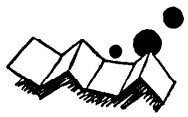
解　説



ドキュメンテーションと用語†

中　井　浩††

1. はじめに

情報をより効率よく流通させるシステムは，企業における MIS，社会における各種情報の供給（科学技術，経済，信用，法令，各種ニュース等），国際的な学術情報の流通等，種々の形で多様な規模で形成されつつある．それらのシステムを構成する多くのサブ・システムは，知的な作業ではあるが労働集約的である．急速に知的に成熟しつつある社会の中で，情報の増加に対応しつつ更に処理精度をあげるために，多くの所でそれらサブ・システムの機械化を試みている．従来，人間の行ってきた作業を機械に代替するためには，その過程の数学モデルが必要である．

2. データベース・モデルと概念論

データベースは，何らかの対象に関する情報の，再び取り出せるように組織化された集まりである．必要なものが，必要としたときに，必要なもののみ取り出せるためには，それを取り出さねばならぬ情報要求について，必要なレコードと必要でないものとが区別できなければならない．情報要求は千差万別であって特定できない．そのため，ファイル内レコード間の区別は，一般的なものでなければならない．

受けとった情報によって，特定のものを他のものと区別するのは，概念論の領域に属する．20世紀の初頭に，新カント派の哲学者 Ernst Cassirer (1874～1945)[1] は，アリストテレス以来の伝統的概念論，ヘーゲルやマルクスの概念論にとらわれることなく，技術論に基づく新しい概念論を提唱した．人間は対象を観察するに当り，その観察の目的に基づきいくつかの観察視点を定める．その視点ごとに観察の結果として与えられる値のスコアによって，対象を規定しようとするものであった．その視点として Cassrer は，その対象のもつ要素機能をとり，対象をそれらの機能の集まりとみなした．それゆえに，彼の概念論は機能概念と呼ばれる．しかし，観察視点は，必ずしも機能である必要はない．データベース理論は，この概念論の考え方を出発点としている．

対象は観察可能量の集まりである．その観察可能量の中で，特定の対象を他と区別するに必要なものを属性と呼ぶ．観察とは，特定の目的のために，いくつかの観察視点を定め，その観察視点ごとに観察可能量に対して値を与えることである．観察によって各属性に与えられる値を，属性値と呼ぶ．

すなわち，観察過程は，対象の集合Oと，属性の集合Aと，属性値の集合Vと，特定の対象について，各属性に対し属性値を与える規則の集合ρによって，次のように記述される．

$$\text{観察}=(O, A, V, \rho)$$
$$\rho: O\times A\longrightarrow V \quad \text{又は} \quad \longrightarrow 2^V \qquad (1)$$

3. ファイルとしての入力文献

ドキュメンテーションとは，主として文献情報をより効率よく流通させるための諸活動である．この流通過程においては，物流における倉庫と同じように，ファイルが重要な役割を果たす．発生した情報は記録され，ファイルに格納される．このファイルは，情報流通過程の中で，幾段階かの構造をもつ．情報要求はファイルへの質問としてファイルへ入力される．そして回答が出力される．すなわち，オリジナル・データのファイルへの入力と質問の入力という．二つの入力経路と，回答という一つの出力をもつ．

入力文献は，それ自身でファイルである．著者の頭の中にある思考対象Oを，属性Bごとに(この属性は，観察者の視点によって多くのバラエティが想定できる．例えば「標題，著者名，抄録，はじめに，本論《節の繰返し》，おわりに，引用文献」という設定も可能である．しかし，全体を「文の集まり」或いは「語の

† Documentation and Vocabulary by Hirosi NAKAI (Technical Coordination office, the Japan Information Centre of Science and Technology).

†† 日本科学技術情報センター技術管理室

列」というような設定も可能である.)属性値を与えることによって構成される．この各属性に与える値は，通常，自然言語或いは数式等の人工言語による文の列である．そして文は語の列であり，語は文字の列である．すなわち著述過程は次のように表わされる．

$$\text{著述過程}=(O,B,U,Q)$$
$$\text{規則}Q:\ O\times B\longrightarrow U$$
$$\text{属性値}U=S^* \quad S:\text{文の集合}$$
$$\text{文}\quad S=W^* \quad W:\text{語の集合}$$
$$\text{語}\quad W=L^* \quad L:\text{文字の集合} \qquad (2)$$

そしてAを任意の集合とするとき，A^*は，Aの要素を並べて造る列の集合を表わす．

そして，その結果得られる文献Dは，各属性について配列順序の指定された属性値の列の集合である．

$$D\subseteq \underset{b\in B}{\pi} U_b, \qquad U_b\subseteq U$$

すなわち，入力文献は各属性ごとの属性値集合U_bの直積の部分集合，言いかえれば属性間の関係である．

4. 属性値変換過程（インデクシング）

ファイルへ文献データを入力する過程での観察をインデクシングと呼ぶ．インデクシングは，入力文献をインデックスに変換するプロセスである．そして，インデックスは，対象（すなわち入力文献）について，属性（すなわちデータエレメント）ごとに与えられた属性値の列である．

インデクシング・システムは，入力文献の集合Dと，属性の集合Aと，属性値の集合Vと，インデクシング・ルールの集合ρによって定まる．

$$\text{インデクシング・システム}=(D,A,V,\rho)$$
$$\rho:\ \times DA\longrightarrow 2^V \qquad (4)$$

そしてインデックスIは次のように表わされる．

$$I=\underset{a\in A}{\pi} 2^{V_a}, \qquad V_a\subseteq V \qquad (5)$$

ここで，Dは（3）の構造をもつ．ゆえに（4）におけるρは，各データエレメント$a\in A$に対し，

$$\rho(u_b,a)=\{v_a{}^i, v_a{}^m, \cdots v_a{}^n\}:\ g(b,a)=1$$
$$=\phi \qquad :\ g(b,a)=0$$
$$a\in A, v_a{}^i, v_a{}^m, v_a{}^n\in Va, u_b\in U_b$$
$$U_b\subseteq U, \qquad V_a\subseteq V \qquad (6)$$

として，$g(b,a)$は，入力文献の属性bが，インデックスの属性aと，全く関係のないとき値が“0”，関係のあるとき値が“1”をとる判別関数である．この(6)は，次のようにも書くことができる．

$$\rho_a(u_b)=\{v_a{}^i, v_a{}^m, \cdots v_a{}^n\}:\ g(b,a)=1$$
$$=\phi \qquad :\ g(b,a)=0 \qquad (6')$$

すなわち，インデクシング・ルールは，入力文献の属性値をインデックスの属性値に変換する関数である．

この変換過程は，次の基本変換或いは，それらの合成変換である[2]．今，α,β,γを属性値の列であるとすると，

（1） 同義：$SYN(\alpha)=\beta$ （7）

属性値列αを，同義な関係にある属性値列βに変換する．これは反射的，対称的，推移的である．

（2） 同値：$EQV(\alpha)=\beta$ （8）

属性値列αを，それと同値な関係にある属性値列βに変換する．これは反射的，対称的，推移的である．

（3） 類・種：$S\text{-}G(\alpha)=\beta$ （9）

属性値の間に，樹型の階層関係（上位—下位，類概念—種概念，Generic-Specific，Broader-Narrower等）があるとき，下位の属性値列αを与えると，その直属上位の属性値列βに変換する．これは反射的，非対称的，推移的である．この類—種の対応は，一般には1対多であり，下位に対して上位は“1”であるが，上位に対して下位は“多”である．ゆえに，（9）の逆は，

$$G-S(\alpha)=\{\beta_1,\beta_2,\beta_3\cdots\} \qquad (10)$$

である．

（4） 部分—全体：$P-W(\alpha)=\beta$ （11）

属性値間に部分と全体の関係があるとき，部分の属性値列αに対して，全体の属性値列βに変換する．これは反射的であり，非対称的である．この部分—全体変換は1対多である．ゆえに，(11)の逆は

$$W-P(\alpha)=\{\beta_1,\beta_2,\beta_3\cdots\} \qquad (12)$$

である．

（5） 定義：$DEF(\alpha)=\beta$ （13）

これは「αは定義によってβである」と読む．あらかじめ与えられている定義によって，αがβに変換される．これは推移的である．

（6） 属性追加：$ADD_\beta<\alpha>=<\alpha\cdot\beta>$ （14）

属性値列αのあとに，新しい属性値列βを追加する．逆に削除するときは

$$DEL_\beta<\alpha\cdot\beta>=<\alpha> \qquad (15)$$

（7） 置換：$STT_{\alpha\beta}<\alpha\cdot\gamma\cdot\beta>=<\beta\cdot\gamma\cdot\alpha>$ (16)

属性値列$<\alpha\cdot\gamma\cdot\beta>$に働いて，$\alpha$と$\beta$を置きかえる．Sort-merge に当る．

5. インデクシング・システムの実例

米国防省の DDC (Defence Documentation Center) で稼動している MAI (Machine-aided Indexing)[3] は，次のプロセスから成っている．

（1） 入力テキストを，語の集まりとみなし，語単位で読み込む．ワークエリアの各語は辞書を照合され，その語の属するカテゴリ属性と，処理指示属性が各語の属性として追加され，暫定レジスタへ格納される．〔属性追加〕

（2） 読み込んだ語に対する処理指示属性が，処理開始を指示したとき，暫定レジスタの中の属性値列に対して，処理指示属性の示す処理を加える．〔定義変換〕

（3） その結果得られた，暫定レジスタの中のカテゴリ属性の並びが，カテゴリ辞書と照合される．同じカテゴリ列のパターンが辞書内にあったとき，カテゴリ属性と処理指示属性を削除して，〔属性削除〕語の並びをインデックス・ターム候補として出力する．〔定義変換〕

（4） インデックス・ターム候補は，DRIT (Defence Retrieval and Indexing Terminology. この中には，約10万のタームより約15,000のタームへ，USE 参照がつづいている．その約15,000のタームについて，Narrower-Broader の階層関係がつけられており，インデックス・タームとなり得る.)というシソーラスと照合され，USE 参照のついている先のインデックス・タームに変換される．〔同値変換〕

（5） その結果は人間の点検をうける．そしてインデックス・タームの追加削除をうける．〔属性追加〕そしてアルファベット順に Sort され，〔置換〕重複を削除して，〔属性削除〕出力編集に送られる．ワークエリアはクリアされ，次の語を待つ．

この DDC の MAI は，一つの例にすぎない．しかし本質的な要素を含んでいる．インデクシングのプロセスにおいては，テキストは語や句や文を単位として読み込まれる．その語や句は辞書と照合されて，辞書内情報を転送される．（すなわち属性追加である.）テキスト内の属性値列は，属性値に対応する処理アルゴリズムによって処理をうける．この過程で，同義・同値・類種・部分全体；定義等の基本変換の或いはその合成変換が行われる．その結果から不要な属性が削除され，適当な置換を行って出力編集される．

6. サポート・システムと，システム冗長度

検索システムにおいて，質問は情報要求のファイル操作言語によって表現されたものである．ファイル操作言語については，これまでに多くの試みがなされた．多値論理の応用，自然言語処理における相関分析，1次ベクトル空間としての属性値空間における Similarity の判定等などある[4]．しかし目下のところ次の方式に落ついている．すなわち属性 $a\in A$ と，その属性値 $v_a\in V_a$ との間に演算 $\theta\in\{<,\leqslant,=,\neq,\geqslant,>\}$ を定義すると，その基本演算

$$a\theta v_a \tag{17}$$

を行って得られる集合，X，Y を集合代数であるとして，その上の集合算 $\odot\in\{+,*,-\}$ によって，

$$X\odot Y \tag{18}$$

によって，回答集合を求める．この基本演算に基づき，リレーショナル・モデルが導びく諸演算，すなわち射影，制限，結合，商は[5]，すべて検索システムにおいて有用である．

この中で，基本演算 $(a\theta v_a)$ を決定すること，すなわち特定のデータエレメントに特定の属性値を対応させるプロセスは，インデクシング・プロセスと本質的には同じである．すなわち，情報要求の質問化は，インデクシング・システム，特にインデクシング・ルールについての知識を前提とする．

インデクシング・ルールは，従来人間の頭脳というブラック・ボックスの中で行われて来た．そのため，その中における手続の解明がまだ殆どなされていない．それゆえに質問化に当って，インデクシング・プロセスに関する情報を明確に供給できない．すなわち，符号化の規則の明らかでない信号の復号化を行うことに相当する．

この欠陥を補うため，従来いくつかのサポート・システムが造られ，ユーザをサポートしている．シソーラス，分類表，その relative index，著者名オーソリティ・ファイル，雑誌オーソリティ・ファイル，化合物オーソリティ・ファイル，インデクサー・マニュアル，等の Tool に加えて，キーワード使用頻度リストのような Pragmatic な tool も用意されている．前述 DRIT の10万語のタームより15,000の USE 参照は自然語から統制語への変換テーブルである．

これらのユーザ・サポート・システムは，インデクシングのサイドから検索のサイドへの情報の供給である．シャノンの通信理論においては，情報源が基本的

な役割を果し，それはアルファベットXと，安全事象系の確率Pによって定まった．

情報源$=(X, P)$ (9)

そして，その情報源モデルから，$-\Sigma p_i \log p_i$ の情報量概念が導びかれた．それを用いて通信系の冗長度の概念が得られ，その冗長度の存在だけを条件として，雑音特性がいかに劣悪な通信系であっても，誤伝送と誤った復号化の可能性を，いくらでも小さくできる符号化の存在を保証する，符号化の基本定理[6]が得られた．これと同様な基本定理が，ファイルを通してのコミュニケーションである，情報流通過程においても成立する．

ファイルを通してのコミュニケーションにおいて，インデクシングが符号化であるならば，質問化は一種の復号化である．実際の検索オペレーションにおいては，そのファイルが対象とする学術分野や産業分野での一般的な知識，ファイルの構造についての知識，種種のサポート・システムの利用法に関する知識によって，インデクシングの欠陥や，インデクシング・ルールの不明確さを補っているのである．これらは，ファイルを通しての通信系のもつ内部情報量とみなすべきであろう．シャノンの通信系において，通信系の内部情報量を利用して歪んだ信号から正しいメッセージを復元するように，ファイルを通してのコミュニケーションにおいてもこれらの内部情報を利用して，精度のあまり良くないインデクシングの結果（例えば，オリジナル・テキストから，Stop Word を除き，残った語をすべてインデックス・タームとするようなレベルのファイルもある．）からも，実際に有効な情報をとり出している．これは，シャノンの符号化の定理を本質的に同じ原理が，ファイルを通してのコミュニケーションにおいても成立っていることを示している．

シャノンの理論によって，完全事象系モデルから情報量概念を導びき，情報量から冗長度を導びき，それによって間接的に内部情報量を規定できる．しかしファイル理論の場合，内部情報量の方を陽に規定するより直接的な方法の可能性を考えている．この問題は別の機会に譲り，稿を改めて議論したい．

7. おわりに

本稿において，インデクシング・システムとインデックスに関する代数モデルを提出した．その中で，入力文献の属性値をインデックスの属性値に変換する．基本変換手続を提出した．これにより，インデクシングのプロセスは一応記述できる．

しかし，これだけでは情報流通システム全体の記述として完全ではない．特に，流通システムにおける流通情報量，特にシステム冗長度或いはシステム内部情報量の概念が定式化できていない．この問題は，別の機会に論じる．

ドキュメンテーションにおける用語の問題とは，属性値変換プロセスにおける諸辞書であり，またインデクシング・プロセスの情報を検索側に伝える諸サポート・ファイルであり，処理のための諸アルゴリズムである．そして具体的には，入力テキストの中の語や句に対して，重要度判定を行うに当って必要な諸情報を与えるための属性をもつ辞書であり，分類表であり，シソーラスであり，インデクサー・マニュアルであり，relative index であり，著者名や機関名や資料に関する諸オーソリティ・ファイルであり，化合物名や構造や特性に関するオーソリティ・ファイルである．そして，これらは，流通システムの内部情報でありこれらにコンサルトすることによって，よりよき符号化であるインデクシング，復号化である質問構成と検索が可能になる．そして流通ファイルのサイズを小さくし，処理速度を高め，情報流通容量を大きくすることができるのである．

参考文献

1) Cassirer, E.: Substanzbegriff und Funktionsbegriff. (1910).
2) Landry, B. C.: A Theory of indexing; indexing theory as a model for information storage and retrieval. OSU-CISRC-TR-71-31 (PB 205829), Comp. Inf. Sci. Res. Center. The Ohio State Univ. ('71).
3) Klingbiel, H.: Machine-aided Indexing of Technical Literature, Inform. Star. Retr. Vol. 9, pp. 79-84 ('73).,
Klingbiel, P. H.: Machine-aided indexing, Report DDC-TR-69-1, AD-696 200 ('69).,
Klingbiel, P. H.: Machine-aided indexing, Report DDC-TR- 71-3, AD-721 875 ('71).,
Klingbiel, P. H.: Machine-aided indexing, Report DDC-TR-71-7, AD-733 800 ('71).,
Jacobs, Charles R.: Machine-aided Indexing, Report DDC-TR-72-4, AD-754 400 ('72).,
McCanley, Ellen V.: Natural Language Data Base, Report DDC-TR-72-1, AD 743 600 ('72).
中井 浩: 機械援助索引 (MAI) について (1), 情報管理 Vol. 3, No. 4 (July, '76).
4) 中井 浩，笹森勝之助: 情報検索システム，日本経営出版会 ('71).

Hillman, D. J. and Kasadra, A. J.: The LEADER retrieval system. AFIPS Conf. Proc. Spring Joint Compt. Conf. Vol. 34 ('69).
Salton, G.: Automated information organization and retrieval. McGraw-Hill ('68).
5) Codd, E. F.: A relational model of data for large shared data banks. Comm. ACM. Vol. 13, No. 6, pp. 377-389 ('70).
6) Feinstein, A.: Foundation of information theory. McGraw-Hill ('58).

(昭和54年6月20日受付)