情報アクセス支援における「特徴単語群の抽出」の利用

概 要

文書検索における対話的なガイダンス機能の一つとして、単語間の関連性を視覚的に表示するシステムを試作した。検索途中の文書集合に特徴的に出現する語群を相対的な頻度（当該文書集合における頻度と全体頻度の比）に基づいて抽出し、さらに特徴語相互間の関連性を共起統計（同じ文書中に使われやすい度合）で検出した。それらをグラフ状に表示することによりユーザは文書集合の傾向を一覧できるようになり、次の検索ステップを考える際に参考にすることができる。日経新聞一年分（約１８万記事）を対象とした実験を行ない、検索絞り込みなどの対話的な検索ガイダンスとして有効に利用できる見通しを得た。

1 諸言

ネットワークの普及により大量のオンライン文書へ手軽にアクセスできるようになったが、求める情報を的確に検索して利用することは必ずしも容易ではない。例えば何千という検索結果を前にして対処に窮するというのも珍しいことではない。そのため対話的、段階的に目的に近づくことを支援してくれるようなインタフェイスが求められている([13, p.1263-4])。

対話的な検索ガイダンスの一例としてはシソーラスの提示が有効である。検索タームの上位語や下位語、関連語などを提示することにより、ユーザは次第に適切なキーワードへ近付くことができる。しかしながらこのようなデータベースを用いる方法ではメンテナンスの問題があり、新語や複合語、また分野依存性の高い固有名詞や専門語などへの対処が課題となっている。

本研究の目的は統計的な手法を用いることにより検索作業と連動して関連語を自動抽出し、ユーザーに提示することにより、低コストで汎用的に使える検索ガイダンスを提供することである。

関連研究

対話的な検索インタフェイスを実現する手法には大きく分けてガイダンス機能とフィードバック機能(Salton [6]) があり、相補的な関係にあると考えれられる。本研究は関連語の提示に関するものでありガイダンス機能の一つである。

ガイダンス機能については近年情報の可視化という観点が注目されている(Rao et al. [5])。大きく分けて文書間の関係に着目する場合と出現単語に着目する場合があり、本研究は後者に属する。関連研究としては Scatter-Gather 法(Cutting et al. [1]) を挙げることができる。これは検索された文書群を自動分類(クラスタリング) して各クラスごとの特徴語を表示するものである。本研究との優劣は現段階では結論を出せないが、クラスタリ

ングは文書数が増えると計算が重くなり、また一般にクラスの特徴語として抽出された語群からそのクラスの性格が把握できる場合はそれほど多くないという問題もあり、本研究のように文書分類をせずに単純に単語間の関係を示すのも一法ではないかと考えている。

統計的関連語とシソーラスの融合については最近、イリノイ大学のグループが同大学で推進中の電子図書館システムの一環として取り組んでいる([2, 8])。まだ両者の融合による相乗効果が出るまでには至っていないように感じられるが重要なテーマだと思われる。

情報可視化のもう一つの流れである文書間の関連性については比較的研究例が多い。[7] の Smart システムは文書間の類似度に基づく文書関係マップを表示し、また Information Visualizer ([3]) では文書の参照関係をビジュアルに表現する。一般に文書の関連性を表示する場合、単に文書番号で示されても内容がつかめないのでユーザーはかえってフラストレーションを感じる恐れがある。少くともタイトルは表示しなければならないがタイトルといえどもそれほどコンパクトなものではないので表示には工夫が必要である。これまでの研究ではその点への配慮が十分でないように思われる。

日本での研究では東大の堀を中心とするグループにより発想支援システムの研究が続けられている([11])。杉本ら [10] は一画面中に文書と関連語を同時に表示する方法を提案しており注目される。またＩＢＭの諸橋らのグループによる Information Outlining に関する一連の研究 [4, 12]、三菱電機(ＲＷＣプロジェクト）の有田らによる情報散策システム [9] など、文書空間の可視化によるブラウジング手法として興味深い提案がいくつか行なわれている。

2　動的共起解析による関連語グラフの生成

2.1　実験方法

本方法による対話的な文書検索の流れを示したのが図1である。検索(上)によって得られた文書群(右)から特徴語と特徴語間の関連を抽出しグラフ状に表示する(下)。ユーザーはそれを参考にしてキーワードの追加など次の検索ステップに進む(上)という作業の繰り返しである。

以下では検索された文書群からの特徴語抽出、特徴語間の関連度計算、特徴語のグラフ表示方法について説明する。

2.1.1　特徴語抽出

特徴語抽出は基本的には検索された文書群における相対頻度(すなわち当該文書群における文書頻度と検索対象全体での文書頻度の比)が上位のものを取るという方針である。(文書頻度は当該単語が現れる文書の数の意味である。以下誤解の恐れがない場合には文書頻度の意味で「頻度」を用いることにする。）全体頻度は、検索対象となっている文書データ全体における当該単語の文書頻度であり、これは予めデータを作っておく。すなわち、全文書を形態素解析し、各単語の文書頻度をカウントする。形態素解析には京都大と奈良先端大が無償で公開しているＪＵＭＡＮシステム ([15]) を利用した。

上記の特徴語抽出方針は「当該文書群における相対頻度が大きい単語はその文書群に特

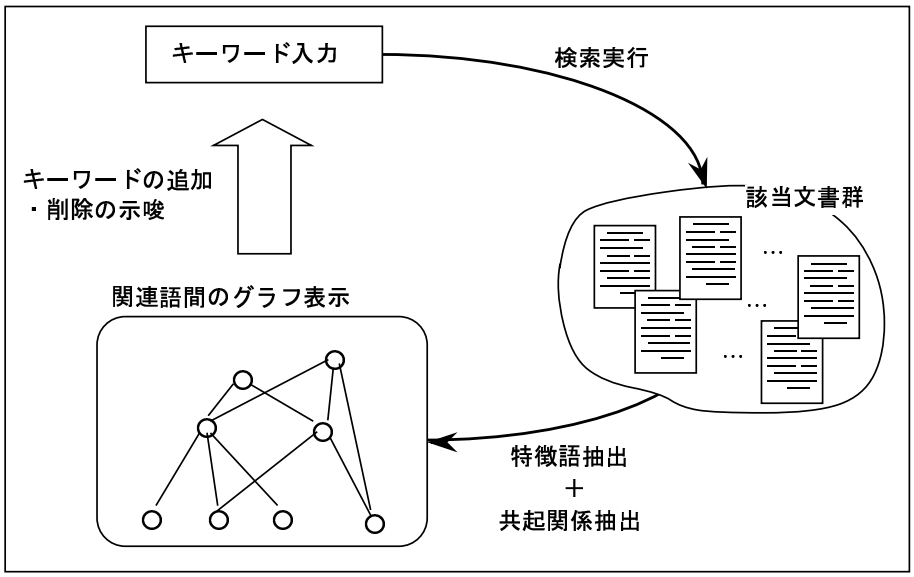


図1. 検索作業と連動した特徴語表示

徴的な語であろう」という直観的に分かりやすい考え方にもとづいているが、単純にこれを適用してしまうとどうしても頻度の低い語に有利になりがちであり、比較的頻度の高い重要語が落ちてしまうという問題点が生じてしまう。

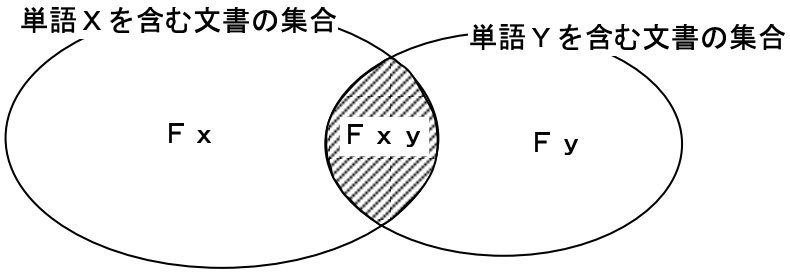
たとえば、「インターネット」で９４年の日経新聞を検索すると、２４９件のヒットがあるが、そららの中で、「パソコン」は９５件の記事に現れるが全体頻度も８５７２件と高いので相対比は０．０１強という値になってしまう。一方、たまたま現れたとしか考えられないような「土足」や「移り気」は全体頻度２に対して当該文書群での頻度が１なので、相対比が０．５となり「パソコン」に比べて約５０倍という非常に大きな値になってしまう。これは明らかに不合理であり「パソコン」の方がより関連性が高いと考えるのが妥当である。

このような不合理を解消するため、本実験では当該文書群に出現する単語集合をそれらの頻度に応じてクラス分けを行ない、それぞれのクラスの中で上記の相対頻度比が上位のものを取ることにした。これにより頻度が大幅に違う単語同士が比較されることがなくなり、低頻度から高頻度の語までバランスよく特徴語が抽出できるようになった。クラス分けのための頻度の閾値は最大頻度から出発して公比が0.5の等比数列とした。

2.1.2　単語間関連度

特徴語は単にリストの形で提示することも可能であるが、特徴語相互間の関連性も示せればより高い一覧効果が期待できる。関連性の高い一群の語は検索された文書群における何かあるまとまった話題に結び付いていると考えられるからである。

本研究では単語間の関連度は共起統計を用いて計測することにした。図2はそれを模式的に示したものである。今、二つの単語ＸとＹに着目しているとして両者が出現する文書の集合を二つの楕円で表している。これらの重なっている部分は単語ＸとＹが共起する文書の全体であり、この部分の占める割合によって両者の関連度を測ろうとする考え方で



$$\text{関連度}(\mathrm{X},\mathrm{Y}) \;=\; \frac{\mathrm{Fxy}}{\mathrm{Fy}}$$

図 2. 共起統計による単語間関連度の計算

ある。

計算式は図の下部に示した式を用いた。分子は単語ＸとＹを共に含む文書の数Ｆｘｙを取り、分母は単語Ｙの文書頻度Ｆｙとした。この値はＸとＹに関して対称ではない。Ｘからみた Ｙの関連度という意味合いで用いている。関連度の計算には、この他にも色々なバラエティがあり、最適な計算式は現段階では不明である。

2.1.3　グラフ表示

特徴語をノードとし、関連性の高い特徴語間にリンクを張ってグラフ表示を行なう。特徴語の選択は 2.1.1 節記載の方法により、指定個数を選択する。個数はユーザが指定できるが、試作システムではデフォールトを３０個に設定した。

次に各特徴語について、自身より (検索された文書群における) 頻度の高い語の各々について共起関連度 (2.1.2 節) を計算し、一番共起関連度の強い特徴語との間にリンクを張ることにした。

以上によりグラフの構造が決まったことになるが、それを平面上に表示するには各ノードのｘ ｙ座標を定める必要がある。これも選択の余地の大きい問題であるが、本実験ではｙ座標については基本的にその単語の文書頻度を取ることにした。ただし値を有限領域に正規化するため、表示される全特徴語の中で頻度がちょうど真中に来るものを基準としてそれとの比を取り、さらに対数を取った。

もう一方のｘ座標については特に意味づけはなく、ノードが重ならないよう以下のように再帰的に計算した。まずはじめに親ノードの無いノードについて、画面に均等になるようにｘ座標を決める。以下ｘ座標の決まったノードのみを親に持つノードの中で、親ノードの集合が共通のものについてそれらを均等に配置するようにｘ座標を決める。

このようにして決められた座標では一般に重なりが生じてしまうので左に配置されるノードから順に見ていき、重なりが生じる場合にはそれより右に来るノードをすべて右へずらして、重なりを避けるようにしている。

2.1.4　実験システム

検索対象としたのはＣＤ-ＲＯＭの日経新聞９４年版 ([14]) の記事約１８万件である。

検索キーワードとしては、全文をＪＵＭＡＮ[15] により形態素解析を行ない、各語の頻度を調べた後、頻度が２以上の単語（約８万語）についてインデックスを作成した。

2.2　実験結果

前節記載の方法に基づいて構築した実験システムを動作させた結果を一例により説明する。図3 ａ～ｆは「ニュートン」に関する検索における絞り込みの例を示したものである。

はじめに図3 ａは「ニュートン」をキーワードとして検索を実行したところである。中段左には検索ヒット件数、中段右にはタイトルが表示されている。タイトルを見ると性格の異なる記事が混在していることが分かる。例えば一番上の記事は航空機の墜落に関するもので、記事の中身(下段) を見るとアメリカのニュートン郡というところで起こった事故であることが分かる。

図3 ｂは上記検索結果(３５件)における関連語を表示したものである。これを見ると中心付近やや右寄りに「コンピュータ」という語が見られ、その周囲の単語群からアップル社の携帯端末「ニュートン」に関する記事が多く存在していることが伺われる。仮にこの利用者がこのニュートンに関心があるとすれば、例えば「コンピュータ」を加点キーワードとして加えることにより、検索の絞り込みを実現することができる。

(本検索システムでは３種類のキーワードを用いている。それぞれ必須キーワード、加点キーワード、減点キーワードと呼んでいるが、必須キーワードはそれらのキーワードのアンドで検索を実行することを意味し、加点キーワーはそれを含む文書のスコアを１点加点し、減点キーワードの場合には１点減点するようにしている。)

関連語の表示画面の中からキーワードを利用する場合には、はじめに使いたいキーワードをマウスでクリックし、利用したいキーワードの種類に応じて、該当するキーワードリスト欄の「＋」ボタンを押すとそこに追加されるようになっている。

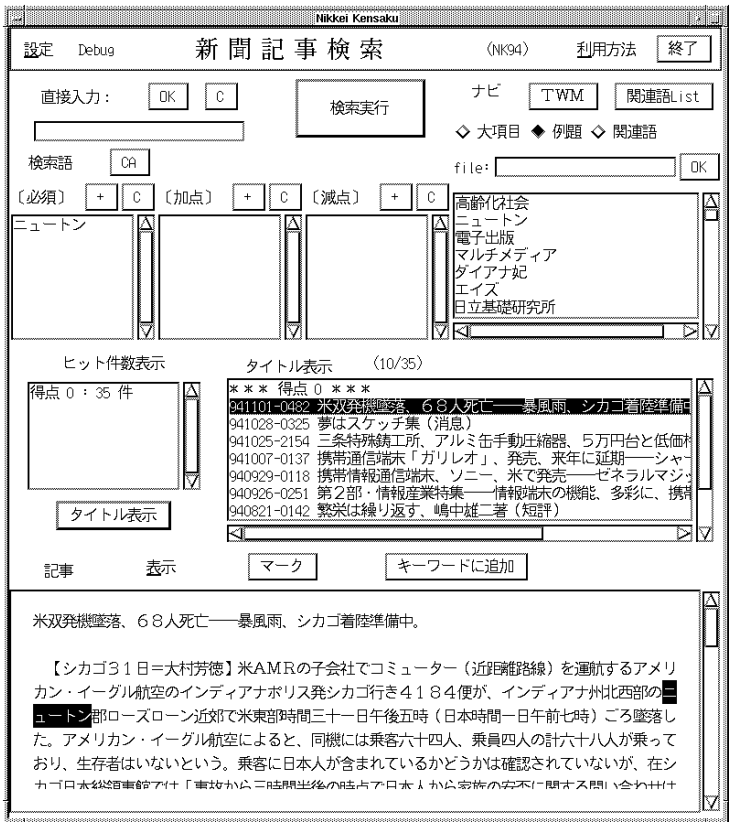


図３ａ 検索例「ニュートン」

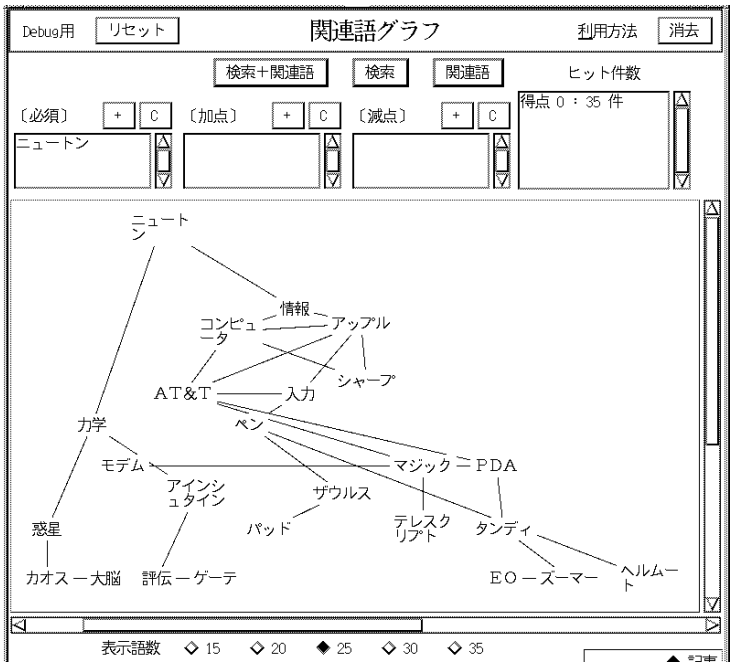


図３ｂ 「ニュートン」の関連語

このような操作により、「コンピュータ」を加点キーワードとして加えて検索を実行した結果が図３のｃである。表示されたタイトルを見ると確かに携帯端末「ニュートン」関

連の記事が選抜されていることが分かる。

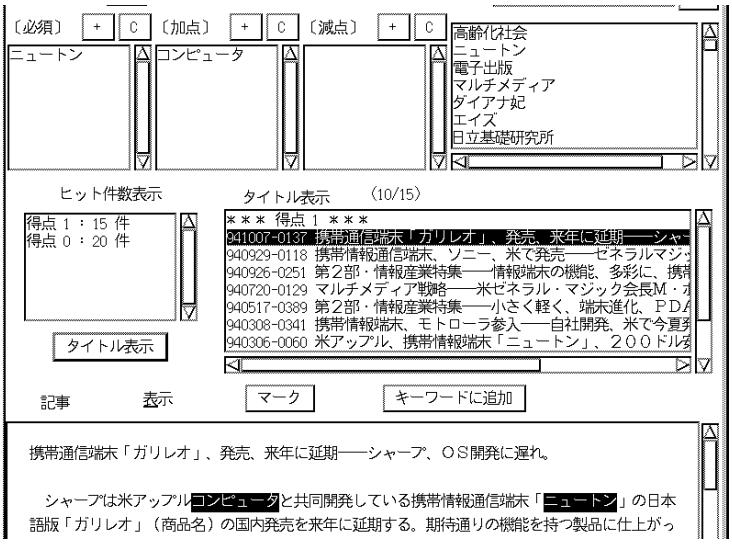


図３ｃ「コンピュータ」による絞り込み

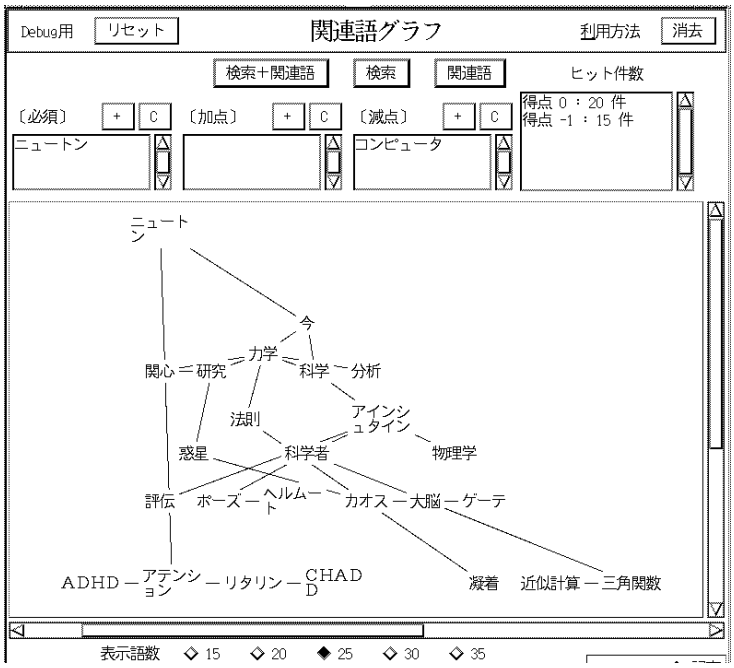


図３ｄ 「コンピュータ」以外のニュートン

次に携帯端末のニュートンではないニュートンに興味がある場合の操作を示す。この場合には「コンピュータ」を減点キーワードとすることによって対処できる。図３ｄは「コンピュータ」を減点キーワードとして関連語を抽出して表示したものである。表示された語群を見ると、今度は科学者あるいはニュートン力学のニュートンが目立ってくることが分かる。もしニュートンの力学法則に興味がある場合であれば「力学」「法則」などを加点キーワードとして検索すれば良い。それを実行したのが次の図３ｅである。４件のヒッ

ト中最初の記事「繁栄は繰り返す...」は一見するとニュートン力学とは関係がなさそうであるが、中身(下段)を見ると経済法則の説明にニュートン力学を比喩として用いている書物の書評であることが分かる。

図３ｅ「力学＋法則」による絞り込み

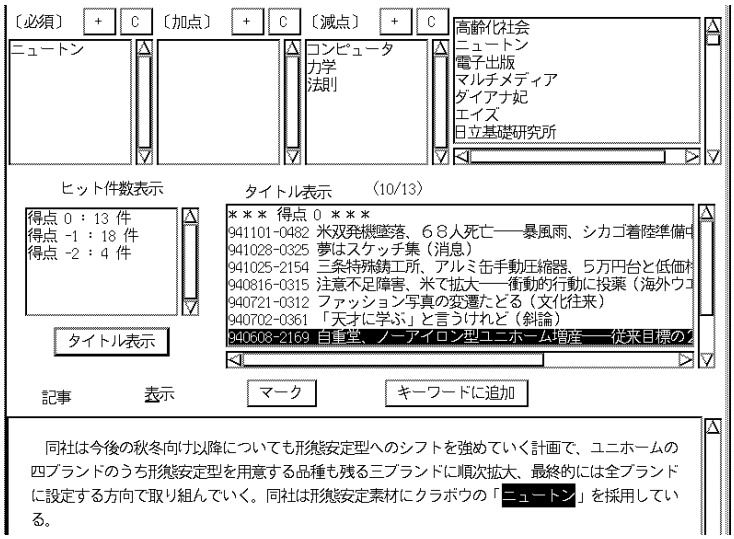


図３ｆ 「その他」のニュートン

さらに「コンピュータ」でも「力学法則」でもないニュートンについて検索したい場合には、これらをすべて減点キーワードとして検索を実行する。その結果が次の図３ｆである。今度は種々雑多なニュートンを含む記事が取れている。さきほどのニュートン郡における飛行機事故の記事もその一つとして現れている。図３ｆの記事表示欄(下段)に示したのは形態安定素材のニュートンに関する記事である。

以上でニュートンの例を離れ、図4(ａ～ｄ)では他の検索例での関連語グラフを列挙し

た。ａとｂは比較的最近の話題として「もんじゅ」と「ダイアナ妃」で検索した場合である(ただし検索対象は平成６年の日経新聞)。また少し長めの検索要求の例として高齢化社会と地球温暖化防止の例を示したのが図4 ｃ～ｄ (次ページ) である。

検索要求は形態素解析に通して単語分割し、それぞれの単語を検索キーワードとして利用する。例えば「地球温暖化防止」は「地球／温暖／化／防止」と単語分割される。(「化」は除外語としているので除かれ、実際にはそれ以外の３語がキーワードとなる。)

これらの結果の良し悪しの評価は今後の課題であるが、直観的にはそれぞれの検索要求と関係のありそうな単語が多く取れている、と言えるのではないかと考えている。

3　結言

・文書検索における対話的なガイダンス機能として、検索された文書群に特徴的に現れる語群を統計的に自動抽出し、さらに特徴語相互間の関連性が見やすいようにグラフの形で表示するインタフェイスを試作した。

・新聞１年分（約１８万記事）を対象にした実験により、本方法による関連語のグラフ表示が検索の絞り込みや意外な記事の発見に有効に利用できる見通しを得た。

・文書集合からの特徴語抽出については、予め頻度によるクラス分けを行ない、各クラスごとに相対頻度比(当該文書集合における頻度と全体頻度との比)が上位のものを取るようにした。これにより低頻度語から高頻度語までバランス良く特徴語が抽出できるようになった。

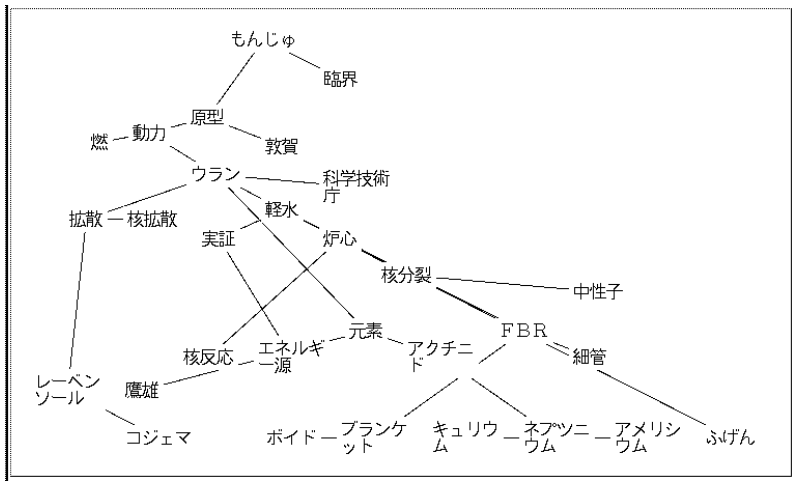


図4a. 「もんじゅ」

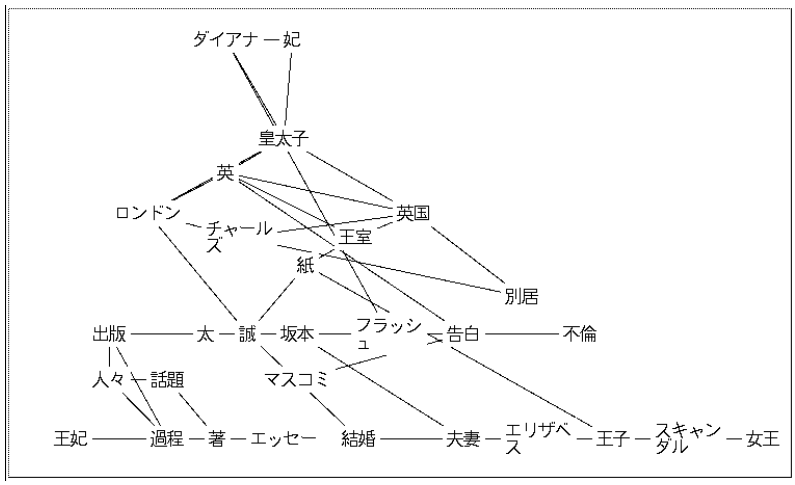


図4b. 「ダイアナ妃」

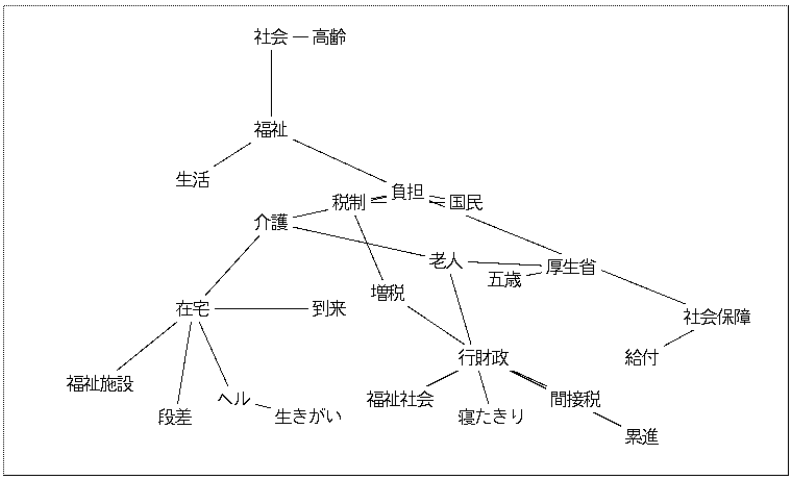


図4c. 「高齢化社会」

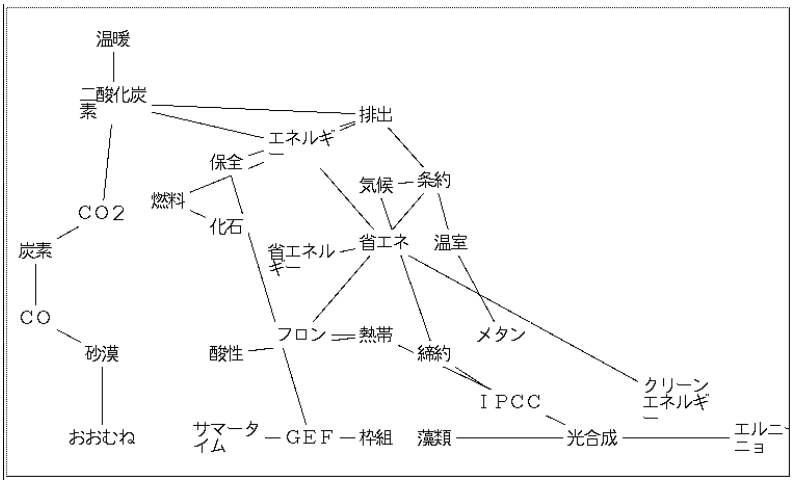


図4d. 「地球温暖化防止」

参考文献

[1] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Ann Int'l SIGIR'92*, pp. 318–329, 1992.

[2] Eric H. Johnson and Pauline A. Cochrane. A hypertextual interface for a searcher's thesaurus. In *Proceedings of the Digital Libraries '95*, Austin, Texas, 1995. Center for the Study of Digital Libraries, Texas A&M University.

[3] Jock D. Mackinlay, Ramana Rao, and Stuart K. Card. An organic user interface for searching citation links. In *Proceedings of ACM CHI'95*, pp. 67–73, Denver, Colorado, 1995. ACM.

[4] Masayuki Morohashi, Koichi Takeda, Hiroshi Nomiyama, and Hiroshi Maruyama. Information outlining - filling the gap between visualization and navigation in digital libraries. In *Proceedings of International Symposium on Digital Libraries*, pp. 151–158, Tsukuba, 1995.

[5] Ramana Rao, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and George G. Robertson. Rich interaction in the digital library. *Communications of the ACM*, Vol. 38, No. 4, pp. 29–39, 1995.

[6] Gerard Salton. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, NY, 1968.

[7] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, Vol. 264, pp. 1421–1426, June 1994.

[8] Bruce R. Schatz, Eric H. Johnson, and Pauline A. Cochrane. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of ACM DL'96*, pp. (to appear), Bethesda, Maryland, 1996. ACM.

[9] 有田英一, 安井照昌, 津高新一郎. 単語集合の自動構造化機能を持つ「情報散策」方式. 情報処理学会・自然言語処理研究会報告, Vol. 95-NL-108, pp. 69–74, 1995.

[10] 杉本雅則, 小山照夫, 堀浩一, 大須賀節雄, 絹川博之, 間瀬久雄. 文書間の関連性を可視化することによる文献検索システム. 情報処理学会・自然言語処理研究会報告, Vol. 96-NL-112, pp. 15–22, 1996.

[11] 角康之, 小川竜太, 堀浩一, 大須賀節雄, 間瀬健二. 思考空間の可視化によるコミュニケーション支援システムＣＳＳ. 電子情報通信学会・信学技報, Vol. TL95, No. 6, pp. 11–22, 1995.

[12] 武田浩一. テキスト情報の可視化による情報検索. 言語処理学会第２回年次大会発表論文集, pp. 121–124. 言語処理学会, 1996.

[13] 藤澤浩道, 絹川博之. 情報検索における自然言語処理. 情報処理, Vol. 34, No. 10, pp. 1259–1265, 1993.

[14] 日本経済新聞社. 日本経済新聞ＣＤ－ＲＯＭ１９９４年版, 1994.

[15] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム juman, 使用説明書 version 2.0, 1994.