトは何をしていますか?」などのように、キーワードが少なく、何を聞いているのかあいまいな文章では検索はうまくいかない。しかし、「マイクロソフトの戦略は?」のような文章では、「マイクロソフト」や「戦略」といった単語から関連語が連想されるため、それなりの検索結果が得られる(図2)。

さらに、検索にあたって連想キーワードを使うことができる(図3)。これは、検索条件から抽出されたキーワードを使って文章との類似度を計算したのち、上位の文書が持つ特徴単語を追加して検索を行うものである。簡単にいうと、検索結果の文書に関連した文書も検索結果にすることができる。また、この関連語を別途、追加することもできる。ただし、任意の単語ではなく、検索条件に対して、関連する単語の中からの選択だけが可能だ。設定で、利用する関連語の関連度のしきい値を設定できるが、それ以下の関連語を選択項目として提示しているのであろう。選択項目とともに関連度が表示されるようになっている。

さらに、この検索条件に加えて、「絞り込み条件」と呼ばれるものを設定できる(図4)。自然文や連想キーワードによる検索では、思いがけない結果が得られるというメリットもあるが、逆にユーザーから見て無関係なものも検索結果に出てくることがある。そこでキーワードで特定される文書だけを対象としたいような場合、たとえばニュースリリースだけを検索対象にしたいので、「発表文」や「ニュースリリース」といった単語を確実に含む文章だけを対象とするような場合に利用できる。これは、内容的な対象文書の指定だが、PDFやDOC(Word文書)などの文書フォーマットそのものを指定しての検索はできない。検索結果リストに文書フォーマットに対応したアイコンが表示されるので区別は可能だ。あまり結果数が多くなければなんとかなる。

もう1つは、文章の要約機能である。検索された各文章に対して、パーセンテージを指定して、文章の要約を作らせることができる。このようにすることで、

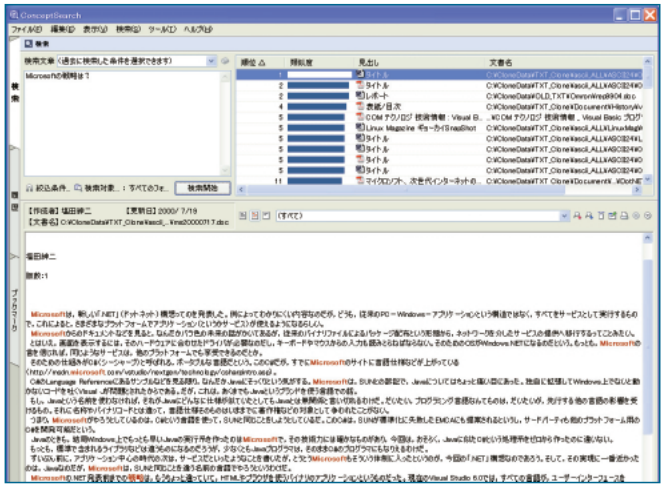

図2　キーワードになりそうな単語が2語以上あれば、まともな検索結果になる。

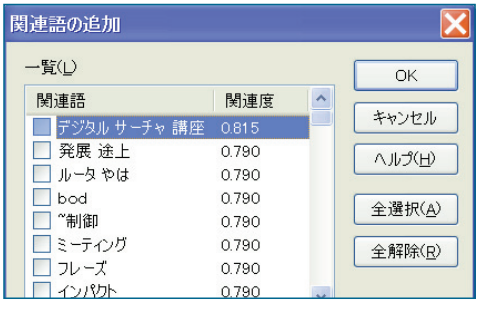

図3　関連語は、別途付加したり、検索時に必ず利用するように設定したりもできる。

図4　絞り込みは、通常の検索条件に加えて、検索結果を制限するために利用できる。

インターネットマガジン／株式会社インプレスR&D
©1994-2007 Impress R&D

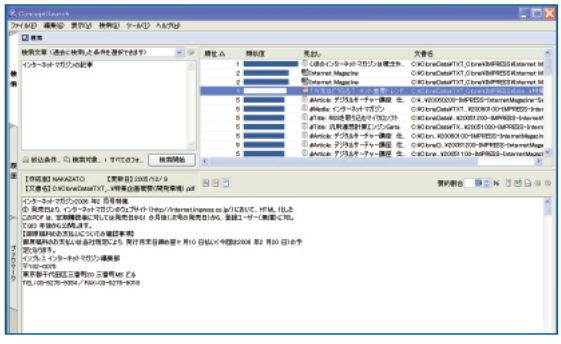

図5　ハイライト表示にすると、検索条件に含まれる単語や関連単語部分が色つきで表示される（上）。要約表示では、抽出したテキストを要約処理した結果が表示され、長文でも短時間に把握ができる（右）。これは、PDFやWordの非テキスト系の文書ファイルでも可能である。

長い文書であっても、ざっと目を通すことが可能だ。

検索結果である文書は、同じウィンドウ内に表示される。オリジナル表示だけでなく、キーワード（連想キーワードを含む）をハイライト表示したものや、前述の要約表示が可能だ（図5）。

また、検索の履歴を保存したり、検索した文書に対してブックマークを付けて記録したりするといった細かい機能もある。検索履歴は、Googleがパーソナライズ検索で対応しているが、ConceptSearch /R.2の検索履歴は、検索条件を保存したもので、選択するたびに再度検索を行っている。ただし、検索対象のフォルダ（ConceptSearch /R.2では、インデックス作成をWindowsのフォルダ単位で行っている）や絞り込み条件などは記録されている。

デスクトップ検索では、さまざまなフォーマットを解釈できなければならない。Google Desktop SearchはプラグインでWindows Desktop SearchはiFilterでこれを処理する。ConceptSearch /R.2もやはり専用のファイルフォーマットドライバー（Content Filter）を持っており、これを使ってインデックス化やハイライト表示などを行う。また、Content Filterは、後からでもインストールが可能なようである。

詳しいドキュメントはないが、ファイルフォーマットを解釈し、テキストを抽出するものだと思われる。ただし、一部、セキュリティー制限がかけられたPDFファイルなどではテキスト抽出ができないようだ。マイクロソフトやGoogleの場合には可能だったので、対応フォーマットの充実と合わせ、このあたりの機能充実は今後の課題だろう。

基本をおさえた使い勝手
さらに明確な差別化が課題

ConceptSearch /R.2のウィンドウは、左側に3つのタブを持つ。タブは、「検索」「履歴」「ブックマーク」になっていて、それぞれの機能を切り替える。

検索タブでは、3つの領域が表示される。それぞれ検索文章、検索結果リストと文書のプレビュー用である。領域にはツールボタンが配置されているため、基本的な使い方であれば、メニュー操作をすることなくクリックだけで利用できる。

また、検索結果リストには、文書タイトルやファイルパスなどのほかに、作成日や更新日を表示することもできる。

文書表示領域は、3つの形式で文書を表示できる。1つは、オリジナルの形式、もう1つは関連度推定の根拠となった単語をハイライト表示したもの（図5）、そして要約表示である。オリジナル形式の表示は、おそらくActiveXによるインプレース表示、つまり、IEのプラグインなどの表示と同じで、表示領域の中で、該当するアプリケーションを動かすものだ。

あとの2つは独自の表示で、抽出したテキストを使ったものだと思われる。

履歴タブでは、検索タブの検索条件領域が縦に延び、検索履歴を表示する場所となる。右側は、上下に検索結果リストと文書プレビューが並ぶ（図6）。

検索では、似たような検索を再度行うことが多い。筆者もよくGoogleの検索履歴を使うし、それを見ると同じ検索を繰り返していることがわかる。検索履歴は、ある意味で、特定の目的に合わせた文書リストでもある。しかも、固定したリストではなく、対象が変化したり、文書が自動的に追加されたりするリストである。その意味で検索に履歴管理は必須と思われる。

ブックマーク（図7）は、検索した文書を階層構造（フォルダ構造）で管理するものだ。使い方としては、ウェブブラウザーのブックマークと同じである。参照頻度の高い文書を登録しておくと、位置にかかわらず、特定内容の文書をまとめておくなどが可能だ。ただし、登録できる文書は、

インターネットマガジン／株式会社インプレスR&D
©1994-2007 Impress R&D

検索結果リストに表示されたもので、1つ1つ登録する必要がある。できれば、検索履歴から検索条件そのものをフォルダなどとして登録できると便利である。というのも、検索履歴は検索した順番の表示となるため、何回も繰り返した検索でも、探すのがちょっと面倒なのである。

ConceptSearch /R.2では、検索対象のインデックスを作成したら、それを検索時に対象として指定する。別の仕事であるなど対象を限定するような場合、それぞれが別のフォルダになっていれば、インデックスを分離できる。また、検索時には、フォルダに対して個別に対象と非対象を設定できるため、必ずしも1つのフォルダになっている必要もない。インデックス更新の効率からいえば、1つのフォルダ以下全体をインデックス化するよりも、更新頻度や用途などを見て、個別のフォルダを指定したほうがいいだろう。また、このインデックス対象フォルダ個別に、インデックス更新のスケジュールが指定できる。これを使うことで、たとえば、更新がほとんどない過去の文書などのフォルダでは自動更新をストップさせておくといった使い方が可能だ。ただ、このスケジュールの最小のくり返し単位は1日で、それ以下では指定できない(図8)。数ギガバイトの文書データに対して、更新処理に数時間ほどかかることもあり、あまり短い時間単位でのインデックス更新はできないようだ。

マイクロソフトやGoogleの検索ソフトがプラグインやiFilterの開発キット(SDK)などを配布してAPIなどを公開しているのに比べると、単体アプリケーションである ConceptSearch /R.2は、少しクローズドな感じがある。もっとも、インターネットだけとはいえ、開発者のサポートは簡単ではなく、Googleやマイクロソフトだからできることともいえる。

原理的には、ConceptSearch /R.2は単なる検索だけでなく、文書のグループ化も行える能力がある。文書を指定し、抽出した自然文を検索条件として指定すれば、関連文書をピックアップできるのだ。だから、検索ソフトの伝統的なインターフェイスにこだわらず、文書の自動分類といった部分を前面に出したほうが、他との差別化で有利な気もする。実際、検索後に「結果の読みこみ」を検索結果リスト上でコンテキストメニューから選べば、文書内容を検索条件とすることができる。たとえば、分類した文書相互の関係などをグラフィカルに表示などできると便利そうなのだが。

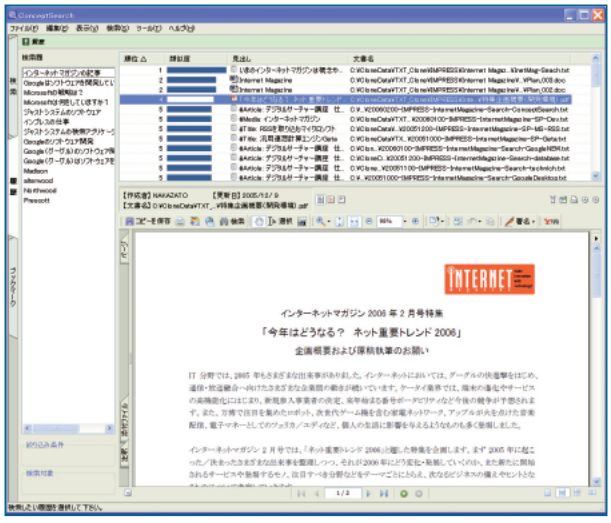

図6　履歴では、項目を選択するたびに再度検索が行われる。文書の追加などを考慮して検索結果そのものを保存しているわけではないようだ。

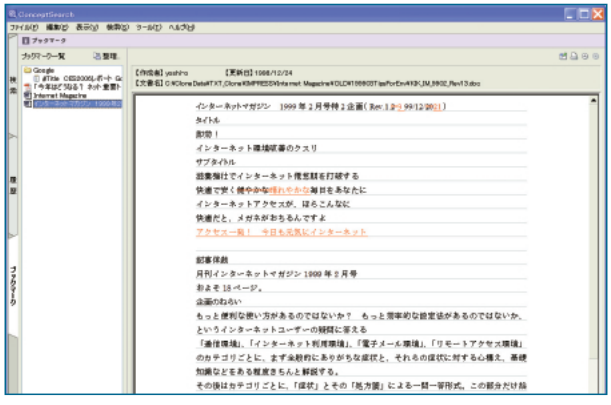

図7　ブックマークは、検索結果リストから選んだ文書をフォルダ構造で管理するもの。文書の格納位置に関係なくグループ化などが行われる。

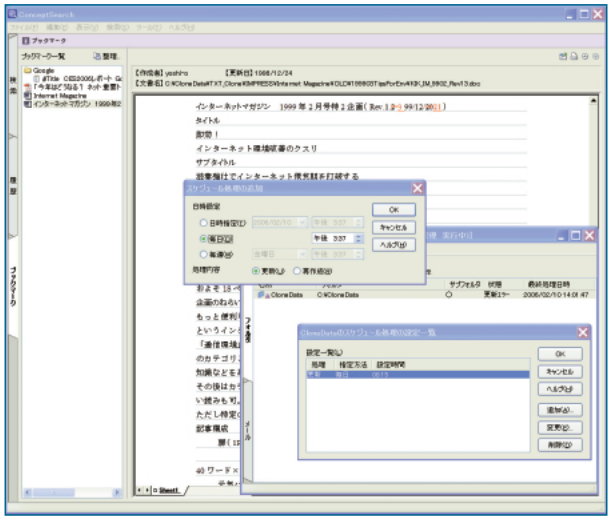

図8　インデックスの再作成や更新スケジュールは、1日、1週間のくり返しと、日時の直接指定で行える。1日以下の設定は、直接時間を入力する必要がある。

©1994-2007 Impress R&D

［インターネットマガジン　バックナンバーアーカイブ］ ご利用上の注意

このPDFファイルは、株式会社インプレスR&D（株式会社インプレスから分割）が1994年～2006年まで発行した月刊誌『インターネットマガジン』の誌面をPDF化し、「インターネットマガジン　バックナンバーアーカイブ」として以下のウェブサイト「All-in-One INTERNET magazine 2.0」で公開しているものです。

http://i.impressRD.jp/bn

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、URL、団体・企業名、商品名、価格、プレゼント募集、アンケートなど）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真の撮影者、イラストの作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は収録されていない場合があります。
- このファイルやその内容を改変したり、商用を目的として再利用することはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用する際は、出典として媒体名および月号、該当ページ番号、発行元（株式会社インプレス R&D）、コピーライトなどの情報をご明記ください。
- オリジナルの雑誌の発行時点では、株式会社インプレス R&D（当時は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

このファイルに関するお問い合わせ先
株式会社インプレスR&D
All-in-One INTERNET magazine 編集部
im-info@impress.co.jp

Copyright © 1994-2007 Impress R&D, an Impress Group company. All rights reserved.