김일성종합대학학보
정보과학
주체108(2019)년
제65권 제3호

조선어교육지원체계에서 실마리어정보를 리용한 본문 및 화상자료의 검색

리 명 일

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《나라의 과학기술을 발전시키자면 과학기술정보사업체계를 정연하게 세워야 합니다.》

조선어교육지원체계에서는 본문뿐아니라 화상, 음성, 다매체자료들을 많이 리용하며 이 자료들은 분산되여 존재한다.

한편 조선어교육지원체계에서 교육내용을 편집하기 위하여서는 편집자가 주제를 설정하고 그에 알맞는 본문과 화상자료들을 수집하여야 하며 이 과정은 실마리어를 리용한 자료검색[1, 3]을 요구한다.

본문자료검색에서는 일반적으로 문서내용의 주요단어들을 색인화하고 실마리어와 색인단어의 호상관계를 리용하여 문서를 검색하는 방법[2]을 리용한다. 이때 문서내용의 주요단어는 단순단어분리에 의한 자연어검색방법을 리용하여 추출된다.

또한 화상자료의 검색에서는 요구하는 표본화상자료와 전체 화상을 크기단위로 검사하고 일치하는 경우 검색결과를 출력한다.

선행한 본문 및 화상자료의 검색방법에서는 본문검색과 화상검색이 서로 다른 방법으로 진행되는것으로 하여 정확도가 낮으며 특히 화상자료검색은 표본화상과 전체 화상을 크기단위로 검사하므로 검색속도가 뜨다.

론문에서는 본문 및 화상자료의 기록구조를 확정하고 실마리어의 무게를 결정하기 위한 한가지 방법을 제기하였다.

1. 본문 및 화상자료의 검색을 위한 실마리어의 무게결정

본문자료의 실마리어모임가운데서 매 실마리어는 본문의 내용을 설명하는데 서로 다른 기여를 하므로 매 실마리어마다 내용을 설명하는 무게도 역시 서로 다르다.

만일 한 실마리어가 모든 본문에 출현한다면 이것으로 어느 한 본문을 식별한다는것은 어렵다.

그러나 몇개의 본문에만 출현하는 실마리어라면 검색적가치가 있다고 볼수 있다.

이러한 원리에 기초하여 본문자료에서 실마리어의 출현빈도수나 문헌위치에 따르는 실마리어탐색방법이 많이 리용되고있다.

문서를 이루는 매개 용어에는 해당 용어가 가지는 무게값을 부여할수 있는데 그러한 무게값은 문서에서 용어들의 출현회수에 관계된다. 용어들에 무게를 할당하는 가장 간단한 방법은 문서 d에서 용어 t들의 출현회수로서 무게값을 할당하는것이다. 이러한 무게결정값을 용어빈도수라고 부르고 tf로 표시한다.

한편 조선어교육지원체계에서는 본문과 화상자료에 대한 기록들을 리용한다.

① 본문에 대한 기록구조

$$RT = <t, author, data, K>$$

② 화상에 대한 기록구조

$$ST = <t, author, data, f\_on>$$

여기서 *t*는 제목, *author*는 저자, *data*는 날자, *K*는 실마리어모임, *f_on*은 새로 창작 혹은 이전에 창작, 창작중인가를 나타내는 기발이다.

모든 화상자료들은 실마리어를 통하여 높은 정확도로 식별할수 있어야 하며 전체 화상뿐아니라 정화상의 일부분, 동화상의 일부 장면들도 리용할수 있어야 한다.

론문에서는 화상자료에 대한 내용이 화상에 대한 해설문에 서술되여있다고 보고 화상자료검색을 본문자료의 검색으로 진행한다.

화상에 따르는 본문자료를 $d_i \in D$, 실마리어를 $k_j \in K$, k_j가 화상에 따르는 본문자료 d_j의 내용을 반영하는 정도를 ω_{ij}라고 할 때 d_j는 실마리어벡토르 $(\omega_{i1}, \omega_{i2}, \cdots, \omega_{ink})$에 의하여 표현되게 된다. 여기서 *D*는 화상에 따르는 본문자료모임, *K*는 실마리어모임, *nk*는 실마리어의 개수이다. 이후부터 화상에 따르는 본문자료모임 *D*를 화상자료모임으로 고찰한다.

이제부터 고찰하는 화상자료라고 할 때에는 매 화상에 따르는 본문자료 즉 제목, 해설문 등을 념두에 둔다.

화상자료모임 *D*에서 화상의 수를 *N*이라고 하자.

실마리어 *k*에 대하여 화상자료모임 *D*에 속하는 본문가운데서 *k*가 나타나는 화상의 수 *N*(*k*)는 모든 $d \in D$에 대하여 다음과 같이 계산한다.

$$N(k) = 1$$

$I = I + 1$: 실마리어 *k*가 화상자료 *d*에 출현

$I = I$: 실마리어 *k*가 화상자료 *d*에 출현하지 않음

여기서 *I*는 화상자료모임 *D*의 요소수를 나타내며 초기에는 $I = 0$이다.

실마리어(*k*)에 대하여 다음과 같은 판정식을 리용한다.

$$H(k) = \log(N / N(k))$$

$N(k) = 0$이라면 *H*(*k*)가 정의되지 않는다. 이것은 실마리어가 들어있는 화상이 존재하지 않거나 실마리어가 잘못 선택되였다는것을 의미한다.

$N(k) \geq 1$이면 $H(k) \neq 0$이다. 즉 *N*(*k*)에 비례하여 *H*(*k*)는 작아진다.

$N(k) = N$이면 $H(k) = 0$이다. 즉 실마리어 *k*가 모든 화상에 출현하며 화상을 검색하지 못한다는것을 의미한다.

실례로 화상에 대한 해설문에서 《이 설계도는 …》라는 표현은 모든 해설문에 존재하므로 《설계도》라는 실마리어는 적합치 않다.

실마리어추출 및 무게결정방법은 다음과 같다.

실마리어는 품사가 명사인 단어, 합성명사, 명사구로 결정하였다.

단계 1 해설문제목에 대한 형태부해석(어간추출)

이 단계에서는 해설문제목에 있는 실마리어가 문서에서 가장 무게가 크다고 보고 추출한다. 정합결수를 1로 한다.

단계 2 해설문내용에서 실마리어추출

여기서는 먼저 문장을 입력한다.

다음 문장의 문법요소(., !, — 등)를 제거한다.

각종 토를 분리하고 제거한다.

어간을 얻은 후 실마리어목록을 리용하여 실마리어를 추출한다.

다음의 식을 리용하여 무게를 결정한다.

$$W = mi \times tf \times H(k)$$

여기서 mi는 정합결수, tf는 용어빈도수, $H(k)$는 실마리어(k)가 들어있는 화상의 수이다.

2. 조선어교육지원체계에서 본문 및 화상자료의 검색성능평가

론문에서는 화상의 주제가 화상에 대한 해설문의 제목과 내용에 반영되여있다고 보고 화상검색을 해설문자료에 대한 검색으로 실현하였다.

한편 실마리어의 리용정도에 따라 실마리어에 정합결수를 설정하였다. 이것은 실마리어에 따라 화상자료를 검색하는 경우 보다 정확하게 검색할수 있도록 하기 위하여 많이 리용되는 실마리어에 높은 무게를, 적게 리용되는 실마리어에 작은 무게를 할당한것이다. 이 결수들은 통계적인 분석에 기초하여 미리 설정한 값이다.

설정된 실마리어에 따르는 정합결수(mi)는 표 1과 같다.

표 1. 실마리어에 따르는 정합결수

번호	실마리어	정합결수(mi)
1	교실	0.9
2	학생	0.9
3	생활	0.8
4	학습장	0.7
⋮	⋮	⋮
20	공원	0.2

다음 화상의 기록구조에서 실마리어항목을 다음과 같이 추가한다.

$$ST = < t, author, data, f\_on, K >$$

여기서 K는 해설문을 리용한 실마리어모임이다.

다음 무게를 결정한다.

$$W = mi \times tf \times H(k)$$

실마리어들에 대한 무게합을 계산한다. 즉 문서에서 질문에 들어있는 실마리어들의 무게합을 계산한다.

화상문서에서 실마리어들의 무게합에 따라 작아지는 순서로 검색된 문서들을 정렬한다.

론문에서 제기한 방법으로 평가한 본문검색률과 화상검색률평가는 표 2와 같다.

표 2에서 보는바와 같이 4차에 걸치는 시험에서 평균검색률은 모두 97.5%이상이며 론문에서 제안한 방법에 의한 검색률은 98%이다.

표 2. 본문검색률과 화상검색률평가

실험차수	본문검색률/%	화상검색률/%	평균검색률/%
1	97	98	97.5
2	99	97	98
3	99	98	98.5
4	98	98	98

맺 는 말

본문 및 화상자료의 기록구조를 확정하고 실마리어의 무게결정방법을 확립하였으며 본문 및 화상자료의 검색성능을 평가하였다.

참 고 문 헌

[1] 백영철; 확률적언어모형, 중앙과학기술통보사, 1～112, 주체92(2003).

[2] Q. N. Rockiman; Natural Language Process, **2**, 2, 32, 2015.

[3] R. Thibaux; Natural Language Process, **3**, 3, 45, 2017.

주체108(2019)년 5월 5일 원고접수

Retrieval of Text and Image Data Using Keyword Information in Computer-Aided System for Korean Education

Ri Myong Il

In this paper we determined the record structure of text and image data and established a weighting method of key words.

Key words: database, image retrieval, language process