総説

情報知識学試論†

藤原 譲††

データ、知識、情報に関する理論や性質に関する学問としての情報知識学について、その生い立ちの背景、とくに情報の資源化および活用の観点から現状、理論体系、問題点、展望を述べる。

1. 序

情報処理や通信の技術や理論の急速な進歩とともに、処理され、流通する情報の量が膨大となり、かつ種類も理系文系の枠を越えて多様化して来た。それに伴いデータや知識は、入力すること自体が容易でないことも明らかになって来た[1,5,8,18]。これらは入力されたデータや知識が、適切に管理され、有効に利用されることが期待されているのに対しシステムの機能が不適合または不十分であるためである[10,11]。

典型的な例を挙げれば、データベース管理システムで、多様化したデータに対しては、最低限の要求であるデータへの正確なアクセス機能すら、特殊な場合を除いて十分に満足させるものはない[6,14,15,16]。

すなわち情報のもつ特性に対し、管理や識別の手法が適切でなければ、管理不能となったり、アクセス精度の低下をもたらすことになる。さらに、高度な思考過程に対応する機能をシステムとして実現するには、数値計算および、符号照合に依存するデータベース検索や演繹推論中心の論理演算を超えた情報の処理方式の確立が必要である[1,3,7,9]。

もちろん帰納推論、学習、連想等に関しての研究も盛んに行われており、その一方でハイパーメディアのように、文書構造化による思考支援の観点から出発して、自己組織化、類推、発想、問題解決のシステム化を目指した技術開発も進み、特殊用途では実用の段階に達している[1,12,13]。

以上のような現状からみて、情報に係る多くの課題を解決し、高度に発展しつつある各種の情報処理や通信の技術と理論を活用して必要な情報を適切に管理し、有効に利用するには、まずデータ、知識、情報の本質を把握し、基礎となる情報自体の理論を体系化し、今後の研究、開発の動向を洞察することが必要である。

情報知識学はこのような研究、開発のために設けられた新しい学問の分野である。現時点で情報知識学の領域に関しては様々な考え方があり、統一された定義があるわけではないが、情報知識学を情報の資源化、およびその活用のための基礎として、情報知識学の体系化を目指した研究分野についての試論を述べる。

2. 情報のキャラクタリゼーション

2.1 定義

ここで用いる用語は以下のように定義されているとする。

情報 (information) とは、ある対象について認識された内容またはその一部をさす。その認識の対象を実体 (entity) と称し、実体のもつ性質 (property) を属性 (attribute) という。

したがって認識内容の記述項目が属性名となり、その最小単位の値およびその集合をデータ (data) と呼ぶ。

知識 (knowledge) は、広義には抽象化された情報を意味するが、狭義にはプロダクションルールや述語論理など一定の構成規則で定式化されて、表現された規則をさす。

データの集合で、それだけで独立した資源とみなされるときデータベース (database) と呼び、知識に対しては知識ベース (knowledge-base) と呼ぶ。

† A Review on Science of Information and Knowledge by Yuzuru Fujiwara

†† 筑波大学電子情報工学系

2.2 情報の性質

情報の概念を規定するために、その特性別に論ずることにする。よく知られているように情報は量、意味、構造、媒体、質、動態の6項目に大別される側面を有している。

2.2.1 情報の量

いわゆる情報理論は情報の量と符号化に関する理論である。情報が符号化されたときその符号の列 S の生起確率 $P(S)$ は構成要素の符号の生起確率 P_i とすると

$$P(S)=\prod_{i=1}^{n}P_i \tag{1}$$

となる。情報量 Iは通常 $P(S)$ の逆数の対数で表わされる。

$$I=\log a\frac{1}{P(S)} \tag{2}$$

a を2とするとスィッチング回路を基本とする計算機などに好都合な単位となりビット (bit) と書く。

これは個別の情報に対して用いることもできるが、通信では情報源に適用することが多く、情報処理においては媒体とくに物理媒体の容量に用いることが多い。従来は利用可能な情報媒体の容量に対して強い制約が課せられていたが、容量に関しては今後は目的に則して必要最小限ではなく、適切な量を検討すべき段階に入っている。

なお生起確率を自然語に対応して Markov 型で取り扱ったり、符号化にエラーチェックなどの機能を組み込ませるなど通信理論、符号化理論は既に長い歴史を有し、確立した分野であるのでここではこれ以上立ち入らないことにする。

ただ、計算機の記憶容量と通信技術の進歩は著しい。従来の印刷物中心の情報の伝達や操作の単位は、表1に示すようによく用いられる論文や専門書がそれぞれ 0.1MB、1MB 程度であるのに対し、光ファイルに代表されるニューメディアは千倍以上の 1GB が普通であり、今後さらに千倍 TB 以上の拡大が予想されている。

このような大きな変化は情報の流通の処理に対し質的な変化をもたらすことになる。とくに、専門領域の網羅的情報の量が 100MB から 1GB 程度と見積られることと比較して、非常に意味があることになる。

表 1 Quantity of Information Flow

Articles for Reading	0.1 ～ 1 MB
Working Articles on a Desk	5 ～ 50 MB
Dictionaries, Handbooks	
Specified Domain Articles	100 ～ 1000 MB

2.2.2 情報の意味および構造

情報の内容すなわち意味するところのものが情報の最も重要な性質であるが、これは同時に機械処理における最も困難な問題でもある[3,5]。

言語学における意味論 (semantics) と関わりも深いが、情報の利用、生産においては構造解析、記述項目、表現方法などはとくに重要である。表現は別項で取り上げるが、記述項目と構造は意味に直結している。

情報空間は現実世界の射影とみなせるので、現実世界のもつ多種多様な意味的関係がすべて情報空間の構造に反映されることになる（図1参照）。そこで、情報の空間が内部に持つ要素間の関係を解析して、構造を把握しようとする構造解析の立場と、典型的な基準構造を枠組にして情報空間を扱うモデリングの立場とがある[5,17]。

図書、動植物などの分類は木構造をモデルとする手法である。また、データベース管理によく用いられる関係モデルにおいては関係は属性の定義域を D_i とすると次式のように D_iの直積の部分集合として表わされる。

$$R(D_1\times D_2\times\cdots\times D_n) \tag{3}$$

しかし、これでは不十分として関係モデルの提唱者 E.F.Codd は5年後に、意味を扱うために関係間の関係を明示的に扱う拡張関係モデル RM/T を提案している。

また、抽象データ型関係モデル (abstract data type relational model) ほかのモデルも、拡張関係型モデルと同様関係モデルの特長を活かしつつ、より複雑な構造のデータに対応できるように制約を緩和しようとする試みである。

CODASYL に代表されるネットワークモデルは木構造モデルの一般化とも言えるが、P.Chen の実体—関係モデル (entity-relationship model:E-R)、意味ネットモデル (semantic net model)、オブジェクト志向データベースモデル (object oriented DB

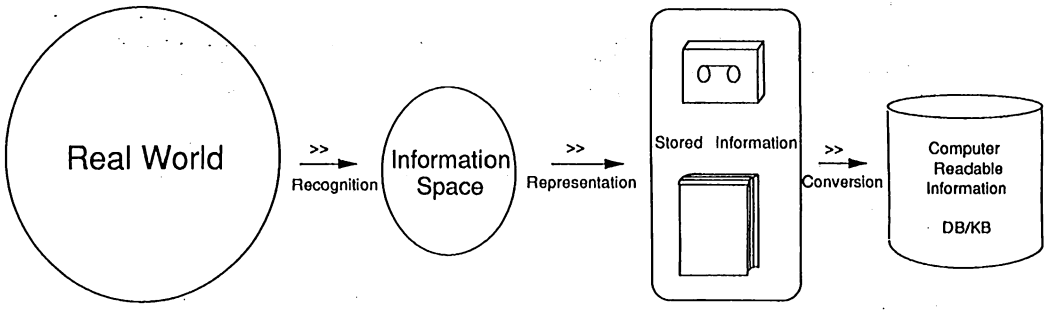

図 1 Relationship among the Real World, Information Space and Databases and Knowledge-bases

model) コネクショニズムモデル (connectionism model) など最近注目されているモデルの多くは、ネットワーク型である[7,9,10,15,17,18]。とくに最近急速に関心を集めているハイパーメディアはノードとリンクからなる柔軟なネットワーク構造を持っている。いずれにしても現在の計算機上で意味を扱うにはデータまたはレコードの構造として表現し、それに応じて処理を行う必要がある。

2.2.3 情報の媒体

情報は伝達、記録されるために媒体上に表現される。媒体は言語のように論理的な媒体と、文字符号、信号のような物理的な媒体とがある。さらに物理的媒体は音声のようなアナログ型と、符号のようなディジタル型がある。情報自身にもアナログ型とディジタル型とあるが、情報の型と媒体の型は一致するとは限らない。

また物理媒体はその次元によって 1 次元、2 次元、3 次元、4 次元、多次元、複合次元などに分類できる。

媒体の型と次元を組み合わせると表 2のようになる。

情報の型と媒体の型が、必ずしも対応しないのと同じように情報の次元と媒体のそれとも一致するとは限らない[18]。例えば 2 次元、または 3 次元のグラフを線形符号化する事により一次元媒体上に表現することは可能である。また、ディジタル情報を電気で通信することもよく知られたディジタル－アナログとその逆変換で実現されている例である。

また上述の DA 、AD 変換のように媒体間の変換は情報側からは同値変換である。同値変換は物理媒体間のみでも音声の符号化のような高度な技術を要するものがある。また、論理媒体間の同値変換の代表は言語間の翻訳である。

2.2.4 情報の質

情報は上に述べたように現実世界の射影であるが、その射影の過程によって現実世界で存在する関係が成立しなくなることがある。そこで、ある基準に従って、その差異を評価することが必要となる。評価基準は情報の目的によって設定されることになり、さらにそれに応じて評価方式が決められる。

与えられた情報を予測結果と比較することが代表的な評価方式である。この方法は数値情報にもよく用いられるが、一般的には意味と係わりが深く、解決困難な問題の一つである。

2.2.5 情報の動態

情報を対象とする機能は記録、伝達、処理が主たるものであるが、伝達は情報の位置的移動であり、記録も長い時間を伴う伝達と考えられるので大きくは伝達に含ませることが出来る。とくにマスメディアや各種通信手段を含めた高速かつ広汎な大量の情報の伝達が行われるようになっているので、情報の流通という表現もよく用いられる。

また情報の処理においては、媒体のところで述べたような同値変換のみではなく、抽象化のような圧縮、外延への展開、さらには類推、連想のような構造、または表現の変換など情報は様々な変態を遂げる。

さらにまた情報は利用頻度利用水準において利用

表 2 Physical Media of Information

Type / Dimension	Analog	Digital
1	Sound Electronic Signal	Sentence,Number
2	Image	Table,Topograph
3	Solid	Stereograph
4	Motion	Digital Video
Composite	Multimedia	Multimedia

者すなわち社会の価値観に応じて変化する。

以上のように情報は空間、時間、内容、表現、利用において変化を受けるものである。これが情報の動態であり、情報のもつ基本特性の一つである。

3. 情報の基礎理論

3.1 情報の記述

情報の内容は何らかの媒体に表現されなければならないが、表現されるべき内容を全て記述することは通常極めて困難であり、また、必要とも限らない。そこで選択されたものが記述項目とその内容である。

これらは情報の利用目的に応じた選択基準に基づいて定められるが、対象概念の利用観点からみた特性が記述項目に含まれる必要がある。

特に重要な記述項目は主要キー属性 (primary key attribute) と呼ばれるもので、実体の同定、識別可能な識別子 (indentifier) を構成する項目であるが、これについては管理のところで言及する。

3.2 情報の表現

情報の記述項目とその内容は、論理媒体を介入として物理媒体上に表現される。記憶、流通、処理に適した表現方式が望ましいが、記憶に関しても収録に便利な簡潔な表現と、高精度アクセスに必要な表現とでは相互に矛盾する要求となってくる。このことはシステムの多様性と同じ程度に表現の多様性があることになり、システム内およびシステム間で内容の変更を伴わない等価な変換および、目的別の内容変化を伴う変換が起こることになる。媒体の型や次元が異なるときは表現法も当然異なってくるが、一つの言語に限定しても多様性が存在する。たとえば、用語には本来同意語、多義性、類似性等があるため表現の多様性がさけられないことになる。

このような用語の問題を中心にして取り扱うのが語彙論 (terminology) である[2,4,16]。

また、2次の情報は、画像はもとよりグラフに関しても表現の線型化、符号化は大きな課題である。

3.3 情報のモデル

既に 2.2.3 で情報の特性の一つとしての構造について述べたように、よく知られた3大モデル（木構造、網構造、関係型モデル）があり、それらの拡張型のものも多数報告されている。木構造と網構造はラベル付き有向グラフとして扱うことができる。

関係型モデルは集合および述論論理によって数学的基盤が与えられている。三大モデル共通の関数従属性の他に多値従属性や結合従属性なども扱える[5]。

またハイパーメディアでは利用者が内容、目的に応じノード間にリンクを張ることによって構造化を行なう[17]。ただしリンク情報のシステム管理がなかったり、または不充分な場合が多く、モデルとしては確立されていない。

全体として現時点のモデルはレコード中心であるが、ハイパグラフを含めて構造情報を扱うには、グラフ理論的観点も今後重要となるであろう。

3.4 情報の管理

図書の管理にみられるように分類は管理上非常に有効な手段である。国際的にも FID を中心にして 10 進分類法である UDC が普及している。

10 進分類法は可算集合または全順序集合である線型集合に対しては、優れた方法である。しかし一般の情報は線型集合ではなく、また情報空間を被覆する半順序関係の集合を決定することも通常は極めて困難である。分類基準の順序付けと、基準カテゴリーの明確化に関する技術によって分類法は支えられている。

また逆に分類記号による情報の記述、分類も試みられている。

情報の可算性に関して、個別的外延 (specific

表 3 Relationship between Manipulation and attributes of Information

Manipulation / Attributes	Collection	Management	Transfer	Use
Quantity	◎	○	○	
Structure (Meaning)		◎		◎
Quality	○			○
Media		○	○	○
Dynamics	○		◎	

extension) と包括的内包 (generic intension) というレベルの異なる概念を同一の枠組で扱うためのモデルとして、抽象データ型、オブジェクト指向型または演繹型のものが数多く報告されている[8,10,14,15]。しかし、これらの概念には相対性 (relativism) があり、また固定的でもない。結局これまでの理論や技術で実際に必要な複雑な情報を扱う柔軟さを持ち適切な管理のできる方法は確立されていない。

3.5 情報資源化

情報の資源化の第一歩である収集においては、まず量が問題となる。実体について必要な項目についての記述が適切な方式で表現されていなければならない。Shannon の通信理論によって情報の量は確率論的に定量化することができる。しかし実際の処理においては、表現値の最小単位としてのデータの個数、またはデータの集合であるレコードの個数として定量化されることが多い。この方式は、情報源または記憶装置の情報量との対応が良く、また後に述べるように管理的にも扱い易いことに依る。

現実に有限のレコードを扱っていく時、この方式はきわめて自然な考え方ではあるが、その問題は情報空間が可算集合でないことから生ずる。可算集合は数学的に明確に定義されている

ある集合の要素 a, b, c に対し、反射、反対称、推移の 3 つの法則が成り立つ関係 Rを半順序関係と呼ぶ。

$$aR_a \tag{4}$$

$$aR_b \ bR_a \rightarrow a = b \tag{5}$$

$$aR_b \ bR_c \rightarrow aR_c \tag{6}$$

ある集合の全ての要素が半順序関係を満足すれば、その集合を線型集合と呼ぶ。さらに、線型集合が自然数と 1 対 1 の対応がとれるとき、その集合を可算集合と呼ぶ。また、ここで 5' に示すように反対称性が対称性に置き換えられると、

$$aR_b \rightarrow bR_a \tag{5'}$$

これは、後で出て来る同値関係と定義される。

推移則が符号の照合を中心とする検索や演繹推論を超える情報処理の手掛りとなるものであることは注目すべきである。これは、たとえば対訳用語等からのシソーラス、すなわち同語間の同値関係、上下関係の抽出に役立つことが示されている。

レコードの集合に対して半順序関係は成立するが、通常可算集合ではない。また、各要素は相互に異なるもの (distinct entity) であるという集合の前提条件が必ずしも成立していない。これは包含関係にある総称表現と個別表現が共存したり、類似の実体が存在することによる。記述、表現の問題も見過ごされ易いが、よく引用されるように 2 次元の画像を伝達や処理し易い線形の言語の世界で扱うとすると、きわめて困難となる。化学グラフのように比較的簡単な構造で記述し易いと思われているものでも、元のグラフの持つ特徴を全て記述する実用的方式は見出されていない。

表現方式における問題としては、同意語や多義性などのために実体と表現の関係が多対多となることが予想外に多いことが知られている。これに対しては、シソーラスの活用や、統制語、標準化などの対策もあるがそれで解決できる範囲は限られており、しかも本質的に非常に厄介な問題である。

また、記述におけると同時に、次元の異なる媒体での表現もまた多様であり、処理方式との関連も大きい。媒体の型と次元を組み合わせると表 3のようになる。

表 4 Levels of Information

Category	Contents	Storage Type
Location	Bibliographic data, Abstracts	Data Bases
Facts	factual Data	
Analysis	Abstracted Knowledge and Rules	Knowledge-Bases
Principles, Laws	Deep Knowledge and Rules	
Learning Analogy and creation Evaluation Problem Solving	Working Hypotheses and Related Information	α
Total	IB=DB+KB+α	Information-Bases

情報の型と媒体の型が必ずしも対応しないと同じように、情報の次元と媒体のそれとも一致するとは限らない。例えば2次元、または3次元のグラフを線形符号化する事により一次元媒体上に表現することは可能である。また、ディジタル情報を電気で通信することもよく知られたディジタル－アナログとその逆変換で実現されている例である。

また情報管理上重要な分類は同値類の集合を作ることが基本であるが、より実用的には10進分類法に対応し半順序関係の組合せで情報空間を被覆する事も考えられる。ただ表現の一意性(uniqueness)と非曖昧性(unambiguity)が成立し難いことが混乱のもととなっている。これも情報量と処理手段の観点、とくに分類法の維持、改訂問題の重視などから、現時点での見直しが必要である。

次に管理の基本となる、実体の同定に関して述べる。この場合は識別子としてキーを設定することが通常用いられる方式である。一項対応(unique)、一義的(unambiguous)かつ正準化可能(canonical)であればキーと実体とは1対1に対応するので、管理・アクセス両面において好都合である。キーの3要素のうち3番目の正準化可能性は同定、識別機能としては必要ではないが、アクセスの効率上は重要である。

また情報へのアクセスのための記述子(descriptor)であるキーワードは索引として古くから定着している手法であるが、問題もある。自然言語には多くの同意語や多義性があるため、前者は再現率(recall)の低下の、後者は適合率(relevance)の低下の、それぞれ要因となっている。しかも両者は逆相関の関係にあるため、通常はtrade offの最適化を行なうことになる。

この点、検索機能を強化するため、自動索引やシソーラス利用の方式とともに、最近次項で述べるsignature fileの面から注目されている。

3.6 情報の管理、アクセス

大量の情報に対して管理の第一は同定と識別、すなわち次のことが決定できればよい。

$$A = A \tag{7}$$

$$A \neq (A) \tag{8}$$

これは表現されたレコードの空間のみに限定すれば極めて明確である。実際には実体と表現の関係は1対1ではなく、表現の多様性すなわち同意語と曖昧性すなわち多義性があるために見掛け上、

$$R(A) \neq R'(A) \tag{9}$$

または

$$R(A) = R'(A) \tag{10}$$

が成立する。ただし、ここで$R(A)$はAの表現とする。また、識別番号のような付番、ないしラベルングの手法も同一レベルの外延のみを扱うのでないため、実体の区別が明確でなく包含関係などから混乱を惹き起こすことになる[16]。図1に示すように情報としては表現の世界ではなく、実体の世界での関係、したがって意味の世界での関係であるので、人間にとっては自然に行えることが計算機には適さない機能となる。

いずれにしてもこのことは、意味理解と深く係わっているので、符号に対してシソーラス[16]、辞書

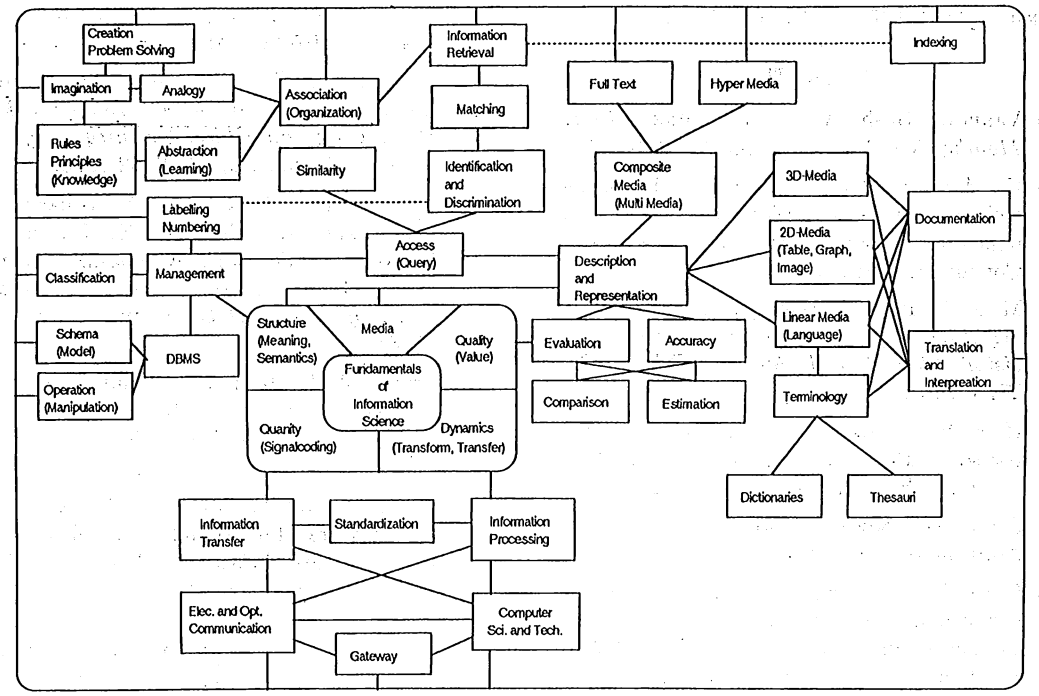

図 2 Structures in Science of Information and Knowledge

で対処する方向、情報の構造化、自己組織化の方向などがある。(表 4参照)

また、アクセスに関しても現在主として用いられているキーワードまたは識別番号などを中心とする方式も計算機向きの signature file の方式なども全文データベース化と関連して検討すべき段階にある[6]。

3.7 情報の変換

翻訳、符号化等の同値変換は既に確立された分野であるのでここでは立ち入らないことにする。

自己組織化、学習は帰納、抽象化等によって進められ、思考にとって重要な展開、発想につながるものである。

類推、連想も情報空間の部分構造同型性判定の問題として扱える。

3.8 情報の流通

通信工学で扱われる問題はそちらに委ねることにして、ここでは標準化に関する問題を取り上げる。

標準化は流通の効率化には必須のものであるが、柔軟性とは往々にして相反するものである。情報動態の観点から標準化の最適化が論じられ得る。とくによく用いられる用語については語彙論 (terminology) としての検討も重要である[2,4]。

4. 情報知識学の展望

データ、知識、情報のもつ基本的課題を明確にし、その解決を見いだすことによって、研究開発や意志決定などの思考活動を支援し得る情報の資源化とその利用の方式が明らかとなる。さらには、これらの機能と人間の思考との比較から、頭脳の構成と機能の理解が進むであろう。

また、これらはデータ、レコード中心の考え方から、意味とそれを表現する構造への方向転換である。そして、構造表現では高度なリンク機能が重視されよう。

情報知識学の全体像は図 2に示される。

5. むすび

大量情報の蓄積、流通によって情報化が急速に進みつつあり、情報洪水とも云われながら、一方では高度な思考活動には支援がほとんどなく、基本的には従来と同じであった。

情報知識学の体系化によって、新しい第二世代の

情報化社会が到来することが期待される。

文 献

1) Vannevar Bush: As we may think, *Atlantic Monthly*, Vol. 176, No. 1, pp. 101-108 (1945).

2) E. Wiister: Die Allegemeine Terminologielehre. Ein Grenzgebiet Zwischen Sprachwissenschaft, *Proc. 3rd Congress of AILA*, Copenhagen (1972).

3) M. A. Minsky: A frame work for representing knowledge, *The psychology of computer vision P. Winston ed., McGrow Hill*, pp. 211-277 (1975).

4) H. Felber: UDC and terminology. A comparison of their classification, *Proc. 41st FID Congress Hong Kong*, pp. 7-8 (1982).

5) John F. Sowa: *Conceptual structures*, Addison-Wesley (1984).

6) Chris Faloutsos: Signature files: Design and performance . Comparison of some signature extraction methods, *Proc. ACM SIGMOD Conf. Austin*, pp. 63-82 (May 1985).

7) Daniel W. Hills: *The connection machine*, MIT Press (1985).

8) Y. Fujiwara, T. Nakayama and N. Ohbo: Computer aided design system for polymeric materials, *CODATA*, Vol. 10, pp. 237-276 (1985).

9) Guy L. Steele Jr. and Daniel W. Hills: Connection Machine LISP: Fine-gained parallel symbolic processing, *Proc. ACM Conf. on LISP and Functional Programming*, Cambridge (1986).

10) Michael J. Carey, et al.: The architecture of the EXODUS Extendible DBMS, *Proc. Object-Oriented Database Workshop Pacific Grove*, pp. 52-65 (Sept. 1987).

11) J. Conklin: Hypertext: An introduction and survey, *IEEE Computer*, Vol. 20, No. 10, pp. 17-41 (1987).

12) B. Campbell and J. M. Goodman: HAM: A general purpose hypertext abstract machine, *Comm. of ACM*, Vol. 31, No. 7, pp. 856-861 (1988).

13) Pankaj K. Garg: Abstraction mechanisms in hypertext, *Comm. of ACM*, Vol. 31, No. 7, pp. 862-870 (1988).

14) M. Stonebraker, B. Rubenstein and A. Gutlman: Application of abstract data types and abstract indices to CAD data, *Proc. Ann. Meeting Database Week Sa Jose*, pp. 107-115 (May 1983).

15) D. Maier, J. Stein, A. Otis and A. Purdy Development of an object-oriented DBMS *Proc. Conf. on Object-Oriented Prog. Sys Lang. and Apple*, pp. 472-482 (Sept. 1986).

16) Y. Fujiwara, W. G. Lee, Y. Ishikawa, T Yamagishi, A. Nishioka, K. Hatada, N Ohbo and S. Fujiwara: A dynamic thesauru for intellignet access to research databases *Proc. of 43rd FID Conf.*, pp. 47-54 (Sept 1988).

17) Frank WM. Tompa: A data model fo flexible hypertext database systems, *ACM Trans. on Inf. Sys.*, Vol. 7, No. 1, pp. 85-10((1989).

18) Y. Q. Luan, N. Ohbo, H. Kitagawa anc Y. Fujiwara: Functional approach to chem ical structure databases, *Proc. of DASFAA Seoul, Korea*, pp. 80-89 (1989).

(1990年 4 月 13 日受付)
(1990年 5 月 14 日採録)