카테고리 정보 활용을 통한 링크 기반 검색의 속도 향상

임정묵, 오효정, 맹성현, 이만호
충남대학교 컴퓨터과학과

Improving Efficiency for Link-based Retrieval by using document categories

Jeong-Mook Lim, Hyo-Jung Oh, Sung-Hyon Myaeng, Mann-Ho Lee
Dept. of Computer Science, Chungnam National Univ.

요 약

본 연구에서는 링크정보를 하이퍼텍스트 문서집합에 대한 정보검색에 이용할 때 링크 카테고리 정보를 사용하여 보다 효율적으로 검색할 수 있는 알고리즘을 제안하였다. 링크정보는 문서간에 내용적으로 관련이 있음을 나타내는 중요한 정보이지만, 일반적으로 적합한 문서와 링크로 연결된 모든 문서가 사용자의 정보요구를 만족시켜 주는 것은 아니다. 따라서 질의어의 카테고리를 추측하여 적합한 문서 집합을 해당 카테고리의 링크 정보만을 이용함으로써 불필요한 계산을 줄여 검색 시간을 단축할 수 있다. 본 연구에서는 23,113건의 문서와 46개의 질의를 갖는 계몽사 자료에 대해 실험한 결과 모든 링크를 사용한 경우에 비해 카테고리 정보를 갖는 링크정보를 이용한 경우 비슷한 수준의 검색 신뢰도임에도 불구하고 검색 시간을 현저히 단축시켰으며, 이를 통해 링크 정보를 검색에 이용할 경우 링크에 의한 확장 문서집합 구성 방법이 검색 성능 개선에 중요한 요인임을 알 수 있었다.

1. 서론

이미 웹을 통한 검색 서비스는 사용자들에게 가장 친숙한 정보 검색 방법이 되었다. 사용자는 정보를 얻기 위하여 문서 집합에 대해 질의와 유사한 문서들을 제공하는 정보 검색 시스템을 이용하는데, 이러한 정보 검색 시스템은 제공하는 정보의 양이 많을수록 실제 사용자의 정보 요구 만족도를 충족 시켜 주지 못하고 있다. 일반적으로 정보 검색 시스템이 질의에 대한 결과로 제공하는 문서의 대부분은 적합성이 매우 떨어지는 문서들이기 때문에 사용자가 필요한 고급 정보를 선택하기란 그리 쉽지 않다.

이러한 문제는 검색이 대부분 문서의 색인정보만으로 이루어지기 때문이다. 정보 검색 시스템은 문서 내용을 대표하는 색인어와 질의와의 일치 정도를 사용하여 검색 여부를 결정하는데 색인어는 일반적으로 문서에 실제 출현하는 단어 중 중요한 것만 선택한 것이다. 그러나 현재 사용자에게 제공되는 디지털 문서에는 색인정보 이외에도 검색에 이용할 수 있는 다른 중요한 요소가 있다. 예를 들어 하이퍼 텍스트는 실제 문서의 내용(text)과, 다른 문서와 연결할 수 있는 링크(link)로 구성된다. 여기서 링크는 두 문서가 내용적으로 관련이 있다는 링크 생성자의 주관적인 판단에 의해 생성되므로[1][2], 색인어 기반 검색의 결과 집합 중 입력 링크의 수가 많은 문서를 좀 더 중요하게 생각할 수 있다.

링크의 이러한 특성을 이용한 검색과정을 일반화하면 다음과 같다. 색인어 기반 검색의 수행 결과를 기본집합이라고 할 때, 이 기본집합을 링크로 확장한 확장집합 내부의 문서간 링크 연결관계를 분석하여 링크가 집중적으로 가리키는 문서에 높은 가중치를 줌으로써 확장 집합의 순위 재랭킹 시 상위에 놓일 수 있도록 한다. 최근에는 색인 정보와 함께 링크 정보를 이용한 검색 방법도 연구되고 있다[1][2][3].

그러나 링크 정보를 이용한 검색방법은 시간적 비용이 많이 소요된다는 단점이 있다. 이러한 문제는 확장집합 안의 링크를 같은 비중으로 고려하므로 링크의 중요도를 계산할 때 모든 링크 종착 문서의 내용을 분석하기 때문이다. 실제로 일반적인 웹 문서의 경우, 링크에 의해 확장된 문서 집합 내에는 질의어와 관련 없는 문서들이 다수 포함될 가능성이 높다. 따라서 이러한 문서들에 대한 링크를 처리하는 데 많은 시간적 비용이 소모될 뿐 아니라 검색의 신뢰도에도 영향을 줄 수 있다.

이러한 단점을 해결하기 위해서 본 연구에서는 문서의 카테고리 정보를 사용하여 링크를 분류한 후 확장된 문서 집합에 구성할 때 질의어와 관련 없는 문서를 효율적으로 제거하는 방안을 제시한다. 이러한 방안의 효과를 판단하기 위하여 카테고리 정보를 이용한 경우와 이용하지 않은 경우의 검색 신뢰도 및 속도를 실험을 통해 확인하였다.

논문의 구성은 다음과 같다. 2절에서는 링크 정보를 검색에 활용한

기존 연구를 설명하고, 3절에서는 카테고리 정보를 갖는 링크 정보의 활용, 4절은 이러한 링크 정보를 이용한 검색 시스템을 구현하고 실험한 내용이다. 마지막으로 5절에서 결론을 맺는다.

2. 링크 정보를 이용한 정보검색

정보검색에 링크 정보를 이용하기 위하여 링크 생성은 기본적으로 다음과 같이 가정한다[1].

[가정 1] 두 문서가 링크로 연결되었을 경우, 두 문서는 서로 관련 있는 내용을 갖는다.
[가정 2] 링크로 연결된 두 문서의 저자가 서로 다른 사람일 경우 링크 생성자는 링크의 종착 문서가 내용적으로 가치가 있다고 판단한다.

이러한 가정을 기반으로 색인 정보와 링크 정보를 이용한 대략적인 검색 과정은 다음과 같다.

[단계 1] 초기 검색결과 집합을 구한다. 초기 검색결과 집합은 질의어 기반 검색의 수행 결과이며, 이 중 상위 일부문서로 제한할 수 있다. 링크에 의해 확장될 문서 집합을 기본 집합이라 한다.
[단계 2] 기본 집합이 포함하는 문서와 링크에 의해 연결된 문서까지 포함하는 확장 문서 집합을 구한다. 링크로 연결된 문서는 기본 집합에 포함된 문서의 출력 링크(Outgoing Link)에 의해 연결된 외부 문서 및 기본 집합에 포함된 문서를 가리키는 입력 링크(Incoming Link)의 시작 문서를 의미한다.
[단계 3] 링크의 가중치를 계산하여 문서를 재랭킹 한다.

링크 정보를 이용한 검색은 각 방법에 따라 링크의 종류를 분류하는 방법이 다르며, 각 링크 가중치를 계산하는 방법도 차이가 있다. 링크의 종류와 영향력의 반영 방법에 따라 다음과 같은 연구가 있다. [1][2]는 단순히 링크의 방향성만을 고려하여 입력 링크, 출력 링크로 분류하고 링크의 수 및 종착 문서에 따라 링크 가중치를 결정한다. 또한 이들은 두 문서 사이의 링크가 두 문서의 관계를 나타낸다는 가정 하에 링크를 이용하여 질의에 대한 문서의 적합성 정도를 계산하였다. 입력 링크가 많은 문서는 높은 Authority값을, 높은 Authority값을 갖는 문서에 대한 링크를 많이 포함하고 있는 문서는 높은 Hub값을 줌으로써 질의어에 적합한 문서를 검색 할 수 있음을 보였다. 특히 [2]는 실제 웹 문서 집합에서 링크 정보만을 이용했을 때 발생되는 오류를 지적하고, 링크 정보뿐만 아니라 링크로 연결된 문서의 내용을 분석함으로써 효과 있는 검색을 할 수 있음을 보였다. [3]은 링크를 방향성에 따라 입력 링크 및 출력 링크로 구분하고, 연결 형태에 따라 문서끼리 직접 연결된 직접 링크 및 직접 연결되지 않은 간접 링크로 분류한 후 각 링크의 효과를 결정한다. 실제로 182,844개의 링크를 갖는 실험집합(ETRI-kyemong)을 대상으로 링크의 방향성과 질의에 의한 직접 링크의 영향과 간접 링크의 영향을 고려한 검색 기법을 제안하였다. 또한 웹 문서 검색에서도 링크를 통한 검색 신뢰도 향상의 가능성을 보였다.

3. 카테고리 정보를 갖는 링크 정보의 활용

기존에 연구된 방법들은 모두 공통적으로 검색 시간이 많이 소요된다는 문제점이 있다. 많은 시간이 소요되는 주된 이유는 [단계 2]에서 모든 링크를 동일한 비중으로 다루기 때문에 사용자가 원하지 않는 문서가 있음에도 불구하고 링크로 연결된 모든 문서가 확장 문서집합에 포함되기 때문이다. 따라서 링크 종착 문서의 유사도, 링크의 수 등으로 링크의 가중치를 계산하는 [단계 3]의 계산량이 크게 증가된다. 실제로 일반적인 하이퍼텍스트 문서 집합에는 가정과 달리 링크에 의해 연결된 두 문서가 내용적으로 관련성이 빈약한 경우가 많이 있기 때문에 확장 문서 집합에는 사용자 질의어와 관련 없는 문서가 다수 포함될 수 있다[2].

이 방법은 사용자가 원하지 않는 문서가 확장 문서집합에 포함된에 따라 [단계 3]에서 검색에 불필요한 계산이 수행될 수 있으며, 의미 없는 문서 때문에 전체적인 검색 신뢰도가 낮아질 가능성이 있다. 즉 링크 정보만을 이용하여 확장 문서집합을 구성하는 경우, 검색 대상이 되는 문서 집합의 중심이 사용자 정보요구와는 다른 쪽으로 이동 할 수 있다. 이러한 문제는 링크 생성자의 의도와는 상관없이 자동으로 생성된 링크가 많은 일반 웹 문서의 경우 더욱 심각해질 수 있다. 따라서 링크 정보를 이용한 하이퍼텍스트 문서 검색 시 확장 문서 집합을 어떻게 효율적으로 구성하는가가 검색 신뢰도를 향상시킬 수 있는 중요한 요인이 된다.

본 연구의 목적은 문서 카테고리[1] 정보를 이용하여 확장 문서집합을 효과적으로 구성하는데 있다. 즉 초기 검색 결과집합을 이용하여 사용자가 요구하는 문서의 카테고리를 추측하고, 해당 카테고리의 링크만으로 확장한 확장 문서집합을 사용하였다. 이 방법은 기존 방법에 비해 확장 문서집합의 크기가 줄어듦으로 해서 [단계 3]의 계산량을 크게 줄일수 있고, 또한 링크에 의해 검색 대상이 되는 확장 문서집합이 확장되더라도 확장 문서집합의 중심이 사용자의 의도와 크게 벗어나지 않는 장점이 있다. 또한 데이터 집합이 일반 웹 문서처럼 링크 종착 문서의 성격을 쉽게 예상할 수 없는 상황에서 효과적으로 링크 정보를 이용할 수 있으며, 전자도서관과 같이 데이터 문서와 링크가 관리자에 의해서 관리되는 경우에는 더욱 효과적이다[4].

카테고리 정보를 이용한 링크 활용 단계는 다음과 같다.

[i] 초기 검색 결과 집합 생성

일반 정보 검색 시스템에서 제공하는 질의에 대한 초기 결과 집합을 생성하며, 초기 결과 집합내의 문서는 질의에 대한 적합도에 따라 순위를 갖는다.

$$S=\{d \mid d\in D, rel(Q,d)>0\}$$

S = 초기 검색 결과 집합
D = 데이터 문서 집합
rel(Q,d) = 질의어 Q에 대해 문서 d의 적합도

[ii] 링크 카테고리 및 확장할 기본 집합 선택

질의어에 따른 링크의 카테고리를 추측하고, 확장 문서집합을 위한 기본 집합을 구성하는 단계이다. 링크 카테고리란 링크의 의해 연결된 종착 문서의 카테고리를 의미한다. 질의어에 따른 링크 카테고리를 추측하기 위하여 S의 일부문서의 집합인 S'을 추출한다. 추출된 S'을 분석하여 사용자가 검색하고자 하는 링크 카테고리를 예상할 수 있다.

1 문서 집합내의 모든 문서에 대한 카테고리 정보가 존재 한다고 가정한다.

$$S' = \{ d \mid d \in S, rel(Q, d) > t \}$$

$$\forall\, d_i \in S'$$

$$C_i = \{ d_j \mid d_j \in S', category(d_i) = category(d_j), i < j \}$$

$$C = \underset{i}{Max} [\, weight(C_i) \,]$$

S' = 질의어에 따른 링크 카테고리를 추측하기 위한 문서집합
t = 적합성 정도의 임계치 (threshold value)
C_i = S'의 문서 중 링크 카테고리 i에 해당하는 문서집합
$weight(C_i)$ = 질의에 대한 링크 카테고리 C_i의 중요도
C = 질의어에 따라 추측된 링크 카테고리

일반적인 경우 적합도를 이용하여 결과 집합을 정렬할 경우 질의어에 따라 적합도가 매우 낮은 문서라도 검색 결과의 상위에 놓일 수 있다. 따라서 적합도의 임계치를 사용하여 링크 카테고리를 추측할 집합 S'을 선택하였다. S'을 분석하는 방법은 순차적으로 만들어진 카테고리 집합(C_i)중에 질의와 가장 가까운 카테고리를 $weight(C_i)$를 통해 선택한다. 질의에 대한 카테고리 중요도를 구하는 방법에는 여러 가지가 있을 수 있는데, 간단하게는 가장 많은 문서를 포함하는 카테고리를 선택하거나 문서의 순위에 따라 카테고리 가중치를 부여하는 방법 등이 있다

카테고리 C를 구한 후 실제로 해당 링크에 의해 확장될 기본 집합(B)을 구한다. 본 연구에서는 B를 S중 상위 n개 문서로 한다.

$$B = \{ d \mid d \in S, rank(d) \le n \}$$

B = 링크에 의해 확장될 기본 집합
rank(d) = S에서 문서 d의 적합도 순위

[iii] 확장 문서 집합 생성

확장 문서 집합(E)는 B와 링크 카테고리 C에 의해 확장된 문서 집합과의 합집합이다.

$$E = \{ d_i \mid d_i = link(d, C),\ d \in B \} \bigcup B$$

$link(d,C)$ = 문서 d로부터 링크 카테고리 C에 의해 확장된 문서

[iv] 링크 효과의 적용

이 단계는 확장 문서 집합의 링크 관계를 고려하여 문서의 순위를 재랭킹하는 단계이다. 본 연구에서는 [3]의 식을 이용한다.

4. 구현 및 실험

본 연구에서 사용한 실험 데이터 집합은 23,113개 문서가 89개의 카테고리로 분류된 계몽사 데이터 집합(ETRI-kemong)이다. 또한 182,844개의 링크를 갖고 있으며 각 링크에는 종착 문서의 카테고리 정보를 저장하였다. 실험에서는 [3]의 실험과 동일한 질의어 46개를 사용하였다.

본 실험에서 사용한 기본 집합(B)과 S의 일부 문서 집합 S', 질의에 대한 카테고리 중요도는 다음과 같이 정의하였다.

$$S' = \{ d \mid d \in S, sim(Q, d) > 0.3 \}$$

$$B = \{ d \mid d \in S, rank(d) \le 30 \}$$

$$C = \underset{i}{Max} \mid C_i \mid$$

sim(Q,d) = 질의어 Q와 문서 d의 유사도

[표 1]은 위와 같이 정의된 기본 집합(B)에서 일부 문서집합(S')을 통해 선택된 카테고리(C)로 연결되는 확장 문서(E)를 대상으로 링크 검색을 적용했을 때와 기존의 링크 검색[3]의 11-point average precision 비교한 결과이다.

	링크 검색[3]	개선된 링크 검색
확장 문서 집합에 포함된 문서 개수의 평균(개)	230	70
속도 비율	10	1
평균 정확도	0.5126	0.5026

[표 1] 11-point average precision 실험 결과

실험 결과를 분석하면 카테고리 정보를 갖는 링크 정보를 이용한 경우 [3]과 유사한 수준의 정확도임에도 불구하고 수행시간은 1/10로 단축되었음을 알 수 있다. 이러한 이유는 확장 문서집합 구성 시 사용자 질의와 관련이 없는 문서를 제외시켜(평균 70%감소) [단계 3]에서 사용하는 색인정보에 대한 통신비용과 직·간접 링크의 효과를 결정하기 위한 DLE와 IDLE[3]를 구하는 계산량을 줄였기 때문이다.

5. 결론 및 향후 연구

본 연구는 하이퍼텍스트 문서집합에 대한 검색 시 링크 정보를 이용할 때 카테고리 정보를 참조하여 보다 효과적으로 확장 문서집합을 구성하는 알고리즘을 제시하였으며, 실험 결과 기존의 링크 검색에 비해 10배 이상의 빠른 수행시간으로 비슷한 수준의 검색 효과를 얻을 수 있었다. 이와 같은 실험 결과는 링크가 문서들 사이에 내용적으로 관련이 있음을 나타내는 중요한 정보이지만, 실제 상황에서 모든 링크가 사용자의 정보요구를 만족시켜 주지 않는다는 가설을 입증한다. 이로써 확장 문서집합의 구성 방법이 링크 정보를 이용한 검색 시스템의 실용화에 있어서 중요한 요인임을 확인할 수 있었다.

향후 연구 방향으로는 검색 신뢰도를 높이기 위하여 사용자가 원하는 문서의 카테고리를 추측하는 방법에 대한 연구와 이 때 사용되는 기본 집합 및 각 변수의 최적치를 찾는 연구를 수행할 예정이다.

참고 문헌

[1] Kleinberg, J., "Authoritative sources in a hyperlinked environment." *Proc. of 9th ACM SIAM Symposium in Discrete Algorithms.* 1998.

[2] Krishna, B. & Monika, R. H., "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", *Proc. of ACM SIGIR*, 1998.

[3] Joo, W. K. & Myaeng, S. H., "Improving Retrieval Effectiveness with Link Information", *Proc. of the International Workshop on IRAL'98*, 1998.

[4] 조은일, 임정묵, 오효정, 이만호, 맹성현, "CORBA와 JAVA를 사용한 에이전트 기반 디지털 도서관 프로토타입 구현", 한국정보과학회 봄 학술대회 발표 예정, 1999