# 문서 내의 주제정보를 이용한 개선된 링크분석 알고리즘+

박기림° 장유진 김민구 박승규
아주대학교 정보통신전문대학원
{mind7°, cyj310, minkoo, sparky}@ajou.ac.kr

# Improved Link Analysis Algorithm Using Document Feature Information

Kirim Park° Yujin Chang Minkoo Kim Seungkyu Park
Graduate School of Information and Communication, Ajou University

**요 약**

최근 인터넷을 대상으로 하는 정보검색의 방법 중 하이퍼링크 정보를 이용한 방법이 각광받고 있다. 그리고 하이퍼링크 정보이외에 문서내에 존재하는 다양한 정보를 이용하여 검색 성능을 향상시키고자 하는 시도가 지속적으로 이루어지고 있다. 본 연구에서는 문서와 문서 사이의 유사도를 이용하여 하이퍼링크의 가중치를 부여하여 검색 성능을 향상시킨 방법을 개선하여 문서내의 주제정보를 추출하고 주제 단위의 유사도를 이용하여 하이퍼링크의 가중치를 새롭게 부여하여 링크분석 알고리즘에 적용하였다. 본 연구에서 제안한 방법이 문서사이의 유사도를 이용한 방법보다 뛰어난 성능을 나타내고 있음이 입증되었다.

## 1. 서 론

인터넷 검색 엔진의 문서 순위 결정 전략은 크게 두 가지로 나눌 수 있다. 하나는 사용자의 질의에 대하여 각각 문서의 단어 색인 정보, 즉 내용 정보를 이용하는 전략이고, 다른 하나는 사용자의 질의에 대하여 각각 문서에 포함된 링크 정보를 이용하는 전략이다. 내용정보를 이용한 방법은 사용자의 질의내용이 자세하고 많은 단어를 포함하고 있을 때 좋은 성능을 나타내고, 링크정보를 이용한 알고리즘은 일반적으로 사용자의 질의에 포함된 단어의 의미가 광범위하고 단어의 수가 적을 때 좋은 성능을 나타낸다. 그러한 장단점을 이용하여 최근에는 위의 두 가지 전략을 결합하여 사용한 다양한 연구가 진행되고 있다. 실제로 두 가지 전략을 결합한 알고리즘들이 각각의 전략을 이용한 알고리즘보다 좋은 성능을 보인다는 것이 입증되고 있다.

본 연구에서는 Kleinberg의 HITS 알고리즘을 문서간의 유사성를 이용하여 개선한 방법을 소개하고, 문서간의 유사성을 이용한 방법의 단점을 지적하며, 문서안의 주제별 유사성을 이용한 방법을 제안한다. 이는 하나의 문서안에는 하나의 주제만 있는 것이 아니라 여러 개의 주제가 있을 수 있다는 가정하에 문서와 문서사이의 유사도가 아닌 주제와 주제 사이의 유사도를 측정하여 Kleinberg의 HITS 알고리즘에 적용하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 하이퍼링크 분석에 관한 관련 연구로서 Kleinberg의 HITS 알고리즘을 소개하고, 문서간의 유사도를 이용한 HITS 알고리즘을 소개한다. 3장에서는 본 연구에서 제안하는 문서내의 주제들 간의 유사도를 이용한 알고리즘을 소개 한다. 그리고 4장에서는 제안된 알고리즘과 기존 알고리즘의 성능을 비교 분석 하기 위해 TREC11 자료를 이용하여 실험 및 분석을 하고, 마지막으로 5장에서는 본 연구의 결론과 향후 과제를 제시한다.

+ 본 논문은 KISTEP의 국가지정연구실 사업의 일환으로 지원받아 수행되었음 (과제번호 M10302000087-03J0000-04400)

## 2. 관련 연구

하이퍼링크 정보에 대한 분석방법은 크게 두 가지로 나누어질 수 있다. 하나는 문서집합 전체를 대상으로 하여 링크분석을 하는 전역적 방법이고, 다른 하나는 기본 검색 시스템이 도출한 결과집합을 대상으로 하여 링크분석을 하는 지역적 방법이다. 본 연구에서는 지역적 방법인 Kleinberg의 HITS 알고리즘에 초점을 맞추어 연구를 진행한다. [1][4]

### 2.1 Kleinberg의 HITS 알고리즘

Kleinberg의 HITS알고리즘은 각각의 문서에 대하여 두 가지 지표가 되는 값을 구하게 되는데 하나는 authority로서 사용자 질의와 관련도가 높은 문서들이 해당 문서에 얼마나 많이 링크를 하고 있는가를 나타내는 척도이고, 다른 하나는 hub으로서 해당 문서가 사용자 질의와 관련도가 높은 문서들에 대한 링크를 얼마나 포함하고 있는가를 나타내는 척도이다. 두 가지 척도는 서로 순환참조하여 서로의 값에 대한 보정을 하게 된다.

HITS 알고리즘은 사용자의 질의에 대하여 초기 검색 시스템(내용기반 검색 시스템)을 이용해 결과 문서 집합을 구하고, 그 문서 집합과 연결된 문서들을 포함하는 확장된 문서집합을 구한다. 그리고 확장된 문서집합에 있는 문서들에 대해서 authority값과 hub 값을 다음과 같이 계산하게 된다.

$$H(p) = \sum_{u \in S | p \to u} A(u), \quad A(p) = \sum_{v \in S | v \to p} H(v)$$

[수식 1] Kleinberg의 HITS 알고리즘

위의 수식을 통해서 각 문서에는 authority값과 hub값이 구해지게 된다. 좋은 authority값을 갖는 문서는 사용자의 질의와 밀접한 관련이 있는 문서이고, 좋은 hub 값을 갖는 문서는 사용자의 질의와 밀접한 관련이 있는 문서들을 많이 링크하고 있는 문서이다. HITS 알고리즘에서는 위의 수식에서 도출된 authority값이 높은 문서들이 일반적으로 사용자의 질의와 더욱 관련성이 높다고 보고 있다. [1]

### 2.2 문서간의 유사도를 이용한 HITS 알고리즘

Kleinberg의 HITS 알고리즘은 두 문서간의 하이퍼링크의 가중치를 특별한 고려 없이 1로 설정하고 알고리즘을 수행한다. 하지만 두 문서간의 연관성이 높은 링크에 대해서는 더 높은 가중치를 주어야 한다. 두 문서간의 링크에 대한 가중치를 계산하기 위해서 두 문서간의 유사도를 측정하여 이를 반영하고 있다.

두 문서간의 유사도는 유클리디언 거리(Euclidean Distance)를 이용하였다. 모든 문서 사이의 링크에 유클리디언 거리를 이용한 유사도를 반영하여 HITS알고리즘을 수행 하였을 경우 기존의 Kleinberg가 제안한 HITS알고리즘보다 월등히 뛰어난 성능을 보이고 있다.

## 3. 문서내의 주제들 간의 유사도를 이용한 알고리즘

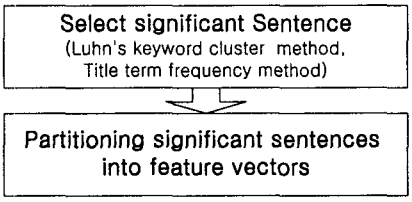
본 연구는 하나의 문서에는 단 하나의 주제만 있는 것이 아니라 여러 개의 주제가 있을 수 있다는 가정에서 출발한다. 그러므로 두 문서 사이의 하이퍼 링크의 가중치를 계산하기 위해서 두 문서간의 단순한 유사도를 측정하는 것이 아니라 두 문서에서의 여러 개의 주제들 중 가장 연관성이 높은 주제들 사이의 유사도를 측정하여 이를 HITS 알고리즘에 반영한다. 문서내의 주제들 사이의 유사도를 이용한 알고리즘은 크게 문서에서 주제들을 찾아내는 부분과 찾아낸 주제들을 이용하여 각 문서간의 링크에 대한 가중치를 부여하고 그에 따라 HITS 알고리즘을 수행하는 부분으로 나누어진다.

### 3. 1 문서내의 주제 추출 방법

문서 내에서 주제를 추출하는 방법은 문서내의 각 문장들 중 의미 있는 문장들을 추출하여 중복되는 단어가 있는 문장들에 대하여 합치는 과정을 통해서 문서내의 주제를 추출하게 된다.

우선 문서 내에서 의미 있는 문장을 추출하기 위해 의미있는 단어를 선정한다. 의미 있는 단어의 선정은 Luhn's Keyword Cluster Method 와 Title term frequency method를 이용한다. 그리고 의미 있는 단어의 개수와 전제 문서내의 문장의 개수를 이용하여 의미 있는 문장의 개수를 도출해낸다. 그리고 의미 있는 단어를 많이 포함한 문장을 의미 있는 문장으로 선정하게 된다.

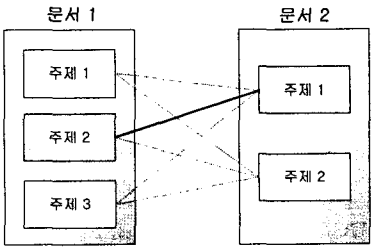
의미 있는 문장들이 선정되게 되면 문장들 간의 중복되는 단어가 있는가를 검사하여 중복되는 단어가 있는 문장들은 하나로 합치는 분할 화 과정을 거치게 된다. 분할 화 과정을 거치고 나면 한 문서 내에서 주제별로 여러 개의 주제 벡터가 생성되게 된다. 전체적인 과정은 [그림 1]과 같다. [2]




[그림 1] 문서내의 주제 추출 방법

### 3. 2 문서내의 주제를 이용한 하이퍼링크의 가중치 부여

문서와 문서 사이에 존재하는 하이퍼링크는 문서내의 주제와 링크 된 다른 문서의 주제를 연결하고 있다고 본다. 그렇기 때문에 하이퍼링크의 의미는 문서와 문서를 연결하는 것이 아니라 한 문서내의 주제와 다른 문서내의 주제를 연결하고 있는 것이라고 볼 수 있다. 이렇게 볼 때 하이퍼링크의 가중치는 주제와 주제사이의 유사도로 주어질 수 있다. 이때 연결된 두 문서 내에는 여러 개의 주제가 있을 수 있기 때문에 주제 사이의 유사도가 가장 높은 것을 두 문서의 하이퍼링크의 가중치라고 볼 수 있다.




[그림 2] 하이퍼링크의 가중치 결정

그림과 같이 두 문서가 연결되어있을 때 문서 1과 문서 2를 연결하는 하이퍼링크의 가중치는 주제간 유사도가 가장 큰 것으로 결정된다.

이와 같이 두 문서를 연결하는 하이퍼링크의 가중치가 결정되면 [수식 2]와 같이 변형된 Kleinberg의 HITS알고리즘을 수행하여 각 문서의 Authority값과 hub값을 결정하게 된다. 여기서 $Fsim$은 위에서 결정한 두 주제 사이의 유사도이다.

$$H(p) = \sum_{u \in S | p \to u} Fsim(p,u) \times A(u)$$

$$A(p) = \sum_{v \in S | v \to p} Fsim(p,v) \times H(v)$$

[수식 2] 문서의 주제간 유사도를 이용한 HITS 알고리즘

## 4. 실험 및 결과분석

본 연구에서 제안한 알고리즘을 검증하기 위해 Web TREC11 자료를 이용하였다. Web TREC11은 18Gb의 텍스트 자료만을 가지고 있는 대용량 자료로서 1,247,753개의 문서를 포함하고 있다. 질의어는 Web TREC11의 Topic Distillation Task의 551번부터 600번까지의 질의를 사용하였다.

실험은 Kleinberg에 의하여 제안된 기존의 HITS 알고리즘과 문서간의 유사도를 이용하여 개선된 HITS 알고리즘, 그리고 문서내의 주제간 유사도를 이용하여 개선된 HITS 알고리즘을 구현하여 실시하였다. 이를 정리하면 [표 1]과 같다.

- HITS : Kleinberg의 HITS 알고리즘
- Dsim HITS : 문서간의 유사도를 이용한 HITS 알고리즘
- Fsim HITS : 문서내의 주제간 유사도를 이용한 HITS 알고리즘

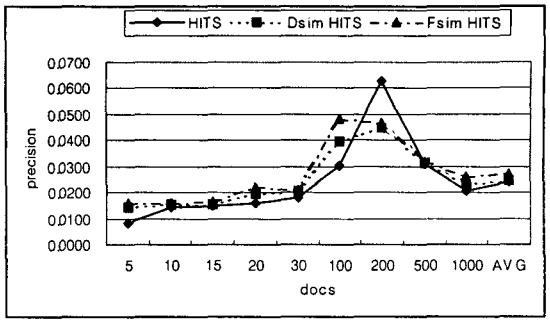
[표 1] 실험된 알고리즘들

위의 알고리즘들을 이용하여 얻은 실험 결과는 다음과 같다.

| | HITS | Dsim HITS | Fsim HITS |
|---|---|---|---|
| 5 | 0.0081 | 0.0143 | 0.0159 |
| 10 | 0.0142 | 0.0152 | 0.0161 |
| 15 | 0.0147 | 0.0153 | 0.0167 |
| 20 | 0.0160 | 0.0196 | 0.0221 |
| 30 | 0.0180 | 0.0204 | 0.0213 |
| 100 | 0.0306 | 0.0395 | 0.0484 |
| 200 | 0.0629 | 0.0449 | 0.0469 |
| 500 | 0.0309 | 0.0315 | 0.0321 |
| 1000 | 0.0208 | 0.0234 | 0.0261 |
| AVG | 0.0240 | 0.0249 | 0.0273 |

[표 2] 실험결과

실험 결과를 측정하기 위해 사용된 방법은 R-Precision 방법으로써 검색된 R개의 문서 내에서의 정확도를 측정하는 방법이다. 이를 그래프로 나타내면 [표 3]과 같다. [3]




[표 3] 실험결과 그래프

실험 결과에서 문서간의 유사도를 이용한 HITS 알고리즘은 기존의 Kleinberg가 제안한 HITS 알고리즘보다 평균적으로 3%정도의 미약한 성능향상을 보였으나, 본 논문에서 제안한 문서내의 주제간 유사도를 이용한 HITS알고리즘은 약 14%정도의 성능향상을 보였다. 그리고 문서간의 유사도를 이용한 HITS 알고리즘보다 문서내의 주제간 유사도를 이용한 알고리즘이 약 10%정도의 성능 향상을 나타냈다.

## 5. 결론 및 향후과제

본 논문은 하나의 문서의 내용이 여러 개의 주제를 포함하고 있음에 착안하여 문서를 연결하고 있는 하이퍼링크의 가중치를 주제 사이의 유사도를 이용해야 한다는 점을 주장했다. 이는 실험을 통해서 입증된 바와 같이 문서간의 유사도를 이용하여 하이퍼링크의 가중치를 설정하여 링크 분석을 한 것보다 문서내의 주제 사이의 유사도를 이용하여 링크 분석을 한 것이 성능의 더 좋게 나타났다.

향후과제로는 문서 내에서 주제를 찾을 수 있는 방법에 대하여 더욱 심도 있는 연구가 필요할 것이며, Kleinberg의 HITS알고리즘을 기반으로 하는 다른 개선된 알고리즘들에 대해서도 동일한 환경에서 테스트 해봄으로써 본 연구에서 제안한 방법의 효용 가치를 측정해 보아야 할 것이다.

## 6. 참고문헌

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

[2] Y. Chang. Conceptual retrieval Based on feature clustering of document. In Proceedings of the ACM-SIGIR Workshop 3 Mathematical/Formal Methods in Information Retrieval, 2002

[3] R. Baeza-Tates, B Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th WWW Conference, 1998.