World Wide Web을 위한 개선된 Threshold HITS 알고리즘

김혜민[O] 김민구
아주대학교 정보통신전문대학원[O], 아주대학교 정보 및 컴퓨터 공학부
{khm[O], minkoo }@ajou.ac.kr

Enhanced Threshold Algorithm for HITS on the World Wide Web

Hyemin Kim[O], Minkoo Kim
Graduate School of Information and Communication, Ajou University
College of Information & Computer Engineering, Ajou University

요 약

링크 구조를 이용하는 대표적인 알고리즘인 HITS는 링크 정보를 이용하여 Authority와 Hub rating을 구하는 알고리즘이다. 그러나 HITS에서는 중요도와는 관계 없이 단순히 링크만을 많이 갖는 page의 Authority와 Hub rating이 비정상적으로 높게 계산되는 문제점이 있어 이를 해결하기 위한 연구들이 있었다. 본 논문에서는 이러한 연구들의 결과를 개선시키기 위해 Authority와 Hub rating의 단순한 합이 아닌, 평균과 priority를 적용하였다. 정확도를 측정하는 실험을 통해 제안하는 알고리즘이 기존의 방법보다 우수한 성능을 나타냄을 알 수 있다.

1. 서 론

웹에서 가치 있는 웹 페이지를 검색하는 일은 정보검색에 있어서 매우 중요한 문제 중의 하나이다. 따라서 많은 웹 서칭 기법들이 연구되었고 이러한 기법들은 주어진 쿼리에 대한 적합한 페이지를 찾기 위해 대부분 링크 구조를 이용하고 있다. HITS(Hypertext Induced Topic Selection) 알고리즘[4]은 웹 페이지 랭킹 알고리즘 중에서도 가장 잘 알려진 방법중의 하나이다. 그러나 이 알고리즘은 Authority와 Hub값을 계산하는데 있어서 문제를 갖고 있다. 즉, 어떤 페이지가 다른 페이지에 대한 링크를 많이 갖고 있을 경우 링크된 페이지의 중요성과는 관계없이 그 페이지는 높은 Hub와 Authority값을 갖게 된다. 또한 여러 페이지가 한 페이지를 링크하는 경우에도 높은 rating을 갖게 된다. 이러한 문제점들을 해결하기 위해 Threshold HITS[2]이 제안되었다. Threshold HITS는 평균 이상의 hub rating을 갖는 page들의 hub rating의 합과 authority rating이 상위 K인 page의 authority rating라는 개념을 도입함으로써 보다 정확한 Authority와 Hub값을 도출한다. 그러나 Threshold HITS 역시 편차가 큰 표본들이 분포할 경우 authority와 hub rating이 낮아지는 문제점이 발생한다. 따라서 본 논문에서는 이 알고리즘을 개선하는 방법에 대해 기술한다.

본 논문에서는 2장 관련 연구에서 HITS의 계산 방식의 문제점과 문제점을 해결하기 위해 제안되었던 방법[2]을 다. 4장에서 테스트 환경 및 결과를, 5장에서 결론 및 향후 연구 과제를 기술한다

2. 관련 연구

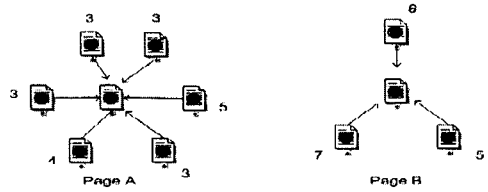
Kleingberg[4] 알고리즘은 authority와 hub rating을 계산하기 위해 링크 정보를 이용 하는 방법이다. 그러나 링크된 page의 authority와 hub rating을 모두 더하기 때문에 많은 page들과 link된 page는 항상 높은 rating을 갖게 된다. 이를 해결 하기 위해 rating의 평균과 상위 page만을 고려하는 방법[2]이 제안되었으나 이 방법대로 계산하게 되면 중요한 page의 authority와 hub rating이 상대적으로 낮아지는 경우가 생기게 된다.

2.1 HITS의 문제점

HITS에 따르면 페이지 i의 authority rating은 i를 링크하고 있는 모든 페이지의 hub rating의 합과 같다. 페이지 i의 hub rating은 i가 링크하고 있는 모든 페이지의 authority rating의 합과 같다. 따라서 [그림 1]에서 page A와 B의 authority rating은 각각 21, 180이다. Page A는 hub rating이 낮은 많은 수의 페이지에 의해 링크가 되어 있는 반면 page B는 hub rating이 높은 적은 수의 page에 의해 링크되어 있다. 이런 경우 B가 더 높은 authority rating을 갖는 것이 바람직하나 HITS에 따르면 A가 높은

본 논문은 과학기술부의 국가지정연구실 사업의 일환으로 지원받아 수행되었음. (과제번호 M10302000087-03J0000-04400)

살펴보고, 3장에서는 본 논문에서 제안한 방법을 설명한 Hub rating역시 단순한 합으로 계산되기 때문에 authority rating을 계산 할 때와 같은 문제가 생긴다.



[그림 1] 많은 링크와 적은 링크를 갖는 경우(authority)

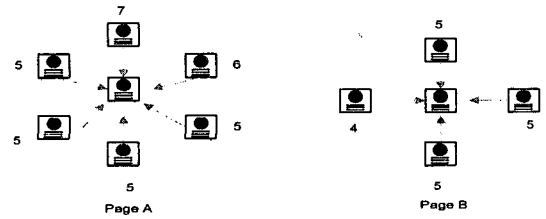
2.2 The Threshold-Kleinberg Algorithm

앞 절에서 언급한 문제점을 해결하기 위해 Borodin[2]은 hub와 authority의 수정된 계산방법을 제안하였다.

Authority rating의 경우, i라는 page를 링크한 모든 page j의 hub값을 더하지 않고 j의 hub의 평균값을 구하여 그 이상인 page의 authority값을 더한다..

Hub rating의 경우에는 i라는 page가 링크하는 page중 상위 K개만을 고려하여 계산한다.

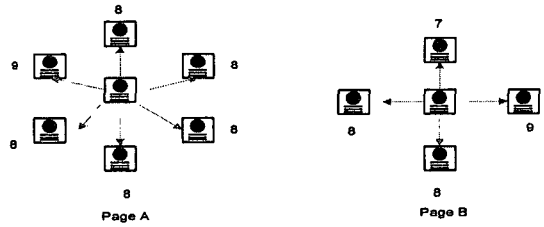
그러나 Authority rating의 경우 평균값 이하의 hub값을 갖는 page는 모두 무시되어 낮은 authority값을 갖게 된다. [그림 2]에서 pageA가 더 높은 authority rating을 가져야 하지만 평균값 이상인 page의 수가 적기 때문에 B가 더 높은 authority rating을 갖게 된다. 그러나 B의 경우 전체적으로 hub rating이 낮은 page에 의해 링크되었으나 평균이 낮으므로 거의 모든 page가 authority계산에 반영되어 높은 authority rating을 갖게 되는 문제가 있다



[그림 2] Threshold Algorithm에 따라 계산할 경우 (Authority Rating)

따라서 이와 같은 경우에는 좋은 많이 링크한 page의 hub값이 HITS의 계산 방식을 따를 때보다 더 낮아지게 된다. [그림 3]에서 page A는 page B보다 hub rating이 높아야 하지만 Threshold Algorithm을 따른다면 page B와 같은 hub rating을 갖게 된다.

authority rating을 갖게 되는 문제점이 있다.



[그림 3] Threshold Algorithm을 적용할 경우 (Hub Rating)

3. 개선된 Threshold-Kleinberg Algorithm

제안된 방법에서는 2장에서 언급한 문제를 해결하기 위해 평균값을 이용하였다. Rating이 평균 이상인 group과 평균 이하인 group으로 나누어 각각의 평균을 더하여 hub와 authority rating을 구하는 방법을 사용하였다. 단, 평균 이하의 rating을 갖는 page에 대해서는 그 page의 영향을 줄이기 위해 priority를 곱하여 평균을 구하도록 하였다. 예를 들어, P라는 page의 hub rating을 구한다고 하면 P가 링크하고 있는 page의 authority rating의 평균을 구한다. 평균의 값이 Avg라고 했을 때 Avg보다 큰 authority rating을 갖는 page에 대한 평균과 Avg보다 작은 authority rating을 갖는 page에 대한 평균을 각각 구한다. 이때 Avg보다 작은 authority rating을 갖는 page Q의 경우에는 priority를 적용하여 authority rating을 다시 계산하게 된다. 즉, Q의 authority rating AQ에 평균으로 나눈 값(AQ/Avg<1)을 곱하여 Q의 authority rating은 AQ보다 낮은 값이 된다. Priority는 원래의 authority rating을 평균으로 나눈 값인데 평균에 가까울수록 원래의 authority rating이 반영되어 계산되게 된다. 반대로 평균보다 작은 authority rating을 갖는다면 원래의 값보다 더 작은 값이 반영된다. Priority가 적용된 값을 이용해 Avg보다 낮은 page들의 평균을 구하고 그 값에 Avg보다 높은 page들의 평균을 합하면 P의 hub rating이 된다.

위에서 설명한 것을 식으로 나타내면 아래와 같다.

$$Hub\ (i) = U_{avg} + L_{avg}$$

$$A_{avg} = \frac{1}{N_j}\sum_{k=1}^{N_j} Authority\ (k)$$

$$U_{avg} = \frac{1}{N_u}\sum_{j=1}^{N_u} Authority\ (j)$$

$$L_{avg} = \frac{1}{N_l}\cdot\frac{1}{A_{avg}}\sum_{j=1}^{N_l} Authority\ (j)^2$$

Authority rating역시 링크하는 page들의 hub rating의 평균을 구하여 두 group으로 나누어 평균 이하인 page에 대해서는 priority를 적용하여 각각의 평균을 구하여 hub rating을 계산할 때와 같은 방법으로 계산 할 수 있다. [표 1]과 [표 2]를 살펴보면 제안된 방법으로 계산했을 경우 더 합리적인 authority rating을 얻을 수 있다는 것을 알 수 있다. Hub rating 역시 [표 4]를 통해 확인 할 수 있다.

	Page A	Page B
HITS	21	18
Threshold	9	13
Enhanced Threshold	7	10.6

[표 1] 그림 1의 authority rating

	Page A	Page B
HITS	33	19
Threshold	13	15
Enhanced Threshold	10.3	8.2

[표 2] 그림 3의 authority rating

	Page A	Page B
HITS	49	32
Threshold	25	25
Enhanced Threshold	16	14.4

[표 3] 그림 4의 hub rating (K = 3일 경우)

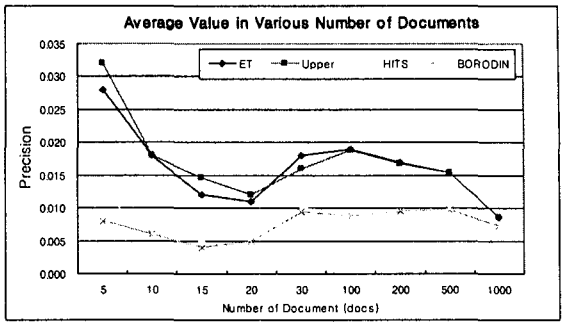
4. 실험 결과

실험을 위해 Text Retrieval Conference (TREC)에서 2002년도에 제공한 WebTREC11 자료를 테스트 데이터로 사용하였다. 자료의 크기는 약 18 Gbytes이며 총 page의 수는 1,247,753개이다. 검색에 사용된 질의는 WebTREC11에 포함된 Topic Distillation Task의 551-600번 질의를 사용하였다. 실험에 이용된 질의어의 예는 [그림 5]와 같다. 실제 검색 시스템에서는 title만을 이용하여 검색하였다. UPPER는 평균 이상인 page만을 고려한 것이고 ET는 Enhanced Threshold algorithm을 뜻한다. [표 4]는 문서의 개수에 따른 precision이다.

```
<top>
<num> Number : 517
<title> titanic what went wrong
<desc> Description:
Find documents that discuss the reasons for or problems leading to the sinking of the Titanic.
<narr> Narrative:
A relevant document will discuss what caused the Titanic to sink.
</top>
```

[그림 5] TREC Data의 예

HITS 알고리즘과 BORODIN, 본 논문에서 제안한 방법을 이용 하여 실험한 결과, precision이 향상된 것을 확인할 수 있었다. [표 5]를 통해rating이 평균 이상인 page만을 고려하는 것이 하이퍼링크 정보 검색성능에 더욱 긍정적인 영향을 미침을 유추해 낼 수 있다.



[표 4] Average Precision in Document Number

5. 결론 및 향후 과제

본 연구에서는 기존의 HITS 기법들이 갖는 문제점을 개선할 수 있는 방안을 제시하였다. 기존의 방식들은 hub와 authority rating의 단순한 연산을 통해 page의 rank를 매겼으나 제안하는 알고리즘은 보다 정밀한 검색을 위해 각 웹 페이지의 hub와 authority rating에 priority를 적용한 평균값을 이용하였다. 그 결과, 향상된 precision을 얻을 수 있었다. 여기에서 더 나아가 더욱 높은 정확도를 얻기 위해 URL 정보나 HTML tag 정보 등을 이용하는 방안은 향후 연구과제가 될 것이다.

참고문헌

[1] D.Cohn and H.Chang. Learning to probabilistically identify authoritative documents. Preprint, 2000.

[2] Allan Borodin, Gareth O.Roberts, Jeffery S.Rosenthal, Panayiotis Tsaparas. Finding Authorities and Hubs From Link Structures on the World Wide Web. In 10th Int. World Wide Web Conference, 2001.

[3] R.Lempel and S.Moran. The Stochastic Approach for Link-structure Analysis and the TKC effect. In 9th International World Wide Web Conference, May 2000.

[4] J.Klenberg. Authoritative sources in a hyperlinked environment. Journal of ACM, 46, 1999.

[5] Krishna Bharat, Monika R. Henzinger, Improved Algorithms for Topic Distillation in a Hyperlinked Environment. Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, 1998.