自己組織型情報ベースシステム

佐藤 誠

SATO Makoto

Self-organizing information base system

5

The latest information related technology

Series

[著者抄録] 科学技術の急速な発達により、科学技術情報の質的・量的な拡大は日々に増大している。これら情報を適切に処理し自己秩序を形成して成長していくという、自己組織化の考え方は、これからの情報分野で重要なコンセプトになると思われる。この自己組織型情報ベースシステムについて、データベースの自動構築、情報の自己組織化機能、自己組織型情報ベースの知的インタフェースの観点より、研究を進めている。

[著者付与キーワード] 自己組織型情報ベースシステム, 知識ベース, マルチメディアデータベースシステム, 人工知能, 可視化, ハイパーメディア, 自己組織型情報検索システム

[Author Abstract] The current rapid pace of development in science and technology has brought a continuous expansion of quality and quantity of scientific and technical information. In order to process such information properly and give it a self-ordering system, and moreover to make the system grow, the approach of self-organization will become a very important concept in the field of information processing. We are carrying out the study of this self-organizing information base system from the viewpoints of an automatic database compilation system, self-organizing function and intelligent interface.

[Keywords by Author] self-organizing information base system, knowledge base, multimedia database system, artificial intelligence visualization, hypermedia, self-organized information retrieval system

* 日本科学技術情報センター (〒100 千代田区永田町2-5-2) Tel. 03(3581)6411
JICST (5-2, Nagata-cho 2-chome, Chiyoda-ku, 100)

1. はじめに

近年の急速な科学技術の発達、科学技術情報の質的高度化、量的拡大、記録媒体の多様化等の変化を引き起こしている。この変化は、情報の流通形態に革命的な変化をもたらし、情報分野の科学技術の発達を更に促進させつつある。

しかし科学技術情報を取りまく状況が多種、多量、複雑に変化する中であって、旧来のデータベースシステムでは対応しきれないという問題が出ており、これへの対応のために、新しい情報ベースの考え方が必要になってきている。

ここでは、科学技術振興調整費による総合研究「創造的研究支援のための自己組織型情報ベースの研究」の中で研究している自己組織型情報ベースシステムについて紹介する。

2. 自己組織型情報ベースとは

情報を適切に処理し、自己秩序を形成して成長していくという、生命それ自体がもっている「自己組織化」の考え方は、これからの情報分野で重要な概念（コンセプト）になると思われる。

とりわけ科学技術情報の高度活用を可能にするために、人間の頭脳をモデルにした人工知能の原理や自己組織化の原理などを取り込み、システム自体が知識を覚え、秩序を作り、学習しながら自己成長していく自己組織型情報ベース（情報ベース）システムの実現が望まれる。また、このシステムは数値、文字、図形、音声、画像等いろいろのメディアに対応できる必要がある。

研究では、このようなコンセプトの下に、「情報ベース」という新しい概念を研究し、情報ベースシステムの構築を行おうとするものである。

3. 自己組織化と学習

情報ベースが自己組織化するということは、まさしく情報ベースが学習することに等しい。例えばコンピュータが与えられた問題を解くために

は、問題固有の知識、常識など多くの知識が必要となる。与えられた問題が解けないということは、さらに知識の範囲を拡大する必要が出るのである。

このように増大するすべての知識をシステムに導入することは、学習機能無くしては不可能である。また、社会的にも、蓄積されるデータが急速に増加しており、これらを格納、組織化、伝達するだけでなく、データベースや情報システムを概念解析によって自動的に知識ベース化することが望まれる。

一方、学習の問題は、情報科学においてはオートマトン理論、確率過程論、数理統計学など、システムのレベルでは、計算機、神経回路網、制御、ロボットなど広い分野で取り扱われている。また、応用面でもニューロコンピュータ、知的ロボットなどで研究が進められている。

機械学習は、人工知能において最も重要で困難な研究領域である。工学的にはある程度の知識を与え自動的に学習するシステムを求め、科学的には人間の学習能力そのものに関する研究が進められている。

しかし期待するレベルにははるかに遠く、現在の人工知能システムでは、ほとんど学習機能を備えられていないか、学習機能が備えられている場合でも、かなり限定された機能となっている。

4. 自己組織型情報ベースの研究開発の方向について

旧来の情報システムの枠組みを見直し、次世代の情報システムとして、今後活用されるメディア、可能になるであろう情報の高度利用、将来の電子技術にも対応可能なシステムの構築を、以下の三つの観点に従って研究を進めている。

(1) データベースの自動構築

既存の情報システムは、概念設計において既存の計算機の処理能力の影響を受けるため、計算機向きのシステム設計になる。すなわち利用者（人間）の発想、研究の手法等の支援よりも、情報の

所在または比較的単純な情報を効率的に蓄積、管理、検索、出力することを主たる目的としているので、形式の定まった情報の一元的な管理には適しているものの、情報の高度化、複雑化、発生形態、流通形態等の変化に対応していくことは困難である。

このような、一元的情報の効率的管理を目的としたシステムは、情報システム、データベースを構築するための情報の加工・投入またはその管理のために多大の知的労働力を必要とするものである。

新たな情報ベースシステムにおいては、多種多様の情報を柔軟に取り込み、処理することが可能な、比較的原情報に近い形式でシステム内に蓄積する。

そのため、情報投入のための処理・加工、蓄積情報を管理する目的の知的労働資源の投入は、最小限とするようにしなければならない。そして多くの情報は自動的に情報ベースに取り込まれ、研究者、利用者は情報ベースの利用に専念することができる。

(2) 情報の自己組織化機能

既存情報システムは、研究活動の高度化に応じて研究者が行う文献調査、データ整理などの情報収集、分析作業量の増大に対応し得るものではない。情報システムの主流はキーワードの検索による所在案内であり、利用者がシステム内の情報に対して十分な知識を有することが必須であるとともに、そのシステムだけで必要十分な情報を得ることは難しい。

新たな情報ベースシステムは、自己組織化機能を付加することによって、研究者が新たな知見を得るため、情報を収集し、分析する作業の大部分を代替し得る機能を備えるものでなければならない。自己組織化機能により、個々の研究者が求める情報、利用傾向等をシステムが理解し、求める情報をフルテキスト、マルチメディア、知識、総説等の必要な形式に加工して情報を出力する。

(3) 自己組織型情報ベースの高度利用

既存の情報システムでは、研究者があらかじめ

質問式を作成し、イエス、ノーで情報の有無を回答することを基本としており、研究者の発想を支援し、あるいは研究者に新たな研究フィールドを示唆し、導くようなシステムとはなっていない。

情報を柔軟に処理・蓄積している情報ベース内においては、フルテキスト、マルチメディア、知識等が蓄積され、その情報は原情報のままで、または利用しやすく加工され、管理されている。

このシステムから情報を効果的、効率的に抽出するために、新たなヒューマンインタフェースの研究が不可欠となる。数値データを解析する、複数文献を比較分析し、その結果から推論を行う等の機能の開発整備と、それらを研究者が自由に利用することを可能にするヒューマンインタフェースの研究を行う必要がある。

5. 自己組織型情報ベースシステムの全体イメージ

ここで、情報ベースの概念図を掲げる(図1)。システム全体は(1)情報獲得部、(2)情報自己組織化部、(3)情報格納部、(4)応用インタフェース部の四つの部分からなる。

(1) 情報獲得部

この部分において、外界からの様々なメディアを通じて情報を獲得する。テキスト情報、表情情報、図形情報、音声情報、動画情報、その他が考えられる。

テキストに関しては表記/用字/用語の揺れや他国語の扱い等の組織化がある。テキスト以外のデータあるいは文書におけるテキストを除いたレイアウト情報を含むデータに対しては、フォーマットや規格化といった表層的構造化を通しての組織化がある。

(2) 情報自己組織化部=内容的組織化機構

この部分では動的シソーラス(ダイナミックに組み換え可能なシソーラス)や汎用/対象別タキソノミー(物質の物性・構造・特徴等を含む論理構造)を利用して、概念階層/論理構造のガイドラインとし、(1)からの入力情報にハイパーリンク

創造的研究開発支援のための自己組織型情報ベースシステム構成図

1
✕

(多種のメディアを結ぶリンク)を張っていくことにより、情報の自己組織化を進めていく。このプロセスを「内容的組織化」と呼ぶことができるだろう。特に、シソーラスを用いて「意味的組織化」を、タキソノミーを用いて「内容的組織化」を行う。

この部分に各種の帰納学習機能を付加してやることにより、様々なタイプの学習や情報の解釈機構を実現できる。当然、動的シソーラスやタキソノミー自身も情報の増加に伴って進化していくことになる。

(3) 情報格納部

この部分に、様々な情報レベル、様々な形態の大量のデータを格納する。テキスト、数値、表、図形、ビットマップ、音声、アニメーションをはじめとして、数式、経験則、法則、理論といったメタ情報(通常の情報全体)もすべて統一的にオブジェクトとして、アクセスおよび操作可能な枠組みを提供している。データの圧縮や正規化、インデキシングの一部もこの部分の役割である。

(4) 応用インタフェース部

ここでは、情報ベースシステムとユーザとのインタフェースを管理する。第1に、ユーザからの問い合わせに対応する。このとき、問い合わせと、自然言語、連想や分脈、図や画像等を用いた問い合わせに答えられる必要がある。第2に、より高次の研究開発プロセスに、より踏み込んだ形の応用に対応できる必要がある。

すなわち、データの解析や解釈、シミュレーションや予測、評価、問題解決、意志決定といった諸問題に的確に対応し、ユーザを支援できる必要がある。このとき、ユーザの認知モデルや思考様式をモデルとして保持した、ユーザ思考のインタフェースが提供される必要がある。

6. 各種自己組織型情報ベースシステム

このような情報ベースというコンセプトの下にこの研究に参加した機関は、11機関あるが、その中からデータベースの自動構築、自己組織型情報

検索、自己組織型情報ベースの高度利用の観点を研究している、三つの研究を紹介する。

6.1 論文自動認識システム

(日本科学技術情報センター)

科学技術情報は、文字・図形・画像といったマルチメディアの情報であるのみならず、情報が高度に組織化可能であることが必要である。このため、情報ベースへの入力に際しては、例えば論文の標題、著者、抄録、本文、図表といったデータ項目に分類することが必要となる。論文誌等の科学技術情報を光学的に入力する際も、こうしたデータ項目を自動識別することが、人手を大幅に軽減できることになる(図2)²⁾。

このため論文誌等においては、入力したいデータ項目がレイアウトに反映されていることに着目して、レイアウト情報による解析により、項目を抽出することが考えられる³⁾。

図3は、レイアウト解析による解析結果を示したもので、レイアウト情報として各項目の識別情報を記録した識別関数モデルと入力画面から得られる画像特徴(連結する黒画素群の位置座標など)を比較し、項目領域を識別する。

レイアウトの固定的な項目については、識別関数モデルを用い、レイアウトが固定的でない図表領域については、画像としての特徴を抽出し、図表領域を識別する手法をとる。

また、文字認識では、文字の書かれている論文を、イメージリーダから読み込んでイメージデータとし、これを1文字ずつ切り出し、文字認識を行う。文字認識では、結果として画像特徴の似ている複数の候補文字が出力される。このとき候補の第1位が正しい文字であるとは限らない。このため文字認識後処理として、認識精度を向上させる単語辞書との照合を試みた(図4)。

今回行った文字認識後処理は、単語辞書と認識候補文字との照合により単語を比較する単語照合処理、抽出した単語間の文章並びの妥当性を調べる接続チェックである。

この研究の自己組織化の取り組みは、識別関数

論文自動認識システム

特徴

- ・レイアウト知識を利用した項目の自動抽出
- ・文章、表、図の自動分離
- ・文字認識
- ・形態素解析を利用した誤認識文字の自動指摘
- ・認識結果はSGML文書に変換して格納

図表分離

固定要素モデル

領域の分離

固定要素領域

不定要素領域

レイアウト解析

文字認識後処理

表・イラストなどを含む文書

イメージリーダー

図 2

図5 自己組織型情報検索システムSIRの環境の全体像

モデルを設定して、出現する位置が固定的なデータ項目に対して自動的に項目の判定を行い、なおかつ画像としての特徴を抽出して図表領域を識別している。さらに文字認識を行い、文字認識候補文字について単語辞書と照合を行い、辞書チェックを行っていることである。

6.2 自己組織型情報検索システム：SIR (電子技術総合研究所)

自己組織型情報検索システムでは、技術マニュアルや文献、技術情報等の分野で使うことを想定し、原情報は豊富にあるが、情報検索のための情報の集積が不十分な状態においても、情報検索を可能にし、使っている間に情報検索のための情報の集積を行えるシステムの実現を目標としている⁴⁾。

このような状況での「情報の自己組織化」の初期段階をサポートする検索機能と特徴を、より具

体的に述べると、図5のようにまとめることができる²⁾。

これは電子化文献の生データからキーワードを抽出しキーワードの出現頻度を解析する。この頻度情報を用いてキーワード層とドキュメント名層の2層相互結合ネットワークの初期荷重を計算する。これに基づき、あるキーワードを入力すると関連度の強いキーワードが引き出され活性化する。同時にネットワークによってリンクされている、ドキュメント名層へも活性化が伝わる(図6)。

これをネットワーク表現ができるようにして情報検索の結果の可視化を行った。図7、図8において、画面左上は活性領域のネットワークを表示するウィンドウ画面で、2層の上段はキーワードノード層、下段はドキュメント名ノード層である。表示されている円は各々のノードの活性化の大きさを表現している。中央側に重要なノードが配列されている。

図6 情報検索のネットワーク表現とコネクショニスト情報検索の原理
(連想活性化伝播)

図7 Unix Command fileの自己組織型情報検索システム：SIR (学習前の状態)

図8 Unix Command fileの自己組織型情報検索システム：SIR (学習後の状態)

6.3 自己組織型情報ベースのための知的インタフェース (CastingNet)

(富士ゼロックス)

研究活動において創造的活動を支援し、かつ情報ベースの利用性を高める知的インタフェースは、多種多様の情報ベースおよび大量の未体系情

報の研究に必要な情報を一元的に表現・操作でき、かつ、情報群を体系的に組織化する過程を支援できる必要がある。

データベースでは、扱いにくい未体系情報の蓄積と組織化には、ハイパーメディア技術が有効となる。ハイパーメディアは、従来のデータベースのようにあらかじめ抽象構造(レコードやクラス)

図9 CastingNetシステム構成

図10 可視化ブラウジング

を定義しなくても、直接、情報を入力できるという点で有効である。この研究では、新たなハイパーメディアモデル「フレーム関係軸モデル」を考案した(図9)⁵⁾。

そしてこれに基づくシステムとして、未体系情報群の主観的な可視化を支援し、それら情報群を様々なチャートによって可視化してブラウジングする機構「ハイパーチャート」を構築した。

これは図10が示すように、表示されているフレームビューの関心のある特定のスロットに基づいて写像されている関係軸を含むハイパーチャートの候補が、メニューにより一覧される。その中から表示すべきハイパーチャートを選択することができる。さらにハイパーチャート上で関心のあるフレームアイコンを指示することにより、該当するフレームビューが表示される。このように次々にハイパーチャートとフレームビューを渡ることができる²⁾。

7. おわりに

紹介してきた研究を含む情報ベースの研究は、まだほんの入り口を入ったばかりであり、いかにして生命体が持っている自己組織化機能を究明し、進行中の研究へ取り込むかが研究の鍵になる。しかも近年、計算機の世界では、数値や文字ばかり

りでなく音声や動画像のマルチメディアの時代に突入している。

こういった分野を深く研究することによって、

情報の氾濫する科学技術の分野で、研究者に役に立つ、自己組織型情報ベースシステムを開発していくことが必要になっている。

参 考 文 献

- 1) 創造的研究開発支援のための自己組織型情報ベースシステムの構築に関する調査成果報告書 p.215—218
- 2) 自己組織型情報ベースシステムの構築と利用に関するシンポジウム 予稿集 p.62
- 3) 西村康, 高橋友一, 小林幸雄: 論文フロントページの構造理解: 電子情報通信学会技術研究報告, PRU87—54 (1987)
- 4) 仁木, 渡辺, 杉崎: コネクショニスト&論理の統合型情報検索システム: CLIP, 電子情報通信学会 1993年春期大会
- 5) 石飛康浩, 植田学, 増田佳弘: 「CastingNet: 大規模情報を分析するためのハイパーメディアアプリケーションフレームワーク」 情報処理学会: 情報学 基礎研究会 30—2 p.9(1993. 7)