문헌자동분류의 효률평가방법

최 영 희

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《콤퓨터가 처음에는 단순한 계산수단으로 출현하였지만 오늘은 방대한 정보량을 처리하는 만능의 정보처리기로 발전하여 사람들의 로동과 생활에서 필수적인 수단으로 되고있습니다.》(《**김정일**선집》 증보판 제20권 352페지)

오늘 세계적으로 과학기술은 매우 빠른 속도로 발전하고있으며 사람들의 생활에서 과학기술의 역할은 날을 따라 더욱더 커가고있다.

콤퓨터기술이 도서관사업의 이모저모에 도입되면서 사람들이 《미래의 도서관》으로 환상적으로만 그려보던 전자도서관이 출현하였으며 종래의 전통도서관들을 점차 전자도서관화, 수자도서관화하는 방향으로 나가고있다.

전자도서관은 전자장서를 자기의 물질적토대로 하여 목록관리, 열람봉사를 비롯한 도서관의 모든 관리운영과 봉사를 콤퓨터에 의하여 진행하는 하나의 통합정보체계라고 말할수 있다.

새 세기의 요구에 맞게 전자도서관의 봉사활동을 더욱 발전시키며 전통도서관을 전자도서관화, 수자도서관화하기 위한 연구사업에서 중요한 문제의 하나는 수집된 문헌들에 대한 자동분류문제이다.

도서관에 수집된 모든 문헌들에 대하여 콤퓨터로 주제분석을 진행하고 그에 기초하여 분류기호를 자동적으로 제공해주는것이 문헌자동분류이다.

자동분류에서는 분류대상인 모든 문헌들이 정확히 분류되였다면 리상적인 분류결과가 얻어졌다고 말한다. 그러나 현실적으로 자동분류에서 이러한 리상적인 분류결과를 얻는다는것은 거의 불가능하다.

아무리 훌륭한 분류체계라고 하여도 사람과 같은 고급한 의미분석능력을 소유하지 못한 조건에서 주제분석과 미등록어처리, 주제개념을 분류기호로 변환하는 과정에 생기는 오유를 완전히 피하기는 어렵다. 따라서 자동분류에서도 정확도가 어느 정도 보장되는가를 평가하는 문제가 제기된다.

다른 모든 사업에서와 마찬가지로 문헌자동분류의 효률평가도 해당 분류체계의 성능을 정확히 파악하고 그것을 부단히 갱신하여 분류정확도를 높이며 그 리용가치를 직관적으로 보여주기 위한 중요한 사업이다.

문헌자동분류의 효률평가문제는 자동분류체계의 부족점을 제때에 포착하고 그것을 갱신하여 분류의 정확도를 높이는데서 매우 중요한 문제로 제기된다. 문헌을 자동분류하였을 때 얼마만한 정확도를 보장하는가를 따지는것이 효률평가이다.

문헌자동분류의 효률평가에서 기본은 자동분류의 질적평가이다.

문헌자동분류는 도서관에서 분류사서들이 진행하는 분류작업을 지원하는 기능으로서 그 효률이 어떤가 하는것을 론할 때에는 여러가지 각도에서 평가지표를 설정할수 있다.

평가지표들을 종합하여 보면 다음과 같다.

① 사전의 용량

② 주제어추출의 정확도

③ 미등록어처리의 정확도

④ 분류의 정확도

⑤ 처리속도

⑥ 절약되는 로력수

⑦ 단위시간동안에 분류되는 문헌량

우의 지표들가운데서 ①~③까지의 지표들인 사전의 용량, 주제어추출의 정확도, 미등록어처리의 정확도는 분류정확도에 영향을 주는 주요인자들이다. 그것은 분류정확도가 사전의 용량, 주제어추출의 정확도, 미등록어처리의 정확도에 비례하기때문이다. 그러므로 사전의 용량, 주제어추출의 정확도, 미등록어처리의 정확도는 자동분류의 질적측면을 반영하는 지표라고 말할수 있다.

지표 ⑥과 ⑦은 처리속도와 관련되는 지표이다.

자동분류에서 처리속도는 단위시간당 분류한 문헌의 량을 측정하는 방법으로 계산한다. 다시말하여 1시간 또는 24시간동안에 얼마만한 량의 문헌들을 분류하였는가를 측정하는 방법으로 계산한다. 처리속도가 높을수록 단위시간동안에 분류되는 문헌량은 많아지며 로력을 더 많이 대신할수 있을뿐아니라 그에 따라 얻게 되는 경제적리득도 크다.

일반적으로 콤퓨터의 처리속도는 콤퓨터의 성능을 평가하는 지표로서 설비조건에 크게 의존한다. 같은 프로그람이라고 하여도 서로 다른 성능을 가진 콤퓨터에서 수행시킬때 처리속도에서는 일정한 차이가 있다. 그러므로 콤퓨터의 처리속도에 관계되는 절약되는 로력수와 단위시간당 분류되는 문헌량도 콤퓨터의 장치환경에 따라 달라진다.

그러나 자동분류에서 분류정확도는 프로그람자체에 내재하고있는 고유한 특성으로 하여 장치환경에 무관계하다. 따라서 자동분류의 효률을 평가할 때 그 질을 평가하는 지표들을 합리적으로 설정하는것이 중요하다.

일반적으로 자동분류의 효률을 평가할 때 적중률과 완전률이라는 지표를 많이 리용한다.

자동분류의 적중률은 실지 분류된 문헌들가운데 정확히 분류된 문헌수의 백분률을 말한다.

$$적중률 = \frac{정확히\ 분류된\ 문헌}{실지로\ 분류한\ 문헌수} \times 100(\%)$$

자동분류의 완전률은 해당한 분류기호에 적합한 문헌들가운데서 실지로 분류된 적합한 문헌수의 백분률을 말한다.

$$완전률 = \frac{실지로\ 분류된\ 적합한\ 문헌}{적합한\ 문헌수} \times 100(\%)$$

검색에서와 마찬자지로 자동분류에서도 리용자들은 대상문헌들이 모두 정확히 분류될것을 요구한다. 즉 100%의 완전률과 100%의 적중률을 지향한다.

그러나 이러한 목표는 어디까지나 리상적인것이고 현실적으로는 도달하기 어렵다.

자동분류실천에서는 대상문헌이 분류되지 않는 현상(루실)이 없지만 오분류현상(소음)은 생긴다.

자동분류에서 소음은 주제어선정과 미등록어처리, 분류기호선정에서 생긴다. 그러므로 자동분류의 정확성은 주제어선정과 미등록어처리, 분류기호선정의 정확성에 의하여 결정된다고 말할수 있다.

자동분류의 효률은 세부적으로 주제어선정과 미등록어처리, 분류기호선정의 정확성견지에서 평가하는것이 더 합리적이라고 볼수 있다. 그것은 자동분류가 주제어선정과 미등

록어처리, 분류기호선정단계를 거쳐 완성되기때문이다. 그러므로 매 단계에서의 효률평가를 하여야 그에 따르는 정확한 대책을 세울수 있으며 나아가서 자동분류체계의 성능을 부단히 높일수 있기때문이다.

자동분류에서는 문헌에서 추출선정된 주제어들을 분류용어사전과 비교하여 분류기호를 도출하기때문에 추출선정된 단어들이 가능한껏 사전적단어들과 일치하여야 문헌의 주제에 맞는 정확한 분류기호를 얻어낼수 있다. 그러므로 주제어선정의 정확성은 본문에 실지 존재하는 단어들가운데서 사전적단어들을 얼마나 많이, 정확히 찾아내는가에 의하여 평가된다.

해당 본문에서 사전적단어들을 얼마나 많이, 정확히 찾아내는가를 완전률과 소음률로 평가할수 있다. 주제어선정의 완전률은 본문에서 사전적단어들을 얼마나 추출선정하였는가를 본문안에 실지 존재하는 사전적단어들과의 비률로 표시한다.

이제 본문에 존재하는 단어들의 모임을 $W_{완}$, 그가운데서 사전적단어들의 모임을 $W_{사}$, 분류기가 찾아낸 적합한 사전적단어들의 모임을 $W_{적}$이라고 하면 $W_{완}$과 $W_{사}$, $W_{적}$의 관계를 다음과 같이 표시할수 있다.

$$W_{적} \subseteq W_{사} \subseteq W_{전}$$

이때 주제어선정의 완전률을 $W_{완}$이라고 하면

$$W_{완} = \left| \frac{W_{적}}{W_{사}} \right| \times 100(\%)$$

라고 표시할수 있다.

소음률은 분류기가 사전적단어라고 추출선정한 단어들가운데 오진한 단어들이 얼마나 되는가를 나타내는 값이다. 분류기가 사전적단어라고 추출선정한 단어모임을 $W_{분}$, 오진한 사전적단어모임을 $W_{오}$, 소음률을 $W_{소}$라고 하면

$$W_{소} = \left| \frac{W_{오}}{W_{분}} \right| \times 100(\%)$$

라고 표시할수 있다.

우의 식에서 보는바와 같이 완전률은 분류기가 찾아낸 적합한 사전적단어가 많을수록 커지고 소음률은 분류기가 오진한 사전적단어개수가 작을수록 작아진다. 따라서 자동분류에서 주제어선정의 정확성을 높이려면 완전률을 최대한 높이는것과 함께 소음률을 최소로 되게 하여야 한다.

자동분류에서 소음은 단어결합과 동의어, 다의어들을 언어학적요구에 맞게 정확히 분석처리하지 못한데로부터 생긴다. 때문에 주제어선정에서 정확성을 높이려면 문헌의 내용을 표현하는 주제어들을 문장론적으로나 의미론적으로 변함이 없이 정확히 분석처리하는데 깊은 주의를 돌려야 한다.

미등록어처리의 정확성도 우와 같은 방법으로 평가할수 있다.

분류기호선정의 정확성은 다른 방법으로 평가할수 있다.

자동분류체계는 선정된 주제어들에 대하여 분류기호를 제시한다. 이때 제시되는 분류기호는 일반적으로 하나이상이며 이 분류기호들이 그대로 대상문헌의 분류기호가 되는것은 아니다.

물론 자동분류체계가 제시한 분류기호가 대상문헌에 알맞는 분류기호로 되는 경우도 있을수 있지만 일반적으로 제시된 분류기호들중에서 1~2개 특이한 경우에 3개까지의 분류기호만이 대상문헌의 분류기호로 선정될수 있다. 그러므로 자동분류체계가 추출선정한 주제어에 대하여 제시하는 분류기호들을 분류기호후보라고 할수 있다.

자동분류체계가 제시하는 분류기호후보들가운데 대상문헌에 부여할수 있는 분류기호가 얼마나 있는가를 나타내는것이 바로 분류기호제시의 정확도이다.

분류기호후보제시의 정확도는 분류기호후보제시능력으로 평가한다.

분류기호후보제시능력은 대상문헌에서 추출선정된 주제어에 대하여 제시하는 분류기호후보들가운데에 정확한 분류기호가 얼마나 있는가를 나타내는 능력이다.

일반적으로 문헌에 대한 분류기호는 하나이지만 경우에 따라 몇개의 기본기호와 반영기호까지 대응될수 있다. 분류전문가에 의하여 문헌에 부여된 분류기호를 하나의 모임으로, 자동분류체계에 의하여 제시되는 분류기호후보들을 다른 하나의 모임으로 보면 분류기호제시의 정확도는 두 모임의 류사도측도로 평가할수 있다.

분류기호제시의 정확도를 모임비교에 가장 많이 리용되는 타니모토측도로 평가하려고 한다.

두 모임 즉 정확한 분류기호모임 X 와 분류체계가 제시하는 분류기호후보모임 Y 가 주어졌을 때 두 모임사이의 타니모토측도는 다음과 같이 표시할수 있다.

$$\frac{|X \cap Y|}{|X|+|Y|-|X \cap Y|} = \frac{|X \cap Y|}{|X \cup Y|}$$

이 식에서 모임에 대한 절대값기호는 해당 모임의 농도를 표시한다. 결국 모임에 대한 타니모토측도는 공통원소개수에 대한 서로 다른 원소개수의 비로 계산된다.

분류기호제시정확도가 1에 가까울수록 자동분류체계는 분류전문가들의 분류작업을 완전히 대신할수 있는 리상적인 분류기호제시능력을 가졌다고 말할수 있다.

분류기호제시의 정확도와 주제어선정의 정확성은 자동분류의 정확도에 결정적인 영향을 주는것만큼 호상 밀접한 련관관계를 가진다. 분류기호제시의 정확도는 주제어선정에서 정확성에 의존한다. 주제어선정에서 정확성이 보장되지 못하면 분류기호제시의 정확도에 대하여 기대할수 없으며 나아가서 분류의 정확도에 부정적인 영향을 준다. 한편 아무리 주제어선정에서 정확성이 보장되였다 하더라도 분류기호제시의 정확도가 보장되지 못하면 분류의 정확도에 대하여 기대할수 없다. 이로부터 문헌자동분류의 정확도는 주제어선정의 정확성과 분류기호의 정확도에 의하여 담보된다고 말할수 있다.

결국 주제어선정의 정확성측면에서나 분류기호후보제시의 정확도측면에서 높은 효률을 보장하는 자동분류체계만이 본문에 대한 분류를 원만히 진행할수 있다고 말할수 있다.

우리는 앞으로 자동분류에서 제기되는 여러가지 문제점들을 해결하기 위한 연구를 더욱 심화시켜 자동분류의 정확도를 부단히 개선해나감으로써 우리 식의 문헌자동분류체계를 완성하는데 적극 이바지하여야 할것이다.