Shannon의 정보이론과 문헌정보

정 영 미*

〈차 례〉

1. 커뮤니케이션

인간의 일상생활은 커뮤니케이션의 연속이며 커뮤니케이션은 말, 글, 손짓, 몸짓, 음악, 그림, 무용, 신호등, 경적 등의 다양한 도구를 통해 수행되고 있다.

커뮤니케이션의 대상이 되는 것은 정보이다. 예를 들어 지나는 택시를 보고 손을 드는 것은 서달라는 정보를 전달하기 위한 것이고 네거리의 파란 신호등은 가도 좋다는 정보를 전달하기 위한 것이다. 여기에서 말하는 정보는 전달되는 모든 형태의 메시지의 내용으로서 반드시 의사결정작업을 위한 것에 국한되는 것은 아니다.

Weaver[1]가 1949년에 발표한 그의 논문에서 지적했듯이, 커뮤니케이션은 정보전달의 정확성과 관련된 기술적인 면, 정보내용의 올바른 해석과 관련된 어의적인 면, 그리고 전달된 정보가 정보 입수자에게 미치는 영향과 관련된 효과적인 면의 세가지 측면에서 고려되어야 한다.

기술적인 면은 단순히 정보를 얼마나 정확히 전달할 것인가 하는 문제

* 연세대학교 도서관학과 조교수

1) Weaver, Warren, "The Mathematics of Communication," *Scientific American*, no. 7 (July 1969) pp. 11~15.

이다. 글을 통한 커뮤니케이션에서 가능한 한 오자 발생을 감소시키기 위해서 몇번씩 교정을 보는 것이나, 말을 통한 커뮤니케이션에서 말소리를 높인다거나 발음을 정확히 한다거나 외부에서 잡음이 섞여들지 않도록 창문을 닫는다든가 하는 방법을 써서 가능한 한 말이 정확히 전달되도록 하는 것등이 여기에 관련된다. 전보나 텔렉스에서와 같이 코우드를 사용하여 정보를 전송하는 경우에는 코우딩 및 코우드 해석의 정확성, 채널의 용량, 송·수신 방법 등이 정보전달의 정확성을 좌우하게 된다.

어의적인 문제는 기술적인 문제만큼 단순하지 않다. 커뮤니케이션 참여자들 간에 공통의 인식이나 의식구조, 공통의 관습 내지는 문화, 공통의 관심사, 공통의 언어 등이 존재하지 않고서는 어의적인 면에서의 완전한 커뮤니케이션은 달성하기가 힘들다.

다음의 대화를 보자.

A : 굉장히 비싸대요.

B : 네?

A가 한 말의 내용이 A와 B의 경험에 있어서 공통되는 부분에 관련된 것이 아니고는 B는 A가 한 말의 내용을 전혀 이해할 수가 없다. 여기에서는 정보가 기술적인 면에서는 정확히 전달되었다고 해도 어의적인 면에서의 커뮤니케이션은 이루어지지 않고 있다. 역사학 전공의 학생들에게 자동제어에 관한 강의를 해보자. 학생들이 강의를 한마디도 놓치지 않고 정확히 들었다고 해도 고급수학이나 공학에 대한 지식이 전혀 없는 상태에서는 실질적인 커뮤니케이션이 달성되기 힘들다.

효과적인 면은 어의적인 면과 밀접한 관계가 있다. 한 엄마가 네살난 아들에게 "밥먹기 전에는 손을 꼭 닦아야 한다"고 항상 말한다고 하자. 아들은 말의 내용을 완전히 이해하고 고개를 끄덕이지만 실천은 안하려고 든다면 효과적인 면에서 볼 때 엄마와 아들 사이에 성공적인 커뮤니케이션이 수행되고 있다고 볼 수 없다. 손을 들어 빈 택시를 잡으려고 할 때 운전사가 손짓을 보고도 그냥 지나가 버린다면 이 때에도 효과적인 면에

서는 커뮤니케이션이 이루어지지 않고 있는 것이다.

A , B , C 세 사람이 똑같은 내용의 메시지를 전달받았다고 해도 각자가 이 메시지에 대해 갖는 이해 정도에 따라 실제로 얻는 정보의 양은 달라진다. 즉, 기술적인 면에서는 똑같은 내용의 정보가 전달되었다고 해도 개개인의 이해의 범주나 경험이 다르기 때문에 실제로 A , B , C가 얻은 정보의 양은 다르게 되는 것이다. 예를 들어 물리학에 관해 전혀 모르는 사람에게 물리학 분야의 박사학위 논문이 전달하는 정보의 양은 극히 적을 것이다. 그러나 물리학에 관해 어느정도 지식이 있는 사람에게는 훨씬 많은 양의 정보가 전달될 것이며 반면에 그 논문 내용을 이미 잘 알고 있는 사람에게는 아무런 정보도 전달되지 않을 것이다. 입수된 정보가 효과적인 단계에서 정보입수자의 결정작업에 하등의 도움을 줄 수 없을때에는 실제로 전달된 정보의 양은 거의 없다고 볼 수 밖에 없다. 기술적인 단계에서 전달되는 정보의 양은 정보의 내용과는 무관한 것이며 어의적인 단계와 효과적인 단계에 가서야 실제로 정보 입수자의 지식체계에서 새롭다고 판단되고 또 이해되는 내용만이 전달되는 것이다. 즉, 어느 정도 예상되면서 동시에 새로운 내용만이 정보로서의 가치를 갖는다.

2. Shannon의 정보이론

2.1 Shannon[2]의 정보이론

정보이론은 텔레커뮤니케이션과 더불어 발전되어 왔고 이것은 여러 커뮤니케이션 시스템을 비교하는 평가 기준을 제공한다. 전신시스템에서의 신호전송과 관련된 Nyquist와 Hartley의 이론을 바탕으로 하여 Shannon의 정보이론이 정립되었다. Shannon의 정보이론에서 텔레커뮤니케이션 시스템은 채널의 신호 전송율에 의해 평가된다. 전송율의 측정에는 엔트로피가 이용되며 정보원의 엔트로피가 H이고 전송채널의 용량이 C인 경우 적

2) Shannon, C. E. and Warren Weaver. *The Mathematical Theory of Communication*, University of Illinois Press, 1949.

절한 코우딩에 의해서 이 시스팀은 C/H에 가까운 전송율을 얻을 수 있다는 것이 이 정보이론의 기본정리이다.

Shannon의 정보이론에서 사용한 정보란 말의 개념은 일반적인 커뮤니케이션 이론에서 말하는 정보의 개념과는 다르다. 공학적으로 해석되는 정보는 메시지 내용이 뜻하는 바와는 무관하다. 텔레커뮤니케이션 채널을 통해 전송되거나 컴퓨터 내에서 처리되는 메시지는 자연어로 된 것이 아니라 코우딩 된 형태의 것이다. 즉, CAT이란 단어는 1100001100000100 10101과 같이 코우딩된다. 결국 정보이론에서 말하는 정보량은 실질적으로는 메시지를 이진수 (0과1) 로 코우딩 할 때 필요한 최소한의 자리수에 해당한다. 즉, 전달 대상이 되는 전체 메시지의 숫자와 각 메시지의 발생확률에 따라 가장 효율적인 코우드를 설계할 수 있는데 이 때 코우드의 평균 길이가 메시지의 평균 정보량이 된다.

Morse의 전신부호는 영어의 26개 알파벳을 · (단음)과 — (장음)으로 코우딩해 준 것이다. 각 글자의 발생빈도에 따라 빈번하게 사용되는 글자에는 짧은 코우드를 주고 사용빈도가 낮은 글자에는 긴 코우드를 주므로써 전송율을 높여주고 있다. 만일 알파벳 26자의 발생확률이 다 똑같다면 코우드의 평균 길이는 $\log_2 26$으로 각 글자당 평균 약4.8bits가 요구된다. 따라서 실제 코우드의 길이는 5자리가 되어야 하나 Morse부호에서는 E,T와 같은 발생빈도가 가장 높은 글자에는 한자리 코우드, 그 다음 발생빈도를 갖는 글자에는 두자리 코우드 등으로 코우딩 함으로써 글자당 평균 3.2 bits가 사용되게 하였다. 즉, Morse코우드에서 각 글자가 갖는 평균 정보량은 3.2 bits가 되는 것이다.

2.2 불확실성 (Uncertainty)

앞에서는 정보이론에서의 정보량이 실질적으로는 메시지를 이진수로 코우딩 할 때 필요한 평균 자리수임을 설명하였다. 오늘날 정보이론이 텔레커뮤니케이션 공학에서 뿐만 아니라 다른 학문분야에서도 응용성을 갖는 것은 정보량의 개념이 공학적인 차원을 넘어서서 적용되어지기 때문이다.

정보이론에서의 정보량은 개념적으로는 전달될 메시지에 대한 불확실성의 정도를 나타낸다. 즉, 메시지 (또는 기호) 가 갖는 평균 정보량은 정보원이 생산해 낼 수 있는 다양한 메시지들로 부터 하나의 메시지를 선택할 때에 부여되는 선택의 자유를 물량화한 것이다.

어느 메시지가 선택될 것인가에 대한 불확실성의 정도는 엔트로피로 표현되며 메시지에 대한 불확실성은 메시지가 전달됨으로써 얻은 정보에 의해 해소된다. 따라서 메시지가 전달하는 정보량은 엔트로피로 측정된다. 즉, 정보는 불확실성 또는 엔트로피를 감소시키는 도구이다.

정보원이 전달할 수 있는 메시지가 "예"와 "아니오"의 두가지 뿐인 경우를 생각해 보자. 이때에 정보입수자는 이미 정보원이 전달할 수 있는 메시지에 대해 50%는 알고 있는 셈이다. 만일 정보원이 전달가능한 메시지가 열개이며 선택확률이 동일한 경우에는 전달될 메시지에 대한 불확실성은 훨씬 커진다. 즉, 선택의 대상이 되는 메시지의 수가 많아질 수록 불확실성이 커지는 것이다.

현대는 정보사회로서 정보의 폭발 내지는 홍수라는 말로 특징지워진다. 정보가 과거에 비해 엄청나게 증가하고 있음을 시사하는데 이것은 우리가 필요로 하는 정보를 막대한 양의 정보원으로 부터 선택해야 함을 의미한다. 특정한 정보를 선택하고자 할 때 우리에게 주어지는 선택의 자유는 점점 커지고 선택될 정보에 대한 불확실성 또한 계속 증가하는 것이다.

2.3 엔트로피(Entropy)

텔레커뮤니케이션에서는 메시지의 엔트로피가 적을 수록 전송속도가 빨라지며 따라서 전송에 드는 경비가 낮아진다. 결혼, 생일, 입학등을 축하하기 위해 축하전문을 보낼 때 이미 만들어져 있는 문례전문을 사용하는 것이 자유롭게 보통문을 만들어 사용하는 것보다 가격이 저렴하다. 문례전문의 경우에는 선택대상이 되는 메시지의 수가 제한되어 있어서 보통문에 비해 엔트로피가 훨씬 적어지기 때문이다.

엔트로피의 산출공식은 다음과 같다.

$$H = \sum_{i=1}^{n} P_i \log_2 P_i$$

H는 n개의 메시지가 갖는 평균 정보량으로 P_i는 i번째 메시지가 선택될 확률을 나타낸다. 이때 엔트로피 H는 각 메시지의 선택확률이 동일할 때, 즉 $P_1 = P_2 = \cdots = P_n$일 때 최대치를 갖는다. <표 1> 에는 졸업을 축하하는 뜻으로 보낼 수 있는 다섯개의 전문약호와 문례가 나와있다. 이때 각 메시지가 선택될 확률이 똑같다면 P_i는 각각 $\frac{1}{5}$이 된다. 따라서 엔트로피 H는 $-5\ (-\frac{1}{5} \cdot \log_2 5)$로 약 2.2의 값을 갖는다.

<표 1>

전문약호	문 례	확률 1	확률 2
조 고	1. 졸업을 축하합니다.	$\frac{1}{5}$	0
조 노	2. 영광된 졸업을 축하하며 앞날의 영광을 빕니다.	$\frac{1}{5}$	0
조 도	3. 형설의 공을 치하하며 더 큰 영광을 빕니다.	$\frac{1}{5}$	0
조 로	4. 교문을 나서는 벗이여, 꿈은 뜻을 성취하소서	$\frac{1}{5}$	$\frac{1}{2}$
조 모	5. 영광의 졸업을 축하한다.	$\frac{1}{5}$	$\frac{1}{2}$

즉, 위의 전문들을 이진수로 코우딩 할 때 00,01,10,11,100의 다섯가지 코우드를 사용할 수 있는데 각 코우드를 구성하는 이진수의 평균숫자는 2.2로서 위 식에서 산출된 엔트로피의 값과 같다. 위에서 각 전문이 선택되어질 확률이 같지 않은 경우에는 엔트로피는 적어진다. 만일 친구의 졸업을 축하하기 위해서라면 전문1, 2, 3이 선택될 확률은 0이고 전문 4와5가 선택될 확률이 $\frac{1}{2}$씩이 된다. 이 경우에 엔트로피 H는 $\log_2 2$로서 1이 된다. 선택대상이 될 수 있는 메시지가 한개라면 이미 선택될 메시지에 대해 알고 있기 때문에 불확실성은 존재하지 않으며 엔트로피는 0이 된다.

엔트로피는 정보원에 대해 갖고 있는 사전지식에 의해 영향을 받는다. 일기예보의 경우 만일 오늘 날씨가 흐리고 무덥다면 내일은 비가 온다는

예보가 내릴 확률이 높다. 이것은 날씨가 흐리고 무더운 날에는 비가 오는 일이 많다는 사실을 이미 알고 있기 때문이다. 정보원에 대한 불확실성은 상당히 감소된 상태로서 엔트로피는 적어지는 것이다.

지금까지는 메시지가 전달된 후에도 정보입수자가 갖는 불확실성은 고려하지 않고 메시지가 전달되기 전에 갖는 불확실성만을 고려하여 정보량을 측정하였다. 즉, 메시지가 전달되고 나서 불확실성이 완전히 제거되는 경우로 정보원이 갖는 엔트로피가 메시지가 전달하는 평균 정보량으로 측정되었다. 그러나 실제로 입수된 정보의 양은 메시지를 전달받기 전과 전달받은 후에 정보원에 대해 갖는 불확실성의 차이로써 측정되어야 한다. 즉, 메시지가 전달한 정보의 양은 메시지가 전달되기전의 엔트로피와 전달된 후의 엔트로피의 차이이며 아래의 공식으로 표현된다.

$$T = H(Q \mid X) - H(Q \mid X')$$

메시지를 받기 전의 엔트로피 H는 잘 정의된 의문 Q와 Q에 대해 메시지를 전달받기 전에 갖고 있는 지식상태 X에 의해 결정되며 메시지를 받은 후의 엔트로피 H는 메시지를 전달받은 후에 갖게될 지식상태 X'에 의해 새로이 결정된다. 다시 말해 지식상태 X에 따라 의문 Q에 대해 가능한 여러 해답에 각각 확률을 부여함으로써 H가 측정된다. 여러 해답중 하나의 메시지가 전해졌을 때 이 전달된 메시지에 의해 Q에 대한 지식상태 X는 X'로 변하게 되고 X'에 의해서 가능한 나머지 해답들에 새로운 확률을 부여하게 된다. 따라서 새로운 값의 엔트로피가 산출되며 실제로 전달된 정보의 양은 두 엔트로피의 차이가 되는 것이다.

예를 들어 전화번호의 마지막 숫자를 기억할 수 없는 경우를 생각해 보자. 가능한 숫자는 0에서 9까지의 10개 숫자로서 10개의 해답을 갖고 있는데 첫번째 숫자인 1을 돌리기 전에 각 숫자에 주어지는 확률은 각각 $\frac{1}{10}$이 된다. 그러나 첫번째 숫자를 돌리고 나서는 확률의 배정이 달라진다. 1이 옳은 번호가 아닌 경우 1에 부여되는 확률은 0이되고 나머지 아홉 숫자에 각각 $\frac{1}{9}$의 확률이 부여되는 것이다. 즉, 1을 돌리기 전의 지

식상태 X와 돌린 후의 지식상태 X'는 다르므로 가능한 해답에 주어지는 확률은 달라지는 것이고 이때 1을 돌려 봄으로 해서 얻은 정보의 양은 두 엔트로피의 차이 만큼인 것이다.[3)]

2.4 잉여정보 (Redundancy)와 잡음 (Noise)

앞에서 각 메시지의 발생확률이 똑같은 경우에 엔트로피는 최대치를 갖는다고 하였다. 실제로는 일기예보의 예에서와 같이 특정한 메시지의 선택확률은 정보원에 대한 사전의 지식상태에 의해 달라지므로 실제 엔트로피는 최대 엔트로피에 미달하는 경우가 많다. 실제 엔트로피를 최대 엔트로피로 나눈 값을 상대 엔트로피라고 하며 1에서 상대 엔트로피를 뺀 값이 잉여정보에 해당한다.

잉여정보는 텔레커뮤니케이션시스템에서 중요한 역할을 한다. 잉여정보에 대한 설명은 글자를 선택하여 메시지를 구성하는 경우에 명확해진다. 메시지의 잉여정보가 50%라는 말은 메시지의 반 정도가 생략되어도 메시지의 내용은 여전히 이해할 수 있음을 의미한다. 즉, 메시지를 구성하기 위해 필요한 글자나 단어의 반은 선택의 자유가 주어지지만 나머지는 언어의 구조적인 특성에 의하여 통제를 받는다는 것이다.

가을은 아름다운 계절이다.

위의 문장에서 가을이란 단어 다음에 올 단어는 가을이 명사이기 때문에 품사에 있어서 제한을 받게 되며 따라서 선택대상이 될 단어의 수는 한정된다. 발음상의 제약도 선택확률에 영향을 미친다. 즉, 가을 다음에 조사가 오되 는이나 가가 올 확률은 0이다. 즉, 메시지를 구성할 글자의 선택확률은 어의적인 면은 제외하더라도 글자자체의 순수한 발생빈도, 구문상의 법칙, 발음상의 법칙등에 의해 결정된다고 볼 수 있다.

---

3) 정영미. "정보 이론과 문헌정보 검색," **정보관리연구**, v. 11, no. 3 (1978. 6) p. 56,

언어에 있어서의 잉여정보는 메시지의 전달과정에서 발생하는 철자상의 오류를 쉽게 수정해 준다. 위 문장에서 가을은으로 쓰여지지 않고 가을음으로 쓰여졌다고 해도 독자는 곧 은이 음으로 잘못 쓰여졌다는 사실을 알게 된다.

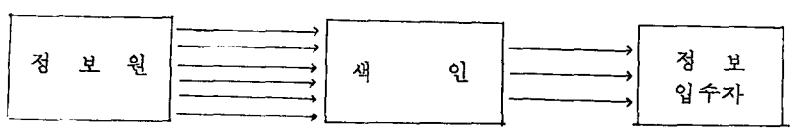
텔레커뮤니케이션에서 신호의 전송이 완벽하게 이루어졌을 경우 메시지의 엔트로피가 H(X)이면 전달받은 정보량도 H(X) 만큼이 된다. 그러나 실제로는 전송과정에서 불필요한 잡음이 섞여든다든가 오류가 발생한다든가 하여 커뮤니케이션에 장애가 생기게 된다. 이런 경우에는 장애로 인해 전송되는 메시지에 대한 불확실성이 높아지고 따라서 전달받은 정보량은 H(X) 에 못미치게 된다. 이와같이 전달과정에서 추가되는 불확실성의 크기를 H(X') 라고 한다면 실제로 전달된 정보량은 H(X)-H(X') 가 된다. 이것이 정보이론의 두번째 기본정리이다. 전송과정에서 발생하는 오류를 자동적으로 수정할 수 있도록 하기 위해서는 코우드가 어느정도 잉여정보를 포함하게끔 설계되어야 하는 것이다.

3. 색인과 엔트로피

3.1 색인 이론

색인 (Index) 의 어원은 라틴어 indicare로서 이 단어는 가르킨다, 지시한다는 의미를 갖는다. 정보를 원하는 사람에게 그 정보의 위치를 지시해 주는 도구가 색인이다. 또한 색인은 방대한 양의 정보로 부터 원하는 정보만을 걸러내어 주는 여과기의 구실을 한다. <그림 1>에서와 같이 정보원과 정보입수자 사이에 위치하여 전달되는 정보를 선별하여 주는 장치 가 되는 것이다.

<그림 1>



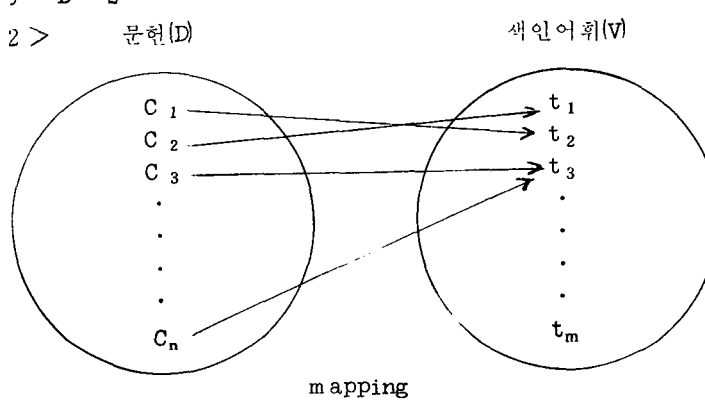
문헌을 색인한다는 것은 문헌이 포함하고 있는 중요한 개념들을 적절한 코우드 즉, 색인어로 변환시켜주는 일을 말한다. 이때의 색인어는 전혀 통제를 받지 않는 자연어 그대로일 수도 있으나 검색의 효율성을 높이기 위해서는 적절한 통제가 요구된다. 따라서 색인어는 기존 색인어휘표에서 선택되어진다.

<그림 2>는 개념 ($C_1, C_2, \cdots C_n$)의 집합인 색인어휘 (V)로 변환시켜주는 과정을 그려준 것이다.

색인작업에 의해서 각 문헌에는 색인어들의 소집합이 배정되는 것으로 색인작업은 아래와 같이 표현된다.

$$f = D \rightarrow 2^V$$

<그림 2>



색인 자체는 색인작업 결과 생산되는 것으로 색인어와 문헌의 식별요소들의 집합이 된다. 문헌의 식별요소란 특정한 색인어 아래 색인된 문헌의 청구번호를 비롯한 서지사항이 되며 기계가독형 파일에서는 흔히 문헌의 등록번호가 색인파일 내에서의 문헌의 식별요소가 된다. 즉, 색인 (I)은 색인어 (t)와 문헌 (D)의 집합이 되며 아래와 같이 표현된다.

$$I = \{ (t, D) \mid t \in V, D \in W \}$$

위 식에서 V는 색인어들의 집합인 색인어휘표이고 W는 문헌들의 집합인 문헌파일을 나타낸다.

검색작업은 이용자의 정보요구를 충족시켜주는 정보문헌을 찾아내는 일

로서 색인작업에서와 같이 정보요구 (Q) 를 구성하는 개념들을 적절한 검색어 (t) 로 변환시켜준 뒤 색인을 통해 그 검색어와 관련된 문헌들을 찾아내는 일련의 과정이다. 검색작업은 아래와 같이 정보요구→검색어, 검색어→해당문헌의 두 단계 변환과정을 거치게 된다.

$$g \cdot f : Q \rightarrow 2^V \rightarrow 2^D$$

정보 요구를 적절한 검색어로 표현하는 일과 문헌의 주제를 적절한 색인어로 표현해 주는 일은 둘 다 무질서한 상태의 정보원에 질서를 부여하는 작업으로 결과적으로는 엔트로피를 감소시키는 일이다.

3.2 색인 작업과 엔트로피

문헌파일을 구성하는 개개의 문헌들은 크거나 작거나간에 내용면에서 차이를 갖는다. 두개의 문헌의 내용이 꼭 일치하는 경우는 복본이 아니고는 존재키 어려우며 설령 있더라도 내용이 똑같은 별개의 문헌을 문헌파일내에 포함시킨다는 것은 장서 구성에 있어서 그다지 효율적이 못된다. 이용자가 어떤 특정한 문헌을 원하게 되는 것은 문헌파일을 구성하는 각 문헌들 간에 내용상의 차이가 있기 때문이다.

대학도서관의 경우에는 인문과학, 사회과학, 자연과학, 응용과학등 온갖 학문분야에 관한 문헌들이 다 소장되므로 문헌파일의 내용이 그만큼 다양해진다. 반면에 특수한 분야의 문헌만을 소장하는 특수도서관의 경우에는 문헌파일의 내용은 어느 정도 한정이 된다. 전자가 후자의 경우보다 선택될 문헌에 대한 불확실성이 훨씬 크다는 것을 알 수 있다.

색인은 문헌파일을 구성하는 문헌들을 내용상 차이가 가장 근소한 것끼리 모아주므로 해서 문헌파일의 다양성과 선택될 문헌에 대한 불확실성을 감소시켜 준다. 색인은 문헌파일을 주제별로 모아 소집단화하는 것으로 주제가 세분되면 될수록 문헌 간에 존재하는 내용상의 차이는 근소해지며 선택될 문헌에 대한 불확실성의 감소를 초래하는 것이다. 정보이론적으로 해석하자면 문헌파일이 소집단으로 세분될수록 한 집단에 속하는 문헌의 수는 적어지며 따라서 선택대상이 되는 문헌의 수가 적으므로 엔트로피가

적어지는 것이다.

N개의 문헌들로 구성된 문헌파일 W를 색인하는 과정을 살펴보자. 문헌이 색인되어 있지 않은 경우에는 문헌파일의 내용에 대해서는 전혀 모르는 상태이므로 각 문헌의 선택확률은 $\frac{1}{N}$이며 이때의 엔트로피는 $H(W) = \log_2 N$이 된다. 이 문헌파일을 색인하게 되면 문헌파일은 각각 W_1, W_2, … W_m개의 소집단으로 나누어지고 각 소집단은 $n_1, n_2, \cdots n_m$ 개의 문헌을 포함한다. 특정한 주제에 관한 문헌은 그 주제 아래 색인된 소집단 W_i 중에서 선택되며 문헌의 선택확률은 $\frac{1}{N}$에서 $\frac{1}{n_i}$ 로 변한다. 따라서 문헌파일 W의 엔트로피는 다음과 같다.

$$H(W) = \sum_{i=1}^{m} H(W_i)$$

각 소집단이 갖는 평균 엔트로피는 $H(W) / m$으로서 엔트로피의 현저한 감소를 보여준다.

문헌파일을 점점 소집단화하게 되면 결국에 가서는 한 집단에 한개의 문헌만이 배정되게 된다. 이때에는 각 집단의 엔트로피는 0이 되며 따라서 문헌파일 W의 엔트로피도 0이 된다.

문헌파일을 더 많은 수의 집단으로 나누는 것은 보다 특정한 개념을 나타내는 색인어를 배정하는 것으로 특정성의 증가를 가져온다. 색인작업은 엔트로피의 감소를 가져오는 동시에 특정성의 증가를 가져온다.

앞에서는 각 집단을 구성하는 문헌들에 동일한 선택확률이 부여되었다. 즉, 각 문헌에 해당 색인어만을 배정해 준 것으로 동일한 집단에 속하는 문헌들 간에 존재하는 내용상의 차이는 표출되지 않은 것이다. 그러나 특정한 색인어 아래 n개의 문헌이 똑같이 색인된다고 해도 각 문헌이 그 주제를 다루고 있는 정도에는 차이가 있다. <그림3>은 문헌과 색인어 의 행열이다.

다음 <그림3>에서 색인어 b가 문헌 D_1과 D_4에 똑같이 주어졌으나 b 라는 개념이 두 문헌 내에서 다루어지는 정도가 반드시 같지는 않다. D_1 에서는 b가 핵심 주제인 반면 D_4에서는 d가 핵심주제이고 b는

<그림 3>

색인어	문헌 D_1	D_2	D_3	D_4	
a		×			
b	×			×	
c	×		×		
d		×		×	
⋮					

약간만 언급되고 있다면 두 문헌이 b라는 색인어 아래 색인이 되었다고 해도 이 주제에 대해 두 문헌이 전달하는 정보의 양에는 차이가 있게 된다

한 문헌이 여러개의 중요개념을 다루고 있을 때 각 개념의 중요도를 상대적으로 표시해 주기 위해 색인어에 가중치를 부여한다. 이 방법은 주로 전산화된 정보검색시스템에서 사용되며 문헌파일을 구성하는 각 문헌에는 색인어와 가중치가 동시에 부여된다. 검색시 이용자는 일정치 이상의 가중치를 갖는 문헌만을 선택할 수 있으므로 이차적인 선별이 가능해진다. 이 때에는 집단 W_i 의 엔트로피 $H(W_i)$ 는 다음과 같이 $\log_2 n_i$ 보다 적은 값이 된다.

$$H(W_i) = -\sum_{i=1}^{n_i} P_j \log_2 P_j$$

즉, 가중치가 큰 문헌은 선택확률이 높아지고 가중치가 적은 문헌은 선택확률이 낮아지므로 정보원 (문헌파일 W_i) 이 갖는 엔트로피는 감소된다.

3.3 색인 검색 시스템의 엔트로피

문헌의 검색효율은 원하는 문헌은 많이 검색되고 원하지 않는 문헌은 적게 검색될수록 높아진다. 정보요구를 충족시킬 수 있는 문헌만이 검색되고 또한 정보요구를 충족시킬 수 있는 모든 문헌이 검색되었을 때 검색효율이 최대가 된다. 검색효율의 측정단위로 많이 사용되는 것은 정확률과 재현율로서 정확률은 검색된 문헌들이 모두 만족스러운 것일 때 100%가 되며 재현율은 정보시스템이 소장하고 있는 정보요구와 관련된 모든 문헌들이 검색되었을 때 100%가 된다. 이러한 검색효율은 정보시스템의 성능

을 좌우하는 가장 중요한 요소로서 시스팀의 경제성 및 신속성과 더불어 시스팀 평가의 기준이 되고 있다.

검색효율을 좌우하는 것은 물론 색인으로 색인이 정확할수록 검색효율은 높아진다. 색인어휘의 특정성, 색인작성의 망라성, 선택된 색인어의 적합성 등의 요소가 색인시스팀의 성능을 결정한다. 색인시스팀의 성능외에도 검색작업과 관련된 요소들이 검색효율에 영향을 미치며 여기에는 정보요구의 정확한 표현, 적합한 검색어의 선택, 검색방법의 적합성 등이 포함된다. 정보 시스팀의 성능은 절대적인 기준에 의해 평가되기 보다는 정보시스팀 상호간에 상대적으로 평가되어진다.

정보시스팀의 문헌파일을 구성하는 문헌들은 이용자의 정보요구가 있을 때 잠정적으로 그 정보요구를 충족시키는 적합문헌과 그렇지 못한 부적합 문헌으로 분류가 되어진다. 색인파일에는 이 정보요구에 상응하는 색인어 아래 그 색인어와 관련있다고 색인자에 의해 판단되는 문헌들이 수록된다. 이상적인 정보시스팀에서는 색인에 의해 검색되는 문헌들은 이용자의 판단에 의해 분류되어지는 적합문헌들만이며 또한 적합문헌 전부여야 한다. 그러나 실제로는 색인과정 및 검색과정과 관련된 여러 요인들 때문에 이러한 이상적인 정보시스팀은 존재하기 힘들다.

이용자 판단에 의해 잠정적으로 분류되는 문헌파일 W(Q), 색인검색시스팀에 의해 분류되는 문헌파일 W'(Q), 색인에 의해 검색된 문헌파일 I(Q) 는 각각 다음과 같이 표현된다.

$$W(Q) = \{ d_R, d_{\bar{R}} \}$$

$$W'(Q) = \{ d_{RS}, d_{R\bar{S}}, d_{\bar{R}S}, d_{\bar{R}\bar{S}} \}$$

$$I(Q) = \{ d_{RS}, d_{\bar{R}S} \}$$

d_R = 적합문헌

$d_{\bar{R}}$ = 부적합문헌

d_{RS} = 검색된 적합문헌

$d_{R\bar{S}}$ = 검색되지 않은 적합문헌

$d_{\bar{R}S}$ = 검색된 부적합문헌

$d_{\bar{R}\bar{S}}$ = 검색되지 않은 부적합문헌

이상적인 색인·검색 시스템에서는 $\{ d_R \} = \{ d_{RS} \}$, $\{ d_{\bar{R}} \} = \{ d_{\bar{R}\bar{S}} \}$ $\emptyset = \{ d_{\bar{R}\bar{S}} \}$, $\emptyset = \{ d_{\bar{R}S} \}$ 의 관계가 성립한다.

문헌파일 W'(Q)에서 문헌이 각 소집합에 속할 확률을 각각 P_{RS}, $d_{R\bar{S}}$, $P_{\bar{R}S}$, $P_{\bar{R}\bar{S}}$로 표시하면 $P_{RS} + P_{R\bar{S}} + P_{\bar{R}S} + P_{\bar{R}\bar{S}} = 1$ 이며 $P_R = P_{RS} + P_{R\bar{S}}$, $P_{\bar{R}} = P_{\bar{R}S} + P_{\bar{R}\bar{S}}$ 가 된다.

문헌파일 W(Q)와 W(Q)의 엔트로피 H(W)는 아래와 같이 산출된다.

$$H(W) = -(P_R \log_2 P_R + P_{\bar{R}} \log_2 P_{\bar{R}})$$

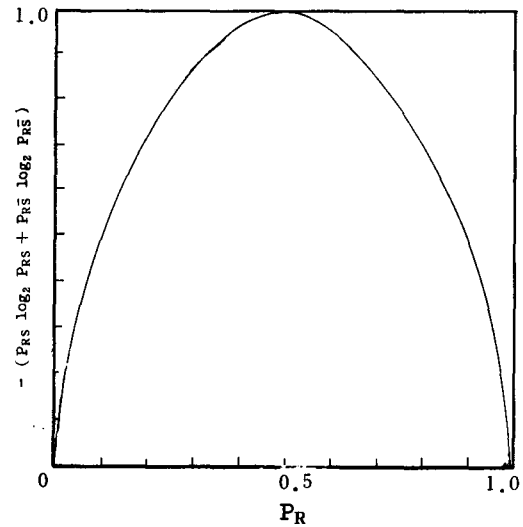
$$H(W') = -(P_{RS} \log_2 P_{RS} + P_{R\bar{S}} \log_2 P_{R\bar{S}} + P_{\bar{R}S} \log_2 P_{\bar{R}S} + P_{\bar{R}\bar{S}} \log_2 P_{\bar{R}\bar{S}})$$

위에서 $H(W) \leq H(W')$로서 완전한 색인·검색시스팀의 경우에는 $P_R = P_{RS}$, $P_{\bar{R}} = P_{\bar{R}\bar{S}}$이므로 두 엔트로피의 값은 같아진다.

검색과정에서 오류가 발생하지 않는한 $d_{\bar{R}S}$와 $d_{R\bar{S}}$는 잘못 색인된 문헌이며 정보원에 대한 불확실성을 추가시키는 잡음의 역할을 한다. 따라서 $H(W') - H(W)$의 값은 색인시스팀에서 색인의 오류로 인해 추가되는 바람직하지 못한 불확실성에 해당된다. $d_{R\bar{S}}$와 $d_{\bar{R}S}$는 검색과정에서도 발생할 수 있으나 색인과정에서 발생하는 오류는 검색과정의 오류처럼 쉽게 수정할 수 없으며 그만큼 결정적이다.

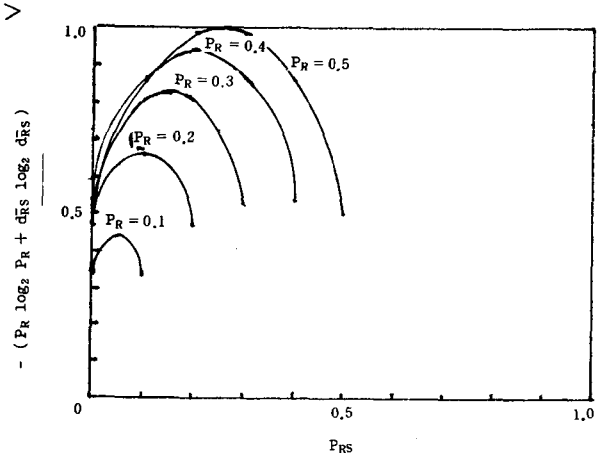
H(W)의 값이 고정되어 있으면 H(W)와 H(W')의 차이가 적을수록 정보시스팀의 성능이 높으리라는 것을 유추할 수 있다. <그림 4>는 P_R과 H(W)와의 관계를 보여준다. 이 그림에서 보는 바와 같이 P_R이 0.5 ± 0.05일때 H(W)는 최대치를 갖는다. 실제 시스팀에서는 특정한 색인어 아래 색인되는 문헌의 수는 전체 문헌파일에 있어서 일부분에 지나지 않는다. 따라서 P_R은 0.5를 넘지 않는다는 가설을 세울 수 있으며 $0 \leq H(W) \leq 1$의 관계가 성립한다.

<그림 4>



<그림 5>는 P_{RS}와 $-(P_{RS}\log_2 P_{RS} + P_{R\bar{S}}\log_2 P_{R\bar{S}})$의 관계를 보여준다. 그림에서 $P_{RS} \geq P_{R\bar{S}}$이면 P_{RS}의 값이 커질수록 $f(P_{RS})$의 값은 적어지며 또한 $P_{\bar{R}S} \geq P_{\bar{R}\bar{S}}$일때도 $P_{\bar{R}S}$와 $-(P_{\bar{R}S}\log_2 P_{\bar{R}S} + P_{\bar{R}\bar{S}}\log_2 P_{\bar{R}\bar{S}})$ 사이에는 같은 관계가 성립되므로 적합문헌의 검색확률이 높고 부적합문헌의 검색확률이 낮아 질수록 전체적으로 H(W')의 값은 적어진다. P_{RS}와 $P_{R\bar{S}}$가 커진다는 것은 상대적으로 시스팀의 잡음에 해당하는 $d_{R\bar{S}}$와 $d_{\bar{R}S}$의 수가 감소함을 의미한다.

<그림 5>



H (W) 의 증가는 H (W') 의 증가를 가져오므로 $0 \leq P_R \leq 0.5$, $P_{RS} \geq P_{R\tilde{S}}$, $P_{\bar{R}\bar{S}} \geq P_{\bar{R}S}$ 의 세 조건이 만족될 때 충분한 수의 정보요구를 표본으로 하여 두 문헌파일의 평균 엔트로피를 측정한다면 이 두 엔트로피의 차이 (E) 는 정보검색효율의 훌륭한 측정단위가 된다.

$$\frac{\sum_{t=1}^{n} H(W' \mid Q_t) - \sum_{t=1}^{n} H(W \mid Q_t)}{n}$$

지금까지 색인작업은 엔트로피를 감소시키는 일임을 설명하였고 또한 정보시스팀 성능평가의 중요한 기준인 검색효율은 이용자 판단에 의해 분류되는 문헌파일과 색인·검색시스팀에서 재분류되는 문헌파일이 갖는 두 엔트로피의 차이로써 측정할 수 있음을 설명하였다. 이와같이 Shannon 의 정보이론은 문헌정보의 커뮤니케이션에 있어서도 적절한 응용성을 갖는다.

Shannon's Information Theory and Document Indexing

Young Mee Chung*

(Abstract)

Information storage and retrieval is a part of general communication process. In the Shannon's information theory, information contained in a message is a measure of uncertainty about information source and the amount of information is measured by entropy.

Indexing is a process of reducing entropy of information source since document collection is divided into many smaller groups according to the subjects documents deal with. Significant concepts contained in every document are mapped into the set of all sets of index terms. Thus index itself is formed by paired sets of index terms and documents.

Without indexing the entropy of document collection consisting of N documents is $\log_2 N$, whereas the average entropy of smaller groups($W_1, W_2, \cdots W_m$)is as small(as $\left(\sum_{i=1}^{m} H(W_i) \right)/m$.

Retrieval efficiency is a measure of information system's performance, which is largely affected by goodness of index. If all and only documents evaluated relevant to user's query can be retrieved, the information system is said 100% efficient. Document file W may be potentially classified into two sets of relevant documents and non-relevant documents to a specific query. After retrieval, the document file W' is reclassified into four sets of relevant-retrieved, relevant-not retrieved, non-relevant-retrieved and non-relevant-not retrieved. It is shown in the paper that the difference in two entropies of document file W and document file W' is a proper measure of retrieval efficiency.

* Assistant Professor, Yonsei University