デジタルサーチャー講座

検索を科学する

塩田 紳二

第1回 Google Desktop Searchの活用

インターネットのサイト検索やワープロの文字検索、置換、そしてデータベースなど、コンピュータを使うときに検索は欠かせない処理といっていいだろう。この連載では、そんな検索処理について、ソフトウェアなどの具体例を交えながら原理や理論を含め総合的に解説していくことにする。

どこにでもある検索

コンピュータを使うとき、「検索」はどこにでもある。もちろん、我々ユーザーが、意識的にソフトウェアの検索機能やインターネットの検索エンジンを使うこともあれば、プログラム内の処理として検索が行われることもある。

たとえばかな漢字変換も、入力した文字から該当する漢字を「検索」しているし、インターネットエクスプローラにURLをキーボードから打ち込めば、キャッシュに該当するページがあるかどうかが「検索」される。

「検索」は、コンピュータの基本処理の1つといってもいいぐらい、頻繁に使われるが、その中身はさまざまだ。インターネットの検索サイトでは、同じキーワードでもサイトごとに検索結果が違ってくることがある。それは、検索のアルゴリズムなどの処理が異なるからだ。

実際、検索機能を使う場合、その内部アルゴリズムによって、たとえば「いいキーワード」と「悪いキーワード」ができてしまう。理想的には、悪いキーワードなどあってはならないのだが、実際にプログラムを作る場合、コストやプログラムサイズなどの問題から、少々手を抜いたプログラム(アルゴリズム)が作られることがある。アプリケーションのちょっとした検索機能などの場合には、それほどコストがかけられないことも多い。

そういうわけで、ここでは、効率的な検索を行うための方法などを解説するとともに、その内部へと多少踏み込んで解説を行うことにする。内部を多少知ることで、その検索システムにとって、いいキーワードとはどういうものなのかを知ることができるからだ。

Google Desktop Search

さて、第1回の今回は、現在日本語版のベータテスト中の「Google Desktop Search」(http://desktop.google.co.jp/から入手できる。以下GDSと略す)を取り上げることにする。これは、パソコンのハードディスクにあるファイルやウェブブラウザーの履歴、電子メールといったローカルファイルの検索を行うプログラムだ(図1)。

検索は、インターネット上のGoogle検索と同じく、キーワードを入れて行う。

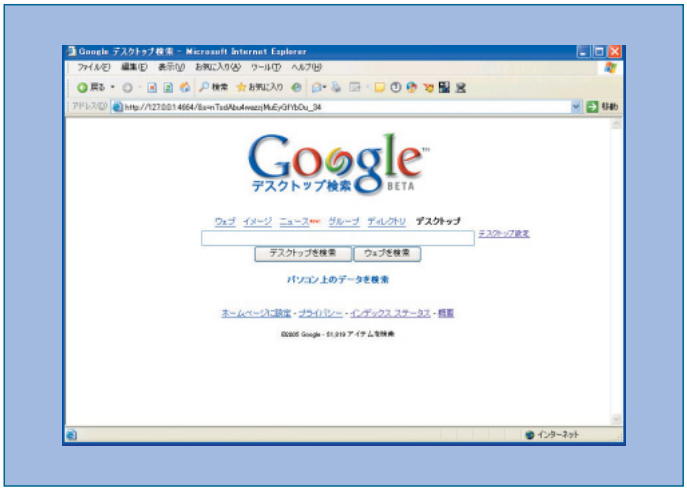

図1 Google Desktop Searchの検索画面。インターネットのGoogleホームページと同じく検索語を入力するだけのシンプルなもの。このほか、インデックス状態や設定ページなどがある。

特に条件などは指定せず、単にキーワードを入れれば、その単語を含むファイルを検索するようになっている。また、GDSを組み込むと、インターネットのGoogle検索時に同時にローカルファイルに対しても検索を行ってくれる。このほか電子メールやPDF、Officeの文書ファイルなども見つけることができる。

なお、以下の文章で、実際にGDSを使った検索結果を示すが、あくまでも筆者の環境で行った検索結果であり、他の環境では違う結果が出ることもあり得ることをご了承頂きたい。

インデックスの作成

GDSの検索はかなり高速で、大量のファイルがあっても瞬時に結果が表示される。これは、予めインデックスと呼ばれるデータが作られているからだ。

GDSのようにデータ内に含まれる任意の文字列について検索を行うことを「全文検索」という。全文検索では、対象となるデータが膨大な量になることが多く、なんの前準備もしないで検索を行うと、データ量に比例した時間が必要となり、通常はかなり長時間の処理が必要となる。メールソフトの検索機能などでは、全文検索が可能なものの、ほとんどのソフトでは前準備をしていないため、検索時に結果が得られるまでに多くの時間が必要なことが多い。

では、GDSはどうしているのかというと、パソコンが利用されていない時間に、検索対象となるファイルを読み込んで、インデックスと呼ばれる単語（正確には文法上の単語ではないので以後インデックス情報内に登録された語を対象語と呼ぶ）とそれを含むファイルを関連付けるデータを作っているのである（図2）。

検索は、このインデックスに対して検索を行う。インデックスは、検索しやすいように、対象語が整列されているため、検索を瞬時に行うことができる。結果として、対象語を含むファイルがどれであるかがわかるので、これを処理して結果として表示するわけだ。

このため、インストール直後には、対象となるすべてのファイルを読み込んで対象語を切り出すという処理を行わねばならない。初回の作業は、パソコンの性能や対象となるファイルにもよるが数時間程度は必要である。しかし、こうした処理を予め行っておくことで、高速な検索が可能になるのである。

ただし、この方法には、インデックス用のデータ領域が必要になるという欠点がある。GDSでは4Gバイト程度の空き領域が必要とされている。最近では、外部記憶装置のコストは低く、パソコンでは数十Gバイト以上の容量は当たり前になっているため、それほど大きな問題にはならないだろう。

もう1つ、事前にインデックスを作るために、プログラムを常に動かす必要がある。GDSをインストールすると、タスクトレー（通知領域）にアイコンが現れる。これは制御のためのアイコンだが、図3の

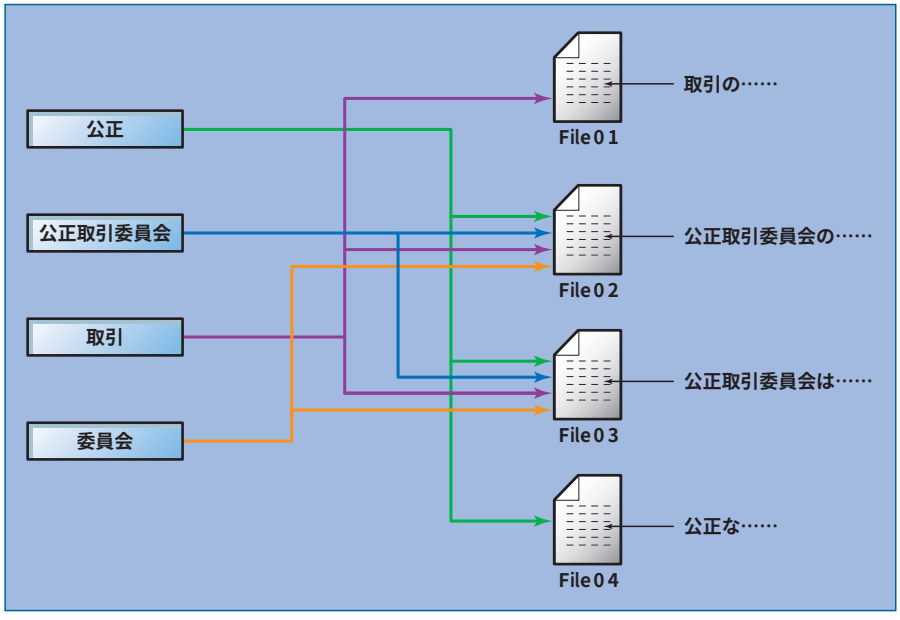

図2　インデックスとは、対象語とそれを含むデータ（ファイル）のリストを組みにしたもの。場合によっては、データ中の対象語の位置も記録される。

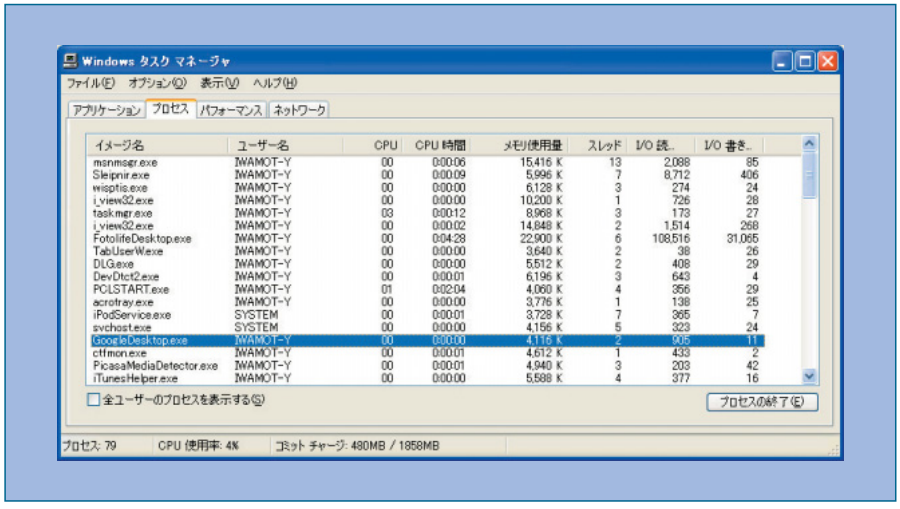

図3　Google Desktop Searchは複数のプログラムからなり、一部がメモリー内に常駐し、インデックスなどの更新を続けるようになっている。

ように同時にインデックスを作るためのプログラムがバックグラウンドで動き続けている。詳細は公開されていないが、ファイルやネットワークアクセスを監視して即座にインデックスに反映させるような構成になっているようだ。

インデックスを作成することは、簡単にいうと、予め対象語での検索を行ってその結果を保存しているのと同じである。このため、実際にキーワードが入力されたときに行うのは、インデックス情報の中からキーワードを探すだけである(図4)。通常、インデックス情報では、キーワードがアルファベットや文字コード順に並べてあるため、検索処理は高速に行える。また、インデックス情報自体が巨大になるような場合には、検索がしやすいようなファイル構造を作ることが多い。

対象語を切り出す

英語では、単語はスペースで区切られているのでデータから単語を取り出すのは簡単な作業だが、日本語の場合、文章から単語を取り出すのは簡単ではない。このような作業を「切り出し」と呼ぶ。

本格的に日本語の文章を文節に分解するには、日本語の単語と品詞、活用形などを考慮した辞書を使い、文法的な処理を行う必要がある。こうした処理を行うのはかなり大変だし、プログラム自体も大きくなってしまう。GDSのプログラムファイルを見るとデータファイルらしきものは512Kバイト程度で、大量の単語が記録されている辞書のようには見えない。

ただし、辞書なしでもある程度の処理は可能。文字コードや、記号、特定の文字といった辞書を調べなくとも判定可能な情報を使った分解方法がある。

「インテルのCPU解説」という語句の一部があるとする。ここには、ひらがなやカタカナ、漢字、アルファベットが含まれていて、これらは文字コードから簡単に判別が可能である。日本語の場合、カタカナやアルファベット、連続した数字などは、かならず独立した単語となるため、まずは「インテル」、「の」、「CPU」、「解説」という4つの対象語に分解ができる。

このような形で対象となるファイルから取り出した日本語文を分解していく。さらに漢字が連続するような単語、たとえば「公正取引委員会」といった単語は、通常は辞書などを使わないと分割ができない。しかし、もし他に「公正な判断」、「取引の結果」、「委員会の結論」といった文字列が文章に含まれていたとしたら「公正」や「取引」、「委員会」という対象語がインデックスに含まれるようになる。そうなると、文章中の「公正取引委員会」は「公正」などの3つの対象語から構成されていると判断できる。ある程度大量のファイルを全文検索の対象とするなら、辞書を使わなくとも、細かい対象語に分割することは不可能ではない。

GDSで検索を行わせてみると「公正取引委員会」の部分文字列である「正取」では検索ができなかった。この結果から、元の単語(この場合は公正取引委員会)から部分文字列を勝手につくりだしてインデックス化するようなことは行われていないと判断できる。また、前方一致となる「公正取」は、「公正取[改行]引委員会」のように間に改行が入る場合のみ検索結果として表示された。

つまり「公正取引委員会」という対象語があったとき、他の場所で単独では現れない「公正取」、「正取」のような部分文字列からなる対象語はインデックス化されていないと推測できる。

これらから考えると分割したときにそれぞれの対象語がインデックスに登録されないような部分文字列(前述の「公正取」のようなもの)は、GDSに対しては「悪いキーワード」であると考えられる。

なお、GDSの詳細は公開されておらず、使用頻度の高い単語のリストや日本語文字列の出現確率のリストを使っての分割という方法が使われている可能性もある。

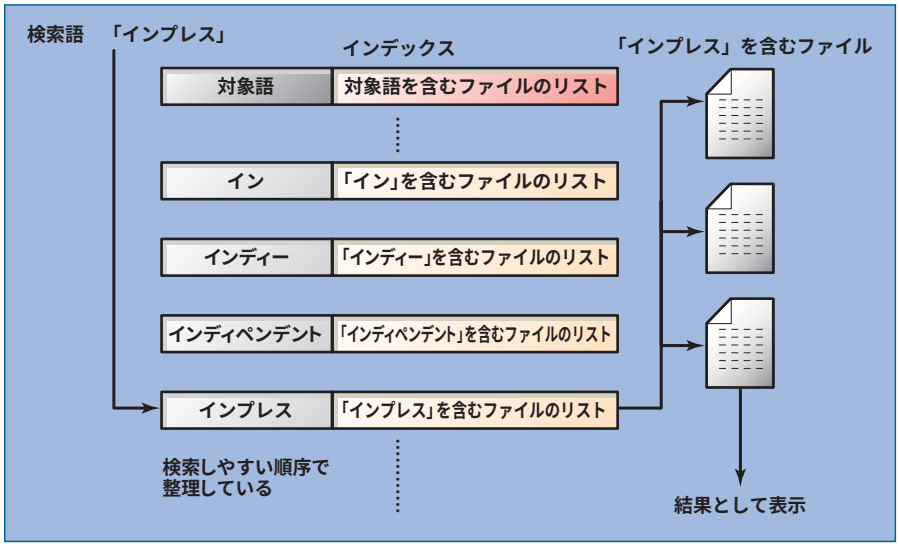

図4　インデックスを使った検索は、検索語をインデックス内の対象語から見つけることである。検索語が見つかれば、それを含むデータ(ファイル)はインデックスが持つ情報から知ることができる。

検索時の処理

検索時にもある程度の対象語による分割処理を行うこともある。もちろん、正直に入力されたキーワードと完全一致するものだけを検索するというのもありえるのだが、これだとあいまいな記憶からの検索では、まったく結果が得られないことがある。なお、あるデータが検索条件

を満たすことを「ヒット」するという。

GDSでは入力されたキーワードをある程度分割して検索を行うようだ。たとえば、「インターネットマガジンの記事」と入力すると、「インターネットマガジン」、「の」、「記事」といった3つの語を同時に含むファイルがヒットする(図5)。「インターネットマガジン 記事」と2つの単語に分割するとすこし検索結果が変わってくる。これでも「インターネットマガジンの記事」という語を検索することができる。違いは、対象語として「インターネットマガジン」、「記事」は含むが「の」を含まないという点である。このように単独のひらがな1文字も対象語となる可能性があるため、場合によっては検索語には助詞のようなひらがな1、2文字は入れないほうがいい場合もあるだろう。この例でいえば、「インターネットマガジンに記事が……その内容は……」という文章があれば、ヒットすることになる。これを禁止するには検索語をダブルクオート(")でくくる「フレーズ検索」を使う。このフレーズ検索は、他の単語、あるいは別のフレーズと組み合わせて使うこともできる。複合語そのものを検索したい場合も、ダブルクオートでくくってフレーズ検索を使う必要がある。

「公正取引委員会」という検索語では対象ファイルに「公正」、「取引」、「委員会」の3つの対象語が含まれれば結果として表示されるようだ。この場合、前述のように辞書を使わないかぎり分割はできない。しかし、「公正取引委員会」という対象語について「公正」、「取引」、「委員会」といった他の対象語への関連を示す情報がインデックス内にあれば、この検索語から「公正」や「取引」といった単語を含むファイルを検索することが可能になる。あるいは、「公正取引委員会」という対象語は、前記3つの対象語への参照としてインデックスファイル内で定義されているのかもしれない(図6)。GDSでは、「公正取引委員会」でも、間にスペースを入れた「公正 取引 委員会」でも結果は同じだった。このように、GDSでは、あまり検索語の区切りなどを気にすることなく入力ができる。

図5 「インターネットマガジンの記事」という検索では、「インターネットマガジン」、「の」、「記事」という3つの語を同時に含むデータがヒットする。太字になっているのが対象語である。

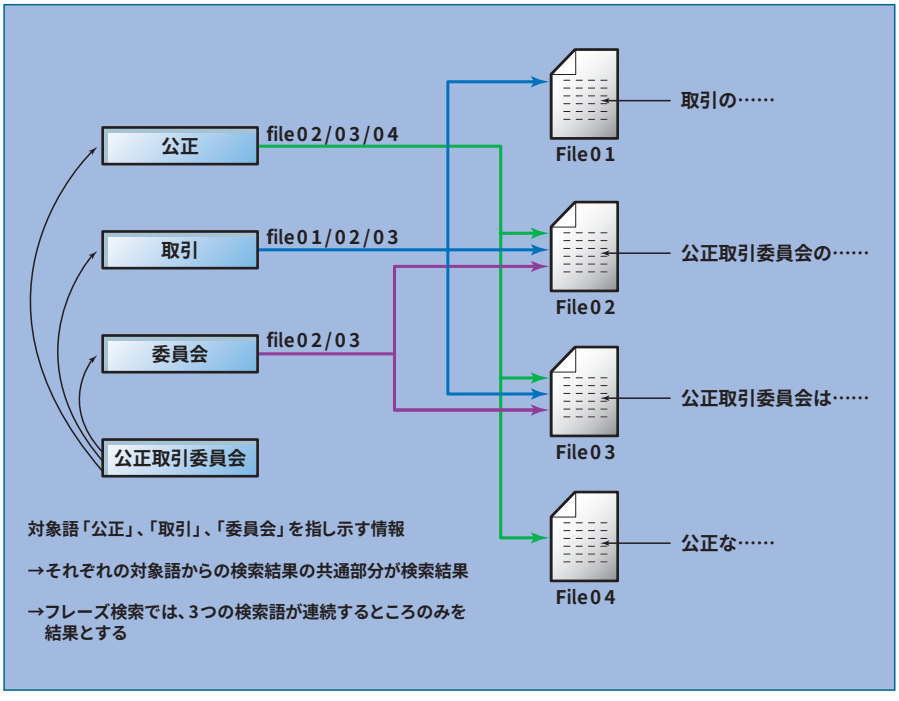

図6 GDSでは、複数の対象語から構成される語は、それぞれの対象語を指し示す情報として格納されている可能性がある。

Google Desktop Searchの日本語版は、ベータ版ながら、高速な検索が可能で、多少あいまいにキーワードを並べても検索が可能だ。日本語に関してもかなりの実用レベルにあると見ていいだろう。いくつか奇妙な動作も見受けられるが、ベータ版でもあるし、これについては正式版に期待しよう。

［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