김일성종합대학학보
정보과학

주체109(2020)년
제66권 제2호

올림말용어의 기록구조에 기초한 표준실마리어사전작성의 한가지 방법

리 명 일

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《우리는 정보기술, 나노기술, 생물공학을 발전시키는데 선차적으로 힘을 넣어야 하며 그중에서도 정보기술 특히 프로그람기술을 빨리 발전시켜야 합니다.》(《**김정일**선집》 증보판 제22권 21페지)

표준실마리어사전은 사회과학, 자연과학의 여러 분야에서 정보자료들을 검색하는데 리용되는 사전으로서 표준실마리어사전관리를 보다 과학적으로 실현하는것은 중요한 문제로 나선다.

표준실마리어사전[1]은 용어들을 단순히 자모순으로 배렬하고 뜻을 풀이하는 일반언어사전과는 다른 특성을 가지고있다.

표준실마리어에 대한 선행연구[2]에서는 용어들의 호상관계문제를 론의하지 않고 단순히 문서에 출현하는 용어의 빈도를 리용하여 표준실마리어를 설정하였다.

론문에서는 우리 나라 올림말용어들의 기록구조를 분석한데 기초하여 표준실마리어사전을 작성하는 한가지 방법을 제안하였다.

1. 올림말용어의 기록구조

일반사전에서는 용어들이 서로 의존하지 않고 독립적으로 존재하지만 표준실마리어사전에서는 모든 용어들이 론리적관계로 서로 련관되여있다.

그러므로 먼저 올림말용어들에 대한 기록구조를 확정하는것이 선차적인 문제로 나선다.

올림말용어에 대한 기록구조는 다음과 같다.

$$Rec = <\mathrm{No}, d, R, S, P>$$

여기서 No는 올림말의 번호, d는 올림말용어($d \in D$, D: 사전의 전체 용어모임), R는 용어들사이의 관계모임($R = \{H_j | j = 0, 1, 2\}$, H_0: 동의어관계, H_1: 하위어관계, H_2: 상위어관계(여기서 동의어관계는 동의어사전을 참고하여 설정하며 상위어와 하위어관계는 의미계층사전을 참고하여 설정)), S는 분야(분야는 사회과학과 자연과학, 경계과학에서의 부문별 분야로 설정), P는 기타이다.

용어사전에서는 단어들의 반복관계를 피하기 위해 자모순으로 정돈한다.

2. 표준실마리어사전작성방법

사용자들의 요구를 충족시키는 자료들에 대한 검색성능을 높이기 위해서는 표준실마리어사전을 잘 작성하는것이 무엇보다 중요하다.

1) 표준실마리어사전작성을 위한 요구

① 사회과학과 자연과학에 대한 연구과정에 리용하는 사전으로서 많은 용어들과 그것들의 의미－론리적관계를 반영하여야 한다.

② 용어들을 자모순으로 배렬하고 용어들사이의 론리적관계를 밝혀주어야 한다.

③ 과학기술의 발전과 함께 부단히 수정, 보충되여야 한다.

④ 사전구조를 동적으로 변경하여 문헌자료기지의 검색효률을 높일수 있어야 한다.

2) 표준실마리어의 선정

언어생활에서 리용되는 모든 용어들을 그대로 실마리어로 리용하는것은 합리적이지 못하다.

그러므로 현재까지 리용된 용어들의 사용빈도, 비표준실마리어들에 대한 확정 등을 리용하여 표준실마리어를 선정한다.

문서를 이루는 매개 용어에는 문서에서 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 tf_{td}로 표시한다.

문서 d에 대하여 무게값모임은 하나의 벡토르로 표현되며 해당 벡토르의 요소들은 개별적인 용어들에 대응된다. 이때 문서에서 용어들의 출현순서는 무게에 영향을 주지 않으며 같은 의미를 여러가지로 표현한 문장이나 문서들에 대한 벡토르는 류사하게 표시된다.

용어빈도수만을 리용하는 경우 모든 용어들에 대한 중요도는 정확히 표현되지 않는다.

실례로 프로그람작성언어와 관련된 문서집합에서 《프로그람》이라는 단어는 거의 모든 문서에 들어있고 《론문》이라는 단어는 모든 과학소론문들에 언급되지만 이러한 단어들은 문장의 사상을 반영하는 중요한 용어라고 말할수 없다.

그러므로 문서들에서 너무 많이 출현하는 용어들의 상관성결정에 미치는 영향을 일정한 정도로 줄이기 위한 방식으로서 높은 무게(tf_{td})값을 가진 용어들의 무게를 낮추는 방법을 리용한다.

이로부터 문서모임에서 용어 t를 포함하는 문서들의 수를 문서빈도수(df_t)라고 부르고 주어진 문서집합에서 많이 출현하는 용어들의 중요도를 평가하는데 리용할수 있다.

일반적으로 문서집합에 들어있는 문서들의 총수를 N이라고 할 때 용어 t의 거꿀문서빈도수를 다음과 같이 정의한다.

$$idf_t = \log \frac{N}{df_t}$$

거꿀문서빈도수는 드물게 출현하는 용어들에서는 높은 값으로 설정되지만 자주 출현하는 용어들에 대해서는 그 값이 낮아진다.

문서에서 나타나는 매 용어들에 대하여 혼합무게값을 부여하기 위하여 용어빈도수와 거꿀문서빈도수를 결합하여 리용한다.

문서 d에서 용어 t에 할당되는 무게값을 다음과 같이 설정한다.

$$tf - idf_{td} = tf_{td} \times idf_t$$

$tf - idf_{td}$ 값은 다음과 같은 특성을 가진다.

① t가 적은 수의 문서들에서 많이 나타날 때 제일 높은 값을 가진다.

② 용어가 문서에서 거의 나타나지 않거나 수많은 문서들에서 나타날 때에는 그 값이 작아진다.

③ 용어가 가상적으로 모든 문서들에서 나타날 때 제일 낮은 값을 가진다.

매개 문서를 하나의 벡토르로 보고 벡토르의 개별적인 성분값으로는 용어들의 $tf-idf_{td}$ 값에 따르는 무게값을 설정한다.

다음 무게에 대한 턱값(T)을 설정하고 턱값보다 큰 무게를 가지는 용어들을 표준실마리어로 선정한다. 즉 관계식

$$tf-idf_{td}>T$$

가 성립하는 용어들을 선정한다.

3) 표준실마리어들사이의 관계

표준실마리어들이 확증된 다음의 공정은 실마리어들사이관계를 얻는것이다.

최근시기 새로운 경계과학들이 출현하면서 실마리어들에 대응한 《분야》항목도 증가하고있다.

따라서 《분야》항목을 동적으로 추가한다.

또한 실마리어들사이의 상위어, 하위어들을 설정하여야 한다. 여기서는 의미계층사전을 통하여 실마리어의 상위어와 하위어들을 설정하는것으로 한다.

올림말용어들의 기록구조에 기초한 표준실마리어사전작성단계는 다음과 같다.

① 올림말을 입력한다. 여기서는 일상생활에서 리용되는 단어들은 올림말로 설정하지 않는것을 원칙으로 하며 《올림말설정프로그람》을 리용하여 올림말을 설정한다.

② 올림말의 기록구조를 작성한다.

③ 표준실마리어사전에 등록한다. 여기서는 표준실마리어사전에 등록되여있는 올림말과의 비교를 진행하며 자모순으로 등록되여있는것을 고려하여 순차적탐색방법을 리용한다.

3. 표준실마리어사전을 리용한 자료검색의 정확도평가

론문에서 취급한 표준실마리어사전을 리용하여 자료검색을 진행한 결과는 다음과 같다.(표)

표. 자료검색의 정확도비교

구 분	선행체계	제안체계
완전률/%	81.00	85.75
정확률/%	89.21	93.55
F값/%	84.90	89.48
오유률/%	25.3	14.58

표에서 보는바와 같이 론문에서 취급한 표준실마리어사전을 리용한 검색체계에서는 완전률과 정확률, F값이 높이 평가되고 오유률은 14.58로서 낮게 평가되였으며 자료검색의 성능이 높아졌다는것을 알수 있다.

맺 는 말

표준실마리어사전은 용어들을 단순히 자모순으로 배렬하고 뜻을 풀이하는 일반언어사전과는 다른 특성을 가지고있다. 론문에서는 우리 나라 올림말용어들의 기록구조를 분석한데 기초하여 표준실마리어사전을 작성하는 한가지 방법을 제안하고 자료검색의 정확도를 높이였다.

참 고 문 헌

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

[2] Paul Piwek, Natural Language Process, 1, 2, 12, 2010.

주체109(2020)년 2월 5일 원고접수

A Method for Standard Keyword Dictionary Creation Based on Record Structure of Term

Ri Myong Il

In this paper, we considered the feature of keyword dictionary and proposed a method for standard keyword dictionary creation based on record structure of term.

Keywords: information retrieval, database, dictionary creation