링크 구조 기반의 순위 알고리즘을 이용한 메타 검색 에이전트

김형욱[0] 김민구 최경희
아주대학교 정보통신 전문대학원
(wizard[0], minkoo, khchoi)@ajou.ac.kr

The Meta Search Agent using Ranking Algorithm with Link Structure Analysis

Hyoung-Uk Kim[0] Min-Koo Kim Kyung-Hee Choi
Dept. of Information & Communication, Ajou University

요 약

하이퍼 텍스트 구조의 특성을 이용한 순위 평가 알고리즘 중의 하나인 HITS 알고리즘은 웹 페이지들의 상호 간에 연결된 링크 정보로부터 웹 문서들의 중요도를 평가하여 순위에 따른 결과를 제시한다. 그러나 초기의 HITS 알고리즘은 문서 내의 링크 빈도 수만을 고려하고, 입력 값으로 주어지는 웹 문서 집합의 특성에 의존적인 단점을 가지고 있다. 본 논문에서는 여러 웹 검색 엔진들로부터 얻어진 문서 집합에 수정된 HITS 알고리즘을 수행하는 메타 검색 에이전트를 설계하여 보다 나은 검색 성능을 구하고, 결과의 지역성을 보완한다.

1. 서 론

이미 무한하다고 할 수 있을 만큼 웹(WWW)이 포함하는 정보의 규모는 방대하고, 나날이 새롭게 변경되거나 확장되어 가고 있다. 이러한 정보의 홍수 속에서 사용자가 요구한 정보를 단순히 나열하기 보다는 사용자의 판단에 앞서서 보다 중요한 정보를 선별하여 제공해야 할 필요성이 커지고 있다. 이를 위해 웹 환경의 근간을 이루고 있는 하이퍼 텍스트 구조를 분석하여 웹문서의 중요도 판단을 돕는 다양한 방법들이 제안되었다.

웹(WWW)을 구성하는 웹 문서에는 문서의 내용과 함께 웹 문서들 상호 간에 서로를 연결하고 있는 하이퍼 링크가 존재한다. 일반적으로 웹 문서의 작성자는 링크를 삽입함으로써, 링크가 가리키고 있는 다른 웹 문서와 자신의 웹 문서 사이의 관련성을 갖게 한다. 웹 문서의 순위 평가 방법 중 하나인 HITS 알고리즘[1]은 특정한 웹 문서가 다수의 웹 문서들로 부터의 링크를 가지고 있을 때, 보다 작은 수의 웹 문서들에게서 참조되고 있는 웹 문서보다 중요한 자료임을 가정한다.

단순히 링크의 빈도 수만으로 중요도를 판단하는 것은 순위 향상을 위한 의도적인 링크 삽입이나 서로 관련이 없는 웹 문서들 사이에서 만들어진 참조의 경우와 같은 문제점을 가지고 있다. IBM의 Clever[4]나 ARC 시스템[5]등은 이를 보완하기 위해서 링크마다 적절한 가중치를 부여하도록 HITS 알고리즘을 수정하여 구현되었다.

그러나 HITS 알고리즘은 질의에 대해 의미 기반의 검색 엔진에서 구한 상위의 웹 문서 집합을 확장하여 사용하기 때문에, 최초의 웹 문서 집합의 특성에 의존적인 한계를 갖게 된다. 예를 들어 검색 엔진이 미처 인덱싱하지 못하고 있는 중요한 문서를 빠뜨릴 수 있고, 전체 WWW 내에서 검색 엔진이 제시한 일부 웹 문서들로 연결되어 있는 지역적인 영역으로 HITS 알고리즘의 결과가 한정된다.

본 논문에서는 다수의 검색 엔진들로부터 상위로 검색된 웹 문서들을 취합하여 초기 집합을 구성한 뒤, 이를 HITS 알고리즘에 적용하여 최종적인 웹 문서들의 중요도 순위를 만드는 메타 검색 에이전트를 설계한다. 또 실험을 통하여 하나의 검색 엔진만을 이용하여 HITS 알고리즘을 수행한 결과와 구현된 메타 검색 에이전트로부터 얻어진 결과를 비교하여, HITS 알고리즘의 지역적인 특성을 보완할 수 있음을 보인다.

2. 링크 분석을 이용한 HITS 알고리즘

HITS 알고리즘은 링크를 분석하여 웹 문서에 대한 authorities와 hubs로서의 두 종류의 속성 값을 계산한다. authorities는 제시된 주제에 대해 중요한 정보를 갖고 있는 문서라는 것을 의미하며, hubs는 authorities들로 연결하는 많은 수의 링크들을 가진 문서라는 것을 의미한다. 이를 위해 웹 페이지들 상호 간에 연결된 링크의 빈도 수로부터 authorities를 결정하며, 역으로 hubs를 알 수 있다.

먼저 질의로부터 의미 기반의 검색 엔진에서 얻어진 검색 결과의 웹 문서 집합 N을 N을 HITS 알고리즘의 입력으로 사용한다. 위에서 $A[n]$는 $n \in N$ 인 웹 문서 n에 대한 authority score이고, $H[n]$는 hub score이다. 위의

연산을 반복적으로 수행하면, $A[n]$와 $H[n]$는 최종적인 authority score와 hub score로 수렴하게 되며, 이는 Kleinberg[1]에 의해 증명되었다.

$$H[n] = \sum_{(n,n')\in N} A[n']$$
$$A[n] = \sum_{(n',n)\in N} H[n'] \quad \cdots(1)$$

3. 수정된 HITS 알고리즘

위의 식과 같이 웹 문서들이 갖는 링크의 빈도 수 만으로 중요도 순위를 결정하는 것은 다음과 같은 경우들을 고려하지 못하여 예상치 못한 결과를 가져온다.

(a) 특정한 웹 문서의 순위를 올리기 위해 의도적으로 많은 수의 링크를 삽입할 수 있다.
(b) 웹 문서내에 특정 회사나 기관을 연결하는 링크가 자동적으로 삽입될 수 있다.
(c) 주제와 관련없는 다른 웹 문서로 연결하는 링크가 순위 결정에 영향을 주거나 다른 주제로의 확장을 유도할 수 있다.

(a)은 하나의 호스트에 존재하는 웹 문서들이 다른 호스트의 특정한 웹 문서로 연결하는 링크를 가질 때 상호 간에 연결하고 있는 링크들에 대한 가중치를 상대적으로 낮게 설정함으로써 해결될 수 있다. (b), (c)은 링크의 주위에 있는 anchor 텍스트나 paragraph 혹은 웹 문서 전체의 내용과 사용자의 질의와의 유사도를 구하여, 의미 기반으로 분석된 결과를 링크의 가중치에 반영함으로써 보완될 수 있다.[2,3,5]

While the vectors H and A have not converged
for all n in N,

$$A[n] = \sum_{(n',n)\in N} H[n'] \times auth\_wt(n',n)$$
$$H[n] = \sum_{(n,n')\in N} A[n'] \times hub\_wt(n,n')$$

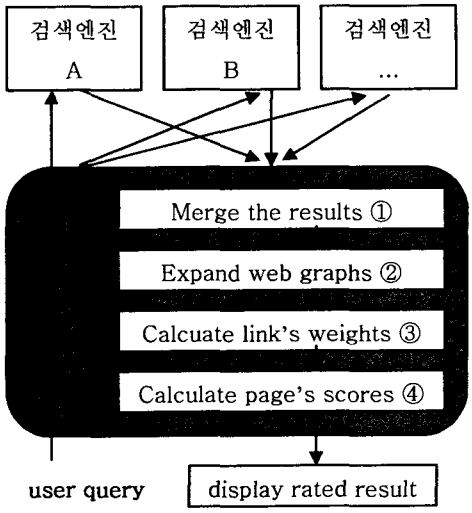
Normalize the H and A vectors $\cdots(2)$

$$auth\_wt(n',n) = \frac{1}{k} \times relevancy\_wt(n',q)$$
$$hub\_wt(n,n') = \frac{1}{l} \times relevancy\_wt(n',q)$$

수정된 HITS 알고리즘에서는 웹 문서 n과 n' 을 연결하고 있는 링크에 가중치를 부여한 뒤, authority와 hub score를 구한다. 두 호스트의 웹 문서들 사이의 링크 수로부터 보정하기 위한 인수 k는 문서 n' 이 존재하는 호스트의 다른 문서들로부터 문서 n을 연결하는 링크들의 개수이고, 인수 l는 문서 n' 이 존재하는 호스트의 다른 문서들로 문서 n이 연결하는 링크들의 개수이다. 두번째 가중치 요소로서 relevancy weight는 사용자의 질의와 웹 문서 사이의 cosine distance 식[6]으로 부터 유사도를 구하여 사용한다.

4. 메타 검색 에이전트의 설계

본 논문에서 메타 검색 에이전트는 사용자로부터 질의어를 입력받은 뒤, 야후, 알타비스타, 심마니, 구글의 4 개의 검색 엔진들에게 요청하도록 PC 환경에서 동작하는 검색 에이전트 어플리케이션으로 설계하였다.



[그림 1 메타 검색 에이전트의 구조]

① 질의에 대해 각 검색 엔진들마다 응답한 웹 문서들 중 상위 25개의 URL들을 순서를 고려하지 않고 병합하여, 총 100개의 URL을 초기 집합으로 사용한다. 검색 엔진의 결과들에서 URL이 서로 중복되는 경우 다음 순위의 URL을 대신 추가한다.
② 초기 집합의 문서가 갖는 링크의 URL과 문서를 가리키는 링크를 갖는 새로운 웹 문서들의 URL을 검색 엔진에 요청하여 받은 웹 문서들로 초기 집합을 확장한다.
③ 확장된 집합 내의 문서들 사이에서 링크에 대한 가중치를 계산한다. 문서 $x \rightarrow y$로 연결하는 링크의 가중치는 전체 100 x 100 크기를 갖는 2차원 행렬 상에서 (x, y)의 위치에 저장한다.
④ authority score와 hub score를 구한다. score가 수렴하기까지 (2)를 10회 반복 실행하였다. 최종적인 score가 0~1 범위의 값을 갖도록 정규화 과정을 거친 뒤 authority score만을 사용하여 내림차순으로 정렬된 문서들의 리스트를 사용자에게 보여준다.

5. 실험 결과 및 평가

실험을 위해 자바, 해외 여행, 인공 지능등 두 단어 이내로 만들어진 20개의 질의어를 임의로 선택하여, 직접 검색 엔진으로부터 얻어진 결과와 메타 검색 엔진을

사용한 결과를 비교하였다. 또 HITS 알고리즘에서 다른 초기 집합을 사용했을 때 구해지는 결과를 비교하기 위해 메타 검색 에이전트에 하나의 검색 엔진만을 연결했을 때와 동시에 4개의 검색 엔진을 이용하였을 때 얻어지는 결과를 비교하였다.

각각의 결과에 대한 평가를 내리기 위해 본 실험에서는 다음과 같은 두 가지 방법을 사용하였다. 첫번째로 3명의 훈련된 실험자에게 검색된 결과가 포함된 웹 문서와 질의어 사이의 관련성과 중요도에 따라 0점(관련 없음), 1점(관련 있음), 2점(관련있으며 유용함), 3점(관련있으며 매우 유용함)으로 점수를 부여하도록 하였다. 만일 질의어의 의미로부터 여러 주제를 포괄할 수 있는 경우 실험자의 주관에 따라 중요도를 판단하도록 했다. 다음으로 첫번째 방법을 사용하여 직접 검색 엔진들로부터 얻어진 결과에서 중요하다고 판단되어진 전체 문서들의 집합을 베이스로 하여 각 검색 엔진들에 대해 메타 검색 에이전트로부터 얻어진 결과들로 상대적인 재현도를 구하였다.

	상위 10개	상위 20개
야후	0.4667	0.5167
알타비스타	0.2667	0.3333
심마니	0.5000	0.4833
구글	0.7333	0.6166

[표 1 10개 질의 결과의 실험자 평가]

표 1에서 야후와 알타비스타는 상위 10개의 문서들에서 보다 상위 20개의 문서들에서의 중요도의 평균값이 더 높게 나왔다. 따라서 사용자가 높게 평가한 중요한 문서가 오히려 낮은 우선 순위를 갖고 있음을 알 수 있다. 비슷한 링크 분석 알고리즘인 PageRank로 구현된 구글 검색 엔진은 상위 10개의 문서들에서 가장 높은 평가를 받았지만, 상위 20개의 문서들에서의 중요도 평균값은 상당히 큰 폭으로 감소되었다.

	중요도 평균	상대 재현도
야후	0.5367	0.52
알타비스타	0.4233	0.43
심마니	0.5233	0.55
구글	0.6667	0.65
4개 종합	**0.6933**	**0.71**

[표 2 메타 검색 에이전트의 수행 결과]

표 2는 각각의 검색 엔진에서 얻어진 초기 집합을 이용하여 메타 검색 에이전트를 수행한 결과이다. 표2에서의 중요도 평균은 실험에 사용한 전체 질의어에 대해 상위 문서 20개에서 구해진 중요도들을 평균한 것이다. 상대 재현도는 각 실험에서 얻어진 문서들 중 질의어와 관련있다고 판단되어진 문서 집합에 상대적인 Recall을 구한 것이다. 표 1과 비교했을 때, 직접 검색 엔진에서 얻어진 결과에 수정된 HITS 알고리즘을 적용하여 사용자의 평가는 전체적으로 높아지지만 검색 엔진에서 직접 얻어진 초기 결과의 성능에 좌우되는 것을 알 수 있다. 표 1에서 가장 나은 성능을 보였던 구글이 HITS 알고리즘을 사용한 후에도 다른 3개의 검색 엔진에 비해 좋은 평가를 받았다. 그러나 하나의 검색 엔진의 결과를 사용하는 대신, 4개의 검색 결과를 종합하여 HITS 알고리즘을 적용했을 때의 메타 검색 에이전트의 결과가 가장 우수한 성능을 보였다.

6. 결론 및 향후 과제

웹 환경에서 문서들의 구조적인 정보를 이용하여 좋은 검색 결과를 보이고 있는 HITS 알고리즘은 의미 기반 검색에서 얻어진 문서들의 상호 연결 관계에 대한 수량적인 정보에만 의존하는 단점을 가지고 있다. 또한 HITS 알고리즘의 결과가 갖는 효율은 일차적으로 검색된 문서 집합들의 전체 웹 환경 내에서의 분포에 의존적이 된다.

본 논문에서는 여러 개의 검색 엔진들을 이용하여 검색된 결과에 수정된 HITS 알고리즘을 사용하여 최종적인 문서들의 순위를 결정하는 메타 검색엔진을 설계하였다. 구현된 메타 검색엔진에 질의를 제시하여 얻어진 결과를 실험자에 의한 평가와 상대 재현도를 측정하였고, 각 검색 엔진들보다 비교적 우수한 성능을 보였다.

향후 과제로는 하이퍼 텍스트의 구조 및 관계 분석과 의미 기반 분석의 방법을 결합하여 보다 좋은 성능을 가질 수 있는 검색 모델을 연구하고자 한다. 또한 순위 향상 알고리즘에서 정확한 비교 평가가 이루어질 수 있는 기법을 연구할 필요가 있다.

7. 참고 문헌

[1] J. Kleinberg, " Authoritative sources in a hyperlinked environment" IBM 1997.

[2] Krishna Bharat, R.Henzinger, " Improved Algorithms for Topic Distillation in a Hyperlinked Environment" 21th ACM SIGIR Conference, 1998

[3] Nick Craswell, David Hawking, Stephen Robertson, " Effective Site Finding using Link Anchor Information" ACM SIGIR' 01, 2001

[4] Soumen Chakrabarti, et.al," Mining the Web' s Lnk Structure" IEEE Computer,

[5] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, et.al, " Automatic resource compilation by analyzing hyperlink structure and associated text" 7th WWW Conference, 1998

[6] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, " Modern Information Retrieval" , Addison-Wesley