하이퍼링크 환경에서 정보검색을 위한 개선 알고리즘

최익규°, 김희수, 이병희, 김민구
아주대학교 정보통신전문대학원
(ikchoi, heemanx, acecult, minkoo,sparky)@madang.ajou.ac.kr

Improved Algorithms for Information Retrieval in a Hyperlinked Environment

Ik-Kyu Choi°, Hee-Soo Kim, Byoung-Hui Lee, Min-Koo Kim
Graduate School of Information and Communication, Ajou University

요 약

하이퍼링크 환경에서의 정보검색은 주로 문서에 존재하는 링크정보를 이용하여 이루어진다. 본 논문은 하나의 문서에 존재하는 여러 개의 하이퍼링크마다 연결되는 문서와의 유사성을 측정하여 차등적으로 링크의 연결정보를 부여하여 기존의 알고리즘을 개선하였고, 관련이 없는 문서로의 하이퍼링크로 인해 발생되는 topic drift 현상을 제거하기 위해 문서와 확장된 질의와의 유사성을 측정하여 문서의 가중치를 계산에 적용하도록 알고리즘을 개선하였다. 개선한 알고리즘의 성능을 확인하고자 TREC10의 web trec 부분에 적용하여 향상된 검색 결과를 얻었다.

1. 서 론

정보 검색분야의 연구는 인터넷의 발달을 계기로 활발하게 전개 되고 있다. 인터넷상에 존재하는 문서의 수는 기하급수적으로 증가하여 그 수가 이미 10억 개를 넘었고, 그 수를 정확하게 알기는 힘들어졌다. 끝없이 쌓여가고, 생성되는 문서들 속에서 자신이 원하는 정보를 찾는 것은 점점 더 중요한 일이 되었고, 이러한 요구를 만족시키고자 많은 연구가 이루어지고 있다.

인터넷은 하이퍼링크를 이용하여 서로 연결된 거대한 문서의 창고이다. 문서들끼리 연결되어 있는 링크정보를 이용하는 연구는 Kleinberg의 HITS(Hypertext-Induced Topic Search) 알고리즘[1]을 계기로 많은 연구가 이루어지고 있다. Kleinberg는 링크의 연결성을 분석하여 좋은 문서들에서 많이 연결되어 있고, 좋은 문서들을 많이 연결하고 있으면 좋은 문서로 판단하였다. 그러나 문서들 상에 있는 링크의 연결정보는 인터넷의 발달로 의미 없고 불필요한 것들이 많이 발생하였다. Bharat은 HITS에 존재하는 몇 가지 문제점을 제기하였고 이를 해결할 수 있는 알고리즘[2]을 제시하였다.

본 논문에서는 Bharat의 알고리즘을 개선하고자 문서들 사이의 유사성을 계산하여 문서에 존재하는 여러 링크들에게 차등적으로 가중치를 주는 방법을 제안하고, 관련 없는 문서로의 하이퍼링크로 인해 발생되는 topic drift 현상을 제거하기 위해 문서와 확장된 질의와의 유사성을 측정하여 문서의 가중치를 계산에 적용하도록 알고리즘을 개선하고자 한다. 또한 새롭게 제안된 알고리즘을 검증하고자 TREC10에 있는 web trec10G 자료를 이용하여 Kleinberg 알고리즘, Bharat 알고리즘과 제안된 알고리즘을 비교하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개하고, 3장에서는 관련 연구의 개선점을 제시하고, 이를 개선할 개선 알고리즘을 제시하고자 한다. 4장에서는 이를 검증하고자 TREC10의 자료를 이용하여 알고리즘을 비교하고, 5장에서는 본 논문의 결론과 향후방향을 제시한다.

2. 관련 연구들

하이퍼링크의 환경에서 문서의 링크정보에 대한 연구는 크게 전역 링크정보 연구와 지역 링크정보 연구로 나눌 수 있다. 전역 링크정보 연구는 현재 Google 검색시스템에서 사용하고 있는 PageRank 알고리즘[3]과 같이 전역의 링크정보를 이용하여 전체 문서에 가중치를 부여하는 연구이고, 지역 링크정보 연구는 Kleinberg의 HITS 알고리즘처럼 사용자의 초기 질의에 대한 기본 검색자료를 이용하여 선정된 기본 문서들과 관련된 지역적인 링크정보를 이용하는 연구이다. 본 논문에서는 지역 링크정보연구에 초점을 맞추어 연구를 진행하고 있다.

2.1 Kleinberg의 HITS 알고리즘[1]

HITS 알고리즘은 각각의 문서마다 hub값과 authority값을 계산한다. 좋은 authority 값을 가지고 있는 문서는 관련 있는 내용을 가지고 있고, 좋은 hub 값을 가지고 있는 문서는 관련 있는 문서를 연결하는 링크를 가지고 있다. 좋은 authority값을 가지고 있는 문서들에 대한 연결링크를 많이 갖는 문서는 좀더 좋은 hub 값을 가지며, 좋은 hub값을 가지고 있는 많은 문서들에 의해 연결되어지는 문서는 좋은 authority 값을 가지게 된다.

HITS 알고리즘은 사용자의 질의를 이용하여 내용검색은 수행한 검색결과에서 출발한다. 내용기반 검색결과

를 시작 집합으로 정의하고, 시작 집합에 들어있는 모든 문서들을 대상으로 문서의 링크정보를 이용하여 1 단계 확장을 통하여 HITS의 기본집합을 구한다. 이때, 문서에 들어오는 인 링크의 개수는 최대 50개로 제한하여 처리하고, 문서에서 나가는 아웃 링크의 개수에는 제한을 두지 않고 기본집합을 구한다.

HITS 알고리즘은 위에서 구해진 기본집합을 대상으로 아래와 같은 방법대로 hub 값과 authority값을 각각 구한다.

(1) Let N be the set of nodes in the neighborhood graph.
(2) For every node n in N. Let H[n] be its hub score and A[n] its authrority score.
(3) Initialize H[n] and A[n] to 1 for all n in N
(4) While the vectors H and A have not converged:
(5) For all n in N. $A[n] := \sum_{(n',n)\in N} H[n']$
(6) For all n in N. $H[n] := \sum_{(n,n')\in N} A[n']$
(7) Normalize the H and A vectors

[방법 1] Kleinberg: Kleinberg의 HITS 알고리즘

2.2 Bharat의 개선 HITS 알고리즘[2]

Bharat의 알고리즘은 Kleinberg의 HITS에 존재하는 몇 가지 문제점을 해결하고자 HITS 알고리즘에 기반을 두고 만들어졌다. Bharat은 Kleinberg의 HITS 알고리즘의 문제점을 아래와 같이 세가지로 요약하였다. 첫째, 호스트들이 서로서로 협조적일 때, 협조의 정도에 따라 특정 호스트로부터 연결되어만 있어도 특정 호스트의 수많은 연결로 인하여 높은 hub값과 authority값을 가질 수 있게 된다. 둘째, 소프트웨어 산업이 발달하여 웹 문서들을 쉽게 만들 수 있는 도구를 제공하면서, 문서 제작 도구는 특별한 의미 없이 기본적으로 문서마다 링크정보를 추가함으로 HITS의 계산을 의미 없게 만든다. 셋째, HITS의 시작집합이나 기본집합에 사용자의 질의에 의미 없는 문서가 좋은 연결성을 가지고 있다면, 가장 높은 hub 값과 authority 값이 사용자 질의와 상관없는 문서가 되는 topic drift 문제가 발생하게 된다.

Bharat은 첫번째 문제를 해결하고자 아래와 같이 HITS 알고리즘을 수정하였다. 여기에서 사용되는 auth_wt()와 hub_wt()는 하나의 문서에 같은 호스트로부터의 연결이 n개 있다면, 1/n의 값이 되어 하나의 호스트는 하나의 의견으로 비중을 낮추었다.

(4) While the vectors H and A have not converged :
(5) For all n in N,
$A[n] := \sum_{(n',n)\in N} H[n'] \times auth\_wt(n',n)$
(6) For all n in N,
$H[n] := \sum_{(n,n')\in N} A[n'] \times hub\_wt(n,n')$
(7) Normalize the H and A vectors

[방법 2] Bharat: Bharat 알고리즘

Bharat은 HITS 알고리즘의 둘째, 셋째 문제를 해결하고자 시작집합에서 상위에 랭킹된 문서들로부터 확장질의를 만들었다. 이렇게 만들어진 확장질의와 시작집합에서 연결정보를 이용하여 확장된 기본집합에 속하는 모든 문서들과 유사도를 계산하였다. 기본집합에 속하는 문서들의 유사도값들 중에 중간 값을 구하고, 중간 값 이하인 문서들은 기본집합에서 제거하고 나서 위와 같이 hub값과 authority값을 구하였다.

3. 문제점과 향상된 알고리즘 제시

3.1 문서와 확장질의의 유사도 적용

Bharat 알고리즘에서 **topic drift**문제를 해결하기 위해 확장질의를 만들고 기본집합에 속한 문서들과 유사도를 계산하여 기본집합에서 관련 없는 문서들을 제거한다. 이 부분에서 본 논문은 확장질의와 문서들 사이의 유사도 정보를 **hub**값과 **authority**값을 구하는데 사용하는 일반적인 계산식을 아래와 같이 제안하여 알고리즘을 개선하였다. 또한 **Bharat**의 알고리즘에서는 문서들이 연결된 **neighborhood** 그래프에서 중간 연결이 되는 문서를 제거함으로 삭제된 문서와 연결되어 있는 문서들간의 연결정보를 상실하는 문제를 해결할 수 있다.

(4) While the vectors H and A have not converged:
(5) For all n in N,
$A[n] := \sum_{(n',n)\in N} H[n'] \times auth\_wt(n',n)$
$A[n] := A[n] \times eq\_wt(n,EQ)$
(6) For all n in N,
$H[n] := \sum_{(n,n')\in N} A[n'] \times hub\_wt(n,n')$
$H[n] := H[n] \times eq\_wt(n,EQ)$
(7) Normalize the H and A vectors

[방법 3] B+ EQ : Bharat알고리즘에 eq_wt 추가

위의 알고리즘에서 EQ는 확장질의를 의미하며, eq_wt(n,EQ)는 문서 n과 확장질의 EQ와의 유사도를 의미한다. 여기에서 eq_wt(n,EQ)의 값이 0에 가까우면, 그문서의 authority값과 hub값에 0에 근접하여 Bharat의 알고리즘에서 그 문서를 제거하는 효과를 준다.

3.2 문서간의 유사도를이용한 차등적인 링크가중치

Kleinberg와 Bharat의 알고리즘에서는 한 문서에 존재하는 링크의 가중치를 동일하게 부여하고 있다. 그러나, 문서 안에 존재하는 링크들 중에서 중요한 링크가 있는가 하면, 중요하지 않은 링크도 존재한다. 또한 연결된 문서가 관련 있는 문서이기도 하고, 관련 없는 문서이기도 한다. 문서의 링크정보에서 현재 문서와 관련 있는 문서로의 연결에 더 높은 가중치를 부여하는 것이 타당하다고 생각한다. 본 논문에서는 연결되어 있는 문서들끼리의 유사도를 계산하여 문서에 있는 링크들에게 차등적인 가중치를 부여하였다.

(4) While the vectors H and A have not converged:
(5) For all n in N,
$A[n] := \sum_{(n',n)\in N} (H[n'] + sim(n,n')) \times auth\_wt(n',n)$
(6) For all n in N,
$H[n] := \sum_{(n,n')\in N} (A[n'] + sim(n,n')) \times hub\_wt(n,n')$
(7) Normalize the H and A vectors

[방법 4] B+ SIM : Bharat알고리즘에 문서 유사도추가

위의 알고리즘에서 sim(n,n')은 문서 n과 n'과의 유사도 값을 의미한다.

본 논문에서는 3.1절과 3.2절에 사용된 알고리즘을 결합하여 아래와 같은 알고리즘을 수정하였다.

(4) While the vectors H and A have not converged:
(5) For all n in N,
$A[n] := \sum_{(n',n)\in N} (H[n'] + sim(n,n')) \times auth\_wt(n',n)$
$A[n] := A[n] \times eq\_wt(n,EQ)$
(6) For all n in N,
$H[n] := \sum_{(n,n')\in N} (A[n'] + sim(n,n')) \times hub\_wt(n,n')$
$H[n] := H[n] \times eq\_wt(n,EQ)$
(7) Normalize the H and A vectors

[방법 5] B+ SIM + EQ : Bharat알고리즘에 eq_wt와 문서 유사도추가

4. 실험

본 논문에서 제시한 알고리즘을 검증하기 위하여 TREC10의 web trec 10G 자료를 이용하였다. 사용한 질의는 adhoc 질의 500-550번의 50개의 질의를 사용하였다. 또한 web trec에서 제공하는 링크정보를 이용하여 문서의 연결성을 분석하였다.

실험은 본 논문에서 언급한 방법 다섯 가지를 구현하여 비교하였다. 이를 정리하면 아래와 같다.

- Kleinberg : Kleinberg의 HITS 알고리즘 방법
- Bharat : Bharat의 개선한 HITS 알고리즘 방법
- B+ SIM : Bharat의 알고리즘에 문서들간의 유사도를 적용하여 차등적으로 링크 가중치 부여한 방법
- B+ EQ : Bharat의 알고리즘에 확장질의와 문서간의 유사도를 계산하여 적용한 방법
- B+ SIM+ EQ : Bharat의 알고리즘에 문서들간의 유사도 정보와 확장질의를 적용한 방법

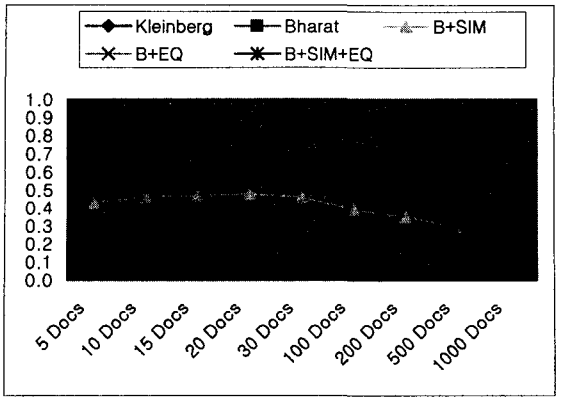
위의 방법에 의거하여 실험을 하여 얻은 결과는 아래에 표에 기술하였다.

	Kleinberg	Bharat	B+SIM	B+EQ	B+SIM +EQ
5 Docs	0.096	0.324	0.432	0.764	0.872
10 Docs	0.102	0.332	0.462	0.714	0.838
15 Docs	0.112	0.337	0.469	0.684	0.835
20 Docs	0.115	0.349	0.479	0.674	0.823
30 Docs	0.122	0.342	0.461	0.665	0.799
100 Docs	0.163	0.375	0.390	0.543	0.707
200 Docs	0.198	0.378	0.352	0.441	0.595
500 Docs	0.209	0.295	0.296	0.304	0.348
1000 Docs	0.149	0.171	0.177	0.167	0.175
AVG Pre.	0.141	0.323	0.391	0.551	0.666

[표 1] 실험결과

실험에서 적용한 평가방법은 R-P 방법으로 검색된 문서 R개에서의 정확율을 계산하였다. 즉 5Docs는 검색된 문서 5개에서 관련 있는 문서의 정확율을 측정한 값이다.

실험의 결과를 그래프로 표현하면 아래와 같다.



[그림 1] 실험결과

실험의 결과를 볼 때, Bharat의 알고리즘은 Kleinberg의 알고리즘보다 개선된 것을 볼 수 있고, 본 논문에서 제안한 방법들이 Bharat의 방법보다 더 좋은 결과를 얻었다. B+ SIM 방법은 21% 향상되었고, B+ EQ방법은 70% 향상되었으며, B+ SIM+ EQ 방법은 100%의 향상을 보았다.

5. 결론 및 향후과제

본 논문은 **Bharat**의 개선된 **HITS**알고리즘에 기반을 두고 확장된 질의와 기본 집합 안에 있는 문서와의 유사성을 계산하여 **hub**값과 **authority**값을 계산하는 일반식을 제시하였고, 문서 안에 존재하는 여러 링크정보에 연결된 문서와의 유사성을 계산하여 차등적으로 가중치를 부여하는 알고리즘을 개발하였다. 개선한 알고리즘을 검증하기 위하여 **TREC10**의 자료를 이용하여 실험하였으며 좋은 결과를 얻었다.

향후과제로는 본 논문에서 제시한 알고리즘에 대한 실험을 실제 웹을 대상으로 진행해보고자 한다. 또한 **HITS** 알고리즘이 안고있는 근본적인 문제인 초기 사용자 질의의 검색 결과인 시작집합이 좋지 않을 경우, 좋은 결과를 얻기가 어려운 문제가 있다. 이 문제를 해결해보고자 한다.

6. 참고문헌

[1] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998

[2] K. Bharat and M. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. ACM-SIRIR, 1998.

[3] The Google Search Engine: Commercial search engine founded by the originators of PageRank. Located at http://www.google.com