문헌자동분류에서 무리짓기리론에 기초한 분류기호의 확정방법

최 영 희

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《콤퓨터를 가지고 여러가지 복잡하고 정밀한 작업을 할수 있게 하여야 하며 그렇게 하는것이 바로 정보산업시대의 요구에 맞게 일하는것입니다.》(《**김정일**선집》 증보판 제21권 41~42페지)

콤퓨터가 출현하고 정보기술이 발전하면서 사람들이 환상적으로만 생각하던 문제들이 현실로 되고있으며 자연을 정복하고 세계를 개조하는 인간의 창조적힘은 더욱더 위력한것으로 되고있다.

이 글에서는 문헌자동분류에서 추출선정된 주제어들을 종합적으로 분석하여 대상문헌에 분류기호를 부여하는 방법에 대하여 서술하려고 한다.

지난 시기 분류기호확정은 주로 이미 분류된 문헌들에서 추출된 주제어들의 특징행렬과 대상문헌에서 추출된 주제어들에 대한 특징벡토르를 작성하여 그 류사성을 판정하는 방법으로 하였다. 이 방법은 모든 주제어들에 대한 특징(해당 문헌에서 출현회수)과 함께 류사성판정을 위한 계산에 품이 많이 든다.

이로부터 현재 도서관실천에서 사용하는 분류－검색어사전에 기초하여 문헌자동분류를 위한 언어적보장수단인 분류용어사전개발이 가능하게 된 조건에서 분류기호확정에 실천에서 흔히 리용하고있는 대상식별 및 인식리론의 한 부분인 무리짓기리론을 적용하려고 한다.

식별 및 인식하여야 할 모든 대상모임에 대하여 본질적인 징표들을 추출한 다음 같거나 류사한 징표를 가지는 대상들을 묶어 단일한 대상으로 묶어놓으면 원래의 대상모임은 작은 대상모임들을 가지는 새로운 모임으로 변환된다. 이때 원래의 대상모임에 속하는 작은 대상모임(부분모임)을 보통 무리 또는 작은 집단이라고 하며 이렇게 대상모임을 보다 작은 부분모임들로 분할하는 과정을 무리짓기과정이라고 부른다. 그리고 개별적대상들이 가지는 성질들을 종합적으로 분석평가하고 미리 정해진 적용기준에 따라 분류하여 전체 대상을 몇개의 부분모임으로 분할하는 기술을 무리짓기기술이라고 한다.

무리짓기를 수학적으로 다음과 같이 모형화할수 있다.

대상모임을 X 라고 하자. 그러면 대상모임 X 는 다음과 같이 표시할수 있다.

$$X=\{x_1,\ x_2,\cdots,x_n\}$$

n 개의 성분들로 된 대상모임 X 의 무리짓기는 모임 X 를 다음의 성질을 만족하는 m 개의 부분모임 $C_1, C_2, \cdots C_m$ 으로 분할하는 과정으로 정의한다.

$$C_i \neq \phi,\ \ i=1,\ 2,\ \cdots, m$$

$$\bigcup_{i=1}^{m} C_i = X$$

$$C_i \cap C_j = \phi,\ \ i \neq j,\ \ i, j = 1,\ 2,\ \cdots, m$$

이 정의는 매 무리 C_i 안의 모든 벡토르들은 어떤 기준(류사성측도)에 대하여 서로 비슷하여야 하며 서로 다른 무리들에 속하는 벡토르들은 될수록 비슷하지 말아야 한다는것을 보여준다.

식으로부터 대상모임 X 가 서로 다른 속성을 가진 원소들로 구성되여있는 경우에는 매 원소 그자체가 무리로 된다는것, 무리짓기를 한 결과 초기의 모임 X 의 매 원소가 개별적인 무리(원소)로만 되고 더 이상 무리짓기가 진행되지 않는 경우에는 주목하는 무리짓기기준척도인 근접성측도에 대하여 부분모임으로의 분할이 불가능하다는것을 알수 있다.

문헌자동분류에서 무리짓기리론에 기초한 분류기호확정은 다음의 단계를 거쳐 진행된다.

첫번째 단계에서는 대상문헌에 대한 주제분석결과들에 대하여 무리짓기를 한다.

추출선정된 주제어들을 색인하면 매 주제어들에 대한 사전정보 즉 분류기호가 얻어진다. 분류기호는 주제어와 1:1대응관계에 있기때문에 이 분류기호들은 추출선정된 주제어모임을 보다 작은 부분모임들로 가르기 위한 기준으로 된다. 결국 분류기호는 같은 속성을 가진 주제어들을 하나로 묶어놓는 기준으로 된다.

두번째 단계에서는 매 무리들이 대상문헌에서 차지하는 비중을 계산한다.

대상문헌에서 개개의 무리들의 비중을 G_i라고 하면 G_i는 다음의 식에 의하여 계산한다.

$$G_i = \frac{|i\text{ 째 무리의 단어개수}|}{|\text{해당 문헌의 전체 단어수}|}$$

세번째 단계에서는 계산된 무리비중과 무리의 분류기호들을 종합적으로 분석하여 대상문헌의 분류기호를 결정한다.

일반적으로 대상문헌에서 추출선정된 주제어들에 대한 무리짓기를 하면 하나이상의 무리들이 형성되며 이때 매 무리들이 대상문헌에서 차지하는 비중은 각이하다. 즉 같은 비중을 가지는 무리들이 있는가하면 서로 다른 비중을 가진 무리들도 있다. 그리고 분류기호도 다양하다.

무리들의 비중과 분류기호의 다양성은 무리비중과 함께 분류기호들사이의 계층관계까지 고려하여 대상문헌의 분류기호를 확정하여야 한다는것을 말해주고있다.

도서관실천에서 분류규칙과 무리들의 비중 및 분류기호들의 다양성을 고려하여 분류기호확정규칙을 다음과 같이 설정할수 있다.(여기서 같은 류문을 가지는 무리들을 합무리, 합무리의 비중을 류문비중, 류문이 서로 다른 무리를 독립무리하고 약속한다.)

첫째로, 무리들이 하나의 류문에 종속되는 경우에는 다음의 규칙을 적용한다.

규칙 1. 무리비중이 서로 다른 경우에는 비중이 제일 큰 무리의 분류기호를 기본기호로 한다.

규칙 2. 비중이 같은 무리가 2개이상인 경우에는 그 웃준위의 분류기호를 기본기호로 한다.

둘째로, 무리들가운데 일부가 다른 류문이거나 모두 서로 다른 류문인 경우에는 다음의 규칙을 적용한다.

규칙 1. 류문개수가 2이면서 표제자체에 두개의 주제를 담고있는 경우

ㄱ. 합무리에 3개이상의 무리를 포함하고있다면 그가운데서 가장 큰 비중을 가진 무리의 분류기호와 독립무리의 분류기호를 기본기호로 선정한다.

ㄴ. 합무리에 3개이상의 무리를 포함하고있지만 비중이 같다면 그 무리들을 포함하는 웃준위의 분류기호와 독립무리의 분류기호를 기본기호로 선정한다.

규칙 2. 류문개수가 2이면서 표제자체에 두개의 주제를 담고있지 않는 경우에는 비중이 제일 큰 무리의 분류기호를 기본기호로, 작은 비중을 가진 무리의 분류기호를 보조기호로 한다. 이때 합무리속에 포함되는 무리가운데 비중이 큰것이 있으면 그 무리의 분류기호를 택하고 무리들의 비중이 같거나 비슷하다면 합무리의 웃준위분류기호를 택한다.

규칙 3. 류문의 개수가 3이상이면서 표제자체에 두개의 주제를 담고있는 경우에는 표제에 준하여 기본기호와 보조기호를 선정한다.

규칙 4. 류문개수가 3이상이면서 표제자체에 두개의 주제를 담고있지 않는 경우에는 비중이 제일 큰 무리의 분류기호를 기본기호로, 작은 비중을 가진 무리의 분류기호를 보조기호로 한다. 이때 합무리속에 포함되는 무리가운데 비중이 큰것이 있으면 그 무

리의 분류기호를 택하고 무리들의 비중이 같거나 비슷하다면 합무리의 웃준위분류기호를 택한다.

셋째로, 형성된 무리의 류문이 모두 서로 다른 경우에는 다음의 규칙을 적용한다.

규칙 1. 문헌의 표제에 두개의 주제를 담은 경우에는 표제에 근거하여 두개의 분류기호를 기본기호로 선정한다.

규칙 2. 제일 큰 무리비중이 0.5이상되는 무리가 없으면서 무리의 개수가 3이상이면 종합성문헌의 분류기호를 선정한다.

분류기호확정규칙에 근거하여 분류기호를 확정하기 위한 알고리듬을 다음과 같이 작성할수 있다.

1. 무리의 개수를 n에 등록한다.

2. 매 무리에 해당한 분류기호들의 류문종속관계를 확인하고 종속관계가 있다면 아래의 조작을 수행하고 없다면 3의 조작에로 넘어간다.

1) 매 무리들의 분류기호들이 하나의 류문인 경우에는 다음의 조작을 수행한다.

(1) 무리들을 비중이 큰 순서로 정돈한다.

(2) 무리들의 비중가운데 서로 같은것이 있는가를 확인한다.

① 만일 같은것이 없다면 비중이 제일 큰 무리의 분류기호를 선택하고 4의 조작에로 넘어간다.

② 만일 같은것이 있다면 해당 분류기호에 따르는 주석들을 참고하여 분류기호를 선택한다. 그리고 4의 조작에로 넘어간다.

③ 무리들의 비중이 같다면 웃준위의 분류기호를 선정한다. 그리고 4의 조작에로 넘어간다.

2) 무리들중 일부가 같은 류문에 종속되는 경우에는 다음의 조작을 수행한다.

(1) 같은 류문을 가지는 무리들로 합무리를 형성하고 무리개수를 m에 보관하고 새로 형성된 m개 무리의 비중을 계산한다.

(2) m개의 무리들을 비중의 크기순서로 배렬한다.

(3) m=2개이면 다음의 조작을 수행하고 아니면 (4)의 조작수행에로 넘어간다.

① 문헌의 표제가 2개의 주제를 담고있는가를 확인한다. 만일 문헌의 표제가 두개의 주제를 담고있다면 두 무리의 분류기호를 기본기호로 선정한다. 그리고 4의 조작에로 넘어간다.

② 문헌의 표제가 두개의 주제를 담고있지 않다면 두 무리의 비중이 같은가 같지 않은가를 확인하고 다음의 조작을 수행한다.

ㄱ. 합무리와 무리의 비중이 서로 다르면 비중이 큰 무리의 분류기호를 기본기호로, 작은 비중을 가진 무리의 분류기호를 보조기호로 한다. 그리고 4의 조작에로 넘어간다.

ㄴ. 합무리와 무리의 비중이 같다면 해당 분류기호에 따르는 주석들을 참고하여 하나는 기본기호로, 다른 하나는 보조기호로 선정한다. 그리고 4의 조작에로 넘어간다.

(4) m≧3이면 다음의 조작을 수행한다.

① 무리들의 비중이 같다면 3개 무리의 분류기호에 따르는 주석들을 참고하여 하나는 기본기호로, 나머지는 보조기호로 선정한다. 그리고 4의 조작에로 넘어간다. 만일 합무리에서 기본기호를 선정한다면 합무리가운데서 가장 비중이 높은 무리의 분류기호를 기본기호로, 합무리에 속하는 무리들의 비중이 같거나 류사하다면 합무리에 속하는 모든 무리의 웃준위를 기본기호로 한다.

② 무리의 비중이 서로 다르면 다음의 조작을 수행한다.

ㄱ. 만일 제일 큰 비중이 0.5이상이라면 다음의 조작을 수행하고 아니면 ㄴ의 조작에로 넘어간다.

(ㄱ) 제일 큰 비중을 가진 무리가 합무리이면 합무리가운데서 무리비중이 제일 큰 무리의 분류기호를 기본기호로, 합무리가 아니면 그 무리의 분류기호를 그대로 기본기호로 선정한다.

(ㄴ) 제일 큰 비중을 가진 무리를 제외한 나머지무리들을 비중의 크기순서로 배렬한다.

㉠ 제일 큰 비중을 가진 무리가 합무리이면 합무리가운데서 무리비중이 제일 큰 무리의 분류기호를 보조기호로, 합무리가 아니면 그 무리의 분류기호를 그대로 보조기호로 선정한다. 그리고 4의 조작에로 넘어간다.

㉡ 무리비중이 같다면 해당 분류기호에 따르는 주석들을 참고하여 보조기호를 선택한다. 그리고 4의 조작에로 넘어간다.

ㄴ. 무리들의 분류기호에 따르는 주석들을 참고하여 분류기호를 선택한다. 그리고 4의 조작에로 넘어간다.

3. n개의 무리들을 비중크기에 따라 배렬한다.

1) n=2개이면 다음의 조작을 수행하고 아니면 2)의 조작에로 넘어간다.

① 문헌의 표제가 2개의 주제를 담고있는가를 확인한다. 만일 문헌의 표제가 두개의 주제를 담고있다면 두 무리의 분류기호를 기본기호로 선정한다. 그리고 4의 조작에로 넘어간다.

② 문헌의 표제가 두개의 주제를 담고있지 않다면 두 무리의 비중이 같은가 같지 않은가를 확인하고 다음의 조작을 수행한다.

ㄱ. 무리들의 비중이 서로 다르면 비중이 큰 무리의 분류기호를 기본기호로, 작은 비중을 가진 무리의 분류기호를 보조기호로 한다. 그리고 4의 조작에로 넘어간다.

ㄴ. 무리의 비중이 같다면 해당 분류기호에 따르는 주석들을 참고하여 하나는 기본기호로, 다른 하나는 보조기호로 선정한다. 그리고 4의 조작수행에로 넘어간다.

2) n≧3이면 다음의 조작을 수행한다.

(1) 표제에 두개의 주제를 담고있는가를 확인한다.

(2) 표제에 두개의 주제를 담지 않았다면 아래의 조작을 수행하고 그렇지 않으면 (3)의 조작에로 넘어간다.

① 무리들의 비중이 같다면 종합성문헌의 분류기호를 택한다. 그리고 4의 조작수행에로 넘어간다.

② 무리의 비중이 서로 다르면 다음의 조작을 수행한다.

ㄱ. 만일 제일 큰 비중이 0.5이상이라면 다음의 조작을 수행하고 아니면 ㄴ의 조작에로 넘어간다.

(ㄱ) 제일 큰 비중을 가진 무리가 합무리이면 합무리가운데서 무리비중이 제일 큰 무리의 분류기호를 기본기호로, 합무리가 아니면 그 무리의 분류기호를 그대로 기본기호로 선정한다.

(ㄴ) 제일 큰 비중을 가진 무리를 제외한 나머지무리들을 비중의 크기순서로 배렬한다.

㉠ 제일 큰 비중을 가진 무리가 합무리이면 합무리가운데서 무리비중이 제일 큰 무리의 분류기호를 보조기호로, 합무리가 아니면 그 무리의 분류기호를 그대로 보조기호로 선정한다. 그리고 4의 조작에로 넘어간다.

㉡ 무리비중이 같다면 해당 분류기호에 따르는 주석들을 참고하여 보조기호를 선택한다. 그리고 4의 조작에로 넘어간다.

ㄴ. 종합성문헌의 분류기호를 선택하고 4의 조작에로 넘어간다.

(3) 표제주제에 따르는 분류기호들을 기본기호로 선정한다.

4. 작업을 끝낸다.

우의 방법은 문헌자동분류에서 분류기호 확정의 한가지 방법에 불과하며 가능한 모든 경우를 다 고려하여 프로그람을 개발한다는것은 결코 쉬운 일이 아니다.

문헌자동분류를 완전히 실현하자면 아직까지 해결할 문제점들이 적지 않다.

우리는 앞으로 문헌자동분류에서 제기되는 세부적인 문제들에 대한 연구를 심화시켜 자동분류의 효률을 부단히 개선해나감으로써 정보산업시대의 요구에 맞게 도서관들을 전자도서관화하는데 적극 이바지해나가야 할것이다.