김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제11호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 62 No. 11 JUCHE105(2016).

주제속성을 리용한 자료검색의 한가지 방법

리명일, 김예화

사용자들이 요구하는 주제에 대하여 널리 분산되여있는 자료모임에서 가장 적합한 자료를 검색하는것은 절실한 문제로 제기된다.

선행연구들[1, 3]에서는 사용자들의 정보요구를 응용, 개발, 리론의 범주측면에 귀착시키고 자료모임을 얻어내기 위한 범주색인사전의 작성방법 및 그것에 기초한 중요잡지출력모형을 제기하였다. 또한 프락탈리론[2]에 기초하여 사용자의 정보요구에 적합한 화상을 얻어내는 전문화상검색방법도 제기하였다.

론문에서는 본문, 정화상, 동화상자료의 특성을 분석한데 기초하여 주제속성을 리용한 자료검색방법을 제안하였다.

1. 주제속성을 리용한 자료검색방법

모든 문서에는 작성자가 주장하는 주제가 있다.

한편 사용자는 문서열람과정에 작성자가 주장하는 주제외에 자기자체의 주제를 설정할수 있다. 그러므로 주제를 문서의 기본내용을 담고있는 기본주제와 사용자자신의 요구에 부합되는 보조주제로 분류한다.

일반적으로 본문자료로는 저자, 잡지(도서)명, 출판사 등과 같은 항목을 리용하며 화상(정화상, 동화상)자료로는 창작가, 창작(촬영)년월일, 입수대상, 사용조건과 같은 보조항목을 리용한다. 이러한 항목들은 자료를 보관, 관리, 리용하는데서 매 자료를 식별하기 위한 중요한 인자로 된다.

화상자료는 문자나 수자로 이루어진 본문자료와 달리 주제를 정확히 반영하기 힘들므로 론의대상을 기본화상과 배경으로 나누어볼수 있다. 기본화상이나 배경에 대한 내용은 본문으로 서술되여있지 않기때문에 주제판단에서 편차가 심하다. 그러므로 주어진 화상자료에서 화상의 주제를 반영할수 있는 기본항목을 화상의 제목과 해설문으로 정한다.

화상자료에서 실마리어는 화상의 제목, 해설문 등과 같은 본문으로 서술된 항목을 원천으로 하여 생성한다. 그러므로 론문에서는 주제속성을 실마리어항목으로 표현하고 본문자료와 화상자료에 대한 검색을 통털어 자료검색으로 취급한다.

만일 한 실마리어가 모든 자료의 본문에 출현한다면 그 실마리어에 대응한 내용을 정확히 검색하는것은 어렵다.

그러나 몇개의 본문에만 출현하는 실마리어라면 검색적가치가 있다고 볼수 있다.

목적하는 자료를 $b_i \in B$, 실마리어를 $g_i \in G$, g_i가 자료 b_i의 내용을 반영하는 정도를 w_{ij}라고 할 때 자료 b_i는 실마리어벡토르 $\vec{g}_j = (w_{i1}, w_{i2}, \cdots, w_{im})$에 의하여 표현되게 된다. 여기서 B는 자료모임, G는 실마리어모임이며 m은 실마리어의 개수이다.

실마리어의 무게를 객관적으로 판정하기 위하여 실마리어에 대한 적합도와 리용도를 받아들인다.

적합도는 그 실마리어를 자료검색에 리용하였을 때 실마리어가 자료를 어느 정도 잘 나타내는가를 반영하는 기준이다.

리용도는 자료의 검색에서 그 실마리어가 실제로 리용될 가능성을 나타내는 기준이다.

그러므로 실마리어의 적합성과 리용성에 기준하여 보다 사용하기 쉬운 실마리어를 리용하게 되면 자료검색의 효과성을 높일수 있다.

실마리어의 적합도와 리용도는 이 실마리어를 포함하고있는 본문과 전체 본문모임에서의 분포정도에 기초하여 통계적으로 수값화된다.

결국 실마리어의 무게는 수값화된 적합도와 리용도의 적으로 계산할수 있다.

화상자료인 경우 매 화상에 따르는 본문자료는 제목, 해설문 등을 념두에 둔다.

자료모임 B에서 자료의 수 N, 실마리어 g에 대하여 B에 속하는 본문가운데서 g가 나타나는 화상의 수를 $N(g)$라고 하자.

이때 실마리어 g의 적합도가 높다는것은 g를 검색실마리어로 리용하였을 때 검색범위를 좁혀나간다는것을 의미한다. 실마리어검색에서는 보통 1개 이상의 실마리어에 의하여 질문식이 작성되는데 실례로 g_1 and g_2와 같은 형태로 질문식이 주어지는 경우 g_1로 먼저 검색하고 다시 g_2로 검색한다.

실마리어 g의 적합도는 다음과 같이 표시된다.

$$H(g)=\log(N/N(g)) \tag{1}$$

웃식을 분석하면 $N(g)$가 0일 때에는 $H(g)$가 정의되지 않는다. 그러나 $g\in G$는 고정된 화상모임 D에서 취한것이므로 항상 $N(g)\geq 1$이며 따라서 항상 $H(g)\neq 0$이다. 그리고 $N(g)$가 N과 같을 때에는 $H(g)=0$이다. 이것은 실마리어 g가 모든 화상에 다 출현한다는 것을 보여주며 결국 검색대상을 전혀 줄이지 못했다는것을 알수 있다. 이로부터 $H(g)=0$인 g는 실마리어로서 적합하지 못하다고 본다.

이제 두 실마리어의 적합도를 대비해보자.

식 (1)로부터 $H(g_1)-H(g_2)=\log[N(g_2)/N(g_1)]$이다.

한편 검색에서 실마리어가 쓰이는 정도를 나타내는 리용도는 검색자의 환경에 많이 의존된다. 그러므로 실마리어 g의 리용도는 매 요인마다 일정한 무게를 주고 이것을 실마리어가 리용된 빈도수에 기초하여 수값화한다.

자료모임에서 g의 사용빈도에 기초한 g의 리용도 $U(g)$는 다음과 같이 표시된다.

$$U(g)=(N(g)/N)^{\alpha} \tag{2}$$

여기서 α는 사용자가 선택하는 파라메터이다.

식 (2)에서 보는바와 같이 실마리어의 리용도는 자료모임에서 많이 출현하는 실마리어일수록 검색에서 리용되기 쉽다는것을 보여준다.

자료모임에서 실마리어 g의 무게 $W(g)$는 다음과 같이 표시된다.

$$W(g)=H(g)\times U(g)=-\log(N(g)/N)\times(N(g)/N)^{\alpha} \tag{3}$$

이러한 실마리어 g의 무게에 턱값을 주고 자료검색에 리용할수 있다. 즉 $W(g)\geq T,\ T>0$을 만족시키는 실마리어를 실마리어모임의 원소로 한다.

우의 과정을 리용하여 자료검색을 진행하는 방법은 다음과 같다.

① 본문, 정화상, 동화상 등의 자료를 화일의 확장자를 리용하여 분류한다.

② 화상(정화상, 동화상)자료에 대하여 제목, 해설문 등을 참고하여 실마리어모임을 구성한다.

③ 실마리어에 대한 적합도와 리용도를 계산한다.

④ 실마리어의 적합도와 리용도를 리용하여 무게를 계산하고 일정한 턱값(T)을 적용한다.

⑤ $W(g) \geq T\,(T > 0)$ 를 만족시키는 실마리어(g)를 실마리어모임(G)의 원소로 한다.

⑥ 실마리어를 포함한 검색질문식을 리용하여 자료의 검색을 진행한다.

2. 실험결과 및 분석

우리는 주제속성을 실마리어로 하여 본문자료뿐아니라 화상자료검색을 진행하였는데 그 결과는 그림과 같다.

그림에서 보는바와 같이 실마리어의 무게는 리용도값이 커지는데 따라 증가하다가 일정한 한계점을 지나면 감소하며 파라메터 α가 증가함에 따라 무게의 값은 작아진다. 그리고 리용도가 대략 0.3이고 $\alpha = 1$일 때 실마리어의 무게는 최대로 된다. 이것은 리용도가 0.3이고 $\alpha = 1$일 때 자료검색효과가 가장 좋다는것을 보여준다.

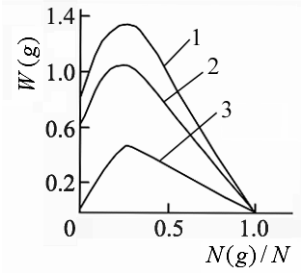

그림. 리용도에 따르는 무게값 그라프
1－$\alpha = 0.4$ 일 때, 2－$\alpha = 0.5$ 일 때, 3－$\alpha = 1$ 일 때

맺 는 말

주제속성을 실마리어로 하여 본문, 정화상, 동화상자료검색을 진행하기 위한 한가지 방법을 제기하고 실험을 통하여 그 효과성을 확증하였다.

참 고 문 헌

[1] M. Ashikhmin; ACM Symp., 217, 2013.

[2] Song Xiaomu; Geoscience and Remote Sensing Letters, 189, 2012.

[3] Wang Xiaobing; Natural Language Process, 2, 475, 2008.

주체105(2016)년 7월 5일 원고접수

A Method for Data Search using Subject Property

Ri Myong Il, Kim Ye Hwa

We realized a method for search of text, image and video using expression of key word about subject property.

Key words: data search, information abstraction