情報システムのシステム特性

——シミュレーションによる文献情報システムの費用・効果分析——

電子計算機の出現により，情報システム，データ・ベース，データ・バンク，情報検索というような言葉が世の中にはんらんするようになった．さらに，計算機の大容量・高速化の進展と，社会における情報の需給の増大との相乗作用の結果として，ますます大規模で高度な情報システムの開発が試みられるようになってきている．

このような大型のシステム開発にあたっては，そのシステムが基本的にもっている固有の特性を，まず理解しておくことが必要不可欠である．しかし，計算機を計算以外の形で利用することに関しては，その歴史が浅く，充分なデータや基礎理論が整備されていないので，情報システムのシステムとしての特性を分析することは容易ではない．

一方，情報システムは計算機の出現と同時に世の中にあらわれたような錯覚をいだきがちであるが，よく考えてみると，文献情報の蓄積や処理を行なう図書館という形で古くから存在し，広く利用されていることに気づく．そして，その歴史の長さに比例する形で，個々の部分的機能の諸特性に関する分析は蓄積整備されている．ただし，情報源から使用者にいたる全体を一つのシステムとして把握することは，計算機の図書館への導入とともに順次なされつつはあるが，充分とはいえない．

そこで，本稿においては，科学技術の学術文献のための情報システムを対象として，ドクメンテーションの分野で個別に得られている諸特性を，シミュレーションモデルという形で統合することにより，情報システムのもつ基本的特性を費用・効果という形で明らかにしたい．したがって，特集テーマであるデータ・ベースを直接的に取扱うことにはならないが，情報システムの特性分析に関するモンテカルロ・シミュレーションというOR的アプローチの一例を提示することにより，この分野におけるORワーカーの活躍が今後増大していくことを希望したい．

なお，本稿の内容は，昭和45年度の大蔵省主計局の科学的財務管理事例研究として援助を受けることにより出発し，以後改良されたものであることを記しておく．

1. モデルと設計パラメータ

情報システムの機能を，収集，貯蔵，処理に分類し，情報源と使用者を環境と考えれば，図1に示すようにモデル化できる．システムは，情報源から一次情報(primary information)を収集し，これを二次情報(secondary information)に処理し，この二次情報を使用者に配布する．そこで，使用者は配布された二次情報を利用して，必要な一次情報をシステムから引き出すことができる．

以上より，情報システムをつぎのように定式化できる．まず，

S_i：情報源に存在する i 種の一次情報の集合
（たとえば，定期刊行物，単行本，テクニカルレポート）

C：収集された一次情報の集合

T_j：収集された一次情報に対する j 種の処理
（たとえば，抄録する，索引をつくる）

N_j：システムにより処理された j 種の二次情報の集合（たとえば，抄録集，索引集）

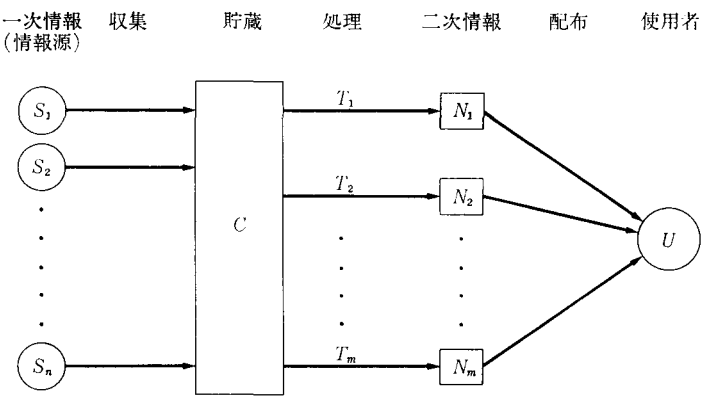

図 1　情報システムのモデル

とすれば，N_j は j 種の処理 T_j の結果として得られるものであるから，T_j は C から N_j の上への写像であると考えられる．すなわち，$T_j : C \longrightarrow N_j$，である．

上記モデルにしたがえば，システムの設計パラメータは，収集する一次情報の種類，その収集率，実行される処理の種類，の三つである．情報の種類と処理の種類の設計パラメータについては，それぞれ，収集される一次情報の種類の集合を $\boldsymbol{I}$，実行される処理の種類の集合を $\boldsymbol{J}$，とすることにより定式化できる．

同種類の一次情報の中で，どの情報から収集していくかの順序については，使用されることの多い順に収集されるという経験則を採用する．これは，同種類の一次情報を使用実績の多い順に，ランク（rank）をつけることにより操作化できる[8]．すなわち，システムがランクにしたがって一次情報を収集すると仮定すれば，収集率の設計パラメータは，システムが収集した一次情報の最後のランクにより決定できる．そして，この最後のランクを，情報源に存在するその種類の一次情報のすべての数で割ることにより，i 種の一次情報の設計パラメータ，ϕ_i を，$\phi_i = W_i/|S_i|$，

W_i：収集された i 種の情報の最後のランク

$|S_i|$：集合 S_i の要素の数

で定義できる．

2. 情報需要の分布

使用者の情報需要の分布を知るためには，おのおのの一次情報が使用者にとって必要であったかどうかを示すデータを入手しなければならない．この種のデータの典型的なものとして，学術論文の引用文献をあげることができる．化学分野の論文の中からランダム抽出により100論文を抽出し，おのおのの論文について i 種の一次情報をいくつ（h_i コ）引用しているかを調べた結果を相対頻度の形で**表 1** に示す．

そこで，R を使用者にとって必要であった一次情報の集合とする．収集率との関係では，R の任意の要素のランクがどうであるかが重要である．

表 1　使用者に必要であった一次情報の数の分布

一次情報の種類 (S_i) ＼ 引用された一次情報の数 (h_i)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
定期刊行物	0.01	0.01	0.02	0.03	0.02	0.02	0.04	0.01	0.03	0.05	0.09	0.05	0.02	0.04	0.05	0.08
テクニカルレポート	0.72	0.15	0.09	0.00	0.02	0.01	0.01									
単行本	0.34	0.27	0.16	0.09	0.05	0.03	0.01	0.00	0.01	0.02	0.00	0.02				
会議情報	0.91	0.06	0.02	0.00	0.01											
個人的接触	0.75	0.20	0.04	0.01												

(S_i) ＼ (h_i)	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
定期刊行物	0.05	0.04	0.05	0.03	0.05	0.06	0.02	0.02	0.01	0.02	0.03	0.01	0.02	0.01	0.01	0.00

これに関しては，ドクメンテーションの分野で，Brandford-Zipf の法則があり，R の任意の要素 r のランク α が W_i よりも低くない確率は，

$$P_r\{\alpha \leqq W_i\} = F_i(W_i) = \log W_i / \log |S_i|,$$

で与えられる[1]．したがって，集合 R の任意の要素 r が，収集率 ϕ_i のシステムにより収集されている確率は，$F_i(\phi_i \cdot |S_i|)$ で計算できる．

二次情報と情報需要との関係においては，つぎのことが重要である．すなわち，二次情報というものは，それ自身が使用者にとって直接に有用なものではなく，そこから一次情報を引き出すという意味においてはじめて価値をもつものである．したがって，種々の種類の二次情報，すなわち種々の処理の効率は，その二次情報によって使用者が必要な情報を引き出すことができる確率により計測することができる．

ASLIB Cranfield プロジェクトにおいては，**表 2** に示すフォームにもとづき，種々の処理についての評価テストが行なわれている[3]．そこで，

n_j：使用者に必要な情報を引き出してくれる j 種の二次情報の集合，

とすれば，R の任意の要素 r が j 種の二次情報により引き出されることは，r が $T_j^{-1}n_j$（T_j^{-1} は T_j の逆写像）に属するという形で定式化できる．そして，その確率は，

$$P_r\{r \in T_j^{-1}n_j\} = a/(a+c)$$

で与えられる．

3. システムの有効性指標と費用分析

使用者がシステムを利用することにより期待することは，必要な一次情報をできるだけ多く知ることである．したがって，システムの有効性は，

表 2 ASLIB テストの評価フォーム

	relevant（必　要）	non-relevant（不 必 要）
retrieved（引き出した）	a	b
not retrieved（引き出さなかった）	c	d

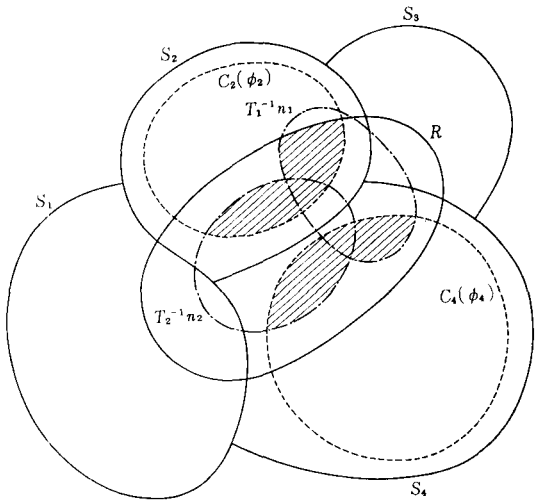

図 2 有効性指標の計算のためのフェン図

情報源に存在する使用者に必要な一次情報の数と，システムを利用することにより得られる一次情報の数との比で指標化できる．

情報システムの設計パラメータ，I, J, ϕ_i が与えられたときに，使用者に必要な情報の集合 R と，システムを利用することにより得られる情報の集合との関係は，**図 2** に示すフェン図で表現できる．図においては，$I=\{2, 4\}$，$J=\{1, 2\}$ であり，収集率 ϕ_2, ϕ_4 で収集される一次情報の集合は $C_2(\phi_2)$，$C_4(\phi_4)$ で示されている．システムの使用により得られる情報の集合は斜線で示されている．各集合の間の関係は，前節で述べたように，すべて確率現象として定式化されている．したがって，モンテカルロ・シミュレーションによりシステムの有効性指標を計算することができる．

図 1 のモデルにしたがえば，情報システムの運用費用は，収集費用，貯蔵費用および処理費用の三つからなる．したがって，設計パラメータが I, J, ϕ_i であるシステムの費用 $C(I, \phi_i, J)$ は，

$$C(I, \phi_i, J) = \sum_{i \in I} \alpha_i \cdot |S_i| \cdot \phi_i + \gamma \cdot A(I, \phi_i) + \sum_{j \in J} \beta_j \cdot B_j(I, \phi_i)$$

となる．ここで，

α_i：i 種の一次情報の収集単価，

γ：貯蔵単価，

$A(I, \phi_i)$：貯蔵数，

表 3　費用分析のための文献調査結果
（単位：ドル）

単価の種類	報告されている数値			出典
収集単価 α_i	定期刊行物	テクニカルレポート	単行本	Williams [9]
	33.71	25.97	36.26	
貯蔵単価 γ	0.135			Fussler [5]
処理単価 β_j	抄録	索引	タイトル索引	Overmyer [8]
	26.0	10.0	2.74	Linder [7]

β_j：j 種の処理単価，

$B_j(\boldsymbol{I},\ \phi_i)$：$j$ 種の処理数，

であり，各種の文献調査により得られた単価を表3に示す．

4.　情報システムの特性

前節に述べた有効性の分析と費用分析とを組み合せれば，情報システムの基本的特性を費用・効果の形で調べることができる．本稿では，以下の設計パラメータの組合せに限定する．収集の設計パラメータの要素については，

$$\boldsymbol{I}=\begin{cases}1\cdots\cdots\text{定期刊行物},\\ 2\cdots\cdots\text{テクニカルレポート},\\ 3\cdots\cdots\text{単行本},\end{cases}$$

とし，処理の設計パラメータの要素については，

$$\boldsymbol{J}=\begin{cases}1\cdots\cdots\text{抄録},\\ 2\cdots\cdots\text{索引},\end{cases}$$

とした．したがって，可能な情報システムの数は，$(2^3-1)\cdot(2^2-1)=21$通りである．収集率については，それぞれのシステムが収集する一次情報のすべての種類について，$\phi_i=0.25,\ 0.50,\ 0.75,\ 1.00$の4段階に限定した．

代表的な情報システムについての計算結果を図3に示す．図において，○印で囲まれた番号はおのおのの情報システムに対応し，設計パラメータとの対応表を表4に示す．さらに，おのおのの情報システムについて，その収集率を変化させることにより，費用・効果曲線が描かれている．

図3の結果を一般化して，情報システムのもつ基本的特性という見地から検討すれば，以下のようになる．

(1) 情報システムを④と⑩およびそれ以外の2つのグループにわけることができる．この区別は定期刊行物を収集しているかどうかに起因する．使用者が研究者であることを考慮すれば，当然の結果といえよう．このことは，収集される一次情報の種類が情報システムの費用・効果を第一義的に決定することを意味している．

(2) 収集率については，一般的には，有効性が

表 4　番号と設計パラメータの対応

番号	収集のパラメータ $\boldsymbol{I}$	処理のパラメータ $\boldsymbol{J}$
1	{ 1, 2, 3 }	{ 1, 2 }
2	{ 1, 2, 3 }	{ 2 }
4	{ 2, 3 }	{ 1, 2 }
7	{ 1, 3 }	{ 1, 2 }
8	{ 1, 3 }	{ 2 }
10	{ 3 }	{ 1, 2 }
13	{ 1, 2, }	{ 1, 2 }
14	{ 1, 2, }	{ 2 }
15	{ 1, 2, }	{ 1 }
19	{ 1 }	{ 1, 2 }
20	{ 1 }	{ 2 }
21	{ 1 }	{ 1 }

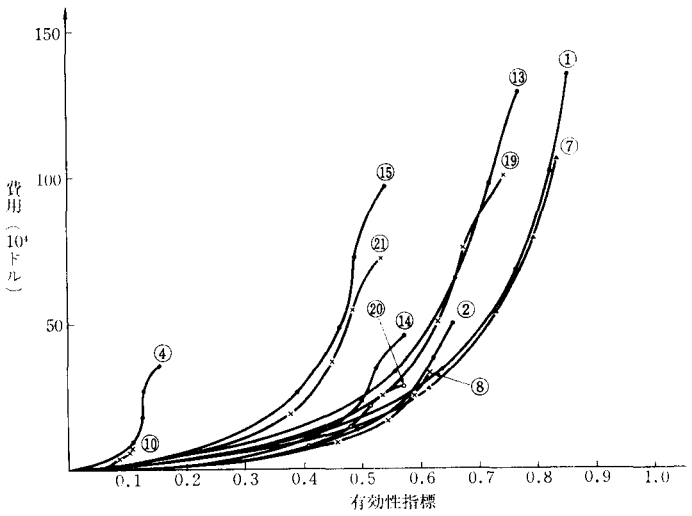

図 3　費用・効果の計算結果

上昇するにつれて費用はほぼ指数的に上昇しているといえる．しかし，システムが図3において左に位置すればするほど，費用の上昇は激しい．このことは，収集する一次情報の種類と実行される処理の種類が適当でない場合にはいくら収集率を上昇させても，有効性に限界があると同時に，収集率を上昇させることのメリットが少ないことを意味している．

(3) 費用・効果という評価基準にしたがえば，図において右下にくるほどよいシステムということになる．したがって，おのおのの費用・効果曲線の包絡線を描くことにより最適システムを設計することができる．結論的には，最適システムの費用・効果曲線は費用・信頼性曲線に類似しており，使用者の要求を100%満たす情報システムは非常にコスト高になる．

以上の結果を，今後の情報システムの開発に関連させれば，つぎのようにいえる．

使用者の要求は多様であるため，情報システムの開発には，多大の労力と費用を投入しなければ使用者にとって有効なものになり得ない．一方，特定の使用者を対象とする小規模で手軽なシステムの設計にあたっては，収集する情報や処理について，使用者の利用パターンを充分調査して，決定していく必要があろう．

むすび

ここでは，特別な例を対象として，情報システムの特性分析へのOR的アプローチの例を示したが，情報システムをとりまく環境や技術は大きく変化しているので，このような研究は持続的になされねばならない．したがって，情報システムを社会システムの一環として分析していくうえで，ORワーカーの果す役割は大きいといえよう．

参考文献

[1] Brookes, B. : The Derivation and Application of the Brandford-Zipf Distribution. *The Journal of Documentation*, vol. 24, No. 4 (1968).

[2] Carter, L. : *National Document-Handling Systems for Science and Technology*. John Willey, 1957.

[3] Cleverdon, C. and Keen, M. : Factors Determining the Performance of Indexing Systems. *ASLIB Cranfield Research Project*, National Science Foundation, 1966.

[4] Coile, R. : Periodical Literature for Electrical Engineers. *The Journal of Documentation*, vol. 8, No. 4 (1952).

[5] Fussler, H. and Simon, J. : Patterns in the Use of Books in Large Research Libraries. *The Univ. of Chicago Library*, 1961.

[6] Landau, B. : The Cost Analysis of Document Surrogation. *American Documentation*, vol. 20, No. 4 (1969).

[7] Linder, H. : Comparative Costs of Document Indexing and Book Cataloging. *Special Libraries*, vol. 56 (1965).

[8] Overmyer, L. : Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists. *American Documentation*, vol. 13 (1962).

[9] Williams, C. : *Library Cost Models*. Western Research Inc., Maryland, 1968.

こだま・ふみお　1941年生
埼玉大学教養学部助教授