정보이론과 문헌정보검색

정 영 미
<延大 圖書館學科 敎授>

1. 서 론

커뮤니케이션이 없는 사회는 존재하기 어렵다. 사회가 지역적 특성, 언어적 특성, 정치적 특성 등 어떤 특성에 의해 구분되어지건 각 사회의 구성원들은 공동의 인식, 공동의 관습, 공동의 관심사 등을 통해 의사를 소통하며 살아간다.

의사소통의 도구는 말이나 글 또는 온갖 몸짓뿐만 아니라 음악, 그림, 무용 등을 비롯하여 마천루 꼭대기에서 반짝거리는 신호등, 비행장 활주로에 켜져 있는 안내등 또는 밀림에서 들리는 북소리 등 실로 다양하다.

본고에서는 일반적인 커뮤니케이션시스팀과 커뮤니케이션시스팀으로서의 문헌정보시스팀에 관해 약술하고 커뮤니케이션의 대상이 되는 메시지가 전달하는 정보량의 측정을 다루는 정보이론을 소개하며 또한 문헌정보검색에 있어서의 정보이론에 관해 생각해 보고자 한다.

2. 커뮤니케이션시스팀

커뮤니케이션의 일반적인 모형은 다섯가지 구성요소로 이루어진다. 즉 정보를 생산해내는 정보원(Information Source)과 정보를 전달할 수 있는 형태로 변환시키는 역할을 하는 송신체(Transmitter), 정보가 전달되는 채널(Channel), 전달된 정보를 받아 다시 원래의 형태로 환원시켜 주는 역할의 수신체(Receiver), 그리고 정보가 최종적으로 도달하는 정보의 수혜자(Destination)가 있다. 그림 1은 커뮤니케이션 시스팀의 모형이다. 전화시스팀을 예로 들어 설명하면 정보원은 통화를 하는 사람이 되고 전달되는 메시지는 말로 표현되는 통화내용이며, 송화기는 송신체로서 말을 변전류로 변환시켜 주고 변전류는 채널인 전선을 통해 수신체인 상대방의 수화기로 보내진다. 전류는 다시 말로 바꾸어져서 상대방에게 전달되는 것이다. 이때에 말이 제대로 들리지 않는다든가 혼선이 되어 다른 통화자의 말이 끼어든다든가 하는 것을 잡음(Noise)이라고 한다.

전화, 전신, 레디오, 텔레비젼 등의 텔레커뮤니케이션시스팀이 아닌 동일한 장소에서 일어나는 커뮤니케이션의 경우에도 위의 모형은 마찬가지로 적용된다. 강연회에 참석한 청중과 연사와의 사이에 존재하는 커뮤니케이션시스팀에서는 연사의 두뇌가 정보원이고 강연내용은 메시지가 되며 메시지는 송신체인 연사의 발성기관에 의해 말로 표현되고 말은 채널인 공기를 통해 음파로서 청중에게 전달된다. 청중의 청각기관은 수신체가 되며 청중의 두뇌가 최종적인 수혜자가 된다. 위의 경우에는 정보원과 송신체, 또 수신체와 수혜자를 구태어 구별지울 필요가 없다.

위의 다섯가지 구성요소를 통해 정보가 전달된다고 해서 커뮤니케이션이 성취되는 것은 아니다. 사람들간에 공통적인 관심이나 흥미 또는 공통적인 목적이나 이해 등이 존재하지 않고서는 공통의 언어를 사용하는 경우라도 의사가 소통

되지 않을 수가 있다. 예를 들어 물리학에 관한 논문은 그 분야를 전공하지 않은 사람에게는 별로 의미를 갖지 않는다. 다시 말해 그 논문의 저자가 전달하려는 메시지를 수혜자는 이해할 수 없기 때문에 그 논문을 읽더라도 실질적인 커뮤니케이션은 달성되지 않는다. 음악이나 미술의 경우에도 마찬가지이다. 고전음악을 들어본 적이 없거나 관심이 없는 사람에게 베에토벤의 태공트리오는 그다지 감명을 주지 않을 것이다. 이러한 공통의 이해나 관심의 문제는 문헌정보를 다루는 정보시스팀에서 특히 고려되어야 한다.

그림 2는 커뮤니케이션시스팀으로서의 문헌정보시스팀모형이다. 정보시스팀에서 저자는 정보검색시스팀의 소장 대상이 되는 온갖 문헌들을 생산해내는 정보원이다. 문헌을 생산해내는 과정, 즉 출판은 저자의 사상을 독자가 읽을 수 있는 형태로 기록하는 것으로 encoding의 단계이며 문헌이 전달되는 채널은 도서관, 정보센터, 정보분석센터, 데이터센터 등의 각종 정보검색시스팀이 된다. 색인은 정보검색시스팀의 일부로서 문헌에 수록된 저자의 사상 즉 메시지를 이용자에게 전달하기 위해 문헌을 분석해주는 것으로 decoding의 단계로 볼 수 있다. 다시 말해 문헌정보시스팀이란 저자와 이용자 사이에 존재하는 커뮤니케이션시스팀인 것이다. 정보검색시스팀에서 소장하는 문헌의 주제는 잠정적인 이용자들이 관심을 갖는 것이어야 하며 문헌의 내용 또한 그들의 이해 범위를 벗어나지 않는 것이어야 한다.

커뮤니케이션시스팀을 통해 메시지가 전달하는 정보의 양은 정보원과 수혜자 사이의 공통의 인식이나 관심 등에 따라 달라진다. 동일한 메시지라 할지라도 전달받는 사람이 갖는 그 메시지에 대한 이해도에 따라 전달되는 정보량이 달라진다. 예를 들어 물리학에 관해 전혀 모르는 사람에게 물리학에 관한 논문이 전달하는 정보의 양은 극히 적을 것이고 물리학에 관해 어느 정도 지식이 있는 사람에게는 훨씬 많은 양의 정보가 전달될 것이며 반면에 그 논문에 관해 이미 잘 아는 사람에게는 아무런 정보도 전달되지 않을 것이다. 음악속에서는 모든 것이 놀랍고 동시에 예상되어지는 것이어야 한다고 말한 베에토벤의 견해는 매우 타당하다. 이미 다 알고 있는 내용을 담고 있는 문헌이나 이해의 범주를 넘

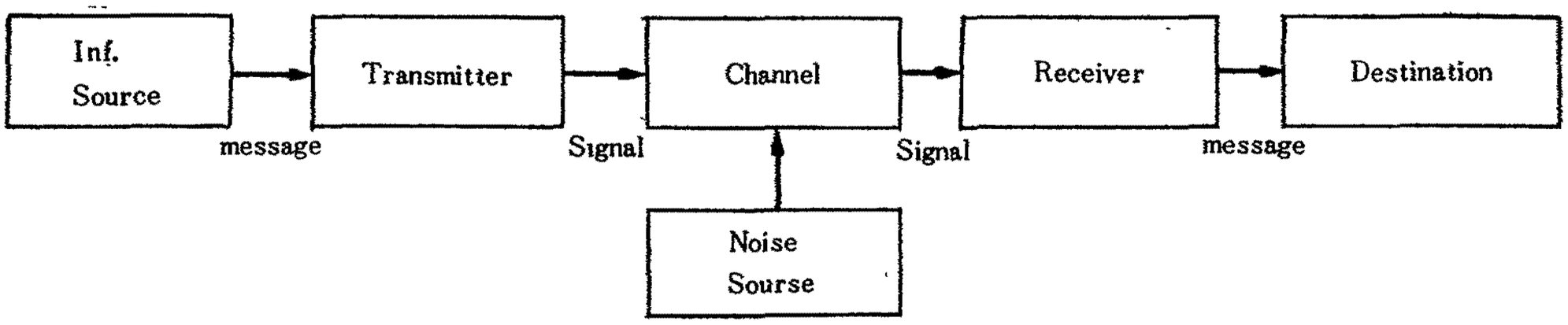

그림 1. Shannon의 커뮤니케이션시스팀모형

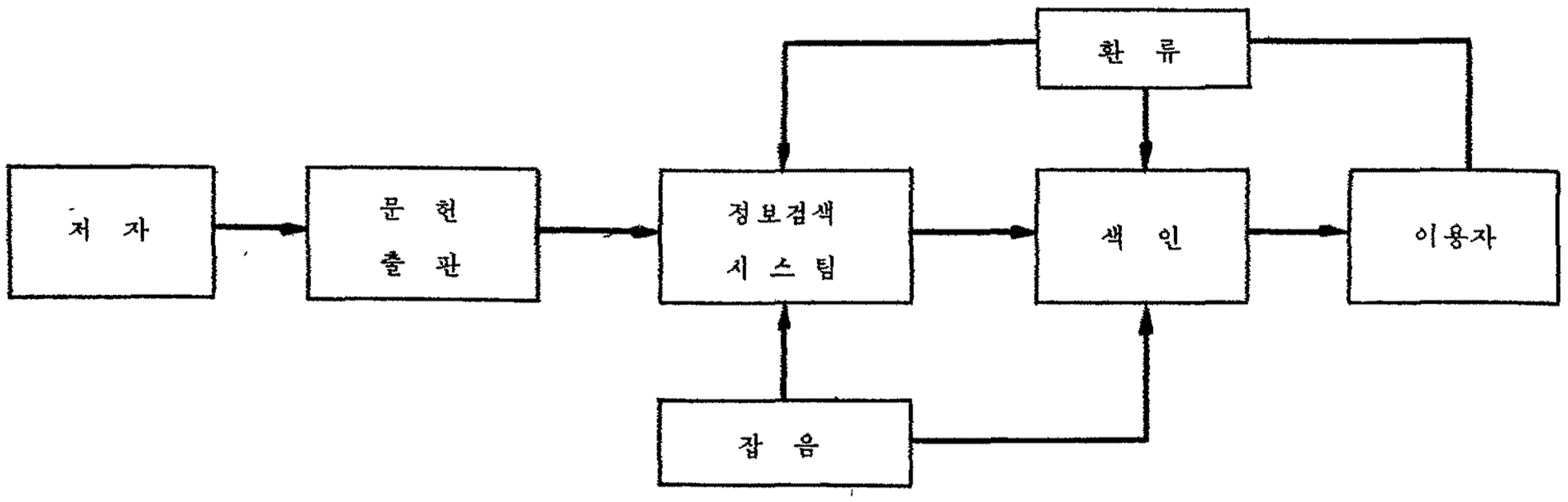

그림 2. 문헌정보시스팀 모형

어설 정도로 너무 예상밖의 내용을 담고 있는 문헌은 독자에게 별다른 가치를 갖지 못하는 것이다.

3. 정보량과 정보이론

메시지의 수혜자가 그 메시지의 내용을 알므로 해서 얻은 정보의 양을 측정하기 위해 Shannon의 정보이론이 사용된다. 정보이론은 텔레커뮤니케이션과 더불어 발전되어 왔고 여러 커뮤니케이션시스팀을 비교하는 평가기준을 마련해 준다. 전신시스팀에서의 신호전송과 관련된 R. V. L. Hartley의 1920년대의 이론을 바탕으로 하여 1948년 C. E. Shannon은 정보이론을 정립하였다. Shannon의 정보이론은 본래는 전달되는 메시지를 구성하는 부호의 발생 확률만을 고려한 것으로 개개의 부호나 메시지가 어떤 의미를 갖는지는 상관하지 않았으나 후에 W. Weaver에 의해 새로이 해석되었다. 즉 커뮤니케이션은 정보 전달의 정확성과 관련된 기술적인 면, 그리고 전달된 정보가 갖는 의미의 올바른 해석과 관련된 어의적인 면, 그리고 전달된 정보가 수혜자의 행위에 미치는 영향과 관련된 효과적인 면의 세가지 측면에서 성취되어져야 한다는 것이며 Shannon의 이론은 기술적인 면과 관련되어 있으나 나머지 두 측면과도 무관하지 않다는 것을 보여주었다. 그후 이 정보이론은 커뮤니케이션공학뿐만 아니라 생물학, 언어학, 심리학 등의 분야에 응용되기 시작했다.

정보이론에서 메시지가 갖는 정보의 양은 정보원이 생산해낼 수 있는 다양한 메시지들로부터 하나의 메시지를 선택할 때에 부여되는 선택의 자유를 측정하는 것이다. 정보이론에서 정보측정의 개념은 열역학에서의 엔트로피 개념과 유사하며 정보원이 갖는 평균 정보량을 엔트로피라고 부른다. 메시지에 의해 전달되는정보량은 메시지 내용에 대한 불확실성이 크면 클수록 증가되는데 엔트로피는 이 불확실성의 크기를 나타낸다. 정보원이 전달할 수 있는 메시지가 "네"와 "아니오"의 두가지 뿐이라면 수혜자는 이미 전달 가능한 메시지에 대해 50%는 알고 있는 것이며, 정보원이 전달할 수 있는 메시지가 열개인 경우보다 전달될 메시지에 대한 불확실성의 크기는 훨씬 적은 것이다. 즉 선택 가능한 메시지의 수가 많을수록 한 메시지가 갖는 엔트로피는 커지는 것이다.

엔트로피는 다음 공식에 의해 산출된다.

$$H = -\Sigma pi \log_2 pi$$

위 공식에서 pi는 i번째 메시지가 선택될 수 있는 확률을 나타낸다. 한 예로 졸업 축하전문을 보내고자 할 때 전문이 갖는 평균 정보량을 계산해 보자. 표 1에는 졸업을 축하하는 뜻으로 보낼 수 있는 다섯개의 전문과 해당되는 메시지가 나와 있다. 각 메시지가 선택될 확률이 똑같다면 pi는 각각 $\frac{1}{5}$이 된다. 따라서 평균정보량은 $H = -5(-\frac{1}{5} \cdot \log_2 5)$로 약 2.2가 된다. 이때 메시지의 엔트로피 H는 메시지를 이진수로 코우딩할 때 필요한 숫자의 평균치를 말하기도 한다.

표 1. 졸업축하전문 및 선택확률

전문	메 시 지	확률
조고	졸업을 축하합니다	$\frac{1}{5}$
조노	영광된 졸업을 축하하며 앞날의 성공을 빕니다	$\frac{1}{5}$
조도	형설의 공을 치하하며 더 큰 영광을 빕니다	$\frac{1}{5}$
조로	교문을 나서는 벗이여, 품은 뜻을 성취하소서	$\frac{1}{5}$
조모	영광의 졸업을 축하한다	$\frac{1}{5}$

즉 위의 전문들을 이진수로 코우딩할 때 00, 01, 10, 11, 001의 다섯가지 코우드를 사용할 수 있는데 코우드를 구성하는 이진수의 평균 숫자는 2.2로서 위에서 산출된 H와 같은 값이 된다.

위에서 각 전문이 선택되어질 확률이 동일하지 않은 경우 엔트로피 H는 적어진다. 만일 한 메시지에 주어지는 확률이 1이고 나머지 메시지들에는 0의 확률이 주어진다면 엔트로피 H는 0이 된다. 즉 이미 선택될 메시지에 대해 알고 있기 때문에 불확실성은 존재하지 않는다. 반면 앞의 경우처럼 각 메시지에 동일한 확률이 주어질 때 엔트로피 H는 극대치를 갖게 된다.

지금까지는 메시지가 전달된 후에 수혜자가 갖는 불확실성은 고려하지 않고 메시지가 전달되기 전에 갖는 불확실성만을 고려하여 정보의 양을 측정하였다. 즉 메시지가 전달되고 나서 불확실성이 완전히 제거되었을 때의 경우이다. 그

러나 실제로 전달된 정보의 양은 메시지를 받기 전과 받은 후의 불확실성의 차이로서 측정된다. 즉 메시지가 전달한 정보의 양은 수혜자의 지식상태에 있어서의 변화를 측정하는 것이라고 볼 수 있다. 즉 메시지가 전달한 정보의 양은 메시지가 전달되기 전의 엔트로피와 메시지가 전달된 후의 엔트로피의 차이로서 측정되며 아래의 공식으로 표현된다.

$$I=H(QIX)-H(QIX')$$

메시지를 받기 전의 엔트로피 H는 잘 정의된 질문 Q와 Q에 대해 메시지를 전달받기 전에 갖고 있는 지식상태 X에 의해 결정되며 메시지를 받은 후의 엔트로피 H는 메시지를 전달받은 후에 갖게 된 지식상태 X′에 의해 새로이 결정된다. 다시 말해 지식상태 X에 따라 질문 Q에 대한 가능한 여러 해답에 확률을 부여하게 되며 H가 측정된다. 여러 해답 중 하나의 메시지가 전해졌을 때 전달된 메시지에 의해 Q에 대한 지식상태 X가 X′로 변하게 되고 X′에 의해서 가능한 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 즉 10개의 해답을 갖고 있는데 첫번째 숫자인 0을 돌리기 전에 각각의 숫자에 주어지는 확률은 각각 $\frac{1}{10}$이 된다 그러나 첫번째 숫자를 돌리고 나서는 확률의 배정이 달라지게 된다. 0이 옳은 번호가 아닌 경우에는 0에 부여되는 새 확률은 0이 되고 나머지 아홉 숫자에 각각 $\frac{1}{9}$의 확률이 부여되는 것이다. 즉 0을 돌리기 전의 지식상태 X와 돌린 후의 지식상태 X′는 달라지므로 가능한 해답에 주어지는 확률이 달라지는 것이고 이때 0을 돌려보므로 해서 얻은 정보의 양은 두 엔트로피의 차이만큼인 것이다.

4. 문헌정보검색에 있어서의 정보이론

그림 2에서 전체적인 정보시스팀을 구성하는 한 요소인 정보검색시스팀은 그 시스팀을 이용하는 이용자그룹의 정보요구를 충족시키기 위해 이용자들의 관심주제와 관련된 문헌들을 소장한다. 각 문헌의 주제 즉 문헌이 담고 있는 정보들은 분석되고 적절한 코우드에 의해 대표된다. 이때의 코우드들은 색인파일을 구성하며 이용자들은 이 색인파일을 통해 원하는 문헌에 접근하는 것이다.

문헌파일을 구성하는 개개의 문헌들은 크거나 작거나 간에 내용면에서 차이를 가진다. 다시 말해 두개의 문헌이 담고 있는 정보내용이 꼭 일치하는 경우는 복본이 아니고는 존재키 어려우며 설령 있다 하더라도 내용이 똑같은 별개의 문헌을 파일 내에 포함시키는 것은 그다지 효율적이라고 볼 수 없다. 이용자가 어떤 특정한 문헌을 원하게 되는 것은 문헌파일을 구성하는 각 문헌들 간에 내용상의 차이가 존재하기 때문인 것이다.

시스팀에 따라 문헌파일을 구성하는 문헌들 간에 존재하는 내용상의 차이가 갖는 범주가 달라진다. 예를 들어 대학도서관의 경우에는 인문과학, 사회과학, 자연과학 및 공학 등 온갖 학문분야에 관한 문헌들이 다 소장되므로 문헌파일의 내용이 그만큼 다양해진다. 반면에 특수한 분야의 문헌만을 취급하는 특수도서관의 경우에는 문헌파일의 내용이 한정된다. 전자의 경우가 후자의 경우보다 선택될 문헌에 대한 불확실성이 크다는 것을 알 수 있다. 분류 및 주제색인은 문헌파일을 구성하는 문헌들의 차이가 가장 근소한 것끼리 모아주므로 해서 문헌파일의 다양성과 선택될 문헌이 갖는 불확실성을 감소시켜 주는 역할을 한다. 즉 문헌파일을 소주제에 따라 여러 그룹으로 나누어주는 것으로 주제가 세분되면 될수록 그룹지어지는 문헌들 간에 존재하는 내용상의 차이는 근소해지며 선택될 문헌에 대한 불확실성의 감소를 초래하는 것이다. 색인파일을 구성하는 개개의 색인레코드는 문헌의 내용을 대표하는 메시지로서 원하는 문헌이 검색될 확률을 지배하게 된다.

그러면 특정한 문헌이 실제로 전달하는 정보의 양에 대해 생각해 보자. 특정한 문헌이 전달하는 정보의 양은 이용자의 지식정도와 가장 큰 관련을 갖는다. 즉 똑같은 문헌을 읽더라도 갑이라는 이용자와 을이라는 이용자가 얻는 정보

의 양이 똑같은 것은 아니다. 또한 정보의 양은 이용자가 그 문헌에 수록된 정보에 대해 갖는 요구정도에 따라 달라진다. 즉 이용자의 정보요구와 관련된 문헌일 때에야 전달된 정보는 정보로서의 가치를 가진다. 따라서 정보검색 시스팀에서 문헌자체가 제공하는 정보의 양은 이용자의 정보요구와 문헌과의 관련성에 근거해서 측정되어져야 하는 것이다. 정보검색시스팀에서 정보요구와 제공된 문헌과의 관련성은 몇가지 요인의 영향을 받는다. 가장 큰 요인은 색인의 성능이다. 색인의 성능에 영향을 미치는 요소로는 색인자, 색인언어, 색인방법 등이 있는데 문헌의 내용이 제대로 분석되어 적절한 색인어가 배정되었을 때 색인은 만족할만한 성능을 발휘하게 된다. 다른 한 요인은 이용자가 자신의 정보요구를 제대로 표현하는 능력이다.

정보요구가 제대로 표현되지 못했을 때 정보요구에 따라 검색된 문헌이 이용자의 잠재적인 정보욕구를 충족시키기는 어려운 것이다.

다음은 정보이론을 문헌의 대용물(Surrogates)이 갖는 정보량을 측정하는 데에 응용한 예로서 J. Belzer(참고문헌 1)가 행한 실험내용이다. 이 실험에서는 의학분야의 연구를 수행하는 사람들로 구성된 이용자그룹이 질문을 형성하고 관련문헌의 검색작업은 실험과 관련된 정보학자들이 행하였다. 검색작업결과 이용자들은 먼저 각자의 질문에 따라 검색된 문헌들의 대용물을 제공받는다. 대용물은 서지사항, 초록, 문헌의 첫째 파라그라프, 문헌의 마지막 파라그라프, 그리고 첫째 및 마지막 파라그라프의 다섯가지로 각 이용자는 주어진 대용물을 보고 질문과 관련있다고 판단되면 "1"을 매기고 관련없다고 판단되면 "0"을 매긴다, 후에 완전한 문헌이 제공되고 이용자들은 제공된 문헌과 질문과의 관련도에 따라 대용물의 경우에서처럼 "1"과 "0"을 각 문헌에 부여한다. 각 문헌대용물과 완전문헌의 관련도 판단결과에 의해 문헌대용물이 갖는 엔트로피 H(P)와 완전문헌이 갖는 엔트로피 H(R)이 측정되고 각 문헌대용물이 실제로 전달한 정보의 양 I(P)가 산출된다. 전달된 정보의 양은 완전문헌의 엔트로피와 각 문헌대용물의 엔트로피를 더한 값에서 완전문헌 및 문헌대용물이 갖는 엔트로피 H(R, P)를 뺀 것으로 다음의 공식들이 사용된다.

$$H(P) = -[p(P)\log p(P) + p(\bar{P})\log p(\bar{P})]$$

$$H(R) = -[p(R)\log p(R) + p(\bar{R})\log p(\bar{R})]$$

$$H(R, P) = -[p(PR)\log p(PR) + p(\bar{P}\bar{R})\log p(\bar{P}\bar{R}) + p(\bar{P}R)\log p(\bar{P}R) + p(P\bar{R})\log p(P\bar{R})]$$

$$I(P) = H(P) + H(R) - H(R, P)$$

위 공식들에서 P와 $\bar{P}$는 각각 문헌대용물에 의해 관련있다고 판단된 문헌과 관련없다고 판단된 문헌을 나타내고 R과 $\bar{R}$은 각각 완전문헌을 본 결과 관련있다고 판단된 문헌과 관련없다고 판단된 문헌을 나타낸다. 또한 PR과 $P\bar{R}$은 각각 문헌대용물에 의해 관련있다고 판단된 문헌이 실제로 관련있다고 평가된 경우와 관련없다고 판단된 문헌이 실제로 관련없다고 평가된 경우를 나타내고 $\bar{P}\bar{R}$과 $\bar{P}R$은 관련없다고 판단된 것이 실제로 관련있는 경우와 그 반대 경우를 각각 나타낸다. p(P), p(R) 등은 각 경우의 문헌수를 백분율로 바꾸어 준 것으로 각 경우의 발생 확률을 나타낸다. 산출된 I(P)의 값은 서지사항이 .0953, 초록이 .1233, 첫째 파라그라프가 .1603, 마지막 파라그라프가 .1659, 그리고 첫째 및 마지막 파라그라프는 .3013으로 나타났다. 즉 완전한 문헌이 갖는 정보량을 1 bit로 했을 경우 서지사항은 평균 .0953bit의 정보를 갖는 것으로 해석된다. 실험에서 사용된 다섯가지 문헌대용물 중에서 서지사항이 가장 적은 양의 정보를 전달하고 다음이 초록으로 예상보다 적은 정보량을 가지며 첫째 파라그라프나 마지막 파라그라프 또는 첫째 및 마지막 파라그라프는 비교적 많은 양의 정보를 전달하는 것을 볼 수 있다. 이 실험적 연구의 결과를 일반화하기 위해서는 더 많은 실험적 및 이론적 연구가 요구되지만 이러한 실험결과는 보다 효율적인 정보검색시스팀을 설계하고 개발하는데 참고할 가치가 있을 것이다.

5. 결과론

정보이론은 1948년 Shannon에 의해 정립된 후 지나간 십수년간 과학 및 공학 분야에서 정보처

리와 관련된 연구에 큰 영향을 끼쳐왔으나 정보학 분야에서는 이렇다할 응용성을 발견하지 못한 채였다. 앞에서 소개한 Belzer의 실험은 정보이론을 순수확률적인 면에서만 해석하지 않고 정보내용과 관련시킴으로써 정보검색에의 응용을 시도한 좋은 예이다. 정보검색에 있어서는 정보의 양적인 면 뿐만 아니라 질적인 면이 동시에 고려되어야 하며, 통신공학에서와 같이 정보를 순수확률적으로만 처리하기 어려운 점 등이 있어 관련연구가 별로 활발하지 못한 실정이다. 문헌정보학이 하나의 과학적인 학문분야로서 확립되기 위해서는 무엇보다도 고유한 이론의 개발이 필수적이지만 타 학문분야의 기존이론을 도입하여 응용하는 연구 또한 무시할 수 없다. 현시점에서는 정보이론에 대한 보다 충분한 검토와 정보학적인 새로운 해석이 시도되어야 하리라 믿는다.

참 고 문 헌

1) Belzer, Jack. "Information Theory as a Measure of Information Content," JASIS, v. 24 no. 4 (July–August 1973) pp. 300~304

2) Cherry, Colin. On Human Communication. 2nd ed. Cambridge, MIT Press, (1966)

3) Lynch, Michael. "Variety Generation–A Reinterpretation of Shannon's Mathematical Theory of Communication, and Its Implications for Information Science," JASIS, v. 28 no. 1 (Jan. 1977) pp. 19~25

4) Pierce, John. "Communication," Scientific American, v. 227, no. 3 (Sept. 1972) pp. 31~40

5) Rapoport, Anatol. "What is Information?," ETC : A Review of General Semantics, v. 10, no. 4 (1953)

6) Rapoport, Anatol. "The Promise and Pitfalls of Information Theory," Behavioral Science, v. 1 (1956) pp. 303~309

7) Shannon, C. E. "A Mathematical Theory of Communication," Bell System Technical Journal, v. 27 (1948) pp. 379~423, 623~656

8) Tribus, M. and E. C. Melrvine. "Energy and Information," Scientific American, v. 183 no. 9 (Sept. 1971)

9) Weaver, Warren. "The Mathematics of Communication," Scientific American, v. 181, no. 7 (July 1969) pp. 11~15