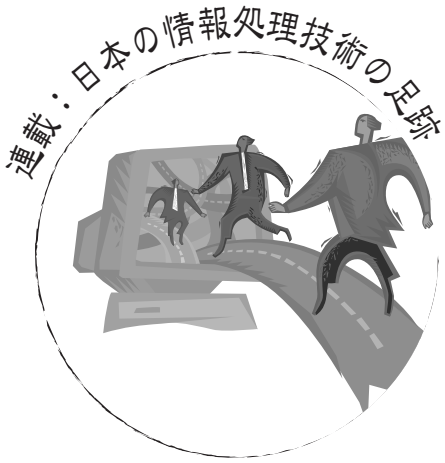

日本語情報処理の諸相：日本語情報検索技術の系譜

藤澤 浩道
日立製作所中央研究所
fujisawa@crl.hitachi.co.jp

絹川 博之
東京電機大学工学部情報メディア学科
kinukawa@im.dendai.ac.jp

1970年代に始まる日立の日本語情報検索技術の系譜について，その当時の問題意識や，研究所における先駆的な取り組みなどを含めて語る．技術的には，情報部門の専門家向けの統制語インデクスを用いたオンライン情報検索サービスの時代，自由な検索語で本文の全文検索が可能となるフルテキストサーチの時代，そして数理的な方法を用いて概念的あるいは連想的な検索が可能となる高度化の時代について，多少の原理を含めて紹介する．

本稿では，日立製作所（以降，日立と表記）における日本語情報検索にかかわる技術開発の系譜を，当時の裏話なども含めながら，述べてみたい．

日立における検索技術の概略史

日本語情報検索の幕開け

国内においてメインフレーム計算機メーカ各社は，1970年代に入ると漢字が扱える情報システムの開発に取り組み始め，コンピュータによる日本語処理が本格化の緒についた．当初はそれぞれ個別のアプリケーションとして開発されたが，1978年には漢字コードがJIS-1978として標準化され，日立は同年12月に日立漢字情報処理システムKEISを製品化した．また，翌年1979年には日本でも「オフィスオートメーション」が着目され始め，日本語情報処理はコンピュータ分野での期待される分野となった．

日立では1970年代に日本特許情報センター（JAPATIC）や日本科学技術情報センター（JICST）などにおける情報検索サービスシステムの開発に参画し，それが日立の日本語情報検索技術の流れの源となった．当時の同社情報システム研究所（現在のシステム開発研究所，以降ではシ研と表記）で絹川らが研究所から参加して取り組みを始めた．後に述べる日本語解析による自動インデクシングの開発に当たることとなった[1]．その後，同研究所では，日本語処理あるいは自然言語処理として研究対象を広げ，機械翻訳や自然語インタフェースの研究も始まった．

この時期の情報検索システムは，主に情報部門の専門家にオンライン情報検索サービスを提供するためのものであった．当時のオンラインサービスは公衆電話回線を用い，通常300bpsから1200bpsの通信速度で，特許や文献のタイトルや発明者・著者などの書誌情報を検索するものであった．たとえば，上記JICSTは1976年にJOIS-Iのサービスを日本語文献についてはカタカナ情報で開始したが，1979年に漢字仮名混じり文で抄録を扱うJOIS-Kを試験運用し，1981年にJOIS-IIとして本稼働させた．当時，JOIS-IIでは，欧米のデータベースを含む約1,200万件の学術文献を蓄積しており，日立漢字プリンタターミナルHT-5217を用いて対話的な検索サービスを提供した．ただし，このときの日本語処理は，検索する際はカタカナで日本語検索語を入力し，結果は漢字仮名混じり文で出力するレベルであり，完全に漢字で検索ができるようになるのはJOIS-III（1990年）からであった．また，この時代の検索対象は書誌的事項に限られ，本文全文が扱えるようになるのは1990年代のフルテキストサーチ（全文検索）を待つ必要があった．

当時の情報検索方式としては，いわゆる統制語を介した検索を基本としていた．すなわち，インデクサと呼ばれる専門家が登録文献を読み，それらのキーとなる概念を表す単語を，統制語リストを参照しながら抽出して，データベースに登録した．この統制語リストはこうした単語を事前に集めたいわゆる辞書であり，シソーラスともいわれる．したがって，検索する場合も，その統制語リストから検索語を選ぶ必要があり不便であった．また，次々と現れる新しい概念を取り込むためのシソーラス保守の手間やその遅れは利便性を損なう問題となっていた[2]．

絹川らが研究を始めた自動インデクシングは，タイトルや抄録の日本語解析を行って，単語を抽出してシソーラスに存在する単語は統制語キーワードとして，存在しない単語は自由語キーワードとしてデータベースに登録

する技術である[3].

新しい流れの始まり

1980年代に入るとワードプロセッサが少しずつ机上に乗り始めた．すべての文書は電子化されるのか，と一部の人々は思い始めた．また，1983年頃には，データを記憶することができる光ディスクが出現して，その応用として文書を画像として蓄積する光ディスクファイリング装置なる製品が各電機メーカから続々と商品化された．これらの変化は，コンピュータの専門家でない一般の人々が文書を自ら作り出したり，それらをコンピュータのファイルから検索したりする時代がすぐにくることを予感させた．

同社中央研究所（以降，中研と表記）では，この流れをとらえて1984年に藤澤らが文書ファイリングの研究を始めた．彼らは文字認識や文書画像処理を専門としており情報検索の研究者ではなかったが，素人が検索ユーザになることの問題点に着目した．当時注目されていた知識工学を応用した概念ネットワークによる知的ファイリング[4]や，文書理解技術を用いたイメージ文書の自動ファイリングなどの研究を始めた．

また，同時に，当時始まりつつあった特許庁ペーパレス計画の存在を知り，審査官による高度な検索が課題になるであろうことを予見した．特に，自由な検索語で検索できること，さらに，同義語や表記のゆれなどをまったく気にせずに検索できることが重要であると考えた．そこで，彼らが日立光ディスクディスクファイリング装置HITFILEのために開発中であったあいまい検索技術を拡張して，大規模文書データベースのフルテキストサーチを実現することを考え始めた．

フルテキストサーチは当時まだ必ずしも常識的なものではなく，米国で一部先進的な取り組みがなされている状況であった．たとえば，Dow Jones社はThinking Machine社のコネクションマシンを2台購入してWall Street Journalのフルテキストサーチを試験的にサービスし始めた，という記事が流れていた．一方，技術論文では，フルテキストサーチは本当に有効なのか，というような議論もまだなされていた．しかし，これは研究としては始めるに大変良いタイミングであった．さらにいえば，単語の切り出しが容易である英語文献検索のアプローチを日本語に直接適用できないことは明らかであり，研究テーマとしてはチャレンジングでもあり魅力的であった．

1986年春に米国特許庁を訪問する機会を得た藤澤はそこで試行実験中のフルテキストサーチを目の当たりにし，その研究意義を大いに確認することができた．同年夏に，大規模特許情報検索をビークルとしたテキストサーチマシンTSM-Iの開発を中研で開始した．しかし一方では，研究所が考えていることと事業部門が特許情報検索システムとして計画していることには当然かなりの距離があり，「あまりフルテキストサーチと繰り返して言わないで欲しい」と現場SEから言われることもあった．次世代の新技術を開発することが使命である研究所と，確実に進歩させていくことが求められる事業部門の役割をバランスよく両方進めることが重要であった．

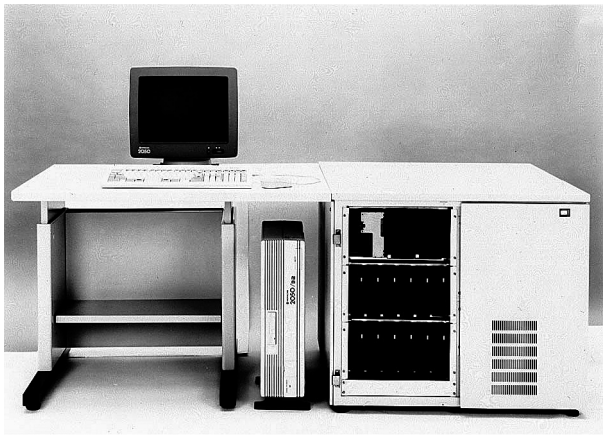

図-1　テキストサーチマシン試作機TSM-I概観

このテキストサーチマシンの設計と試作は加藤が約15名の技術者を率いて研究所内で徹夜をしながら成し遂げたものである[5]．20年間分の特許明細書1,000万件の全文検索をサーチマシン100台で実現しよう，という構想のもとで，その1台の設計と試作に入った．

試作機の原理は，並列に接続した12台の5インチ小型磁気ディスクから毎秒20MBの速度でテキストを読み出して文字列照合を行う．検索語は1,024語まで一括照合できるアルゴリズムを採用して，同義語や異表記の問題を解決する．また，2段階のサロゲーションを用いて，すべてのテキストファイルを読まずに済ませる，というものであった．1988年の末には試作機TSM-I（**図-1**）が完成し，システム的には100MB/秒のテキスト照合が実現し，1件当たり平均1万文字の日本語文書2万5千件を5秒で検索することが可能となった．

実用期に入った全文検索とそれに続く新しい技術

翌年から顧客へのアピールを中研で始めたが，その試作機のままでもよいから購入したいという見学者も現れ，製品化方針が決定された．当初，試作機と同様，ハードウェア製品とすることを検討したが，徐々に高性能化していくワークステーションを見ながら，ソフトウェア製品とする方針変更を行った．第1号の製品は限定版製品であったが，1991年に神奈川県議事録検索システムとし

て納入した．翌年，正式版がソフトウェア製品「高速全文検索システムBibliotheca/TS」（ビブリオテカ）として世に出た．命名は森秀司（当時コンピュータ事業本部副技師長）による．

しかし，特許庁で全文検索システムとして正式稼働したのは1999年である．それは研究を開始してから実に13年後で研究者にとってはきわめて長い道のりであったが，一方では加藤らによる継続した技術開発の積み上げの成果であったともいえる．

1990年代に入ると，プロセッサの能力が格段に向上して，かなり高度なことまでがソフトウェアで実現されるようになり，Bibliothecaも新しい概念検索などの新機能を実現させた．また，1985年に新設された日立の基礎研究所（以降，基礎研と表記）から新しい連想検索などの技術の芽も生まれた．

これは双対型連想検索システムDualNAVIとしてかたちを表すことになるが，自然言語処理で語彙空間の可視化を研究していた丹羽，情報のクラスタリングを研究していた岩山，そして実装の西岡を束ねて，リーダであった高野が具体的な方向性を示して実現したものである．その研究は1995年頃始まり，翌年にはその概念の具体化と並行して，Science誌やStanford大学との共同研究が浅井基礎研所長のリーダシップのもとで始まり，翌々年にはプロトタイプが完成した．DualNAVIは，1999年に日立デジタル平凡社「ネットで百科・デュアル連想検索」に採用され，初めて世に出た．その後は東京大学医科学研究所ヒトゲノム解析センター「分子生物学関連データベースBACE」，日立PatentRetrieverなどで実用に供されている．

また，このシステムの基本となる連想演算を高速に行う汎用連想計算エンジンは1999年の情報処理振興事業協会の独創的情報技術育成事業の1つとして開発され，現在，GETA（Generic Engine for Transposable Association）として無償で公開されている[6]．このGETAエンジンは，すでに国立情報学研究所の図書ナビゲータWebcat Plus等で利用されている．

以下の章では，いくつかのトピックスについて少し詳しく紹介する．

先駆的な取り組み

本章では，必ずしも製品適用には至らなかったものも含めて，技術の流れに影響を及ぼしたと考えるいくつかの先駆的な取り組みについて紹介する．

表層格構造に基づくロール付きキーワード抽出

国内でコンピュータを用いた新聞制作は1975年に朝日新聞社で最初に始まったが，この流れをとらえたシ研の絹川・木村らの自動インデクシングの研究は自然語解析を基本とする先駆的なものであった．同年に始めたこの研究は国際政治関連の新聞記事を対象に「ロール付きキーワード」を自動的に抽出することを狙ったもので，後の情報抽出技術の1つともいえる[3]．

一般に文を構成する要素である単語は，文の意味を構成する枠組みでそれぞれ役割を担っているが，その役割をロールと呼ぶ．たとえば，「19XX年A国が世界中の石油を支配した」という文の場合，「A国」は「主体」というロールを担っており，「19XX年」は「時」というロールを担っている．この研究では，主体，客体，時，場所，活動，その他の主題の6種類のロールを定義した．

その解析の原理は，**図-2**に示すように，形態素解析を行って自然文の構成語を同定し，表層格構造認定により得た文型に照らし合わせながら，構成語にロールを付与する方法である．形態素解析では動詞約5,600語を含む自立語13万語，付属語700語の単語辞書を用いている．

表層格構造認定では，まず用言の連用中止形，接続助詞の付接等を基に複文を単文に分解する．その後，係助詞，使役文や受身文中の格助詞，および連体形単文の「ノ」をそれぞれ正規化するとともに省略格助詞を同定補填する．次に各用言に対する文型を参照し，体言と用言の係り受け関係を認定する．連体形単文では，文型から係り得る格の1つが用言の修飾先であるとして係り受け関係を認定するとともに，体言同士の修飾関係も認定する．これらの認定結果を上記文型に照らし合わせながら，各単文の構成語にロールを付与するものである．

ここで文型とは，表層格フレームにおける格パタンのことで，約1,700文型からなる．動詞については格パタンの違いで約600に分類する一方，名詞については，組織体，人名，資料名，地名，動作，抽象概念，時の7種類に分類し，格パタンとの照合の精度を高めた．

この研究のロール付与のための表層格構造認定は，日本語文解析の多くの基本的課題を扱っているとともに，解析用の言語知識も1970年代後半としては他に類を見ない大規模なものとなっている．

評価実験は新聞記事281件（1,225文）を対象に行い，その結果，単語辞書の語彙カバー率と表層格フレームの格パタンカバー率が各々90%であれば，ロール付きキーワードをほぼ7割正しく抽出できる見通しを得た．ここで得られた基本的な技術は日立漢字情報処理システムKEISの日本語情報検索システムORIONに適用された．

概念ネットワークによる知的ファイリング

文書ファイリングの研究で藤澤らは，ファイリングに

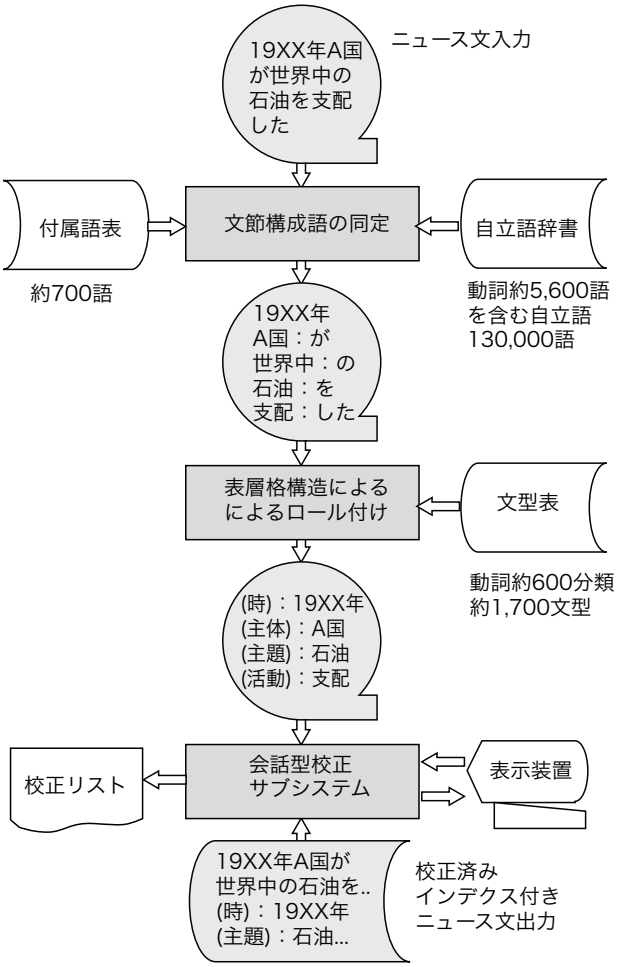

図-2　ロール付きキーワード自動抽出方式

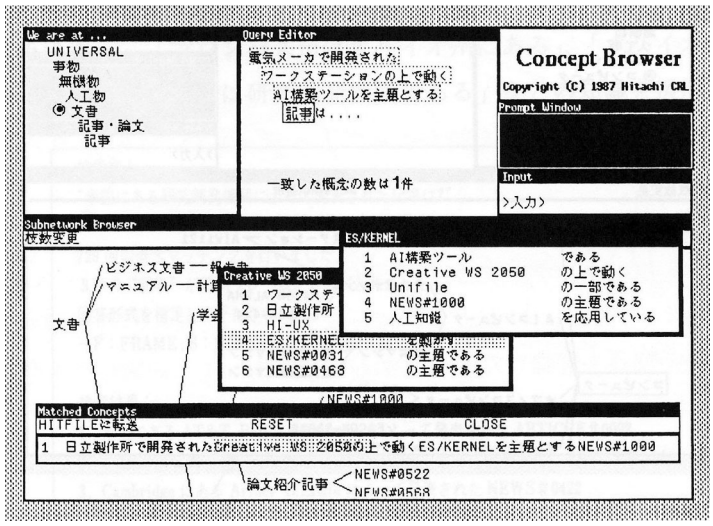

図-3　概念ネットワークを用いた知的検索ConceptBrowser

おける検索は文献検索と異なり，自分で一度見たことのある「あの文書」の検索が重要であり，認知心理学的な考察が重要であることを意識した．たとえば，知っていることを明示的に述べることは難しいが，具体的に示せばそれを知っているか否かは容易に言える．あるいは，具体的な名称は忘れやすいがその上位概念は覚えていることが多い，といった人間の記憶の特性が，情報検索でも重要であることを考えた．これに基づき，概念のネットワークを図式的にコンピュータで示して，その空間をブラウズしながら検索条件を対話的に編集したり，上位概念から意味的な検索を行ったりすることができる対話型推論検索システムConceptBrowserを1987年にXeroxのリスプマシン上に試作した（**図-3**）．

この研究では，約450件の新聞記事とそれに出現する総計3,500の概念と7,000の関係を登録した[4]．アプローチの面白さ，有効さについては認めてもらえたものの，実際に使う場面においての概念の登録と整理が現実論としては困難であると評価された．

そこで，概念ネットワークの基本部分だけでも自動構築して，それを検索時の検索条件作成の支援に用いるというアプローチの研究を1994年から中研の梶・森本が開始した．単語の共起性を用いてテキストコーパスから関連シソーラスを自動構築する技術と，出現頻度の高い単語のクラスタリングを行ってトピックスを抽出する技術を研究した[7)☆1]．その成果はソフトウェア製品「探索支援ライブラリ」として実用化された．このように研究アプローチは日立においても，1980年代のAI的アプローチから1990年代にかけての大規模コーパスを前提とした工学的アプローチへと大きく変わった．

文書自動分類

日本特許庁のペーパレスシステムは1990年2月に稼働し，電子出願の受付が始まった．毎年40万件を超える特許が出願されているが，すべてがフルテキスト情報としてデータベース化されるようになった．シ研の間瀬・絹川らは1990年から8年にわたり特許庁での審査業務の短縮化・効率化のための出願特許の自動分類の研究を行った[8)]．

その目標は，出願された特許明細書を38の上位分類カテゴリのいずれかに，また，その下位の2,815の詳細カテゴリのいずれかに自動分類するというものである．その方式は以下のようである（**図-4**）．まず，大量の分類済み特許明細書サンプルと，各カテゴリの適用範囲を規定した分類マニュアルを教師データとして用いて「分類知識」を自動生成する．そして，その分類知識を用いて，新規明細書を上位・詳細の各カテゴリに分類する．知識生成は，まず学習サンプルとしての各明細書の「発明の名称」「請求項」などを解析して，特定の表現に含まれるキーワードを抽出する．次に，キーワードの位置

☆1　この研究は情報処理振興事業協会と日本情報処理開発協会の「次世代電子図書館システム研究開発事業」の支援を受けた．

とそれが出現した項目の数によって重み付けし，分類カテゴリ名と組を作る．また，分類マニュアルから抽出したキーワードはその出限頻度で重み付けし，当該カテゴリ名とで組を作る．これらの重み付きキーワードとカテゴリ名との組を統合して分類知識とした．分類したい新規明細書は分類知識と照合して類似度の高いカテゴリを分類先とした．

実験では，専門家が分類した1993年から4年間の特許公報の明細書サンプル32万件を用いた．そのうち31万件を教師データに，1万件を分類実験に用いた．分類では，上位3つのカテゴリを答えとして付与してよいこととした．その結果，上位分類では96%，詳細分類では83%に正解が含まれることが分かった．また，分類知識生成には，約1,000件／カテゴリが必要であることが分かった[8]．

1997年に行ったこの実験規模は，近年のTREC（Text Retrieval Conference）やNTCIR（NII Test Collection for IR）等に劣らないが，実用化に際しては，さらなる精度の向上が望まれる．

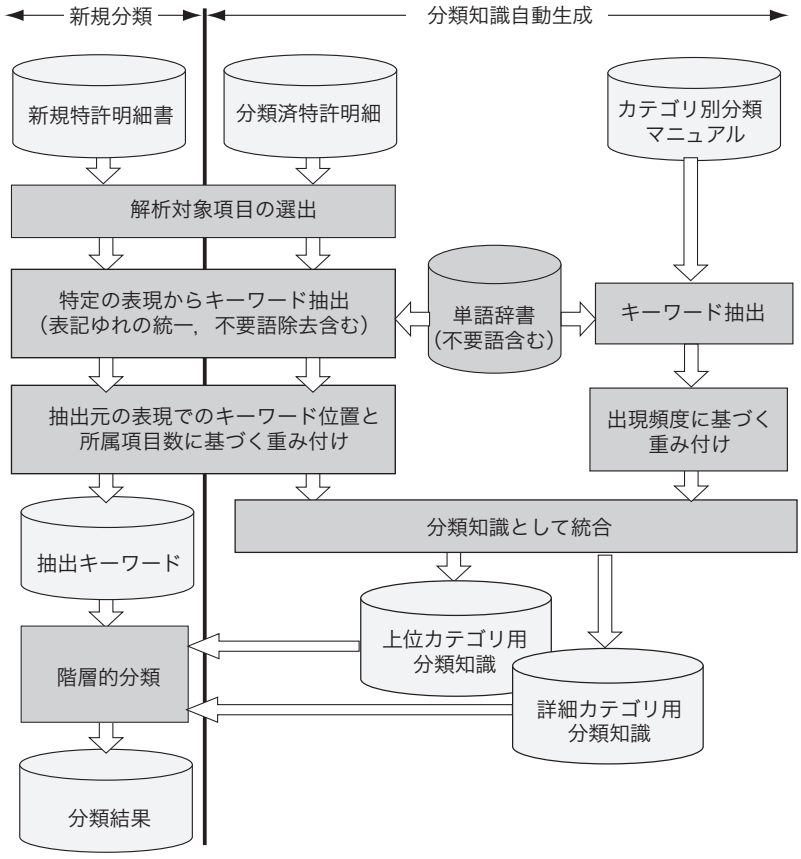

図-4　特許自動分類方式

フルテキストサーチ

あいまい検索とテキストサーチマシン

情報処理の素人が文書の登録や検索をする光ディスクファイリングでは，それまでの検索手法が採用できないことは明らかであった．藤澤・川口・畠山は，特に，新旧漢字や外来語など表記のゆれや同義語・類義語による検索漏れを問題視し，「あいまい検索」の開発を行い，1988年頃ファイリング装置HITFILEに搭載した．

あいまい検索は，検索語の文字列，たとえば「バイオリン」を解析して，事前に準備した異表記生成ルールを当てはめて，内部的に検索語「ヴァイオリン」を自動生成し，これら両方を検索する．これによって検索漏れをなくす．このための異表記生成ルールとしては，たとえば「バ」と「ヴァ」，「ホ」と「フォ」は等価である，というルールを約1,300項目，国語審議会報告の外来語表記勧告を参照して作成した．このような機能は今や主なワープロソフトで「あいまい検索」として常識的である．

その実現には，複数文字列を一括して照合するAho-Corasick法を改良した「先行フェール型オートマトン」という文字列照合アルゴリズムを考案して，専用ハード「サーチエンジン」を開発した．先に述べたように，1986年にテキストサーチマシンに取り組むことになったが，それにはニーズの先取りということもあったが，実はこれらあいまい検索とそのための文字列照合の技術を開発中であったことがそのきっかけとして大きい．

テキストサーチマシンの技術課題は明らかに大規模化とそれに伴う高速化であった．大規模化については，当時は分散システム化の流れの中で，サーチマシンを多数台併置して並列検索することを考えたが，一方の高速化が大きな課題であった．最初は，すべてのファイルを読み出さないで済ませるために，書誌事項検索で絞り込むことを考えたが，それではフルテキストサーチのメリットを殺す．そこで，テキスト情報を凝縮して代表ファイル（サロゲーション）を作るサロゲート法を考えた．

高速化のためのアイディアは以下の3つである．

- 並列小型ハードディスク：12台で読み出し速度20MB/s
- 専用サーチエンジン：文字列走査速度20MB/s
- サロゲート法：2段階のシステム的な高速化

その結果，オーバヘッドを含めても全体として平均100MB/sの等価走査速度を実現した．

サロゲーションには「文字成分表」と「凝縮本文」を階層的に用いた．文字成分表は，1文書2048ビットのビットベクトルを文書の数だけ並べたビットマップである．文書に現れるすべての文字コードそれぞれに特殊なハッシュ関数を適用して0から2047までの値を得て，ビッ

トベクトルのその位置にフラグ1を立てる．また，凝縮本文は，接続詞や助詞などの付属語と重複して現れる文字列とを本文から除外したテキストファイルである．これらを最初に照合することにより，最終的に読み出すべき本文テキストの量を平均2%に抑えることができた．

高速化でこだわったことは，複合条件も文字列走査時に処理することであった．複合条件とは，複数の検索語の間に論理条件（AND，OR），近傍条件（単語間の距離），文脈条件（たとえば，同一文中で共起するという条件）などの演算子を用いる検索条件のことである．これらの複合条件をオンザフライで判定するハードウェアを実現し，たとえば，検索条件“計算機〔S〕検索”を投入すると，“計算機”と“検索”が同一の文に現れる文書を検索することができるようになった[5]（**図-5**）．

完成したテキストサーチマシンTSM-Iは何度か展示会に出展した．LSI化をしていないハードウェアは大きかったが，フルテキストサーチのデモ効果は絶大であった．いよいよ製品化の方針が出て，LSIの設計を始めたが，ハードウェア規模は思いの外大きく，機能の整理を進めたが，その間かなり時間がかかった．至急に欲しいという顧客も現れ，ソフトウェア型の限定版製品を出荷することとなった．

超高速全文検索システムBibliotheca

一度限定版が製品として出ると，当初は文書データ規模もさほど大きくなく，ソフトウェア製品でもよいのではないかという気運が高まり，方針を変更した．そして1992年にUNIXワークステーション3050で走るソフトウェア製品「高速全文検索システムBibliotheca/TS」が誕生した．

その後，1990年代半ばを過ぎると，フルテキストを収めた特許情報CD-ROMも公開され，膨大な文書件数を対象とするニーズも増え，更なる高速化が求められた．これに応えるため高性能化の研究を加藤らが行い，1994年には特許100万件のデータベースを1秒で検索する超高速全文検索システムBibliotheca v2を出荷した．

ここで開発した技術は「ハッシュレス文字成分表方式」と呼ぶもので，文字成分表を見るだけで検索ができる方式である．連接する2文字の組合せは，製品が対象とする約12,000字種の漢字に対して約1億4千万に上るが，実際に調べてみるとそのうちのわずか5%にも満たないことが分かった．そこで，ハッシングせずに文字成分表を効率良く記憶するデータ構造を工夫し，本文サーチを不要として，上記の検索速度を達成した．

その後，SGMLやHTMLといった構造化文書（現在はXML文書）の普及に伴い，論理構造を利用したきめ細かな検索も求められるようになった．文書1件当たりの容量の拡大や文書数の爆発的な増大もあり，多田らが新たな技術の開発に当たり，1997年に「インクリメンタルn-gramインデクス方式」と呼ぶ新検索方式を採用したBibliotheca2 TextSearchを製品化した．

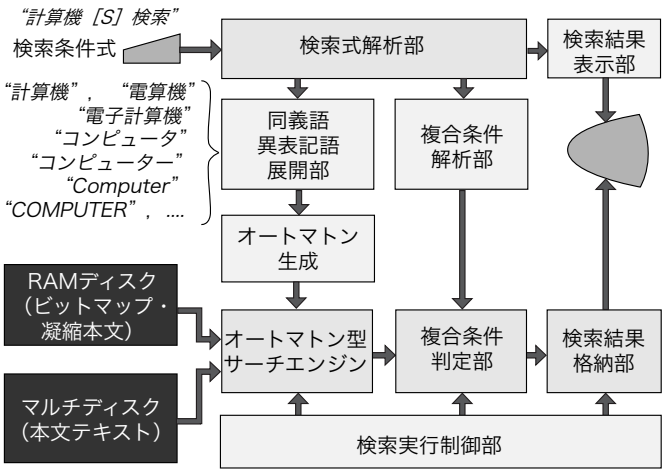

図-5　テキストサーチマシンの構成

一般にn-gramインデクス方式は，連接するn文字（n-gram）に対して，各々がどの文書のどの位置に出現するかをインデクスとして記憶する．たとえば，本文に「半導体」があった場合，その2文字連接（bigram）「半導」と「導体」をbigramインデクスに登録する．検索時には同様にこれらの連接文字列を検索語から抽出して，同インデクスに登録されているか否かを調べ，次にそれらが単語「半導体」を構成する位置関係にあるか否かを調べて同単語の存在を検出する．この方式は検索性能に優れているが，高速化のためにはn-gramのnを大きくする必要があり，インデクス容量の増加と登録性能の劣化が問題であった．

そこでインクリメンタルn-gramインデクス方式では，文書の登録時に各nに対するn-gramインデクスの容量を参照し，検索性能劣化の要因となる容量の大きいn-gramに対してだけ，自動的に文字列長を延長したインデクスを作成する（**図-6**）．たとえば，「東京」から作られた「東」「京」のunigramインデクスの一部を除去して，新しく「東京」というbigramインデクスを作る．これにより，インデクス容量ならびに登録性能を著しく劣化させることなく，高速なフルテキストサーチを可能とした[9]．

概念検索と連想検索

N-gramに基づく概念検索

フルテキストサーチは指定された検索語そのものを含む文書を検索するには適しているが，検索要求があいま

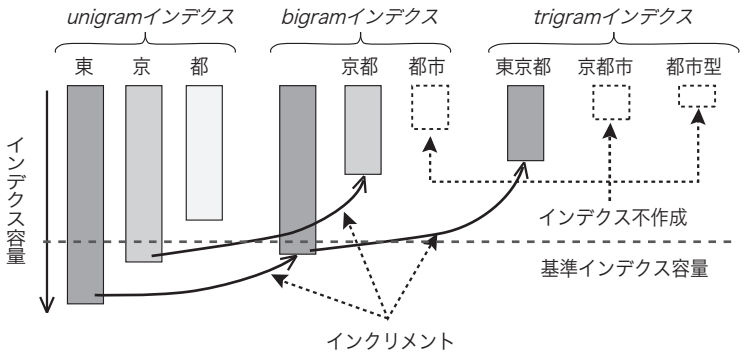

図-6　インクリメンタル*n*-gramインデクス方式

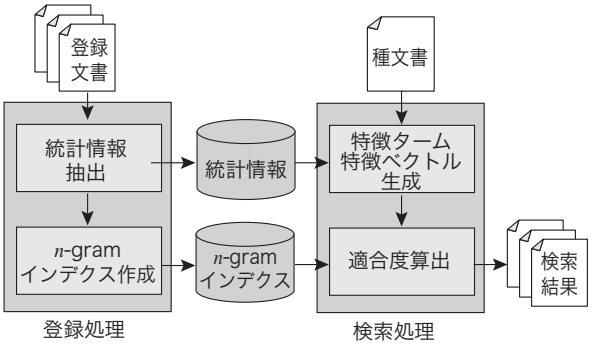

図-7　*n*-gramインデクスによる概念検索

いな場合には，どんな検索語が良いかの判断が難しい．特に，絞り込み検索で，どんな検索語を除去してどんなものを追加すべきかが難しい．このような問題を解決するため，探したい文書を説明した自然文や類似した文書のテキストを条件として検索できる「概念検索」をシステム開発本部の多田・松林らが開発した．

概念検索は通常，文書内容を表す特徴ベクトルを日本語解析で抽出した単語をベースに作成して，検索条件を表す同様の特徴ベクトルとの類似度の評価で検索を実現する．彼らは日本語解析を行って単語を切り出す概念検索システムは未知語の問題があると考え，*n*-gramインデクスを基本とする方式を開発した[10]（**図-7**）．

しかし，*n*-gramを用いるアプローチは未知語の問題はないが，ノイズとなる連接文字列が問題となる．たとえば，「水不足」という複合語で「水不」という連接文字列が特徴タームとなることを避けねばならない．そのために，単独文字や連接文字列の出現確率を事前に統計情報として抽出して，この例の場合，「水」と「不足」という特徴タームのみを得る．

この技術による概念検索は1999年に日立スケーラブルデータベースサーバHiRDBの文書検索用プラグインとして製品となった．これにより，フルテキストサーチで得られた候補文書の自然文を概念検索の条件にして再検索したり，その逆に，概念検索の結果をフルテキストサーチで絞り込んだりするなど，柔軟な検索が可能となった．特に，検索結果に対して，検索意図に「合致している」「合致していない」という評価を入力して再検索する適合性フィードバックが可能となった．

双対型連想検索システムDualNAVI

DualNAVIは1996年頃に基礎研で研究着手し翌年原型が完成した検索システムで，文書の空間と単語（索引語）の空間の双対性を前面に出したシステムである．

双対という言葉はもともと数学用語であるが，ここでは半ば比喩的に「双方向的な2面性」の意味で用いている．文書に語を対応させるのが索引付けであり，語に文書を対応させるのが検索であるので，どちらも新しくはないが，この2方向の関係を双対としてとらえるということはこれまで意識されずにきた．「検索における双対性」は抽象レベルの概念であるが，以下に説明するような複雑な対話機能をシンプルに実現するキーであり，今後，重要性が認識されるようになると思われる．

その特長である双対ビューに立脚した多面的なフィードバック機能について説明する（**図-8**）．

画面は2面からなり，左に検索結果である文書の世界を，そして右にその概要としての語の世界を表示する．これらは相互に密接に関連づけた形で同時に見ることができるので，1つ1つの文書を読まなくても全体像が把握できる．たとえば，気に入った文書をいくつでもチェックして選択すると，それらの文書を特徴づける単語が単語空間の中でハイライトされる．逆に，単語を選択するとそれらの単語に対応する文書にチェックマークが自動的に入る．また，このような相互参照機能に加えて，文書や単語を選んだ上で再検索（適合性フィードバック）することが可能である．

このような検索インタフェースを実現する基本演算は大規模な行列を対象とする連想計算である．この計算を処理パワーの面で支え，かつ双対性を美しく実現したのが汎用連想計算エンジンGETAである．連想計算では，入力として行の重み付きリストを与えると，出力としてそれに対応した列の連想リストを返すが，行と列を入れ替えても同じことができる．行を文書，列を単語とすれば，行から列への連想が特徴語抽出（索引付け）となり，列から行への連想が検索ということになる．

両者を合成すれば文書で文書を検索する関連文書検索が実現できる．DualNAVIで文書を選んでフィードバックをかけた場合には，このような2段階の連想計算が行われて関連文書が検索される．また，連想計算によって検

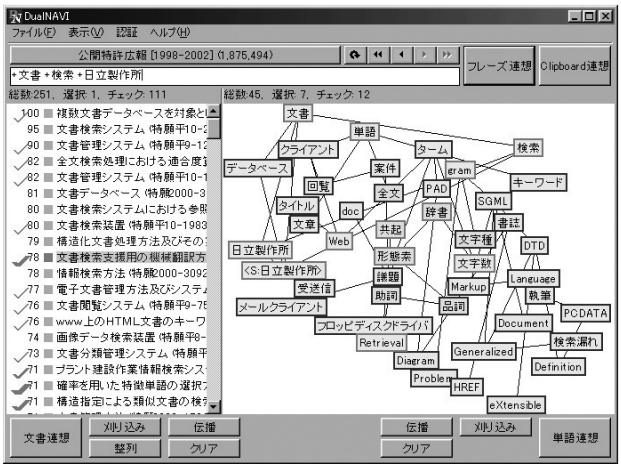

図-8　DualNAVIの双対ビューインタフェース

索結果を得たあと，それに対応する右側の単語のネットワークを得るには，検索結果の文書群からもう一段階単語側へ連想計算を行えばよい．このように，多様な機能がすべて１種類の連想計算とその転置の組合せとして実現できていることが特長である．

そして将来へ

1970 年代に始まった日本語情報検索技術への日立の取り組みについて紹介した．自然言語処理，統計的情報処理，ソフトウェア実装などの専門性が重要であるが，一方で専門家でない研究者の発想も重要である．過去の流れにとらわれずに，素朴に問題を直視することが時として有効である．

この分野はまだまだ解決されていないことが多い．個人の情報環境をみても，まだ検索の問題は解決したとは言えないであろう．やることはいろいろありそうである．たとえば，セマンティックWebなる取り組みもあるが，上位概念からの意味的な検索はまだ実現できていない．大いに今後の技術的な進歩を期待したい分野である．

謝辞　本稿を執筆するに当たり，日立製作所の旭寛治，鳥居哲郎，今城哲二，多田勝己，丹羽芳樹，森本康嗣，間瀬久雄，永野勝也，細矢良智，日立プリンティングソリューションズの小池建夫の各氏にご協力をいただいた．ここに感謝の意を表したい．

国内	日立社内
1970：特許庁特許情報検索システム開発開始 1972：日本特許情報センター（JAPATIC）：特許情報検索システムPATOLISサービス開始 1976：日本科学技術情報センター（JICST）オンライン文献検索JOIS-I稼働（カタカナ情報） 1978：JAPATIC PATOLIS日本語検索サービス開始 1979：日本科学技術情報センター（JICST）：漢字が使えるオンライン文献検索JOIS-K実験サービス開始	1970：漢字ドキュメント編集システムHDESの開発 1971-74：速記を漢字仮名混じり日本文に変換するシステムの開発 1975：自動インデクシングの研究開始 1978：日立漢字情報処理システムKEIS製品化
1981：JICSTオンライン文献検索サービスJOIS-II稼働 1983：東京大学文献情報センター設立 1984：特許庁特許検索システムFターム検索採用 1986：学術情報センター設立	1983：光ディスクファイリング装置HITFILE製品化 1984：知的ファイリングの研究開始 1986：テキストサーチマシンTSM-I試作開始 1988：HITFILE650 あいまい検索の搭載テキストサーチマシン試作機TSM-I完成
1990：特許庁ペーパレスシステム稼働・電子出願開始 JICSTオンライン文献検索サービスJOIS-III稼働 1991：神奈川県議事録検索システム稼働 1999：特許庁：全文検索システム稼働	1992：高速全文検索システムBibliotheca製品化 1994：超高速全文検索システムBibliotheca v.2 製品化 1997：大規模高速全文検索システムBibliotheca2 1999：文書管理システムDocumentBroker製品化 HiRDB概念検索プラグイン製品化 日立デジタル平凡社「ネットで百科」 1999-2001：汎用連想計算エンジンGETAの開発（情報処理振興事業協会（IPA）「独創的情報技術育成事業」）

表-1　年表

参考文献

1）絹川博之：情報検索のための日本語解析，情報処理，Vol.20, No.10, pp.907-910（Oct. 1979）.
2）藤澤浩道，絹川博之：情報検索における自然言語処理，情報処理，Vol.32, No.10, pp.1259-1265（Oct. 1993）.
3）絹川博之，木村睦子：日本語文構造解析による自動インデクシング方式，情報処理学会論文誌，Vol.21, No.3, pp.200-207（May 1980）.
4）藤澤浩道：概念ブラウザと個人情報ベース－概念知識の体系化のためのメディア・スペース，コンピュータ科学，Vol.2, No.1, pp.39-45（1992）.
5）加藤寛次，藤澤浩道，川口久光，大山光男他：大規模文書情報システム用テキストサーチマシンの研究，情報学基礎 14-6，pp.1-8（July 1989）.
6）高野明彦，西岡真吾，丹羽芳樹他：汎用連想計算エンジンの開発と大規模文書分析への応用，IPA 2001 年度成果報告集．
7）梶　博行，森本康嗣，相薗俊子：テキストコーパスのトピック階層の抽出，情報処理学会論文誌，Vol.44, No.2, pp.405-420（Feb. 2003）.
8）間瀬久雄，辻　洋，絹川博之，石原正博：特許テーマ分類方式の提案とその評価実験，情報処理学会論文誌，Vol.39, No.7, pp.2207-2216（July 1998）.
9）川下靖司，岡本卓哉，多田勝己他：構造化文書対応全文検索システムBibliotheca2/TextSearchの開発（1）～（4），情報処理学会第 55 回全国大会，4N-3 ～ 6，pp.3-107 ～ 3-114（Sep. 1997）.
10）松林忠孝，多田勝己，菅谷奈津子他：知識指向文書管理基盤の開発（5）n-gramに基づく概念検索，情報処理学会第 59 回全国大会，5P-11，pp. 3-145 ～ 3-14（Sep. 1999）.

（平成 15 年 10 月 6 日受付）