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

2W－1

早川 和宏　井上 孝史　大久保 雅且　田中 一男
NTT ヒューマンインタフェース研究所

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。検索エンジンでは、1回の検索で数万件の検索結果が得られることも珍しくない。その検索結果を、より利用者の検索意図に近づけつつ、数百～数十件まで絞り込んでいくことが利用者の重要な利益になる。

著者らは、(1) 検索結果の文書と各検索語との関連、(2) 検索結果の文書と検索語以外の語との関連、(3) 検索語間および文書間の関連、という情報を可視化することにより、絞り込みに必要な、検索結果件数の確認、検索語の追加という作業を支援しようと考えた。また、同時に従来の情報可視化手法 [1] であまり重視されていなかった、AND、OR など検索に伴う各種の抽象的な概念を直接操作として行なうことができるインタフェースを目指した。

2　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。

従来の手法では、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提示することができる。しかし、「文献と単語や他の文献との画面上での距離」という量的な情報は検索の AND や OR という論理的な条件とは直接関連しないし、「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提示することを考えた。マトリクスを表の形で提示すれば、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。

しかし、マトリクスの大きさは数千文献×数百単語にもなるので、これをすべて提示することは実装上も実用上も現実的ではない。従って、このマトリクスの行（文献数）と列（単語数）をなるべく小さくすることが必要になる。

そこで、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。これは「新しい検索語を用いた時、検索結果の件数がどう変化するかをユーザに提示する必要がある」が、「件数の数値は完全に正確である必要はない」からである。

単語数を減らすためには、意味のない単語を除き、なるべく重要そうな単語だけを選ぶ必要がある。著者らはまず一般的な $tf*idf$ と呼ばれる指標で単語の重みづけを行なった。i 番めの文献の j 番めの単語の重みは $tf_{ij}*log(D/df_j)$ 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。従って数少ない特定の文献だけに頻繁に出現する語が、高い値を持つことになる。

このようにして作成した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こととした。予備的な実験では、分散の 70 %を説明するためには上位 10 ～ 20 までの主成分が必要で、単語としては 30 ～ 40 個程度が抽出された。

主成分分析を行なうことで、類似する文献、共起している単語をある程度まとめることができる。単

A Visual User Interface for Search Engine Results.
Kazuhiro HAYAKAWA, Takafumi INOUE,
Masaaki OHKUBO and Kazuo TANAKA
NTT Human Interface Laboratories

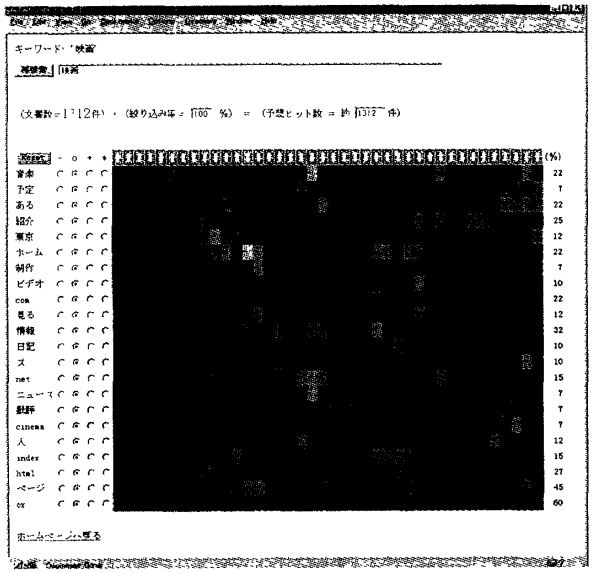

図 1: インタフェース全体像

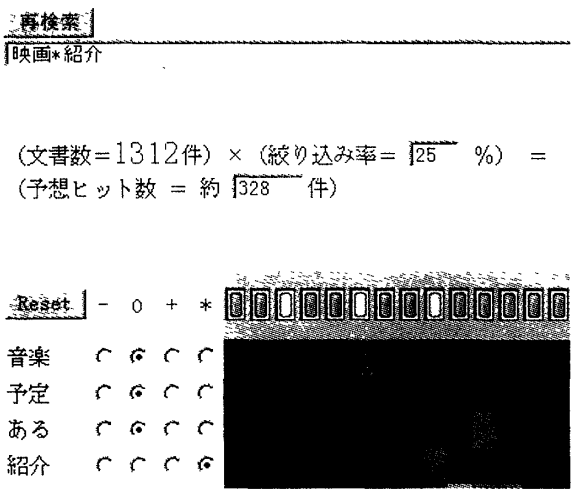

図 2: キーワードの追加

語は主成分の順で並べ、文献は最も重みの大きい主成分で分類すれば、単語同士の関連と文献同士の類似性も提示できる。

3 絞り込み検索用インタフェースの試作

以上のアイデアを元に、NTT DIRECTORY[2][3]（約 15 万件）を検索するインタフェースを試作した。

図 1は、インタフェースの全体像である。キーワード検索を行なうと、文献のサンプリングと重要語の抽出が検索サーバで行なわれ、文献と単語の関係の表など絞り込みに必要な情報がクライアント側に送られ、図 1のように表示される。

縦方向は絞り込み用検索語候補で、その右に並ぶ -,0,+,* の 4 つのボタンは NOT、何もしない、OR、AND の検索を指定する。単語の出現頻度は行を横に見れば視覚的に分かるが、行末には実際の出現率の数値も出力されている。

横方向はサンプリングされた各文献が並んでいる。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ソートしてあるので、内容が近いものは近くに並ぶ。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。

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「映画」の検索結果をさらに「紹介」で AND 検索する指定を行なった所である。「映画」で検索した結果件数は 1312 件で、うち「紹介」を含むものは 25% あるので、「映画 AND 紹介」で検索すると、328 件程度の結果が得られるであろうことがわかる。

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。

4 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という問題について述べ、絞り込みを支援するためにマトリクス形式の文献空間可視化を提案し、その実現上の技術として検索結果件数の確率的推定と文献空間の主成分分析の利用について述べた。

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、検索結果件数推定の正確さの評価、検索語候補のより正確な抽出法、絞り込み操作の効率化の検証を行なっていく予定である。

参考文献

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

[2] http://navi.ntt.co.jp/

[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.