김일성종합대학학보
(자연과학)
주체105(2016)년 제62권 제11호

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

웨브페지열람을 위한 웨브페지처리의 한가지 방법

문 명 옥

경애하는 **김정은**동지께서는 다음과 같이 말씀하시였다.

《인민경제의 현대화, 정보화를 다그쳐 나라의 경제를 지식경제로 전환시켜야 합니다.》(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》 단행본 47페지)

인터네트가 인간생활과 뗄수없이 밀접히 련관되여있는 오늘 웨브페지는 중요한 정보원천으로서 각이한 목적을 위한 정보의 축적과 교류에 널리 리용되고있다.

한편 웨브페지의 대부분은 문자정보이므로 웨브페지의 검색에 벡토르모형과 같은 본문자료를 대상으로 하는 정보검색기술이 적용되고있다.[1]

일반적으로 사용자는 웨브페지를 읽어보면서 자기가 요구하는 정보를 담고있는 부분을 찾는 방법으로 웨브페지를 열람한다.

사용자가 대상하는 문서가 웨브페지일 때 다음과 같은 특징을 가진다.

① 웨브페지는 보통 1개 이상의 정보내용을 포함하고있는 경우가 많다. 이런 경우 사용자는 웨브페지의 처음부터 자기가 요구하는 정보를 찾을 때까지 읽어야 한다. 이러한 부족점을 해소하자면 웨브페지를 개별적인 정보내용을 담고있는 부분들로 분할한 다음 사용자가 요구하는 정보내용을 담고있는 부분을 표식해주어야 한다.

② 동일한 정보내용이 여러 부분으로 갈라져있을수 있다. 이런 경우 사용자가 찾으려는 정보가 웨브페지에는 있지만 정보가 갈라져있는것으로 하여 사용자의 정보탐색이 실패할수도 있다. 이 문제를 해결하자면 같은 내용을 담고있는 부분들을 표식하여 사용자에게 제공하여야 한다.

이상과 같은 특징으로 하여 사용자가 필요한 정보를 모두 정확히 찾을수 있다고 담보할수는 없다. 또한 웨브페지에 대한 사용자들의 정보요구는 개별적인 단어나 단어렬, 문장, 본문으로 충족될수 있으므로 웨브페지를 개별적인 정보를 담고있는 부분(블로크)들로 분할하여 제시하는것이 편리하다.

론문에서는 웨브페지의 분할에 기초하여 사용자의 정보탐색을 지원하기 위한 방법을 제기한다. 그 방법을 보면 다음과 같다.

웨브페지의 첫번째 특징에 대처하기 위하여 정보블로크분할방법[2]으로 웨브페지를 개별적인 정보내용을 포함하고있는 블로크들로 분할한다.

블로크는 웨브페지에서 개별적인 정보내용을 포함하고있는 부분이며 블로크의 정보내용은 영어나 조선어 등의 문자로 표현된다.

이로부터 웨브페지의 두번째 특징에 대처하기 위하여 검색체계의 벡토르공간모형을 리용한다. 벡토르공간모형은 블로크렬에서 같은 내용을 담고있는 블로크들을 찾을 때와 웨브페지

가 포함하고있는 여러 내용들중에서 사용자가 요구하는 정보의 위치를 표시하는데 적용한다.

웨브페지의 블로크분할과 벡토르공간모형에 기초한 처리는 개별적인 웨브페지를 단위로 하여 다음의 알고리듬에 따라 진행한다.

① 검색결과로 얻어진 웨브페지를 정보블로크분할방법[2]에 따라 분할하여 블로크렬 $B_1, B_2, \cdots, B_N$ 을 얻는다.

② 모든 $B_i(1 \le i \le N)$ 에 대하여 다음의 처리를 진행한다.

먼저 블로크에 포함되여있는 본문을 추출하고 본문으로부터 색인어[1]를 추출한 다음 벡토르공간모형[1]에 기초하여 색인을 진행한다.

③ 모든 $B_i(1 \le i \le N)$ 에 대하여 다음의 류사블로크찾기를 진행하여 류사블로크쌍 렬들의 모임 BS 를 얻는다. 이를 위해 먼저 B_i 를 하나의 질문으로 하여 질문벡토르 q 를 만든 다음 ②에서 구축한 벡토르공간모형에서 질문하여 가장 류사한 블로크 B_i' 를 찾고 류사도값이 γ 이상이면 류사블로크쌍 (B_i, B_i') 를 만들고 $BS = BS \cup \{(B_i, B_i')\}$ 로 한다.

④ 블로크무리짓기를 다음과 같이 진행한다.

$BS = \phi$ 이면 ⑤로 간다. 그리고 $(B_i, B_i') \in BS$ 에 대하여 블로크 B_i' 를 쌍의 오른쪽에 포함하는 모든 쌍들의 모임 PB 를 얻고 $PB = PB \cup \{(B_i, B_i')\}$ 로 한다.

다음 $\forall p \in PB$ 에 대하여 P 의 왼쪽에 있는 블로크들을 모두 모아 하나의 무리로 하고 $BS = BS \setminus PB$ 로 한 다음 ④로 간다.

⑤ 블로크무리 $BC_1, BC_2, \cdots, BC_M$ 의 매 무리 BC_i 에 대하여 다음의 처리를 진행한다.

우선 BC_i 의 블로크들을 웨브페지에서 출현하는 순서대로 련결하여 하나의 블로크로 만들거나 HTML꼬리표를 리용하여 표식한다.

다음 블로크의 본문내용을 벡토르공간모형으로 색인한다.

⑥ 사용자의 질문렬을 ⑤에서 만든 벡토르공간모형에 대한 질문으로 하여 류사도가 큰 순서로 블로크의 본문들을 출력하거나 표식한다.

알고리듬에서는 블로크분할을 통하여 웨브페지를 개별적인 정보내용을 포함하는 부분들로 분할하고 동일한 정보내용을 포함하는 블로크무리를 형성하여 갈라져있는 공통의 정보를 하나로 모은 다음 사용자질문과의 류사도를 계산하고 가장 류사한 순서로 무리들을 배렬(또는 표식)하여 사용자에게 제공한다.

류사블로크쌍을 형성할 때 γ 는 림계값으로서 류사도가 γ 보다 높은 쌍만을 하나의 무리로 형성한다.

벡토르공간모형에서 블로크가 가장 류사한것으로 선택되였다고 해도 류사도가 너무 작을수 있기때문에 론문에서는 림계값을 통하여 무리정보의 일관성을 높이도록 한다.

맺 는 말

열람대상인 웨브페지들을 개별적인 정보내용을 포함하고있는 부분들로 분할하고 사용자의 정보요구에 적합한 부분들을 추출하여 사용자에게 제시함으로써 정보검색의 편리성을 높일수 있게 하였다.

참 고 문 헌

[1] 문명옥 등; 정보기술, 2, 38, 주체104(2015).

[2] Ricardo Baeza Yates et al.; Modern Information Retrieval, ACM Press, 345～468, 2010.

주체105(2016)년 7월 5일 원고접수

A Method of Processing a Web Page for Web Browsing

Mun Myong Ok

We proposed one way for enhancing the convenience of information retrieval by splitting the browsed web pages into parts that included individual information pieces and presenting the extraction of proper parts according to the information request of users.

Key words: information retrieval, web page, web browsing