ヒューマンインタフェース 76－5
情 報 メ デ ィ ア 33－5
（1998. 1. 29）

# WWW検索サービスにおける検索結果絞り込み用インタフェースの開発

早川 和宏　　大久保 雅且　　田中 一男
{hayakawa, ohkubo, tanaka}@aether.hil.ntt.co.jp
NTT ヒューマンインタフェース研究所
〒239-0847 横須賀市光の丘1－1

## 概要

WWWの検索エンジンでは、1回の検索で数万件の結果が得られることが珍しくない。しかし、その過半数が検索意図に適合しないのが普通である。この検索結果を利用者の検索意図に近づけつつ数百件程度まで絞り込んでいくことが重要である。本稿では、検索結果の絞り込みを支援するための、検索結果の可視化と追加検索語候補提示を行うユーザインタフェースについて述べる。

# A Visual User Interface for Refining Search Engine Results

Kazuhiro Hayakawa　　Masaaki Ohkubo　　Kazuo Tanaka
{hayakawa, ohkubo, tanaka}@aether.hil.ntt.co.jp
NTT Human Interface Laboratories
1-1 Hikarinooka, Yokosuka, Kanagawa 239-0847 Japan

## Abstract

We often get thousands of URLs by sending a query to a WWW search engine. However, it is very often that most of the search results does not match our purposes. Providing that the user's cost of information retrieval is the time spent for browsing information which does not match the user's purpose, the allowable cost for general users may be less than hundreds of URLs.
Thus it is clearly important for the users of the WWW search engines to reduce the search results to less than a hundred URLs. The task of refining query to reduce the search results is the most important step of the process of using search engines. In this paper, an user interface which helps refining query is discussed. Our prototype interface is also described.

## 1 はじめに

現在、WWWの全文検索エンジンのような大規模な検索エンジンでの重要な問題は、「絞り込み作業の効率化」である。利用者の意図を正確に酌んでくれる検索方法がない現在では、大規模な検索エンジンでは絞り込みを対話的に行う作業が必要不可欠である。

下山ら[1]は検索作業の専門家であるサーチャと一般ユーザの検索質問を比較し、一般ユーザはサーチャと比較すると「適当なキーワードをうまく選ぶことができず」、「検索式を作るのが下手」であると分析している。下山らの観察では、一般ユーザは検索方法が悪いために検索結果が1件も得られないまま検索が終わってしまう場合が多い。一方、現在の全文検索エンジンでは逆に検索結果が多すぎてうまく絞り込めないままになってしまう場合の方が多いと思われる。しかし、一般ユーザの検索方法が上記二つの点で問題であることは変わりないであろう。

検索を行う利用者の大半が一般ユーザである現在、「適当なキーワードの選択」と「検索式の作成」の二点を支援し、絞り込み作業を効率化するユーザインタフェースが必要であると考えられる。本稿では、このインタフェースについて検討し、著者らの試作したインタフェースについて述べる。

## 2 絞り込み作業の支援

絞り込み作業を支援するために、どのような方法が考えられるだろうか。著者らは、現在のWWW検索エンジンにおいて、利用者が絞り込み作業を行なう過程はおよそ以下のようであろうと考えた。

**1.** 探すものに関連しそうな検索語を考える。
**2.** 検索を行う。
**3a.** 検索結果の件数が十分少なければ絞り込みを終わり、結果を検討する。
**4a.** 件数が多すぎる場合、検索結果の内容を見て、絞り込みに役立ちそうな語を探す。
**5a.** その語を前回の検索語に加えて、**2.** へ。
**3b.** 検索結果を見て、意図した情報が得られていれば検索を終わる。
**4b.** 意図した情報が得られなかった場合、別の検索語を考える。
**5b.** その語を新しい検索語として、**2.** へ。

ここで**2.** から**5.** の作業は試行錯誤的に何度か繰り返されるものであり、著者らが支援しようとしている作業である。

明らかに改善の余地があるのは**4a.** の絞り込みに役立つ語の探索である。この点については、既にいくつかの研究が行われている[6]。一方、**1.** や**4b.** のように、利用者しか知らない「検索の目的」に直接依存する作業は、計算機による支援は難しい。この部分については、[7]のように検索ログを分析するなどして、検索の目的そのものの典型例を収集し、あらかじめメニュー化して提示するのが現実的と思われる。

本研究では、**3a.** の検索結果が十分少なくなるかどうかの判断と、**4a.** の絞り込みに役立つ語の探索を含む作業ループを支援することを考える。そのためには

「ある語を検索式に追加すると検索結果件数がどのように変化するのか」

を提示することにより、

「提示された追加検索語候補の中で、どの語を追加するのがより適当か」

を判断しやすくすることが有効であろうと考えた。

すなわち、検索結果件数の変化が容易に推測できるような手がかりを、追加検索語候補と共に提示することで、追加検索語の選択をより適切に行うことができるのではないかということである。この背景には、絞り込みの進み具合の評価基準として、少なくとも「結果の件数」と「結果の内容」という二つの基準があるのではないかという仮説がある。たとえば、最初の検索の結果件数が2万件であり、ある単語はそれを5千件に絞り込むが、別の単語はそれを50件に絞り込むとすれば、後者は前者よりも利用者にとってより好ましい単語であると言えるのではないか。もちろん、どちらの単語も利用者の検索意図に沿っているという前提が必要だが、それは利用者本人にしか分からないので計算機から提示することができない。一方、結果の件数は客観的な基準であり、絞り込み作業支援の手がかりとして計算機から提示することができる。

## 3 絞り込み作業支援UIの要件

上に述べたように、著者らは追加検索語候補から候補を選択する際に、検索語候補と検索結果件数との関係を提示することで、絞り込み作業を効率化することができるのではないかと考えた。これを具体化するユーザインタフェースには、二つの要件が考えられる。一つは検索語候補と検索結果件数との関係をわかりやすく提示することであり、もう一つは追加検索語の追加、すなわち検索式の変更を容易に行えることである。

第一の提示に関しては、検索語と検索結果文献集合との関係を図示する研究がいくつか行われている。これらの研究は情報可視化（Information Visualization）として大きくまとめることができる [2]。

情報可視化で用いられるもっとも直接的な認知モデルは、空間メタファを用いるものである。たとえば、検索語のひとつひとつと検索結果の各文献との関係を、ある空間の中での位置関係と考える。すると、文献同士も検索語を媒介としてある位置関係を持つ。このようにして作られる文献空間（あるいはもっと一般的に情報空間）を可視化することによって、文献探索・情報探索を支援する手法は10件以上も提案されている。これまでに提案された可視化手法の典型は、文献間、ないし文献と検索語との関係をネットワークとして、または散布図として図示するものである [4][8]。可視化結果からは、ある文献に関連する文献、ある文献に関連する語、ある検索語に関連する文献等を獲得することができる。文献や検索語を地図的に配置することにより、人間の「場所に基づく記憶」や「配置に関する記憶」を有効に利用することができる。

その一方、これまでの可視化手法は、第二の要件である検索式の変更手段をあまり提供していない。空間メタファに基づくシステムでは、「キーワード検索」や「AND 検索」「OR 検索」といった、視覚的に表現しにくい抽象的・論理的な操作を、可視化された空間へマッピングすることが難しい。たとえば、現在の検索式に、ある検索語を AND で追加するという操作を、可視化された空間内での操作にうまく対応づけられない。したがって、初心者には扱いが難しい、論理演算子を用いた検索式を隠蔽することができない。

これは、たとえば Macintosh Finder がフォルダや書類、ごみ箱といったアイコンの移動で、ファイルの「移動」「削除」といった操作を無理なく表現していたことと対照的である。集合同士の関連を表現するためによく用いられる Venn 図を用いた InfoCrystal[9] は数少ない例外であると思われるが、Venn 図は検索語が3つまでの場合は検索条件をうまく可視化できるが、4つ以上になると視覚化されない検索条件が現れてくる点で問題がある。

そこで著者らは、上記のような検索に伴う各種の抽象的な操作を、自然に行わせるような視覚的アフォーダンスを備え、同時に「ある検索語は検索結果件数をどのくらい絞り込めるか」「現在の検索結果と、どの検索語がどのように関連しているか」という情報を可視化する能力を持つインタフェースの開発を試みた。

## 4 マトリクス形式の文献空間可視化

検索結果の文献 $d_1..d_m$ と、それを形態素解析して得られる単語 $w_1..w_n$ について、$w_1..w_n$ をそれぞれ次元と考え、$d_i$ の中で $w_j$ が出現する回数を $tf_{ij}$ とすれば、文献 $d_i$ を $(tf_{i1},..,tf_{in})$ というベクトルで表すことができ、そのベクトルを並べて $m$ 行 $n$ 列のマトリクスができる。このマトリクスが、理論的には利用者が操作しうる文献空間になる。たとえばこの検索結果集合に対してさらに単語 $w_j$ で絞り込み検索を行なうとは、$tf_{ij} > 0$ であるような行 $d_i$ のみを集めることになる。

従来の可視化手法は、この文献ベクトルを何らかの形で2次元平面あるいは3次元空間へ投影し、各文献の空間内での布置を散布図やネットワークとして表現する。この方法はある文献間の関連や文献と単語の関係などを空間的に提示することができるが、「ある条件を満たすような行（あるいは列）だけを選ぶ」という絞り込み検索のシンプルなモデルが失われてしまっている。

そこで、著者らは単語や文献の空間内での布置を提示するのではなく、単語と文献を行と列に持つマトリクスを直接ユーザに表の形で提示

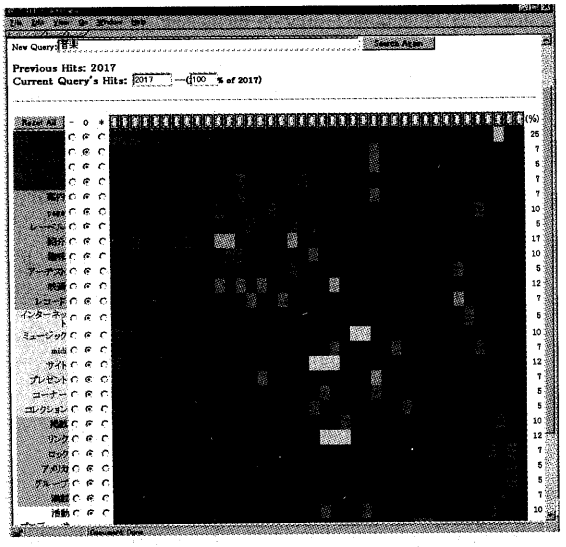


図 1: インタフェース全体像

すれば、さまざまな検索操作を行や列に対する直接操作として実現できると考えた。そのためには、数千文献×数百単語にもなるマトリクスをなるべく小さくすることが必要になる [5]。文献数を減らすために、検索結果の文献の一部だけをサンプリングし、その中での単語の出現率から、絞り込みの結果が何件程度になるかを推定することにした。単語数を減らすためには、一般的な $tf * idf$ と呼ばれる指標を絞り込みという目的に合うように少し変更して単語の重みづけを行う。通常の $tf * idf$ は一つの文献にしか出現しない単語がもっとも大きな重みを得る。しかし、絞り込みにおいては特定の文献を探しているのではなく、文献集合中のある部分集合を探している。従って、ある部分集合に共通して出現する単語に大きな重みを与えるべきであると考えられる。著者らの用いた指標では、$i$ 番めの文献の $j$ 番めの単語の重みは $tf_{ij} * log(D/df_j) * log(df_j)$ となる。$D$ は総文献数、$df_j$ は単語 $w_j$ が含まれる文献の総数である。この指標では、数少ない、ただし一つよりもある程度多い文献だけに出現する語が、高い値を持つことになる。

このようにして小さくしたマトリクスを見やすくするために、文献ベクトルを主成分分析し、上位の主成分に対して高い重みを持つ単語を、絞り込み用単語候補として利用者に提示する。

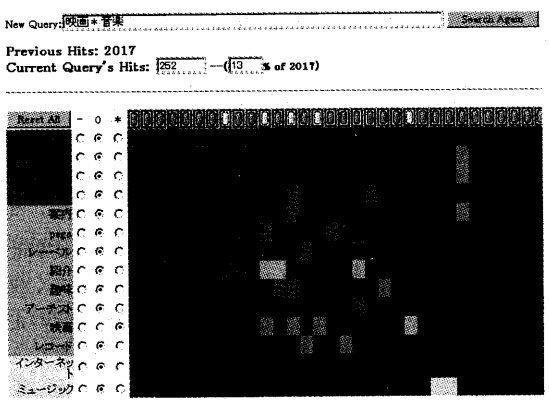


図 2: キーワードの追加

主成分分析を行なうことで、類似する文献、共起している単語同士を分類して、関連の深い単語集合と文献集合の組を強調できる。

## 5 絞り込み検索用ＵＩの試作

図 1は、以上に基づいて試作した NTT DIRECTORY [3] を検索するインタフェースの全体像である。図はキーワード「音楽」で検索を行なったところである。各行の先頭は絞り込み用検索語候補で、その右に並ぶー、０、＊の３つのボタンは NOT、無視、 AND の検索を指定する。行末にはサンプル内でのその語の出現率の数値も出力されている。

縦一列がサンプリングされた文献一つを表す。サンプルはこの図では 40 件である。マトリクスのセルの明るさは、縦の文献における横の単語の重みを表している。単語と文献は主成分に従って分類してあるので、内容が近いものは近くに並ぶ。図では特定の文献集団に特定の単語が集中して出現していることが見て取れる。

単語の横のボタンを押すと、新しい検索条件、その条件で元の検索結果集合の何％がヒットするか、その条件で再検索を行なった場合何件程度がヒットするかが自動的に表示される。

図 2は「音楽」の検索結果をさらに「映画」で AND 検索する指定を行なった所である。「音楽」で検索した結果件数は 2017 件で、うち「映画」を含むものは 13% あるので、「映画 AND 音楽」で検索すると、 252 件程度の結果が得ら

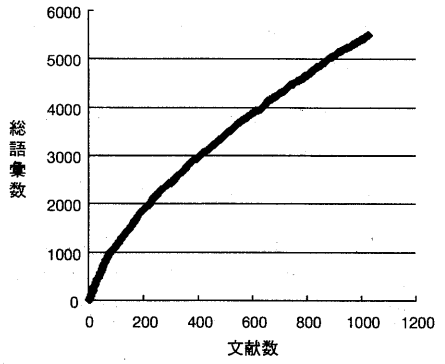


図 3: 総語彙数の増大

| n = 50 | n = 100 | n = 150 | n = 200 | n = 219(ALL) |
|---|---|---|---|---|
| おみくじ | おみくじ | homepage<br>おみくじ<br>イラスト | homepage<br>おみくじ | homepage |
| インターネット<br>オンライン<br>カード<br>ゲーム | インターネット<br>カード | インターネット<br>カード<br>ゲーム | インターネット | インターネット<br>ゲーム |
| サービス<br>サイト<br>タロット | コーナー<br>サービス<br>タロット | コーナー<br>サービス<br>サイト<br>タロット<br>プレゼント | コーナー<br>サービス<br>タロット<br>プレゼント | コーナー<br>サービス<br>タロット<br>プレゼント |
| メール<br>リンク<br>運勢<br>鑑定 | メール<br>リンク<br>運勢<br>鑑定<br>無料<br>結婚 | ページ<br>ホーム<br>メール<br>リンク<br>運勢<br>鑑定 | メール<br>リンク<br>運勢<br>鑑定 | メール<br>リンク<br>運勢<br>鑑定 |
| 結婚<br>情報<br>占星術 | 紹介<br>情報<br>星座 | 紹介<br>情報 | 紹介<br>情報<br>星座<br>占星術<br>販売 | 結婚<br>紹介<br>情報<br>星座<br>占星術<br>販売 |
| 無料<br>電子 | 無料 | 通信<br>無料 | 通信<br>無料 | |
| 11/17 | 12/17 | 13/17 | 15/17 | 17/17 |
| 11/17 | 12/16 | 13/21 | 15/18 | 17/17 |

図 4: サンプル数に対する検索語候補の変化

れるであろうことがわかる。最上部のランプは、点灯していればその文献が現在の検索条件に合致することを示し、クリックすれば文献の中身が表示される。

以上のように、このインタフェースでは、ある単語を使って絞り込みを行なうとどの程度結果を減らすことができるかが即座に分かる。また、インタフェース部分の動作はサーバで再検索を行なわずにクライアント側だけで行なえるので、軽快な操作を行なうことができる。

## 6 考察

本システムでは、検索質問に適合する文書の集合の一部をサンプリングして、集合全体の中での単語の出現頻度を推定しようとしている。従って、サンプルが元の文献集合をどの程度良く反映するのか、また推定された単語の出現頻度はどの程度正確であるのかを検討する必要があろう。

図 3は、ある検索質問に適合する文献集合において、集合全体に対するサンプルの割合と、サンプルの中に含まれる総語彙数の関係を示したものである。直観的には、あるトピックに関する文献に現れる語彙を調べているのであるから、サンプルがある程度多くなれば語彙数は収束していくように思われる。しかし実際には語彙数は収束せず、サンプルの増加に対してある割合で増加していく。

しかし、この結果はサンプリングを行うことが不適当であることを示しているのではない。重要なのは総語彙数の変化ではなく、提示される検索語候補がどのように変化するかである。図 4は、「占い」という単語で検索した場合に本システムで提示される追加検索語候補が、サンプルの量によってどのように変化しているかを示したものである。右端は文献集合全体を用いて求めた追加検索語候補であり、それ以外はそれぞれ 50、100、150、200 個の文献をサンプリングして求めた追加検索語候補である。下の 2 行は上から順に、文献集合全体を用いて求めた追加検索語候補を正解とした場合の、再現率と適合率を表す。網掛けになっている語は不正解の語である。

この図を見ると、全体の 1/4 弱をサンプリングした n=50 の場合でも、65 %程度に正解を再現しており、図 3から想像されるほど悪くはない結果が得られている。不正解だった単語もそれなりに「占い」との関連は認められ、まったく見当違いな検索語候補を提示してしまう可能性は小さいといえるであろう。これは、文献をサンプリングすると総語彙数は大きく減ってしまうが、減ってしまった語のほとんどは出現頻度が非常に小さいために、もともと追加検索語候補として選択される可能性が小さかったためであると考えられる。

次に、単語の出現確率の推定の精度について考える。これはサンプルをランダムに選ぶのであれば、サンプルの個数だけから計算できる。n 個のサンプルをとった場合に、そのうち R 個に単語が出現していたとする。すると、真の出現確率が R/n のまわりのある一定範囲に存在する確率を計算することができる。たとえば真の値が含まれている確率が 95 %の区間は

$$1.96\sqrt{R/n(1-R/n)/n} * 100(\%)$$

となる。

R=10, n=40 の場合、この値は 11 %、つまり確率 95 %で 14 ～ 36 %くらいの範囲に真の値があると言える。確率 90 %なら、1.96 の代わりに 1.645 を使って、誤差の区間は ± 7.7 %程度になる。R=4 であった場合には、± 9.3 %の区間に 95 %、± 7.8 %の区間に 90 %の確率で真の値が存在する。ただし、複数の単語を検索式に追加していくと誤差が累積していくので、この推定が意味があるのは単語 2 ～ 3 個の追加までであろう。

なお今後行う必要のある評価項目としては、単語および文献の分類の適切さの度合い、本インタフェースを用いることにより検索過程がどのように変化したかの 2 点がある。これらの定量的な評価手法はいまだ確立されていないが、引き続き検討していきたい。

## 7 まとめ

本稿では検索システムにおける絞り込み作業の効率化のためのインタフェースについて述べた。本インタフェースでは文献のサンプリングを行っているが、サンプリングしたことによる追加検索語候補の抽出精度、単語出現頻度の算出精度の低下について定量的に評価した。今後は、現在のプロトタイプのインタフェースの改善と共に、より効果的な検索語候補の抽出法、絞り込み操作の効率化の評価を行なっていく予定である。

## 参考文献


[1] 下山, 富士, 松井：サーチャのノウハウに見る検索インタフェース, 情処研報 92-HI-41,pp.9-16, 1992.

[2] Card, S. K. : Visualizing Retrieved Information: A Survey, IEEE Computer Graphics and Applications, Vol. 16, No. 2, pp.63-67, 1996.

[3] 田中 : InfoBee 検索エンジンを用いたディレクトリ検索サービス, NTT 技術ジャーナル, Vol. 8, No. 8, pp.24-27, 1996.

[4] 早川, 福永, 鈴木：ユーザの利用履歴に基づく WWW サーバの地図型ディレクトリ, 情処研報 97-HI-70,pp.17-24, 1997.

[5] 早川, 井上, 大久保, 田中：検索結果の文献集合を視覚的に提示するインタフェースの提案, 第 55 回情処全国大会, 1997.

[6] 井上, 杉崎, 早川, 田中：追加検索語候補提示に関する一検討, 第 55 回情処全国大会, 1997.

[7] 大久保, 杉崎, 早川, 田中 : WWW 検索ログに基づく情報ニーズ傾向の把握, 第 55 回情処全国大会, 1997.

[8] 舘村 : DocSpace: 文献空間のインタラクティブ視覚化, インタラクティブシステムとソフトウエア IV, 近代科学社, 1996.

[9] Spoerri, A.: Visual Tools for Information Retrieval, Proceedings IEEE Symposium on Visual Languages, pp. 160-168, 1993.