

# 학술 커뮤니케이션의 수량학적 분석에 관한 연구

金 賢 姬\*

## <目 次>

- |                     |                        |
|---------------------|------------------------|
| I. 서 론              | III. 주제 문헌의 질적 여과 알고리즘 |
| II. 학술 커뮤니케이션의 수리모형 | IV. 결 론                |
| A. 정보중복             | 참고문헌                   |
| B. 종 합              | 부 록                    |
| C. 역 동              |                        |

## I. 서 론

학술 커뮤니케이션(scientific communication)은 과학자가 연구에 대한 아이디어를 얻는 데서 시작해서 이러한 연구 결과들에 대한 학술정보가 학술 지식의 구성요인(constituent)으로써 받아들여질 때까지의 정보의 生産·傳播·利用에 관련된 모든 활동을 포함한다. 간단히 말해서, 학술 커뮤니케이션은 과학자간의 정보 교환활동을 의미한다.<sup>1)</sup> 학술 커뮤니케이션은 공식 채널과 비공식 채널을 통해서 수행되는데, 비공식 커뮤니케이션은 구두(word of mouth)로 수행되며 공식 커뮤니케이션은 기록(written word)을 통해서 이루어진다.<sup>2)</sup>

일반적인 학술 커뮤니케이션양식(pattern of scientific communication)을 살펴보면, 연구를 막 완성한 과학자는 연구 결과들에 대한 정보를 비공식 연

\* 명지대학교 도서관학과 조교수

1) W. D. Garvey, *Communication: The Essence of Science* (Oxford: Pergamon Press, 1979), ix.

2) W. Goffman & K. S. Warren, *Scientific Information Systems and the Principle of Selectivity* (N.Y.: Praeger, 1980), pp.1~2.

구집단(invisible college), 학술회의 등 비공식 채널을 통해 전파하여 수신자와의 정보 교환과정을 거쳐서 정보를 확고한 것으로 다듬어가며, 최종적으로 단행본, 학술잡지 등 공식 채널을 통해서 발표한다.<sup>3)</sup> 각 과학자는 개인으로서 과학에 기여하며, 또한 과학은 상호 비판에 의해서 개정되고 명확하게 된다. 과학이란 우리들 한 사람 한 사람이 같은 시대의 사람들과 경쟁하거나 협력하여서 선배들의 업적위에 서서히 쌓아놓은 고도의 相互依存的 활동으로써, 과학자간의 정보 흐름인 학술 커뮤니케이션에 크게 의존하고 있다.<sup>4)</sup> 이에 따라 학술 커뮤니케이션은 과학의 본질(essence)이라고 불리운다.

과학의 발전을 촉진시키기 위해서는 커뮤니케이션과정의 수행을 그 기능으로 하는 메카니즘(mechanism)인 도서관, 온라인 검색 시스템 등 정보 시스템의 효율적인 설계 및 운영이 필수적인데, 이를 위해서는 커뮤니케이션 과정에 대한 충분한 연구가 이루어져야 한다.

커뮤니케이션과정은 한 對象에서 다른 대상으로 정보의 전달을 가져다 주는 사건들의 연속으로, 정보학은 이러한 모든 커뮤니케이션과정과 그 과정과 관련된 정보시스템의 행태(behavior)를 지배하는 일련의 기본적인 규칙 설정에 관계하는 학문이다.<sup>5)</sup> 정보학의 핵심 주제는 정보 교환과정인 커뮤니케이션으로서, 정보학에서 의미하는 정보는 수신자의 지식 구조를 변경시킬 수 있는 송신자에 의해서 구성된 기호의 집합으로 정의할 수 있으며,<sup>6)7)</sup> 이러한 개념 속에는 기록정보뿐만 아니라 구두정보도 포함된다. 따라서, 정보학은 기록 커뮤니케이션(공식 커뮤니케이션)과 구두 커뮤니케이션(비공식 커뮤니케이션)을 모두 다루는 학문 분야라고 할 수 있다.

이와 같은 정보학 연구 영역에서 주제문헌의 書誌를 수량학적으로 분석하여 現象으로서의 정보와 過程으로서의 커뮤니케이션의 행태를 연구하는 분

3) W. D. Garvey, *op. cit.*, pp.165~183.

4) W. Goffman & K. S. Warren, *op. cit.*

5) W. Goffman, "Information Science: Discipline or Disappearance?," *Aslib Proceedings* 22 ; 11(1970) : 589~595.

6) N. J. Belkin & S. E. Robertson, "Information Science and the Phenomenon of Information," *JASIS* 27 ; 4 (1976) : 197~204.

7) B. C. Brooks, "The Foundations of Information Science. Part 1, Philosophical Aspects," *J. of Information Science* 2(1980) : 125~133.

야는 계량서지학(bibliometrics)이란 명칭하에 고유한 연구 영역을 창출해 내고 있으며, 이 분야는 정보학에 관한 하나의 학문 영역으로 다듬어져서 과학적 학문으로서의 정보학의 위치를 굳혀주고 있다.<sup>8)</sup> 계량서지학은 계량사회학(sociometrics), 계량경제학(econometrics) 등과 동일하게 계량정보학(informetrics)으로 불리기도 한다.<sup>9)</sup>

그러나 좀 더 엄밀하게 학문 명칭의 語源을 살펴보면, Bibliometrics 라는 용어에서의 metric 이란 ‘測定’을 의미하는 것이다.<sup>10)</sup> 그러므로 계량서지학은 주제문헌의 서지관계를 측정하는 학문이며, Informetrics 는 정보관계를 측정하는 분야인 것이다. 따라서 공식·비공식 커뮤니케이션과정을 측정하는 학문 분야라고 할 수 있는 계량정보학은 문헌을 통한 공식 커뮤니케이션 과정만을 측정하는 계량서지학보다는 좀 더 광범위한 학문 분야로 정의될 수 있다. 공식 커뮤니케이션의 정적 과정은 물론 동적 과정에 대한 계량서지학적 연구는 활발히 진행되어 오고 있다.<sup>11)12)13)14)15)</sup> 이러한 연구 결과들은 정보시스템의 효율적인 설계 및 운영에 응용되고 있으며, 또한 계량과학사(sciento-metrics)를 연구하는 한가지 수단이 되어오고 있다.

本稿에서는 학술 커뮤니케이션과정의 행태를 기술하고 예측할 수 있는 수리모형들(mathematical models)을 살펴본 후, 이 모형들을 효율적인 정보시스템 설계와 정보서비스를 위한 기초 자료를 얻을 수 있는 주제문헌의 질적 여과 알고리즘(quality filtering algorithm) 설정에 적용해 봄으로써 계량정보학의 기초적인 이론체계를 定立하는 데에 目的을 두고자 한다.

8) T. Saracevic, "Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science," *JASIS* 26 ; 6(1975) : 323.

9) *Ibid.*

10) 柳志星, 計量經濟學原論 (서울: 博英社, 1985), p.1.

11) H. G. Small, "Co-citation Context Analysis and the Structure of Paradigms," *J. of Docu.* 36 ; 3(1980) : 183~196.

12) E. L. Logan & W. M. Shaw, Jr., "An Investigation of the Coauthor Graph," *JASIS* 38 ; 4(1987) : 262~268.

13) W. Goffman, "A Mathematical Method for Analysing the Growth of Scientific Discipline," *J. of ACM* 18 ; 2(1971) : 173~185.

14) W. Goffman & K. S. Warren, *op. cit.*

15) M. L. Pao & L. McCreery, "Bibliometric Application of Markov Chains," *Information Processing & Management* 22 ; 1(1986) : 7~17.

## II. 학술 커뮤니케이션의 수리모형

### A. 정보중복(Information Overlap)

임의의 과학자 전집의 구성원들이 서로들 정보를 전달한다면 구성원간의 정보중복이 존재할 수 있는데,  $t$  시각에 구성원  $X_i$ 에 대한 구성원  $X_j$ 의 정보중복의 측정  $R_{ij}(t)$ 는 조건확률(conditional probability)을 이용하여 다음과 같은 공식으로 구할 수 있다.<sup>16)</sup>

$$R_{ij}(t) = \frac{\bar{X}_i \cap \bar{X}_j}{\bar{X}_i} \quad \dots\dots(1)$$

여기서  $R_{ij}(t)$ =구성원  $X_i$ 가 커뮤니케이션기호인  $\bar{X}_i$ 를 전달한다는 조건하에  $t$  시각에 구성원  $X_j$ 가 커뮤니케이션기호  $\bar{X}_j$ 를 전달하게 될 확률.

$\bar{X}_i$ =구성원  $X_i$ 가 전달한 기호의 수.

$\bar{X}_j$ =구성원  $X_j$ 가 전달한 기호의 수.

$\bar{X}_i \cap \bar{X}_j$ =구성원  $X_i$ 와  $X_j$ 가 공통으로 전달한 기호의 수.

$R_{ij}(t) \neq R_{ji}(t)$ (정보중복이 0인 경우는 제외)

$0 \leq R_{ij}(t) \leq 1$

정보중복의 개념은 구두로 수행되는 비공식 커뮤니케이션에는 물론 문헌을 통한 공식 커뮤니케이션에도 적용시킬 수 있다. 문헌을 통해서 이루어지는 정보중복을 측정할 경우에는 과학자 전집은 한 학문 분야에 관계하는 저자 전집  $N$ 으로 보며, 커뮤니케이션기호의 집합은 저자 전집  $N$ 에 의해서 산출된 문헌의 집합  $N$ 로 볼 수 있다. 따라서 저자 전집  $N$ 의 구성원간의 가장 간단하고 좁은 형태의 정보중복은 저자간의 직접 접촉으로 생산된 공저 문헌(coauthored paper)으로 측정되며, 또한 저자동시인용법(cocitation), 저자동시참조법(coreferencing) 등과 같이 좀 덜 제한된 方法에 의해서 측정

16) W. Goffman & K. S. Warren, *op. cit.*, pp.23~28.

되기도 한다.

주제문헌의 다른 중요한 매개변수들(parameters)에는 논문과 학술잡지가 있는데, 논문 전집의 구성원간의 정보중복은 공통 주제표목, 공통 키워드 등과 같은 의미론적 내용에 기초하거나 또는 동시인용법과 같은 인용관계에 기초하여 측정할 수 있으며, 학술잡지 전집의 구성원간의 정보중복은 동시인용법에 의해서 측정할 수 있다.

한편 한 학문 분야에 관련된 저자·논문·학술잡지 전집의 構成元간의 정보중복의 측정  $R_{ij}(t)$ 를 기초로 하여 다음과 같은 정의들이 얻어진다.<sup>17)</sup>

- (정의 1) 전집이  $N$ 이고 구성원  $X_i$ 와  $X_j$ 가 있다고 가정할 때, 만약  $R_{ij}(t) > T$ 이고  $\phi \leq T \leq 1$ 이면  $X_i$ 는  $X_j$ 와  $t$  시각에 담화한다고( converse) 말해지며, 담화(conversance)는  $X_i \subset X_j$ 로 표시한다.
- (정의 2) 전집  $N$ 의 구성원들의 연속이  $X_i \subset X_{i+1} \subset \dots, X_{j-1} \subset X_j$ 처럼 존재하면  $X_i$ 는  $X_j$ 와  $t$  시각에 커뮤니케이트한다고(communicate) 말해진다. 커뮤니케이션(communication)은  $X_i \subset X_j$ 로 표시하며, 이러한 체인을 커뮤니케이션체인이라고 부른다.
- (정의 3) 만약에  $X_i \subset X_j$ 이고  $X_j \subset X_i$ 이면  $X_i$ 와  $X_j$ 는  $t$  시각에 상호 커뮤니케이트한다고(intercommunicate) 말해지며, 상호 커뮤니케이션(intercommunication)은  $X_i \subset X_j$ 로 표시한다.

위에서 정의한 담화·커뮤니케이션·상호 커뮤니케이션관계를 수학적 용어를 빌어서 표현해 보면, 담화관계는 재귀적(reflexive)이지만 이행적(transitive)·대칭적(symmetric)이지 않으며, 커뮤니케이션관계는 재귀적·이행적이지만 대칭적이지 않기 때문에 前秩序(preorder)를 규정한다. 또 상호 커뮤니케이션관계는 재귀적·이행적·대칭적이기 때문에 전집  $N$ 의 구성원들에 대하여 대등한 관계를 설정한다.<sup>18)</sup>

17) Ibid.

18) 'a는 b에 관계된다'를  $aRb$ 로 나타낸다면 집단  $A$ 의 모든 요소에 대하여  $aRa$ 인 경우 집단  $A$ 내의 관계  $R$ 은 재귀적이라고 하며,  $a$ 와  $b$ 가  $A$ 의 요소이고  $aRb$ 일때  $bRa$ 라하면 집단  $A$ 내의 관계  $R$ 은 대칭적이라고 하며  $a, b, c$ 가 모두  $A$ 의 요소이고  $aRb$ 와  $bRc$ 가 곧  $aRc$ 를 의미한다면 집단  $A$ 내의 관계  $R$ 은 이행적이라고 한다.

주어진 쓰레쉬홀드(threshold)  $T$ 에서, 상호 커뮤니케이션관계는 각 구성원이 그 집단에 있는 적어도 다른 하나의 구성원과 상호 커뮤니케이트하고 있을 때, 구성원의 집합에 의해서 형성되는 이접 상호 커뮤니케이션 집단들(disjoint intercommunication classes)로 전집  $N$ 을 분할시킨다. 이 때 쓰레쉬홀드  $T$ 가 1에 접근할 수록 전집  $N$ 은 각 한 구성원만을 갖는 이접 상호 커뮤니케이션 집단들로의 분할에 접근하며, 반대로  $T$ 가  $\phi$ 에 가까울 수록 전집  $N$ 은 전집  $N$  그 자체가 하나의 상호 커뮤니케이션 집단인 구조로 접근한다.

상호 커뮤니케이션 집단  $X_i$ 에 속한 그 어느 구성원도 상호 커뮤니케이션 집단  $X_j$ 에 속한 그 어느 구성원과 서로 상호 커뮤니케이트하지 않는다 하더라도 그러한 구성원들이 담화하거나 커뮤니케이트한다고 할 수 있다. 따라서 정의 1, 2, 3은 다음의 정의 4, 5로 연장된다.<sup>19)</sup>

(정의 4) 상호 커뮤니케이션 집단  $X_i$ 의 요소인  $a$ 와 상호 커뮤니케이션 집단  $X_j$ 의 요소인  $b$ 가 존재하고  $a$ 가  $b$ 와 담화한다면  $X_i$ 가  $X_j$ 와 담화한다고 말해진다.

(정의 5)  $X_i$ 로부터  $X_j$ 까지 이르는 일련의 집단들이 존재하고 인접한 집단들이 담화한다면  $X_i$ 는  $X_j$ 와 커뮤니케이트한다고 말해진다.

〈표 1〉

참고문헌 코딩표

$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
500	500	500	500	529
521	570	529	521	570
570	573	531	536	583
573	578	536	570	629
578	595	570	573	
583		573	575	
584		575	584	
		583	595	

19) W. Goffman & K. S. Warren, *op. cit.*

〈표 2〉 커뮤니케이션 행렬  $D$

		$D_1$	$D_2$	$D_3$	$D_4$	$D_5$
$D =$	$D_1$	1.00	.57	.57	.71	.29
	$D_2$	.80	1.00	.60	.80	.20
	$D_3$	.50	.38	1.00	.63	.38
	$D_4$	.63	.50	.63	1.00	.13
	$D_5$	.50	.25	.75	.25	1.00

다음으로 문헌을 통한 정보중복을 공식(1)을 이용하여 측정하고, 이를 기초로 하여 위에서 논의한 정의들을 구체적인 예를 들어 설명해 보고자 한다. 만약 5개의 논문들로 이루어진 논문 전집  $N$ 이 있고, 각 논문이 인용한 참고문헌(bibliography)을 숫자로 코딩한 결과가 표 1과 같으며 표 1의 데이터를 정보중복의 측정공식(1)에 대입하여 표 2의 커뮤니케이션행렬(comm-unication matrix)  $D$ 를 얻었다고 가정하자.

표 2에 있는 커뮤니케이션행렬  $D$ 의 제 1행 제 2열에 있는 원소  $D_{12}$ 가 .57이 된 과정을 설명하면 다음과 같다.

$$R_{12}(t) = \frac{\text{논문 1과 2가 공통으로 갖고 있는 참고문헌수}}{\text{논문 1이 갖고 있는 총 참고문헌수}}$$

$$= 4/7 = .57$$

쓰레쉬홀드  $T$ 가 .5인 상태에서 이 특정 논문 전집  $N$ 을 이접 상호 커뮤니케이션 집단들로 분할하기 위해서 먼저 커뮤니케이션행렬  $D$ 의 제 1행의 원소들  $D_{1j}(j=1, 2, 3, 4, 5)$ 를 체크해 보면,  $D_{11}$ 은 對角원소(diagonal element)이기 때문에 제외하고 그 다음 제 2열의 원소  $D_{12}$ 가 .5보다 크기 때문에 일단 선정한 후, 다시 행과 열을 바꿔서 제 2행 제 1열의 원소  $D_{21}$ 을 체크해 보았더니 .5보다 컸다. 따라서, 이 경우는  $D_1$ 이  $D_2$ 와 담화하며 동시에  $D_2$ 가  $D_1$ 과 담화한다고 말해지는데, 이를  $D_1 \text{ I } D_2$ 로 표시해 둔다. 이 과정을 되풀이한 결과  $D_1 \text{ I } D_2$ ,  $D_1 \text{ I } D_4$ ,  $D_3 \text{ I } D_4$ 를 얻게 되었다. 이에 따라 이 논문 전집  $N$ 은 2개의 이접 상호 커뮤니케이션 집단들로 분할되며

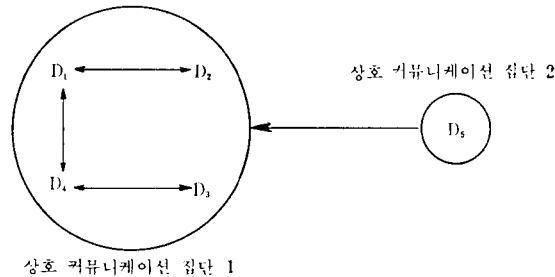


그림 1. 커뮤니케이션 네트워크

논문  $D_1, D_2, D_3, D_4$ 가 모여서 상호 커뮤니케이션 집단 1을 구성하고 논문  $D_5$ 는 집단 2를 이루는데, 이 때 집단 2의 구성원  $D_5$ 가 집단 1의 구성원  $D_3$ 와 담화하고 있기 때문에 상호 커뮤니케이션 집단 2가 집단 1과 담화한다고 말해진다(그림 1 참조).

그림 1에서 담화관계를 체크해 보면  $D_1 \subset D_2, D_1 \subset D_4, D_4 \subset D_3, D_3 \subset D_4, D_4 \subset D_1, D_2 \subset D_1$ 이며, 커뮤니케이션관계는  $D_2 \bar{\subset} D_4, D_2 \bar{\subset} D_3, D_1 \bar{\subset} D_3, D_3 \bar{\subset} D_1, D_3 \bar{\subset} D_2, D_4 \bar{\subset} D_2$ 이다. 또한 상호 커뮤니케이션관계는  $D_2 \bar{\subset} D_4, D_2 \bar{\subset} D_3, D_1 \bar{\subset} D_3$ 가 된다.

위의 정의들(1~5)을 기초로 하여 최종적으로 관련성(relevance)을 정의하면 다음과 같다.<sup>20)</sup>

(정의 6) 주어진 전집  $N$ 과 스펙트럼  $T$ 에서 만약 구성원  $X_i$ 가  $X_j$ 와 커뮤니케이션한다면  $X_i$ 는  $X_j$ 에 관련된다고 말해진다.

결국, 한 학문 분야와 관련된 저자·논문·학술잡지 전집은 관련성에 기초하여 정의된 구조를 갖게 된다.

### B. 종합(Synthesis)

어떤 학문 분야와 관련된 저자·논문·학술잡지 전집의 구성원간의 직접 접촉이나 공저·인용관계에 의한 정보중복의 결과는 구성원을 나타내는 절

20) *Ibid.*



점(node)의 집합과 구성원간의 커뮤니케이션 채널(communication channels)을 나타내는 가지(edge)의 집합으로 구성된 커뮤니케이션 그래프(communication graph)를 형성한다.<sup>21)</sup> 연결 그래프(connected graph)는 그래프 내의 모든 상이한 절점들에 대해서 이 들을 잇는 통로(walk)가 있을 때 형성되며, 그렇지 않은 경우는 구성요소(component)라고 불리우는 최대한으로 연결된 서브그래프(subgraph)의 집합으로 구성된 불연결 그래프(disconnected graph)가 된다.<sup>22)</sup> 이때 서브그래프는 앞에서 논의한 상호 커뮤니케이션 집단과 동일하게 볼 수 있다. 커뮤니케이션 그래프의 각 구성원이 그래프 구조내에서 질서를 생산하는 정도를 연관된 커뮤니케이션 과정에 대한 공헌의 척도로 간주하고, 이에 Brillouin의 정보측정공식을 이용하여 커뮤니케이션 그래프의 각 구성원의 공헌도를 측정할 수 있다.<sup>23)24)25)</sup>

1. Brillouin의 정보측정공식

Brillouin은 S개의 다른 유형(type)으로 나뉘어지는 N개의 기호로 구성된 하나의 메시지에 있는 총정보량(I)을 산출하기 위해서 다음의 공식을 이용한다.<sup>26)</sup>

$$I = K \cdot \ln \frac{N!}{n_1! n_2! \dots n_i! \dots n_s!} \dots \dots \dots (2)$$

여기서 I=N개의 기호로 구성된 메시지에 의해서 전달되는 총정보량을 비트(bit)로 표시함.

$$K = 1/\ln 2$$

$n_i = i$  번째 기호의 출현빈도수

21) W. M. Shaw, Jr., "Statistical Disorder and the Analysis of a Communication-Graph," *JASIS* 34 ; 2(1983) : 146~149.  
 22) *Ibid.*  
 23) M. L. Pao, "Co-authorship as Communication Measure," *Library Research* 2(1981) : 327~338.  
 24) W. M. Shaw, Jr., *op. cit.*  
 25) W. M. Shaw, Jr., "Entropy, Information and Communication," *Proceedings of the ASIS* 16(1979) : 32~37.  
 26) *Ibid.*

Shannon의 엔트로피공식<sup>27)</sup>은 위의 Brillouin의 정보측정공식(2)에서 모든  $n_i$ 가 1보다 클 때  $I/N$ 에 의해서 주어진 기호당 평균정보로부터 끌어내어질 수 있다.<sup>28)</sup> 공식(2)는 통계적 열역학에서 물리계(physical system)의 엔트로피를 측정하는 볼츠만공식과 형태상 일치하며, 또한 공식(2)는  $S$ 개의 다른 종(species)으로 구분되는 個體數가  $N$ 인 동물 전집의 통계적 무질서인 다양성을 측정하기 위해서 사용되기도 한다.<sup>29)</sup> 상대적 다양성  $I_{rel}$ 은  $I_{rel}=I/I_{max}$ 로 주어지며 여기서  $I_{max}=K\ln N!$ 이다.  $N$ 개의 절점이  $S$ 개의 서브그래프에 분포되어 있는 불연결 그래프의 다양성과 상대적 다양성은 함수  $I$ 와  $I_{rel}$ 에 의해서 각각 측정될 수 있다.<sup>30)</sup>

## 2. 응용

비공식 연구집단, 학술회의 등 비공식 채널과 공저·동시인용 등 공식 채널을 통해서 형성된 커뮤니케이션 그래프의 각 구성원의 공헌도는 Brillouin의 정보측정공식을 이용하여 측정할 수 있는데, 구체적인 예로 공저관계에 의해서 형성된 커뮤니케이션 그래프의 한 구성원의 공헌도를 측정해 보고자 한다. 4개의 논문들과 각 논문에 관여한 저자들이 다음과 같다고 하자.

논 문	공 저 자
A	1, 2, 3
B	1, 4
C	1, 5
D	6, 7

여기서 모든 공저쌍(coauthor pair)의 집합은 아래와 같이 구성된다.

27) Shannon의 엔트로피공식

$$H = -\sum_{i=1}^n P_i \cdot \ln P_i$$

28) 김현희, "엔트로피와 커뮤니케이션", 圖書館學論集, 창간호 (명지대학교 도서관학과, 1987): 9~24.

29) E. C. Pielou, *Mathematical Ecology* (N.Y.: Wiley-Interscience, 1977), pp.299~307.

30) W. M. Shaw, Jr., "Statistical Disorder and the Analysis of a Communication-Graph," pp.146~149.

그룹 1 : (1, 2), (1, 3), (2, 3), (1, 4), (1, 5)

그룹 2 : (6, 7)

이로부터 그림 2와 같이 2개의 서브그래프들로 이루어진 불연결 그래프가 형성된다.

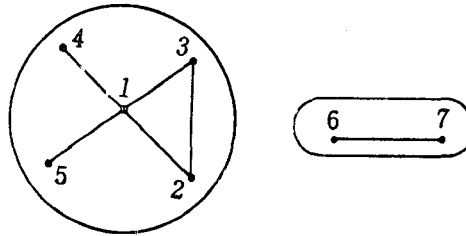


그림 2. 커뮤니케이션 그래프

그림 2에서 저자 1의 공헌도를 측정하기 위해서 먼저 모든 구성원이 존재했을 때 커뮤니케이션 그래프의 다양성  $I_i$ 를 측정해야 하는데, 공식(2)를 이용하여 구하면 다음과 같다.

$$I_i = K \cdot \ln \frac{N!}{n_1! n_2!}$$

$$= \frac{1}{\ln 2} \cdot \ln \frac{7!}{5! 2!} = 4.3923 \text{ bits}$$

다양성  $I_i$ 를 구한 다음에는 측정하고자 하는 구성원이 그래프에서 제거된 후의 그래프의 다양성  $I_f$ 를 구해야 하는데, 이 경우는 그림 2에서 저자 1을 제거하면 그림 3과 같이 4개의 서브그래프들로 구성된 불연결 그래프가 형성된다.

그림 3에 있는 불연결 그래프의 다양성  $I_f$ 를 공식(2)를 이용하여 구하면 다음과 같다.

$$I_f = \frac{1}{\ln 2} \cdot \ln \frac{6!}{1! 1! 2! 2!} = 7.4919 \text{ bits}$$

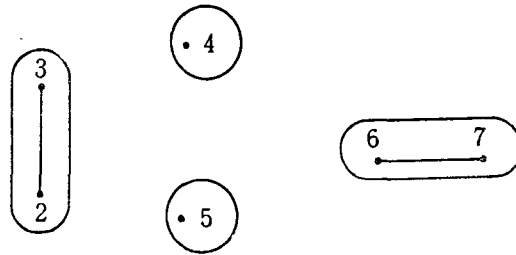


그림 3. 커뮤니케이션 그래프

다양성 측정 결과가  $I_f - I_i > \phi$ 이 되어 한 저자가 존재함으로써 질서를 만든다면 이를 종합적(synthetic)이라고 할 수 있으며,  $I_f - I_i < \phi$ 이 되어 무질서를 만든다면 분석적(analytic)이라고 할 수 있다.<sup>31)</sup> 또한  $I_f - I_i = \phi$ 인 경우는 한 저자가 커뮤니케이션 그래프의 구조에 아무런 영향을 미치지 못한다. 앞에 든 예에서는  $I_f - I_i = 7.4919 - 4.3923 = 3.0996$  (bits) 이므로 저자 1은 이 그래프에서 종합적이라고 할 수 있다.

Brillouin의 정보추정공식을 이용하여 많은 다른 사람들과 개인적인 접촉을 가졌거나 혹은 공저했거나 또는 많은 다른 사람들에 의해서 인용되어진 해당 주제 분야의 핵심 종합자들(key synthesizers)을 식별할 수 있는데, 만약 이러한 핵심 종합자들이 그래프에서 제거되면 남아있는 구성원간의 커뮤니케이션이 극심하게 방해받게 되는 것이다.<sup>32)</sup>

### C. 역 동(Dynamics)

이제까지는 고정된 視點에서의 구성원간의 정보중복을 살펴보았다. 그러나 커뮤니케이션 그래프의 구성원간의 정보중복은 시간에 따라서 변하는 시간의존적 현상이기 때문에, 그래프의 구조는 해당 학문 분야의 활동의 강도에 따라서 팽창되거나 붕괴된다. 다시 말해서 강력한 발전 시기에는 그 분야가 하나의 팽창된 연결 그래프에 의해서 표현되어지며, 활동이 부진한 시

31) M. L. Pao, "Co-authorship and Productivity," *Proceedings of the ASIS* 17(1980) : 279~281.

32) W. Goffman & K. S. Warren, *op. cit.*, pp.31~32.

기에는 핵심 종합자들의 이탈에 의해서 야기된 축소 상태에서 수축된 그래프들로 표현되어질 것이다.<sup>33)</sup> 이에 따라 커뮤니케이션 그래프의 전체 역동(gross dynamics)을 고려해야 한다.

커뮤니케이션 그래프의 구성원간에 思想이 전파되는 동적 과정을 기술하기 위해서, 결정적 모형(deterministic model)인 로지스트生長式(logistic equation)과 확률적 모형(stochastic model)인 마코브연쇄(Markov chain)를 이용하고자 한다.

### 1. 결정적 모형

#### ① 로지스트生長式

생태학에서 일반적으로 어떤 공간을 점유하고 있는 동일한 개체의 집합체인 개체군은 적합한 환경에서는 기하급수적 生長을 하다가, 개체군의 密度가 높아지면 먹이의 부족, 생활 공간의 협소, 수명의 단축 등 환경저항(environmental resistance)에 부딪히기 때문에 증가율이 감소되어 기하급수적 生長이 무한히 계속되지는 못한다. 이와 같이 증식율이 감소되는 개체군의 生長型은 로지스트곡선(logistic curve)을 나타내며(그림 4 참조) 로지스

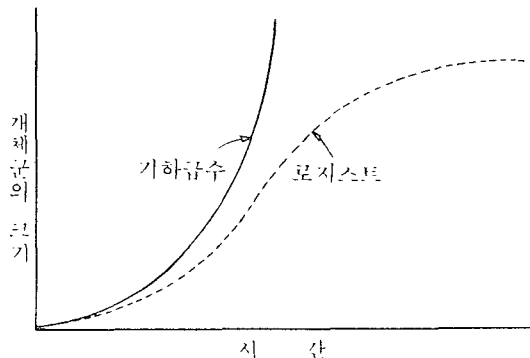


그림 4. 개체군의 두 生長型(이상환경에서는 기하급수적 生長이, 제한환경에서는 로지스트 生長이 이루어진다)

33) *Ibid.*, p.14.

트곡선을 나타내는 식은 아래의 공식(3)과 같다.<sup>34)35)36)</sup>

$$dX/dt = aX \cdot \left( \frac{K-X}{K} \right) \quad \dots\dots(3)$$

여기서  $a$ =증가율

$X$ =개체수의 크기

$K$ =上限漸近線으로 주어진 환경에서의 개체군의 收容能力(carrying capacity)을 나타낸다.

위의 로지스트성장식(3)은 결국 일정한 증가율을 갖는 때의 증가율에 이 증가율이 실제로 나타나는 정도를 곱한 것이라는 뜻이며, 개체수의 크기인  $X$ 가 상한 접근선인  $K$ 와 같게 되면  $\frac{K-X}{K} = \phi$ 가 되기 때문에 생장은 정지되게 된다. 위의 공식(3)을 변형하면 다음과 같다.

$$\begin{aligned} dX/dt &= aX \cdot \left( \frac{K-X}{K} \right) \\ &= aX \cdot \left( 1 - \frac{X}{K} \right) \\ &= \left( a - \frac{a}{K} \cdot X \right) X \\ &= (a - bX) X \quad \dots\dots(4) \end{aligned}$$

여기서  $b = a/K$ 로 생물의 생존능력을 제한하는 환경요인인 限定要因

전염의학에서 사용하는 용어를 빌면, 어떤 사상에 흥미를 갖고 있는 독자를 미감염자(susceptible), 어떤 사상에 관한 논문을 발표하고 있는 저자를 감염자(infective), 그리고 그 사상에 관한 논문을 더 이상 발표하지 않는 사람을 퇴거자(removal)라고 할 수 있다.<sup>37)</sup>

먼저 어떤 특정 분야에 흥미를 갖는 사람, 즉 미감염자수는 무제한일 수 없으며 따라서 일정기간 동안 고정되어 있다고 가정할 수 있다. 둘째, 앞에

34) 崔基哲 等, 一般生物學 (서울: 鄉文社, 1986), pp. 372~382.

35) 玄在善, 一般生態學 (서울: 集賢社, 1986), pp. 206~215.

36) E. C. Pielou, *op. cit.*, pp. 20~23.

37) W. Goffman, *op. cit.*

서 언급했듯이 과학은 고도의 상호 의존적인 활동이기 때문에 특별한 경우를 제외하고는 대부분 미감염자 집단에서 감염자가 발생한다고 볼 수 있다.<sup>38)</sup> 따라서, 각 학문 분야의 감염자수의 변화는 미감염자수가 제한되었기 때문에 제한 환경하에서 개체군의 성장과정을 묘사하는 로지스트성장식(4)로 기술할 수 있다고 가정할 수 있다.

한 학문 분야의 감염자들의 성장과정을 로지스트성장식(4)를 이용하여 기술하는 방법은 공식(4)에서 전집을 떠나는 퇴거자들을 고려함으로써 쉽게 일반화되어진다. 한편 일반 과정은 다음의 미분방정식(5)에 의해서 기술되어질 수 있다.

$$\begin{aligned} dX/dt &= (a-bX)X - cX \\ &= (a-c)X - bX^2 \\ &= c_1X - c_2X^2 \end{aligned} \quad \dots\dots(5)$$

여기서  $dX/dt$  = 감염자수의 변화율

$a$  = 증가율

$b (=c_2)$  = 환경저항 즉 미감염자수의 제한으로 인한 限定要因

$c$  = 퇴거율

$c_1$  = 증가율 - 퇴거율

② 응용

공식 커뮤니케이션의 동적 과정을 기술·예측하는데에 위의 공식(5)를 적용하기 위해서 다음과 같은 方法을 이용할 수 있다.

- (1) 일정한 기간 동안의 한 주제 분야에 대한 완전한 서지 데이터를 수집한다.
- (2) 샘플 서지에서 첫 논문을 발표한 해에 그 저자는 감염자가 되고, 서지에서 마지막 논문의 발간일 1년후에 퇴거자가 된다.
- (3) 샘플 서지를 5년 간격으로 나누어 해당 기간내의 논문을 발표한 감

38) 미감염자가 감염자로 변형되기 위해서 필요한 시간간격인 잠복기가 필요로 한다.

염자수( $X$ )와 감염자수의 변화율( $\Delta X$ )을 측정하여  $X$ 축에는 기준년도를  $Y$ 축에는 각 기준년도에서의 감염자수를 표시하여 그래프로 그려준다.

- (4) 공식(5)의 계수  $c_1$  과  $c_2$  를 구하기 위해서 공식(5)를 방정식  $Y=c_1X - c_2X^2$  로 놓고 각 기준년도에서의 감염자수의 변화율( $\Delta X$ )에 관한 데이터를 이용하여 최소자승법(least square method)으로 2원 1차 연립방정식을 구한 다음, 소거법(elimination method)을 사용하여 2원 1차 연립방정식을 풀어서 계수  $c_1$  과  $c_2$  를 구한다.
- (5) 공식(5)에다 제 4 단계에서 구한 계수  $c_1$  과  $c_2$  의 값을 대입한 후, 외삽법(extrapolation)에 의해서  $dX/dt$  의 값을 구하면 그래프 곡선의 미래 양상을 예측할 수 있다.

구체적인 예로 1852~1962년 사이의 schistosomiasis 분야의 샘플 서지들

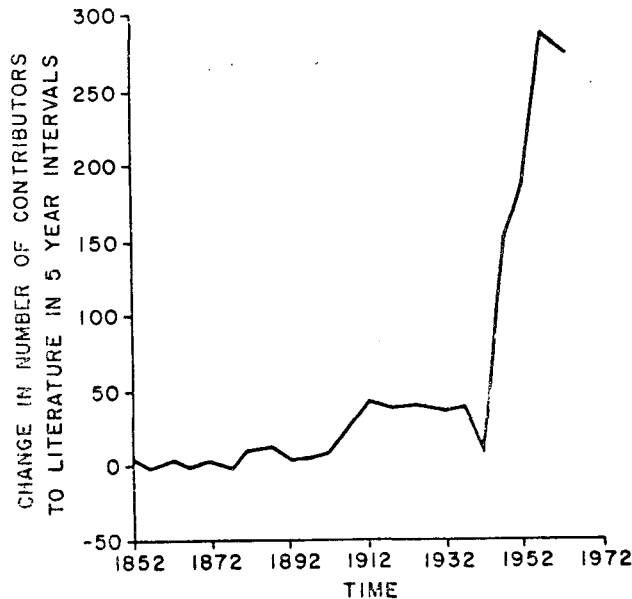


그림 5. 감염자수의 변화율



〈표 3〉 감염자수 · 감염자수의 변화율의 관측치와 예측치

시 간	감염자수의 관측치 ( ) 속은 예측치	감염자수의 변화율의 관측치 ( ) 속은 예측치
1898~1902	46(46)	—
1903~1907	72(65)	26(19)
1908~1912	115(98)	43(33)
1913~1917	155(139)	40(41)
1918~1922	195(186)	40(47)
1923~1927	234(231)	39(45)
1928~1932	272(271)	38(40)
1933~1937	312(301)	40(30)
1938~1942	323(315)	11(14)
1943~1947	434(439)	111(116)
1948~1952	620(641)	186(202)
1953~1957	902(898)	282(257)
1958~1962	1,174(1,180)	272(282)
1963~1967	(1,456)	(276)
1968~1972	(1,697)	(241)
1973~1977	(1,871)	(174)
1978~1982*	(1,950)	(79)
1983~1987	(1,901)	(-49)
1988~1992	(1,695)	(-206)
1993~1997	(1,302)	(-393)

분석하여<sup>39)</sup> 공식 커뮤니케이션의 동적 과정을 로지스트생장식으로 기술하고 예측할 수 있다는 가설을 검증하고자 한다. 표 3은 1898~1997년 사이의 감염자수와 그 변화율의 관측치와 예측치를 표시한 것이며, 그림 5는 1852~1962년 사이의 감염자수의 변화율을 그래프로 그린 것이다.

로지스트생장식(5)로 과학자간의 커뮤니케이션의 동적과정을 기술할 수 있다는 가설을 검증하기 위해서 먼저 그림 5에서 성장과정(epidemic process)에 있는 완전한 순환을 체크해야 한다. 왜냐하면, 로지스트생장식의 계수들이 각 순환에 따라서 달라질 것이 기대되어지기 때문이다. 그림 5를 체크해 보면 1902~1942년 사이에 한번의 순환이 있었으며, 1942년부터 또 다

39) W. Goffman & K. S. Warren, *op. cit.*

른 순환이 시작되어 진행되고 있음을 알 수 있다.

이제, 1942년부터 시작되는 순환과정을 로지스트생장식으로 기술할 수 있는지를 알아보기 위해서 1942~1962년 사이 감염자수의 변화율에 대한 관측치를 이용하여 먼저 최소자승법으로 2원 1차 연립방정식을 구한 다음 Gauss-Jordan 소거법으로 2원 1차 연립방정식을 풀어서 계수  $c_1$  과  $c_2$  를 구한다. 이 값들을 공식(5)에 대입하면  $dX/dt=(26.2354-.6070X)X$ 를 얻게 되며, 이 공식의  $X$ 에 5, 10, 15, ..., 50, 55를 대입하여 1947~1997년 사이의 감염자수의 변화율( $dX/dt$ )에 대한 예측치를 얻는다.<sup>40)</sup> 표 3에서 1947~1997년 사이의 감염자수의 변화율에 대한 관측치와 예측치를 비교해 본 결과 관측치와 예측치가 거의 비슷하였다.

다음에는 1902~1942년 사이의 순환도 로지스트생장식으로 기술할 수 있는지를 검증해 보기 위해서 위와 똑같은 절차로 1902~1942년 사이의 감염자수의 변화율에 대한 예측치를 구하여 관측치와 비교해 본 결과 별 차이가 없었다. 따라서 공식 커뮤니케이션의 동적 과정을 로지스트생장식으로 기술하고 예측할 수 있다는 가설은 검증된 셈이다.

그림 6은 1942~1962년 사이의 감염자수의 관측치를 그래프로 표시한 후

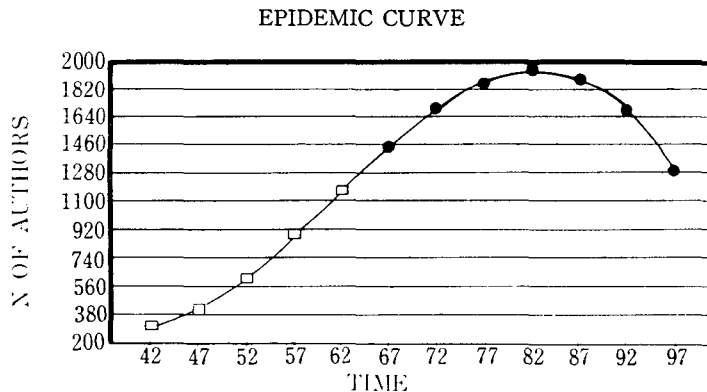


그림 6. 감염자수의 변화

40) 부록 1 (컴퓨터 프로그램).

1967~1997 년 사이의 감염자수의 예측치를 계속 그래프에 그려준 것이다. schistosomiasis 분야는 1982 년에 최고점에 다다를 것으로 예측된다.

## 2. 확률적 모형

### ① 마코브연쇄

우연적으로 일어나는 현상의  $t$ 이 시간의 함수로서 나타내어지는 경우가 있다. 이를테면 어떤 주어진 기간동안 나날의 쌀값의 변동을 연구한다고 할 때,  $t_i(t_i=0, 1, \dots)$ 째 날의 쌀값  $X(t_i)$ 와 액체의 표면에서 멋대로 浮遊하는 두 粒子가 있을 때 시간구간  $[a, b]_{\text{sec}}$ 에 속하는 순간  $t_{\text{sec}}$ 에 있어서의 두 입자 사이의 거리 등이 그 예이다.<sup>41)</sup>

첫번째 예의 현상은 조사의 時點  $t_i$ 에 대응하는 확률변수(쌀값)의列  $\{X(t_i)\}$ 에 의해서 설명되는 경우이다. 이 때  $t_0, t_1, t_2, \dots, t_n$ 은 離散的이며, 이와 같이 시점  $t_i$ 가 이산적인 경우의 확률변수의列  $\{X(t_i)\}$ 를 離散·時 확률과정(discrete-time stochastic process)이라고 한다. 이 때  $\{X(t_i)\}$ 의 각 항의 값을 상태(state)라고 하고,  $X(t_0)$ 의 값을 초기상태(initial state),  $X(t_n)(n=1, 2, 3, \dots)$ 의 값을  $n$  번째 상태라고 한다. 그러나 두번째 경우의 현상은 시각  $t$ 가 구간  $[a, b]$  안에서 연속적으로 변하는 것이므로 첫번째 경우와 같은  $X(t)$ 의列로 나타낼 수 없다. 이 때 확률변수  $X(t)$ 의 집합  $\{X(t) | t \in [a, b]\}$ 를 連續·時 확률과정(continuous-time stochastic process)이라고 한다. 이산·시 확률과정과 연속·시 확률과정을 함께 가리켜 확률과정이라고 한다.<sup>42)</sup>

이산·시 확률과정에서  $n+1$  번째 단계 상태의 확률이  $n$  번째 단계 상태의 영향만 받고 그 이전 단계 상태의 영향을 받지 않을 때, 이 확률과정을 마코브연쇄라고 한다.<sup>43)</sup> 따라서 마코브연쇄는 이산·시 확률과정의 특별한 경우이다. 예를 들어서 설명해 보기로 하자. 현재 세 개의 암 연구기관  $a, b, c$ 가 있고 이 분야에 종사하는 네 명의 연구원  $A, B, C, D$ 가 있다고 가정

41) 鄭英鎭, 수리통계학; 확률과 통계 (서울: 喜重堂, 1987), pp. 323~340.

42) 전계서.

43) 최희찬·이대영·신철재, 이공계 학생을 위한 응용수학 (서울: 한신문화사, 1981), pp. 509~515.

해 보자. 또한 각 연구원이 반드시 세 연구기관중 한 기관에 고용되며, 연초마다 각 연구자는 자유롭게 현재 속해 있는 기관에서 다른 기관으로 옮겨가거나 혹은 그대로 같은 기관에 머무를 수 있다고 가정해 보자. 여기서 각 연구원이 선택할 수 있는 기관을 상태(state)라고 하고 상태의 집합  $\{a, b, c\}$ 를 상태공간(state space)이라고 할 수 있다.<sup>44)</sup>

보기로 든 이 특정 경우를 상태공간이 유한집합인 유한 마코브연쇄(finite Markov chain)로 기술하기 위해서는 다음과 같은 요건(要件)들이 충족되어야만 한다.

- ① 이산·시 확률과정이어야 한다.
- ② 상태공간이 유한 집합이어야 한다.
- ③ 각 상태의 예측된 확률은 바로 이전 단계의 상태에만 의존해야 한다.

이 경우 연구자들의 기관(상태) 사이의 이동이 제한된 간격, 즉 1년에 한 번씩 발생하기 때문에 이산·시 확률과정이며, 상태 공간이  $\{a, b, c\}$ 로 유한 집합이며, 또한 앞 연구처럼 급격하게 발전하는 분야에서 연구자들의 현재의 성과가 연구기관에 고용되는데 유일한 고려의 대상으로 생각될 수 있기 때문에, 각 연구자의 다음 단계의 상태는 현 상태에 의해서 결정된다고 할 수 있다.

위의 예를 마코브연쇄로 완전히 기술하기 위해서는 앞에서 논의한 세가지 요건들이 충족되어지는 것 외에 각 상태의 초기확률(initial probability)과 상태  $i$ 에서  $j$ 로 바뀔 추이조건확률(conditional probabilities of transition)  $P_{ij}$ 를 알아야 한다.<sup>45)</sup> 각 상태의 초기확률은 첫 해에 각 연구기관에 의해 고용된 연구자수를 총연구자수로 나눈 값으로 얻어지는데, 표 4에서 보면 첫 해에는 기관  $a$ 에 연구원  $A$ 가 기관  $b$ 에 연구원  $C, D$ 가 그리고 기관  $c$ 에 연구원  $B$ 가 고용되어 있었기 때문에  $a, b, c$ 에 대한 초기확률은  $1/4, 2/4, 1/4$ 이다. 다음으로 추이조건확률  $P_{ij}(1)$ 은 한 단계를 거친 후 이전의 상태가  $i$ 였을 때 상태  $j$ 로 이동할 확률이다. 여기서는 총 9개의 추이조건확률

44) M. L. Pao & L. McCreery, *op. cit.*

45) W. Goffman & K. S. Warren, *op. cit.*, pp.87~91.

〈표 4〉 4년 동안의 연구자들의 이동과정

States	<i>a</i>	<i>b</i>	<i>c</i>
	$A_1$	$C_1$	$B_1$
	$C_2$	$D_1$	$D_2$
	$A_2$	$B_2$	$D_3$
	$A_3$	$B_3$	$D_4$
	$C_3$	$B_4$	$A_4^*$
	$C_4$		

\*이 표에서 연구자 *A*의 경우, 기관 *a*에서 3년 동안( $A_1, A_2, A_3$ ) 근무한 후 4년째( $A_4$ )에 기관 *C*로 옮겼음을 알 수 있다.

〈표 5〉 정 방 행 렬

		<i>a</i>	<i>b</i>	<i>c</i>
<i>P</i> =	<i>a</i>	$P_{aa}$	$P_{ab}$	$P_{ac}$
	<i>b</i>	$P_{ba}$	$P_{bb}$	$P_{bc}$
	<i>c</i>	$P_{ca}$	$P_{cb}$	$P_{cc}$

이 있는데 이들을 표 5와 같이 행의 수와 열의 수가 같은 正方形行列(square matrices)로 정렬하는 것이 편리하다.

표 5에 표시한 정방행렬의 제 1행  $P_{aa}, P_{ab}, P_{ac}$ 는 *a*라는 상태에서 각각 *a, b, c*라는 상태로 옮겨가는 확률을 보여주고 있다. 상태 *a*에서 한번 이동할 때 *a, b, c* 중 어느 한 상태로 옮겨가게 되므로 반드시  $P_{aa} + P_{ab} + P_{ac} = 1$ 이 성립된다.<sup>46)</sup> 이 경우에  $P_{aa}$ 의 조건 확률은 상태 *a*에서 상태 *a*로 이동할 확률이기 때문에, 기관 *a*에 계속 2년동안 연구자가 머무는 경우를 계산함으로써 얻어질 수 있다. 이 경우는  $A_1 \rightarrow A_2, A_2 \rightarrow A_3, C_2 \rightarrow C_3, C_3 \rightarrow C_4$ 로 총 4번이며 기관 *a*에서 *b*로 옮겨간 사람은 아무도 없으므로  $P_{ab}$ 는  $\phi$ 이다. 기관 *a*에서 *c*로 옮긴 경우는  $A_3 \rightarrow A_4$ 로 단 한번 있다. 따라서 상태 *a*에서 *a, b, c*로의 추이조건확률은  $4/5 (=0.80), \phi, 1/5 (=0.20)$ 이 된다(표 6의 제 1행 참조).

46) 최희찬·이대영·신철재, 상계서.

〈표 6〉 추이행렬  $P$ 

$P =$	$a$	0.80	0	0.20
	$b$	0.25	0.50	0.25
	$c$	0	0.33	0.67

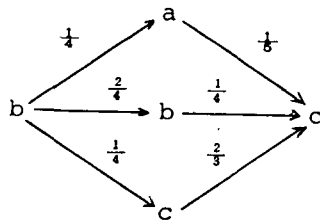
〈표 7〉 추이행렬  $P^2$ 

$P^2 =$	$a$	.640	.067	.293
	$b$	.325	.333	.342
	$c$	.083	.389	.528

〈표 8〉 고정확률행렬  $L$ 

	$a$	$b$	$c$	
$L = P^2 =$	$a$	.3335	.2662	.3996
	$b$	.3332	.2664	.3997
	$c$	.3328	.2667	.3997

한 상태에서 다른 상태로의 추이는 한 단계 이상에서 실행되어질 수 있다. 구체적으로, 두 단계를 거치는 상태  $b$ 에서 상태  $c$ 로의 추이는 세개의 가능한 통로들로 수행되어질 수 있는데, 이를 열거하면 다음과 같다.



합 계

$$1/20 = 0.05$$

$$1/8 = 0.125$$

$$1/6 = 0.0167$$

---


$$0.342$$

따라서, 두 단계를 거쳐서 상태  $b$ 에서 상태  $c$ 로 이동할 조건확률  $P_{bc}(2)$ 은 0.342가 된다. 같은 방법으로  $P_{aa}(2)$ ,  $P_{ab}(2)$ ,  $P_{ac}(2)$ ,  $P_{ba}(2)$ ,  $P_{bb}(2)$ ,  $P_{ca}(2)$ ,  $P_{cb}(2)$ ,  $P_{cc}(2)$ 를 구해 행렬을 만들면 표 7이 된다. 주어진 추이행

렬  $P$ 를 제공해도  $P(2)$ 를 얻는다. 즉  $P(2)=P^2$ 이다. 따라서,  $n$  단계를 거쳐 상태  $i$ 에서 상태  $i$ 로 추이할 조건확률  $P_{ii}(n)$ 의 계산은 추이행렬  $P$ 를  $n$ 승한  $P^n$ 과 동일하다.<sup>47)</sup>

만약 추이행렬  $P$ 를  $n$ 승한  $P^n$ 의 모든 원소가 양수일 때  $P$ 를 정규 마코브연쇄(regular Markov chain)라고 하며, 정규 마코브연쇄에서는 각 상태가 자기 다른 상태로부터 접근되어질 수 있다. 따라서, 추이행렬의 대각원소  $P_{ii}$ 가 만약 1 이라면( $P_{ii}=1$ ) 자기 상태에서 이탈하지 못하는 경우로 정규 마코브연쇄가 아니라 하겠다.

마코브연쇄이론에 의하면 만약 추이행렬  $P$ 가 정규 마코브연쇄이면 여러 단계를 거친 후 각 행에 있는 확률값들의 분포가 똑같은 고정 확률행렬  $L$ 을 얻게 되는데, 일단 고정 확률행렬  $L$ 에 이르면 행렬의 확률값들은 시간에 따라 더 이상 변화하지 않는다.<sup>48)</sup> 이 경우는 5 단계를 거친 후 고정 확률행렬  $L$ 에 이르게 되는데 이 최종행렬에서 몇 년 후에 연구자들의 33%, 27%, 47%가 기관  $a, b, c$ 에서 각각 발견될 수 있음을 추측할 수 있다(표 8 참조).<sup>49)</sup> 초기 확률과 고정 확률을 비교하면 기관  $c$ 로 연구자들이 몰려올 것을 예측할 수 있는데, 이는 기관  $c$ 의 고정 확률이 40%로써 초기 확률인 25%보다 훨씬 높기 때문이다.

## ② 응용

마코브연쇄이론은 물리학·사회과학 분야 등에서 많이 응용되고 있는데, 도서관학·정보학 분야에 적용시키면 먼저 마코브연쇄는 한 주제 분야의 하위주체(subarea)간의 저자의 이동을 기술하고 예측하는 데 응용된다.<sup>50)51)</sup> 예를 들어서 정보학이라는 한 주제 분야가 있고 이 분야의 하위주체들에는 정보검색, 도서관 자동화, 계량정보학 등이 있는데, 이 분야를 연구하는 학자들이 발표하는 논문의 토픽이 정보학의 한 하위주체에서 다른 하위주체로

47) 전계서.

48) 전계서.

49) 부록 2 (컴퓨터 프로그램).

50) W. Goffman & K. S. Warren, *op. cit.*, pp.87~91.

51) M. L. Pao & L. McCreery, *op. cit.*

변화해 가는 과정을 유한 마코브연쇄를 이용하여 기술하고 예측할 수 있다. 그러나 유한 마코브연쇄로 어떤 동적 과정을 기술하기 위해서는 앞에서 언급한 세 가지 요건들이 충족되어야 하므로, 이에 입각하고 있는지를 살펴 볼 필요가 있다. 첫째, 한 저자가 두 개의 논문을 동시에 출판하지 않고 보통 각 출판간에는 시간 차이가 있기 때문에 이산·시 확률과정이라 할 것이며 둘째, 한 주제 분야는 고정된 하위주제의 집합으로 구성되기 때문에 각 하위주제를 상태로 놓으면 상태공간은 유한 집합이 되며 셋째, 한 하위주제에서 다른 하위주제로의 저자의 이동은 저자의 현재의 관심 분야에 의존하기 때문에 저자의 다음 단계의 상태는 현 상태가 결정한다고 볼 수 있다. 따라서 이 경우는 세가지 요건들을 충족시킨다고 볼 수 있으며, 또한 한 저자가 한 하위주제에서 다른 하위주제로 자유롭게 옮겨갈 수가 있기 때문에 그 이동 과정을 정규 마코브연쇄로 기술할 수 있다. 이에 따라서 최종적으로 고정 확률행렬  $L$ 을 얻을 수 있게 된다. 행렬  $L$ 을 통해서 각 하위주제에 대한 과학자들의 미래의 흥미도에 대한 확률치를 구하게 되는데, 이러한 자료는 도서관 이용자들의 현재, 그리고 미래의 관심 분야를 파악할 수 있게 하므로 효율적인 정보 서비스 및 문헌 구입을 위한 기초 자료가 될 수 있다.

유한 마코브연쇄를 이용하여 한 주제 분야의 하위주제간의 저자의 이동을 기술하고 예측하는데는 다음과 같은 절차가 필요하다.

- (1) 한 주제 분야와 그 주제 분야를 구성하는 하위주제들을 결정한 후 해당 분야에 관한 서지 데이터를 수집한다.
- (2) 각 하위주제(상태)의 초기 확률값을 구하기 위해서 샘플 서지를 분석하여 각 저자가 처음으로 쓴 논문들을 수집하고 주제를 분석한 후 각 논문에 적합한 하위주제를 할당한다. 각 하위주제에 할당된 저자수를 샘플 서지의 총저자수로 나눈 값을 각 하위주제의 초기 확률값으로 이용한다.
- (3) 하위주제  $i$ 에서  $j$ 로 바뀔 추이조건확률  $P_{ij}$ 를 구하기 위해서 먼저 각 저자가 쓴 논문들을 날짜순으로 분류한 후 하위주제가 할당되지 않은



논문들은 그 주제를 분석하여 적합한 하위주제를 할당한다. 이러한 자료를 이용하여 하위주제간의 저자의 이동 과정을 열거한 표(표 4 참조)를 작성한 후 추이행렬  $P$ 의 각 원소를 구한다. 이 때 한 주제에 대하여 한 문헌만을 발표한 저자는 다른 하위주제로 추이를 할 수 없기 때문에 제외된다.

- (4) 전 단계(3)에서 얻어진 추이행렬  $P$ 를 이용하여 컴퓨터 프로그램에 의해서 고정 확률행렬  $L$ 을 얻는다.

또한 유한 마코브연쇄는 저자들의 인용 행태(citation behavior)를 기술하고 예측하는데에도 응용된다.<sup>52)</sup> 이 때는 각 저자가 상태가 되며, 서지 데이터를 통해서 그들의 인용 행태를 분석하여 고정 확률행렬  $L$ 을 얻음으로써 미래에 각 저자가 저자 전집의 모든 저자에 의해서 인용되어질 고정된 확률치들을 구하게 되는데, 이러한 자료들은 저자들을 질적으로 여과하는 데에 이용될 수 있다.

유한 마코브연쇄를 이용하여 한 주제 분야의 저자 전집에서 각 저자의 인용 행태를 기술하고 예측하기 위해서는 다음과 같은 절차가 필요하다.

- (1) 한 주제 분야에 관한 일정 기간 동안의 모든 이용할 수 있는 비평 논문들(review articles)을 수집한다.
- (2) 각 저자의 질적 평가치(quality number)는 각 저자를 인용한 비평 논문의 수에 의해 결정되는데, 질적 평가치가 임의의 기준치를 넘는 저자 그룹을 식별한다.
- (3) 인용 색인지인 SCI 나 SSCI 를 이용하거나 또는 수작업으로 데이터를 분석하여 식별된 저자 그룹内の 각 저자쌍(each pair of authors)에 대한 일정 기간 동안의 인용 빈도수를 측정한다. 각 저자쌍에 대한 인용 빈도수를 이용하여 저자  $j$ 가  $i$ 에 의해서 인용되어질 확률인 추이 조건확률  $P_{ij}$ 를 아래의 공식을 통해서 구한다.

52) Seung-hee Sohn, "A Quality Filtering Method for Biomedical Literature," (Ph. D Diss., Case Western Reserve Univ., 1985), pp.1~23.

$$P_{ij} = \frac{\text{저자 } j \text{가 저자 } i \text{에 의해서 인용된 횟수}}{\text{저자 } j \text{가 저자 그룹내에 있는 모든 저자에 의해서 인용된 횟수}}$$

- (4) 전 단계(3)에서 얻어진 추이행렬  $P$ 를 이용하여 컴퓨터 프로그램에 의해서 고정 확률행렬  $L$ 을 얻는다.

### Ⅲ. 주제문헌의 질적 여과 알고리즘

앞장에서 기술한 학술커뮤니케이션의 수리모형들을 이용하여 성장 과정에 있는 주제 분야의 핵심 저자군을 식별하는 방법은 아래와 같다. 이 때 식별된 핵심 저자군은 유망한 과학자들을 선정하여 재정적으로 지원하는 과학자 후원기관의 정책 결정에 이용될 수 있으며, 또한 주제 분야별 데이터 베이스에 핵심 저자군으로 접근하여 관련 문헌을 검색하는 정보 검색 분야에 응용될 수 있다.

- (1) 일정 기간 동안의 관심있는 주제 분야들의 서지 데이터를 수집·분석하여 감염자수의 변화율을 그래프로 표시하고 그래프 곡선의 미래 양상을 예측하여 앞의 그래프에 연속으로 그려준다.
- (2) 감염자수의 변화율( $\Delta X$ )이 양수인 성장 과정에 있거나 성장이 예측되는 주제 분야를 선정하여 이 분야에 관한 모든 이용할 수 있는 비평 논문들을 수집한다.
- (3) 비평 논문에 게재된 모든 참고문헌의 저자, 제목, 학술잡지, 출판년도, 인용한 비평 논문들을 표시하여 컴퓨터에 입력시킨다.
- (4) 인용된 각 저자의 질적 평가치는 각 저자를 인용한 비평 논문의 수에 의해 결정되는데 질적 평가치가 임의의 기준치를 넘는 저자 그룹을 식별한다.
- (5) 선택된 저자 그룹을 공저 관계에 기초하여 커뮤니케이션 그래프로 표시하고 Brillouin의 정보측정공식을 이용하여 각 저자의 종합력을 측정정한 후 종합력이 높은 순서대로 등급을 매긴다. 공저 관계가 거의 없

는 주제 분야의 저자 그룹은 유한 마코브연쇄를 이용하여 각 저자를 인용 빈도수가 높게 예측되는 순서대로 등급을 매긴다.

다음은 학술 커뮤니케이션의 수리모형을 이용하여 한 주제 분야의 핵심 논문군과 핵심 학술잡지군을 체계적으로 식별하는 방법을 설명한다. 이 때 식별된 주제문헌의 핵심 논문군은 그 분야에 흥미를 갖고 있는 과학자들을 위한 정보 시스템을 설계하는데 이용될 수 있으며, 또한 핵심 잡지군은 정보 시스템의 학술잡지 구입에 대한 기초 자료로 이용될 수 있다.

- (1) 해당 주제 분야에 관한 모든 이용할 수 있는 비평 논문들을 수집한다.
- (2) 비평 논문들에 게재된 모든 참고문헌의 저자, 제목, 학술잡지, 출판년도, 인용한 비평 논문들을 표시하여 컴퓨터에 입력시킨다.
- (3) 인용된 각 논문이나 학술잡지의 질적 평가치는 각 논문이나 학술잡지를 인용한 비평 논문의 수에 의해 결정되는데, 질적 평가치가 임의의 기준치를 넘는 논문 그룹과 학술잡지 그룹을 식별한다.
- (4) 선택된 논문 그룹과 학술잡지 그룹은 동시인용법에 기초하여 커뮤니케이션 그래프로 표시하고 Brillouin의 정보측정공식을 이용하여 각 논문이나 학술잡지의 종합력을 측정한 후 종합력이 높은 순서대로 등급을 매긴다. 또한 학술잡지 그룹은 유한 마코브연쇄를 이용하여 각 학술잡지를 인용 빈도수가 높게 예측되는 순서대로 등급을 매길 수도 있다.

위의 주제문헌의 질적 여과 알고리즘에서 일반 논문 대신에 비평 논문을 이용하였는데 그 이유로는 첫째, 한 주제에 대한 완전한 서지 데이터를 수집하는 것은 많은 시간과 노력이 들며 둘째, 저자의 질(quality)과 비평 논문에 의한 인용과의 관계는 저자의 질과 일반 논문에 의한 인용과의 관계보다는 좀 더 강한 상관성이 있는 것으로 알려졌다.<sup>53)</sup>

53) W. Goffman & K. S. Warren, *op. cit.*, p.148.

## V. 결 론

과학자간의 정보 교환활동을 다루는 학술 커뮤니케이션에 대한 연구는 정보학의 핵심적 연구 영역으로 과학 및 기술의 발전을 도모하는데 그 궁극적인 目的이 있다. 본 논문에서는 학술 커뮤니케이션의 정적 과정과 동적 과정을 측정할 수 있는 수리모형들의 이론체계 및 응용 영역을 살펴본 후, 이 모형들을 효율적인 정보 시스템 설계와 정보 서비스를 위한 기초자료를 얻을 수 있는 주제문헌의 질적 여과 알고리즘 설정에 적용해 봄으로써 학술 커뮤니케이션을 수량학적으로 다루는 학문 분야인 계량정보학의 기초적인 이론체계를 定立해 보고자 했다.

논의된 주제문헌의 질적 여과 알고리즘은 실지 정보 검색 분야에 적용시켜 그 효율성을 테스트해 볼 수 있을 것이다. 또한 본 연구에서는 주로 공식 커뮤니케이션의 수리모형들을 다루었는데, 비공식 채널을 통해서 전달되는 정보의 행태를 분석할 수 있는 수리모형에 대해서도 적극적인 연구가 있어야 할 것이다.

공식·비공식 채널을 통해서 수행되는 학술 커뮤니케이션을 관련성에 기초하여 구조적으로 분석할 수 있는 수리모형들을 설정하고 이 모형들을 효율적인 정보 제공이라는 궁극적인 목표를 갖는 정보 시스템의 설계 및 운영에 응용함으로써, 정보의 폭발로 인해서 과학자와 대중간에 있어서뿐만 아니라 과학자간에 있어서까지도 매우 심각해진 커뮤니케이션의 문제를 해결해 나갈 수 있을 것이다.

과학자가 정보를 찾고 사용하고 전파하는 학술 커뮤니케이션양식에 대한 연구는 수량학적 측면은 물론 심리학적 측면 등 다각적인 관점에서 충분히 연구되어야 할 것이다.

## 참 고 문 헌

- 김현희. “엔트로피와 커뮤니케이션.” 圖書館學論集, 창간호 (명지대학교 도서관학  
과, 1987) : 9~24.
- 柳志星. 計量經濟學原論 (서울 : 博英社, 1985)
- 鄭英鎭. 수리통계학 ; 확률과 통계 (서울 : 喜重堂, 1987)
- 崔基哲 等. 一般生物學 (서울 : 鄉文社, 1986)
- 최희찬 · 이대영 · 신철재. 이공계 학생을 위한 응용 수학 (서울 : 한신문화사, 1981)
- 玄在善. 一般生態學 (서울 : 集賢社, 1986)
- Belkin, N. J. & Robertson, S. E. “Information Science and the Phenomenon of  
Information.” *JASIS* 27 ; 4(1976) : 197~204.
- Brooks, B. C. “The Foundations of Information Science. Part 1, Philosophical  
Aspects.” *J. of Information Science* 2(1980) : 125~133.
- Fairthorne, R. A. “Empirical Hyperbolic Distributions(Bradford-Zipf-Mandelb-  
ort) for Bibliometric Description and Prediction.” *J. of Docu.* 25 ; 4(1969) :  
319~343.
- Garvey, W. D. *Communication: The Essence of Science* (Oxford: Pergamon  
Press, 1979)
- Goffman, W. “A General Theory of Communication.” In: T. Saracevic ed.  
*Introduction to Information Science* (N.Y.: Bowker, 1970)
- \_\_\_\_\_ “Information Science: Discipline or Disappearance?” *Aslib Procee-  
dings* 22 ; 11(1970) : 589~595.
- \_\_\_\_\_ “A Mathematical Method for Analysing the Growth of Scientific  
Discipline.” *J. of ACM* 18 ; 2(1971) : 173~185.
- Goffman, W. & Warren, K. S. *Scientific Information Systems and the Principle  
of Selectivity* (N.Y.: Praeger, 1980)
- Logan, E. L. & Shaw, Jr., W. M. “An Investigation of the Coauthor Graph.”  
*JASIS* 38 ; 4(1987) : 262~268.
- Pao, M. L. “Co-authorship as Communication Measure.” *Library Research* 2  
(1981) : 327~338.
- \_\_\_\_\_ “Co-authorship and Productivity.” *Proceedings of the ASIS* 17  
(1980) : 279~281.
- Pao, M. L. McCreery, L. “Bibliometric Application of Markov Chains.” *Infor-  
mation Processing & Management* 22 ; 1(1986) : 7-17.

Pielou, E. C. *Mathematical Ecology* (N.Y.: Wiley-Interscience, 1977)

Prichard, A. "Statistical Bibliography or Bibliometrics?" *J. of Docu.* 25 ; 4(1969)  
: 348~349.

Saracevic, T. "Relevance: a Review of and a Framework for the Thinking on  
the Notion in Information Science." *JASIS* 26 ; 6(1975) : 321~343.

Seung-hee Sohn. "A *Quality Filtering Method for Biomedical Literature.*"  
(Ph. D Diss., Case Western Reserve Univ., 1985)

Shaw, Jr., W. M. "Entropy, Information and Communication." *Proceedings of  
the ASIS* 16(1979) : 32~37.

\_\_\_\_\_ "Statistical Disorder and the Analysis of a Communication-  
Graph." *JASIS* 34 ; 2(1983) : 146~149.

Small, H. G. "Co-citation Context Analysis and the Structure of Paradigms."  
*J. of Docu.* 36 ; 3(1980) : 183~196.

# A Study on the Quantitative Analysis of Scientific Communication

Hyun-hee Kim

## Abstract

Scientific communication is an information exchange activity between scientists. Scientific communication is carried out in a variety of informal and formal ways. Basically, informal communication takes place by word of mouth, whereas formal communication occurs via the written word. Science is a highly interdependent activity in which each scientist builds upon the work of colleagues past and present. Consequently, science depends heavily on scientific communication.

In this study, three mathematical models, namely Brillouin measure, logistic equation, and Markov chain are examined. These models provide one with a means of describing and predicting the behavior of scientific communication process. These mathematical models can be applied to construct quality filtering algorithms for subject literature which identify synthesized elements (authors, papers, and journals). Each suggests a different type of application. Quality filtering for authors can be useful to funding agencies in terms of identifying individuals doing the best work in a given area or subarea. Quality filtering with respect to papers can be useful in constructing information retrieval and dissemination

systems for the community of scientists interested in the field.

The quality filtering of journals can be a basis for the establishment of small quality libraries based on local interests in a variety of situations, ranging from the collection of an individual scientist or physician to research centers to developing countries.

The objective of this study is to establish the theoretical framework for informetrics which is defined as the quantitative analysis of scientific communication, by investigating mathematical models of scientific communication.



## 부록 1

]LIST

```
50 REM THE LEAST-SQUARE METHOD
100 DIM X(20), Y(20), F(20,20), W(20), A(20,20), B(20), C(20)
110 M=2
120 N=5
130 FOR I=1 TO N
140 READ X(I), Y(I), W(I)
150 NEXT I
160 DATA 0, 0, 1
161 DATA 5, 111, 1, 10, 186, 1.2, 15, 282, 1.2, 20, 272, 1.2
180 FOR I=1 TO N
190 F(I, 1)=X(I)
200 F(I, 2)=X(I) * X(I)
215 NEXT I
220 GOSUB 1000
221 FOR I=1 TO M
222 FOR J=1 TO M
223 PRINT "A("I", "J")=" ; A(I, J)
224 NEXT J, I
226 GOSUB 1200
227 FOR I=1 TO M
228 PRINT "B("I")=" ; B(I)
229 NEXT I
320 END
1000 REM SUBROUTINE
```

```
1010 L=M
1020 FOR I=1 TO M
1030 FOR J=1 TO L
1050 A(I,J)=0
1055 FOR K=1 TO N
1060 A(I,J)=A(I,J)+F(K,I) * W(K) * F(K,J)
1070 NEXT K,J,I
1200 REM SUBROUTINE
1220 FOR I=1 TO M
1240 B(I)=0
1250 FOR K=1 TO N
1260 B(I)=B(I)+F(K,I) * W(K) * Y(K)
1270 NEXT K,I
1300 RETURN

]NEW
]LOAD GAUSS
]LIST

10 REM SOLUTION OF SIMULTANEOUS EQUATION BY GAUSS-
   JORDAN METHOD.
20 DIM A (15,16), B(15,15)
70 READ N
80 PRINT "NUMBER OF EQUATIONS TO BE SOLVED=" ; N
90 M=N+1
100 FOR I=1 TO N
110 FOR J=1 TO M
120 READ A(I,J)
130 PRINT SPC (2) A(I,J) ;
```

```
140 NEXT J, I
205 K=M-1
210 IF K=1 THEN 370
220 JJ=1
230 BIG=ABS (A(I, 1))
240 FOR I=2 TO K
250 AA=ABS (A(I, 1))
260 S=SGN (BIG-AA)+2
270 ON S GOTO 280, 300, 300
280 BIG=AA
290 JJ=I
300 NEXT I
310 SS=SGN(JJ-1)+2
320 ON SS GOTO 330, 370, 330
330 FOR J=1 TO M
340 TEMP=A(JJ, J)
350 A(JJ, J)=A(1, J)
355 A(1, J)=TEMP
360 NEXT J
370 FOR J=2 TO M
380 FOR I=2 TO N
390 B(I-1, J-1)=A(I, J) - A(1, J) * A(I, 1)/A(1, 1)
400 NEXT I, J
410 FOR J=2 TO M
420 B(N, J-1)=A(1, J)/A(1, 1)
430 NEXT J
440 M=M-1
450 FOR J=1 TO M
```

```
460 FOR I=1 TO N
470 A(I,J)=B(I,J)
480 NEXT I,J
490 Z=SGN(M-1)+2
500 ON Z GOTO 205,510,205
510 FOR I=1 TO N
520 PRINT SPC(4)"C(" ; I ; ")" SPC(4)A(I,1)
530 NEXT I
540 DATA 2,880,14800,14130,14800,263500,228630
550 END
```

```
]LIST
```

```
10 S=323
20 FOR I=1 TO 11
30 READ X
40 Y=(26.2354-.6070 * X) * X
50 PRINT Y
60 S=S+Y
70 PRINT S
80 NEXT I
90 DATA 5,10,15,20,25,30,35,40,45,50,55
```

```
]RUN
```

```
116.002
439.002
201.654
640.656
256.956
```

897.612  
281.908  
1179.52  
276.51  
1456.03  
240.762  
1696.792  
174.664  
1871.456  
78.2159999  
1949.672  
-48.5820003  
1901.09  
-205.73  
1695.36  
-393.228  
1302.132

## 부록 2

]LIST

```
10 DIM A(3,3)
60 FOR I=1 TO 3
70 FOR J=1 TO 3
80 READ A(I,J)
90 PRINT A(I,J) ; " ";
100 NEXT J,I
```

```
120 LET S=0
130 FOR M=1 TO 3
140 FOR J=1 TO 3
150 FOR I=1 TO 3
160 S=S+A(I,J) * A(M,I)
170 NEXT I
180 A=M : B=J : C(A,B)=S : S=0
190 NEXT J,M
200 FOR I=1 TO 3
210 FOR J=1 TO 3
220 PRINT C(I,J)
230 A(I,J)=C(I,J)
240 NEXT J
250 PRINT
260 NEXT I
270 DATA 0.80,0,0.20,0.25,0.50,0.25,0,0.33,0.67
280 INPUT N$
283 IF N$="Y" THEN GOTO 120
290 END
```