웹 환경에서의 홈페이지 검색 시스템

장중식°, 박의규, 나동렬, *장명길
연세대학교 전산학과
*한국전자통신연구소
[lunch,ekpark,dyra]@magics.yonsei.ac.kr, *mgjang@etri.re.kr

Homepage Retrieval System in Web Environment

Jung-Sik Jang, Eui-Kyu Park, Dong-Yul Ra, *Myung-Gil Jang
Dept. of Computer Science, Yonsei University
*Electronics and Telecommunications Research Institute

요 약

최근 웹 환경은 홈페이지 단위로 구축되는 사례가 보편화 되어 있으며, 사용자가 단순한 웹 문서가 아닌 홈페이지를 요구하는 경우도 빈번하다. 그러나 기존의 웹 환경 검색 시스템의 결과는 이러한 질의에 대한 결과로는 적절하지 않기 때문에, 본 논문에서는 홈페이지 검색을 위한 새로운 방법을 제시한다. 웹 문서 검색을 위하여 먼저 기존 검색 방법을 이용하여 결과를 얻은 후 웹 문서에 포함된 링크가 주는 정보를 추가하여 결과를 확장하는 두 가지 방법을 제시한다. 확장된 결과에서 홈페이지의 엔트리 포인트에 해당하는 웹 문서를 출력 리스트의 상위에 두기 위한 순위 재조정 알고리즘을 소개한다.

1. 서론

최근 웹 환경에 산재해 있는 정보의 양이 방대해짐에 따라, 웹 환경에서의 정보 검색 시스템에 대한 요구가 다양해 지고 있다. 인터넷으로 대변되는 웹 환경에서의 웹 문서들은 홈페이지의 한 부분으로 존재하여, 홈페이지에서 다루는 주제에 대한 일부분을 설명하는 형식을 취하는 추세이다. 이러한 사실을 바탕으로, 사용자는 홈페이지에 포함된 하나의 웹 문서를 요구하는 것이 아니라, 홈페이지의 엔트리 포인트를 요구하기도 하는데, 이러한 요구에 대한 답은 질의와의 유사도(similarity)만으로 결정되기에는 부족한 점이 많다[1].

홈페이지를 구성하는 웹 문서들은 링크를 이용하여 연결되어 있다. 따라서 사용자가 홈페이지에 접근한 이후에는 그 이하 상세 정보에 대한 검색은 홈페이지 구성자의 분류 및 배치 방법을 따르게 되므로, 홈페이지 검색 시스템은 유사도에 의해서 내림차순으로 정렬된 문서집합이 아니라, 홈페이지의 첫 문서 즉 엔트리 포인트가 상위에 위치한 결과집합을 제공하여야 한다.

본 논문에서는 홈페이지 검색 시스템의 전체적인 구조를 알아보고, 링크 정보를 이용하는 방법과 엔트리 포인트 추출을 위한 URL기반 순위 재조정 알고리즘을 을 소개한다.

2. 시스템 개괄

정보 검색 시스템은 문서 집단을 대표하는 색인구조를 저장하는 색인 서브 시스템, 그리고 사용자 질의에 연관된 문서를 검색하는 검색 서브 시스템으로 구성된다. 색인 시스템은 검색의 대상이 되는 웹 문서를 입력으로 받아 각 문서의 정보를 색인구조에 저장하고, 검색 시스템은 사용자 질의를 분석하여 색인구조를 탐색하여 결과를 출력한다.

2.1 전체적인 구성

색인 및 검색 서브 시스템은 기능적으로 세분화된 단계를 포함하고 있다.

색인 시스템은 (1)색인 과정의 일관성을 유지하기 위한 입력 문서 전처리, (2)태그(HTML tag)를 고려한 텍스트 및 링크 추출, (3)추출된 텍스트의 언어 처리를 통한 색인어 추출, (4)문서 정보를 색인구조에 저장하는

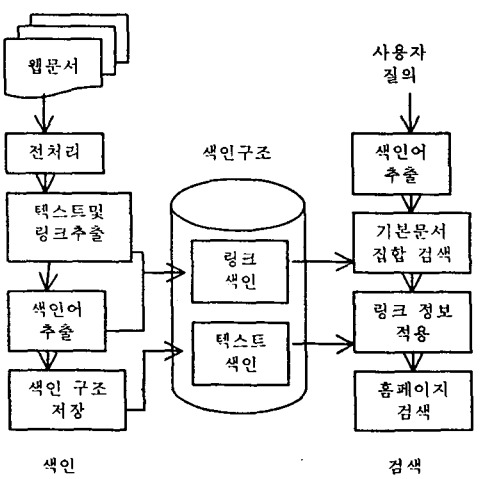

그림 1 : 색인 및 검색 시스템 구성도

단계로 구성되어 있다. 검색 시스템은 (1)입력된 사용자 질의의 언어 처리 및 색인어 추출, (2)색인구조에 저장되어 있는 전체 문서 집합을 대상으로 질의와 연관된 문서인 기본 결과집합(base set) 검색, (3)링크 정보의 적용, (4)홈페이지의 엔트리 포인트 검색을 위한 순위 재조정의 단계로 구성되어 있다(그림 1).

2.2 문서 집단

본 연구에서는 TREC 학술대회의 Web track에서 제공하는 문서 집합을 이용하였다. 각 문서는 HTML화일로서 웹에서 추출하여 화일로 저장된 형태이며 문서이름(document id) 및 URL 등을 넣은 형태이다[1].

웹 문서는 통상적으로 HTML을 기반으로 저장되어 있다. HTML은 이미지와 같은 객체들을 HTML 내부에 포함시키지 않고 위치(URL)로 참조하기 때문에 실제 사용되는 HTML 문서의 크기는 통상적으로 수십 또는 수백 Kbyte 정도를 넘지 않는다. 이러한 작은 크기의 HTML 문서를 개별적으로 관리하는 것은 대량의 문서를 다루어야 하는 정보 검색 시스템의 관점에서는 쉬운 일이 아니다. 따라서 TREC의 웹 데이타에서는 입력 문서에 해당하는 작은 HTML 문서 다수를 일정 크기의 화일 단위로 묶어 제공한다.

같은 화일에 포함된 각 문서들을 구별하기 위해서 HTML에서 사용되지 않는, 임의로 정의한 확장 태그를 사용한다. 또한 색인된 전체 문서 집단에서 문서를 구별하기 위하여 식별자를 추가로 부여하였다.

2.3 입력 및 출력

홈페이지를 검색하기 위한 사용자의 질의는 기관이나 단체의 이름등과 같은 고유명사가 포함된 경우가 많다. 예를 들어 'Hunt Memorial Library'라는 문자열은 실제로 사용되는 홈페이지 검색의 질의문이다. 이러한 문자열이 검색 시스템의 입력으로 주어진다.

위의 질의에 대한 홈페이지는 다수의 웹 문서로 이루어져 있고, 서로가 링크로 연결되어 있다고 가정하자(그림 2). 홈페이지를 이루는 웹 문서들은 질의와의 유사도를 가지게 되지만, 사용자가 요구하는 결과는 유사도가 높은 문서가 아니라 홈페이지 구조에서 최상위 문서인 엔트리 포인트(entry point)이다. 이상적인 출력은 질의의 홈페이지로 간주되는 엔트리 포인트 하나 또는 소수로 이루어진 문서 리스트이며, 각 문서는 URL로 표현하여 사용자가 접근하기 용이하게 하는 것이다.

3. 링크 정보를 포함한 색인

링크는 문서 사이의 관계에 대한 정보를 포함하고 있다. 개념적으로, 링크에 사용된 텍스트는 그 대상이 되는 문서를 외부에서 표현하는데 사용된다. 따라서 이러한 정보는 웹 문서를 대상으로 하는 웹 기반 검색 시스템에서 유용하게 사용되어질 수 있다[2,3].

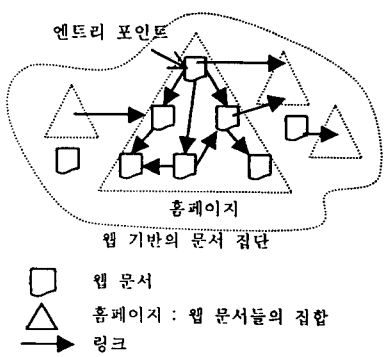

그림 2 : 웹 기반에서의 문서 집단

3.1 문서의 색인어 및 링크 정보 추출

웹 문서는 문서의 표현 방법을 기술하는 태그와 다른 문서를 참조하도록 하는 링크 태그 즉 앵커(anchor) 태그들이 문서 텍스트와 혼합되어 있는 형태이다. 앵커 태그는 하나의 문서 내에서 특정 단어 또는 단문이 다른 문서를 참조하도록 하는 것으로써, 하이퍼 링크(hyper link)를 적용하는 태그이다. 이때 링크가 적용된 단어 또는 단문을 링크 텍스트(link text)라고 하고, 링크에 의해 참조되는 문서를 참조 문서라고 한다.

HTML에서는 문서나 객체의 위치를 URL로 지정하므로 링크에서 참조 문서는 URL의 형태로 표기된다. 태그는 문서의 각 요소들(문장, 그림, 표)의 표현 방법을 수식하는 역할을 담당하기 때문에 실제 색인에 사용될 텍스트에는 포함되지 않아야 한다. 따라서 태그와 텍스트를 분리하는 과정이 선행되어지며, 링크도 태그의 형태로 표현되므로 이 과정에서 함께 추출된다.

태그를 제거하며 텍스트를 수집하는 과정을 수행할 때, 색인 시스템의 부하를 줄이고, 색인 구조에 저장되는 문서 정보의 양을 줄이기 위하여 특정 태그의 섹션에 위치한 텍스트만 추출하는 방법도 있다. 예를 들어 '<TITLE>' 태그의 섹션에 위치한 텍스트들은 문서를 잘 표현한다. 이와 비슷하게 '<H1>'이나 '<H2>'와 같이 단락의 제목이나 중요한 단어의 강조에 사용되는 태그들의 섹션에 포함된 텍스트들만 색인에 참여 시키기도 하나, 본 시스템에서는 태그를 제외한 모든 문서 텍스트를 색인의 대상으로 정의하였다.

추출된 텍스트는 언어 처리부를 거치며 색인어들의 나열로 변환된다. 이 과정에서 불용어 처리 및 스태밍 등의 부분적인 알고리즘을 선택적으로 적용하기도 한다.

링크 추출의 단계에서는 링크 텍스트와 참조 문서 URL을 추출한다. 링크 텍스트로부터도 문서 텍스트의 경우와 동일한 과정을 거쳐 색인어가 추출된다. 이렇게 추출된 링크 텍스트의 색인어들은 검색시 링크 정보를 이용하기 위하여 별도로 저장한다.

3.2 색인 구조

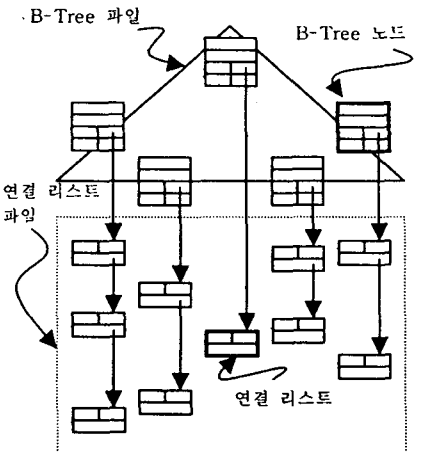

그림 3 : 텍스트 색인구조

색인 구조는 색인 시스템의 성능뿐만 아니라 검색 시스템의 성능에도 영향을 미치므로 입력 문서 집단의 형태와 검색 알고리즘, 검색 결과 출력 등에 용이한 구조를 유지해야 한다. 본 연구는 웹 문서 기반 검색을 주목적으로 하고 있기 때문에 링크에 대한 정보를 저장할 별도의 구조까지 포함한 색인 구조를 가진다.

색인어는 문헌 빈도수(document frequency)와 함께 B-tree에 저장되어 있다. 이 색인어가 나타난 문서들은 연결리스트에 저장되는데 이 연결 리스트의 헤드에 대한 포인터는 색인어와 함께 B-tree에 저장되어 있다. 즉 색인어가 출현한 문서들에 대한 정보는 B-tree에서 이 색인어의 탐색을 통해서 가능하다. 색인어에 대한 문서 연결리스트의 각 노드에는 이 색인어가 나타난 문서

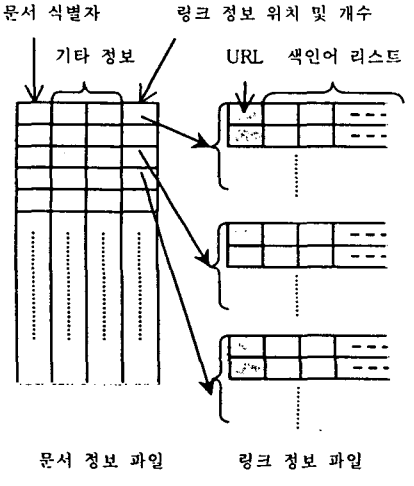

그림 4 : 링크 정보의 색인구조

의 문서번호 및 그 문서에서 나타난 그 색인어의 빈도수(term frequency)를 담고 있다(그림 3).

웹 문서에 포함되어 있는 링크는 문서 내에서 링크가 적용된 단어 혹은 구문인 링크 텍스트와 링크가 참조하는 대상 문서의 URL로 구성되어 있다. (그림 4)에서 볼 수 있듯이 한 문서에는 다수의 링크가 포함되어 있으며, 각각의 링크의 링크 텍스트는 색인어 추출기를 거쳐 얻어진 색인어 리스트로 대체되어 저장된다.

4. 검색 시스템

색인의 과정을 거친 문서들은 검색에 용이한 형태로 요약되어 색인 구조에 저장되어 있다. 검색 시스템은 이러한 색인 구조를 탐색하여 사용자 질의와 연관이 있는 문서들을 수집하고 정렬하는 과정을 수행한다. 문서와 질의의 연관과 비연관을 판단하거나 연관 정도(유사도)를 계산하는 여러 가지 검색 모델들이 연구되어 왔다. 이 중에서 전형적인 검색 모델인 벡터 모델은 문서와 질의간의 연관 정도를 0과 1사이 값으로 계산해 내는데, 연관 정도가 높은 문서를 검색결과의 상위에 위치시킬 수 있기 때문에 널리 사용된다.

본 연구에서는 벡터 검색 모델을 이용하여 수집된 결과 문서 집합에 링크 정보를 추가로 적용한다. 링크 정보의 적용 방법은 여러 가지가 발표되었다. 벡터 검색 모델의 결과 집합을 기본검색 결과집합(줄여서 기본집합)이라 부른다. 이것은 링크 정보를 적용하여 결과를 확장하거나 홈페이지 엔트리 포인트 검색을 위해 순위를 재조정하는 데 이용된다.

4.1 기본집합 검색

벡터 검색 모델은 질의나 문서의 색인어에 가중치를 부여하여 질의에 대한 문서의 유사도(similarity)를 계산한다[7]. 벡터 모델에서 문서 d_j와 질의 q는 다음과 같이 벡터로 표현된다.

$$d_j = (w_{1,j}, w_{2,j}, w_{3,j}, \ldots, w_{t,j})$$

$$q = (q_1, q_2, q_3, \ldots, q_t)$$

$w_{i,j}$는 색인어 k_i와 문서 d_j사이의 가중치이고, q_i는 색인어 k_i와 질의 사이의 가중치를 의미한다. t는 시스템 내의 전체 색인어 수를 나타낸다.

문서 d_j와 질의 q는 t차원의 벡터로 표현되며, 문서와 질의의 유사도는 두 벡터의 상관도로 구할 수 있다. 이 상관도의 예로 두 벡터간 사이각의 코사인 값으로 다음과 같이 정량화할 수 있다.

$$Sim\ (d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} \quad (1)$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} q_i^2}}$$

시스템 내의 총 문서 수를 N, 색인어 k_i가 출현한 문서 수(document frequency ; df)를 n_i라고 하고, 문서 d_j에서의 색인어 k_i의 출현 빈도수(용어 빈도수, term frequency : tf)를 $freq_{i,j}$라고 할 때, 가중치 $w_{i,j}$는 다음과 같다.

$$w_{i,j} = f_{i,j} \times \log\frac{N}{n_i} \quad (2)$$

이 때, $f_{i,j}$는 문서 d_j에서의 용어 k_i의 정규화 빈도라고 하며, 다음과 같다.

$$f_{i,j} = \frac{freq_{i,j}}{freq_{\max,j}} \quad (3)$$

max는 문서 d_j에서 용어 빈도수가 가장 큰 색인어를 의미한다. 같은 방법으로 질의 q에서의 색인어 k_i의 가중치 q_i는 다음과 같다.

$$q_i = (0.5 + \frac{0.5 freq_{i,q}}{freq_{\max q}}) \times \log\frac{N}{n_i} \quad (4)$$

(단 $freq_{i,q}$는 질의 q에서의 색인어 k_i의 빈도수, $freq_{\max,q}$는 질의 q에서 가장 큰 빈도수)

4.2 링크 정보를 이용한 검색 방법

앞 절에서 살펴본 벡터 모델에 의해서 기본집합(base set; BS)이 얻어진다. 이 집합은 그 자체로도 일반 검색의 결과가 될 수 있으나 웹 문서 검색에 적합한 결과를 얻기 위하여 링크 정보를 추가적으로 적용한다[4,5,6].

우선, 홈페이지에 있는 웹 문서들은 서로 링크에 의해서 접근할 수 있도록 구성되어 있다. 따라서 기본집합에 속하지 않은 홈페이지 구성 문서들을 포함시키기 위해서 링크 정보를 이용하여 기본집합을 확장시키고, 이렇게 얻어진 집합을 확장집합이라고 한다[2,3].

(그림 5)와 같은 상황에서 링크 $l_{i,j}$는 링크 텍스트 $lt_{i,j}$를 앵커로 하여, 문서 d_j를 가리킨다. 이때, d_j의 검색에 링크 텍스트 $lt_{i,j}$가 기여하도록 하는 것이 본 연구의 링크 정보의 이용에 대한 주요 개념이다.

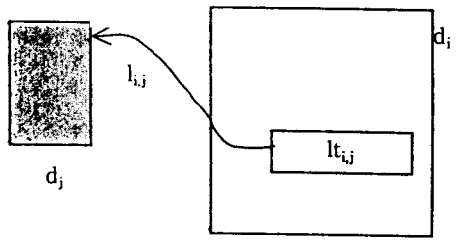

그림 5 : incoming link의 검색에 대한 기여

문서 d_i에 포함된 링크는 링크 텍스트 $lt_{i,j}$와 대상 문서 d_j를 연결하고, 이를 이용하여 확장집합(ES)을 다음과 같이 정의한다.

$$ES = \{ d_j \mid d_i \in BS,\ Sim(lt_{i,j}, q) \neq 0 \} \cup BS$$

즉, 확장집합 ES는 기본집합 BS에 속한 모든 문서 d_j에서, 그 안의 링크들의 링크 텍스트와 질의가 연관이 있을 때, 그 링크의 대상 문서 d_j을 확장 문서에 포함시킨다. 그러나 새로이 포함되는 문서가 이미 기본집합에 존재할 때는 다음에 소개되는 순위값 변경에는 이용을 하나, 기본집합 확장 시에는 중복 시키지 않으며, 새로이 추가된 문서에서 추출된 링크에는 위와 같은 방법을 적용하지 않는다.

기본집합의 확장과 동시에, 링크 정보에 의해서 문서의 유사도, 즉 순위값을 변경시켜 출력 순서에 변화를 주게 되는데, 이 때 사용하는 순위값 변경 방법 두 가지를 소개한다.

4.2.1 방법 1

첫번째 방법은 기본 유사도에 값을 추가해 나가는 방법이다. 확장집합을 구하는 방법에서 링크 텍스트와 질의간의 유사도를 계산하게 되는데, 이 값을 대상 문서와 질의와의 유사도에 추가하여 순위값을 산출한다.

$$PSV(d_j) = Sim(d_j, q) + \sum_i (\sum_k Sim(lt_{i,j}^k))$$

($lt_{i,j}^k$: d_i에서 d_j로 연결되는 k번째 링크의 텍스트)

4.2.2 방법 2

두번째 방법은, 링크 정보를 추가하여 문서와 질의 사이의 유사도를 재계산 하는 방법이다. 링크 텍스트에 속한 색인어를 마치 링크가 가리키는 대상 문서의 색인어인 것처럼 참여 시켜 유사도를 변경하는 방법으로, 첫번째 방법과는 달리 0과 1사이의 형태를 유지한다.

링크 텍스트에 의해서 문서 d_j에 색인어 k_i가 p번 추가되었다고 가정하면, 용어 빈도수 $freq_{i,j}$는 p만큼 증가되어 식 (3)에 대입된다. 또한 식 (2)에 의해서 색인어 k_i에 대한 가중치가 재계산 되며, 변경된 가중치를 바탕

으로 식 (1)에 의해 유사도를 재계산한다.

소개된 두 가지 방법은, 링크 텍스트를 이용하여 기본 검색 모델에 의해 결정된 순위를 재조정 한다. 이때 사용되는 링크 텍스트는 개념적으로 그 의미에 의해서 다른 문서를 가리키게 되므로, 링크 텍스트가 질의와 연관이 있다면 링크의 대상이 되는 문서는 문서 자체의 유사도와는 별도로 추가적인 점수를 받게 되어 결과 리스트에서의 위치가 상승하는 효과가 일어나게 된다.

4.3 URL 스트링 기반 순위 재조정

링크 정보를 이용하여 검색된 결과는 웹 문서들의 집합이다. 웹 문서들은 URL로써 구분되기도 하는데, 같은 홈페이지에 포함되는 문서들은 URL의 일부분이 같은 패턴을 가지게 되는 경향이 있다. 이 사실을 바탕으로 검색 결과 집합에서 홈페이지의 엔트리 포인트를 추측하게 된다. 본 연구에서 소개하는 URL 기반 순위 재조정 방법은 다음과 같다.

단계 1 : 결과 집합 R 수집
단계 2 : 문서 $d_j \in R$의 URL 문자열 U_j에 대해서.
단계 3 : 문서 d_i ($d_i \in R, i \neq j$)의 URL 문자열 U_i와 비교
단계 4 : 만약 U_j가 U_i의 부분 문자열이면
문서 d_j의 유사도를 일정한 값만큼 증가
단계 5 : 모든 $d_i \in R$ 에 대해 단계 3,4 반복.
단계 6 : 모든 d_j에 대해 단계 1부터 반복

이 알고리즘을 적용할 결과 집합 R은 기본집합 또는 링크 정보를 적용한 확장집합 모두 가능하다. 홈페이지를 구성하는 각 웹 문서들은 계층적인 디렉토리구조를 그대로 따른다. 엔트리 포인트가 되는 웹 페이지는 이런 구조의 상위에 위치하는 경향이 있으므로 단계 4의 부분 문자열 비교를 통한 순위값 추가 방법을 사용한다. 엔트리 포인트들의 URL은 "index.htm" 또는 "index.html"과 같이 끝나는 경우가 많은데 이 부분을 제거하고 나서 위의 알고리즘을 적용한다.

5. 실험 결과

홈페이지 검색이 아닌 일반 웹 기반의 검색의 성능을 평가하기란 어려운 일이다. 방대한 양의 웹 문서 중에서 질의와 연관을 가진 집합을 정의하기가 어렵기 때문인데, 이와 비교해서 홈페이지 검색은 그 대상이 비교적 명확하다. 따라서 본 장에서는 링크 정보를 적용한 일반 웹 문서 검색의 결과는 제외시켰으며, 홈페이지 검색의 결과만을 다루도록 하겠다.

일반 검색 결과는 질의와의 연관성이 존재하면 그 자체로 정답으로 간주 할 수 있으나, 홈페이지 검색 결과는 연관성만으로는 정답 여부를 판단하지 않으며, 검색 결과에서 정답이 나타나는 순위를 중요시하였다. 결과 집합을 상위 100개의 문서로 제한하여, 100위 이내에 정답 문서가 존재하지 않은 질의에 대해서는 검색 실패로 처리하였다.

실험은 TREC에서 제공된 웹문서 170만개 (10GB)를 대상으로 하였으며, 링크 정보를 포함하여 색인하였다[1]. 검색은 145개의 홈페이지 검색에 적합한 질의에 대해서 실행하였으며, URL 기반 순위 재조정 방법은 500위 내의 문서들에 적용하여 그 결과에서 100위까지를 최종결과로 추출하였다.

표-1은 각 방법에 대한 결과에서 100위,50위,10위 등의 순위 내에 정답이 포함된 질의의 수를 나타낸 것이다. 링크 정보는 질의에 대한 정답 문서를 100위 이내로 포함시켜주기는 하나 최상위에 위치한 정답 문서를 오히려 끌어 내리는 현상도 보이고 있다. URL 기반의 순위 재조정 알고리즘의 적용은 결과를 전체적으로 크게 향상시키고 있다.

실험 결과를 보면 URL 스트링 순위 재조정 알고리즘의 효과가 매우 좋음을 알 수 있다. 즉 1위 만을 보는 경우 4~5 배의 성능 향상을 보이고 있다. 링크 정보의 이용은 5순위 이내의 경우에는 별 효과가 없으나 10위권 이상을 보는 경우에는 결과가 좋아지는 것이 관찰되었다.

순위 내	base	Base url	Link1	Link1 url	Link2	Link2 url
100	118	127	121	138	114	136
50	109	127	111	136	103	134
10	65	124	69	126	66	127
5	55	121	55	119	53	118
1	20	95	22	81	22	77

표 1 : 각 순위 내의 정답 빈도 – base, link1, link2는 각각 base set, 링크 정보 이용 방법 1, 링크 정보 이용 방법 2를 의미하며, url은 url 기반 순위 재조정 알고리즘 적용 결과를 나타냄.

6. 결론

인터넷을 기반으로 급속히 확산되고 있는 웹 환경은 웹 문서들이 기본을 이루고 있다. 이러한 웹 문서들은 홈페이지를 단위로 구축되고 있으며, 이들 사이의 링크 정보는 중요한 의미를 내포하고 있다. 그러나 링크 정보의 다양한 적용이 실제 검색 성능을 크게 향상시키지 못한다는 연구가 발표되고 있다. 본 연구에서는 링크 정보를 홈페이지 검색의 특수한 경우에 적용해 보았으며, 더불어 홈페이지를 구성하는 웹 문서들의 URL 특성을 기반으로 하는 순위 재결정 알고리즘을 소개하였다.

링크의 이용은 큰 성능 향상을 가져오지 않는다. 그러나 순위 10~100위 사이에 대해서는 약간의 향상을 보이고 있다. URL기반 순위 재조정은 매우 효과적인 개념이라는 것이 실험을 통하여 밝혀졌다.

참고 문헌

[1] Peter Bailey, Nick Craswell and David Hawking. "Engineering a multi-purpose test collection for Web Retrieval experiments," DRAFT Accepted, subject to revision, by Information Processing and Management. Text Retrieval Conference. June 2001.

[2] Jeong-Mook Lim, Hyo-Jung Oh, Sung-Hyon Myaeng, Maann-Ho Lee. "Improving Efficiency with Document Category Information in Link-based Retrieval," IRAL Conference, 1999.

[3] Won-Kyun Joo, Sung-Hyoun Myaeng. "Improving Retrieval Effectiveness with Link Information," IRAL Conference, 1998.

[4] Kleinberg. I., "Authoritative sources in a hyperlinked environment." Proc. Of 9th ACM SIAM symposium in Discrete Algorithms. 1998.

[5] Wessel Kraij, Thijs Westervel. "TNOUT at TREC-9: How different are Web documents?" In the Ninth Text Retrieval Conference (TREC-9), National Institute for Standards and Technology, 2000.

[6] Franco Crivellari, Massimo Melucci. "Web Document Retrieval using Passage Retrieval, Connectivity Information, and Automatic Link Weighting – TREC-9 Report," In the Ninth Text Retrieval Conference (TREC-9), National Institute for Standards and Technology, 2000.

[7] G. Salton, Automatic Text Processing, Addision Wesley, 1989.