デジタルサーチャー講座

検索を科学する

塩田 紳二

第5回 連想検索を可能にするGETA

多くの検索では、ユーザーが指定したキーワードを、対象データの中から直接探し出す。このため、キーワードの選択が検索結果を直接左右する。つまり、正しいキーワードを使えば正しく結果を得られるが、キーワードが正しくなければ正しい結果を得られない。
こうした問題を解決できそうな技術の1つが連想検索である。今回は、この連想検索について、計算エンジンの「GETA」を例に解説する。

連想検索とは？

連想検索とは簡単にいえば、入力された言葉から連想できるもの、あるいは関連性が高い情報を含むものを探す方法である。また、キーワードではなく、文章や文書全体を指定し、内容が似た文書を探すのも一種の連想検索である。

ただ、連想といっても、人間の連想の中には、個人的な記憶や経験によるもの、たとえば「犬→怖い」といったものがあるが、実際の検索では、こうした連想はあまり役立たない。何らかの意味的なつながりがあったり、特定の分野での頻度が高いもの同士といった論理的な連想だけが検索では有効だ。最近のいい方をすれば「○○つながり」が、検索分野でいうところの連想である。たとえば、「ニワトリ」をユーザーが指定したときに、ニワトリは含まないが「ヒヨコ」を含む文書を検索できるというのが連想検索である。

ただし、この連想検索だが、キーワード検索に比べると実現が難しい。コンピュータは人間とは違うため、何の情報もなしに連想してはくれない。連想のための辞書を作ればいいのではないかとも思えるが、実際にはこうしたものを作るのはかなり大変だ。「ニワトリ」と「ヒヨコ」といった連想は簡単に浮かんでも、「砂袋」(砂嚢の俗称。いわゆる砂肝)なんて単語がニワトリと関係するなんてことは、なかなか思いつかない。これをおよそ考えられるキーワードすべてに対して作るなんてことは簡単な作業ではない。また、キーワードを特定することも困難だろう。労力に対して、効果がほとんどないといえる。

文書が提示されたときに、それに似た文書を探すという連想も簡単ではない。人間は、文書全体を読まなくともタイトルや最初の一部を読んだだけで、文書の類似性(あるいは文書同士が無関係なこと)を簡単に判断できる。しかし、コンピュータがこれを行うのは簡単ではない。

コンピュータには、人間が持っているような全体を一度に把握するといった機能がなく、たとえば、デジカメで撮影した2つの写真がまったく同じビットマップを持つかどうかでさえ、ドット単位で全部調べなければ一致を判定できない。同じものを別々に撮影した2つのデジカメ写真データが同じものを撮影したのかどうかは、コンピュータにとっては、さらに判定が難しい。

こうした問題があるため、連想検索を実現するためにはいろいろと工夫が必要になる。

連想計算

連想検索で、キーワードと連想される単語(複数)は、立場を変えてもその関係が成り立つはずである。つまり、連想は一方通行ではなく、双方向であり、お互いに連想できる単語がグループ(集合)を構成していることになる。

では、このグループはどうやって出来上がるのかというと、これは検索対象の文書が作り出すことになる(図1)。通常、文書が含む用語のうち特徴的なものは、そ

の文書が記述している事柄に対して相互に連想が可能な集まりとなる。たとえば、ニワトリについての文書であれば、ニワトリに関するキーワードを多数含み、関係のない事柄を表すキーワードは少ないはずである。

辞書を作らなくとも、検索される文書自体から連想のための単語を取り出せばいいわけだ。対象の文書から取り出した単語を使って、連想を行えるような集合を作り出せば、結果的に連想検索できるはずである(図2)。なんだか、答えを見てから問題を作るような感じだが、世の中の検索システムはすべて、検索される対象(文書)は特定されていて、そこから何らかの情報を取り出しておくことは検索の常道手段ともいえる。

どんな文書も単語の利用頻度には偏りがある。すべての単語を使う文書というのはありえない。また、ある程度の長さであれば、異なる2つの文書の単語とその頻度が完全に同一になるということもない。つまり、文書の単語の出現頻度は、文書固有のパターンを持っているといえる(図3)。

そのうちのいくつかの単語は、特定の事柄について記述した文書にしか現れないなどの特徴を持つ。

これを利用すれば、ユーザーが入力したキーワードを含む文書があったとき、その単語が文書の中でどの程度重要なのかを調べることが可能だ。キーワードが、文書の特徴となるような単語と一致すれば、その文書は、ユーザーが探している文書である可能性が高い。この処理は、あらかじめ、文書から出現単語とその頻度の表を作っておけば、すぐにでも見つけることができる(図4)。

また、この文書の特徴を表すような単語が得られれば、今度は、これから似たような文書を探すことが可能になる。これも、文書同士を直接比較するのではなく、出現単語の頻度表を使って処理することが可能だ。

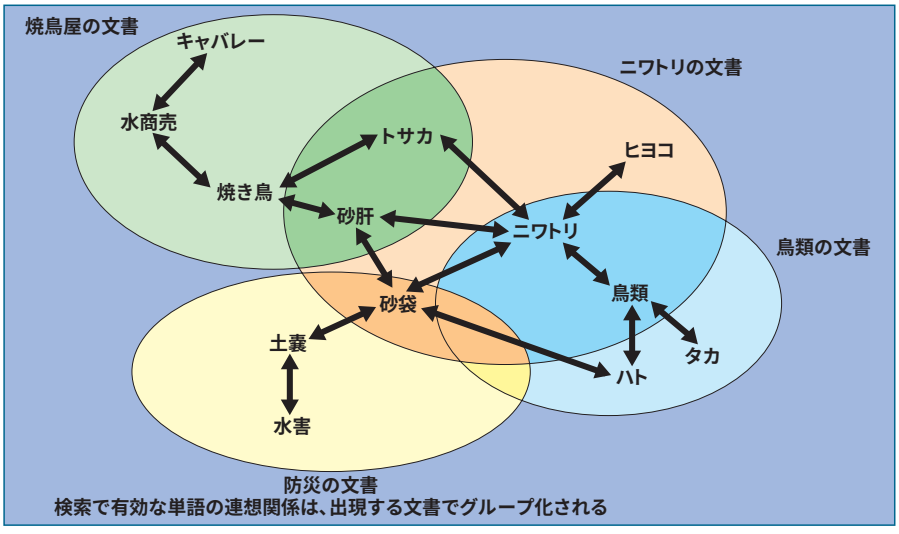

図1　検索として有用な連想は、文書に対応してグループを作る。

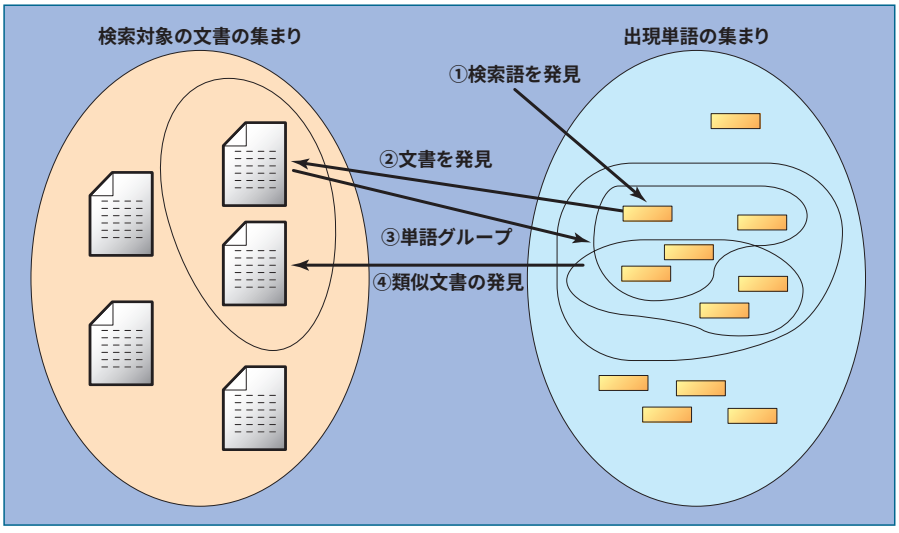

図2　文書と出現単語のグループは対応しており、キーワードを使って文書を見つけることができる。さらにその文書特有の単語から、似た文書を見つけることが可能。この文書は、最初のキーワードを含んでいる必要はない。

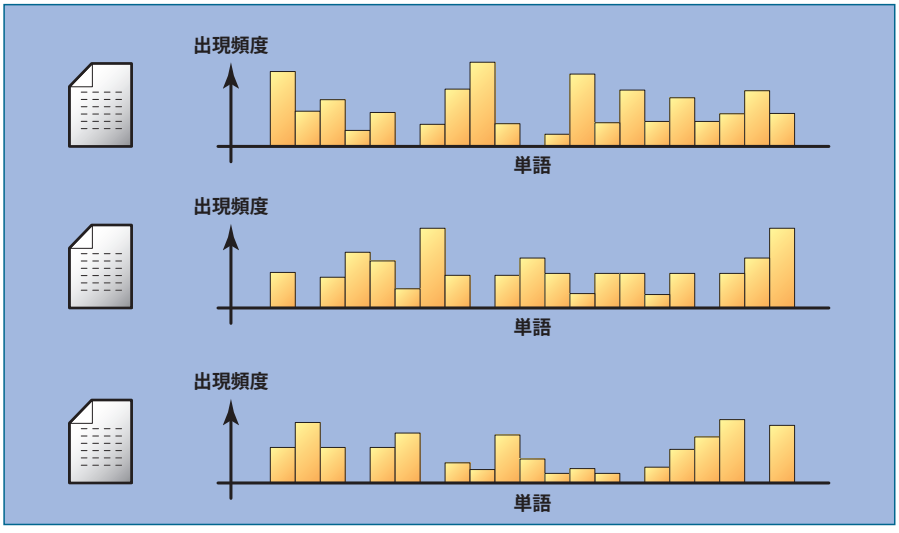

図3　文書の単語の使用には偏りがあり、単語の出現頻度を表にすると、文書ごとに違ったパターンを作る。

このようにすれば、ユーザーが入力したキーワードを直接含まない文書を探すことが可能になる。

計算エンジン「GETA」

今回は、10月号の特集でも紹介したGETA（Generic Engine for Transposable Association：国立情報学研究所）で、連想検索の仕組みを探ってみることにする。概要などについては10月号の特集を参照頂きたい。GETAは、前述のような連想検索システムを作るための基本的な演算機能（連想計算）を提供するプログラムライブラリーである。すでにいくつかのサイトの検索機能で利用されており、その場合、GETAのロゴが付けられている（図5）。

実際にGETAが提供するのは、文書から抽出した単語の頻度リストと、指定された単語や単語群を使って連想計算を行い、最も近い文書や、それらが持つ特徴単語を取り出すことである。

簡単にいうと、連想計算は、文書別に単語とその頻度を表にしたデータ（出現頻度表）を作り、これに対して、ユーザーが指定した単語群との類似を判定するものだ。

ただし、一般に文書を構成する単語の数は多く、表は大きなものとなってしまう。GETAは、こうした大きな表の計算を効率的に行うような仕組みを持ち、一般的なPCクラスの性能で、1000文書を対象にした連想計算を2～3秒で行うことができる。

また、GETAは、連想検索以外にも応用が可能だ。たとえば、文書の類似性を判定できることを利用すれば、多数の文書を類似性でグループ化することができる。このような処理をクラスター化、クラスタリングと呼ぶ（図6）。簡単にいえば、文書が分野ごとに整理されて棚に並んでいるようなものである。こうした処理を自動化することもできるわけだ。

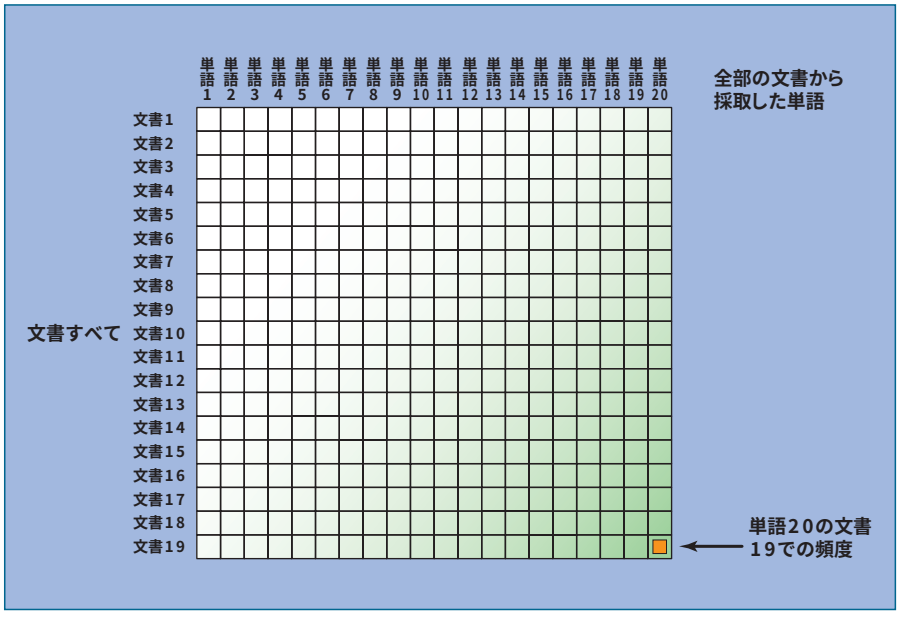

図4　GETAでは、全文書から収集した単語を各文書内での出現頻度の表にして連想計算を行う。この表は、大きなものになる可能性がある。

図5　GETAを使ったサイトの例。これは国立情報学研究所が運営する図書検索サービスの「Webcat Puls」
http://webcatplus.nii.ac.jp/

類似性の計算

もう1つ、GETAは、この単語群の類似性を判定する計算をユーザーが定義できるという特徴を持つ。

GETAは、この類似性を計算する式を使って、ユーザーが指定した単語と表データから、各文書に対して評価値（スコア）を計算する。そして、評価値の高いものを候補として挙げるわけだ。

この計算式にはいくつかの種類があるが、GETAでの基本的な考え方は、検査対象の文書の中での単語の重要性と文書全体の中での単語の重要性を使って、単語（複数の可能性もある）と文書の類似性を計算することである。ただし、このとき、文書の長さなどを考慮する必要がある。というのは、文書が長ければ長いほど多くの単語を含むことになるため、短い文書に比較して有利になってしまうからである。このような現象を抑制すること

を「正規化」という。正規化することで、重要度の判定を文書の長さなどに依存せずに評価することが可能になるわけだ。

GETAの動作

キーワードを1つだけ指定した場合、GETAによる連想計算は、重要度を考慮したキーワード検索になる。すでに各文書の出現単語の頻度表が出来上がっているので、該当のキーワードを含む文書を探すのは簡単に行える。その上で、単語の重要性を考慮すれば、キーワードが重要な単語となっている文書を選ぶことができる。

これで1つの文書が選択されたら、今度は、その文書の頻度表を使えば、キーワードから連想される単語を取り出すことができる。この場合にも、文書に含まれる各単語の重要度を評価することで、関連が深いと思われる単語だけを抽出することが可能だ。

さらにこれらの単語を使ってもう一度他の文書との関連性を探すことも可能だ。この処理は、結局、最初の処理で単語が複数になっただけに過ぎない。

実は、GETAが行うのは、単語と頻度表を使って重要度を計算するだけなのである。ただし、この計算は簡単ではない。

最も大きな理由は、頻度表が巨大なものになるからである。頻度表は、すべての文書から採集された単語とその頻度を記録したものなので、実際には、ほとんどの項目がゼロとなる。このような表形式のデータを「疎行列」と呼ぶ。

GETAは、巨大な配列を圧縮し、オンメモリーで計算できるようするなどのさまざまなテクニックを使って、こうした巨大な行列の計算を高速に行えるようにしたプログラムライブラリーなのである。

また、大量の文書を検索対象とするとき、GETAは、頻度表を分割して複数のマシンで並列に連想計算を行わせることも可能になっている。

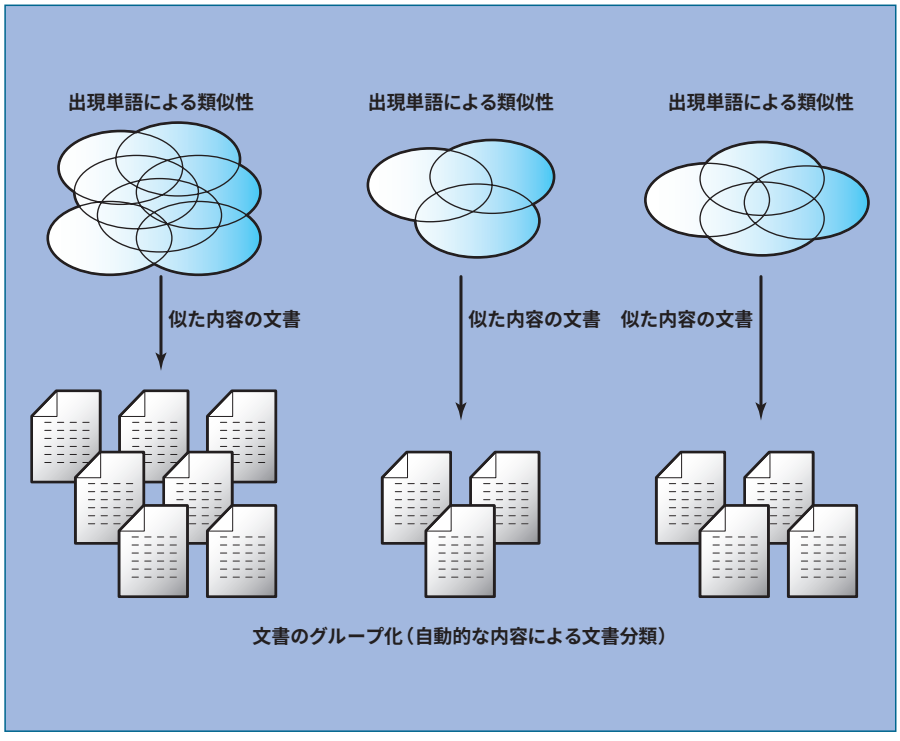

図6　文書同士の類似度を連想計算で求めることができるため、文書を類似度の高いものでグループ化することができる。

GETAに含まれないもの

ただし、GETAは、検索エンジンではなく、検索エンジンを作るための素材である。たとえば、ユーザーが入力したキーワードを検索できるように前処理したり、入力された文章からキーワードを抽出したりするような処理などは含まれない。日本語では、英語のように簡単には単語への分解が行えない。しかし、オープンソースとして入手可能な「茶筅」などの「形態素解析ソフト」を使うことで、文章を解析してキーワードを取り出すことが可能になる。形態素解析とは、辞書などを使って、文を言語として意味を持つ最小のまとまりである形態素に分解して品詞を見分けることをいう。

同様に頻度表を作るためには、検索対象データから単語を抽出し、それを数えなければならない。こうしたデータを作る部分もGETAには含まれない。

ただ、連想検索を行う検索エンジンを作ったときに最も重要な連想計算自体は、GETAが高速に行ってくれるため、小さな労力でパフォーマンスの高い検索エンジンを作ることが可能だ。GETAを使うインターネットのウェブサイトには、GETAの開発チームが直接関与したところもあるが、それ以外にも、配布されているドキュメントとソースコードから独自の検索エンジンを作り上げたところもあるという。

GETAを使ったサイトで検索を行ってみると、その検索結果が的確であることがわかる。もちろん、GETAといえども万能ではない。単語だけを入れるよりも、複数単語に分解される文章を入れたほうが、よりよい結果を得られるようだ。

ウェブページを見ていて気になったことを調べるときに、キーワードをどうするかを考えるのでなく、そこから抜きだした文章やテキスト全部を入れてしまえば検索できるのがGETAを使った連想検索の強みだ。また、最初の検索で得られた重要単語を使ってGoogleのような広範囲をカバーするキーワード検索を行うこともできるだろう。

［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