博士学位申請論文

概念体系に基づく視点情報を用いた文書整理支援システムに関する研究

指導教官 石塚 満 教授

東京大学大学院 工学系研究科
電子情報工学専攻 博士課程
67119 高間 康史

目 次

図目次

表 目 次

内容梗概

インターネットに代表される情報環境の急速な整備・拡大により，研究や仕事などに必要となる情報を収集する過程はますます容易になりつつある．その反面，入手可能な情報量が人間の情報処理能力を越え，かえって効率が低下するという，いわゆる「情報過多(information overflow)」が問題となってきている．すなわち，今までは情報不足が知的活動の足かせとなっていたのが，今後は収集した情報をいかに生かしきるかが死活問題になりつつある．特に，従来我々が最も親しみ，利用して来た文書情報を活用する方法論・支援システムの研究・開発は緊急課題であるといえよう．

本研究の目的は，大量文書情報をユーザが整理・活用する過程を支援するシステムの開発であり，そのための基盤技術として，ユーザの視点を動的に反映した文書の特徴づけおよび関連性の計算手法である**Fish Eye マッチング**を提案した．Fish Eye マッチングは電子化辞書の概念体系を利用して概念単位の単語特徴の選択，縮退を行うものであり，対象に依存した知識をあらかじめ用意する必要がない．また，概念体系と視点を関連づけて抽出するため視点外化能力も備えており，ユーザが考えを整理する上で有効な刺激・情報を与える効果も期待できる．

ユーザによる文書の分類結果から視点を抽出するアルゴリズムについても提案し，これを用いて文書検索実験を行ったところ，視点を反映した文書検索能力および，視点外化能力の両面においてその有効性が確認された．

この Fish Eye マッチングを基盤技術として，個人向けの文書整理支援システム **Fish View** を開発した．このシステムはユーザの作成した図解から視点を抽出し，ユーザに提示するとともに，視点に基づく文書検索，図解への情報追加を行う事ができる．また，最終的に得られた図解を元に，文書ディレクトリを構築，HTML 化して出力する事が可能である．

このシステムを実際のユーザに使用して評価してもらったところ，今までにこの様なシステムの利用経験がないため，とまどうユーザも目立ったが，図解に基づく文書整理支援という形態については殆どのユーザが有効であると感じており，論文の整理や人物評価，新聞スクラップの整理など，日常の活動の様々な用途に実際に使用してみたいとの評価を受

けた.

本研究で提案した文書整理支援システムの機能を高め，洗練させる事により，日常生活においてシームレスな支援を実現する事ができれば，今後ますます発展するであろう情報環境において，我々が知的創造活動を行う上で欠かせない存在となるであろう.

Chapter 1

序論

1.1 はじめに

インターネットに代表される情報環境の急速な整備・拡大により，研究や仕事などに必要となる情報を収集する過程はますます容易になりつつある．その反面，入手可能な情報量が人間の情報処理能力を越え，かえって効率が低下するという，いわゆる「情報過多(information overflow)」が問題となってきている．すなわち，今までは情報不足が知的活動の足かせとなっていたのが，今後は収集した情報をいかに生かしきるかが死活問題になるといえよう．

データベースからの知識発見はデータマイニング，数値データは統計的処理，といったように，大量データを扱う方法論についてはいろいろ研究されているが，我々にとって最も身近な情報源である文書情報が大量に入手された場合，これを活用する方法論については十分な研究がなされているとはいえないのが現状である．

本研究の目的は，大量文書情報をユーザが整理・活用する過程を支援するシステムの開発であり，そのための基盤技術として，ユーザの視点を動的に反映した文書の特徴づけおよび関連性の計算手法を提案した．この手法は **Fish Eye マッチング**と呼ばれ，電子化辞書の概念体系を利用して概念単位の単語特徴の選択，縮退を行うものであり，対象に依存した知識をあらかじめ用意する必要がない．また，視点を反映したマッチングを行えるだけでなく，視点を外化することが可能となるため，ユーザが考えを整理する上で有効な刺激・情報を与える効果も期待できる．

また，Fish Eye マッチングを基盤技術とした個人向けの文書整理支援システム **Fish View** を開発した．このシステムはユーザの作成した図解から視点を抽出し，ユーザに提示するとともに，視点に基づく文書検索，図解への情報追加を行う事ができる．また，最終的に

得られた図解を元に，文書ディレクトリを構築，HTML 化して出力する事が可能である．

1.2 本論文の構成・表記

本論文は 7 章からなり，その構成は以下の通りである．また本論文では，図・表・式は章を単位として通し番号を付す．

1 章：序論

2 章：WWW 上の情報検索・活用に関する従来の研究
情報検索・活用に関する全般的な概要および，近年その発展が著しい WWW 空間上での情報検索・活用に関する研究について紹介する．

3 章：発想法・発想支援に関する従来の研究
発想を創造的問題解決プロセスとしてとらえる事により，従来錬金術的なイメージで見られがちであった発想法を工学的に説明すると共に，これまでに提案・利用されて来た発想法および，計算機を利用した発想支援システムに関する研究について紹介する．

4 章：テキストマイニングの新展開〜ナレッジマネジメント
企業の経営効率化の切札として最近注目を浴びつつあるテキストマイニングについて，その要素技術を概略しながら，本研究で想定する文書整理プロセスについて説明する．

5 章：Fish Eye マッチング： 概念体系に基づく視点を考慮した文書マッチング機構
本研究で提案する，視点情報を活用した文書マッチング機構 **Fish Eye マッチング**の定義，視点抽出アルゴリズムの提案および形態素解析を用いた日本語文書への適用について述べるともに，文書検索に関する評価実験結果を示し，考察する．

6 章：文書整理支援システム Fish View
提案する文書整理支援システム **Fish View** の機能や，これを用いた文書整理プロセスの概要について紹介するとともに，実際のユーザによる評価実験について示し，考察する．

7 章：結論

Chapter 2

WWW 上の情報検索・活用に関する従来の研究

2.1 情報検索・活用とは

現代は情報中心の社会であると言われる．これは，計算機や通信技術の進歩・普及，インターネットなどの情報インフラの整備が進んだ結果，社会に流通する情報量，伝達速度が飛躍的に増大・向上したことに代表される．しかしその本質は，情報のデジタル化にあると言えるだろう．

すなわち，情報と言うものは本来，テキストや音声，画像，映像など，さまざまなメディア[1]，モダリティで表現，伝達される．また，メディアを記録する媒体にも様々なものが存在し，異なるメディアは異なる媒体に記録することが通常であった．

ところがデジタル技術の登場により，ほとんど全ての情報はメディアによらず，0あるいは1のビットの集まりとして表現，記録，伝達することが可能となった．これにより，情報の記憶・伝達だけでなく，情報の再利用・管理までもが格段と容易になり，これが高度情報化社会の実現につながったのである [81].

情報のデジタル化に伴う情報化社会の到来は，利益だけをもたらしたわけではない．確かに，個人レベルではテレビや雑誌，インターネットなどによって各自の嗜好，知的好奇心を満足するだけの情報を得ることは容易になった．企業においても，顧客や製品に関する情報を保持，分析することによって顧客のニーズにあった商品，サービスを提供したり，社内情報を電子的に管理することにより経営コストを削減することが可能となった．また，我々研究者においても，論文作成や実験データ分析，シミュレーションが容易になったり，

[1]「メディア」は記録媒体を指す場合もあるが，ここでは「情報の物理的な表現形式」の意味で用いる．

関連する論文が大量に入手できるようになるなど，情報環境から受ける恩恵は計り知れないものがある．

しかし，人間が処理できる情報量には限界があるため，情報を入手すればするほど活用できると言うわけではない．反対に，情報処理能力の限界を越える情報を扱おうとして混乱し，かえって効率，質が低下する可能性もある．インターネットや大容量記憶装置の普及した現在では，この様な「情報過多」が実際に問題になりつつある．すなわち,「いかにして情報を入手するか」から,「いかにして入手した情報を活用するか」に，我々の活動の焦点が移りつつあると言えよう．人工知能の文脈から見れば，情報不足による判断，認識等の誤りは広義のフレーム問題であるのに対し，情報過多は情報に対する処理の不足であるととらえる事ができる [32]．従って，計算機による情報収集・活用過程の支援，処理の自動化は，情報過多問題を解決する上で非常に有効と考えられる．

一般に，情報検索システムは図2.1の様に示すことができる [79]．情報発生源としては実験データ，プログラムリスト，製品・顧客情報や，最近ではインターネット上のWWW(World Wide Web) なども巨大な情報発生源であると捉えられる．情報検索システムでは目的に応じて情報を収集し，索引付け（インデキシング）などの必要な処理を行って，情報利用者の検索要求に対し必要な情報を利用しやすい形態で提供できる様にする．システム管理者は検索システム運用上の責任者であり，情報が的確に収集され，正しく蓄積・保守されているか，あるいは正当な情報利用者の利益が保護されるように，セキュリティにも気を配る必要がある．個人レベルではシステム管理者と情報利用者は同一である場合も多いが，扱うデータ／情報が大量になるほどシステム管理者の役割は大きくなり，専用のスタッフが必要となる．

中原らは，情報検索システムは情報利用者による要求パターンの多様さと加工処理の複雑さによって以下の4カテゴリに分類されるとしている [79]．

Business Use タイプ： 要求パターンは決められており，固定的な加工処理で十分なシステム．

例． 銀行システム，生産管理システム，顧客情報データベースなどの一般事務処理

Engineering Use タイプ： 複雑かつ多種の加工処理が要求されるが，要求パターンは決まっているシステム．

例． CADデータベース，画像データベースなどの科学技術システム．

文献検索タイプ： 様々な検索要求が発生するが，加工処理は簡単なシステム

例． WWWサーチエンジン，図書館システム，特許情報管理システムなどの文献情報システム．

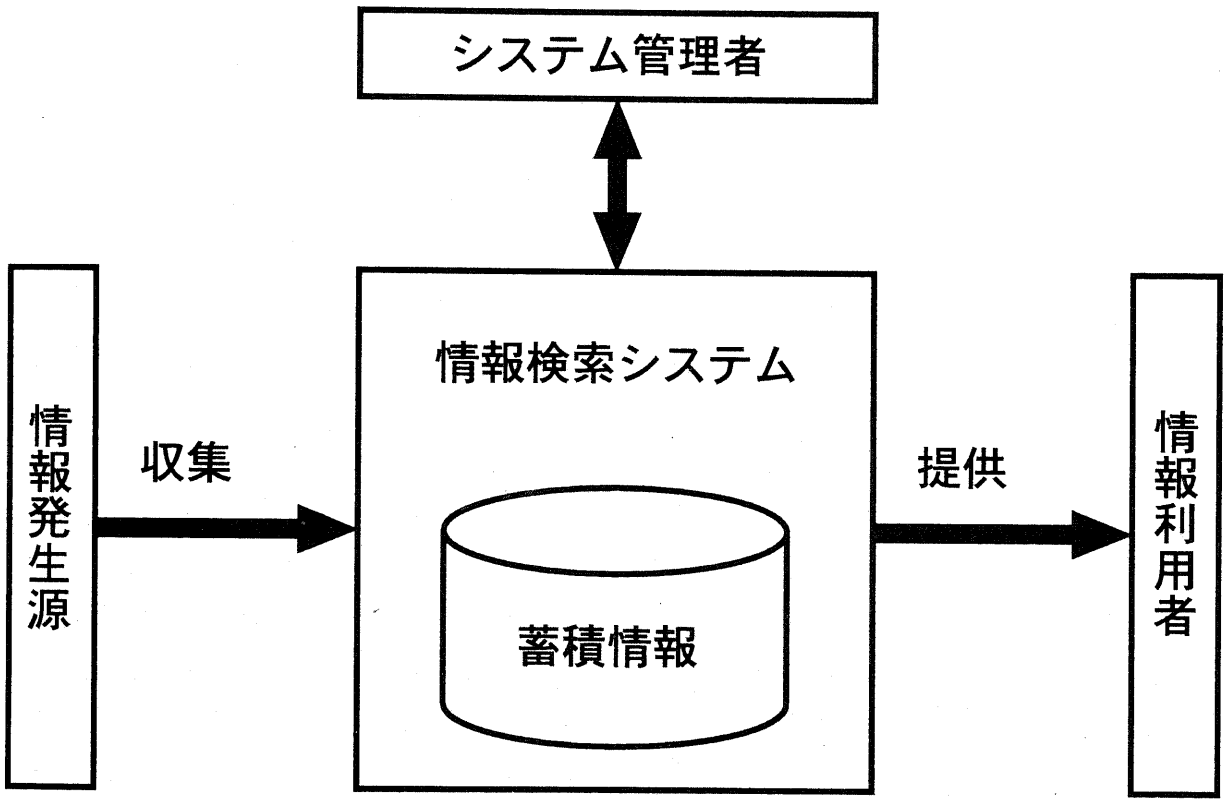


図 2.1: 情報検索システムの構成

Management Scienceタイプ：要求パターン，加工処理ともに複雑なシステム
例．経営意志決定システムなど

加工処理の複雑さに関連して，情報検索におけるデータの蓄積形式も重要な要素である．すなわち，Business Useタイプで主に扱われるデータは，項目値の集合であり，階層型データベースあるいは表形式のリレーショナルデータベースで管理することができる．例えば顧客データベースであれば，個人に関するデータは，「名前（性，名）」，「年齢」，「住所」，「電話番号」，「職業」，「購入商品」，「購入数量」，「購入日」などの項目から構成されるであろう．この場合，各項目の値が持つ意味はシステムにとって明確であり，データの蓄積および，ユーザの検索要求に合わせた加工処理は比較的容易であるといえよう．さらに最近では一歩進んだ活用支援として，顧客の購入履歴データなどから，今後の購入傾向などに関する予測情報を相関ルールの形で抽出するなどの，データマイニングに関する研究が盛んに行われている．これも，蓄積されたデータ形式がシステムにとって扱いやすい（可読である）ことが大きな要因となっている．

一般に，データが数値として扱える場合には，統計的手法による分析が有効であり，ユーザに有用な情報を提供することができる．また，数値で表現できない場合であっても，さきほどの顧客データベースの例のように表形式で扱えるものであれば，検索演算，データマイニングなどの情報検索に伴う加工処理を行うことは容易である．この意味で，これらは計算機の側に近いデータの表現形式であると言えよう．これに対し，Engineering Useタイプや文献検索で扱われるデータには，テキストや画像など，我々人間にとって親しみやすいメディアが多い．この場合，人間にとってはその画像，文書の持つ意味を認識・把握することは比較的容易であるが，計算機にとってはそうではない．すなわち，画像，文書にただIDを与えて記録しただけでは，個々のデータの持つ意味，特徴と言うものを理解することができず，ユーザの検索要求に適切に応えたり，データ間の関係を把握することができない．

Engineering Useでは一般に，ユーザの利用形態および検索要求が比較的パターン化されている場合が多いとされている．この場合は，詳細な要求分析を事前に行い，その結果に基づいてデータの加工処理および蓄積形式を設計する事によって対処することができる．反対にテキスト情報を対象とした文献探索システムでは，システムによる文書の特徴付けはキーワードなど，比較的表層的なものにとどめ，情報の詳細な吟味，分析はユーザ自身の手に委ねる形をとる．文書の要約など，より詳細な加工処理を行う必要がある場合には，対象文書のタイプ，分野を限定し，特別な知識を用意して加工処理を行う事が多い[71]．この場合はEngineering Useと同様であるとみなせるだろう．

特にテキスト情報は我々にとって親和性が高く，歴史も古いメディアであるため，これ

らを用いて表現される情報は多岐に渡り，専門知識を利用する方法は本質的な解決とならない．しかし大量のテキスト情報を活用するには，現在のようなユーザに処理の大部分を委ねる形態では不適切であろう．したがって，情報活用に向けて計算機が果たす役割はますます重要になりつつあり，その研究・開発が急がれているのが現状である．

本章ではテキスト情報の収集・活用に関する研究事例として，近年爆発的に普及しているインターネットの WWW 上の情報検索についてとりあげる．創造的問題解決のために利用される発想法・発想支援，社内文書データの活用を目指すナレッジマネジメントについては続く 3，4 章で概観する．

2.2 WWW 上の情報収集・活用とその支援

インターネットのルーツをたどると，1969 年にアメリカでスタートした初期の ARPAnet までさかのぼるとされる [77]．当初は，防衛上の理由から，設備の集中を避け，リスクを分散させる事を目的としてネットワーク化が試みられ，1983 年に始めて現在のインターネットと同様の形態，すなわち「個々に独立した組織のネットワーク化による統合」という形態が取られる事になる．

誕生当初は軍関係あるいは研究機関のみが利用していたインターネットも，現在では様々な企業，個人が参加し，ありとあらゆる情報が机の前にいながらにして即座に，しかも大量に手に入るようになって来ている．また，利用形態も多様化しており，メールやオンラインショッピング，ライブ中継，ニュース配信，はてはサイバースペースと呼ばれる仮想社会まで出現するに到っている．新しい商売を目指す企業，自分をアピールする場を得た個人など，インターネットに参加する人々は日毎に続々と増えており，日本国内におけるインターネット利用者は 1998 年 10 月現在で 1150 万人にものぼる [31]．

インターネットの普及に最も貢献したのは，WWW(World Wide Web) および Netscape や Internet Explorer に代表されるブラウザの存在であろう．WWW は音声，画像，テキストなど様々なメディアに対し統一的にアクセスする手段を提供しており，ユーザはハイパーテキストで表現された情報源を渡り歩く（ブラウジングする）事ができる．

WWW を異種情報の混合した巨大なデータベースとして捉えた場合，以下の様な問題点が一般に指摘されている．

- 希望する情報が本当に存在するかどうかわからない
- 例え存在したとしても，どこに，どの様に構造化されて存在するかわからない
- 情報の更新が頻繁に起きるため，以前は存在した情報が移動したり，消滅したりする．

従って，Engineering Use タイプにおける詳細な要求分析を事前に行うことは不可能に近く，システムは関連のありそうな情報を，あまり加工せず余分めに提供し，有効／無効の判断および加工処理はユーザ自身の手に委ねざるを得ない．それでも，単なるブラウジングよりも効率的に所望の情報へ到達できる様，以下の様な情報流通・検索技術が提供されている．

- 要求駆動
  - 情報検索
    - サーチエンジン（goo[28]，infoseek[42] など）
    - ディレクトリサービス（yahoo[131] など）
  - ブラウジング支援技術 [1, 4, 48, 63, 93, 95]
- データ駆動
  - 情報フィルタリング（information filtering）[29, 53, 67, 68, 74]
  - プッシュ技術 [59]
  - 情報の可視化 [26, 34, 35, 65, 100, 104, 105, 125]

要求駆動とは，ある情報に関する要求が発生した時点でユーザが行う情報収集形態を指す．このうち情報検索は，利用者が自分の欲しいものについてはっきりわかっている場合に有効な形態であり，サーチエンジンやディレクトリサービスなどを利用して希望するページを直接入手しようとする場合などがこれにあたる．サーチエンジンは，ユーザがキーワードを入力すると，そのキーワード（場合によっては関連語）を含むホームページを検索して返すものであり，goo[28] や infoseek[42] などがその代表的なものである．ディレクトリサービスは yahoo[131] に代表されるように，ユーザはあらかじめ分類され，階層的に整理されたホームページ群から所望の情報を探し出す．一般に，サーチエンジンでは WWW ロボットなどを用いてホームページを自動的に収集するため大量のホームページを扱うことができる反面，ページ内容の吟味，選択作業の大部分はユーザに委ねられることとなる．これに対しカテゴリサービスでは，データベース（カテゴリ）への登録は手作業で行われるのが一般的であるため，所望のホームページを捜し出すのは比較的容易となるが，データベースの構築に手間がかかるため，情報更新が遅れると言った欠点が存在する．サーチエンジン，ディレクトリサービスを用いた情報検索は，WWW の最も一般的な利用形態と言っても良く，インターネットへのポータル（入口）サイトとして，ビジネス業界の注目を浴びている．

これに対しブラウジングはサーフィンとも呼ばれ[2]，文字通りWWWブラウザなどを用いて，Webページ内のリンクを辿っていく形態がこれにあたる．この場合，ユーザは自分の欲しい情報について明確に理解している必要はなく，反対に，ブラウジングの過程を通じて（リンクによる接続関係などから），対象に関する概念を獲得していくことが期待できる．また，目的以外の思いがけない情報を発見する事があるのもブラウジングの利点である．反面，欲しいものがはっきりしている場合には探索効率が悪いという欠点もある．

上記の様な能動的情報収集形態に対し，メールやニュースなど，大量に到着する情報の中から自分にとって適切な，興味のある情報を取り出したいという欲求も存在する．この様な形態の情報収集を行うのが情報フィルタリングであり，データ駆動かつ受動的な形態であると言える．技術的には，情報検索と同様の手法を適用する事によって情報フィルタリングを実現できる [3]

プッシュ技術とは，ユーザが操作しなくてもサーバ側から自動的に情報をクライアントに送る技術である [59]．プッシュ技術により提供される情報は，テレビに例えて「チャンネル」と呼ばれている．現在の利用形態としては，米 PointCast 社の PointCastNetwork に代表されるニュース配信や，ソフトウェアのアップデート，ドキュメントや業務データの配布などがあげられる．

情報可視化とは，WWW 空間におけるページ間のリンク関係などを図解などを用いて表示することにより，大量情報が複雑に絡み合った WWW 空間内で，ユーザが文字通り「迷子」にならないようにする技術である．

要求駆動で扱われるユーザの嗜好は一過性のものであると言えるが，フィルタリングシステムなどのデータ駆動においては要求（嗜好）を固定化し，比較的長期間，継続して使用されるのが普通である．

以下では，ブラウジング支援，情報フィルタリング，サーチエンジンの利用，情報可視化技術，ベクトル空間モデルに関する研究事例を取り上げ，もう少し具体的に見ていく事にする．

2.3 ブラウジングにおける情報活用支援

前述の様にブラウジングは，あらかじめ欲しい情報が明確に定まっている場合の検索活動ではなく，ハイパーリンクを辿りつつ，徐々に欲しい情報を明確にしながら近付いていくプロセスであるといえる．場合によっては，検索要求をあらかじめ持つ事なく，興味の赴くままにリンクを辿る事によって新たな興味・視点を得る事もある．これは，我々の日

[2]情報検索と対比して，「情報散策」と訳されることもある．

常生活にとってめずらしいことではなく，デパートや大学キャンバスなどの案内システムとの対話進行にも同様の傾向が見られる [120].

WWWは，世界中のホームページがハイパーリンクでつながっており，それらを辿る事によって，膨大な情報空間をブラウジングする事が可能である．しかしリンクの向こう側に何があるかは辿ってみなくてはわからず，また全体的な構造空間を把握する事ができないため，リンクを辿っている内に「迷い子」になりやすい事も指摘されている．ブラウジング支援とは，ユーザの興味にあったページや人気のあるページなどへ素早くたどり着けるように，ユーザにどのリンクを辿るべきかについての指標を与える事である．以下ではブラウジング支援を行う代表的なシステムとして，WebWatcher[4] について詳しく取りあげる．

WebWatcher[4] は，ユーザがある目標を持ってブラウジングしている時に，各ページ内に含まれるリンクの内，それを辿ることによってユーザの目標達成に近付くことのできると思われるリンクの推薦（recommendation）を行うシステムである．WebWatcherは，サーバへのリンクを辿ることによって起動される．その後，表示されるフォームに探索目標を入力すると，それ以降のブラウジング過程において表示されるページは，WebWatcherによって加工され，推薦するリンクにはそれを示すマーク（図2.2中の「目」）が付加される．また，関係ありそうなページへのリンクをページの上部に付加する機能も持っている．この様に，ユーザに提示するページに必要な改変を施す方法を採用しているため，ブラウザには依存しない．

WebWatcherの学習アルゴリズムは次の通りである．まず，学習すべきものは，

$$UserChoice? : Page \times Goal \times Link \rightarrow [0,1]$$

としている．すなわち，同一の目的を持ったユーザがそのリンクを選択する傾向の予測を表している．この様に，個人の識別を行わないことにより，システムを利用する全世界のユーザの利用結果から十分な訓練例を得ることができ，十分な学習が行える事が期待できる．

同様に，ブラウジング過程におけるユーザ支援を行いつつ，学習を行うシステムには他に，Letizia [63] や Syskill & Webert [1, 95] （図 2.3），LAW[29] などがある．LetiziaはWebWatcherと異なり，特定のサーバを必要とせず，完全なクライアントとして動作する．また，Syskill & Webertは，ユーザごと，トピックごとの学習を行う．システムはユーザに対し，現在見ているページに対する評価を４段階で要求する．すでに見たページへのリンクにはその評価に対応するマークが付加され，まだユーザが辿っていないページへのリンクには，ユーザがそのページを好むかどうかの予測値が付加される．

また，Naviplan[93] では，あるページを読んだときに得られる概念に対応する単語を効果語，そのページを理解するために知っていなければならない概念に関する単語を条件語

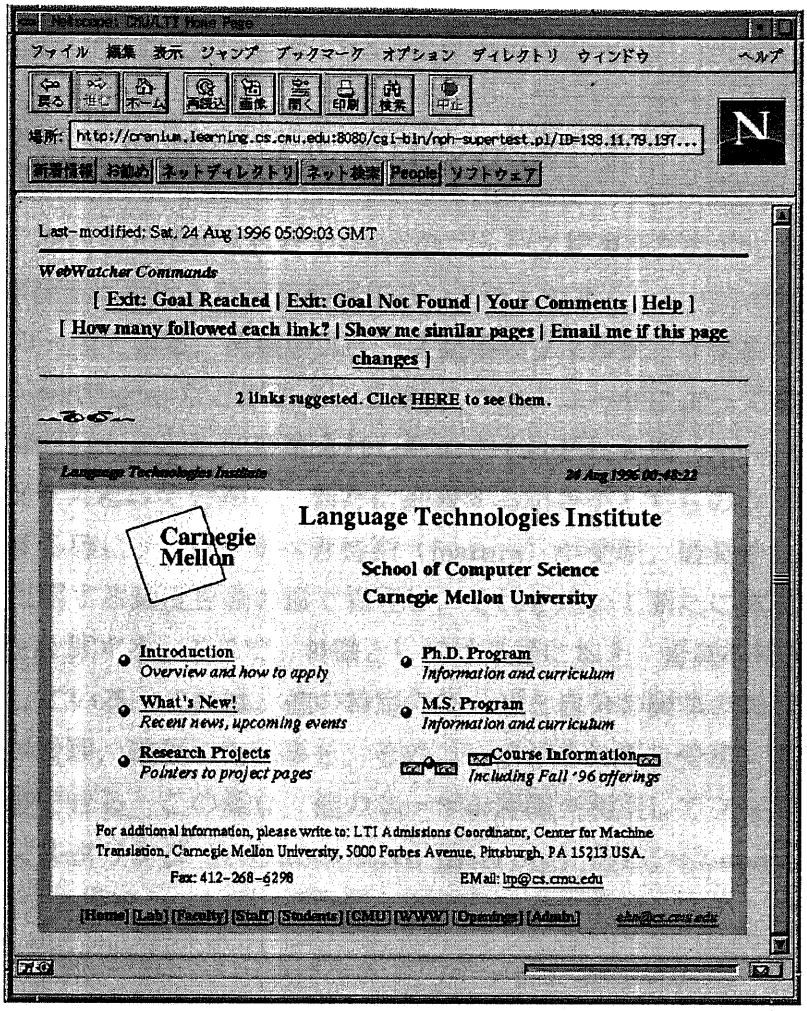


図 2.2: WebWatcher 利用時の画面例

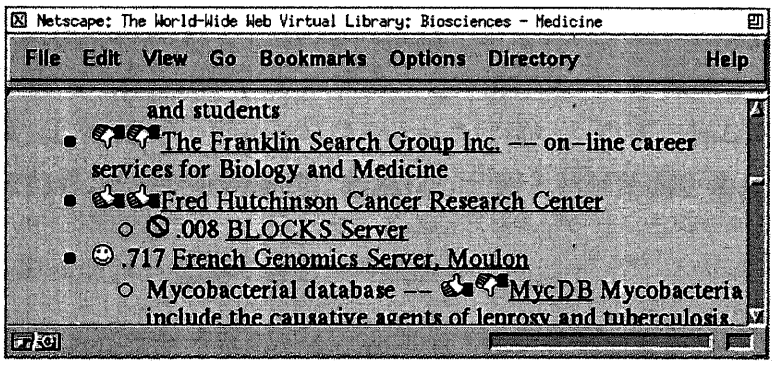


図 2.3: Syskill & Webert の画面例

として，目標概念を理解するために読むべきページ系列をプランニングにより求める手法を提案している．

2.4 INFOS：情報フィルタリング

INFOS[74]は，不適切と予測されたUsenetのニュース記事を自動的に削除することにより，ユーザの記事探索における負担を軽減するシステムであり，上述の情報フィルタリングを行うシステムの一つである．同様のニュース記事のフィルタリングを行うシステムとしてはIAN[29]，NewsWeederなどがある．INFOSでは，ユーザによって興味あると判断された記事と，興味ないものとして破棄されたものをそれぞれ受理（Accepted）クラス，破棄（Rejected）クラスとして分類し，新たに到着する記事がどちらのクラスに属するかを予測する．判断する際にチェックすべき特徴（feature）が受理，破棄それぞれのクラスに属する記事内に出現する頻度を表の形で管理し，それを用いて新たに到着した記事と各クラスとの類似率を計算する．ここで，特徴としては単語に加え，協調的特徴（collaborative feature）を採用している．これは，他の特定のユーザと自分の興味の類似性を利用するものであり，自分が受理，破棄した記事を，そのユーザがそれぞれ受理，破棄する頻度を予測に利用する事ができる．この様な，他のユーザの評価を利用してフィルタリングを行うことを**協調的フィルタリング（collaborative filtering）**あるいは**social filtering**と呼び，WWWページの推薦[101]だけでなく，音楽や映画の推薦システムなどでの利用が期待されている．

情報フィルタリングを行って，ユーザにとって興味があると予測された記事を集めることにより，特定ユーザ向けの新聞が作成できる[7, 29]．例えばAARON[29]は，ニュース関連のドキュメント（記事）を供給する指定サイト群からHTMLドキュメントを集め，これらからユーザが興味を持つであろうドキュメントを選択し，一連のローカルなHTMLページとして個人向け新聞を構成する．集められた各ドキュメント間の類似性を求め，一般に5-20ドキュメント程度のクラスタに分類する．その後，ユーザが興味を持たないと予測されるクラスタを除去した後で，ヒューリスティックに従い，以下の様な三つのセクション見出しのもとにHTMLページを階層的に組織化し，個人向け新聞を構成する．

headings 長いドキュメントを多く含むクラスタ

interest headingsとみなされなかったもののうち，ユーザが興味を持つと予測されるクラスタ

gazette 残ったクラスタから少数をランダムに選択したもの

また，Balabanovic[6] らは，毎日ユーザにとって興味深いと予測した記事を 5 ページ表示し，その記事に対するユーザの評価をフィードバックとして受け取って学習し，その結果を元に次の日にまた新しく提案する，というサイクルを繰り返す実験を行っている．一般的な形態とは異なるが，これも情報フィルタリングの一形態であり，ユーザの嗜好の変化に追随することを重視したアプローチであると言えよう．

この他，情報フィルタリングを電子メールなどに対して行う知的エージェントも研究されている．この場合特に問題となるのが，エージェントに対する信頼性の問題である．すなわち，エージェントの予測が外れ，ユーザにとって重要なメール（記事）が破棄されてしまうと言った問題点である．この様な，知的エージェントにおける自律性と信頼性の間のトレードオフに関して，Maes らは予測結果の確信度に基づくエージェントの行動の選択（提案のみか，実行するか）を行うことを提案しており [67, 68]．この考えは MAGI[29] にも採用されている．

2.5 サーチエンジンのクエリー拡張：AIアプローチの応用

サーチエンジンは，ある目標を持ってブラウジングを行う際に，適切なブラウジング開始点を選択する上でも重要な要素である．サーチエンジンを利用した検索に関する研究は，大別すると次の二つに分けられる．

1. サーチエンジンの改良
2. 適切なクエリーの生成

サーチエンジンの改良とは主に，クエリーとのマッチングに関して柔軟性を与えることにその目的があると言える．すなわち，既存のサーチエンジンは，入力されたキーワード列の全て（AND）あるいは少なくとも一つ（OR）をインデックスに持つページをデータベースから検索し，その全てを返すものがほとんどである．この様なマッチングを **boolean(exact) match** と呼ぶ．これに対し，質問中の全てのキーワードが用いられていない場合でも，その類似性を元に判断する方法を **partial(best) match** と呼ぶ．partial match の代表的手法としては，**ベクトル空間モデル**（Vector Space Model : VSM）がある [21, 78]．VSM では，各ページに対するインデックスおよびクエリーの両者を後述する特徴ベクトル（feature vector）として表現し，クエリーとページインデックス両者のベクトルの内積をとって類似度を判断する．この場合，再現率（Recall : 表 2.3）を自由に決定できるというメリットがある．すなわち，類似度をスコアとして各ページをソートすることにより，ユーザに提示するページ数を上位 n ページに制限する，と言ったことが容易に行える．VSM につい

ては 2.7 節で詳しく触れる．他には，**shingle** と呼ぶ連続した単語列を特徴として，集合論的に文書間の類似性を求める手法が Border らにより提案されている [9]．この手法は文書間の構造的類似性を捉えるのに有効であり，修正されたページの検出などに応用できるとしている．

一方，後者は既存のサーチエンジンの有効利用を考えたものであり，データベース構築などのコストが不要であるという利点がある．一般に，サーチエンジンへ送るクエリー（Query）を適切に生成することは難しい問題である．これは，語と概念間の写像関係が一対一でないことが主な原因であると考えられ，自分の希望したものと事なる検索結果が得られることもしばしばである．この様な問題に対し，ユーザの要求（興味のある，なしなど）に適合するページ集合を例として，その集合に適合するキーワード列を学習し，クエリーとして用いるといったアプローチ [15, 16, 95] は，一つの解決策を提案していると言える．この様なクエリー学習は，パターン認識や概念学習問題として，AI の領域などで従来から研究されてきた成果が利用できると考えられ，多くの AI 研究者が WWW 上の情報検索に取り組むようになった．

インターネット上に構築される情報環境は，トイワールドから実世界への飛躍を目指す AI 研究者にとってもチャレンジしがいのある世界である．これは，形式的には（電子化という）計算機にとって扱いやすいが，内容的には人間を対象とした情報，知識であるという WWW 上の情報源の性質が，AI が次に取り組むべき世界としてふさわしい位置にあるという点と，人間（ユーザ）と計算機が協力して初めて有効に活用できる世界であるという二点がその主な理由であろう．

情報収集における AI 的手法の利用に関しては，現在のところ以下の二つに大別できるようである．

- ドキュメント分類問題を，概念学習問題としてとらえるアプローチ
- 語に関する知識を利用しようとするアプローチ

前者は，ドキュメントがあるクラスに属するかどうかを判断する問題は，そのクラスの概念を学習することと等価であるとみなすものである．この場合，ニューラルネットや帰納学習など，AI 分野で長年研究されてきた成果を利用することができる．

これに対し後者は，語の表層だけではなく，それの持つ意味を積極的に利用しようと言うものである．

概念獲得に用いられる手法は，学習の結果得られる出力のタイプによって，以下の三つに分けられる．

ベクトル・ベース ある概念（カテゴリ）を表す特徴ベクトルの各要素の値（重み）を学習する方法

ルール・ベース ドキュメントに含まれるキーワード列を条件部，カテゴリを結論部とするルールを学習する方法

インスタンス（ケース）・ベース 未知のドキュメントのカテゴリを，既存のドキュメント（インスタンスあるいはケースと呼ぶ）との類似度に基づいて決定する方法

従来の適合フィードバック（relevance feedback）で用いられている各種の重み更新アルゴリズムやニューラル・ネットワークなどがベクトル・ベースにあたる．また，Bayesian classifier[96] の様な確率モデルも，各属性の事前確率を重みと考えればこの範疇に属すると言える．

ルール・ベースで得られるルールの条件部であるキーワード列は，そのまま既存のサーチエンジンのクエリーとして使用する事が容易であるという利点がある [16]．この様なルールを学習する代表的なパラダイムとして帰納学習（インダクション）があり，代表的なシステムに ID3，C4.5[99]，CN2，FOIL などがある．

また，インスタンス・ベースの代表的な手法には IBPL[29] や Nearest Neighbor[18] などがある．

Pazzani[95] らは，いくつかの学習アルゴリズムの性能比較を行っており，予測正当率（Predictive Accuracy）に関して表 2.1 の様な結果が得られている．ここで予測正当率とは，正しく分類されたテスト例の数を，テスト例の総数で割った値であり，機械学習の分野で一般に用いられている[3]．

この結果を見ると，帰納学習である ID3 の性能が明らかに他より劣っている．これは，ID3 の持つ，一般化へのバイアスが強いという性質が原因であると考えられる．これは FOIL などにおいても同様であり，負例を一つも含まないために必要十分な条件しか持たない，シンプルなルールを好む傾向にある [16]．Pazzani らの実験によれば，3–5 の語の有無に関する条件しか持たないルールが生成された事が報告されている [95]．しかし，Web ページやオンラインニュースをユーザの興味に従って分類するようなタスクにおいては，一般に正例（ユーザが好むドキュメント）は負例（それ以外のドキュメント）に比べて非常に少ないといわれている．Pazzani らによると，全てのドキュメントをユーザにとって興味ないものとして分類した結果，予測正当率が 57–74%に達している [95]．従って，一般化されたルールはそれだけカバーするドキュメント数が多くなってしまい，結果として予測正当率が下がると考えられる．この問題に対し，Cohen らは通常のプロセスの事後処理として，

[3]学習結果の一般的な評価尺度については表 2.3 にまとめて記す．

各ルールに,「そのルールによってカバーされる正例全てにおいて真（true）となる条件を元の条件部に追加」する事により，特殊化（specialization）へのバイアスを加えている [16].

2.5.1 関係学習によるドキュメント分類

上述した学習機構は全て，命題表現に基づく学習であるといえる．すなわち，ある属性（語）のドキュメント中における有無のみを扱ったものであり，ドキュメント中の語の出現順序などの特性は無視している．この様な関係情報は一階述語論理を用いれば表現可能であり，帰納論理プログラミングによって学習可能である．Cohen らは，FOIL6 を用いて，語順等の関係を利用したドキュメント分類を試みている [15]．FOIL6 は例と背景知識から関数を含まない Prolog の述語定義を学習するシステムであり，空の述語定義に節を一つずつ追加していくことにより定義を拡張していく．背景知識としては，次の様な関係を用意している．

- $w_i(d,p)$... 単語 w_i がドキュメント d 内の先頭から数えて p 番目に存在すれば真.
- $succ(p_1,p_2)$... $p_2 = p_1 + 1$ の時に真.
- $near1(p_1,p_2)$... $|p_1 - p_2| \leq 1$ の時に真.
- $near2(p_1,p_2)$... $|p_1 - p_2| \leq 2$ の時に真.
- $near3(p_1,p_2)$... $|p_1 - p_2| \leq 3$ の時に真.
- $after(p_1,p_2)$... $p_2 > p_1$ の時に真.

このほかの工夫として，VSM における特徴選択（2.7.2 節参照）に対応するものとして，関係選択（relation selection）を行い，出現頻度の少ない単語についての関係を除去している．また，多くのサーチエンジンでは否定（単語の不在）を扱っていないことを考慮し，否定リテラルを扱わない事にしている．これらはともに，探索空間の削減にも効果があると考えられる．

このシステムの学習結果と命題学習機構である C4.5，Bayesian probabilistic classifier の学習結果を比較した結果を表 2.2 に示す [15]．これより，関係選択を行わない場合は他と相補的（他手法は相対的に高適合率，低再現率で，FOIL6 は高再現率，低適合率）で，関係選択を行うと適合率を改善することができる．また，命題版として FOIL6 を使用した場合[4]よりは，適合率，再現率ともに良くなっていることが分かる．

[4]背景知識として $w_i(d)$ のみを使用した場合.

表 2.1: 学習アルゴリズムの性能比較

アルゴリズム	平均正当率	標準偏差
naive Bayesian classifier	77.1	4.4
backpropagation	75.0	3.9
Nearest Neighbor	75.0	5.5
ID3	70.6	3.6

表 2.2: 命題学習アルゴリズムと述語学習アルゴリズムの性能比較

アルゴリズム	再現率	適合率
FOIL(関係選択なし; 命題版)	0.4652	0.3023
FOIL(関係選択なし; 述語版)	0.4698	0.3537
FOIL(関係選択あり; 命題版)	0.4899	0.5302
FOIL(関係選択あり; 述語版)	0.4973	0.5449
C4.5	0.3367	0.8023
Bayesian probablistic classifier	0.3732	0.6648

2.6 情報空間の視覚化

ブラウジング支援のところでも述べたように，WWW には膨大な量のホームページが存在し，相互にリンクしあって複雑な情報空間を空間を形成しているため，目的の情報へたどりつくのが困難なだけでなく，ブラウジング途中で迷子になったような感覚に陥る場合も多い．これは多くのユーザが経験していることだろう．これは，WWW 空間では距離の感覚があいまいであるため，実空間における空間認知能力が生かせない事も要因の一つであろう．従って空間・距離の概念をブラウジング支援に持ち込むのは自然な考えのように思われる．

この様な，各情報間の関係を空間（平面）的配置を用いて表現し，情報同士の大局的な位置関係と，局所的詳細な情報検索を両立することによってユーザのブラウジングにおける認知負担を軽減する研究は**情報の視覚化（Visualization）**あるいは可視化と呼ばれ，基本的にはディレクトリ構造など，階層構造，木構造で表現できるデータ集合を対象として様々な視覚化手法が提案され，それに基づいたシステムが開発されている [34, 35, 55, 100]. Information Cube[100] では，半透明の箱が入れ子状になった形で階層構造を表現する．また，Cat-a-Cone[34] では，MEDLINE の検索結果をコーンツリーを用いて階層的に視覚化している（図 2.4）.

視覚化の魅力は空間的配置だけではなく，ユーザ自身の手で，見やすいように空間の変形操作を行える，インタラクティブ性も大きな魅力の一つであると言える．PadPrints[35] では，ユーザのブラウジング履歴を視覚化し，ユーザは Zoom 機能を使って希望の縮尺で履歴を眺める事ができる（図 2.5）.

また，DocSpace[125] では，バネモデルにより，ユーザの操作に追従してインタラクティブにオブジェクトを配置する．ここで，キーワードの重みをバネの力として表現することにより，ユーザの視点を視覚化に反映する事ができる．納豆 View[104, 105] では，WWW 空間を納豆のメタファで捉える．すなわち，あるノード（URL）をドラッグし，図解上で持ち上げると，そのノードとつながったページもひきずられるようにして持ち上がる（図 2.6）．これにより，ユーザは自分が興味あるページ周辺をひろげて眺める事が可能となっている.

視覚化技術を適用する上で問題となるのが，ディスプレイの解像度と，画面描画に関する計算機の処理速度である．この面に関して，最近の計算機はパソコンレベルでも十分な処理能力と高精細なディスプレイを備えており，納豆 View や Cat-a-Cone に見られる様な 3 次元視覚化や，PadPrints や DocSpace などに見られるインタラクティブ性の実現も容易になっている.

上述したように，WWW 空間の視覚化はブラウジング支援として今後ますます期待さ

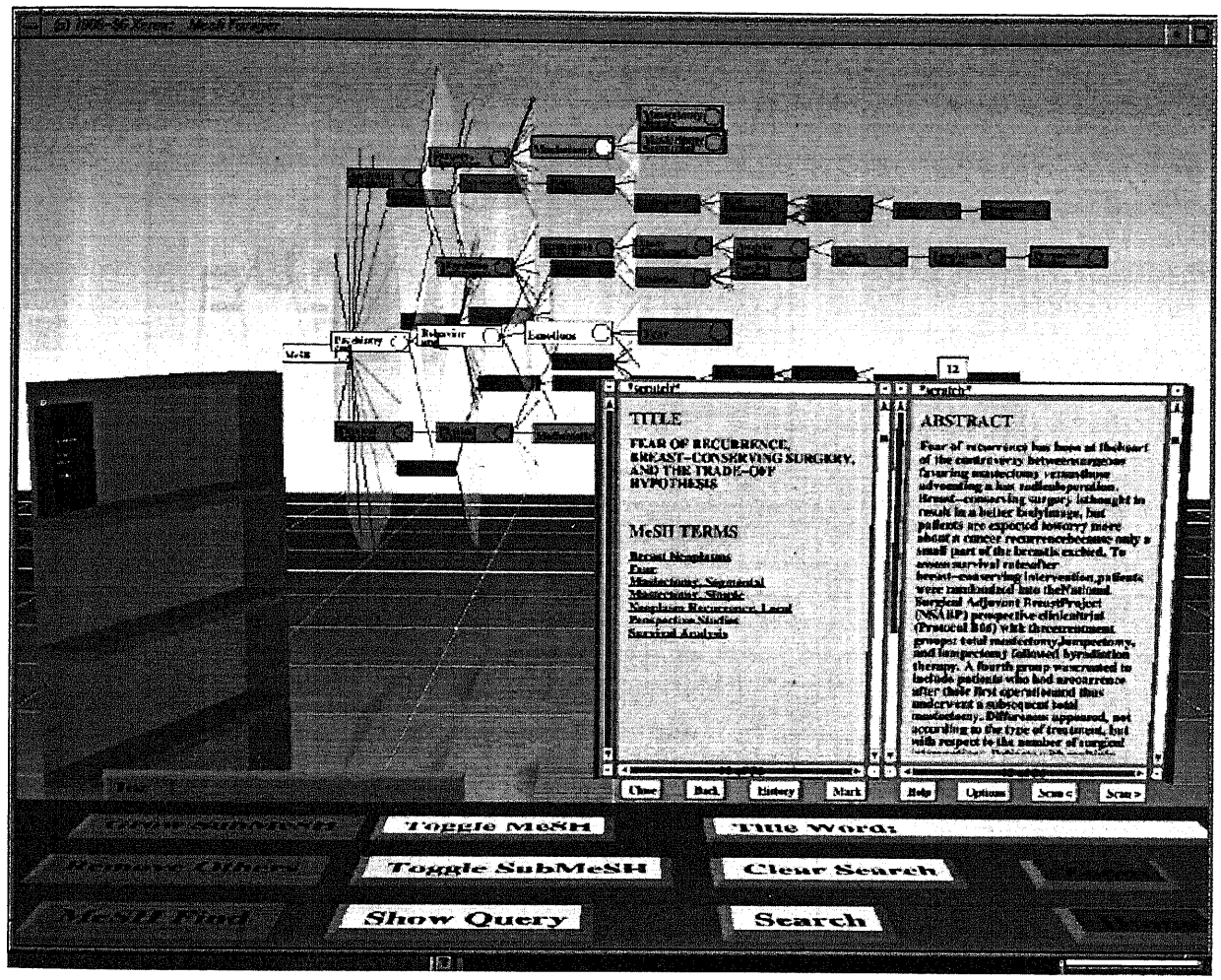


図 2.4: Cat-a-Cone の画面例

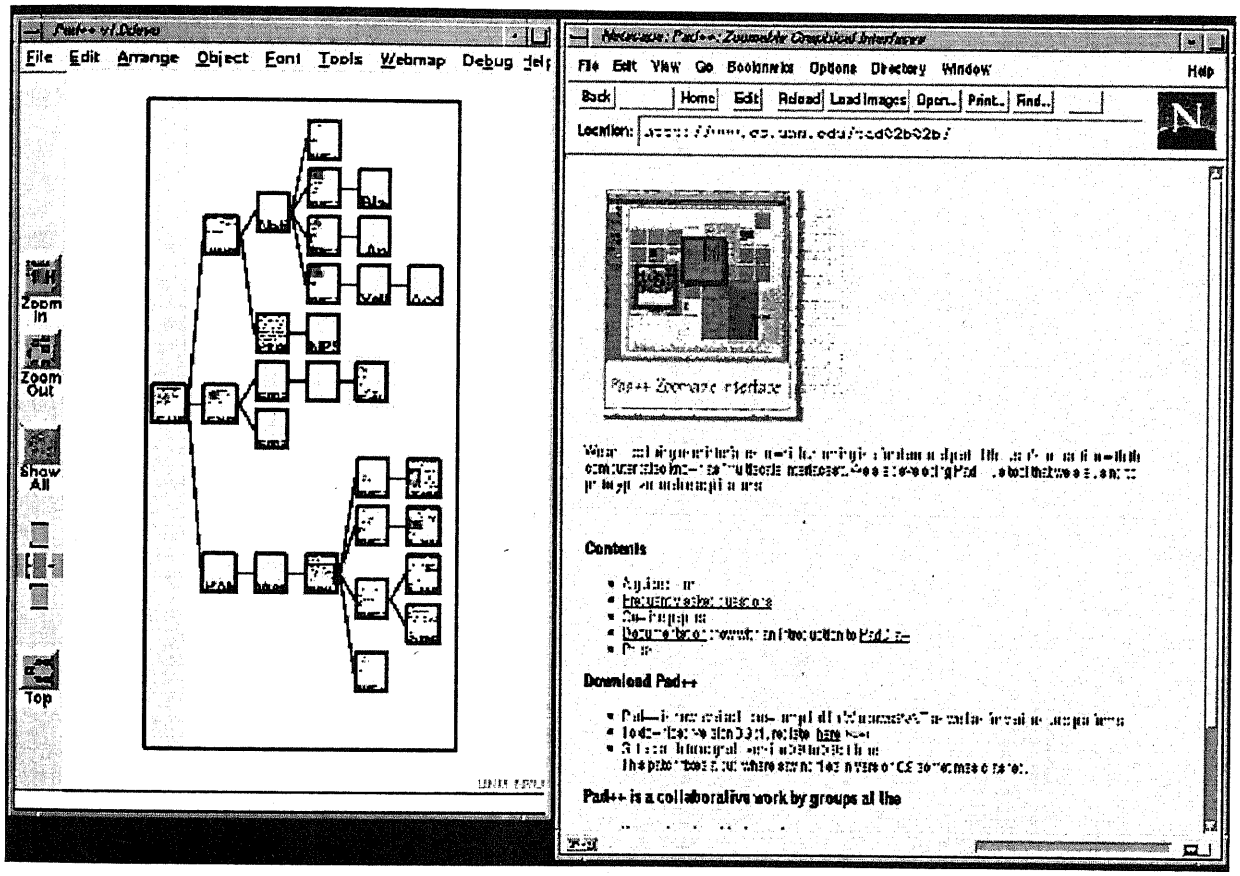


図 2.5: PadPrints の画面例

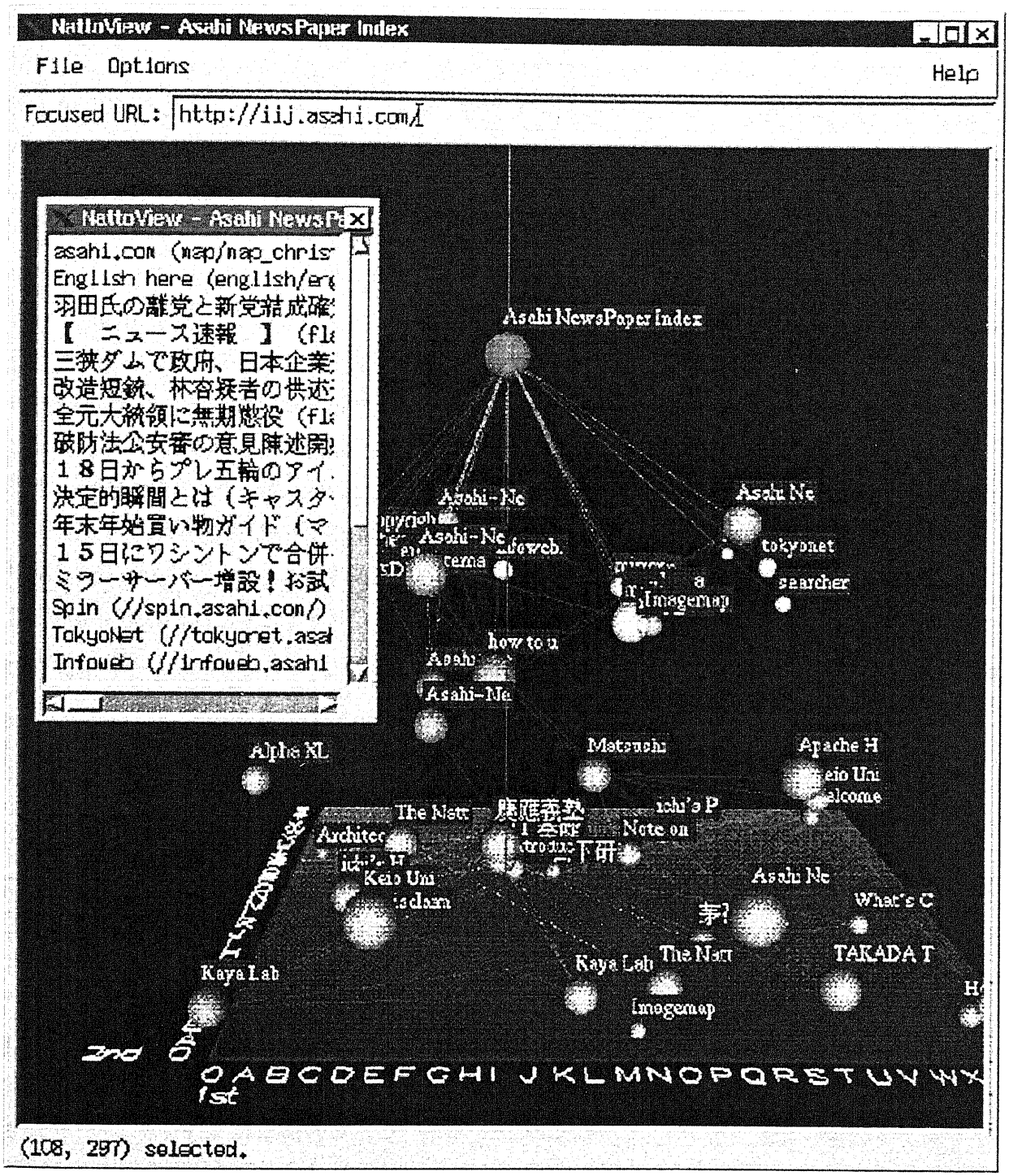


図 2.6: 納豆 View の画面例

れる技術であるが，現行のブラウザに置き換わるためにはいくつかの課題があると考えられる．情報視覚化研究が従来対象として来たディレクトリやデータベースと WWW 空間との違いは，WWW 空間に蓄積された情報量が膨大であり，かつその構造や蓄積情報が頻繁に変化することであろう．従って，情報源全体をあらかじめ構造化しておく事は実用上不可能であると考えられる．PadPrints ではユーザのブラウジング履歴を対象としているが，履歴は漸進的に現在の空間構造に追加していく事が容易に行えると考えられる．また，Cat-a-Cone の様にユーザの検索結果を視覚化する場合も，空間を構成する情報（文書）が一括して得られるため，これも視覚化に向いていると言えよう．

しかし，ブラウジング支援と言う面で視覚化を考えた場合，興味あるページへたどり着くために，**次に**どういう経路，ページを辿るべきかについての指標をユーザは望んでいるわけであり，これを視覚化する事は意外と困難であると考えられる．すなわち，ユーザが今後辿る経路，情報源はユーザのブラウジング行動によって動的に変化するため，可能性のある空間をあらかじめ視覚化できる形に構造化しておく事は不可能に近い．

この問題に対して，WebCutter[65] は有望な解決策を提案している．WebCutter は遺伝子データベースの分野などで利用されているツールであり，ユーザが興味あるキーワード，探索したいサイト，マップ構築にかけてよい時間，マップサイズなどを指定すると，ユーザの興味に限定した限定マップ（Restriction Map）を作成してくれるものである．ユーザの興味に限定する事により，対象とする情報量を削減できるため，視覚化された空間が見やすくなるだけなく，ブラウジング直前あるいは最中に先読み探索して構造化する事が可能になると考えられる．また，この手法は 2.4 節で紹介した情報フィルタリングとも相性がいいと考えられ，今後の研究が期待される分野であると言えよう [26]．

2.7 VSM：ベクトル空間モデル

ベクトル空間モデルでは，文書およびユーザのクエリー（検索質問）を索引語（単語）の集合として扱う．すなわち，各索引語を互いに線形独立なベクトルとしてベクトル空間を構成し，文書やクエリーをこの空間上のベクトルとして表現する．この，文書やクエリーを表すベクトルの事を**特徴ベクトル**（**feature vector**），ベクトル空間を構成する要素を**特徴**と呼ぶ．特徴は索引語だけとは限らず，協調的特徴（collaborative feature）を採用する場合もある [74] ．

索引語 $T_i(i = 1, \cdots, t)$ に対応する特徴を表すベクトルを V_i とすると，文書 r の特徴ベクトル D_r は次式で表される．

$$D_r = \sum_{i=1}^{t} a_i^r V_i \tag{2.1}$$

一般には，V_i として直交成分を用いるのが普通である．ここで，a_i^r は文書 r における索引語 T_i に対応する成分の値であり，以下のように様々な重み決定方法が存在する．

2値モデル T_i が文書中に存在すれば 1（正），しなければ 0(-1) とする．

TF(Term Frequency)[29] 文書 r における T_i の出現回数 tf_i^r を重みとする．

TFIDF(TF × Inverse Document Frequency)[7] T_i を含んだ文書数（df_i）の，全文書（n）中における割合の逆数を TF にかけることにより，希少性の高い特徴の重みを大きくする．

$$a_i^r = tf_i^r \cdot \log \frac{n}{df_i} \tag{2.2}$$

TR(Term Relevance) 文書がすでに分類されている場合に，あるカテゴリのみに頻繁に出現する特徴の重みを大きくする．

以上のようにして求められた特徴ベクトル D_i, D_j 間の類似性は，以下の式を用いて計算する．ここでは，各特徴を表すベクトル V_i は互いに直交するとしている．

$$sim(D_i \cdot D_j) = D_i \cdot D_j = \sum_{k=1}^{t} a_k^i a_k^j \tag{2.3}$$

$$sim(D_i \cdot D_j) = \cos(D_i, D_j) = \frac{D_i \cdot D_j}{|D_i||D_j|} \tag{2.4}$$

これより，式 (2.3) はベクトルの内積，式 (2.4) は余弦（cosine）の値を計算していることがわかる．一般に，文書が長い程含まれる単語数が増えるため，内積をとった場合には大きな値が得られる．これに対し，余弦は正規化後のベクトルについて内積を計算しているため，文書長に影響を受けにくいという特徴がある．

また，検索性能に関する評価指標について表 2.3 に記す．**適合率（precision）**と**再現率（recall）**を用いるのが一般的であるが，機械学習（Machine Learning; ML）の手法を用いた場合には学習性能の一般的な評価指標である**予測正当率（predictive accuracy）**が用いられる場合もある．

一般に，適合率と再現率はトレードオフにあると言われる．例えば，再現率を 1.0 にしたければ，検索結果として全文書を返せばよいが，この場合には適合度はかなり低くなってしまう．反対に，該当文書と判断する類似度の閾値を高くすればするほど適合率は高く

できると考えられるが，検索文書数が減少する程，再現率も低下してしまう．従って，実用的な検索性能について評価するためには，適合率，再現率の両者のバランスを考慮する必要がある．これに関して，両評価指標を考慮した式 (2.5) なども提案されている [15]．ここで，P は適合率，R は再現率である．式 (2.5) において，$\beta = 0$ とすれば適合率と同じになり，値を大きくする程再現率の影響が大きくなる．$\beta = 1$ の時，両指標を同じ比重で用いていることになる．

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2.5}$$

2.7.1 適合フィードバック

サーチエンジンに入力されるユーザの検索要求をクエリーという．キーワードを用いたboolean match の場合は，自分の興味を反映した適切なキーワードを選択する困難さは別として，クエリーの入力に関しては思いついたキーワードをタイプするだけなのでさほど難しくない．しかし，ベクトル空間モデルにおいては，ユーザの要求はクエリーベクトルとして表現される．すなわち，ユーザの興味を反映した特徴ほど，大きな重みを持つように設定しなくてはならない．一般に，ベクトル空間の次元は膨大なため，ユーザ自身の手で重みを調整するのは困難というよりも非現実的と言えよう[5]

従って，計算機によって適切なクエリーベクトルを求める必要が生じて来るが，一般に用いられるのがここで紹介する適合フィードバック（relevance feedback）[78] である．

適合フィードバックでは，同一の興味に関して検索を繰り返す事が前提条件となっている．すなわち，検索された文書それぞれに対し，ユーザが正否の判断を下し，その結果を元に，ユーザの興味ある文書へ近付くようにクエリーベクトルを更新していく（図 2.7）．

t 回目の検索後の，クエリーベクトル Q_t の Q_{t+1} への更新については，以下の式に従う．

$$Q_{t+1} = Q_t + \alpha \frac{1}{|P|} \sum_{D_i \in P} D_i - \frac{1}{|N|} \sum_{D_j \in N} D_j \tag{2.6}$$

ここで，P, N はそれぞれ正例（Q_t により検索された文書のうち，ユーザが興味があると判断した文書）集合，負例（検索された文書中のそれ以外の文書）集合，α は正負例のバランスをとる係数である．

[5] boolean match と同様に，二，三のキーワードに対応する特徴成分の重みだけ大きくすれば良いという考えもあるかも知れないが，それではVSMを用いる利点があまりないと考えられる．

表 2.3: 検索性能の一般的な評価尺度

評価尺度	定義
適合率（Precision）	$\frac{\text{検索された該当文書数}}{\text{検索された文書数}}$
再現率（Recall）	$\frac{\text{検索された該当文書数}}{\text{該当文書の総数}}$
予測正当率（Predictive Accuracy）	$\frac{\text{検索された該当文書数}}{\text{全文書数}}$

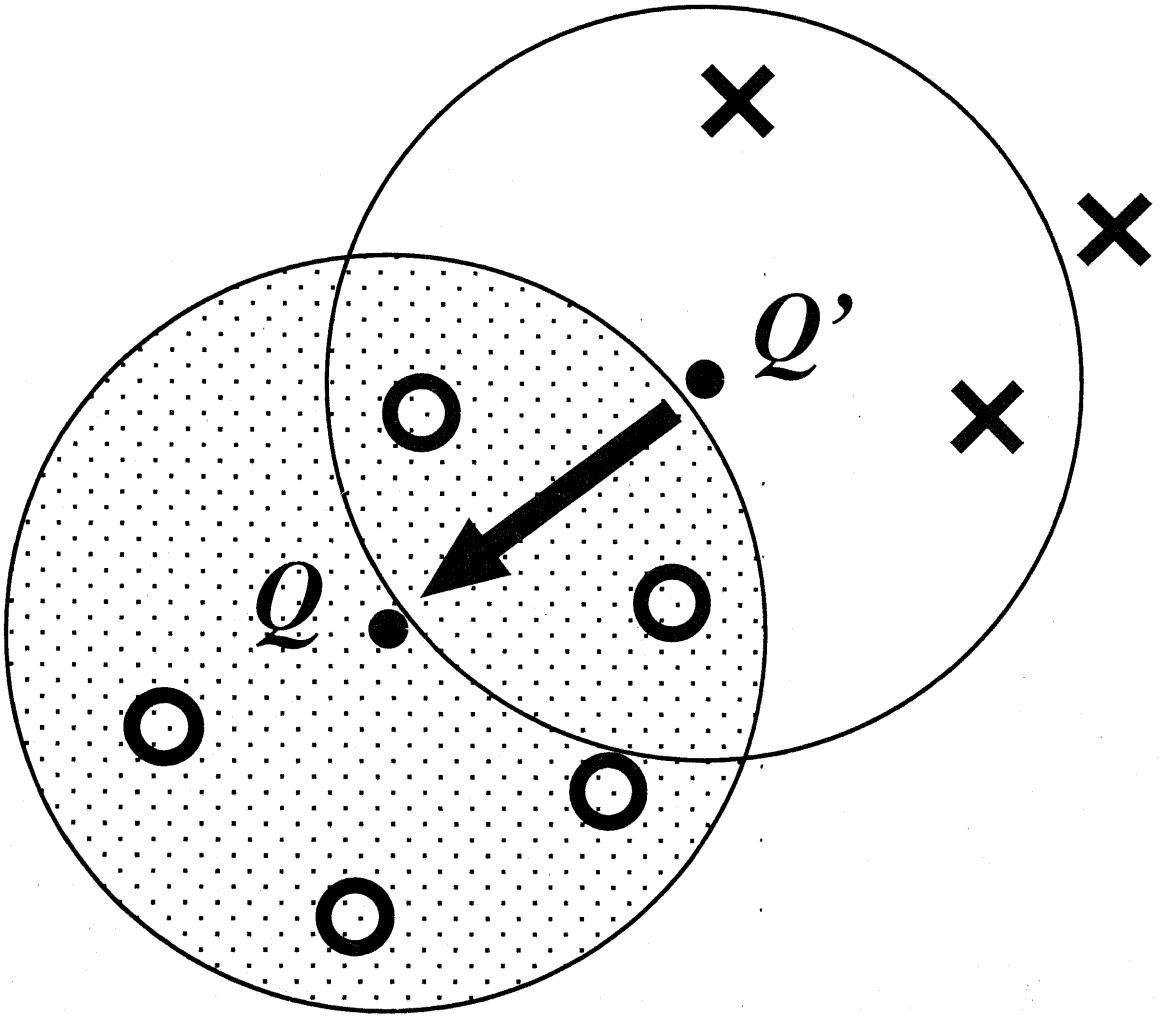


図 2.7: 適合フィードバックのイメージ

2.7.2 ベクトル空間モデルの拡張

本節では，ベクトル空間モデルに対してなされたいくつかの拡張について紹介する．ベクトル空間モデルは比較的シンプルな枠組であるため，様々な拡張が今までになされて来たが，ここでは特徴の削減，ベクトル空間の低次元化に関する視点から，いくつかの研究を取り上げることとする．

ベクトル空間モデルを用いる際の問題点として，索引語として何を用いるか，という問題点がある．すなわち，文書中に含まれる単語を全て索引語とすると，ベクトル空間の次元数は膨大なものとなってしまい，特徴ベクトルを記憶するのに膨大なディスク容量を要したり，類似度の計算に時間がかかりすぎるといった問題が発生する．

また，単語の中には英語の 'the' や 'of' などの様に，明らかに索引語として不要なものも存在するため，この様な単語までを含めて特徴ベクトルを生成したのではこれらがノイズとなり，検索性能は低下すると考えられる．従って，ベクトル空間の次元数削減やノイズカットを目的として，以下の様な特徴選択が行われる場合が多い．これらの手法は排他的なものではなく，いくつかを組み合わせて用いるのが普通である．

- ステマ（stemmer）[17, 98] による語尾変化などの削除
- HTML 構造を利用（要約レベルの特徴）[4, 7, 29]
- 低エントロピーの語の除去
  - **ストップワード（stop word）**や，HTML マークアップタグの除去 [29]
  - TFIDF や TF の値による選択 [29]
  - mutual information の利用 [4, 95]

これらの特徴選択の手法は互いに排他的ではなく，組み合わせて用いることが可能であるものが多い [7, 95]

ステマは例えば，'computer', 'computers', 'computing', 'computability' などを全て 'comput' に "stem" するものであり，これによって語形変化や単・複数などの違いによる表現上の冗長性などを除去することができる．

HTML 構造を利用した特徴選択は，タイトルや見出し，ニュース記事のサブジェクトや投稿者など，識別上有意味であると思われる部分を，構文知識に基づいて検出し，そこに含まれる語を特徴として選択する方法である．

ストップワードは，'a' や 'the' などの様に，あまりに一般的過ぎて識別上無意味な語であり，これらのリストをあらかじめ定義しておき，除去する．

TFIDF や TF の値による選択は，ある閾値を越えるもののみ使用したり，値の大きい上位 n 語のみを使用したりする．

mutual information とは，個々の特徴が観測データを正しく分類することのできる度合を表す一般的な統計尺度であり，一般に次式で求められる．

$$E(W,S) = I(S) - [P(W=present)I(S_{W=present}) + P(W=absent)I(S_{W=absent})] \quad (2.7)$$

$$I(S) = -[p(S_+)\log_2(p(S_+)) + p(S_-)\log_2(p(S_-))] \quad (2.8)$$

ここで，$p(W=present)$ は語 W がドキュメントに存在する確率，$S_{W=present}$ は語 W が少なくとも一つは含まれるドキュメントの集合をあらわす．また，S_+, S_- はそれぞれ正例集合，負例集合を表している[6]．

上記手法とは異なり，パターン認識や多変量解析などの分野で確立された統計的手法を用いて，特徴の縮退・クラスタリングを行う研究も存在する．Wulfekuhler らが提案されている**特徴クラスタリング** [130] では，各ドキュメントを次元とする空間において，k 平均分割クラスタリングによって特徴（単語）をグループ化し，特徴空間を低次元化することが報告されている．これにより，同様の話題・文脈において使用される単語はグループ化され，一つの特徴に縮退されるが，この研究では純粋に統計的手法により，ボトムアップに単語をグループ化するため，ほとんどの単語が含まれるグループが得られてしまったり，人が見て視点が読みとれないグループも多く選ばれるなどの問題点がある．

Latent Semantic Indexing(LSI)[13, 19] では，ベクトル空間モデルの問題点は語の多義性や類義語による表記の揺れなどの語彙問題（Vocabulary Problem）[23, 24, 27] にあるとし，索引語を成分とした通常のベクトル空間から，より低次元の空間に写像することによって語彙問題に対処できるとしている．

具体的には，単語（単語数 t）を行，文書（文書数 o）を列とする行列 X に対して特異値分解を行い，3 行列の積の形に変形する（式 (2.9)）．

$$X = T_0 \cdot S_0 \cdot O_0' \quad (2.9)$$

$$\approx T \cdot S \cdot O' \quad (2.10)$$

ここで，S_0 はランク r の対角行列となるが，さらに上位 k 個のランクのみを使用した行列 S で近似することにより，低次ベクトル空間に単語，文書を写像することができる．すなわち，行列 O' の列成分が，写像後の特徴ベクトルを表している．

[6]ここでは，ドキュメント分類タスクで一般的な二グループへの分類についての式を示しているが，三グループ以上への分類についても同様に定義できる．

LSIでは，単語の自己相関行列から共起関係に基づき，各単語の線形和の形で新空間の軸を求めている事になる．この考えは多変量解析でよく用いられるものであり[73]，文献検索の分野においても，同様の手法で文献や語句の連想検索を行う研究が行われている[72, 83].

Chapter 3

発想法・発想支援に関する従来の研究

「発想法」や,「発想支援システム」という言葉を聞いて，錬金術の様な，なにかうさん臭いものの様に感じる人も多いであろう.「アイデアなんて，思いついた本人でさえ，どうやってひらめいたかなんてわからないはず」,「いつでもアイデアを生み出せる方法があったら苦労はしない」,「そもそも，アイデアを思いつくプロセスに一般法則などあるのか」などが，主な反論であろうか.

ニュートンがリンゴの実が木から落ちるのを見て万有引力を発見したとか，アルキメデスが浴槽から溢れ出たお湯を見て王冠の金の純度を量る方法を発見した話などが，発想 (ひらめき) の例として有名であるが，確かにこれらの話から，発想に関する一般法則を見出すことは一見困難に思われる.

しかし，ニュートン，アルキメデスとも，落ちるリンゴやあふれるお湯を見ただけで，突然アイデアが浮かんだわけではなく，日頃考えに考えを重ねたという土壌があった上で，これらの出来事がきっかけとなり，発想へのブレークスルーとなったのであろう．すなわち，十分な思考を行った上で，なんらかの刺激によりひらめきが発生する事がわかる．すなわち発想とは，ひらめきだけを対象とするのではなく，それ以前の思考プロセスまでを含めるべきものなのである.

また，発想とはこの様な歴史的発見のみを指すのではなく，我々の日常生活でも頻繁に見られるものである．この点に関して Boden は，創造的なアイデアとは，新規で驚くべきものであり，かつ価値のあるもの（興味深い，有用である，美しいなど）でなくてはならないが，新規性にはその個人にとって新規の場合と，歴史上新規である場合があると指摘し，前者に該当する創造性を **P-creativity**，後者を **H-creativity** と呼んで区別している [8]．この考えに基づけば，料理人が新しい料理のレシピを考える場合だけが創造的なのではなく，冷蔵庫の中にあるものを見て今晩のおかずを考えるのも同じく創造的な活動であ

ると言えるのである.

「アイデアとは既存の情報の新規な組み合わせである」という意見もあるほど，入手できる情報の質・量が発想する上で重要である．しかし，インターネットに代表される情報環境の整備・拡大や，情報公開に向かう社会的潮流により，入手可能な情報に関する個人間の差はなくなりつつあるといえる．これからは，情報を入手する事のみに専心するのではなく，それらを元にしていかに思考し，アイデアを得るか，が重要な問題となってくるであろう．従って，発想法，ひいては発想支援が果たす役割は，情報化時代を向かえますます重要度を増すと言える.

本章では，続く 3.1 節で発想を創造的問題解決としてとらえ，そのプロセスについて整理し，以降での理解を助ける．その上で，3.2 節で発想法，3.3 節で発想支援システムについて，代表的なものを紹介する．最後に，発想法の今後の展望についてまとめる.

あらかじめ断っておくが，本稿では芸術，デザインにおける発想および支援システムについては対象外とする．これらにおける発想プロセスも，根底は他の発想プロセスと変わらないという意見もあるが，話が発散するのを防ぐためここでは言及しない.

3.1 創造的問題解決のプロセス

前述したように，発想はひらめきのプロセスだけでなく，そこに到るまでの思考プロセスも含めて考えるべきである．日本の代表的な発想法，KJ 法の提唱者である川喜田は，自らの研究プロセスを分析し，以下の **W 型問題解決プロセス**を提案している[1][52].

問題提起 問題が何であるか明らかにする.

現状把握 提起された問題に関連する情報を網羅的に収集する.

本質追求 関連情報の奥に隠された本質を読みとる.

仮説評価・決断 問題の本質を評価し，解決するための仮説を採択する.

構想計画 採択された仮説に関する構想計画を立てる [38].

具体策 構想計画を現状に合った具体策に展開する.

手順の計画 具体策を実施手順に展開する.

[1] W 型という名前は，このプロセスが思考，経験の両レベル間を行き来しながら進行することに由来する.

実施 手順の実行.

検証 仮説の検証を行う

総括・味わい 結果の吟味

このプロセスに従えば，せまい意味での発想（＝ひらめき）は仮説評価・決断プロセスにあたると考えられる．また，問題解決は，人工知能研究において，推論やプランニングといった文脈で研究されて来たが，発想の文脈における問題解決は，解決すべき問題自体が明確でない場合が多かったり，問題から仮説に到る道筋が自明でないなどの違いがある．従って，通常の問題解決プロセスと区別するために**創造的問題解決プロセス**と呼ばれる [39, 50, 102]．仮説採択後のプロセスについては，通常の問題解決プロセスに近いと考えられる．

創造的問題解決プロセスに関しては，様々なモデルが提案されているが，國藤 [61]，杉山 [110, 112] によるモデルを，前述した川喜田のモデルと対応づけてまとめると表 3.1 の様になる．以降では，國藤の提案したモデルに従って話を進める．

さきほども指摘したが，創造的問題解決プロセスの特徴は，アイデア結晶化までの段階にある．このため，発想支援システムや発想法ではアイデア結晶化までのプロセスを対象とするものが主流であり，それ以降のプロセスや環境まで含めて総合的に支援する場合には，**思考支援システム** [61] や**創造性支援 (creativity support)**，**創発メディア環境** [110, 112] などといった用語が使われている．本稿では狭義の発想プロセスとして，発散的思考プロセス，収束的思考プロセス，アイデア結晶化プロセスに焦点をあてて論じる．

発散的思考プロセスで重要なのは，問題を多方面から分析し，把握することであり，そのためには問題に関連する情報を，あらゆる視点から網羅的に収集することが重要である．

収束的思考プロセスでは，発散的思考プロセスで収集された情報を元に考えを整理し，問題を解決するアイデアを見つけ出すことが目的となる．一般に，発散的思考プロセスで収集された情報の中に，解決策となるアイデアがすでに含まれていることはまれであり，収束的思考プロセスにおいて収集情報を整理する事により，既存の概念の上位レベルにあるメタ概念の存在に気づいたり，情報が形成する概念空間上で強い影響力を持つ属性，コンセプトを把握したり，あるいは概念空間上に存在するホール（空白地帯）を発見したりすることによって，問題解決につながるアイデアを発見する事ができる [64]．発散的思考プロセスで解決策に直結するアイデアが得られた場合であっても，そのアイデアが本当によいものか，他にもっと良いアイデアはないか，さらに改善する余地はないかなど，収束的思考プロセスにおいて吟味・分析する事は大切であろう．

例えば，新商品を生み出したい場合，「何か新しいものはないか」と漠然と考えるよりも，まず過去のヒット商品やライバル会社の商品，購買者アンケートの結果や他分野のヒット

表 3.1: 創造的問題解決プロセス

W 型問題解決	國藤モデル	杉山モデル
問題提起	発散的思考	情報生成／収集
現状把握		
本質追求	収束的思考	情報整理／組織化
仮説評価・決断	アイデア結晶化	情報表現
構想計画	評価・検証	情報評価
具体策		
手順の計画		
実施		
結果の検証		
総括・味わい		

商品，流行に関する話題など，様々な情報を入手することから始め，それらを元に，ヒット商品に共通するコンセプトは何か，今までの商品の欠点は何か，などを分析した上で新商品を考えたほうが，よりよいアイデアが生まれると考えるのが自然であろう．この，情報収集に相当するのが発散的思考プロセスであり，分析に相当するのが収束的思考プロセスである．

この様な作業は，わざわざ言われるまでもなく，我々は日常行っているものである．なぜ，発想法や発想支援システムが必要となるのだろうか．この理由として，人間が思考する際には，無意識に常識などのフィルタリングがかかってしまう事があげられる．通常の思考においては，限定された時間の中で合理的な結論，行動を行うために有効であるが，新しいアイデアを得ようとする場合には，このフィルタリングが思考の妨げとなると言えよう．また，関連する情報量が多くなった場合にも，考慮できる範囲に情報量を限定するため，このフィルタリングが機能し，新しいアイデアにつながる情報が切り捨てられ，常識的な関係に収まる情報や，偏った視点において関連する情報のみが選択されてしまう危険性があると考えられる．これは，関連情報が容易に入手可能となった現代においては大きな問題といえるだろう．

以上の考察より，発想法や発想支援システムの目的は，次の二点に要約できよう．

発散的思考プロセス 常識や偏った視点にとらわれず，関連情報を網羅的に収集するための方法論，支援 [61].

収束的思考プロセス 大量の関連情報全てを考慮し，あらゆる角度，可能性から分析，評価するための方法論，支援 [111].

すなわち，常識などの枠組をいったん取り払い，十分に思考を発散させ，対象情報量を増加させた後で，再びまとめ上げ，概念空間を再構築することが，発想法，発想支援システムの目的と言える．もちろん，思考プロセスがこの２プロセスに完全に分離できるわけではなく，発散的思考プロセスにおいても系統だった発想を行うためには収束的思考の要素も必要であろうし，収束的思考プロセスにおいても，整理しながら新たな情報の必要性を感じたり，新たなコンセプトに属するアイデアを思いつく事も多いであろう．また，発散的思考プロセスと収束的思考プロセスは一方向の流れではなく，収束的思考プロセスを経た結果をもとに再び発散的思考プロセスに戻るといった事も有効である．

3.2 発想法

3.2.1 発散的思考プロセスに関連する発想法

前述の様に，発散的思考プロセスの目的は，常識や固定観念などに縛られず，問題に関連しそうな情報をできるだけ多く，網羅的に集めることである．これは，世界で最も有名な発想法である**ブレーンストーミング**の4つのルールにも表れている（図3.1）[40].

情報収集は，机上で行うべき作業ではなく，日常生活のいたるところで関連情報に出会う可能性がある．また，アイデアもいつ浮かんで来るか，予測できないものである．このため，メモやカメラを持ち歩くなどして，常に情報収集に努めるとともに，創造的問題解決プロセスにおいてすぐ利用できるように，カードやデータベースなどを利用して収集した情報を整理しておくことが重要であると多くの識者が指摘している．しかし，十分な情報を事前に全て収集しておくことは困難であり，発散的思考プロセスにおいて，さらに情報量を増加させる必要がある．事前に収集された情報をもとに，さらに情報量を増加させるために利用される方法論に従い，発散的思考プロセスに関連する発想法を分類すると以下のようになる．

連想を刺激する カタログ法，チェックリスト法，クルーカード法

網羅的数え上げ 属性列挙法，希望点／欠点列挙法，目的発想法

情報の組合せ 形態分析法，△○□発想法，ポジショニング法

類比・類推を利用 NM法，シネクティクス，バイオニクス，システム・アナロジー

発想転換 逆設定法，擬物化法，仮想状況設定法

グループによる発想 ブレーンストーミング，ピン・カード法，ブレーンライティング法

以下では，それぞれの分類について，代表的な方法論を中心に説明する．ここで紹介する以外にも，様々な発想法が提案されているが，それらについては[12, 40]に詳しい．

連想刺激法

人間の脳は，できの悪いデータベースだという意見がある．これは，覚えて（記録されて）いるのに思い出せない（検索できない）情報が多く存在することに由来する．脳における連想構造の非対称性は認知科学でも報告されており[14]，ふとしたきっかけで頭の片

隅にある情報にアクセスすることができたりする．その様なきっかけとなる刺激を与えるのが，連想刺激法の特徴である．

カタログ法では，自然の風景，都会の雑踏，あるいは雑誌の広告などの写真をファイルしてカタログを作り，それを見ながら視覚イメージの刺激を誘発して発想を試みる．チェックリスト法では,「他に使いみちは？」,「拡大したら？」などの，発想を導く手がかりとなるリストを用意し，それらの指示に従って発想やイメージを膨らませて行く．オズボーンのチェックリストや，語呂合わせで暗記しやすくした「ださくにたおち」，“SCAMPER” などがある（図 3.2）．クルーカード法も，同様に発想の糸口となるイラストやメッセージを書いたカードを用意し，ほんの少し発想に方向性をつける事を狙っている．

属性列挙法

属性列挙法は，商品改良や新商品アイデアを考えるのに適した発想法といわれており，まず第一に対象となる課題商品の属性（性格・特徴）を全て列挙し，各属性を検討しながらアイデアを発想しようというするものである．歴史は古く，1930 年代にロバート・クローフォードによって開発された方法である．この方法の根底には,「全てのアイデアは，それ以前にあったアイデアを何らかの方法で修正したものである」という考えがある [41]．具体的なステップは以下の三つに分かれている．

1. 課題とその解決目標を記述する

2. 全ての考えられる属性を列挙する

3. 課題の解決目標に適合する様に，各属性を修正できないか検討する

この方法のポイントは，いかにして多角的に，属性をもらさず列挙するかにかかっているが，一般に属性はその商品の素材・形態・仕様・機能などに関する**直接的属性**と，歴史・意味・イメージなどに関する**間接的属性**に分類できると言われている．直接的属性に関しては，**名詞的属性**（部分・材料・製法など），**形容詞的属性**（性質・状態など），**動詞的属性**に分類して列挙することが提案されている．

属性列挙法は，網羅的数え上げを行うための代表的な発想法であり，オーソドックスな手法のため，様々な変形／類似バージョンが存在する．希望点／欠点列挙法や属性連想法などがこれにあたる．また，他の発想法の前段階として用いられることも多い．発想の転換を行う逆設定法では，属性列挙法によってリストアップされた属性それぞれについて，設定を逆転し，それを元にしてアイデアを発想する．

1. 批判一切お断り
2. 自由奔放
3. 量を求む
4. 組み合わせ・改善

図 3.1: ブレーンストーミングのルール

だ 代用できないか
さ さかさまにしたら
く 組み合わせたら
に 似たものはないか
た 他の用途はないか
お 大きくしたら
ち 小さくしたら

S Substitute?
C Combine?
A Adapt?
M Modify?
Magnify?
Minify?
P Put to other uses?
E Eliminate?
R Reverse?
Rearrange?

図 3.2: チェックリストの例

形態分析法

この方法は，1940 年代初めにフリッツ・ズイッキーが考案した方法である．形態分析法も，属性列挙法と同様に問題を構成する要素を列挙するところから始めるが，次に各要素ごとに，とりうる全ての変化（独立変数）を洗いだす．そして，各要素を次元とする多次元チャート（morphological chart）を作成することにより，全独立変数の組み合わせを一つ一つ検討して行く．

例えば，問題を構成する要素が 3 つあり，各要素ごとにとりうる独立変数が 4 つずつ存在する場合，3 次元のチャートが形成される．独立変数の可能な組み合わせは $4^3 = 64$ 通りであるが，既存の商品や常識的なアイデアはこのうちの一部であり，中には今まで思いもよらなかった組み合わせがあるはずである．それらの中から，有効なものを検討の上で見出すのが形態分析法の狙いである．この様な，今までにない組み合わせ，位置付けによるアイデアを分析により見出す方法論としては，他に**ポジショニング法**などがある．

類比・類推の利用

他の分野の仕組みや事例を当てはめる事により，類似する要素からアイデアを得る事が期待できる．バイオニクスは，生物の機能を工学的に活用する形態の発想法であるが，ガラガラヘビの器官の研究から熱を関知して追尾するミサイルが開発されたのは有名な話である．

類比を活用した発想法としては，シネクティクスと NM 法が有名である．シネクティクスでは，類比に関して次の二つのアプローチを利用する [89, 90].

異質馴化 新規なものを既知なものの同類とみなす.

馴質異化 既知のものに対し新しい見方をすることによって新たな側面を見出す

NM 法は，日本では KJ 法に並んで有名な方法であり，問題のイメージを膨らます手段として類比を用いている．

ブレーンストーミング

この手法は，1930 年代後半にアレックス・オズボーンによって発案されたものであり，発想法の元祖という点でもこの手法の持つ意味は大きい．複数の人間が共同作業でアイデアを生み出して行く際の方法論であり，図 3.1 の様な簡単なルールに従うだけでよく，会議などに応用できるため，多くの企業などで実践され，さまざまな改良版が存在するのもこの方法の特徴である．

発散的思考プロセスでは，多角的な視点からもれなく関連情報を集めることが重要であり，このためにも一人で発想するよりも，視点の異なる人達が集まってアイデアを出し合うほうが有効であるといえる．しかし，やみくもに人数だけ集めても，よいアイデアが出るわけではなく，充実した議論に導くための方法論が必要になる．

ブレーンストーミングの実行は制約が少なく容易である分，参加メンバーの質や司会の力量に左右されやすいという欠点もある．また，日本や欧米では国民性の違いによる会議スタイルの違いがあるとも言われており，各企業などの実情に合わせた変更を加えることも必要となる．例えば，ドイツでは発言する代わりに紙に記入するスタイルの**ブレーンライティング**が考案されている．他には，ブレーンストーミングの途中にテーマを絞るプロセスを加えたり，グループに分けてアイデアを競い合わせたり，本当の課題をあえて知らせず，幅広い議論を期待するなどの改良がブレーンストーミングに加えられている．

3.2.2 KJ 法：収束的思考プロセスに関連する発想法

収束的思考プロセスでは，発散的思考プロセスを経て網羅的に収集された関連情報全てを考慮し，あらゆる角度，可能性から分析，評価を行う．このためには，頭の中だけで思考するよりも思考を外化する方がよく，あらゆる関係の吟味手段としては，線形な文書よりも，図解の方が有効であることが指摘されている．従って，収束的思考プロセスに関しては図解の作成を扱う発想法が主流である（KJ 法，7×7 法など）．

日本を代表する発想法として知られる **KJ 法** [51, 52] は，収束的思考プロセスに重点をおいた発想法であるといえる．これは，文化人類学者であった川喜田が，自身の研究，特に収集したデータの整理に関して生み出した方法論である．彼は科学を，書物（人の手によって整えられた情報）を読むことが中心の書斎科学，整えられた環境で実験・検証を行う実験科学，未整理な環境で検証を行う野外科学の三つに分類することを提案しており，野外科学における方法論として発想法（KJ 法）が威力を発揮すると述べている．

KJ 法は，以下の 4 ステップに従って行う．このうち，方法論としてより整備されているのはラベル作りから A 型図解までであり，狭義の KJ 法という場合にはここまでのプロセスのみを指す場合が多い．

ラベル作り プリミティブなデータの作成 (図 3.3(a))

グループ編成 ラベルのボトムアップな編成 (図 3.3(b),(c))

A 型図解 空間的配置による関係の整理 (図 3.3(d))

B 型文章化 図解から文章の作成

この４ステップを何度も繰り返して行う場合もあり，これは累積 KJ 法と呼ばれる．川喜田は，提案した W 型問題解決モデルにおいて 6 度の累積 KJ 法を行う事が最も有効であると主張している．

KJ 法ではそのプロセスの表面だけをなぞるのは危険であり，研修などを通してその精神から理解する必要があると言われている．例えば，「データをもって語らしめよ」とか，「土の香りをたたえた一行見出しを作れ」などの感覚的な言い回しが多用されている．本稿の限られたスペースでその様な精神を詳しく説明する事は不可能であるので，詳細については [51, 52] を参照されたい．

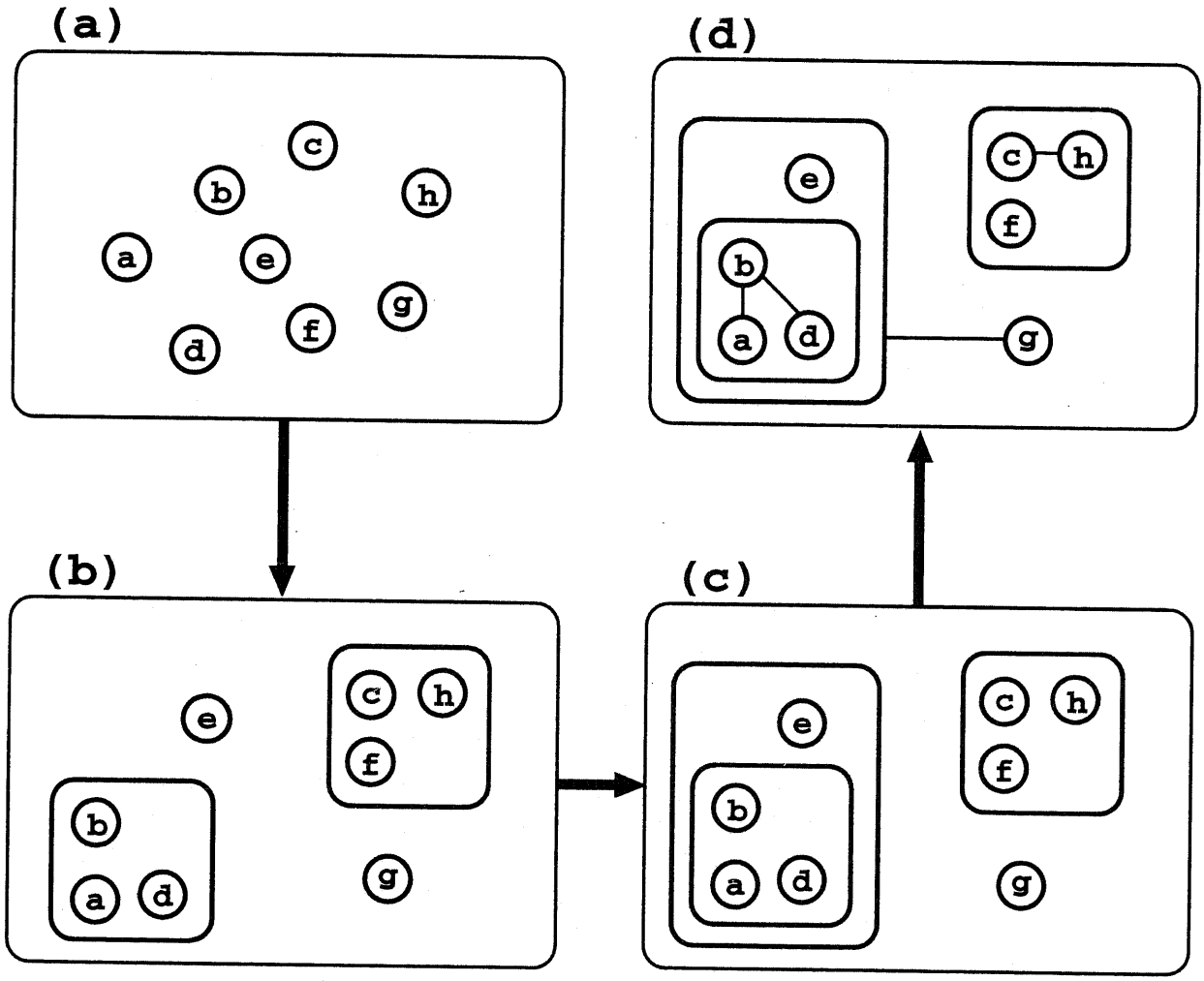


図 3.3: KJ 法 (A 型) のプロセス

3.3 発想支援システム

発想支援システムは，3.1 章で示した人間の創造的問題解決プロセスを文字通り支援するものである．従って，発想支援システムはどのプロセスを支援するかによって分類できるが，何を支援するか，あるいはどうやって支援するかによってさらに分類できる．ここでは，Young によって提案された，レベルによる分類を紹介する．これは何を支援するかによる分類と言える [89, 90].

秘書レベル 計算機を便利な文房具として捉え，概念操作にまつわる雑用を肩代りさせる．

枠組–パラダイムレベル ユーザの思考の構造化，流れに関して適切な枠組を提供する．

生成レベル 新たなアイデアを生成して提供したり，関連を発見したりする．

この三つのレベルは，後のものほど支援の度合い，洗練度が高くなっているが，高いレベルの支援が必ずしも効果的な支援であるとは限らないとされている．この分類に従えば，前章で紹介した発想法のほとんどは枠組–パラダイムレベルに相当することになり，これらの方法論を計算機上で再現したものもこのレベルにあたるとみなすことができる．

図 3.4 は，縦軸にレベル，横軸にプロセスをとって，代表的な発想支援システムを分類したものである．実在するシステムのほとんどは，単一の思考プロセス，レベルしか考慮していないわけではなく，あるいは用途によってどちらの思考プロセスの支援にも用いることができるものもある．従って，この分類はあくまでどの要素が強いかに従っていることに注意されたい．以下では，これらのうちのいくつかのシステムについて紹介する．

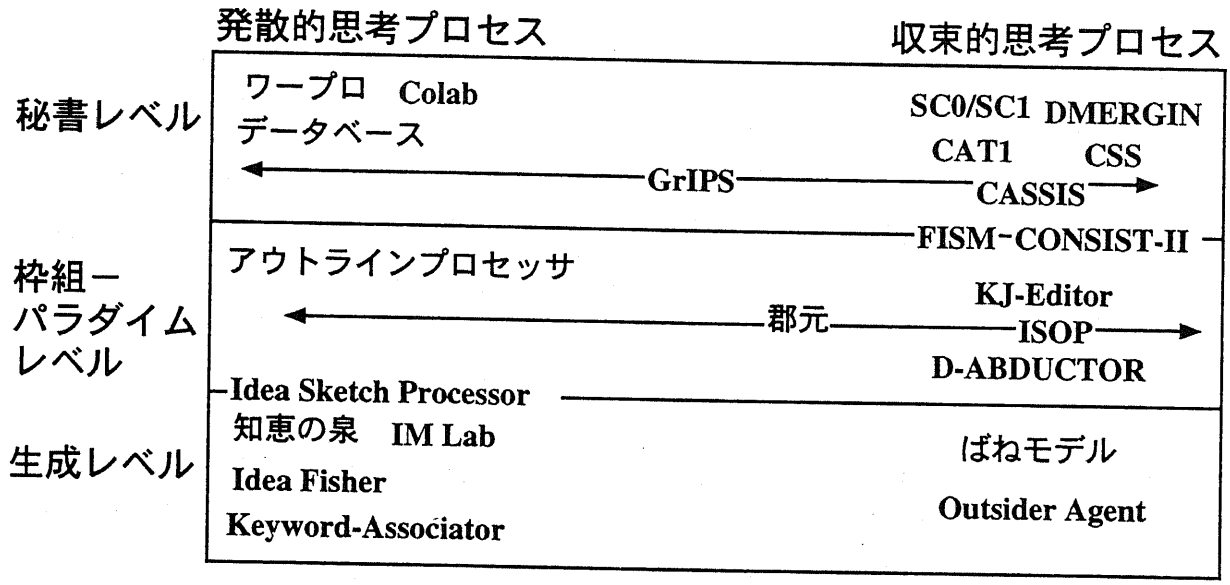


図 3.4: 発想支援システムの分類

3.3.1 秘書レベル

発散的思考支援システム

秘書レベルの支援は，ユーザが試行錯誤しやすい環境を提供する事で十分である．その意味では，紙と違って並べ換えや削除などの修正が容易なワープロは，十分に発散的思考プロセスを支援している．これはワープロに限ったことではなく，既存の発想法を計算機上に実装するだけで，秘書レベルの支援につながるケースも多い．

Colab に代表されるグループウェアは，参加メンバーが時間・場所を共有しなくてはならなかった既存の会議を，遠隔・非同期を可能にしたことが秘書レベルの支援，とくにグループでの発想支援につながるといえる．GrIPS[54] では，後述する Keyword-Associator と D-ABDUCTOR を利用しており，両思考プロセスをカバーした支援を行える．

知的生産に関する議論で，情報をカードに記入して管理することの重要性は以前から指摘されてきたが，情報を計算機上で管理するデータベースも，情報の保守作業を助ける点で秘書レベルの支援システムとみなせる．

収束的思考支援システム

収束的思考プロセスでは，図解により思考を外化し，それらを試行錯誤しながら整理して行く手法をとるのが一般的かつ有効であることが知られている．これは KJ 法などの発想法においても同様であるが，やはり紙の上で行うよりも，計算機上で行った場合のほうが，配置を気楽に変えやすいといったメリットがある[2][22]．秘書レベルの支援としては，この配置を自動化する事が考えられる[3]．

一般に，図解に配置されるオブジェクト間の関係は多次元であると言える．すなわち，オブジェクトの持つ属性（オブジェクトがテキストの場合はキーワード）の数が，空間の次元数に対応する．これに対し画面に表示できるのはせいぜい三次元がせいぜいであろう．これは画面上の制約だけでなく，人間が関係を把握するためにも多次元空間を低次空間（一般には二次元）に削減することは重要である．CONSIST-II[103] では，オブジェクト間に存在する関係を分類しておき，同時に二種類の関係を画面上にマッピングしている．SC0/SC1[106] や CSS[117]，CAT1[114]，CASSIS[2] では，統計的手法である多次元尺度構成法（multi-dimensional scaling）や双対尺度構成法（dual scaling）などを用いて，オブジェクトや属性を二次元空間に配置している．

---

[2]その反面，画面サイズによる一覧性の制約の存在が指摘されている．

[3]後述する KJ 法エディタも，狙いとしては同様であるが，KJ 法と言う発想法の枠組に従う関係上，本稿では分類して扱った．

最終的に得られる図解は，ユーザの視点を反映したものと言えるはずである．したがって，同じオブジェクトを用いて図解を作成した場合でも，ユーザごとの視点，意見の違いが図解上に表れるはずである．複数の図解をマージすることにより，互いの視点の共通点，相違点を図解化し，コミュニケーション支援につなげようという研究が最近目立つようになってきている [88, 113, 115, 116, 117]．今後は，現在のブラウジングに見られるような，あらかじめ検索要求が明確でない場合の情報検索にも応用が考えられている [107]．

3.3.2 枠組–パラダイムレベル

発散的思考支援システム

ワープロに対し，文章の構造編集機能を強化し，「章立て → 節立て → 内容」というトップダウンな文書作成を支援するツールがアウトラインプロセッサであり，市販製品もいくつかある．これは，トップダウンに木構造に展開しながら系統立てて列挙して行くための枠組といえる．また，高野らは発想支援向けの対話モデルを提案し，発想を支援する自然言語インタフェースを開発しているが [121]，これも対話の流れに枠組みを想定したものと言えよう．この他にもユーザにアイデアの発想を促す支援ツールは存在するが，生成レベルに分類すべきものがほとんどである．

KJ エディタ

KJ 法は日本を代表する発想法であり，かつ図解を計算機上で扱うメリットは大きいため，日本で開発された収束的思考支援システムは，KJ 法の影響を受けたものが多い [86, 111]．ここではこれらをまとめて KJ エディタと呼ぶことにする．KJ-Editor や D-ABDUCTOR は，狭義の KJ 法（A 型図解）を忠実にツール化したものであるが，なんらかの拡張を施したものも多い [109, 111]．例えば，郡元はグループウェアとして実装されており，遠隔ユーザ間で用いた場合のコミュニケーションに関して実験が行われている [76, 108]．

図解による思考の外化は，収束的思考プロセスだけでなく，コミュニケーションにも有効であるのは前述の通りだが，D-ABDUCTOR は KJ 法のみの用途ではなく，完成度の高い図解化ツールとして，Keyword-Associator といった他のシステムと組み合わせるなどして，様々に利用されている [85, 112]．

3.3.3 生成レベル

発散的思考支援システム

3.2.1 節で発散的思考プロセスに関する発想を紹介したが，連想や類比を用いるものが比較的多数存在した．生成レベルでの支援では，この連想や類比を利用して，関連キーワードやテキストの提示をするものが主流である [41].

Keyword-Associator は，電子ニュース記事を元に連想辞書を自動生成し，以下の検索を行うことができる [128].

- 単語をキーとして関連単語を検索
- 単語をキーとして関連テキストを検索
- テキストをキーとして関連単語を検索
- テキストを入力して関連テキストを検索

これにより，D-ABDUCTOR などの KJ エディタと組み合わせた場合に，項目のグループ化やラベル付けなどの支援を行うことができ，より効力を発揮する．

IM Lab[41] は，ロシアで提案された発想法である TRIZ をベースとした商用の発想支援ツールである．このツールは，過去の発明，特許などの事例分析に基づいて作成された，イラストを多用した豊富な知識ベースを持っており，自分の専門外の分野の原理・事例を提供することにより発想を触発するものである．解決すべき問題により，以下の三つのモジュールを使い分ける．

IM Effects 必要な機能を実現するための，科学的法則・原理の提供

IM Principles 矛盾する問題を解決するために，矛盾対比表を元に問題解決につながる原理を提供

IM Prediction 製品の飛躍的改善のために，次段階予測情報を提供

知恵の泉 [91] は類推に基づき，ベース領域にあって目標領域にないコンセプトを提示し，ユーザに対しそのコンセプトの名前を要求する形で発想を促す．

変わったところでは，デザインスケッチの作成を支援する Idea Sketch Processor[84] では情報の組合せによる発想を扱っており，ユーザの視覚によるスケッチの分類と，スケッチに付与されたキーワードを元にしたシステムによる分類結果を組み合わせることにより，新たなコンセプトをユーザに気づかせ，スケッチを描かせる形の支援を行う．このシステ

ムは，一回の分類数を3つに限定したり，作業フローを厳密に設定している点で枠組–パラダイムレベルの支援ともいえる．

収束的思考支援システム

3.3.1 節で紹介した図解の自動配置においては，多次元空間を低次（二次）空間に変換する際にゆがみが生じる．例えば，CASSIS[2] では式 (3.1)，ばねモデル [122] では式 (3.2) で多次元空間上での距離（非類似度）と距離空間上の距離との誤差関数 E を定義している．ここで，d_{ij} はオブジェクト i,j の距離空間上の距離（d_{Ave} は平均），s_{ij} は非類似度（s_{Ave} は平均），l_{ij} は i,j 間の類似度を元に計算されたばねの自然長である．

$$E = \sum_{i=1}^{N_{obj}} \sum_{j=i+1}^{N_{obj}} \left(\frac{s_{ij}}{s_{Ave}} - \frac{d_{ij}}{d_{Ave}} \right)^2 \tag{3.1}$$

$$E = \sum_{i=1}^{N_{obj}} \sum_{j=i+1}^{N_{obj}} \frac{1}{2} k (d_{ij} - l_{ij})^2 \tag{3.2}$$

CASSIS や CAT1 では，ユーザの配置を初期配置として，最急降下法を用いて誤差が極小となるように自動配置を行う．最急降下法では最適解が得られることは保証されていないため，システムによる配置はユーザによる配置に依存する．従って，ユーザは配置を様々に変更することによって，システムが提示する配置も変化し，それによって刺激を受け，思考の整理に役立つ効果が期待されている．

さらにばねモデルの場合には，ユーザによる配置からシステムによる配置に変化する過程をアニメーションによって示すことにより，さらなる刺激をユーザに与えることができるとしている [122]．

Outsider Agent[83] は，図解中から抽出したキーワードを元に，連想行列を用いて連想キーワードを求め，それを元に記事を検索し，図解に追加する．これにより，作業が膠着状態に陥った際に刺激を与える効果を期待している．これは，グループ作業における他者の視点が持つ効果を，計算機が代行するものといえる．

3.4 まとめ

本章では，発想を創造的問題解決のプロセスとしてとらえ，対応するプロセスに関して発想法を分類し，紹介した．また，発想支援システムについては，さらに支援レベルによる分類も行った上で代表的なシステムを紹介した．

前述のように，創造的問題解決は特別な作業ではなく，我々の日常生活で一般に見られるものである．時代の変化が速くなるにつれ，企業や研究者にとって創造的問題解決プロセスを円滑に進めることのできる方法論，支援システムの整備は急務であるといえよう．

今後の発想支援研究が向かうべき方向性として以下の三点があげられる．

- 情報検索技術との連係
- 通常作業に対するシームレスな支援の実現
- 評価方法

インターネットに代表される情報環境の整備・拡大により，問題解決に関連する，十分な量の情報が集められるようになりつつある．従って発散的思考プロセスにおける問題は，収集した情報の量をいかに増幅させるかから，膨大すぎる情報源から関係のありそうなものをいかに効率良く見つけ出すかに変わりつつあると考えられる．これは情報検索の問題にもつながる．既存のサーチエンジンやデータベースでは，探索前にユーザの要求が明確になっていることが前提となっているが，創造的問題解決プロセスにおいては，ユーザは問題すらも明確に捉え切れていない場合が多い．この様な場合，図解編集などの作業を通じて視点を整理しながら，情報源にアクセスすることが必要となるであろう．この点で，我々の提案している **Fish Eye マッチング**は，既存の概念体系を利用して図解からユーザの視点を抽出・表現できるため，情報検索と創造的問題解決プロセスを融合することのできる有効な技術といえよう．

これと関連するが，創造的問題解決プロセスは各プロセス毎に分離したものではなく，相互の関連の度合は非常に高いだけでなく，情報検索や文書作成など，日常の他の作業とも連係したものである．しかし今までの発想法および発想支援システムでは，一部のプロセスのみを扱ったり，日常作業との隔離を強いるものが多かった．今後は創発メディア環境[110, 112]で提案されているように，個々の支援システム，ツールを組合せ，日常の作業とシームレスな形で支援環境を用意する方向に向かうであろう．このためには，各ツールの完成度・洗練度を高め，秘書レベルに近い形で支援する必要があると考える．

最後に評価であるが，現在のところ要素技術に関する定量的評価か，アンケートや作業分析による定性的（主観的）評価を行うしかない．これはヒューマンインタフェース研究と同じ問題を抱えているが，各支援システムを共通の土台で比較できるような，評価指標の制定が望まれる．

Chapter 4

テキストマイニングの新展開～ナレッジマネジメント

4.1 ナレッジマネジメントの概要

近年，企業の生産性向上の切札として，**ナレッジマネジメント**というスローガンが注目を集めている [33]．これは，欧米の経営者の間で 90 年代後半から使われ始めた用語であり，日本ではまだまだ十分に認知・普及されているとは言えないようであるが，今後経済界で注目を浴びるスローガンである事は間違いないであろう．

ナレッジマネジメントを簡単に説明すると，「社内情報・知識の共有，集約による，伝達・利用の効率化および，新知識の発見・創発」という事になる．ここでいう社内情報・知識とは，社内文書・資料や各種データベースに蓄えられたものだけではなく，各社員・エキスパートの持つノウハウといったいわゆる暗黙知も含まれる．従来は，情報・知識の共有は社内のごく一部の組織内だけに限られ，暗黙知に到っては人づてで探さなくてはならないなど，全社的な視点で見れば十分に活用できているとは言えない状況であった．これに対しナレッジマネジメントでは，全社レベルでの情報共有を目指し，組織として持っている情報・知識の一層の有効利用を図るものである．

情報を集約する事のメリットはそれだけではない．一般に，新しいアイデア・発明のほとんどは，既存の概念の組み合わせや，他分野からの類推により得られているという指摘がある通り，集約された情報量が増える程，そこから創発により生み出されるアイデア・知識も増えると言えよう．例えば，小売業界で近年成長の著しいコンビニエンスストアでは，POS を用いて集約された各店舗毎の販売情報をもとに，売れ筋商品の把握や新企画立案を行い，成功している．

ナレッジマネジメントを支える要素技術をあげると，以下のようになる．

エージェントアプローチ 柔軟な情報流通・管理の実現

オントロジー・シソーラス 情報理解・伝達における共通の背景知識としての利用

データマイニング データベースに整理された情報からの知識発見

テキストマイニング 要約・抄録の自動生成，文書整理，視覚化など

現代社会の変化の速度は非常に速く，集中・固定的な知識・情報集約の手法を用いたのでは社会変化に十分追従する事ができない．従って，各組織・部門毎に情報・知識を分散管理しつつ，全社的な共有を図るべきであることが指摘されており，これを実現する上でエージェントアプローチが注目を集めている [33, 82].

エージェントは情報空間におけるユーザの代理人となり，情報の収集，伝達などの作業をユーザに代わって行う．この時，エージェントにユーザの意図（収集したい情報，伝えたいメッセージなど）を伝える時，あるいはエージェント間の情報伝達（コミュニケーション）を行う際に，意志疎通のためには共通の理解基盤・背景知識が必要となる．オントロジー・シソーラスはこの様な用途に用いる事ができ，近年研究が盛んになっている．

また，社内情報の多くは文書の形で蓄積されているものが多く，これらは 2.1 節で述べた様に，従来は計算機による収集・活用が難しいとされていたものである．しかしナレッジマネジメントの観点からは，この様な文書化された情報も集約・有効利用の対象としなくてはならない．データマイニングが計算機にとって扱いやすい形で蓄積されたデータベースから，有効な知識を発見する手法であるのに対し，人間が自由な形式で記述した文書情報を整理したり，有効な情報・知識を抽出する手法をテキストマイニングと呼ぶ．

以下では，オントロジー・シソーラスに関する研究および，テキストマイニングについて，もう少し詳しく触れる事にし，これを通じて我々の提案する個人情報の整理支援について考察する．

4.2 オントロジー・シソーラスに関する研究

オントロジー・シソーラスの有効性は，上述したようなエージェント・ユーザ間，およびエージェント間の意志伝達だけではない．人間同士であっても，同じ概念を異なる用語を用いて索引づけしたり，あるいは人によって同じ用語で異なる概念を指すなどの語彙問題（Vocabulary Problem）が存在する事が指摘されている [23, 24, 27].

国語辞典や英和辞典などの各種辞典には，我々の概念に関するコンセンサス，いわば常識が体系化され，整理されていると考える事ができるが，これを人間だけでなく計算機も含めて共通の理解基盤・背景知識として利用するためには，計算機にも理解可能なフォーマットで体系化，電子化しなくてはならない．計算機に理解可能な辞書，シソーラスとしては，英単語については WordNet[129]，Roget のシソーラス，日本語については EDR[20]，分類語彙表 [56] などが存在し，テキストマイニングに関する多くの研究において活用されている [5, 25, 30, 69, 74].

本研究でも使用している EDR 電子化辞書は，計算機による先進的な言語処理を目的として開発されたものであり，図 4.1 に示した辞書から構成される．単語辞書は，各単語を意味（概念）毎に別々のレコードとして格納しており，各単語レコードは見出し情報（語幹，活用語尾，読み，発音など），文法情報，意味情報及び運用・その他の情報（頻度など）から構成される．意味情報として記述される概念識別子は，概念辞書へのリンク情報となっており，これら識別子間の関係は概念辞書にて整理されている．

概念体系辞書は，概念間の上位下位関係を記述したものであり，直接上位下位関係にある識別子のペア毎にレコード化されている．EDR 電子化辞書においては，概念は他の概念との相対的関係によって定義されるが，各概念を説明する見出し情報については概念見出し辞書に記述される．概念記述辞書には，他の概念との意味的関係（動作主，道具，場所など）が記述されている．

4.2.1 オントロジー・シソーラスの生成

前節で紹介したシソーラス，電子化辞書には一つの欠点がある．それは，固定化された体系であるため，流行語や最先端の用語などへの対応ができない，あるいは遅れてしまう事である．そのため，ニュース記事などの対象分野に関する文書を収集し，それから専門分野におけるオントロジー・シソーラスを生成する研究も多く行われている [14, 47, 49, 127]. それらの研究で主に用いられるのは，文書中における語の共起関係である．

浦本らは，コアとなるシソーラスに，分野依存のテキストから抽出した未知語を追加する事によるシソーラスの拡張を図っている [127]．具体的には，ある未知語について，分野依存のコーパス中に出現する係り受け関係がそれと最も近い，すなわち用法の近い単語を既存のシソーラスから探し，そこに未知語を追加する形でシソーラスの拡張を行う．

笠原らは，広辞苑など複数の国語辞典の語義文から各見出し語間の関係を抽出し，概念ベースを構築している [44, 49]．この手法により構築された概念ベースにおいては，単語，文書ともに，単語を特徴とする同一の特徴ベクトル空間上に表現する事ができるため，文書クラスタリングに適用した場合，クラスタを代表する見出し語（クラスタ特徴語）の抽

単語辞書 単語と概念（意味）との対応関係，文法的特性を記述

- 日本語単語辞書 ... 25 万語
- 英語単語辞書 ... 19 万語

対訳辞書 日本語と英語の単語見出し間の対応関係を記述

- 日英対訳辞書 ... 23 万語
- 英日対訳辞書 ... 16 万語

概念辞書 概念についての知識を記述

- 概念体系辞書 ... 40 万概念
- 概念記述辞書 ... 40 万概念
- 概念見出し辞書 ... 40 万概念

共起辞書 言葉の言い回しに関する情報を 2 項関係で記述

- 日本語共起辞書 ... 90 万句
- 英語共起辞書 ... 46 万句

専門用語辞書（情報処理） 情報処理分野に関する専門辞書群

- 日本語専門用語単語辞書 ... 12 万語
- 英語専門用語単語辞書 ... 8 万語
- その他（概念体系，対訳，共起各辞書）

EDR コーパス 形態素レベルから構文・意味レベルまで解析して得られる言語データ

- 日本語コーパス ... 22 万文
- 英語コーパス ... 16 万文

図 4.1: EDR 電子化辞書の構造

出などが容易に行えるといった特徴がある [60].

chen らは，文書中における単語の共起関係を元に意味ネットワークを自動構築し，これをシソーラスとして利用している [14]. 一般的なシソーラスにおけるツリー構造の様な，厳密な体系化を行わない点では西田らが提案する弱構造化オントロジー [47, 82] も同様であり，計算機による文書からの自動獲得が容易であるというメリットがある.

4.2.2 オントロジー・シソーラスの利用

上述した様に，エージェント間の共通の理解基盤・背景知識としてオントロジーを用いる研究もあるが，ここでは情報収集・抽出などにシソーラス・オントロジーを用いる研究についていくつか紹介する.

2 章では，文書マッチングの手法として，キーワードを含む文書を検索する boolean match や，ベクトル空間モデルを紹介したが，これらの手法においては，単語（キーワード）はあくまでシンボリックな特徴として扱われており，例えば同義語や類義関係，上位下位関係など，語が本来持っている重要な情報を利用できていない. この様な語の持つ意味・概念を活用するために，シソーラス・オントロジーを利用する事ができる [5, 25, 30, 36, 46, 47, 58, 60, 66, 69, 82].

英語文書を対象としたシステムでは，WordNet を利用する研究が多く，Green は文書中に存在する類義語の鎖を WordNet を用いて複数発見し，各段落における鎖の密度分布を元に段落に書かれた内容の類似性を判断，リンクを張ってハイパーテキスト化している [30]. Mani は，文書中の語の隣接関係と，WordNet から求めた語の意味的関係を用いて文書をグラフ表現し，比較を行っている [69]. また，Bagga らは文書からの情報抽出ルールを意味ネットワークの形で取り出し，WordNet を利用してルールの一般化を行っている [5].

2.4 節で紹介した INFOS[74] でも，WordNet を利用している. INFOS では，WordNet に登録されている動詞と名詞を利用して，メモリ階層を構築する. ニュース記事中にある全ての名詞と動詞の意味定義に基づいて記事にインデックスをつけ，語の多義性を解消してその関連する概念を決定した後で階層構造化する. メモリ階層のノードは概念を表し，ノルムとケース，子ノードへのポインタを持つ. ケースはその概念に分類される記事を指す. この階層構造は ISA 階層と同様であり，親ノードのノルムは子ノードに継承される. INFOS では，まず最初に全体概念として入力された記事がユーザにとって興味深いものかどうかを GHC と呼ばれる判定方法で決定し，どちらとも判断しかねたものについてはメモリ階層中の各ノードとの適合度を計算し，最も適合する概念に属しているケース（記事）を用いて再び GHC を実行して判定する. このメモリ階層を利用した分類は CBR(Cased-Based Reasoning) の一種であり，これを組み合わせることによる分類精度の向上が報告されて

いる.

西田らは，弱構造化オントロジーを，文書ディレクトリのブラウジングやインターネットからの情報収集など，様々な情報活動に利用する事を提案している [46, 47, 58, 66, 82]. 例えば情報収集において，ユーザが入力したキーワードの類義語・関連語をオントロジーから探しだし，関連語を含む文書も収集対象とする事を提案している [46]. この様な，シソーラスを利用したいわゆるクエリー拡張（Query Expansion）に関する研究は他にも行われている. 藤崎らの提案するキー概念に基づく検索では，ユーザが入力したキーワードを表層的なシンボルとして用いるのではなく，それの指す意味・概念を特定し，概念をキーとした検索を行う [24, 25]. 具体的には，キーワード（K_{w_0}）が図 4.2(a) の様に異表記同義を持つ場合，検索式として $K_{w_0} + K_{w_1} + \cdots + K_{w_m}$ を用いる. 一方，図 4.2(b) の様に K_{w_0} が同表記意義を持つ場合は，着目する概念 C_0 以外の各概念に対応するキーワード集合 $T_1, \cdots, T_n$ を除去し，$T_0 \cdot \overline{T_1} \cdot \ldots \cdot \overline{T_n}$ を検索式とする.

関連する研究として，Ho らはラフ集合を用いて情報検索を行っており，共起関係を用いて同値とみなす集合 tolerance class を決定しているが，これをシソーラスを用いて計算する可能性も示唆している [36]. ここで紹介したクエリー拡張は，2.5 節の AI アプローチによるクエリー学習の研究がキーワードをシンボルとして扱っていたのと比較して非常に対象的な手法であると言えよう.

4.3 テキストマイニング

企業には，多くの情報・知識が文書の形で存在する. これは，文書が人間にとって最も扱いやすい表現形式の一つであるためだけでなく，ノウハウなど，数値などで表現しがたい情報も多く存在する事も一因である. この様な文書情報を有効利用するために，文書から情報・知識を取り出す技術をテキストマイニングと呼ぶ. 従来は，人手で索引づけ，分類などを行う手法が主流であったが，計算機の進歩により高度な処理を高速で行う事ができるようになって来た事から，今後は自然言語処理が重要な役割を果たす事が期待されている [75, 80]. また，今後は計算機で管理・利用する事が前提となるため，XML などの構造化文書による記述の統一も普及する事は間違いないであろう.

以下では，個々の要素技術ではなく，今後テキストマイニングの適用が期待されるいくつかの分野について概略する.

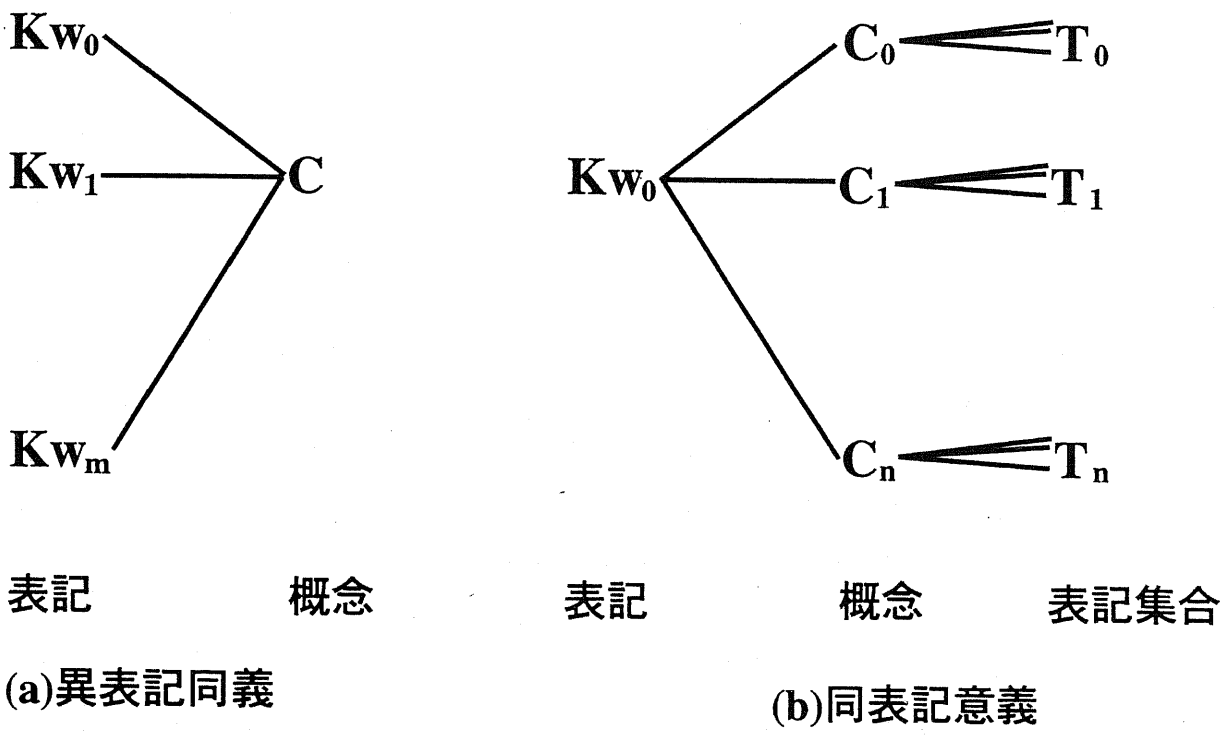


図 4.2: キー概念によるキーワード拡張

4.3.1 要約・抄録の生成

大量の文書が存在する場合，それらの中から自分にとって必要な情報を探し出す事は容易ではない．文書に付加されているタイトルやキーワードは，自分の興味に関連がありそうな文書を限定するには有効であるが，必要かどうかは実際に読んでみないとわからない事が多い．この様な場合，文書の閲覧性を高めるために，内容を簡潔にまとめた要約・抄録を自動生成する技術が有効であると考えられる．

現在，文書からの要約・抄録自動生成に関する研究は，以下に大別できるとされる．

1. キーワードに基づく方法

2. スクリプトなどの世界知識を利用する方法 [71]

3. 文の言い回しなどの表層的知識を手がかりとする手法 [118, 123, 124]

最初の手法は，文書中からキーワードを含む文を抽出し，羅列する手法であり，非常に容易に実現できるが，あくまで文の羅列であって文章として理解しづらい事，必ずしも重要文が抽出されない事などが指摘されている．また，二つ目の手法では対象分野に依存した知識が必要であるため，用途が限定されると言った問題点がある．最後の手法は，人間が文章を書く時，言い換える場合には「すなわち」，対比したい場合は「一方」などの言い回しを用いて文章間の関係を表す事実などを利用しており，文書の内容に比較的依存しないで利用できると言うメリットがある．住田らは，この様な修辞構造と，段落などの書式構造を組み合わせて文書を構造化し，抄録を自動生成する手法を提案している [118]．また，竹下らは重要文を抽出して抄録を生成するのではなく，話題構造を抽出，提示する事によりユーザの文書閲覧を助ける手法を提案している [123, 124]．

4.3.2 既存のテキスト情報源の活用

今までに発生した質問・問題と，それに関する回答・解決策をまとめたものを FAQ (Frequently Asked Question) といい，オンラインニュースの過去の記事などは FAQ として整備され，WWW などで公開されている．また，企業のユーザサポート部門などでも，過去のトラブルなどが FAQ の形で整理され，活用されている．しかし，大量の FAQ が存在する場合，ある問題に関連する記事を検索する事はなかなか容易ではない．この様なデータベースの検索を行うシステムは非常に便利であり，企業においてもユーザサポートに要するコストの削減にもつながる．

FAQ Finder[10, 11] はまさにこの様な用途のために開発されたシステムであり，入力された質問のタイプ，特徴ベクトルを元に，その質問と最も近いと思われる質問を探し出し，

それに対する回答を提示する．この研究においても，類義語関係を認識し，検索精度を高めるために WordNet を利用している．

また，yahoo[131] などの様に，文書をカテゴリ毎に分類して管理している場合も多い．この様にカテゴリ分けして整理されたものを文書ディレクトリと呼ぶ．文書ディレクトリは人手で作成される事が多く，分類自体は比較的高精度で行われているものが多いが，ユーザのイメージするディレクトリ構造と，実際の文書ディレクトリが異なる事により，文書検索が効率よく行えない問題が指摘されている．これは上述した語彙問題や，対象文書に関する知識状態はユーザの習熟度によって異なる事などにより生じる．この様な問題に対し，Index Navigator[94] はユーザが逐次入力したキーワード列を元にユーザの検索意図（興味あるカテゴリ）を推測し，それに対応するインデックス（キーワード）を提示する事によって，ユーザに自分の概念とディレクトリ構造との対応を学ばせるアプローチをとっている．

4.3.3 文書整理：文書ディレクトリの自動生成

上述したように，従来文書ディレクトリは人手で作成する場合が多く，コストと時間のかかる作業であったが，これを自動化しようという研究も行われている [60, 62, 132]．具体的には，パターン認識の分野で研究されて来たクラスタリング（Clustering）やカテゴライゼーション（Categorization）の技術を利用する．ここで，カテゴライゼーションは既存のカテゴリに文書を割り振っていく手法であり，クラスタリングは文書の類似性のみを基準として文書をいくつかのグループに分ける手法である．Lam らは，Bayesian Network[97] を用いてカテゴライゼーションを行っている [62]．姚らは特徴ベクトルを用いて階層的なクラスタリングを行っている [132]．

クラスタリングの場合には，各クラスタがどの様な話題を表しているかが明らかではない．このため，4.2.1 節で触れた，概念ベースを用いたクラスタ特長語抽出 [60] の様に，生成したクラスタに何らかの索引づけをする事も研究されている [57]．これに関連して，大澤らが提案している KeyGraph は，文書を代表するキーワードを抽出する手法である [92]．KeyGraph は，キーワードの文書中における共起関係を元に作成されるグラフ構造であり，これを論文に関する意味の構造物とみなした時に，柱となる単語をキーワードとして抽出する手法である．単に頻度や $TFIDF$ の高い単語に比べ，文書の意味を十分に反映した単語を抽出できる事が報告されている．また，上述の抄録と同様，文書の閲覧容易性を高める用途も期待できよう．

また，生成された文書ディレクトリを効率よく探索（ブラウジング）する手法も重要であり，これには 2.6 節で紹介した視覚化技術を利用する事ができる．

4.4 個人レベルでの文書整理の支援

文書ディレクトリ生成は，情報提供のための文書整理技術と捉える事ができる．すなわち，整理する者（ディレクトリ生成者）と，利用する者（情報検索・利用者）が異なるのが通常である．

しかし，整理整頓を行った者が，必要なものを最も容易に探し出せる様に，整理は対象に対する理解を深める．また，生成したグループ（カテゴリ）にふさわしいキーワードを考える事も，考えを整理したり，アイデアを生み出す上で有効な刺激となるであろう．WWW上のホームページやCD-ROMといった形態で，大量の文書情報を容易に収集する事ができるようになった現代の我々にとって，個人レベルで入手情報の整理を支援し，収集情報の理解・有効活用を可能とする支援ツールは今後必要不可欠な存在となるであろう．すなわち個人レベルでの作業プロセスの中心は，関連する情報の収集や文書の編集といったプロセスから，大量のドキュメントをいかに読みこなし，全体を把握し，目的にあった情報，アイデアを取り出すか，に移ったといってよいだろう．

この様な考えに基づき，個人レベルでの情報整理を支援するシステムの開発も行われている [66, 87, 126]．PAN-WWWは，カードをメタファとした情報整理支援ツールである [87]．また，CM-2[1]は前述の弱構造化の考えに基づく**連想構造**と呼ぶデータ構造を用いて雑多な情報をゆるやかに関連づけ，個人的な視点に基づく情報検索・統合を実現している [58, 66]．

個人レベルの情報整理においては，図解などを用いて試行錯誤しながら整理を行う事が有効であると考えられるが，これは3節で紹介した発想支援，特に収束的思考支援システムで重要視されて来た考えである．従って，ナレッジマネジメントの文脈からの発想支援研究も今後ますます盛んになっていくと考えられる．杉山らが提案する創発メディア環境 [112] も，ナレッジマネジメントの考えに近いと考えられる．

個人レベルの情報支援に関する上記の研究においては，メモの様な細切れの情報を対象とし，ユーザの手でそれらを入力していく形態を想定しているものが多い．これは，KJ法に代表される発想法・発想支援システムにおいても同様である．しかしこの様なメモは，収集された文書情報をユーザが実際に読み，理解した上で作成できるわけであり，現実には収集された文書を読みこなす事自体が，文書量の増加に伴い困難になって来ている．

本研究の目的は，この様な収集された大量文書の読解プロセスの支援にある．すなわち，個人レベルで収集した文書情報を読みこなす過程を支援するために，情報整理の考えを用いる点が既存の研究との大きな違いであると言える．

[1]現在はCoMeMoに名称が変更されている．

発想法・発想支援に関する研究で見て来た通り，論文やニュースなどを読んで考えをまとめたり，新たなアイデアを得ようとする場合には，図解を編集しながら考えを整理することが有効である．特に，扱う文書量が増えて来た場合には，図解作成や情報検索などの，計算機が得意とする能力を利用する事により，ユーザはより知的な作業，すなわち思考に没頭することができると考えられる．これが，我々の目指す文書整理支援システム，「文書の整理を通じた熟読を支援するシステム」の概要である．

具体的には，以下の3プロセスの繰り返しにより文書整理プロセスは進んでいくとする．

読書 今までに読んだドキュメントや，頭の中の知識との関連を意識しながらドキュメントを読む．

図解作成 今まで読み進めて来たドキュメント群から図解（**局所図解**）を作成し，視点を整理する (図 4.3(b)(d))．

検索 得られた視点をもとに，次に読むべき文書を決定する (図 4.3(c))．

すなわち，大量の文書を一度にまとめて読もうとするのではなく，部分的に，十数文書くらいずつ読んでは考えを整理し，また読み進めていくといったプロセスを繰り返し，漸進的に考えをまとめていく．従って図解作成プロセスでは，KJ 法などのように全ての情報をまとめて考慮して図解を作成するわけではなく，現在までに読んだ文書のみを対象とする．その意味で，我々の図解を**局所図解**と呼んで区別している．

KJ 法に代表される収束的思考プロセスでは，断片的な情報をまとめあげるのに図解を利用しているのに対し，本研究で対象とする文書は，それぞれの持つ情報量が大きいため，図解編集前に全ての文書に目を通しておくというのは非現実的である．すなわち，収集された大量の文書を片端から読んでいくことは，考えただけで気の滅入る作業であり，精神的な苦痛を感じるだけでなく，一つ一つを雑に読んでしまうと考えられる．図解編集は，考えをまとめるだけでなく，この様な読解作業を助ける上でも重要な役割を果たすとしているのが，我々の提案する文書整理支援システムの特徴である．

局所図解には，ユーザのその時点での視点が反映されていると考えることができるが，この図解から視点情報を抽出することができれば，システムによる能動的支援を行う事ができ，システムの有効性をより高めることができよう．

以上より，文書整理支援システムには，次の二つの要素技術が必要となる．

- 思考整理，読解のガイドとしての図解編集に関する技術
- ユーザの視点に関する情報を抽出し，利用する技術

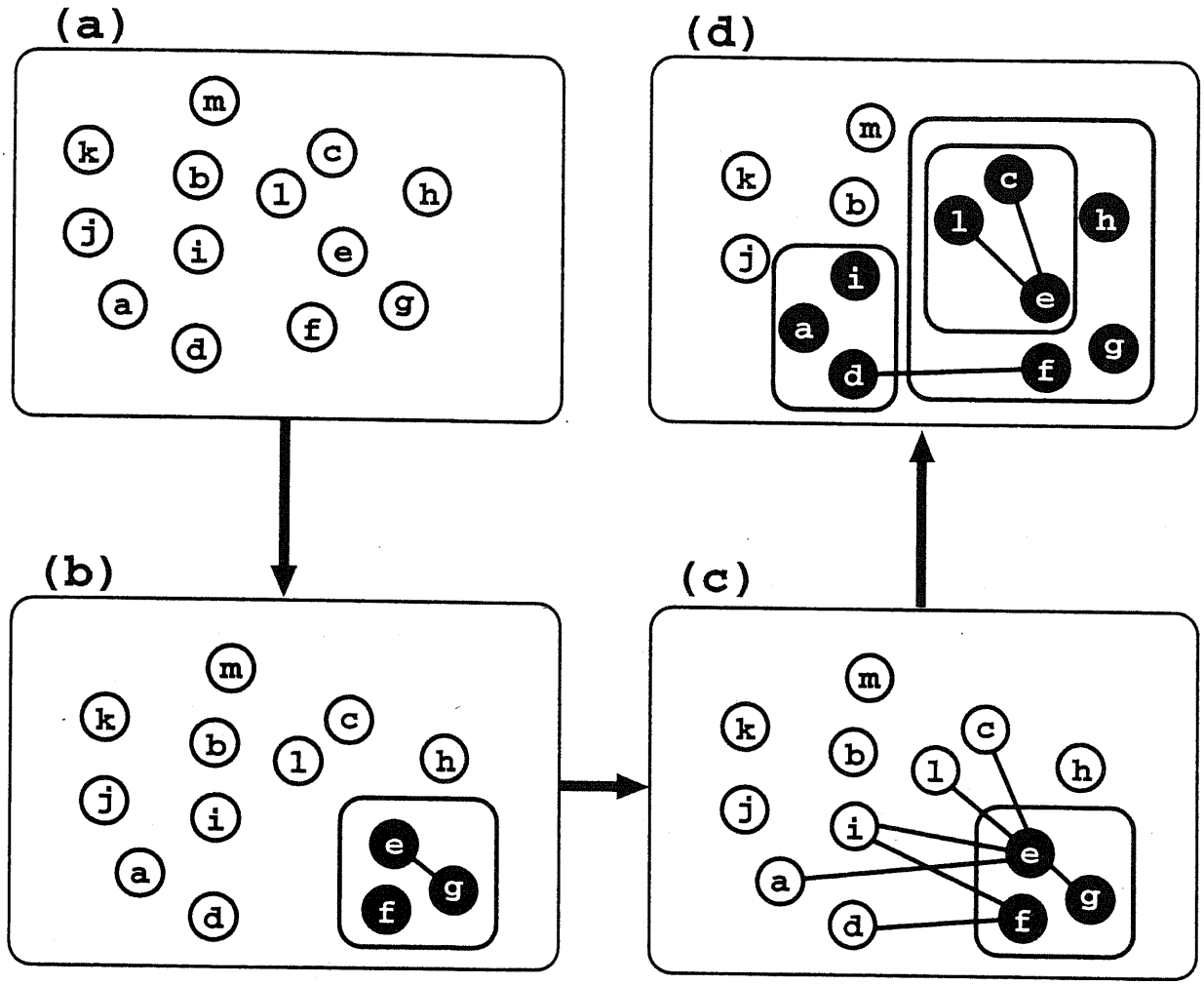


図 4.3: 文書整理プロセス

後者に関して，我々は視点情報を扱うようにベクトル空間モデルを拡張した **Fish Eye マッチング**と呼ぶ，新しい動的文書マッチング機構を提案している．また，前者については，この Fish Eye マッチングを基盤技術として用いた文書整理支援システム Fish View を開発した．Fish View においては，ユーザが作成する局所図解から視点情報を抽出し，利用できるだけでなく，最終的な図解を HTML 化して文書ディレクトリを生成することができ，整理結果をサーベイなどとして有効に活用する事ができる．Fish Eye マッチングについては 5 章，Fish View については 6 章で詳しく触れる．

Chapter 5

Fish Eye マッチング：概念体系に基づく視点を考慮した文書マッチング機構

2.7 節で紹介したベクトル空間モデル（VSM）でユーザの視点を扱う場合には，2.7.1 節に示した適合フィードバックなどにより，クエリーベクトルとして，すなわち各成分に対する重みとしてユーザの視点を反映する手法が一般的である．

しかしこの方法ではユーザの興味がブラックボックスで扱われるため，自分の興味・視点が何であるかをユーザが明示的に知る事は難しい．また，通常のベクトル空間モデルでは，各軸（単語）間の直交性が仮定されるため，ある視点から見て共通の特徴とみなせる単語群も，常に別々に扱われてしまうという問題がある．例えば,「車」という単語を含む記事と「自転車」という単語をそれぞれ特徴として持つドキュメントがある場合，乗物に関する記事という点では関連があるにも関わらず，ベクトル空間上では両ドキュメント間の関係を見出す事はできない．2.7.2 節で紹介した LSI なども，この様な語彙問題を扱う一つの有望なアプローチではあるが，ブラックボックスな操作であることには変わりない．

これに対し，我々が提案する **Fish Eye マッチング**では，電子化辞書の概念体系から抽出した，意味の類似した単語集合（意味グループ）を元に，視点に合わせた特徴を動的に生成してベクトルを構成し，マッチングを行う．背景知識として概念体系を用いることにより，視点に合わせた粒度で文書間の関係を捉えられるだけでなく，ユーザにとってブラックボックスでない操作を実現している．

本章では，始めに 5.1 節にて Fish Eye マッチングの定義について示した後，5.2 節で Fish Eye ベクトル生成演算子の特徴を示す予備実験を行った結果を記す．Fish Eye マッチングではユーザの視点情報を引数としてベクトル空間を動的に構築するが，視点情報の自動抽出アルゴリズムについては 5.3 節にその概要を示し，評価実験については 5.4 節にま

とめる．

当初，Fish Eye マッチングは英語文書を対象としてシステムを構築したが，猪股らによって日本語文書へ対応できる様になった．日本語文書への対応方法および，日本語文書を対象とした評価実験については 5.5 節にまとめる．最後に，関連研究との比較検討は 5.6 節で行う．

5.1 Fish Eye マッチングの定義

Fish Eye マッチングでは，ユーザの視点を反映したドキュメントベクトルを生成するために，通常の特徴ベクトルに対して単語の意味的関係を考慮した拡張を施して特徴ベクトルを動的に生成する．すなわち，ドキュメント中に含まれる単語をそのまま特徴として用いた**基本特徴ベクトル（Basic Feature Vector）**から，ユーザの視点において同様とみなせる単語は一つの特徴に縮退したり，その視点に関係のある単語のみを選択したりといった操作を加えることにより，ユーザの視点・興味が変化する毎に，それに応じた特徴ベクトル空間を動的に再構成する（図 5.1）．この，ユーザの視点を反映して動的に生成される特徴ベクトルのことを **Fish Eye ベクトル**と呼ぶ．

視点については，「そのドキュメントを読む際に仮定する，その属する話題，分野」であると考える．一つのドキュメントの属する話題が一意に決まるということは稀であり，たいていは複数の視点から読む事が可能であると考えられる．また，ある種の話題間には包含関係が存在するため，同じ角度からの視点であっても，専門的な狭い話題として捉えたり，より一般的な，広い枠組で捉えたりする事も可能である．

話題としてどういったものを考えるかであるが，本研究では EDR 電子化辞書の概念体系を利用する．すなわち，全ての話題には対応する概念が電子化辞書中に存在するものとし，その概念の範疇に含まれる単語を，特徴を構成する要素として採用する．

例として，図 5.2 について考えてみよう．ここでは 5 つのキーワードが存在しているが，概念体系に基づいて階層化すると，‘apple’ と ‘lemon’ は ‘fruits’（果物）という概念の範疇に含まれ，これに ‘tomato’ を加えた三つの単語は ‘vegetable’（野菜）という概念に含まれる．これらとは別に，‘bicycle’ と ‘car’ は ‘vehicle’（乗り物）という概念の範疇に含まれている[1]．ここで，野菜に関する話題か，あるいは乗物に関する話題のどちらに属するかが区別できればいい程度の粗い視点であれば，‘apple’，‘lemon’，‘tomato’ を一つの特徴 ‘vegetable’ に縮退し，‘car’，‘bicycle’ も一つの特徴 ‘vehicle’ に**縮退**して Fish Eye ベクト

[1]ここで例示している概念体系は例のためのものであり，EDR 概念体系辞書から実際に抽出したものではない．

ルを生成すれば良い．反対に，野菜の話題に集中し，野菜と果物に関する話題を区別して扱いたい場合には，'apple'，'lemon' を一つの特徴 'fruits' に縮退し，これと 'tomato' を直交する特徴として用いれば良い．

話題に関係ある特徴のみをそのままとりだし，関係ない特徴を切り捨てる事は，概念体系のうちの興味ある部分だけをルーペで**拡大**してみることに相当する．従って，Fish Eye ベクトルは，基本特徴ベクトルに**縮退（Shrink）**と**拡大（Magnify）**という二つの操作を施して計算される．

単語集合を $W=\{w_1,w_2,\ldots,w_n\}$ とすると，これから Magnify，Shrink の各操作によって得られる Fish Eye ベクトルの特徴集合 $S(g_1,\ldots,g_n|W)$，$M(g_1,\ldots,g_n|W)$ はそれぞれ次式のように定義される．

$$S(g_1,\cdots,g_n|W) = \{f_i|f_i=\{w_j|w_j\in g_i\wedge W\},1\leq i\leq n\} \tag{5.1}$$

$$M(g_1,\cdots,g_n|W) = \{w_i|w_i\in(g_1\vee\cdots\vee g_n)\wedge W, \quad 1\leq i\leq|(g_1\vee\cdots\vee g_n)\wedge W|\ \} \tag{5.2}$$

$$\tilde{S}(g_1,\cdots,g_n|W) = S(g_1,\cdots,g_n|W)\cup\overline{M}(g_1,\cdots,g_n|W) \tag{5.3}$$

$$\overline{M}(g_1,\cdots,g_n|W) = \{w_i|w_i\in W\wedge\neg g_1\wedge\cdots\wedge\neg g_n, \quad 1\leq i\leq|W\wedge\neg g_1\wedge\cdots\wedge\neg g_n|\} \tag{5.4}$$

ここで，g_i はある話題に関する単語のグループ（**意味グループ**）を表しており，特徴生成の際に背景知識の役割を果たすといえる．意味グループの計算方法については 5.1.1 節に記す．また，$\tilde{S}(g_1,\cdots,g_n|W)$ は，Shrink 操作の対象とならなかった W 中の単語についてはそのまま特徴として用いる操作であり，5.4 節以降で利用するため便宜上定義した．前述の例に関する各操作について，図 5.2 に記してある．この時，$W=\{apple,lemon,tomato,bicycle,car\}$ としている．

以上の操作により求められた特徴に基づいて，基本特徴ベクトル $O_v(v_1,\cdots,v_n)$ から Fish Eye ベクトルを計算する際には，特徴 f に属する単語の，O_v における値の総和を対応する値 v_f とする．ここで，O_v の各要素の値については，TFIDF[78] などにより求められているものとする．

$$v_f=\sum_{w_i\in f}v_i \tag{5.5}$$

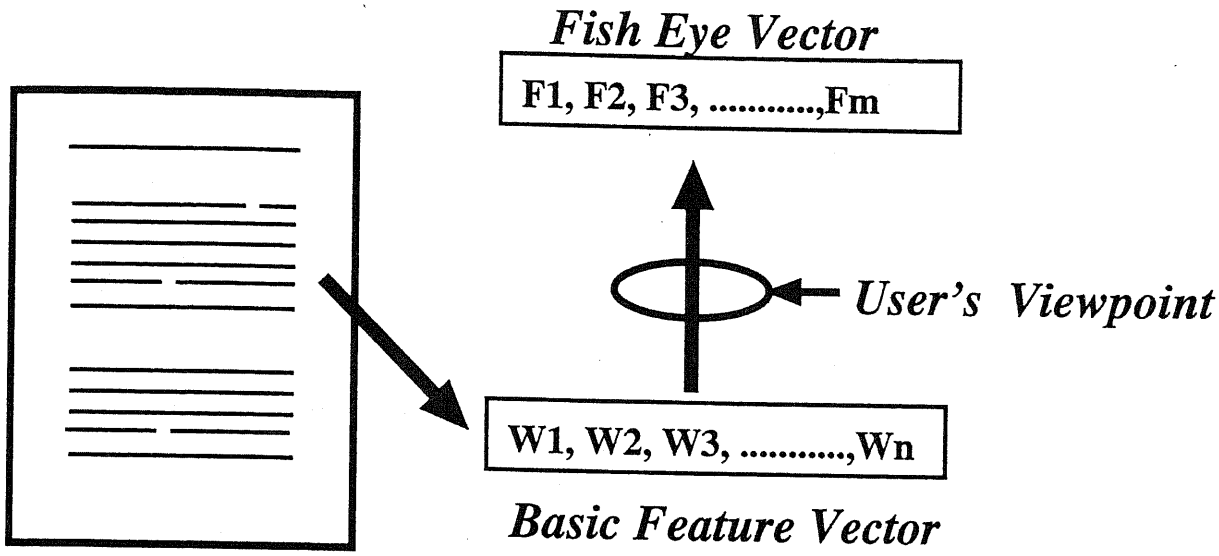


図 5.1: Fish Eye ベクトル生成のイメージ

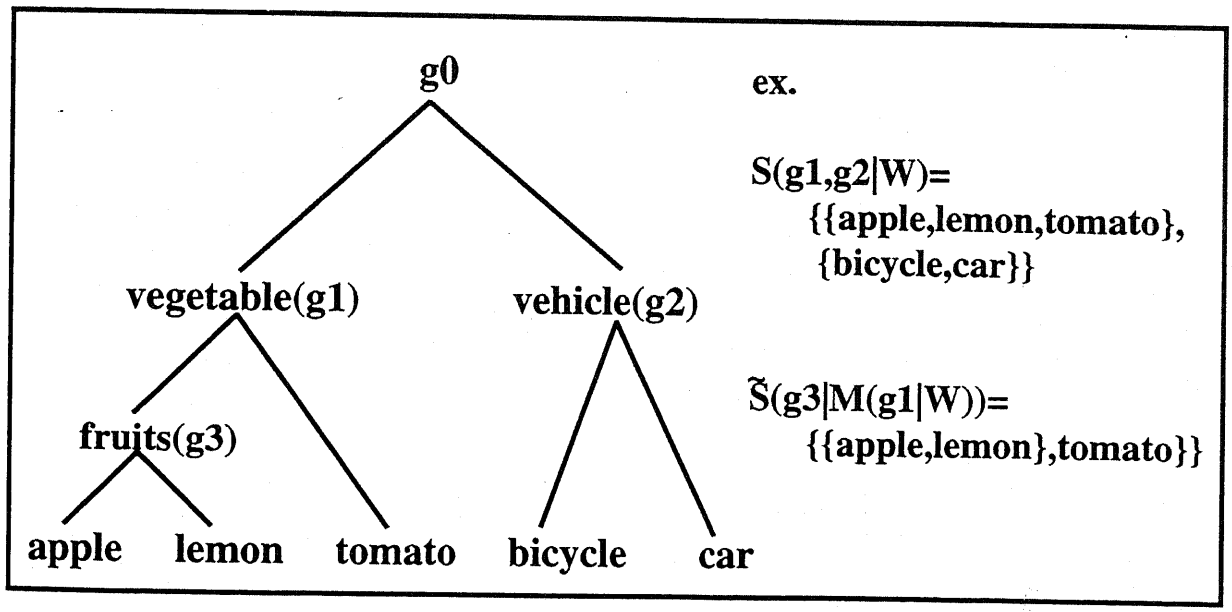


図 5.2: 概念グループの例

また，Fish Eye ベクトルに基づきマッチングを行う際には，通常の特徴ベクトルにおける場合と同様の指標を用いることができる．例えば次式の様に，ドキュメント d_i, d_j 間の類似度 $Sim(d_i, d_j)$ は両ベクトルの内積に基づいて，以下のように定義できる．

$$Sim(d_i, d_j) = \frac{1}{2}\left(1 + \frac{\sum_{k=1}^{n} f_{ik} \cdot f_{jk}}{Mag(d_i) \cdot Mag(d_j)}\right), \quad 0 \leq Sim(d_i, d_j) \leq 1 \tag{5.6}$$

$$Mag(d) = \sqrt{\sum_{i=1}^{n} f_i^2} \tag{5.7}$$

5.1.1 概念体系辞書からの意味グループ計算

本研究で使用している EDR 電子化辞書では，概念に関する辞書として，概念間の上下関係を規定する概念体系辞書と，概念間のそれ以外の関係を規定する概念記述辞書，および各概念を言葉で説明する概念見出し辞書が用意されている．このうち，本研究では概念体系辞書から意味グループを計算している．概念体系辞書によって整理される概念体系は，一般のシソーラスなどと同様に，各概念をノードとし，上下関係のあるものをノードで結んだ木構造の形に体系化される[2]．

前述したように，Fish Eye ベクトル生成演算子の引数として指定できる話題は，対応する概念が EDR 電子化辞書中に存在するとする．基本特徴ベクトルを構成する全単語について，概念体系を下層から上層へ辿りつつ，各概念ノードに対応する単語集合を求め，単語数が 2 から 256 の範囲に納まる概念ノードを意味グループとして用いている．この処理により，どの意味グループにも属さない単語が存在する可能性があるが，この場合には例外的に，単語数が 1 の意味グループとして扱う．

電子化辞書に記述された概念体系は，我々の常識を体系化したものと考えることができるので，話題に応じた特徴を生成するために単語をグループ化する際の制約として適していると考えられる．

5.2 予備実験：Shrink と Magnify の特徴

Fish Eye ベクトル生成演算子 Shrink, Magnify の定義（式 (5.1),(5.2)）より，両演算子は表裏の関係にあるといえる．すなわち，同じ意味グループに対し，ドキュメント間の大まかな関係を概略視するには Shrink，ある視点・興味のみについて詳細に比較したい場合

[2]多重継承を許すため，厳密には木構造にならない．

には Magnify を行えばよいと考えられる．このことを検証するための予備実験を行ったのでここに記す．

実験では，オンラインニュース 3 記事について，同じ意味グループを指定して Shrink および Magnify 操作を行い，各文書間の類似度を計算した．ここで，類似度の計算は式 (5.6) ではなく，内積の値をそのまま用いている．これは，今回の実験では負の重みを持つ特徴がなく，内積の値が負となる事がないため，式 (5.6) の様な $[0, 1]$ の範囲に収める補正を行う必要がないと判断したためである．

以下は，実験に用いた 3 記事の見出しである．記事 a, b, c は全て通信関連の内容であり，さらに記事 b, c は行政関連の話題も含んでいる．

a. Kobe Steel announces entry into international telecom field.

b. Posts Ministry Oks linking home phones,. trunk lines by radio.

c. IDC, others to join Asia Multimedia Forum led by NTT.

これら 3 つの記事に関して，まず最初に通信関連の話題に焦点を当てるため，表 5.1 に示されている意味グループのうち 1fa130 から 2f295a までを指定して Magify, Shrink をそれぞれ行った結果を表 5.2 に示す．これより，どの記事間も，基本特徴ベクトルから類似度を求めた場合と比べて類似度が上昇していることがわかる．記事 a と記事 c に関しては，元から類似度が高かったため，変形後もそれほど類似度が上昇しなかったのであろう．

次に，行政関連の意味グループ ed85d も追加して，同様の実験を行った結果を表 5.3 に示す．表 5.2 と比較すると，記事 b と c に関して，Shrink 操作を施した場合は類似度が唯一上昇しているが，反対に Magnify 操作では著しく低下している．

これは，Shrink では関連する単語を同じ特徴として扱うのに対し，Magnify では関連する単語同士を直交する特徴として扱うという違いが反映されたものと考えることができる．実際，記事 b と c は両方とも行政関連ではあるものの，行政関連の視点から比較した場合にはそれほど類似していない．これより，Shrink, Magnify の両操作に関して，以下の様な違いがあることが確かめられた．

Magnify 詳細な関係の発見

Shrink 大まかな関係の概略視

表 5.1: 実験に用いた意味グループの例

概念 ID	説明 (概念見出し)	所属単語
1fa130	通信手段	networ, telegr
443d4a	cooperator	access, connec
3cfb69	所縁	connec, link
2f352d	その他 (情報処理)	lan, servic, system
2f26b0	システム	isdn, teleco
443cde	audio commun. device	line,teleph
3bdeeb	電話器	phone, teleph
2f295a	通信装置	ethern, fax, ntt
ed85d	行政組織	diet, floor, custom,law, minist

表 5.2: 通信関連に視点を当てた実験結果

	Original	Magnify	Shrink
Sim(a,b)	0.4067	**0.7557**	**0.7493**
Sim(a,c)	0.5645	0.5839	**0.6251**
Sim(b,c)	0.4187	**0.6091**	**0.6371**

表 5.3: 通信＋行政関連に視点を当てた実験結果

	Original	Magnify	Shrink
Sim(a,b)	0.4067	0.5804	0.6397
Sim(a,c)	0.5645	0.5653	0.5903
Sim(b,c)	0.4187	**0.4529**	**0.6849**

5.3 視点を表す意味グループの抽出

Fish Eye マッチングでは，演算子の引数として指定する意味グループの選択が重要である．また，ユーザにとっては一つの視点であっても，対応する意味グループは複数存在するのが普通であり，複数の意味グループを指定する必要がある．しかし，5.1.1 節によって EDR 辞書から求められる意味グループ数は膨大であり，人手で適切な質・量の意味グループを選択し，指定することは困難であると考えられる．従って，ユーザの興味を反映した文書の分類結果や図解などから，視点を表す意味グループを抽出できることが望ましい．また，抽出された意味グループをユーザに提示することにより，今まで意識していなかった視点に気づいたり，漠然としていた考えが明確になるなどの効果も期待できる．この様な視点の外化も，システムによる能動的支援の一環として重要であると考えられる．

現在我々は，Shrink 操作を元にしたグリーディな意味グループ抽出アルゴリズムを用いている．本節ではアルゴリズムの詳細について説明し，評価実験については次節で報告する．

ここで，ユーザの興味・視点に関連するテキスト集合を D_P，関連しないテキスト集合を D_N とする．これらは，今までに読んだドキュメント集合に対するユーザの分類といった形で得ることができる．また，初期化として，基本特徴ベクトル空間を構成する全単語をリスト $Wlist$ に登録する．以下では，v_{ji} はドキュメント d_j に対する基本特徴ベクトル O_{d_j} 中の，単語 $word_i$ に対応する要素の値とする．

1. $Wlist$ 中の各単語 $word_i$ について，適合フィードバック（式 (2.6)）と同様の次式により重み w_i を計算．

$$w_i = \alpha \frac{1}{|D_P|} \sum_{d_j \in D_P} v_{ji} - \frac{1}{|D_N|} \sum_{d_j \in D_N} v_{ji} \tag{5.8}$$

2. $Wlist$ 中より，正かつ最大の重み w_k を持つ単語 $word_k$ を取り出す．なければ終了．

3. 意味グループ集合 $G_k = \{g_i \mid (word_k \in g_i) \wedge (\forall word_j \in g_i, word_j \in Wlist) \wedge (\forall word_j \in g_i, w_j \geq 0)\}$ を求める．これは，$word_k$ を含むグループのうち，すでに抽出された他のグループと単語を一つも共有せず，かつグループ中に負の重みを持つ単語を含まないものを選択している．

4. （3）で求めた G_k 中の各グループ g_i について，重み W_{g_i} を次式にしたがって計算する．$G_k = \emptyset$ の場合には（6）へ．

$$W_{g_i} = \frac{1}{|g_i|} \sum_{word_j \in g_i} w_j \tag{5.9}$$

5. G_k 中で，重みが最大のグループ g_l を抽出.
   $\{Wlist\} = \{Wlist\} - \{word_i | word_i \in g_l\}$ として（2）へ.

6. $\{Wlist\} = \{Wlist\} - \{word_k\}$ として（2）へ.

ここで式 (5.9) により，各グループに属する単語の重みの平均をグループの評価値（重み）としている．これは，ベクトル空間モデルにおける「ユーザの興味に従った分類において，有効な指標となる重要な単語の重みは大きくなる」という仮定を拡張した,「ユーザの興味を表す概念には，重要な（重みの大きい）単語が多く属している」との仮定に基づいている．すなわち平均値をとることにより，重要でない（重みの小さい）単語が多く集まった意味グループより，少数でも重要な単語のみが集まった意味グループを優先して抽出している．

5.4 評価実験

5.4.1 実験設定

本稿で提案した Fish Eye マッチングについて評価実験を行った結果を記す．この実験の目的は，Fish Eye マッチングが，通常の特徴ベクトルと同程度の検索精度を維持しつつ，視点の外化を行える事の検証である．このため，インターネット上で公開されている，英語で記述されたオンラインニュース記事の中から，医学に関する記事を検索するタスクを行い，その適合率（Precision）（表 2.3 参照）について調べるとともに，抽出された意味グループについて考察した[3]．なお，ニュース記事が医学に関するかどうかの判定基準としては，その記事が公開されているニュースサイトにおける分類に従った．

検索方法であるが，正例および負例となるニュース記事をいくつか与え，5.3 節に示したアルゴリズムにおいて抽出された意味グループ集合を引数として $\tilde{S}(g_1, \cdots, g_n | W)$ 操作を行い， Fish Eye ベクトルを生成するとともに，抽出段階で得られた各グループに対する重み W_{g_i}（単語は w_i）を各要素の値としたものをクエリーベクトル q とし，q と各文書 d_i の類似度 $Sim(q, d_i)$ を式 (5.6) より計算し，値が上位のものから順にユーザへ返すとした．

対象としたニュース記事の総数は 218 であり，そのうち該当文書である医学に関するニュース記事は 100 含まれている．これらの記事に含まれる単語からストップワードを除去し，EDR 英単語辞書に名詞の意味で登録されている単語を基本特徴ベクトルの構成要素

[3]適合率と再現率の両指標を用いるのが一般的であるが，本章の目的は通常の適合フィードバックとの検索精度の比較であり，検索文書数に大きく左右される再現率での比較は不要であると判断した．

として抽出したところ，1588 単語が抽出された．さらに，得られた単語集合から 5.1.1 節で示した様にして意味グループを計算したところ，655 グループが得られた．

5.4.2 実験結果 (1)：正例のみを与えた場合

正例のみ 5 から 15 記事まで変化させて与えた場合について，検索された（ユーザに返された）文書数と適合率との関係を図 5.4 に示す．グラフ中に示してある直線は，各条件における平均値を示している．ここで，横軸は検索された文書数，縦軸は適合率をとってある．ここで，“pxny(Fish-eye)” は，正例 x，負例 y 文書を与えて Fish Eye マッチングを行った結果である事を示している．また，比較のため，正例 10 記事，負例なしの条件で通常の適合フィードバックを用いてクエリーベクトルを作成し，検索した場合を “p10n0(normal rf)” として示している．

また，視点として抽出された意味グループ（今後，視点意味グループと呼ぶ）の特徴として，平均グループ数および 1 グループあたりの所属単語数について表 5.4 に記す．システムの応答速度に関してであるが，Sun Ultra2（メモリ 320MB）上で実験を行った結果，意味グループの抽出および文献検索までを含めても 5 秒以内であった．

図 5.3 より，通常の適合フィードバックより適合率は若干落ちるものの，正例を増やすにつれて適合率は上昇し，適合フィードバックの性能に近づいていくことがわかる．この時，正例の増加に連れて抽出される意味グループ数は増加していくのに対し，意味グループ当たりの平均所属単語数は減少していくことが表 5.4 より確認できる．この事は，正例の増加に伴い，5.3 節のステップ (1) によって計算される単語の重みがユーザの興味を正しく反映するようになるだけでなく，抽出される意味グループが，一般化しすぎたものから適度な粒度のものに分割され，精度向上に貢献する事を表していると考えられる．これより，概念体系から抽出された意味グループの，特徴生成に関する制約としての妥当性，および 5.3 節で示した視点意味グループ抽出アルゴリズムの有効性が確認できる．

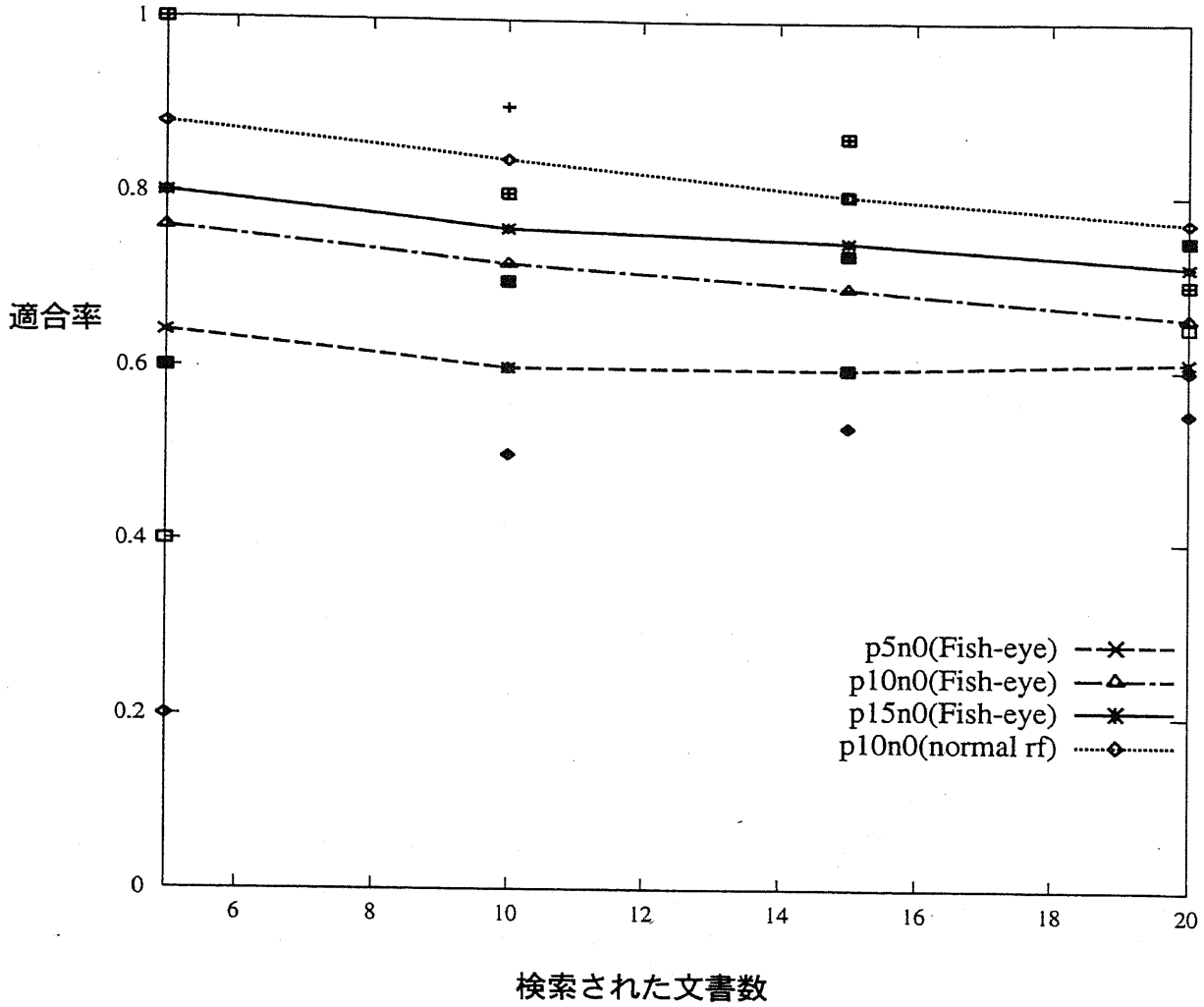


図 5.3: 実験結果 (1)：正例 5–15, 負例 0

表 5.4: 視点意味グループの特徴：正例のみ与えた場合

正例数	平均グループ数	グループあたり単語数
5	23.4	9.2
10	32.8	8
15	38.4	7.2

5.4.3 実験結果 (2)：正負例ともに与えた場合

前節の実験において，通常の適合フィードバックより性能が劣った理由として主に考えられるのは，次の二点である．

1. 不適切な意味グループを特徴として抽出してしまったための誤差（意味グループによるノイズ）

2. 意味グループ内に，不適切な単語が含まれてしまったための誤差（単語によるノイズ）．

（1）に関しては，EDR 電子化辞書から抽出された意味グループ中に適切なものが存在しなかった場合も考えられるが，それ以外の要因として，単語を意味グループに縮退する際に，一般化へのバイアスの方が強いことが考えられる．そこで，正例は一定数（10）に固定し，負例を 0 から 10 記事まで変化させて与えて同様の実験を行ったところ，図 5.4 に示される様な結果が得られた．ここで，適合フィードバックによる検索結果だけは，図 5.3 との比較を容易にするために，負例なしの場合について示している．これより，負例を与えた事によって予想通り適合率が上昇し，負例なしの場合の適合フィードバックより性能がよくなることが確認された．

また，この時の視点意味グループの特徴は表 5.5 に示す通りである．これより，意味グループ当たりの平均所属単語数については，正例のみの場合と同様に負例の増加につれて減少したが，今回は抽出される意味グループ数も減少することが確認された．これより，負例の増加による適合率の上昇は，意味グループの過度な一般化が抑制される効果に加え，5.3 節のステップ (3) において不適切なグループが排除される効果も反映されたものと考えることができよう．後述するように，文書整理支援システムの基盤技術として Fish Eye マッチングを用いた場合には，負例が与えられるという前提は妥当であると考えられる．

残る（2）に関しては，単語の持つ意味の曖昧さや品詞の違い，単語それ自体の意味と，複合語（フレーズ）として使用された場合との意味の違いなどが原因であると考えられる．これについては，複合語を特徴として考慮したり，基本特徴ベクトルの計算時に構文解析などの技術を用い，品詞の区別や曖昧性の解消処理 [23] などを行ってから抽出するなどの対策が考えられる．これについては，次節の形態素解析を用いた日本語文書への対応のところで考察する．

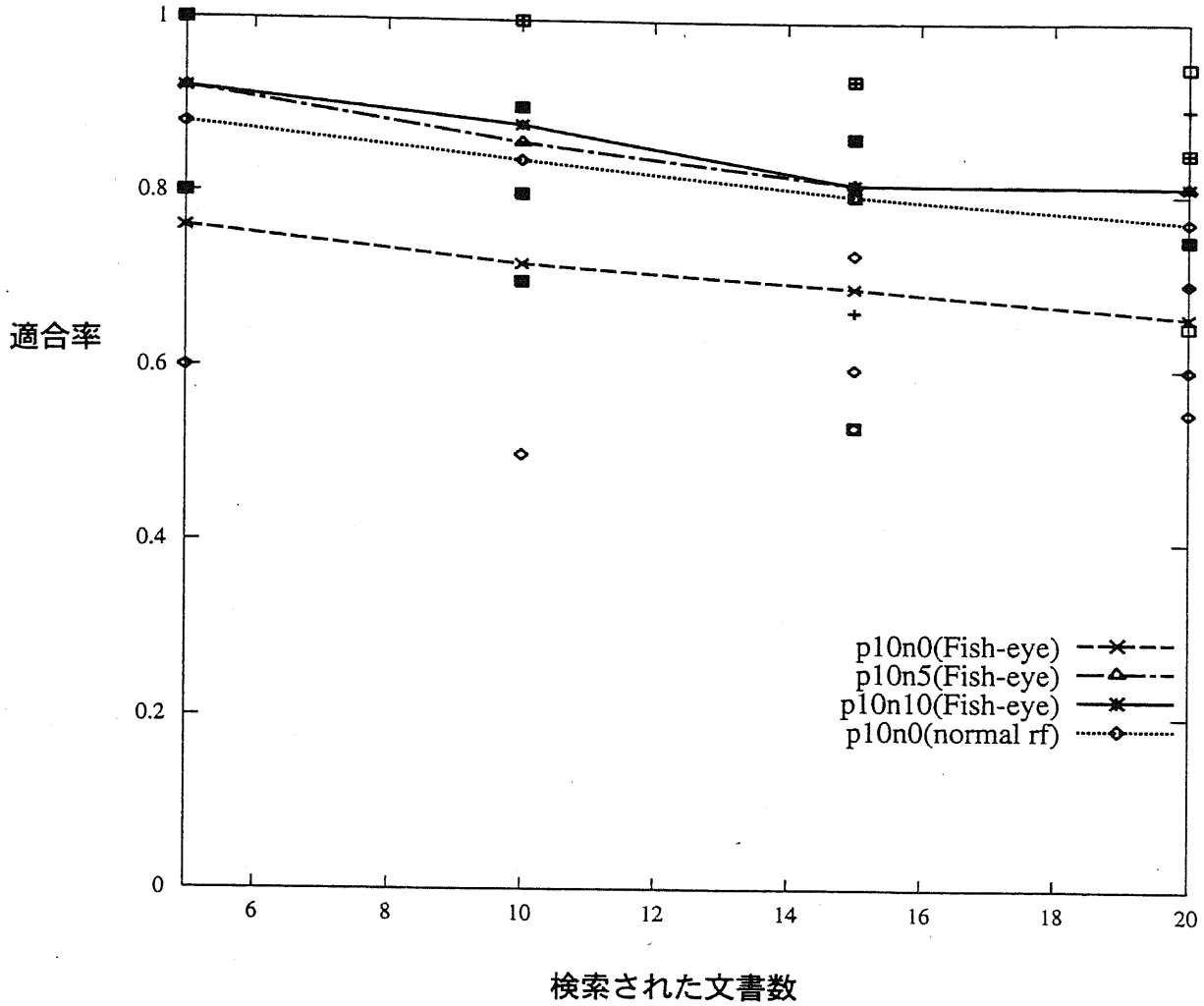


図 5.4: 実験結果 (2)：正例 10，負例 0–10

表 5.5: 視点意味グループの特徴：正負例ともに与えた場合

負例数	平均グループ数	グループあたり単語数
0	32.8	8
5	26.2	6.8
10	22.2	5.9

5.4.4 抽出された視点意味グループ

本実験において高頻度で抽出された意味グループの一部について，その見出し情報と，それに属する単語の一部を表 5.6 に示す．本研究では，ステマとして単純に各語の 7 文字目以降を切り捨てており，この表においてもステマ後の語幹を記している．これら医学関係の概念に対応する意味グループが高頻度で抽出される事からも，5.3 節で示したアルゴリズムの有効性が確認できる．

また，抽出された見出し情報や単語集合をユーザに提示することにより，現在，どの様な視点・話題によってドキュメント間の関係を捉えようとしているかを知る手がかりをユーザに提供することができる．従って，視点意味グループとして不適切な意味グループが抽出された場合であっても，ユーザが判断して視点から削除したり，視点にふさわしい意味グループを新たに追加したりといった編集作業が行いやすいと言えよう．

この様な視点情報の可読性，編集の容易性については，従来の研究では十分に考慮されていたとは言いがたい．6 章で紹介する 文書整理支援システム Fish View では Fish Eye マッチングの持つこの特徴を生かし，視点情報の提示，編集機能の提供を行っているが，これは文書整理支援システムへ適用するにあたり， Fish Eye マッチングの持つ非常に大きなメリットであると言えよう．

表 5.6: 医学関係の記事から抽出された意味グループの例

概念 ID	説明	所属単語
3f98b3	健康状態の値	blindn condit diseas disord sickne death health form habit health infert sex weakne
444506	生体構成物質	protei immuno choles dna
30f6da	臓器	eye part brain heart intest liver lung muscle nerve ligame back blade cartil marrow cornea knee prosta bill
3f969e	病気	sympto syndro aids allerg anemia anesth arteri arthri blindn bulimi cancer cold compli conges dement diabet diseas epilep flu hepati neural osteop poison
44479c	医薬品	drug medica medici vaccin laxati acid stimul patch
30f6f7	医療器具	bandag cathet glasse contac patch
3f9618	身体機能の状態捉えた人間	dead halt blind carrie case patien inpati vegeta
3f9636	嗜好品	cigare plug coffee tobacc drug smoke
44471e	睡眠に関する生理現象	sleep sleepi awaken
444b0d	生命体が内部の物質を外部に出す	period birth childb delive genera reprod run diarrh

5.5 形態素解析による日本語文書への適用

前節までは，英語文書を対象としたシステムを構築し，評価実験などを行った結果について示した．しかし，Fish Eye マッチングの本質的部分は言語に依存するものではなく，文書整理支援システムの構築と言う我々の目標からすれば，日本語の文書へも対応できるようにすることが好ましい．そこで我々は，日本語文書を対象として Fish Eye マッチングを行えるシステムの開発を行った [43]．

概念体系として利用している EDR 電子化辞書は，4.2 節で紹介したように，英単語と日本語の両方について，同一の概念体系に対応している．従って，日本語文書へ適用するためには，文書から単語を抽出する部分のみを変更すれば良い．この概念体系の共有が Fish Eye マッチングにもたらすものは，システム構築上のメリットだけではない．Fish Eye マッチングは概念単位で特徴を生成／選択する事によりベクトル空間を構成するため，異なる言語を用いた文書間でも類似度を計算し，関係を見出す事が比較的容易に実現できると考えられる．従って本研究では考慮しないが，将来的に日英両言語による文書を同時に扱うシステムを構築する事も可能であろう．

5.5.1 形態素解析を利用した単語抽出

日本語の文書では，英語と異なり単語間の区切りが明示的に存在しないため，単語抽出のためには何らかの自然言語処理を行う必要がある．そこで本研究では，奈良先端科学技術大学院大学の松本らによって開発されている形態素解析ツール「茶筅」[4] を利用して文書からの単語抽出を行った．

形態素解析の目的は，文を適切な形態素（意味を持つ最小の要素）に分割する処理であり [78]，茶筅では以下のアルゴリズムを用いている [70]．

- ある特定の位置から始まる全ての可能な形態素を辞書引きによって得る．
- 辞書引きによって得られた個々の形態素に対して，その直前に位置する全ての形態素との連接可能性のチェック，およびコストの計算を行う．

コストとして，個々の形態素に与えられているコスト（**形態素のコスト**）と，二つの形態素の連接に関するコスト（**連接コスト**）の二種類のコストを用いる．例えば図 5.5 の例では，コスト最小のパスとなる「さかな（名詞） だ（判定詞） よ（助詞）」という分析結果が返される事となる．

[4] http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

文法データや形態素データに関して，茶筅は専用の辞書が用意されている．しかし本研究では，EDR の概念体系を利用するため，単語についても EDR 単語辞書にエントリがあるものを利用したい．そこで，EDR 日本語単語辞書，日本語専門用語単語辞書中の名詞エントリについて，茶筅の形態素辞書の形式へ変換して利用した [43].

茶筅形態素辞書における名詞エントリは図 5.6 の様に記述されるので，具体的には見出し語と読みに関する情報を EDR 日本語単語辞書の単語エントリからとりだせばよい．

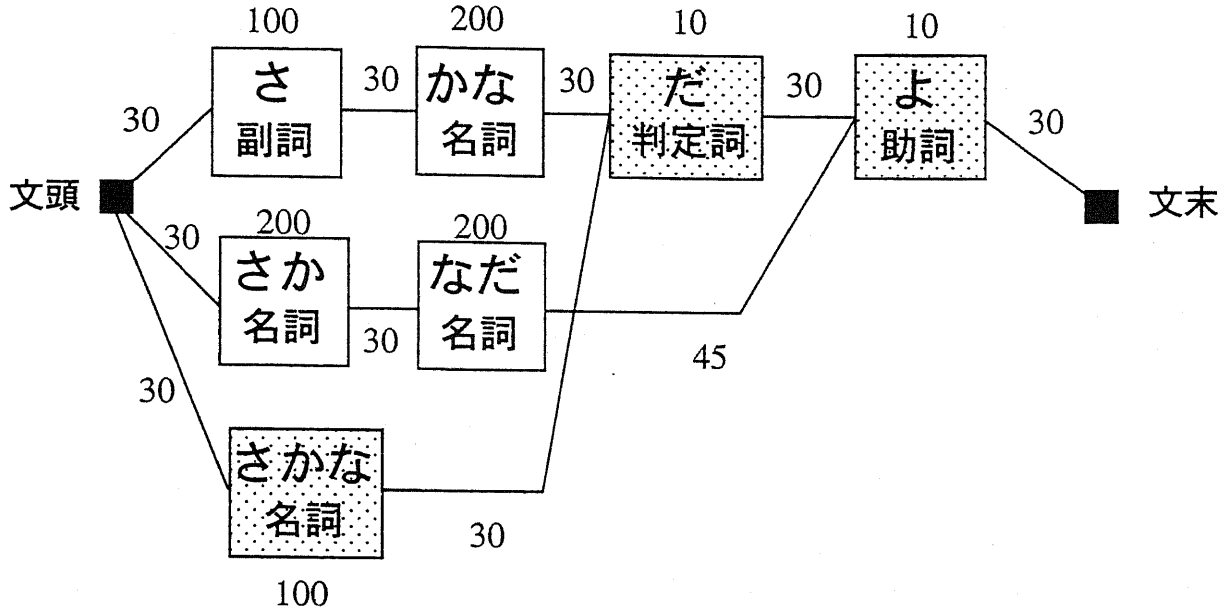


図 5.5: 茶筅による解析結果のグラフ構造

```
(名詞
 (<名詞細分類名>
  ((見出し語 <見出し語情報>)
   (読み <読み情報>))
 )
)
```

図 5.6: 茶筅用形態素辞書における名詞定義の記述

5.5.2 評価実験結果・考察

実験内容については，対象文書を日本語とした以外は 5.4 節で行ったものと同様である．対象文書であるが，英語記事を収集したニュースサイトの日本語ページに掲載されていた，同一見出しについて日本語で書かれた記事を収集し，利用した．従って対象文書数は英語記事の場合と同様に 218 であるが，他の特徴については表 5.7 に記す．

この実験において高頻度で抽出された視点意味グループについては 表 5.8 に示す．ここで，「医薬品」，「身体機能の状態捉えた人間」，「病気」は，表 5.6 においても抽出されている事がわかる．これからも，英語文書と日本語文書で同一の概念体系を共有できる EDR 辞書の長所が確認できる．

図 5.7，表 5.9 には，正例のみ，5 文書から 15 まで変化させた場合の実験結果について，適合率と視点意味グループの特徴についてそれぞれ示してある．同様に，正例を 10 文書に固定し，負例を 0 文書から 10 文書まで変化させた場合の実験結果については図 5.8，表 5.10 に示す．

通常の適合フィードバックを含めた全般の傾向として，英語文書を対象とした時よりも適合率が向上している事が読み取れる．これは，日本語文書の方が英語文書と比較して全体的に文書が短かった事なども影響していると思われるが，やはり形態素解析の利用により，英語版のような，スペースで単に区切っただけの手法よりも，文書の特徴として，意味のある単位で単語を抽出することができているからであろう．

また，Fish Eye マッチングと通常のフィードバックの比較では，英語版ほど適合率の差はなかった反面，正例や負例を増やした場合については，英語版で見られた様な適合率の上昇などの明確な傾向は確認できなかった．また，視点意味グループの特徴に関しては，正例のみの場合のグループあたり平均所属単語数を除き，英語記事の場合と同様の傾向が確認された．

正／負例数による適合率変化の傾向が表れなかった原因として，特徴ベクトルレベルでの日本語文書と英語文書の違いが考えられる．すなわち，日本語文書長が全体的に短い事，かつ表 5.7 に見られるように単語数が英語文書より増えているという事実から，単語を特徴とした場合の特徴空間がスパースになっている事が想像される．今回，実験で用いた $\tilde{S}$ 操作は縮退（Shrink）を基本としたものであり，これにより生成した特徴は文書検索に関して以下の長所／短所を持つと考えられる．このため，特徴空間がスパースな場合には，縮退による影響が，良い意味でも悪い意味でも出やすく，それが適合率の不安定な変化を引き起こしたのであろう．

長所 表層的に異なる単語が使われていても，ある視点において同一にみなせるもの同士の関連を利用し，類似度を高められる．

短所 グループによるノイズ（5.4.3 節 (1)），単語によるノイズ（5.4.3 節 (2)）の影響を受け，精度が低下する

グループによるノイズを低減するための方策の一つには，視点意味グループ抽出の精度を高める事があげられる．5.3 節で紹介した抽出アルゴリズムでは，重みの大きな単語から順にグリーディな探索を行っており，グローバルな視点からの最適性は考慮されていない．この点に関して，現在我々は視点意味グループ抽出に仮説推論 [45, 119] を用いることを検討している．仮説推論では背景知識が必要となるが，EDR 概念体系と仮説推論における背景知識はともに（厳密ではないが）木構造であり，概念体系から背景知識への変換は比較的容易に行えると考えている．視点意味グループ抽出アルゴリズムの改良の他には，構文解析や意味解析などのより高次な自然言語処理を行い，同じ語彙でも意味単位で区別して扱うなどの改善策が考えられるが，処理時間の低下など，ベクトル空間モデルを拡張した現手法の良さが失われてしまう可能性があるため慎重な検討が必要であろう．

単語によるノイズの問題をさらに詳しく分けると，以下の様になる．

- 概念体系において，滅多にあり得ない意味すら網羅的に定義されている
- 抽出時に異なる意味の区切りで抽出してしまう
- 未知語の存在

一つ目の例として，「子供」という概念に，「砂利」という単語が含まれていた．確かにこの様な意味で用いられる場合もあるだろうが，ニュース記事を対象とした今回の実験の様な場合には，この意味で用いられる事はほぼないと言って良いだろう．これより，（対象文書／分野においては）滅多に使われない語義については意味グループから削除するなどの，概念体系の整備を行う事が精度向上のために必要であると考えられる．これについては，4.2.1 節で紹介したオントロジー・シソーラス生成技術が適用できるであろう．

二点目については，単にスペースを区切りとしていた英語版と比較して，形態素解析を用いた事により改善されたと考えられるが，文法上の語の単位としての「形態素」と，文章の特徴としての「キーワード」が必ずしも一致しない，と言う事が考えられよう．すなわち，英語の熟語のように，それを構成する単語の意味と，熟語全体としての意味が全く異なってしまう様な現象が，文章の特徴としてみた場合には日本語でも発生するのではないだろうか．特に，専門用語や流行語などは，一般語の組合せで構成されているものも多く，それらを形態素単位で分割することは精度低下につながるであろう．これは，三点目の未知語の存在とも関連している．すなわち，EDR 単語辞書には，専門用語，新語，固有名詞などのエントリが少ないため，その様な語の多くが未知語となり，特徴として利用で

表 5.7: 英語記事と日本語記事の比較

	英語記事	日本語記事
記事数	218	218
単語数	1,588	1,831
意味グループ数	655	485

表 5.8: 医学関係の記事から抽出された意味グループの例（日本語記事より）

概念 ID	説明	所属単語
44479c	医薬品	ワクチン 漢方薬 薬 薬剤 薬品 薬物 目薬 下剤 新薬 鎮痛剤 サリドマイド
3f96a0	病気になる	病 感染する 潜伏
30f6df	循環器	心臓 心肺 人工心臓 人工心肺 リンパ節 動脈 毛細血管 冠上動脈 大動脈
f9720	生命	生 生命 命
e557b	因果関係	結果 功 作用 功績 効果 刺激 因子 機 原因 死因 種 種子 誘発する 影響 傷 記録 足跡 鍵
3f9618	身体機能の状態捉えた人間	妊婦 患者 入院患者 障害者 痴呆
3f961f	身体	ボディー 体 顔 筋肉 口 手足 首 足 体内 皮膚 面 患部 脚 肩 尾 姿 人体 生体 肉 指 包皮 乳 手 腕 あごひげ ひげ
3f969e	病気	症状 病状 慢性 うつ病 つわり てんかん アルツハイマー病 アレルギー インフルエンザ エイズ 遺伝病 火傷 感染症 肝炎 関節炎 近視 血栓 疾患 障害 食中毒 心臓病 精神病 多発性硬化症 痴呆 潰瘍 伝染病 糖尿病 頭痛 脳卒中 肺 白内障 発作 鼻炎 貧血 不妊症
44457b	天然食品	ニンニク 野菜 果物 なし コメ
4446fb	雌の生殖器官	卵管 乳 乳房

きていない．例えば，「サイバー」という単語の場合，EDR 単語辞書にエントリがなかったため，「サイ」と「バー」に分けて認識されてしまっていた．従って概念体系だけでなく，単語辞書を整備する事も，精度向上のために必要であると結論づけられる．

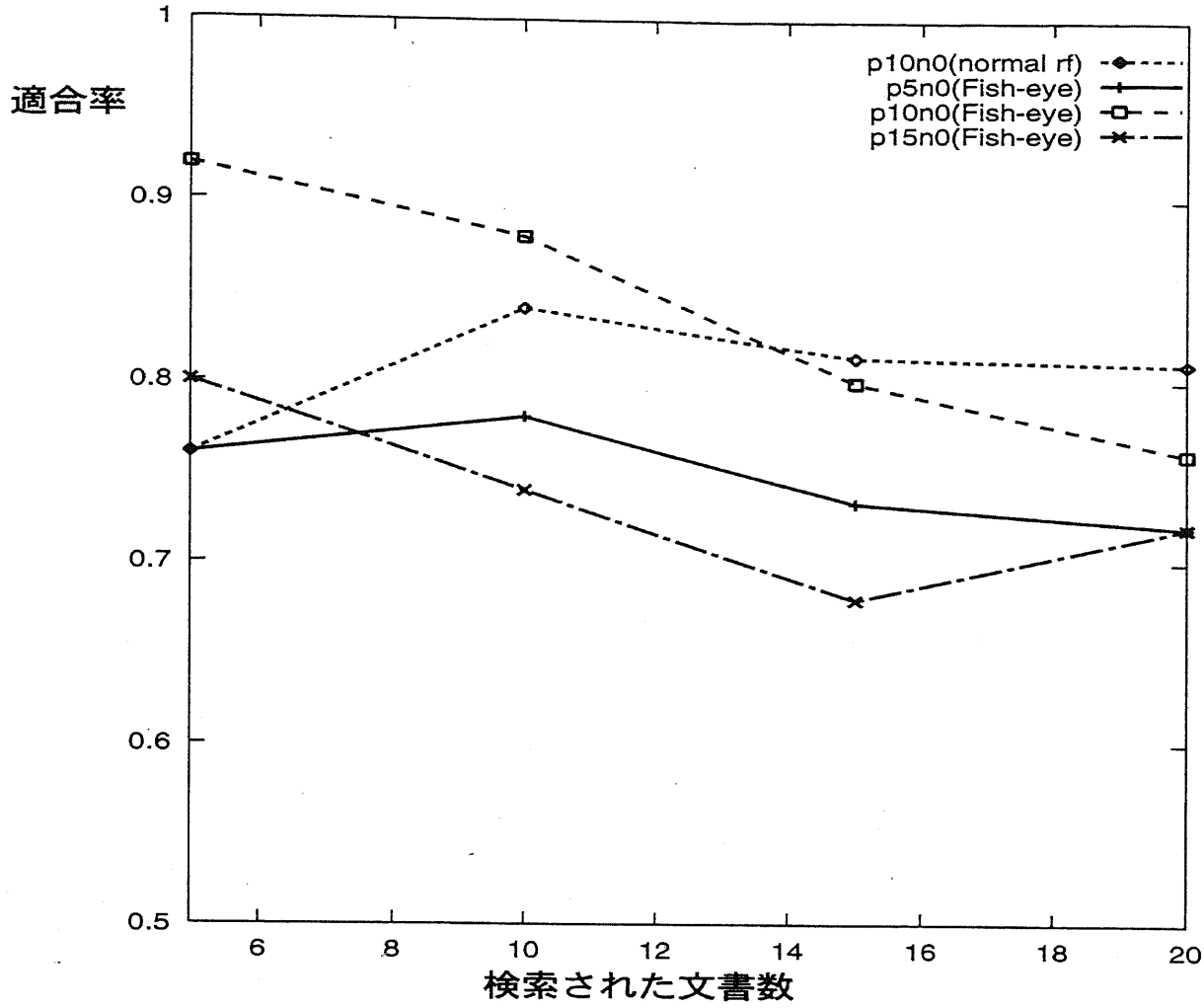


図 5.7: 日本語記事に関する実験結果 (1)：正例 5–15，負例 0

表 5.9: 視点意味グループの特徴：正例のみの場合（日本語記事より）

正例数	平均グループ数	グループあたり単語数
5	31	8.81
10	42.6	9.27
15	48.2	9.19

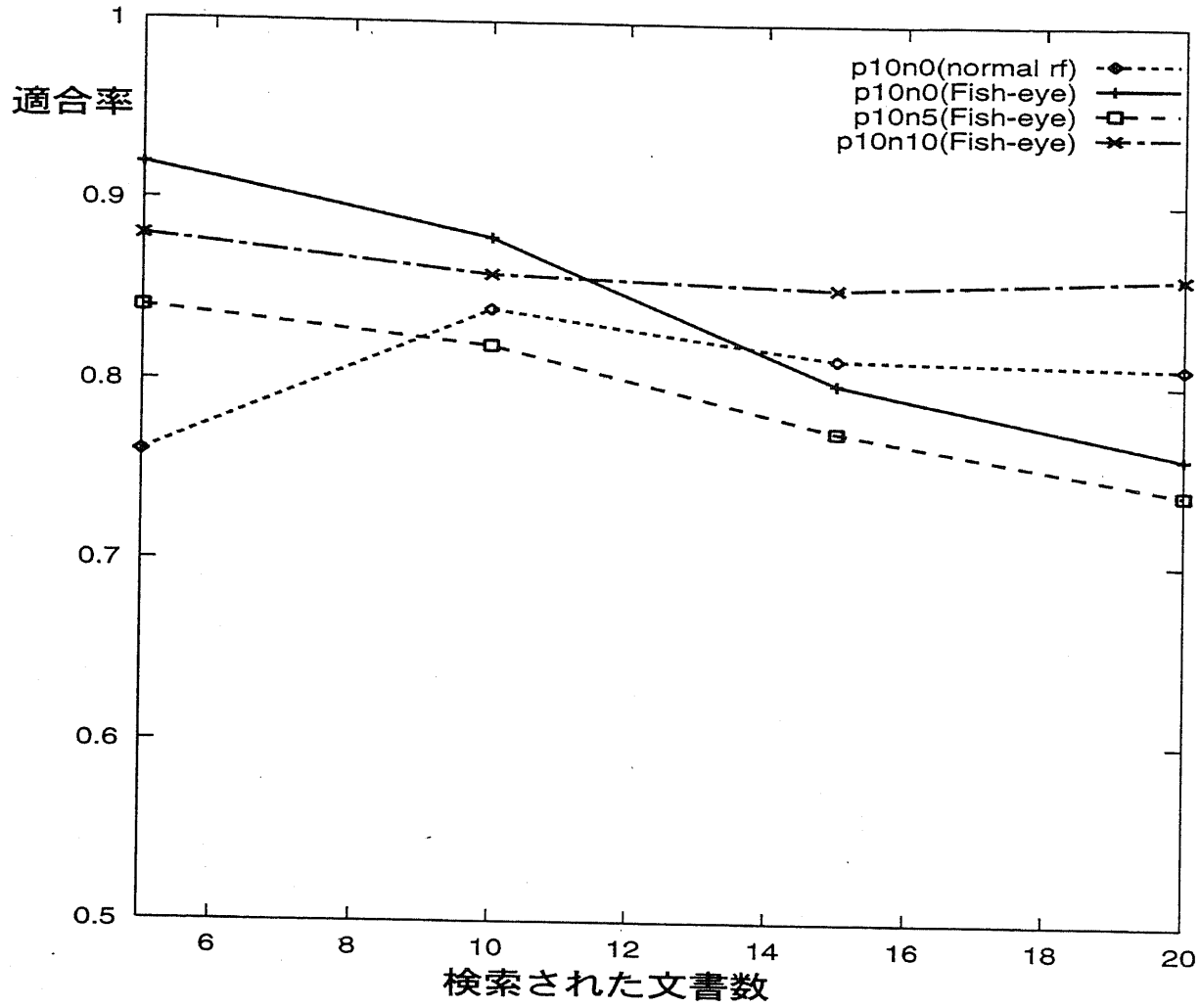


図 5.8: 日本語記事に関する実験結果 (2)：正例 10，負例 0–10

表 5.10: 視点意味グループの特徴：正負例ともに与えた場合（日本語記事より）

負例数	平均グループ数	グループあたり単語数
0	42.6	9.27
5	35.4	7.36
10	29.6	5.82

5.6 関連研究との比較

最後に，Fish Eye マッチングと関連研究との比較検討を行い，Fish Eye マッチングの持つ特徴，利点を考察する．

ユーザの視点を抽出し，可視化すると言う点では，多次元尺度構成法 [114] やバネモデル [122] などを用いて，画面配置上の位置関係の形で表現する方法がある．これらの方法は，本来多次元であるはずのドキュメント間の関係を，二次元空間での距離に変換する際に，その誤差が極小となる様に配置する．ここで，ユーザの初期配置によって最終的に達する局所解（配置）が異なる，という性質を利用することにより，ユーザの視点を画面配置上に反映している．この方式は，ドキュメント数が増えると画面上が繁雑になりわかりづらいという欠点を持つのに対し，Fish Eye マッチングを用いた場合では，画面上での配置だけでなく，Fish Eye ベクトル生成時に使用された意味グループからも視点を読みとることが可能であるといった利点がある．

また，特徴クラスタリング [130] による単語のグループ化は，Fish Eye マッチングの Shrink 操作と類似していると考えられるが，純粋に統計的な操作であるため，ほとんど全ての単語を含むグループができてしまう場合がある事などが知られている．また，得られたグループがどの様な概念・意味を表しているかが自明ではないという問題もある．これは，LSI[13, 19] などの統計的手法を用いたベクトル空間の再構築でも同様である．

この様な問題は，上記手法による単語のグループ化は一種のクラスタリングであるところに起因すると考えられる．これと対比して，Fish Eye マッチングにおけるそれは，既存の概念体系の各カテゴリに単語を適切に割り振る，一種のカテゴライゼーションとして捉える事ができる．すなわち，概念体系から抽出された意味グループは人間にとって理解可能な概念に対応しているため，グループ計算時に概念体系がトップダウンな制約として有効に機能するだけでなく，得られた意味グループからユーザが意味を把握しやすいといったメリットがある．

我々は，Fish Eye マッチングを文書整理支援システムの基盤技術として用いる事を想定している．この場合，高い検索能力に加え，上述の様な視点の可読性，編集容易性を備えている点は非常に有効かつ必要なものであると考えられる．

Chapter 6

文書整理支援システム Fish View

本章では，前章で紹介した Fish Eye マッチングを用いた文書整理支援システム Fish View について述べる．最初に，文書整理支援に Fish Eye マッチングを適用する事に関して考察する．文書整理支援に適用するためには，前章で紹介した手法に多少の修正を加える必要があり，これに関して行った予備実験についてもここで触れる．

その後，開発したシステム Fish View について，各機能の紹介および，このシステムを用いた場合の作業プロセスについて説明し，Fish View を実際のユーザに使用してもらった評価結果について最後にまとめる．

6.1 Fish Eye マッチングの文書整理支援への適用

文書整理支援については 4.4 節で触れた通りであるが，もう一度ここで簡単に振り返る事にする．

WWW などを通じて大量に収集されたドキュメントを活用するには，各ドキュメントを個々に読むよりも，全体としての構造，関係を捉えながら読み進めた方が，個々のドキュメントの深い理解につながると同時に，最終的な目的，サーベイ／レポートの作成や新しいアイデアの生成に有利であることは 4.4 節で述べた通りである．文書整理支援とは，この様な，「大量文書を整理し，全体構造を把握しながら漸進的に読み進める」作業プロセスの支援である．ここでもう一度文書整理プロセスを振り返ると，

読書 今までに読んだ文書や，頭の中の知識との関連を意識しながら読み進める．

図解作成 今まで読み進めて来た文書群から図解（**局所図解**）を作成し，視点を整理する (図 4.3(b)(d)).

検索 得られた視点をもとに，次に読むべき文書を決定する (図 4.3(c)).

これらのプロセスにおいて，計算機により有効に支援を行えると思われる部分は次の通りである．これらの支援について，Fish Eye マッチングを用いていかに実現するかについて考察する．

1. 局所図解から，ユーザの視点を把握・提示する．

2. 次に読むべき文書を検索し，提示する．

3. 図解において，ユーザが見落としている文書間の関係を指摘する．

(1) は，(2)，(3) の支援を有効に行う上で非常に重要な支援である．すなわち，ユーザの作成した図解には，現在までに読み進めて来た結果としてのユーザの視点，興味といったものが反映されており，図解を作成する事によって，ユーザは自らの視点・興味を把握するとともに，次に読み進むべき文書を決定すると考えられる．

しかし，ユーザが自らの視点や興味について，全て明示的に把握しているとは限らないであろう．むしろ，入手した情報を読み進めている途中段階においては，「なんとなく，この文書とこの文書は興味深いな」くらいの感覚で図解を作成していると思われる．この様な段階で，通常のサーチエンジンを用いた文書探索の様に，ユーザが明示的に興味を入力するのでは，その様な曖昧な視点・興味を無視してしまうことになる．従って，ユーザが局所図解の中から現在興味のある文書に関連する部分を「これに関係する文書が読みたい（検索したい）」といった形で指摘し，それらからシステムがユーザの興味・視点を抽出，提示する事ができれば，(2), (3) においてより能動的かつ有効な支援を行う事ができると考えられる．この様な図解からの視点抽出は，5.3 節で紹介した視点意味グループ抽出アルゴリズムを用いる事により，Fish Eye マッチングで行う事ができる．

また，Fish Eye マッチングでは抽出した視点情報をユーザに可読な形で提示する事ができるので，図解に込められた曖昧な視点を明示的な形でユーザに提示できると言う，いわば**視点の外化**効果も期待できる．

もちろん，ユーザが明示的に視点・興味を指摘したい場合もあるだろうし，システムの推論による視点抽出がユーザの思惑とは異なってしまう場合も考えられよう．従って，システムが推論し，提示した視点情報をユーザが修正・編集する機能も用意する必要がある．前述した様に，Fish Eye マッチングではユーザに可読な形で視点情報を扱うため，この機能の実現は容易である．

(2), (3) の支援については上述の通り，図解中の一部をユーザが指定し，それから抽出（場合によってはユーザにより編集）された視点・興味情報をもとに，その視点に関連が深

いと思われる文書を検索したり，図解中に存在する文書間の関係を再吟味したりする．すなわち，指定された視点意味グループを引数として Shrink，Magnify 等の操作を全対象文書に施して Fish Eye ベクトルを計算し，マッチングを行う事によって新規文書の検索，および図解中の文書間の類似性の再計算・図解修正を行う．

(3) に関しては，ユーザが作成した図解から部分的に抽出した視点をもとに，新たな情報を追加することになる．これについて説明する前に，まず本研究で開発するシステム Fish View で採用する図解について定義する．

Fish View では KJ 法と同様に，ユーザがある視点・興味において近い，関連があると思った文書をグループ化することによって表現する[1]（図 6.1）．多数のオブジェクト（文書・キーワード）の関係をディスプレイ上に図示する方法としては，この他にもオブジェクト間の類似度が近い程，近くに配置する方法 [2, 106, 113, 115, 117] などがある．しかしこの方法では，扱うオブジェクト数が増えるにつれて画面上が繁雑になり，関係を正しく把握する事が困難になる．また，二次元のディスプレイ上では同時にたかだか二種類の関係しか表現できないが，グループ化であれば様々な視点に基づくグループを混在させる事が可能である．また，(1) についても，グループ単位で視点・興味を抽出する形で図解中の部分的な興味・視点を容易に指摘できるため，本研究ではグループ化による図解作成を採用した．

[1] KJ 法では，グルーピング作業が終った後でグループ間を線で結ぶ作業を行うが，本研究ではリンクは他の用途に用いる

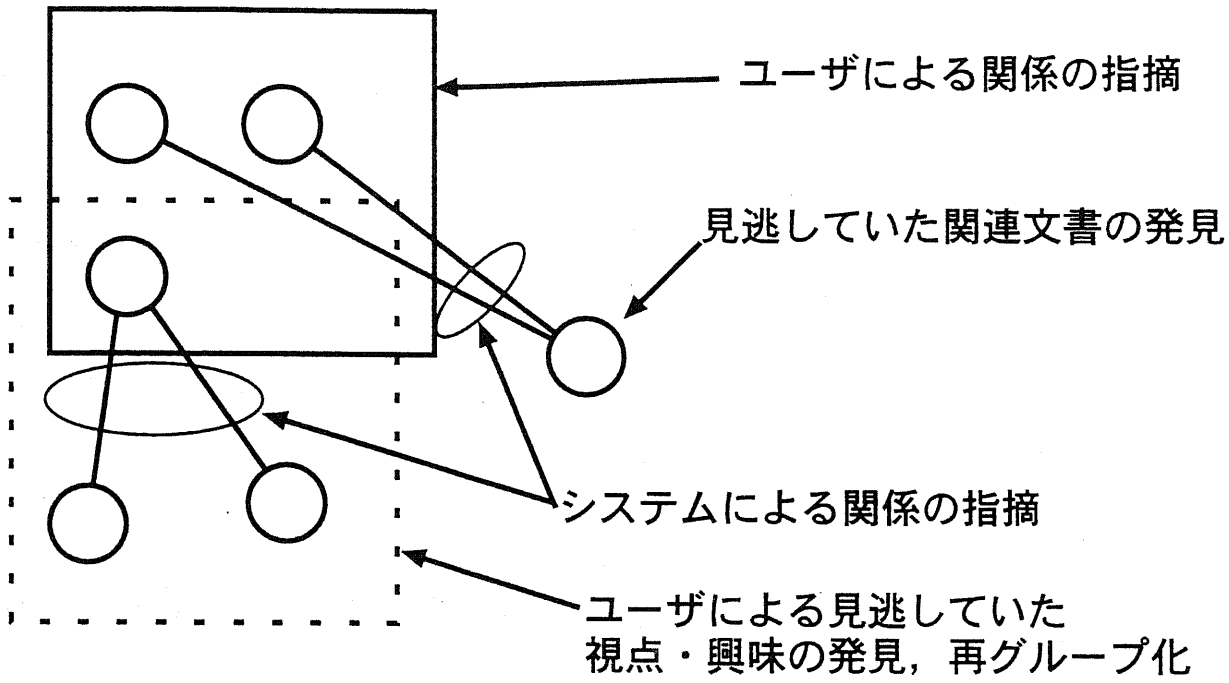


図 6.1: Fish View における図解の概要

ユーザによる文書間の関係の指摘はグループ化によって表現するのに対し，システムによる関係の指摘はリンクで表現する事により区別する（図 6.1）．すなわち，Fish Eye マッチングにより式 (5.6) で計算した類似度の値が閾値を越える文書間にはリンクを張って表現する．リンクによりシステムが指摘する情報は，ユーザにとって以下の二つの意味で有効であると考えられる（図 6.1）．

- 見落としていた関連文書の発見
- 見逃していた視点・興味の発見

前者については，ある視点・興味のもとに文書をグループ化した場合，本来はそのグループに含まれるべき文書を見逃してしまう事は，対象とする文書が大量になればなるほど起こりやすいと考えられる．ユーザが指定した視点情報に基づいてシステムがリンクを張る事により，この様な文書間の類似性の見逃しを指摘する事が期待できる．

後者に関しては，ユーザが想定していなかった（グループ化していなかった）文書間の関係がリンクで指摘された場合に，そのリンクの持つ意味，表す関係を考察する事により，ユーザは文書のグループ化に関する新たなアイデアを得る事が期待できる．

6.2 視点厳選に関する予備実験

前章で紹介した Fish Eye マッチングにおける視点意味グループ抽出アルゴリズムにおいては，重みの大きい単語から順にグリーディな探索を行うことにより，全ての単語についてそれを含む意味グループを検索し，抽出していた．このアルゴリズムによって抽出される意味グループ数は 20 ～ 50 程度であり，これら全てを視点情報としてユーザに提示するには多すぎると考えられる．また，検索／類似度計算の精度の面でも，特に shrink 計算においては単語を縮退する事による誤差の影響が，縮退によるメリットを越えて大きくなってしまう事も考えられる．

以上の考察より，抽出可能な視点意味グループを全て利用するのではなく，視点の可読性，および検索精度向上の両者において有効な意味グループのみを利用する事，すなわち「**視点の厳選**」を行う事が好ましい．そこで，5.3 節で提案したアルゴリズムのステップ (2) において，*Wlist* 中の重みが正の単語がなくなるまで繰り返すのではなく，「重みが正かつ上位 n の単語」についてのみ，意味グループを抽出する事にする．この様な，視点意味グループ抽出処理の対象とする単語を「**核単語**」と呼ぶ事にする．

核単語数を変えて，5.5.2 節と同様の実験（日本語記事）を行った結果を図 6.2, 6.3 に記す．図 6.2 は正例のみを 10 文書与えた場合，図 6.3 は正例 10，負例 10 文書を与えた場

合の適合率について示している．両図において，‘core0’ は意味グループの抽出を行わない，すなわち通常の適合フィードバックと同様の場合を指し，‘All Groups’ は前章の様に抽出可能な意味グループを全て用いる場合，‘core20’ は核単語数を 20 とした場合を指す．

これより，核単語を減らしても適合率の低下は見られない．また，検索文書数が増加しても適合率があまり変化しないという傾向も確認された．これは，検索文書数を増して再現率を高めた時に問題となる適合率の低下を抑える事ができると言う点で好ましい性質であると言えよう．

処理時間に関しても，抽出可能な意味グループを全て抽出した場合は，通常の適合フィードバックと比較して平均 12 〜 13 倍程度時間がかかるのに対し，核単語数を 20 にした場合は 4 〜 5 倍程度に収まっている．

以上より，視点情報としてのわかりやすさ，検索精度，処理時間いずれにおいても，核単語数を限定し，視点を厳選することによるメリットは大きいと判断できる．従って Fish View では核単語数 20 として視点意味グループ抽出を行う事にする．

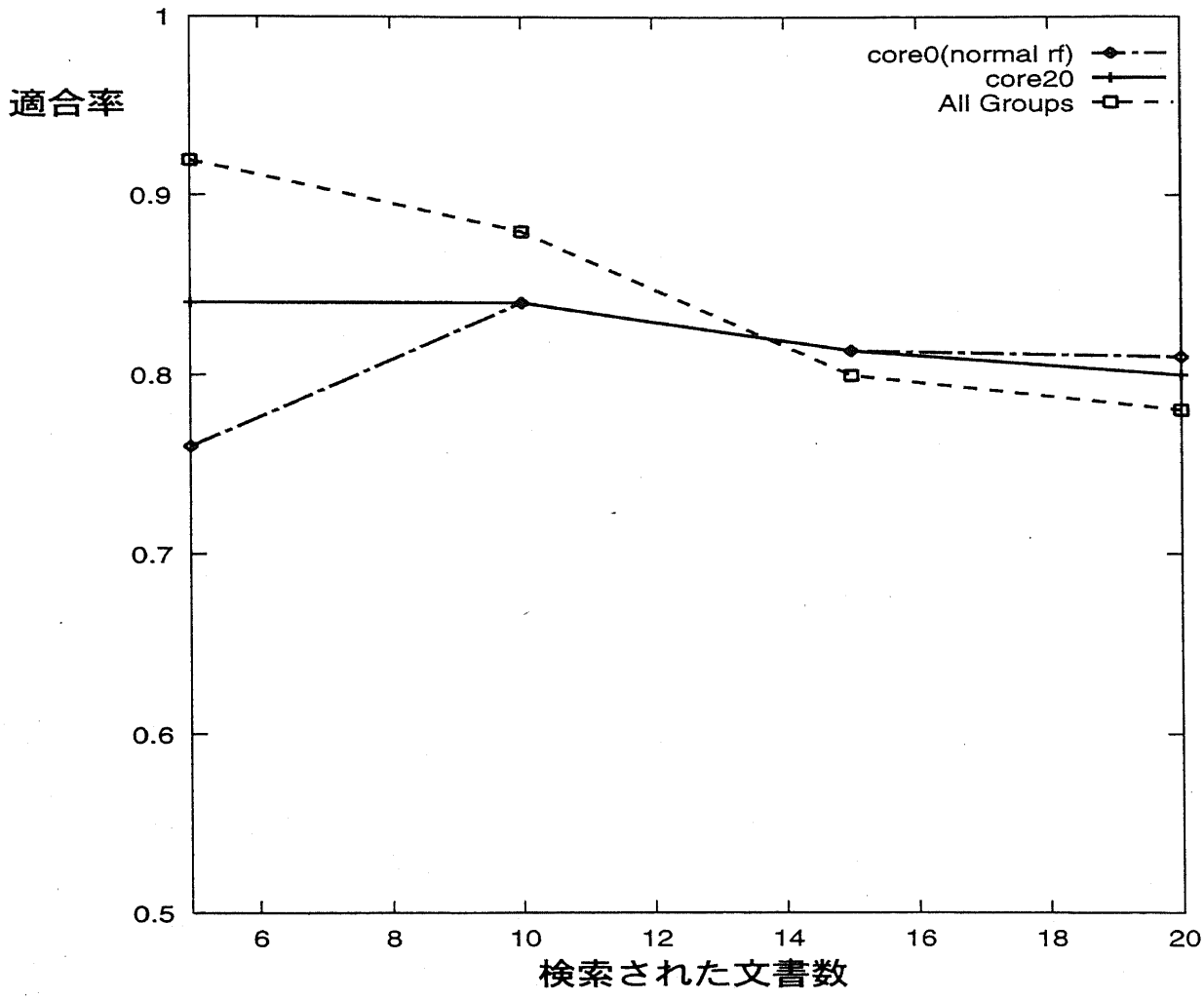


図 6.2: 核単語数による検索精度の比較結果：正例のみを与えた場合

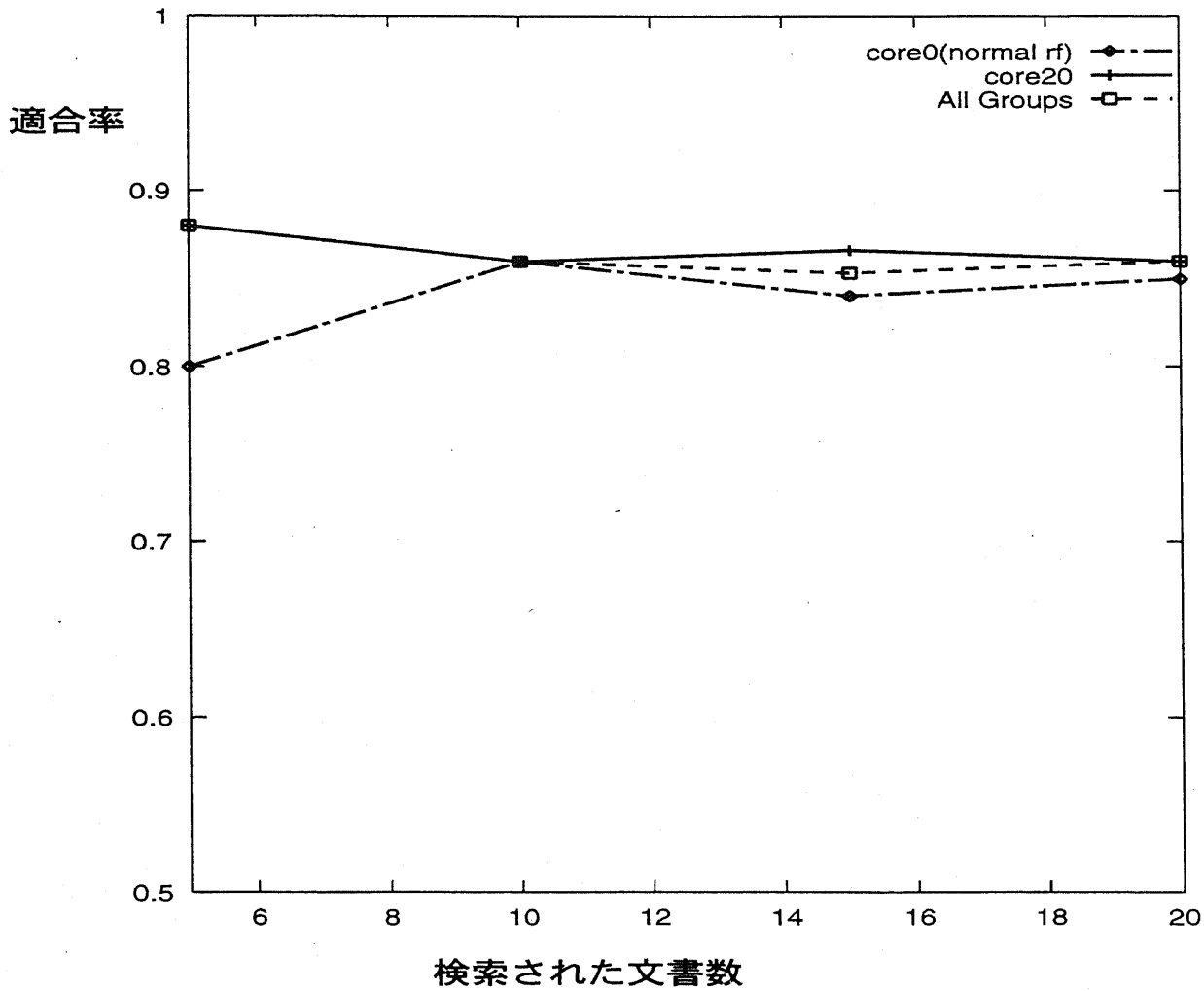


図 6.3: 核単語数による検索精度の比較結果：正例，負例ともに与えた場合

6.3 文書整理支援システム Fish View の概要

図 6.4 は，我々が開発した文書整理支援システム Fish View の全体像である．Fish View はクライアント・サーバ型として開発されており，クライアントは Tcl/Tk8.1 を用いて実装され，Windows，UNIX の両プラットフォームで実行可能である．サーバは C 言語を用いて UNIX 上に実装され，EDR 電子化辞書から求めた単語，概念（意味グループ）に関するデータベースおよび，文書データベースを持っている．従って，図解作成はクライアントで，Fish Eye マッチング，視点抽出などの演算はサーバ側で行われる．

一番大きいウィンドウは Main Window と呼ばれ，このウィンドウ内のキャンバスと呼ばれる領域中に円で表される文書オブジェクトを配置し，図解を作成する．また，現在の視点に関する情報は Main Window 右側に表示され，ここで編集を行うことができる．

図 6.4 左側の一番背後にあるウィンドウは Document List Window と呼ばれ，整理対象となる文書のリストを表示する．リスト中での順番は，ユーザの視点との関連度に基づいてソートする事が可能である．

左側にあるもう一つのウィンドウは Group Retrieval Window と呼ばれ，単語をキーとした意味グループの検索を行う．同様に，右側にあるウィンドウは Group Structure Window と呼ばれ，意味グループをキーとして，それらと関連する意味グループの検索を行う．最後に，図 6.4の中央下部にあるウィンドウは Contents Window と呼ばれ，Main Window のキャンバス上に表示されている文書の内容（本文）を表示する．

以下では，実際の作業プロセスに沿った形で，各ウィンドウの機能説明および利用方法を説明する．

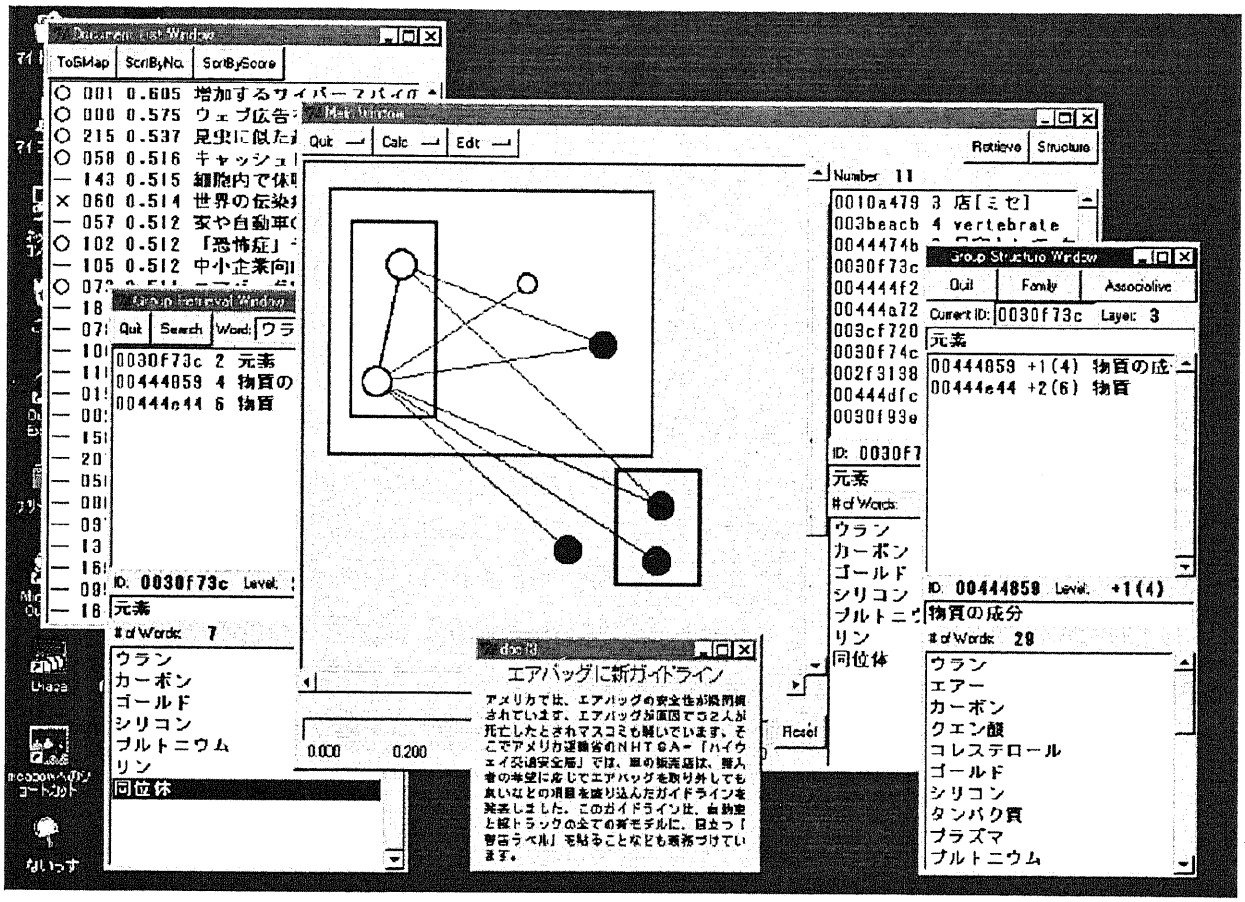


図 6.4: 文書整理支援システム Fish View の概観

6.3.1 図解作成：Main Window

前述した通り，文書整理プロセスにおいてユーザは，文書を読み進めながら漸進的に図解を作成していく．

まず始めに，Document List Window（図 6.8）から次に読む文書を選択する．このウィンドウに関しては次節でくわしく説明する．選択された文書は，ノード（円）として Main Window 中のキャンバスと呼ばれる領域へ追加される（図 6.5）．

ノード内にマウスカーソルを入れると，タイトルバルーンが表示される．また，文書の内容（本文）については，ノードを右クリックして表示されるメニューから Contents Window を開く事により確認できる（図 6.6）．

キャンバス中に現れた各文書の内容を確認したら，自分の興味・視点に従って，関連のあるものごとにグループ化を行う．また，内容を確認した結果，興味なし・不要と判断した文書については，対応するノードをキャンバスから削除する事ができる．これも，ノードを右クリックして表示されるメニューから行う．

グループの作成についてはいくつかの制約がある．まず第一に，グループは入れ子構造をとることができる．これにより，大きな話題でグループ化した後，より細かい話題に従って分類するといったような，話題・興味の階層構造を表現する事ができる（図 6.7）．

また，グループの交差は許していない．すなわち，一つの文書は一つの話題にしか分類してはならないとしている．これは，「一つの文書は複数の視点から読む事が可能である」という事実と矛盾するようにとられるかも知れないが，文書を整理する場合には，現時点でもっとも関連が深いと思われる話題の元にグループ化することが，思考の発散を抑え，考えをまとめる上で好ましいと考える．この考えは KJ 法でも同様である．従って，他の視点・話題に分類したい場合には，一旦設けたグループを取り除き，新たにグループを生成する．このためにも，Fish View では各グループについて表 6.1 に示す二つの状態を用意している．

すなわち，UNFIX 状態はユーザが図解として関係を表現しただけであり，システムへはまだ視点・興味を伝えていない，仮の状態である．これに対し FIX 状態は，正式なグループとして固定した状態であり，システムはユーザの現時点における視点・興味に関する情報として認識，利用する．図 6.5 中では白色に塗りつぶされているグループが FIX 状態にある事を示している．これより，前述の様に文書のグループを変更したい場合には，FIX 状態にあるグループを一度 UNFIX 状態に戻して修正すれば良い．また，入れ子構造に関しては，内部に子グループを含むグループを FIX 状態にする場合には，それらの子グループがあらかじめ全て FIX 状態になっていなくてはならない．

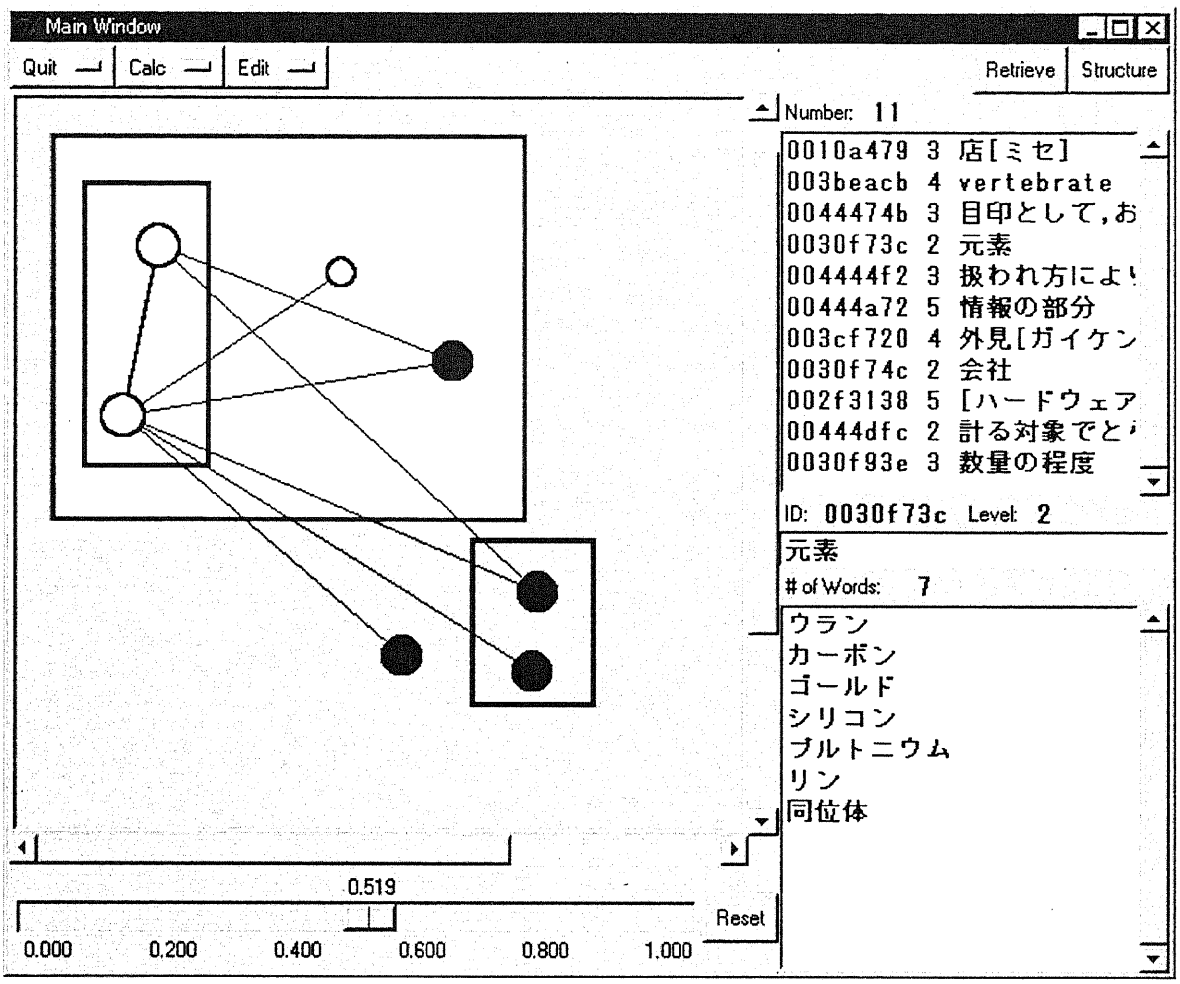


図 6.5: Main Window

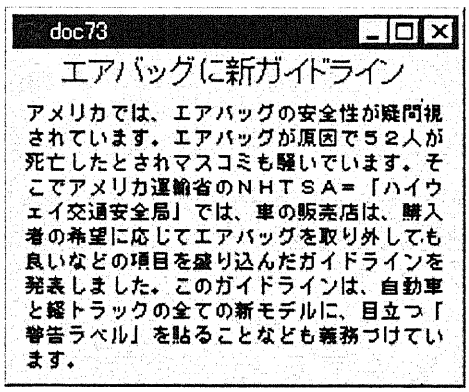


図 6.6: Contents Window

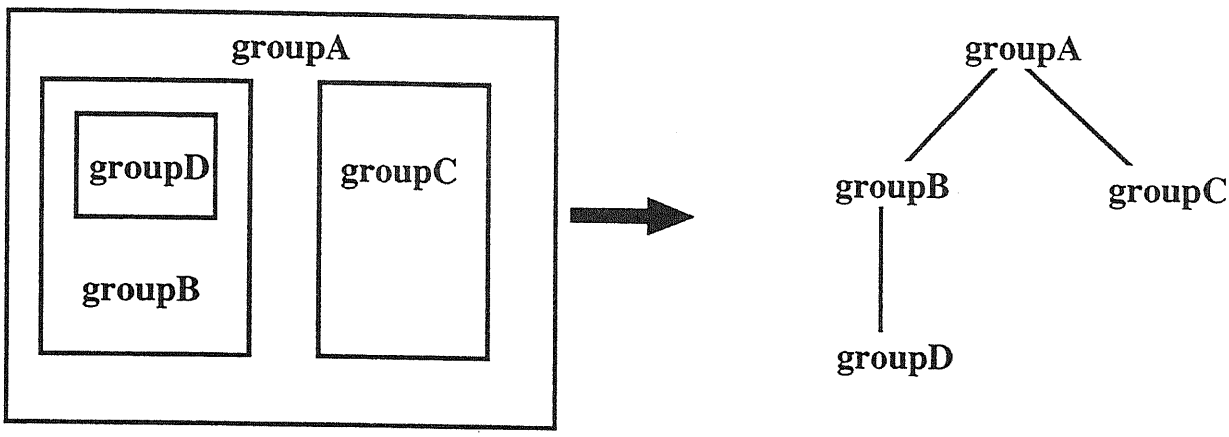


図 6.7: グループの入れ子構造と話題の階層構造の対応

表 6.1: グループ状態の比較

	FIX 状態	UNFIX 状態
視点情報	固定	未定
視点に基づく検索	可能	不可
移動・変形	移動のみ可	両方とも可
ノードの追加	不可	可能

FIX 状態のグループから，視点意味グループを抽出する場合について説明する．5.3 節で説明した視点意味グループ抽出アルゴリズムにおいては，入力として正例・負例文書集合をそれぞれ与える必要があった．そこで Fish View では，ユーザが FIX 状態にするグループを指定した場合，指定されたグループ内部に対応するノードが存在する文書を正例，グループ外に対応するノードが存在する文書あるいはキャンバスから対応するノードが削除された文書を合わせて負例文書集合として視点意味グループを抽出する．また，入れ子構造の場合には，内部子グループによる分類結果を尊重するために，5.3 節で示したアルゴリズムのステップ (5) で意味グループを選択する際に，内部子グループの視点意味グループとなっている意味グループについては式 (5.9) の値を 2 倍し，抽出されやすくしている．

6.3.2 視点に基づく文書検索：Document List Window

新規文書検索は，前節で紹介した FIX 状態にあるグループをクリックし，Main Window 上部にあるメニューボタンから Shrink，Magnify 操作のいずれかを選択する事によって行われる．これにより，指定されたグループに対応する視点意味グループを引数として Magnify，Shrink 操作を行い， Fish Eye ベクトルを生成，マッチングを行う．ここで検索精度を考え，Shrink 操作の場合は純粋な S 演算ではなく，$\tilde{S}$ 演算を用いている事は前章で触れた通りである．また，クエリーベクトルの各要素の値は，意味グループに対応する特徴の値は式 (5.9) で計算した値，単語に対応する特徴については式 (5.8) で計算した値をそれぞれ用いる．ここで，正例，負例文書集合については前節で述べた通りである．

クエリーベクトルと各文書との類似度を式 (5.6) に基づいて計算し，その値順に文書を並べたリストが Docment List Window に返される（図 6.8）．Document List Window における文書の順番は，ユーザの好みに応じ，類似度に基づいて並べる事もできるし，文書番号順に並べる事も可能である．

Document List Window 中の各文書エントリは，文書状態，文書番号，類似度の値，タイトルによって構成される．文書状態とは，その文書をすでに読んだかどうかなどを示すためのものであり，“○” はキャンバス中に存在している状態，“×” は一度は読まれたが現在はキャンバスから削除されている状態，“—” は一度も読まれていない状態を示す．

ユーザはこのリストに基づき，“×” あるいは “—” 状態にある文書の中から次に読みたい文書を複数選択する．ユーザの現在の興味に近い文書程，リストの上位に位置するはずなので，通常はリストの上位から順に，数文書ずつ選択する事になろう．選択された文書は，Document List Window 上部にある “ToGMap” ボタンを押す事により，対応するノードがキャンバス中に追加され，キャンバス中の既存ノードとの関係（類似度）が Fish Eye マッチングによって計算され，リンクが張られる．リンクの太さには 3 種類あり，文書 d_i, d_j 間

のリンクは以下の計算によって決定する．ここで，$sim(d_i, d_j)$ は式 (5.6) で計算される値である．また，Th はしきい値であり（$0 \leq Th < 1$），Main Window 下部のスケールバーによって調整可能である．

1. $sim(d_i, d_j) < Th$ ならば，リンクを張らない．
2. $x = \dfrac{(sim(d_i, d_j) - Th)}{1 - Th}$ を求める．
3. $x < 0.25$ なら細いリンク，$x < 0.5$ なら普通のリンク，それ以上なら太いリンクとする．

また，文書 d に対応するノードのサイズ（半径）$r(d)$ は，文書の特徴ベクトルとクエリーベクトルとの類似度に基づき，以下の計算によって決定する．ここで，$sim(d)$ は式 (5.6) で計算される，文書 d とクエリーとの類似度を表すとする．また，$MaxR$ はノードの最大半径を表し，あらかじめ決められているとする．これより，$sim(d) = 0$ で最小値 $(1 - Th)MaxR$，$sim(d) = 1$ で最大値 $MaxR$ をとる．

$$r(x) = \{1 - (1 - sim(d))Th\}MaxR \tag{6.1}$$

指定された視点意味グループや，演算などの影響によって，キャンバス中のノードサイズやリンクの本数などは大きく変化し，リンクだらけで図解が非常に見づらくなる場合や，反対にほとんど関係が指摘されない場合などもありうる．この様な場合には上述した様に，Main Window 下部にあるスケールバーを調整する事によって上記計算におけるしきい値 Th を変更する事が可能となっている．

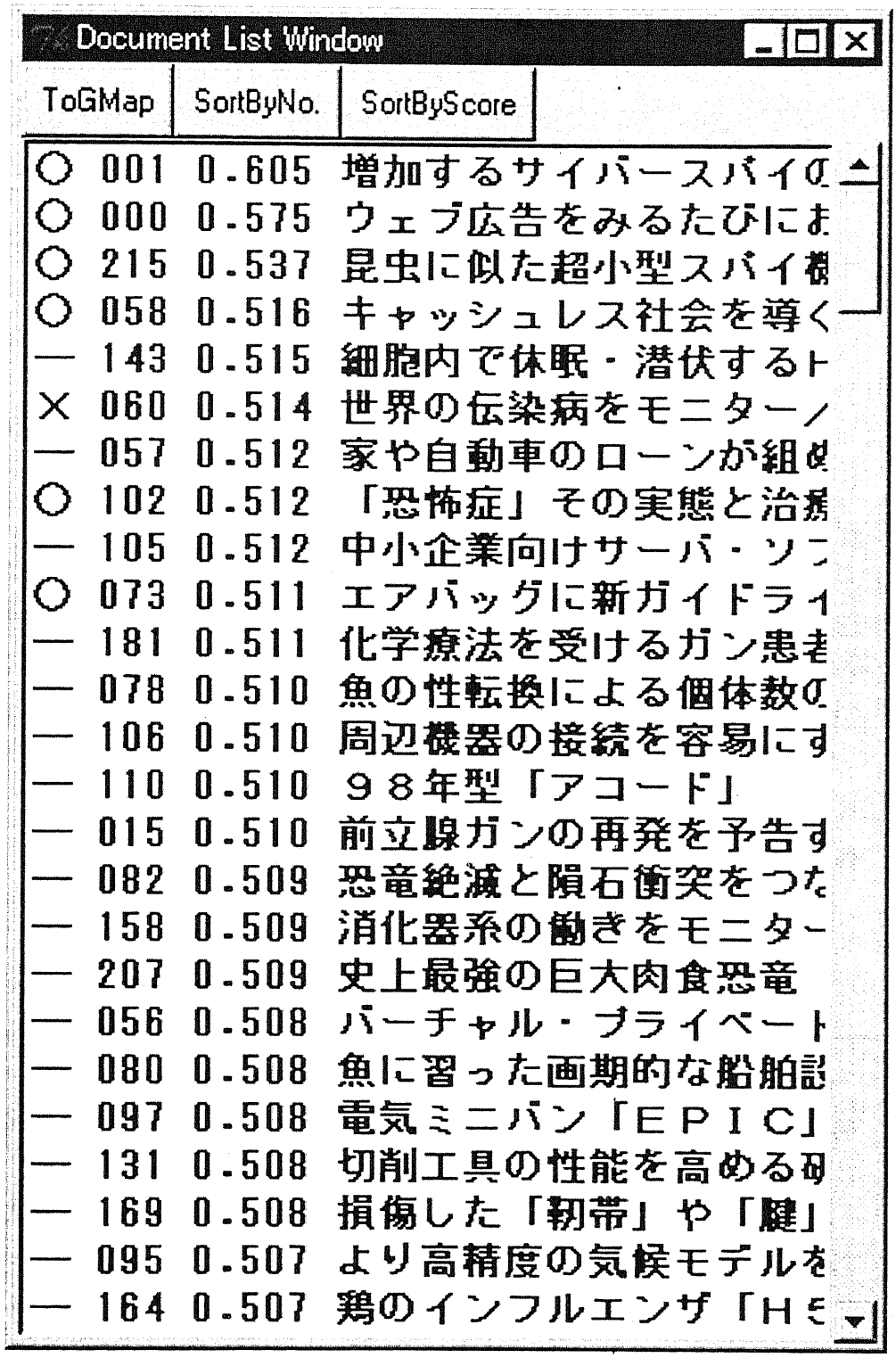


図 6.8: Document List Window

6.3.3 視点情報の編集：Group Retrieval/Structure Window

上述した通り，ユーザによる視点の指定は，関連文書を FIX グループ化するだけであり，後はシステムが自動的に視点意味グループを抽出してくれる．これは非常に便利である反面，システムによって示された視点が不適切であったり，ユーザの思惑とは異なってしまうケースもあり得る．こういったケースに備え，Fish View では視点情報を修正／編集する機能を用意している．

ユーザが視点情報に関して行いたい修正／編集としては，以下のようなものが想定される．

意味グループの削除 不適切，無意味なグループを削除したい

意味グループの入れ換え ある単語を含んでいる意味グループを，その単語を含む他の意味グループに変更したい

意味グループの追加 他の意味グループを新たに視点として追加したい

Main Window 左側にある視点情報リスト中には，各視点意味グループの ID，レベル，見出し情報が表示されている．ここでレベルとは，対応する概念の，EDR 概念体系辞書における階層を意味する．最下位概念は 1 で，上位概念になるほどレベルは大きくなる．また，5.1.1 節の意味グループ計算において，どの意味グループにも属さない単語に関して作成された，所属単語数 1 の意味グループに関してはレベル 0 として扱われる．

また，視点情報リスト中の各エントリをクリックすると，下部に所属する単語リストが表示される．これらの情報により不要と判断した意味グループについては，右クリックして表示されるメニューより削除する事ができる．

意味グループの追加，入れ換えのために，Fish View では Group Retrieval Window（図 6.9），Group Structure Window （図 6.10）を用意している．

Group Retrieval Window では，単語をキーとして，それが含まれる意味グループを検索する事ができる．キーとなる単語はユーザが直接入力する事もできるし，視点情報リスト，Group Retrieval Window あるいは Group Structure Window それぞれの単語リスト中から選択する事もできる．

これに対し Group Structure Window では，意味グループをキーとして，以下の二種類の意味グループを検索する事ができる．キーとなる意味グループは，視点情報リスト，Group Structure Window あるいは Group Retrieval Window それぞれの意味グループリストの中から選択する事ができる．

FAMILY 概念体系において上／下位関係にある意味グループ

ASSOCIATIVE 上位3レベル以内で上位グループを共有する意味グループ（関連意味グループ）

Group Retrieval Window, Group Structure Window ともに，表示内容・形式は Main Window 中の視点情報リストとほぼ同様である．これらのウィンドウから新たに意味グループを視点情報リストに追加したい場合には，その意味グループを右クリックして表示されるメニューから行う事ができる．本研究では，一つの単語は一つの意味で用いられる，すなわち同時に一つの意味グループにしか属さない，という前提を置いている．従って新たな意味グループを視点に追加する際，同一の単語を含む意味グループがすでに視点として選択されている場合には，意味グループの入れ換えを行って良いかを確認するメッセージが表示される．この様な競合する意味グループが存在しない場合には自動的に視点に追加される．

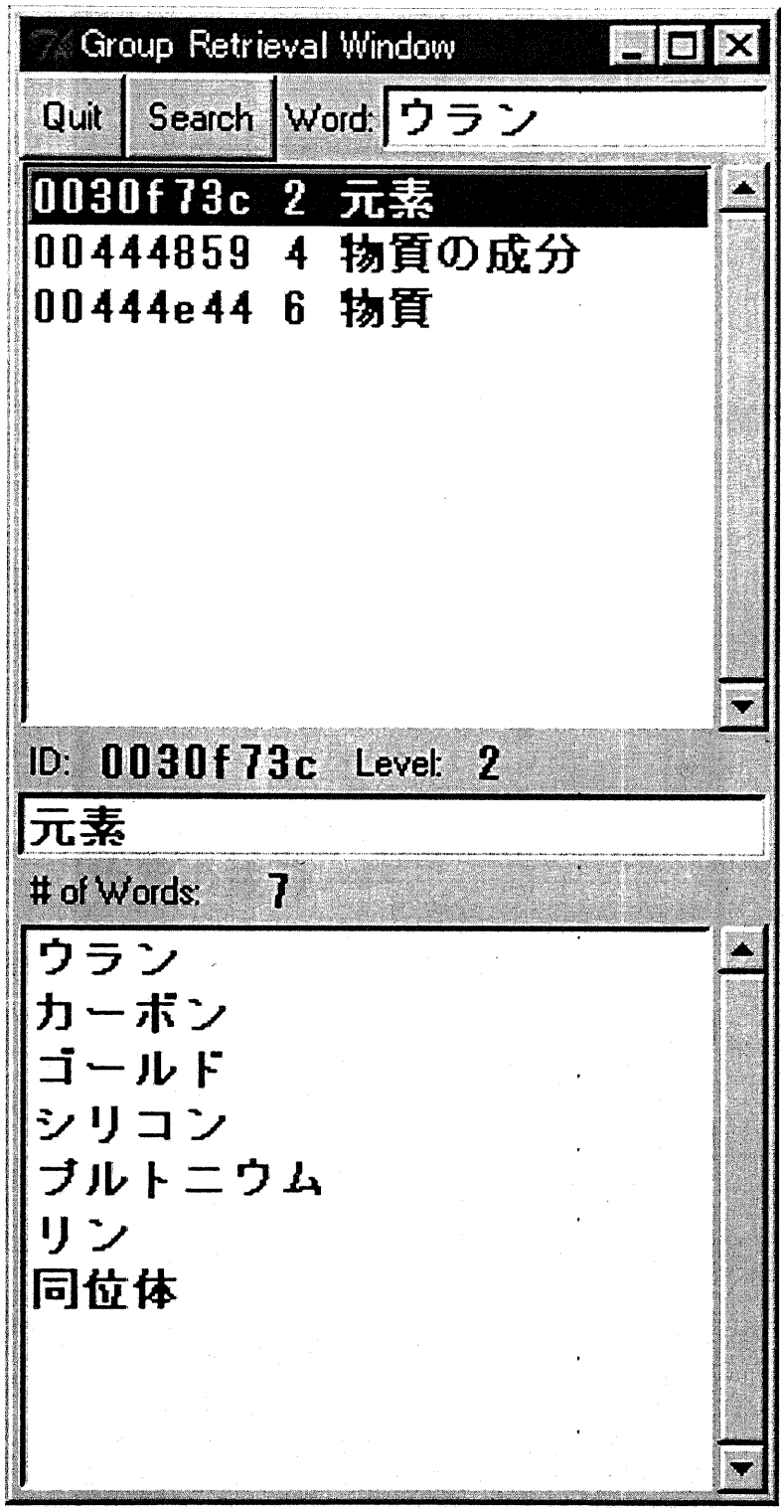


図 6.9: Group Retrieval Window

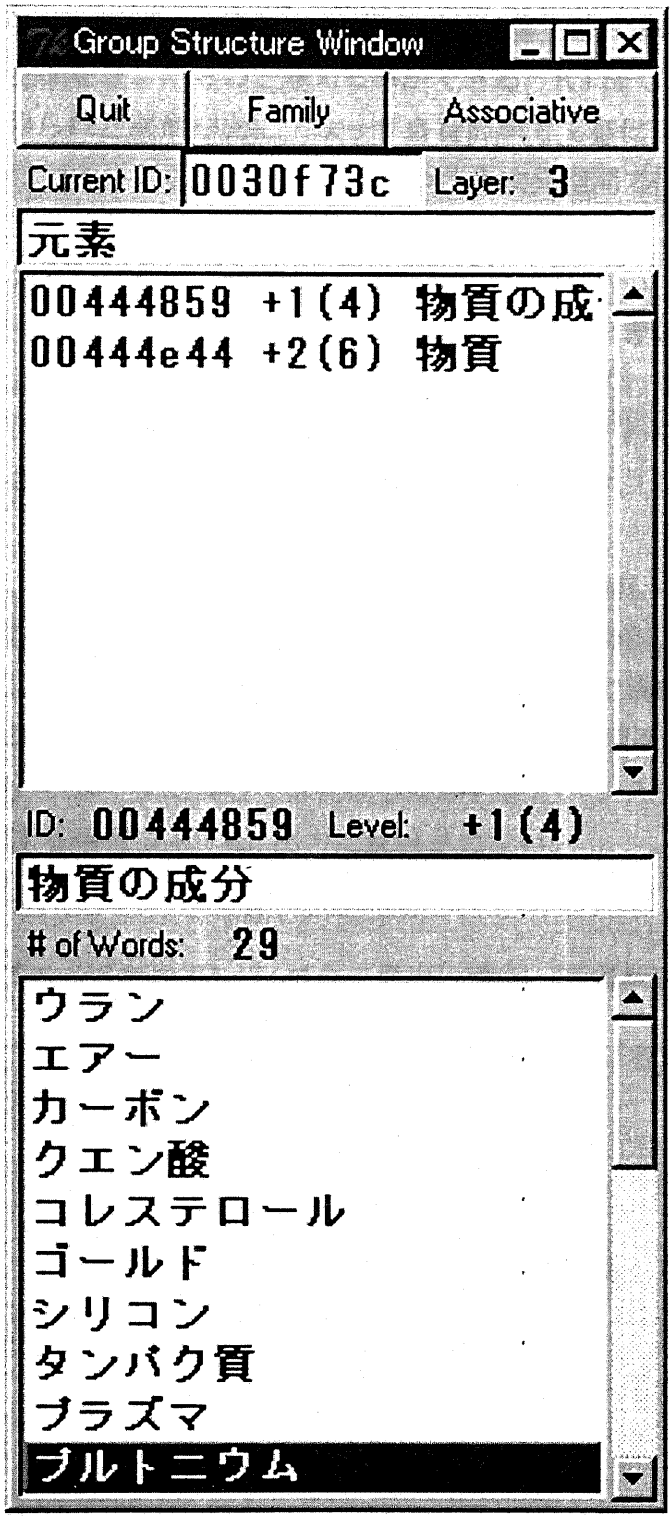


図 6.10: Group Structure Window

6.3.4 整理結果の出力

文書整理プロセスにおいてユーザが作成した図解は，文書を読み進めた結果得られたユーザの思考空間，概念空間を表したものと考える事ができる．KJ 法では，KJ 法 A 型（狭義の KJ 法）によって得られた図解を元に，ユーザ自身の手で文書化する KJ 法 B 型というプロセスが推奨されている．この作業は，図解という多次元空間で行われた思考結果を文章という線形空間に変換する事により，曖昧性や無駄なものを排除し，本質のみを追求する上で重要であるとされている．

しかし図解の文章化という作業はユーザにとって結構負担の大きい作業であるし，研究のサーベイや簡単なレポート代わりとして利用する時など，わざわざ文章に直す必要がない場合もあるだろう．また，計算機で扱う場合には線形性にこだわらず，相互参照を自由に設定でき，読み手に読解順序を委ねる事のできるハイパーテキストの方がむしろ便利であると考えられる．

そこで本研究では，ユーザが最終的に完成させた図解を，HTML ページに変換して出力できる様にする．これは，4.3.3 節で紹介した文書ディレクトリの個人専用版として捉える事ができる．具体的には図 6.11 に示すように，各グループ単位で構成し，階層構造を生かして構造化する．すなわち，各グループに記述される情報は，子関係にあるものであり，孫関係以上にある文書，グループに関しては記述しない．また，システムによって張られた文書間のリンクに関する情報は HTML 出力に反映させない．

各グループ毎に記述される情報は以下の通りである．

視点情報 視点意味グループ毎に ID，レベル，見出し情報，所属単語

内部子グループ 子グループ毎にグループ番号，内部文書数および各文書番号，内部グループ数および各グループ番号

内部ノード（文書） 文書毎に文書番号，タイトル，本文

文書番号，グループ番号からは，対応する文書，グループに関する実際の情報が記述されている部分へのリンクを張り，ハイパーテキスト化する．

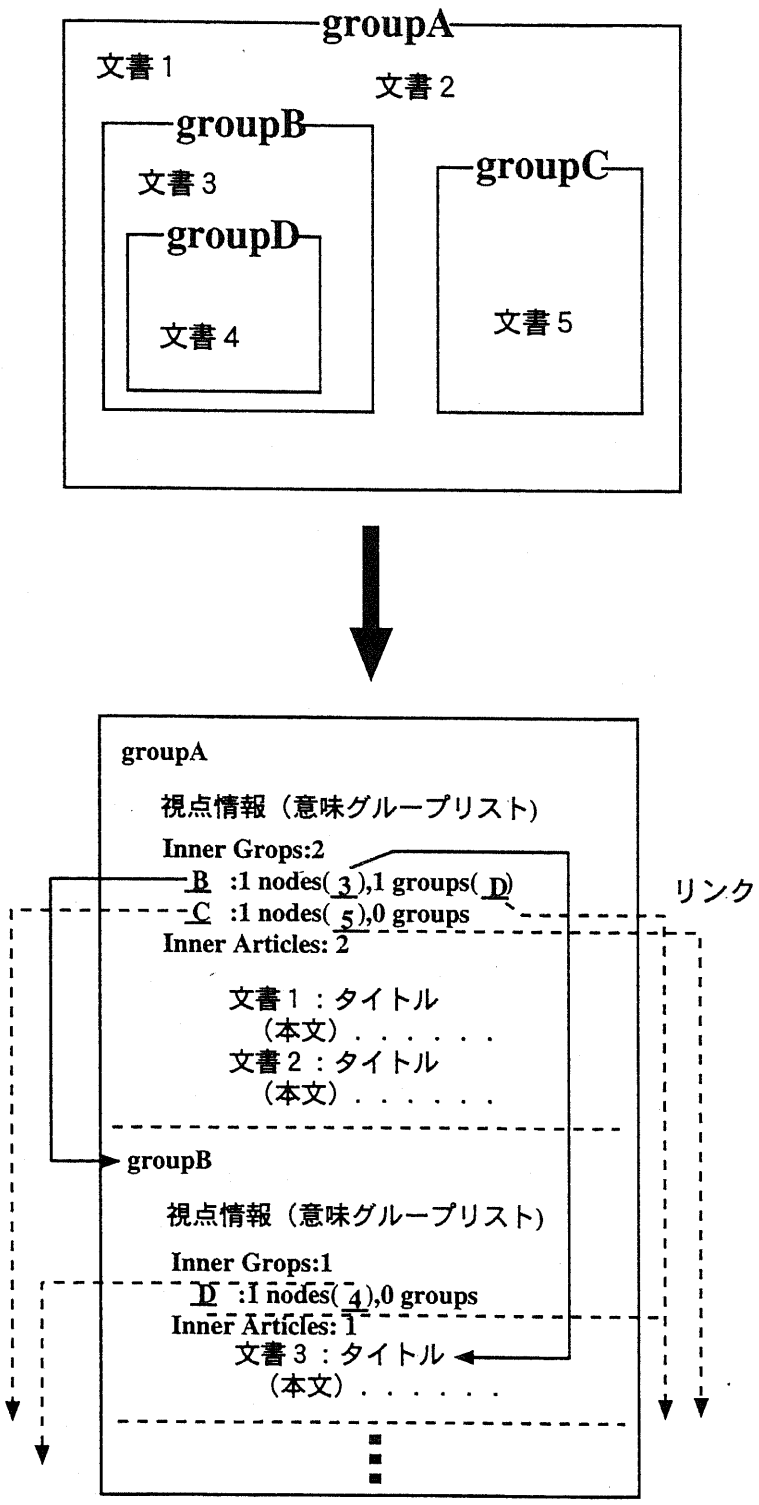


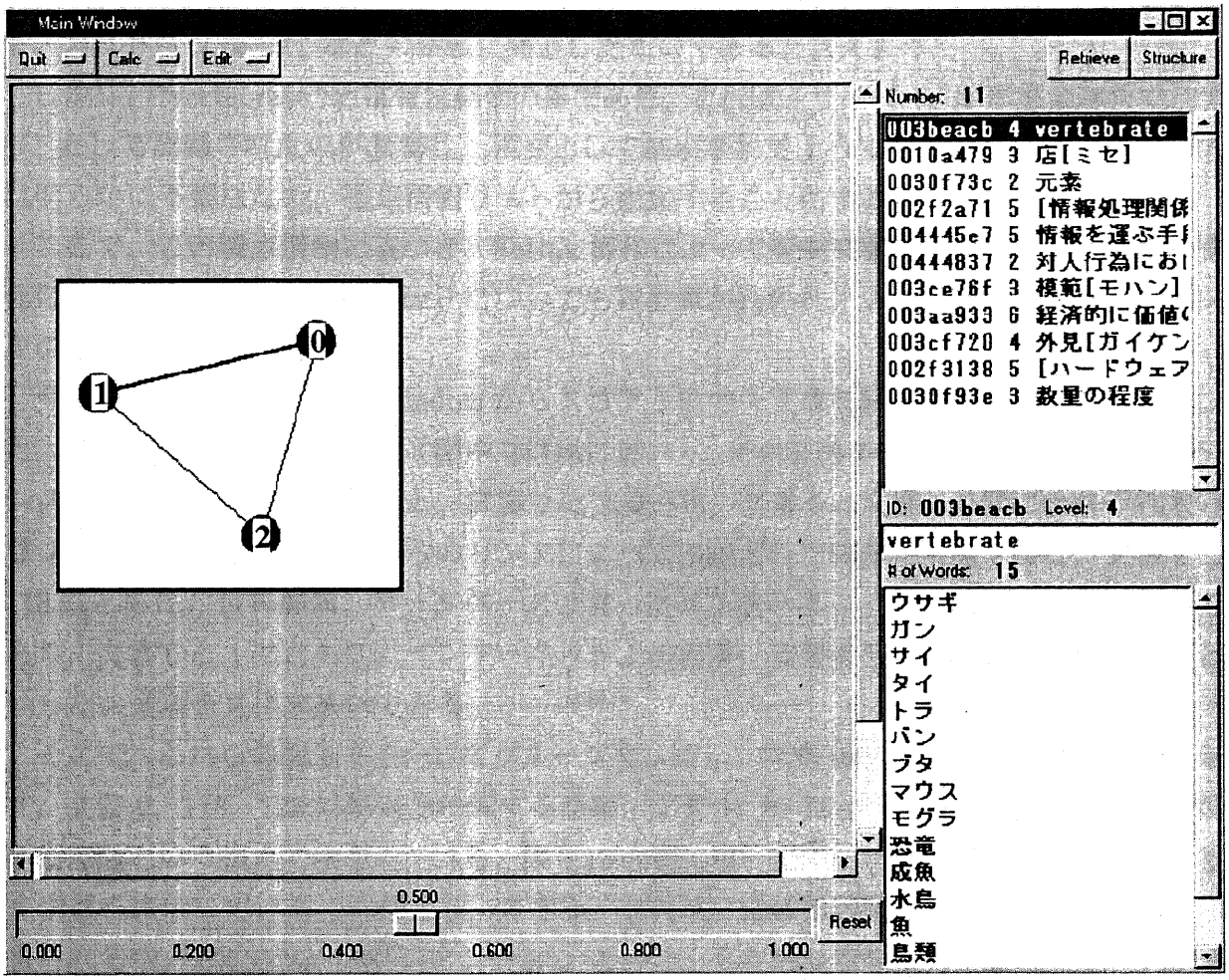
図 6.11: 図解からの HTML 文書の生成

6.4 Fish View を用いた文書整理プロセス

Fish View を用いた文書整理プロセスについてのイメージを把握してもらうために，ここではFish View を用いて実際に文書整理を行った結果を，順を追って紹介する．ここでは，5 節の評価実験で用いたニュース記事（日本語）を例題として用いている．ここで示す文書整理支援プロセスは非常に簡潔なものであり，Fish View に用意されている機能を全て使ったものではない事に注意されたい．

ステップ 1

まず始めに，最初に読みたい文書をいくつか，Document List Window から選択する必要がある．ここでは，文書番号の小さい方から 3 文書を選んでキャンバスに追加し，内容を確認した．その結果，3 文書ともインターネット関連である事が判明したためグループ化し，FIX 状態にする事によって視点情報を抽出した．この結果を図 6.12 に示す．ここで，図の下に記しているのは各文書のタイトルである．また，ノードが対応する文書については，カーソルをノード内部に入れる事によりタイトルを確認できるが，図中では代わりに文書番号を付記している．



0: ウェブ広告を見るたびにお金が入る「サイバーゴールド」
1: 増加するサイバースパイの実態に迫る
2: 情報価値の高いウェブサイトを案内：ブリタニカ・インターネット・ガイド

図 6.12: FIX グループ化による視点抽出直後

ステップ 2

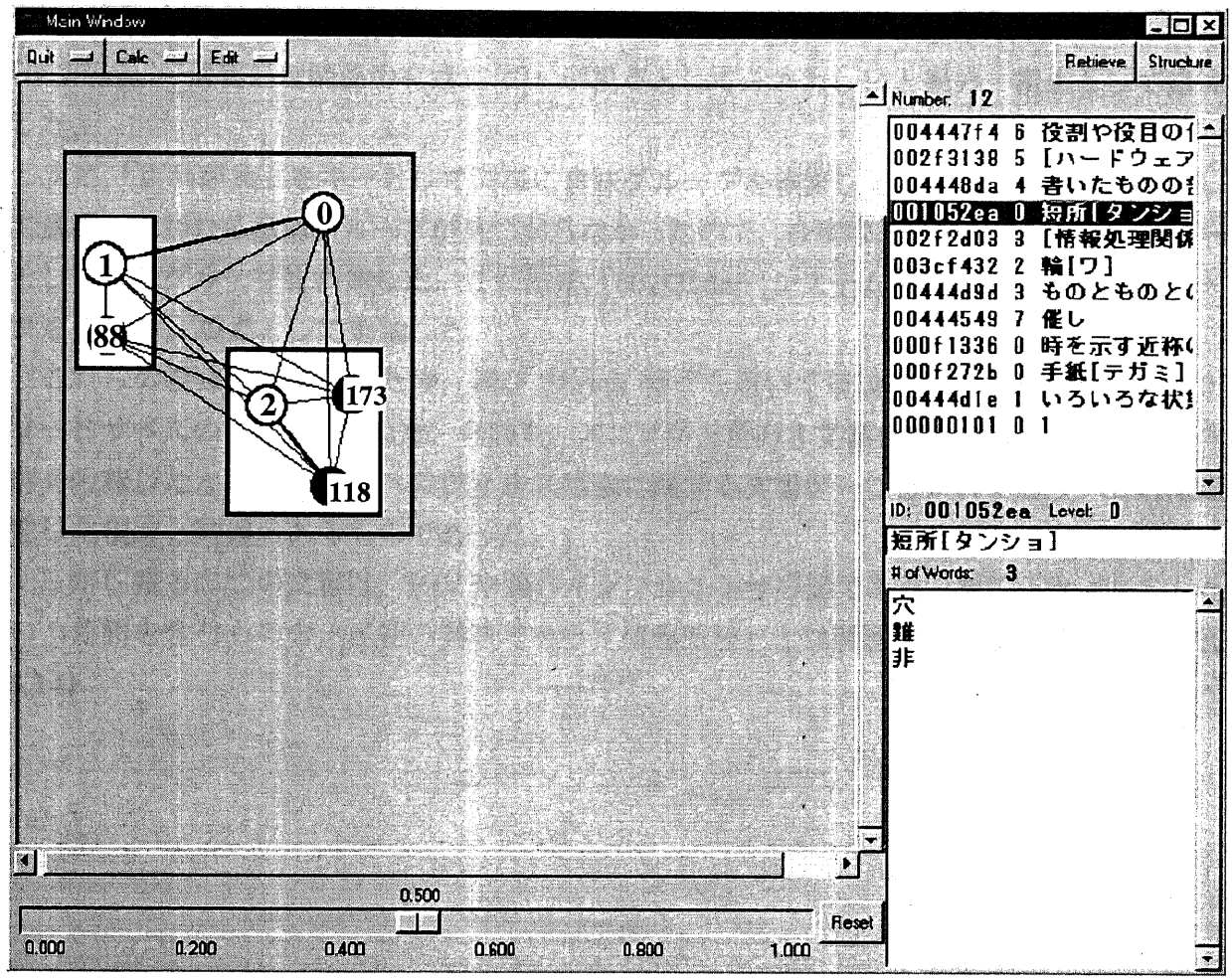
図 6.12 の右側に表示された視点情報リストを見ると,「002f2a71 : [情報処理関係の場所]」,「004445e7 : 情報を運ぶ手段やシステム」,「002f3138 : ハードウェア操作」などは視点に関連していると言える. また,「003aa933 : 経済的に価値のあるもの」や「00444837 : 対人行為における役割で捉えた人間」なども,「サイバーゴールド (文書 0)」や「インターネット・ガイド (文書 2)」などの話題に関連したものであると言えよう.

しかし, "vertebrate" とは脊椎動物の事であり, これは「サイバー」が形態素解析で "サイ" として認識されたのが影響し, 視点として抽出されたようである. この意味グループは明らかに不要なため, 視点情報リストから削除することにする.

ここで, この視点情報に基づき, Shrink 操作によって新規文書の検索を行った. この操作により, キャンバス内のノードについても再計算が行われ, ノードサイズの変更, リンクの張り直しなどが行われる.

今回は, Document List Window からスコア上位の 3 文書を新規にキャンバスへ追加した. この時, 新規ノードは青 (図中では黒に近い), それ以外のノードは白で区別される. システムによるリンクを見ると, 文書 2 と文書 118, 文書 0 と文書 1 の間にそれぞれ太いリンクが張られている. これらのリンクによって表現された関係を吟味すると, 文書 2 と 118 はともに, 百科辞典, データベースと言ったインターネット上の情報源に関する話であり, 文書 0 と 1 はともに, インターネット上の経済, 価値あるものに関する話であるといった共通項を見出す事ができる.

そこで, これらの関係を生かしてグループ化し直し, 図解を作成した結果を図 6.13 に示す. 文書 2, 118, 173 は情報源に関する話題, 文書 1, 88 はインターネット上の犯罪に関する話題という視点でまとめている. 図 6.13 右側には, 文書 1, 88 のグループに関する視点情報を表示している. これより, さきほどと同様に情報処理関係の視点意味グループが見られる中,「001052ea : 短所 [タンショ]」という, 否定的な意味合いの意味グループが選択されてされているのが興味深い.



88: ウェブの悪質な情報の対策を検討
118: アレルギーに関する情報を満載，ウェブサイト「アレデイズ」
173: コンピュータ・ウィルスの情報を提供

図 6.13: 犯罪に関連する文書グループからの視点抽出直後

ステップ 3

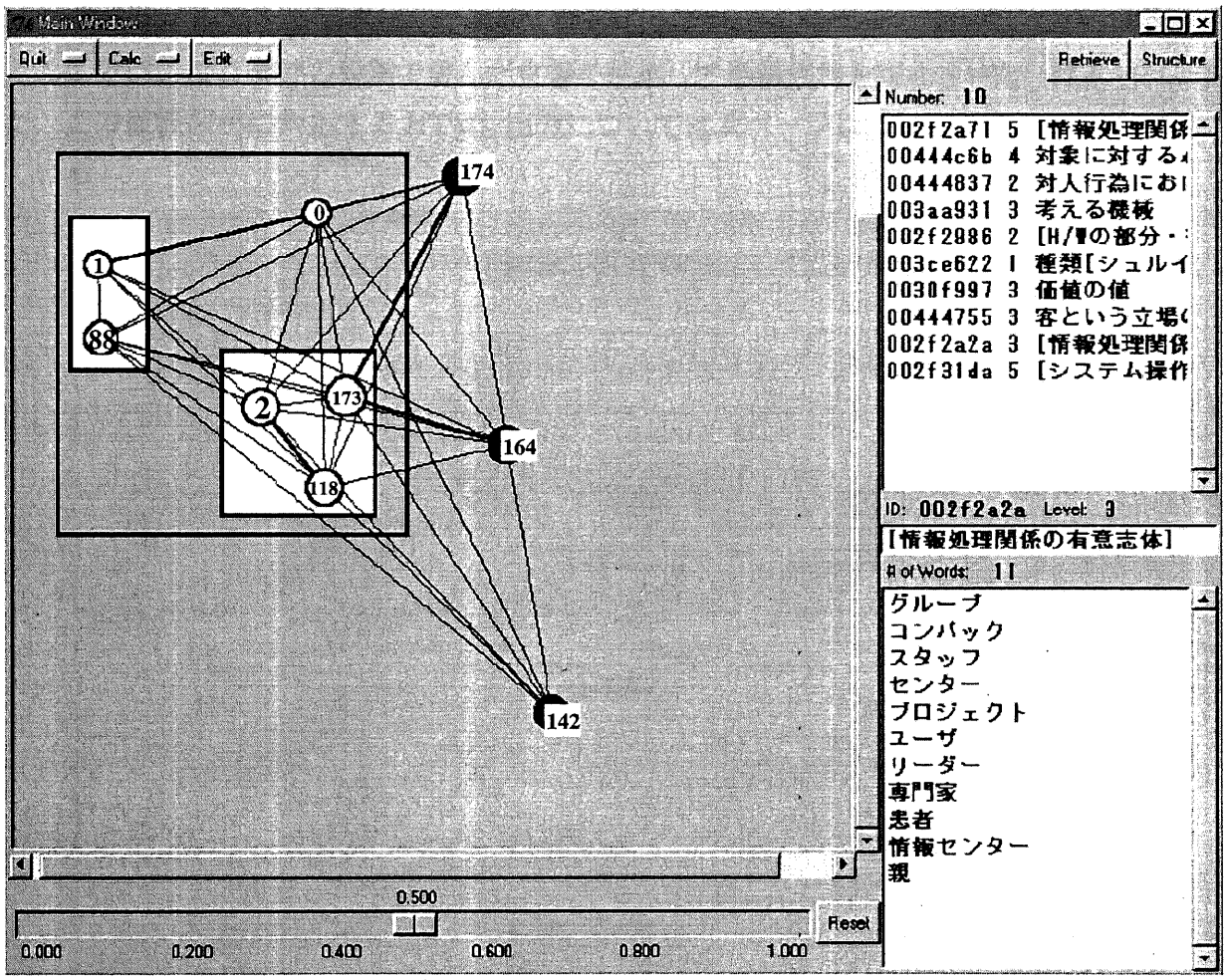
次に，文書 2，118，173 のグループを FIX 状態にし，抽出された視点を吟味したところ，当初の意図である「情報源」に関連する意味グループが少ない様に感じられたので，新たな視点の追加，入れ換えを行う事にする．

まず始めに，Group Retrieval Window に単語「専門家」を入力し，これを含むグループを検索したところ，いくつかの意味グループが検索結果として返された．この中から「002f2a2a：[情報処理関係の有意志体]」が視点としてふさわしいと思い，視点として追加した．

次に，「百科辞典」をキーとして同様に意味グループを検索したところ，現在の視点として適当と思われる意味グループが見つからなかったので，今度は「提供」という単語をキーとして意味グループを検索した．この結果，「002f31da：[システム操作]」という意味グループを選択し，視点として追加した

これらの視点編集を行った後，再び Shrink 操作を用いて新規文書の検索およびリンク，ノードサイズの再計算を行った．今回も，スコアが上位の 3 文書をキャンバスに追加し，内容を吟味したところ，文書へのウィルス感染に関する文書が一つ含まれており，これは指定した視点と関連のあるものであった．

しかし残る二つの文書は，HIV や鶏のインフルエンザの話であり，同じ「ウィルス」という用語を含むものの，当初の興味である「情報処理」とは関連のないものであった（図 6.14）．



142: 細胞に魔法をかけて侵入するHIV／HIVの狡猾で複雑な性質
164: 鶏のインフルエンザ「H５N１」型の研究
174: 安全地帯を設けてウィルス感染を防ぐ「Eセーフ・プロテクト」

図 6.14: 「情報源」に関する文書検索

ステップ 4

前回の検索において，当初の興味とは異なる文書が上位で返されていた．これは，情報源に関する文書についてはデータベース中に未読のものが存在しなくなったためかも知れないし，あるいは視点意味グループが適切ではないためかも知れない．従って，視点情報を再び編集し，検索をやり直す事もできるし，あるいは情報処理におけるウィルスと生物におけるウィルスとの関連などに興味を移してもよいだろう．

ここでは，ここまでの段階で読んだ文書を整理して最終図解を完成させた（図 6.15）．また，この最終図解に対応する HTML 出力については次ページ以降に付す（リンクは省略している）．

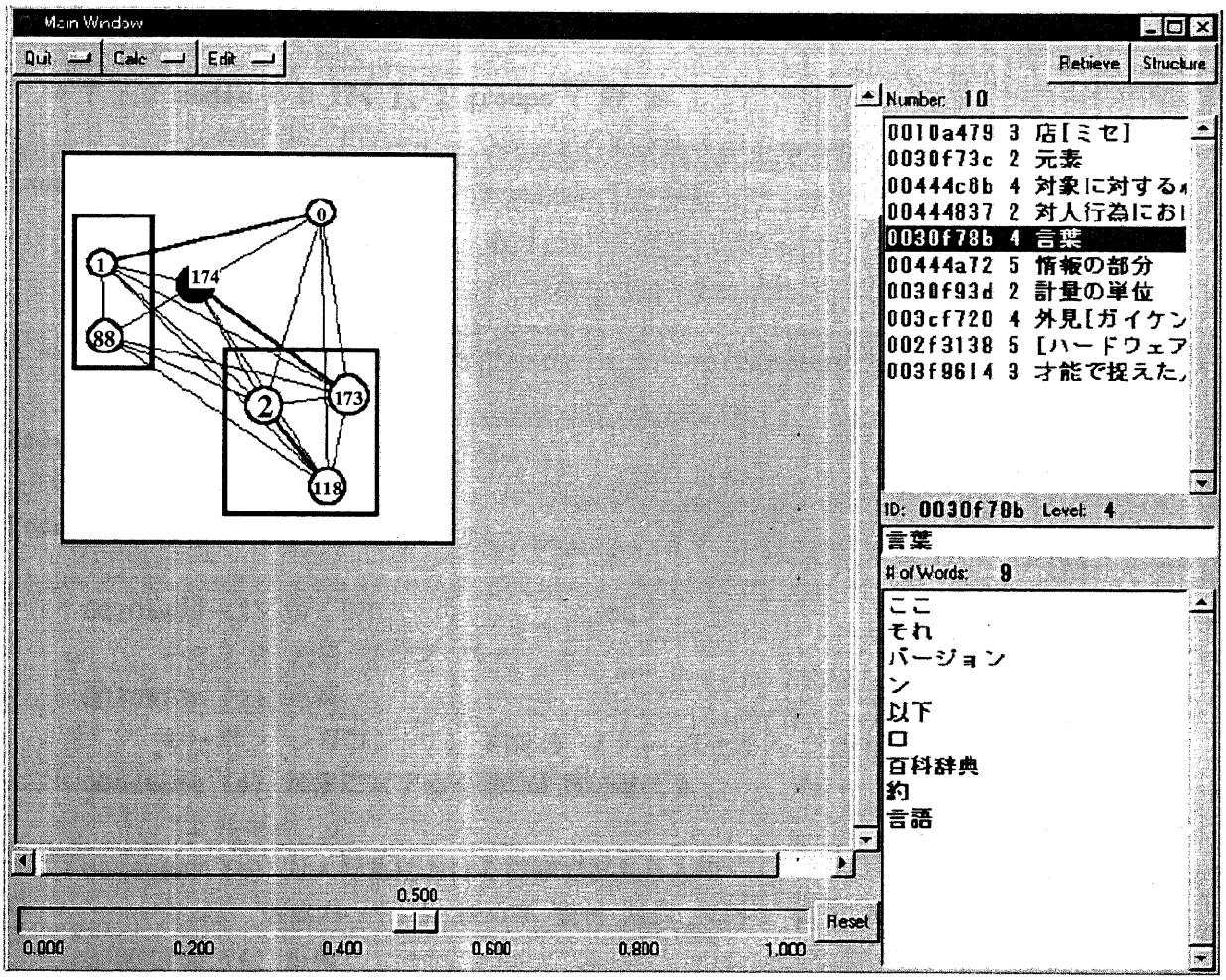


図 6.15: 最終的に得られた図解

最終図解の HTML 出力

作成日 : Mon Jan 11 14:33:16 1999

Group: all

Top of All Groups

Inner Groups: 1

* 7 : 2 nodes (0 174), 2 groups (27 26)

Inner Articles: 0

Top

Group: 7

Focus: 10

0010a479 (3) 店 [ミセ]
キオスク 売店 クラブ バー
0030f73c (2) 元素
カーボン シリコン リン 同位体 ウラン ゴールド プルトニウム
00444c6b (4) 対象に対する心理的距離が離れる
アレルギー 分
00444837 (2) 対人行為における役割で捉えた人間
ガイド 道案内 バック 決め手 受付 窓口 代わり 兵士 兵隊
0030f78b (4) 言葉
百科辞典 ン ここ それ 以下 バージョン 口 約 言語
00444a72 (5) 情報の部分
以上 右 言葉 ケース タイトル テーマ 口座 ヘッド ユニット 話 話題
0030f93d (2) 計量の単位
カーン カロリー キロ コード サイクル セクション フレーム マッハ
メートル 月 度 年 歩 満 目 ドル 両
003cf720 (4) 外見 [ガイケン]
案内 スタイル 構造 手 姿 フレアー 影 形状 骨格 姿勢 肉 アップ カット 髪形

002f3138 (5) [ハードウェア操作]
グラウンド 振動 シールド 充電 切断 フィードバック 超伝導 アクセス
コピーする 指示 スキャン 拡大 排出 液晶ディスプレイ 表示 ビーム
003f9614 (3) 才能で捉えた人間
成人 政治家 人 プロ 通 名手 人物 卵

Inner Groups: 2

* 27 : 2 nodes (1 88), 0 groups
* 26 : 3 nodes (2 173 118), 0 groups

Inner Articles: 2

0 : ウェブ広告をみるたびにお金が入る「サイバーゴールド」
「サイバーゴールド」は、インターネット上の広告を見る度に、所定の金額が振り込まれるサービスです。 お金は「サイバーゴールド」の個人口座に振り込まれ、広告を見る度に残高が増えていきます。 そのお金は、インターネットでの買い物や慈善事業への寄付、あるいはクレジットカードへ移すことができます。
174 : 安全地帯を設けてウイルス感染を防ぐ「Eセーフ・プロテクト」
この新しいウイルス撃退プログラムは、コンピュータ内部に「サンドボックス（sandbox)」と呼ばれる安全地帯を設けることにより、オンラインでやりとりする全てのプログラム、添付文書、アプリケーションをこの地帯に留置、ウイルスの有無を監視します。有害な動きが察知されると、コンピュータの安全を保ちながら、ユーザに警告を発し、標識づけを行います。

Parent Group Top

Group: 27

Focus: 12

004447f4 (6) 役割や役目の値
スパイ キャラクター 善玉 センター 呼び出し コントロール スタッフ
リーダー 会員 外部 代表 教授 員 分子 大統領 会長 長 兵士 宿主
002f3138 (5) [ハードウェア操作]

グラウンド 振動 シールド 充電 切断 フィードバック 超伝導 アクセス コピーする 指示 スキャン 拡大 排出 液晶ディスプレイ 表示 ビーム

004448da (4) 書いたものの部分

ユニット テーマ

001052ea (0) 短所 [タンショ]

穴 難 非

002f2d03 (3) [情報処理関係の抽象物]

音楽 言語 クラス 類 ターゲット プロトタイプ 目標 プロセス 過程 計画 目的 音 音声 声 原子 視覚 内容 問題 利点 モデル 戦略 聴覚 視力 現実世界

003cf432 (2) 輪 [ワ]

リング 輪 ハードディスク

00444d9d (3) ものとものとの関係にかかわる行為

連絡する 結合 合成する 組み換え 統合する 混同する 選 対 配布する 一緒 結合する 弁 スプレー 吹き付け 接続する 付け 締め 根絶する 増加する 入力する 補

00444549 (7) 催し

クラブ ショー ファイト 呼び物 ゲーム シリーズ エキスポ シンポジウム 会議 学会 国際会議 総会 競技 タイトル 場所 争い 式 連邦議会

000f1336 (0) 時を示す近称の指示代名詞

ここ これ

000f272b (0) 手紙 [テガミ]

書 消息 状 通信

00444d1e (1) いろいろな状態

オンライン リスク 危険性 花 休眠する 自動 実態 地理 天然 波 疲労 幕 慢性 無重力 流れ

00000101 (0) 1

1 一 一つ

Inner Groups: 0

Inner Articles: 2

1 : 増加するサイバースパイの実態に迫る

個人情報へのアクセス、口座の改ざん、あるいはハードディスクの情報の破壊などを行うサイバー・スパイが増えています。ウェブ・サーバを持つ人は特に危険です。そこで「ＩＳＳ」社などのメーカーでは、「アダプティブ・セキュリティ・マネジメント・ソフトウェア」を販売しています。この種のソフトは、社内ネットワークや、ファイアーウォール

を通る通信をモニターし、疑わしいアクセスに対しては即座に管理者へ連絡するか、自動的に接続を切ってしまいます。

88 : ウェブの悪質な情報の対策を検討

ここ数年、インターネットは、子供に有害な情報をもたらしているとして非難を浴びるようになりました。そこでウェブの悪質な情報に対して立ち上がった企業の一つが、デジタル・イクイップメントです。今回は、同社が有害な情報の規制をテーマに主催した会議の内容をお伝えします。

Parent Group Top

Group: 26

Focus: 10

002f2a71 (5) [情報処理関係の場所]
サイト 位置 場所 領域 ルート 穴 施設 世界 接触 内部 半分 最後 道 パス 軌道 契約 局 項目

00444c6b (4) 対象に対する心理的距離が離れる
アレルギー 分

00444837 (2) 対人行為における役割で捉えた人間
ガイド 道案内 バック 決め手 受付 窓口 代わり 兵士 兵隊

003aa931 (3) 考える機械
コンピュータ ハード ハードウェア パソコン 人工知能 プリンタ ロボット 産業用ロボット 液晶ディスプレイ

002f2986 (2) [H/W の部分・部品]
スイッチ モーター 脚 端子 チップ トラック ヘッド 案内 ワイヤ 膜

003ce622 (1) 種類 [シュルイ]
型 種類 段階 類 カテゴリー タイプ 部門 比

0030f997 (3) 価値の値
価値 値 ゼロ 幹 根本 質 首 重さ 重要性 数 生命 花 生命線

00444755 (3) 客という立場の人
ビジター 消費者 旅客 クライアント

002f2a2a (3) [情報処理関係の有意志体]
親 グループ コンパック センター プロジェクト ユーザ リーダー 患者 専門家 情報センター スタッフ

002f31da (5) [システム操作]

イメージング コミュニケーション 介入 識別 提供 インストール シミュレーション テスト モニタリング 管理 供給 検証 使用 実現 終了 準備 制御 制御する 設計 設置 劣化 サービス 処理 選別 検索 インストールする 操作 情報サービス 通信サービス 選択 送信 通信 ブロック 交換 情報交換 データ伝送 ファックス 転送 AND アンド アルゴリズム マッピング 配置

Inner Groups: 0

Inner Articles: 3

2 : 情報価値の高いウェブサイトを案内 : ブリタニカ・インターネット・ガイド

「ブリタニカ大百科辞典」のブリタニカ社が制作・管理する「ブリタニカ・インターネット・ガイド」では、特定のトピックごとに何百ものサイトから選りすぐられた最も有益で専門的なウェブサイトの情報を提供しています。現在、１４のカテゴリーで６万５千以上のサイトを案内しています。

173 : コンピュータ・ウイルスの情報を提供

シマンテック社では、危険なコンピュータ・ウイルスの流行を防ぐため、コンピュータ・ウイルスの被害防止に役立つ情報を提供する「アンチウイルス研究センター＝ＳＡＲＣ（サーク）」を開設しました。その情報には、１万種類以上のウイルスについて解説した世界最大のコンピュータ・ウイルス百科辞典なども含まれています。

118 : アレルギーに関する情報を満載、ウェブサイト「アレデイズ」

およそ４千万人のアメリカ人が、季節性のアレルギー性鼻炎に悩まされていると言います。そこで、「アレデイズ」と呼ばれるウェブサイトでは、アレルギーに悩む人々のために様々な情報を提供しています。「アレデイズ」のビジターが、このサイトに詳細な質問を送ると、著名なアレルギーの専門医が答えてくれる上、花粉情報、避けるべき植物の種類、家の中からアレルゲンを無くす方法などのセクションもあります。

Parent Group Top

6.5 評価・考察

ここでは，Fish View を実際に複数のユーザに使ってもらった結果について報告する．WWW から，映画の批評に関する文書を 101 用意し，これを用いて文書整理を行ってもらった．ユーザの負担を避け，実験という事をあまり意識せず気楽に行ってもらうため，各自の研究室で好きな時間に実験を行ってもらった．このために留意したのは以下の点である．

- ホームページからクライアントプログラムをダウンロードできるようにした．
- UNIX(X Window) を利用可能なユーザに対しては，本研究室に実験用アカウントを用意し，Tcl/Tk やクライアントプログラムをダウンロードしなくても実験が行えるようにした．
- ホームページにオンラインマニュアルを用意した[2]．
- サーバ側でユーザとの通信ログをとり，ユーザが行った操作を後で分析できるようにした．
- ユーザには，実験結果として最終出力の HTML 文書と，アンケートへの回答を提出してもらった

電気・電子・電子情報工学科の学生 10 人程度に依頼し，実験を行ったところ，7 人から有効な回答が得られた．表 6.2 は，アンケートの回答結果をまとめたものである．ここでアンケートの各項目については，1 を最低，7 を最高とする 7 段階で回答してもらい，1–3 を “×”，5–7 を “○” として集計してある（Magnify, Shrink どちらの操作を主に用いたかについては 1(Magnify)—7(Shrink) として回答してもらっている）．以下ではこのアンケート結果および HTML 出力，ログを元にして，図解作成，視点情報（可読性，検索性能，リンク生成能力）などに分けて考察する．

[2]http://www.miv.t.u-tokyo.ac.jp/~takama/fview_manual.html

表 6.2: アンケート回答結果

項目名	有効回答数	○ (5–7)	× (1–3)	平均値
応答速度	7	4	3	4.86
図解操作はしやすいか	7	3	4	4.29
入れ子構造が作成できるのは便利か	7	**5**	0	**5.57**
視点編集はやりやすいか	4	1	1	4
視点の自動抽出は便利か	7	3	1	4.43
視点情報はわかりやすいか	7	2	4	3.29
抽出された視点情報は適切か	7	2	**3**	3.86
文書検索結果は適切か	6	**3**	1	**4.83**
Magnify,Shrink どちらを主に用いたか	6	1	4	3.33
システムによるリンクは役立ったか	7	**4**	2	4.43
リンクは適切か	5	2	3	4
HTML 出力は読みやすいか	7	**5**	0	**5.29**
HTML 出力は便利か	7	**7**	0	**6.29**
今後も使ってみたいか	7	**5**	2	**5.29**
実験データはふさわしいものだったか	7	4	1	4.71

6.5.1 図解作成に関する考察

アンケート結果を見ると，図解操作のしやすさ，応答速度に関しての評価は賛否両論にわかれてしまっている．応答速度に関しては，ユーザ毎に通信経路を含め実行環境が異なる事も大きく影響しているだろう．

実験終了後に話を聞いたところ，ほとんどのユーザがこの様なシステムを使用した経験がないため何をしていいかわからなかったとの感想を述べていた．事前説明としてオンラインマニュアルの提供しか行わなかった事も，これに拍車をかけた様である．

GUIとしての設計上の不備や機能不足だけでなく，この様な事前説明不足，ユーザの経験不足もこの評価に影響を与えていると考えられ，反省すべき点である．

作成された図解の特徴は，以下の様であった．

図解中のノード数 平均17.14，最大29，最少5

図解中のグループ数 平均5.29，最大9，最少3

入れ子構造の深さ 平均1.57，最大4，最少1

最大グループに含まれるノード数 平均5.00，最大8, 最少2

作成された図解の一例を図6.16，図6.17に示す[3]．図中の数字は文書ノードを表す（イタリックで示された数字はグループ番号である）．これらの図解作成に要した時間は大体1時間前後であった．

アンケートの結果を見ると，入れ子構造を用いて図解を作成できる事についての評価は非常に高い事がわかるが，図6.16の様にかなり高階層の入れ子構造を作成したユーザは少なかった．また，グループのオーバラップができるようにして欲しいとの意見もあった．

また，図解をHTML化して出力する機能は便利さ，わかりやすさともに評価が非常に高いこともアンケート結果より確認された．

[3]この図はユーザから回収したHTML出力から図に直したものであり，画面写真ではない

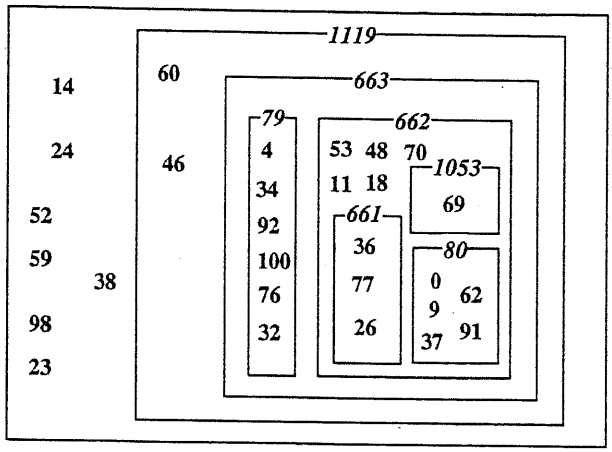


図 6.16: 作成された図解の例 (1)

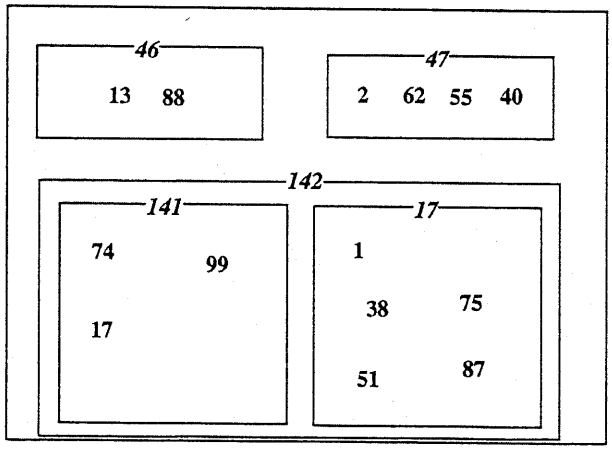


図 6.17: 作成された図解の例 (2)

6.5.2 視点情報に関する考察

視点情報に関する評価として，ここでは (1) 視点に基づく検索，(2) システムによる文書間のリンク生成，(3) 視点の外化（可読性）に分けて考察する．

視点に基づく検索

アンケート結果を見ても，文書検索結果は視点を適切に反映していたという評価をしたユーザの方が多い．これはアンケート結果だけでなく，実験終了後の感想で，「視点に基づく検索」を便利な機能としてあげているユーザが多かった事からも確認できる．

システムによるリンク生成

システムによるリンクが文書整理を進めていく上で，ガイドとして役立ったと評価した人が，そうでない人よりも多かったが，リンクの適切さについては文書検索ほど高い評価が得られていない．特に Magnify 演算を用いた場合の評価が低かった様である．これは，Magnify は視点意味グループに含まれる単語のみに特徴を限定するため，ベクトル空間が極端に低次元化する場合があるためだろう．従って，視点の厳選は視点情報の可読性や，Shrink 操作による類似度判定には有効であった反面，Magnify には逆効果であったといえる．

視点の外化・可読性

検索・リンク生成と比較して，外化された情報としての視点意味グループの評価が低いユーザの方が多かった．この評価には，

- 視点情報の表現方法にユーザが慣れていなかった．
- 視点として抽出された意味グループが不適切であった．

の両要因が考えられる．前者について，本研究では視点を意味グループの集合体として表現しているが，この様に表現された視点情報と，自分の頭の中にある視点との対応がわかりづらかったようである．これは前述の様に，ユーザが文書整理支援ツールなどの使用経験がほとんどなく，事前説明も不足していた事も大きく影響しているだろう．これは，アンケートにおいて視点編集機能を利用したと回答したユーザが少なかった事とも関連しているであろう．

しかし，実験ログを分析したところ，視点編集機能を多用し，使いこなしているユーザもいることが確認された．このユーザの行動を詳しく見てみると，「自分が大事と思う単語

を含む，新たな意味グループを検索し，視点に追加」したり,「ある単語を含む意味グループを他の意味グループに入れ換える」などといった我々が想定した編集作業の他に,「視点とは関係ないキーワードを視点情報から偶然見つけ，それに関する意味グループを検索する」などといった,「視点に関するブラウジング」とでもいうべき活動も行っていた．この様な視点編集の利用方法は我々の想定を越えたものであり，概念体系を対応づけ，可読な形式で視点情報を提供する我々の提案手法の可能性を感じさせる点で，非常に勇気づけられるものである．他のユーザについても，システムの利用経験を重ねていけば，システムの提示する視点情報の読解に慣れると同時に，視点編集機能も多用する様になっていく事が期待される．

後者の，抽出された視点意味グループの不適切さについては，大きく分けて次の 4 つの問題点が存在すると考えている．これらは前章の考察において触れた問題点と同様である．

- 視点抽出アルゴリズムが不十分である
- 話題・視点に対応する適切な概念が概念体系辞書中に存在しない
- 単語抽出のミス
- 未知語の存在

視点抽出アルゴリズムの改良については 5.5.2 節 でも触れた様に，仮説推論を用いた視点意味グループの抽出について検討中である．ここでは残り 3 点について考察する．

適切な概念が存在しなかった例として，今回の実験では映画の批評記事を用いたため，ホラーやサスペンスなどのジャンルでグループ化するユーザが多かったが，その様なジャンルにうまく対応する概念が存在しない場合もあった．例えば，サスペンスやアクションは，見出し情報としては意味合いが若干異なるものの，それらを単語として含む意味グループが存在する．また，怪獣映画（ゴジラなど）については,「軍事組織」や「神話や伝説や人々の心の中でのみ存在する擬似人間や擬似生物」などの意味グループの組合せでうまく表現する事ができるが，コメディなどは対応する概念（群）が存在せず，視点編集を行ってもうまく表現できない様に感じられた．

単語抽出のミスとしては，6.4 節にあったような,「サイバー」を一語としてではなく,「サイ」までで区切って認識・抽出してしまっているなどが代表的なものである．5.5.2 節でも考察した通り，単語抽出のミス，未知語の存在に関しては，概念・単語両辞書の編集・整備を行う必要があるだろう．

6.5.3 最後に

ユーザには，表 6.2 に示した項目だけではなく，便利／不便に思った機能や，システムの利用方法に関する提案などについても回答してもらった．

前述した様に，便利な機能としては視点に基づく検索や，図解を HTML 化し，出力する機能などがあげられており，Fish Eye マッチングを基盤とした視点抽出による支援の有効性，および図解を活用した文書整理，サーベイ作成といった著者の提案に対するユーザの支持が得られたと考えて良いであろう．

反対に不便な機能としては，Undo 機能がないことや，グループのオーバラップや FIX 化等に関する不満など，システム実装上の問題が多く，インタフェース部分の完成度の重要性を痛感した．

アンケートによると，Fish View の様なシステムを今後も使ってみたいと言う人が結構多く，以下にあげる様な多様な用途の提案が得られた．

- 論文の分類・整理・検索
- 新聞スクラップの整理
- ビデオ・ゲームの分類・整理
- 図解を用いた，芸能人などの印象，アピールポイントの分析

前にも述べた様に，文書整理支援システムなど，ナレッジマネジメント・知的活動支援に関するシステムはまだまだ一般に普及しておらず，さらには事前の説明不足もあってシステムの機能を十分に使いこなせてもらえなかった事が最も悔やまれる事である．しかしそもそも，事前に十分な説明，訓練が必要なシステムでは一般ユーザの日常的作業の支援ツールとして利用してもらえるはずもない．WWW ブラウザ並みにわかりやすく，直観的に利用してもらえるツールを実現して初めて，知的創造活動のシームレスな支援につながるであろう．

Chapter 7

結論

本研究の目的は，個人レベルでの情報整理支援システムとして，文書情報を整理し，理解・活用につなげるプロセスを支援するシステムの開発であった．

WWW 空間などの普及により，机の前にいながらブラウザを通じて世界中の情報にアクセス可能であり，あるいは学会などから CD-ROM に収録された形式で大量の論文が提供される様になった結果，自分の興味に関する文書が容易に，(個人で扱うにしては) 大量に入手可能となっている．その結果，とりあえず入手したものの全然目を通す事なく，いわゆる「積ん読」状態になっている文書がかなりの量に達するようになってしまった．

一般の個人レベルにおいて，入手した情報量と，その活用度は反比例の関係にあると考えられる．コピーがなく，必要な情報は本を購入するか，書き写すしかなかった時代には，情報入手のコストが高い分，入手文書を真剣に読み，利用したと考えられる．また，「書き写す」と言う行為時代が，内容を理解する上で有効であったとも言える．これに対し，コピーや電子メディアが普及した現在，情報入手にかかるコストは劇的に低下し，とりあえず入手しただけで満足してしまいがちである．この様な現状に対する筆者自身の反省および改善への要求が，本研究の出発点であったと言える．

本研究で提案した解決案は，「情報の整理を通じた熟読が，入手した大量文書情報の活用につながる」というものである．従来の情報整理支援に関する研究では，文書ディレクトリをシステムが自動的に生成してくれたり，あるいは断片的な，メモ的な情報をユーザが入力し，それを図解を用いて整理する過程を支援するものなどが多かった．しかし，前者では上述のコピーにおける問題と同様，ユーザ自身が整理・理解した気になってしまい，読まずに満足してしまう可能性がある．後者においても，入手文献を読み，理解すると言う行為はシステム利用前にユーザ自身が行うべき行為であって支援対象外となっている．これに対し本研究では，「入手文献をユーザ自身が実際に読む事なしには，理解・活用は不可

能である」，および「実際に整理した者が，最も効率よく利用できる者である」という考えに基づき,「整理過程を通じた文書の熟読」を支援する事によって，大量文書情報を活用できる事を主張している．

上記目標を達成するためには，動的に変化するユーザの視点を捉え，有効に活用する事が必要との考えから，本研究では Fish Eye マッチングと呼ぶ，概念体系に基づいて視点情報を明示的に扱う事のできる文書マッチング手法を提案し，その有効性を文書検索実験によって確認した後，これを基盤技術として用いた文書整理支援システム Fish View を開発した．実際のユーザによる評価実験の結果，システムの完成度や，使用した概念体系の整備などについてまだまだ洗練不足の点も見られたが，提案した文書整理プロセスおよび，それを支援する Fish View の様なシステムの必要性，将来性についてはほとんどのユーザが評価するものであった．

今回の評価実験でも明らかになった事だが，本研究で提案したような，個人レベルでの情報整理支援システムの認知度，普及度はともに低く，今後の発展が期待される．より多くのユーザに認められ，利用されるシステムを開発するには，現在爆発的に普及している NetScape や Internet Explorer などの WWW ブラウザが参考になるであろう．すなわち，誰にでも使いやすく,（マニュアルなどを見なくても）わかりやすい操作を実現する事，および特別なツールを用いたり，特別な作業をしている様な印象をユーザに与えない事が重要であろう．机上で行う日常的な作業の一環として，特別な意識をしたり，肩に力を入れる事なく自然に使用してもらえる様になって初めて，知的活動全般におけるシームレスな支援の実現，およびユーザを取り囲む情報環境の最大限の活用につながるであろう．

謝辞

本研究を進めるにあたり，多くの方から御指導・御鞭撻を賜わりました．

大変御多忙の身でありながら熱心に御指導して下さいました，石塚満教授に深く感謝致します．石塚先生には，卒業研究，修士・博士課程と合わせて 6 年間もの長きにわたりお世話になりました．当初は人工知能どころか，研究というもの自体，全くわかっていなかった私でしたが，6 年間御指導頂いた結果，今後研究を続けていく上での基盤，自信といったものを多少なりとも持つ事ができました．研究者としてまだまだ未熟者ですが，今後の研究成果が先生へのご恩返しにつながるものとして頑張っていく所存です．今後とも，引続き御指導の程，宜しくお願い致します．

伊庭斉志助教授には，一年ほどの短い間でしたが，研究内容に関わる話だけでなく，論文査読の心構えなど，研究者としてためになるお話をいろいろして頂きました．深く感謝致しますと共に，今後とも御指導の程，宜しくお願い致します．

土肥浩助手には，6 年間の長きにわたり，研究生活全般について様々なアドバイスを頂きました．特に，修士課程では共同研究者としていろいろ御指導頂きました．深く感謝致します．

秘書の藤田メイコさんには，様々な事務手続きをして頂いただけでなく，研究室での生活を明るいものにして頂きました．研究室で過ごしたこの 6 年間は，とても楽しかった思い出としていつまでも忘れる事はないと思います．ありがとうございました．

研究室の皆さんには，研究だけでなく，計算機に詳しくない私のために色々助けて頂きました．多くの先輩，後輩達に囲まれて，楽しく，かつ意義のある 6 年間を送ることができた事に感謝致します．

猪股健太郎君，飯島光晴君は共同研究者として，システムの完成度，能力を高めるために御協力頂きました．感謝致します．

最後に，研究以外の様々な面で私を支えて下さった，家族，友人達に深く感謝致します．

参考文献

[1] M. Ackerman, et al. Learning Probabilistic User Profiles. *AI Magazine*, Vol. 18, No. 2, pp. 47–56, 1997.

[2] 相原健郎, 堀浩一, 大須賀節雄. 断片的な情報の集まりから知識を構築する過程の支援. 人工知能学会誌, Vol. 11, No. 3, pp. 432–439, 1996.

[3] G. Amati, F. Crestani, and F. Ubaldini. A Learning System for Selective Dissemination of Information. In *Proc. of 15th Int'l Joint Conf. on Artificial Intelligence(IJCAI97)*, pp. 764–769, 1997.

[4] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web. In *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[5] A. Bagga, J. Y. Chai, and A. W. Biermann. The Role of WordNet in The Creation of a Trainable Message Understanding System. In *Proc. of AAAI-97*, pp. 941–948, 1997.

[6] M. Balabanovic and Y. Shoham. Learning Information Retrieval Agents: Experiments with Automated Web Browsing. In *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[7] E. Bloedorn, I. Mani, and T. R. MacMillan. Machine Learning of User Profiles: Representational Issues. In *13th Nat'l Conf. on Artificial Intelligence (AAAI-96)*, Vol. 1, pp. 433–438, 1996.

[8] M. A. Boden. Creativity and Artificial Intelligence. In *Proc. of 15th Int'l Joint Conf. on Artificial Intelligence(IJCAI97)*, pp. 1563–1566, 1997.

[9] A. Z. Broder, S. C. Glassman, and M. S. Manasse. Syntactic Clustering of the Web. In *6th Int'l WWW Conference*, http://www6.nttlabs.com/HyperNews/get/PAPER205.html, 1997.

[10] R. D. Burke, et al. Question Answering from Frequently Asked Question Files. *AI Magazine*, pp. 57–66, 1997.

[11] R. D. Burke, K. J. Hammond, and E. Cooper. Knowledge-based Information Retrieval from Semi-structured Text. In *AAAI-96 Workshop on Internet-Based Information Systems*, pp. 9–15, 1996.

[12] 日経ビジネス編（編）. 日経ビジネス テーマスペシャル「ヒットを生み出す発想法」. 日経 BP, 1998.

[13] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual Information Retrieval: A Comparative Evaluation. In *Proc. of 15th Int'l Joint Conf. on Artificial Intelligence(IJCAI97)*, pp. 708–714, 1997.

[14] H. Chen and K. J. Lynch. Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 22, No. 5, pp. 885–902, 1992.

[15] W. W. Cohen. Text Categorization and Relational Learning. In *12th Int'l Conf. on Machine Learning*, pp. 124–132, 1995.

[16] W. W. Cohen and Y. Singer. Learning to Query the Web. In *AAAI–96 Workshop on Internet–Based Information Systems*, pp. 16–25, 1996.

[17] W. B. Croft and J. Xu. Corpus-Specific Stemming using Word Form Co-occurrence. In *Proc. 4th Annual Symposium on Document Analysis Information Retrieval*, pp. 147–159, 1995.

[18] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.

[19] S. T. Dumais, G. W. Furnas, and T. K. Landauer. Using Latent Semantic Analysis to Improve Access to Textual Information. In *CHI'88, Conf. on Human Factors in Computing*, pp. 281–285, 1988.

[20] EDR. http://www.iijnet.or.jp/edr/.

[21] D. Ellis (細野公男監訳). 情報検索論：認知的アプローチへの展望. 丸善, 1994.

[22] 遠藤聡志, 大内東. 統合型発想支援システム：FISM. 人工知能学会誌, Vol. 8, No. 5, pp. 611–618, 1993.

[23] 藤井敦. コーパスに基づく多義性解消. 人工知能学会誌, Vol. 13, No. 6, pp. 904–911, 1998.

[24] 藤崎博也ほか. キー概念に基づく情報検索システム方式の高度化 (2) – キーワードの同表記異義の処理 –. 情報処理学会第 57 回全国大会, pp. 3–239–240, 1998.

[25] 藤崎博也ほか. キー概念に基づく情報検索方式の高度化 (3) – キーワードに基づく方式との比較 –. 情報処理学会第 57 回全国大会, pp. 3–231–232, 1998.

[26] 福島伸一, 石塚満. WWW 情報空間におけるコアページの抽出と弱い情報化. 情報処理学会第 57 回情処全国大会, pp. 3–398–399, 1998.

[27] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communication. *Commmunications of the ACM*, Vol. 30, No. 11, pp. 964–971, 1987.

[28] Goo. http://www.goo.ne.jp.

[29] C. L. Green and P. Edwards. Using Machine Learning to Enhance Software Tools for Internet Information Management. In *AAAI-96 Workshop on Internet-Based Information Systems*, pp. 48–55, 1996.

[30] S. J. Green. Using Lexical Chains to Build Hypertext Links in Newspaper Articles. In *AAAI–96 Workshop on Internet–Based Information Systems*, pp. 56–64, 1996.

[31] 萩原雅之. Internet Survey Watching. http://www.mars.dti.ne.jp/ hagi/.

[32] 橋田浩一, 松原仁. 複雑性の海へ, Stage6: 情報の部分性. NTT 出版, 1994.

[33] 幡鎌博, 津田宏, 益岡竜介. ナレッジマネジメントへむけて – 知識検索・整理および基盤技術 –. 人工知能学会誌, Vol. 13, No. 6, pp. 912–919, 1998.

[34] M. A. Hearst and C. Karadi. Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. In *Proc. 20th Annual Int'l ACM/SIGIR Conf.*, 1997.

[35] R. R. Hightower, L. T. Ring, J. I. Helfman, B. B. Bederson, and J. D. Hollan. Graphical Multiscale Web Histories: A Study of PadPrints. In *ACM Conf. on Hypertext*, 1998.

[36] T. B. Ho and K. Funakoshi. Information Retrieval Using Rough Sets. 人工知能学会誌, Vol. 13, No. 3, pp. 424–433, 1998.

[37] 帆足啓一郎, 青木圭子, 松本一則, 橋本和夫. 類似検索における単語寄与度に基づく重要語選択手法の検討. 第 57 回情報処理学会全国大会, pp. 3–235–236, 1998.

[38] K. Hori. Concept space connected to knowledge processing for supporting creative design. *Knowledge-Based Systems*, Vol. 10, No. 1, pp. 29–35, 1997.

[39] K. Hori. Preface: Information Technology Support for Creativity. *Knowledge-Based Systems*, Vol. 10, No. 1, pp. 1–2, 1997.

[40] 星野匡. 発想法入門. 日経文庫, 1989.

[41] IM Lab. TRIZ 応用の発明発想支援ソフト IMLab 評価記事. http://www.sun-inet.or.jp/ igata/imlab.html.

[42] Infoseek. http://www.infoseek.co.jp.

[43] 猪股健太郎, 高間康史, 石塚満. 概念体系に基づく情報整理支援ツールの日本語化. 情報処理学会第 56 回全国大会, pp. 3–102–103, 1998.

[44] 石川勉, 井澤潤次朗, N. V. Ha, 笠原要. 単語の意味に関する概念ベースの類似性判別能力からの最適構成. 人工知能学会誌, Vol. 13, No. 3, pp. 470–479, 1998.

[45] 石塚満. 知識の表現と高速推論. 丸善, 1996.

[46] 岩爪道昭, 武田英明, 西田豊明. 弱構造化オントロジーを用いたインターネットからの情報獲得. 信学技報 AI95–32, pp. 79–86, 1995.

[47] 岩爪道昭, 白神謙吾, 武田英明, 西田豊明. インターネットからの情報収集・分類・統合化のためのオントロ ジー獲得. 第 10 回人工知能学会全国大会, pp. 553–556, 1996.

[48] T. Joachims, T. Mitchell, D. Freitag, and R. Armstrong. WebWatcher: Machine Learning and Hypertext. In *Fachgruppentreffen Maschinelles Lernen*, Dortmund, Germany, 1995.

[49] 笠原要, 松澤和光, 石川勉. 国語辞書を利用した日常語の類似性判別. 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1283, 1997.

[50] N. Kato and S. Kunifuji. Consensus-making support system for creative problem solving. *Knowledge-Based Systems*, Vol. 10, pp. 59–66, 1997.

[51] 川喜田二郎. 発想法. 中公新書, 1967.

[52] 川喜田二郎. 続・発想法. 中公新書, 1970.

[53] 橘高博行, 佐藤直之, 鈴木英明. 分散化リコメンドシステムの提案. 情報処理学会第57回全国大会, pp. 3–155–156, 1998.

[54] 神田 陽治他. グループ発想支援システム：GrIPS. 人工知能学会誌, Vol. 8, No. 5, pp. 601–610, 1993.

[55] 小池英樹. bit 別冊：ビジュアルインタフェース–ポスト GUI を目指して–, 2.1. ビジュアライゼーション. 共立出版, 1996.

[56] 国立国語研究所編. 分類語彙表, 1964.

[57] 小中裕喜. 分類された文書集合における特徴的なキーワードパターンの抽出. 情報処理学会第57回全国大会, pp. 3–217–218, 1998.

[58] 糀谷和人, 前田晴美, 西田豊明. 弱構造知識メディアを用いた情報ベース構築支援. 信学技報 AI95–30, pp. 63–70, 1995.

[59] 小松原健. 特集：プッシュ技術が実用期へ. 日経インターネットテクノロジー, No. 10, pp. 92–111, 1997.

[60] 熊本睦, 飯田敏幸. 想起型情報検索システムにおける文書のクラスタ化. 情報処理学会第57回全国大会, pp. 3–215–216, 1998.

[61] 國藤進. 発想支援システムの研究開発動向とその課題. 人工知能学会誌, Vol. 8, No. 5, pp. 552–558, 1993.

[62] W. Lam, K. F. Low, and C. Y. Ho. Using a Bayesian Network Induction Approach for Text Categorization. In *Proc. of 15th Int'l Joint Conf. on Artificial Intelligence(IJCAI97)*, pp. 745–750, 1997.

[63] H. Lieberman. Letizia: An Agent That Assists Web Browsing. In *Proc. of 14th Int'l Joint Conf. on Artificial Intelligence(IJCAI95)*, pp. 924–929, 1995.

[64] B. Liu, L. P. Ku, and W. Hsu. Discovering Interesting Holes in Data. In *Proc. of 15th Int'l Joint Conf. on Artificial Intelligence(IJCAI97)*, pp. 930–935, 1997.

[65] Y. S. Maarek et al. WebCutter: A System for Dynamic and Tailorable Site Mapping. In *6th Int'l WWW Conference*, 1997.

[66] 前田晴美, 糀谷和人, 西田豊明. 連想構造を用いた情報収集・整理支援. 第 10 回人工知能学会全国大会, pp. 565–568, 1996.

[67] P. Maes and R. Kozierok. Learning Interface Agents. In *Proc. of AAAI-93*, pp. 459–465, 1993.

[68] P. Maes. Agents that Reduce Work and Information Overload. *Communications of the ACM*, Vol. 37, No. 7, pp. 31–40, 1994.

[69] I. Mani and E. Bloedorn. Multi-document Summarization by Graph Search and Matching. In *Proc. of AAAI-97*, pp. 622–628, 1997.

[70] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村 友明. 日本語形態素解析システム「茶筅」Version 1.5 使用説明書. Technical report, NAIST-IS-TR97007, 1997.

[71] 松尾利行, 武田英明, 西田豊明. 技術情報空間の構築と探訪の知的支援に関する研究. 信学技報 AI95–33, pp. 87–94, 1995.

[72] 宮原隆行, 清水康, 北川高嗣. 意味の数学モデルによる意味的連想検索の高速化アルゴリズムとその実現方式. 情報処理学会論文誌, Vol. 38, No. 7, pp. 1399–1411, 1997.

[73] 水野欽司. 多変量データ解析講義. 朝倉書店, 1996.

[74] K. J. Mock. Hybrid Hill-Climbing and Knowledge-Based Methods for Intelligent News Filtering. In *13th Nat'l Conf. on Artificial Intelligence (AAAI-96)*, Vol. 1, pp. 48–53, 1996.

[75] 諸橋正幸, 那須川哲哉, 長野徹. テキストマイニング：膨大な文書データからの知識獲得 –意図の認識–. 情報処理学会第 57 回全国大会, pp. 3–75–76, 1998.

[76] 宗森純. 発想支援ツール：グループウェアからの接近. 第11回人工知能学会全国大会チュートリアル講演テキスト, 1997.

[77] 村上健一郎. インターネット. 岩波書店, 1994.

[78] 長尾真. 自然言語処理, 第11章. 岩波書店, 1996.

[79] 中原啓一, 三次衛. 情報の検索とデータベース. 電子情報通信学会, 1986.

[80] 那須川哲哉, 諸橋正幸, 長野徹. テキストマイニング：膨大な文書データからの知識獲得 –概要–. 情報処理学会第57回全国大会, pp. 3–77–78, 1998.

[81] N. Negropnte （西和彦監訳）. ビーイング・デジタル：ビットの時代. アスキー出版局, 1995.

[82] T. Nishida. The Knowledgeable Community: Towards Knowledge Level Communication. In *Int'l Forum on Frontier of Telecommunication Tech.*, 1995.

[83] K. Nishimoto, Y. Sumi, and K. Mase. Toward an outsider agent for supporting a brainstorming session — an information retrieval method from a different viewpoint. *Knowledge-Based Systems*, Vol. 9, No. 6, pp. 377–384, 1996.

[84] H. Noguchi. An idea generation support system for industrial designers (idea sketch processor) . *Knowledge-Based Systems*, Vol. 10, No. 1, pp. 37–42, 1997.

[85] 野口裕史, 國藤進. データベースからの知識発見法を用いた発想支援システムの研究. 人知研資 SIG-J-9602, pp. 76–81, 1996.

[86] H. Ohiwa, N. Takeda, K. Kawai, and A. Shiomi. KJ editor: a card-handling tool for creative work support. *Knowledge-Based Systems*, Vol. 10, No. 1, pp. 43–50, 1997.

[87] 大見嘉弘ほか. インターネット上の情報を利用できるカード操作ツール PAN–WWW. 情報処理学会論文誌, Vol. 37, No. 1, pp. 154–162, 1996.

[88] 女部田武史, 國藤進. 複数の KJ 法図解の差異や共通部を可視化する思考支援システムの実現と評価. 人知研資 SIG-J-9602, pp. 28–33, 1996.

[89] 折原良平. 発想支援システムの動向. 情報処理学会誌, Vol. 34, No. 1, pp. 81–87, 1993.

[90] 折原良平. 発散的思考支援ツールの研究開発動向. 人工知能学会誌, Vol. 8, No. 5, pp. 560–567, 1993.

[91] 折原良平. 発想支援システム「知恵の泉」. 人工知能学会誌, Vol. 9, No. 2, pp. 248–257, 1994.

[92] 大澤幸生, N. E. Benson, 谷内田正彦. 共起グラフを用いたキーワード抽出. 情処情報基礎論研究会, 第 115-16 巻, 1996.

[93] 大澤幸生, 山田誠二. WWW 上で文献検索プランニングを行うソフトウェアエージェント NaviPlan. 第 12 回人工知能学会全国大会, pp. 376–379, 1998.

[94] 大澤幸生, 須川敦史, 谷内田正彦. ユーザの変化する興味を理解し表現する文献検索支援システム Index Navigator. 人工知能学会誌, Vol. 13, No. 3, pp. 461–469, 1998.

[95] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying interesting web sites. In *13th Nat'l Conf. on Artificial Intelligence (AAAI-96)*, Vol. 1, pp. 54–61, 1996.

[96] M. Pazzani, L. Nguyen, and S. Mantik. Learning from hotlists and coldlists: Towards a WWW information filtering and seeking agent.
In *Proc. of 7th Int'l Conf. on Tools with Artificial Intelligence*, 1995.

[97] J. Pearl. Bayesian Decision Methods, Uncertain Reasoning.
Morgan-Kaufmann, pp. 345–352, 1990.

[98] M. F. Porter. An algorithm for suffix stripping. *Program*, Vol. 14, No. 3, pp. 130–137, 1980.

[99] J. R. Quinlan（古川訳）. AIによるデータ解析. （株）トッパン, 1995.

[100] J. Rekimoto and M. Green. The Information Cube: Using Transparency in 3D Information Visualization. In *Proc. 3rd Annual Workshop on Info. Tech. & Sys. (WITS'93)*, pp. 125–132, 1993.

[101] 斎藤逸郎, 土肥浩, 石塚満. WWW におけるグループ経験の共有を図るメディエータエージェントの構築. 情報処理学会第 53 回全国大会, pp. 4–239–240, 1996.

[102] K. Schmid. Making AI systems more creative: the IPC-model. *Knowledge-Based Systems*, Vol. 9, No. 6, pp. 385–397, 1996.

[103] 篠原靖志. 知識整理支援システム CONSIST-II — CONSIST の評価と改良について—. 人工知能学会誌, Vol. 8, No. 5, pp. 593–600, 1993.

[104] 塩澤, 西山, 松下. 協調検索型ハイパーメディアの WWW による実現. 情処グループウェア研究会報告 95–GW–13, pp. 13–18, 1995.

[105] 塩澤秀和, 相馬隆宏, 野田純也, 松下温. 切り取り操作による柔軟な情報選択ができる WWW 視覚化. 情処ヒューマンインタフェース研究会報告 97-HI-71, 1997.

[106] M. Sugimoto and K. Hori asund S. Ohsuga. Method to assist the building and expression of subjective concepts and its application to design problems. *Knowledge-Based Systems*, Vol. 7, No. 4, pp. 233–238, 1994.

[107] M. Sugimoto, K. Hori, and S. Ohsuga. A system to visualize different viewpoints for supporting researchers' creativity. *Knowledge-Based Systems*, Vol. 9, No. 6, pp. 369–376, 1996.

[108] 杉野陽一, 由井薗隆也, 宗森純, 首藤勝. 発想支援グループウェア郡元を用いて異なる大学間で行った KJ 法の結果の検討. 人知研資 SIG-J-9602, pp. 64–69, 1996.

[109] K. Sugiyama et al. Emergent Media Environment. 人知研資 SIG-J-9602, pp. 19–24, 1996.

[110] K. Sugiyama, K. Misue, I. Watanabe, K. Nitta, and Y. Takada. Emergent media environment for idea creation support. *Knowledge-Based Systems*, Vol. 10, No. 1, pp. 51–58, 1997.

[111] 杉山公造. 収束的思考支援ツールの研究開発動向 – KJ 法を参考とした支援を中心にして –. 人工知能学会誌, Vol. 8, No. 5, pp. 568–574, 1993.

[112] 杉山公造. 発想支援ツール：ツール群の開発と統合化の試み — 創発メディア環境—. 第 11 回人工知能学会全国大会チュートリアル講演テキスト, 1997.

[113] Y. Sumi, K. Nishimoto, and K. Mase. Facilitating Human Communications in Personalized Information Spaces. In *AAAI-96 Work Shop on Internet-Based Information Systems*, pp. 123–129, 1996.

[114] 角康之, 堀浩一, 大須賀節雄. テキストオブジェクトを空間配置することによる思考支援システム. 人工知能学会誌, Vol. 9, No. 1, pp. 139–147, 1994.

[115] 角康之, 堀浩一, 大須賀節雄. ソフトウェアの要求モデル構築における発想支援とモデル生成. 第 9 回人工知能学会全国大会, pp. 439–442, 1995.

[116] 角康之, 西本一志, 間瀬健二. 個人の視点を伝え合うことによる協同発想支援. 人知研資 SIG-J-9602, pp. 70–75, 1996.

[117] 角 康之他. 思考空間の可視化によるコミュニケーション支援手法. 電子情報通信学会論文誌 A, Vol. J79-A, No. 2, pp. 251–260, 1996.

[118] 住田一男, 知野哲朗, 小野顕司, 三池誠司. 文書構造解析に基づく自動抄録生成と検索提示機能としての評価. 電子情報通信学会論文誌 D-II, Vol. J78-D-II, No. 3, pp. 511–519, 1995.

[119] 高間康史, 大澤幸生, 石塚満. 知識の実行時リフォメーションに基づく仮説推論の高速化手法. 人工知能学会誌, Vol. 10, No. 6, pp. 913–920, 1995.

[120] 高間康史, 土肥浩, 石塚満. 擬人化エージェントにおける音声対話を通じての協調的応答戦略の自動学習. 人工知能学会誌, Vol. 12, No. 3, pp. 456–465, 1997.

[121] 高野敦子, 平井誠, 北橋忠宏. ユーザの発想を支援する自然言語 I/F の実現. 人知研資 SIG-J-9602, pp. 52–57, 1996.

[122] 高杉耕一, 國藤進. ばねモデルを用いたアイデア触発システムの構築について. 人知研資 SIG-J-9602, pp. 34–39, 1996.

[123] 竹下敦, 井上孝史, 田中一男. テキストの概要把握支援のための話題構造抽出. 情報処理学会論文誌, Vol. 37, No. 11, pp. 1941–1949, 1996.

[124] 竹下敦, 井上孝史, 田中一男. モノローグに対するブラウジング支援のための話題構造抽出. 情報処理学会論文誌, Vol. 37, No. 11, pp. 1919–1927, 1996.

[125] 舘村純一. DocSpace: 文献空間のインタラクティブ視覚化. 日本ソフトウェア科学会 WISS'96, インタラクティブシステムとソフトウェア IV, pp. pp. 11–20, 1996.

[126] M. C. Torrance. Active Notebook: A Personal and Group Productivity Tool for Managing Information. In *AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, 1995.

[127] 浦本直彦. コーパスに基づくシソーラス—統計情報を用いた既存のシソーラスへの未知語の配置. 情報処理学会論文誌, Vol. 37, No. 12, pp. 2182–2189, 1996.

[128] 渡部勇. Keyword Associator による情報の構造化支援. 人知研資 SIG-J-9602, pp. 13–18, 1996.

[129] WordNet. http://www.cogsci.princeton.edu/~wn/.

[130] M. R. Wulfekuhler and W. F. Punch. Finding Salient Features for Personal Web Page Categories. In *6th Int'l WWW Conference*, http://www6.nttlabs.com/HyperNews/get/PAPER118.html, 1997.

[131] Yahoo! Japan. http://www.yahoo.co.jp.

[132] 姚左軍, 濱田喬. 類似検索における特徴ベクトルのインデックスおよび関連の検索に関する一手法. 情報処理学会論文誌, Vol. 37, No. 11, pp. 2056–2064, 1996.

発表文献

学会誌論文

高間 康史, 大澤 幸生, 石塚 満:
知識の実行時リフォメーションに基づく仮説推論の高速化手法,
人工知能学会誌, Vol. 10, No. 6, pp.913-920, 1995.

高間 康史, 土肥 浩, 石塚満:
擬人化エージェントにおける音声対話を通じての協調的応答戦略の自動学習,
人工知能学会誌, Vol. 12, No. 3, pp. 456–465, 1997.

高間 康史, 石塚 満:
Fish Eye マッチング：概念体系を利用した視点抽出に基づく文書整理支援機能,
人工知能学会誌, Vol. 14, No. 1, pp. 93–101, 1999.

国際会議論文

Yasufumi Takama, Hiroshi Dohi and Mitsuru Ishizuka:
A Visual Anthropomorphic Agent with Learning Capability of Cooperative Answering Strategy through Speech Dialog,
Asia–Pacific Computer and Human Interaction (APCHI'98), pp. 260–265, 1998.

Yasufumi Takama, Mitsuru Ishizuka:
Fisheye Matching: Viewpoint–Sensitive Feature Generation Based on Concept Structure,
Proc. of 16th Int'l Joint Conf. on Artificial Intelligence(IJCAI99), 1999（投稿中）.

研究会論文

高間 康史, 石塚 満:
概念体系を用いた Fish Eye ベクトルの情報整理支援ツールへの応用,
人知研資 SIG-FAI-9702, pp. 97–102, 1997.

高間 康史, 石塚 満:
概念体系を用いた Fish Eye マッチングによる視点を考慮した文書整理支援機能の実現,
信学技報 OFS98-17, AI98-26, pp. 37–44, 1998.

大会論文

高間 康史, 大澤 幸生, 石塚 満:
論理の多段化による知識リフォメーションに基づく仮説推論の高速化手法,
情報処理学会第 48 回全国大会, pp. 3-211–212, 1994.

高間 康史, 大澤 幸生, 石塚 満:
論理の多段化による知識リフォメーションに基づく仮説推論の高速化手法,
第 8 回人工知能学会全国大会, pp. 59–62, 1994.

高間 康史, 土肥 浩, 石塚 満:
擬人化エージェントシステムにおける協調的応答戦略の獲得機構,
情報処理学会第 51 回全国大会, pp. 6-249–250, 1995.

高間 康史, 土肥 浩, 石塚 満:
協調的応答を学習する擬人化エージェントシステムの実現,
情報処理学会第 52 回全国大会, pp. 6-203–204, 1996.

原田 晋, 高間 康史, 土肥 浩, 石塚 満:
擬人化エージェントにおけるマルチモーダルな案内応答の生成,
情報処理学会第 52 回全国大会, pp. 6-205–206, 1996.

高間 康史, 土肥 浩, 石塚 満:
協調的応答戦略の学習機構を備えた擬人化エージェントの開発,
第 10 回人工知能学会全国大会, pp. 481–484, 1996.

久保田 仙, 高間 康史, 土肥 浩, 石塚 満:

擬人化エージェントにおけるWWWブラウザを用いたマルチモーダルインタフェース,
情報処理学会第54回全国大会, pp. 4-11–12, 1997.

高間 康史, 石塚 満:
情報整理・発想支援システムのための概念体系に基づく特徴ベクトルの動的生成・マッチング機構,
第11回人工知能学会全国大会, pp. 372–373, 1997.

高間 康史, 石塚 満:
情報整理支援システムのための概念体系に基づく特徴ベクトルの動的生成,
情報処理学会第55回全国大会, pp. 3-216–217, 1997.

高間 康史, 石塚 満:
概念体系を用いたFish Eyeマッチングによるユーザの視点の抽出,
情報処理学会第56回全国大会, pp. 3-201–202, 1998.

猪股健太郎, 高間 康史, 石塚 満:
概念体系に基づく情報整理支援ツールの日本語化,
情報処理学会第56回全国大会, pp. 3-102–103, 1998.

高間 康史, 石塚 満:
概念体系を用いたFish Eyeマッチングによる視点を考慮した情報の整理,
第12回人工知能学会全国大会, pp. 394–395, 1998.

高間 康史, 石塚 満:
Fish View: 概念体系を用いたFish Eyeマッチングによる視点を考慮した文書整理支援ツール,
情報処理学会第57回全国大会, pp. 3-183–184, 1998.

高間 康史, 石塚 満 :
Fish View: 文書整理支援における視点情報の活用,
情報処理学会第58回全国大会, 1999 (投稿中).

その他

高間 康史 著, 佐倉 統 監修:
図解 人工生命を見る,
同文書院, 1998.

高間 康史 他, 司会 阿部明典:
第 12 回人工知能学会全国大会パネルディスカッション：トイワールドからの脱皮— 新しい時代の新しい実問題のための新しい AI —, 1998.