제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

내용정보와 링크정보의 결합을 통한 검색 시스템의 성능향상 방법

박기림*, 김민구*, 박승규*
*아주대학교 정보통신전문대학원
e-mail:mind7@ajou.ac.kr

Improved Algorithm for information retrieval system with combining content and link information

Ki-rim Park*, Min-koo Kim*, Seung-Kyu Park*
*Graduate School of Information and Communication, Ajou University

요 약

월드 와이드 웹을 기반으로 한 검색 시스템에 대한 최근의 연구들은 링크정보가 내용정보와 함께 검색 시스템의 성능 향상에 커다란 도움을 주고 있다는 것을 증명하고 있다. 본 연구에서는 링크정보를 이용한 과거의 연구들을 살펴보고, 링크정보와 내용정보를 결합한 알고리즘들을 분석한다. 그리고 결합 알고리즘들을 크게 두 가지로 분류하고, 일반화된 모델을 제시한다..

1. 서론

인터넷 검색 엔진의 문서 순위 결정 전략으로 크게 두 가지가 사용되어져 왔다. 하나는 사용자의 질의에 대하여 각각 문서의 단어 색인 정보를 이용하는 전략이다. 이는 문서의 내용이 그 문서의 가중치를 결정짓는 요소가 되는 전략이다. 다른 하나는 사용자의 질의에 대하여 각각 문서에 포함된 링크 정보를 이용하는 전략이다. 이는 문서의 링크가 그 문서의 가중치를 결정짓는 요소가 되는 전략이다.

내용정보를 이용한 알고리즘은 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 링크정보와 내용정보의 결합을 통한 개선 알고리즘들을 분석하고 결합 알고리즘에 대한 일반적인 모델을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 링크정보를 이용한 알고리즘의 대표적인 알고리즘인 Kleinberg의 HITS 알고리즘[2]과 Google 검색엔진의 PageRank 알고리즘[3][4]을 중심으로 한 관련연구들을 소개하고, 3장에서는 위의 두 가지 알고리즘을 이용한 링크정보와 내용정보를 결합한 알고리즘들을 분석할 것이다. 그리고 4장에서는 결합 알고리즘에 대한 일반적인 모델을 제시 하고, 5장에서는 본 연구의 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

링크 정보를 이용한 연구는 크게 지역적 링크정보 연구와 전역적 링크정보 연구로 나눌 수 있다. 지역적 링크정보 연구로는 Kleinberg의 HITS 알고리즘[2]이 대표적이다. HITS 알고리즘은 검색시스템이 사용자의 질의에 대한 결과로 제공하는 초기 문서

집합을 이용하여 확장된 문서 집합을 구성하고 그 안에서의 링크정보를 이용해 문서들의 가중치를 계산하는 알고리즘이다. 전역적 링크정보 연구로는 Google systems의 PageRank 알고리즘[3][4]이 대표적이다. PageRank 알고리즘은 검색시스템이 검색하게 될 대상이 되는 전체 문서 집합에 대한 링크정보를 이용하여 사용자의 질의와 관계없이 문서들의 가중치를 계산하는 알고리즘이다.

2.1 Kleinberg의 HITS 알고리즘 [2]

Kleinberg의 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 hub값을 다음과 같이 계산하게 된다.

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

[수식1] Kleinberg : Kleinberg 의 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.

2.2 Google 검색엔진의 PageRank 알고리즘 [3][4]

Google 검색엔진의 PageRank 알고리즘은 사용자가 인터넷상에서 임의로 문서를 열어 보는 것으로 가정하고 있다. 현재 사용자가 열어보고 있는 문서에 있는 링크를 통해서 다른 문서를 열어볼 확률을 q, 문서 내에 있는 링크와 관계없이 사용자가 직접 주소를 입력하여 다른 문서를 열어볼 확률을 1 - q라고 하고 $C(a)$는 어떤 문서 a가 가지고 있는 링크의 개수이며 문서 a는 문서 p_1부터 p_n까지를 링크하고 있다고 가정할 때, 어떤 문서 a의 PR값은 다음과 같이 정의 된다.

$$PR(a)=q+(1-q)\sum_{i=1}^{n}PR(p_i)/C(p_i)$$

[수식2] Google : Google의 PageRank 알고리즘

PageRank 알고리즘은 위의 수식을 이용하여 문서 전체에 대하여 PR값을 모두 구하고, 사용자의 질의에 대하여 초기 검색 시스템(내용기반 검색 시스템)을 이용해 검색된 결과를 PR값의 가중치에 따라 사용자에게 돌려주게 된다.

3. 링크정보와 내용정보를 결합한 알고리즘

링크정보와 내용정보를 결합한 알고리즘은 매우 다양한 시각에서 접근되고 있다. 크게는 내용정보 검색 시스템에 링크정보를 이용해 가중치를 조절하는 방법과 링크정보 검색 시스템에 내용정보를 이용해 가중치를 조절하는 방법으로 구분 할 수 있다.

매우 다양한 종류의 결합 알고리즘들이 연구되고 있지만, 본 연구에서는 결합방법 중 대표적인 몇 가지만을 소개한다. 내용정보 검색 시스템에 링크정보를 이용해 가중치를 조절하는 방법으로는 Kraaij가 제안한 방법[5]과 Silva가 제안한 방법[6]이 있다. 링크정보 검색 시스템에 내용정보를 이용해 가중치를 조절하는 방법으로는 Chakrabarti가 제안한 방법[7]과 Richardson이 제안한 방법[8]이 있다.

3.1 내용정보 검색에 링크정보를 이용한 알고리즘

내용정보 검색에 링크정보를 결합시킨 알고리즘은 대부분 내용정보 검색에서 검색된 문서들을 순위화하기 위해 사용되는 가중치에 링크정보를 이용한 알고리즘을 접목하고 있다.

3.1.1 Kraaij의 알고리즘[5]

Kraaij의 알고리즘은 Ponte가 제안한 Language Model[10]을 기반으로 하고 있다. Kraaij는 문서가 사용자의 질의와 관련 있을 확률 값을 문서를 연결하고 있는 링크의 개수를 이용하여 조정한 방법을 이용하고 있다.

Hiemstra가 제안한 Language Model에서는 문서가 질의와 관련될 확률을 다음과 같은 수식으로 표현하고 있다.

$$P(D\mid Q)=\frac{P(D)P(Q\mid D)}{P(Q)}$$

[수식 3] Hiemstra : Language Model

Kraaij는 문서가 질의와 관련 있을 확률 값에 대

하여 문서를 링크하고 있는 문서의 개수와 문서 URL의 깊이를 이용하고 있다.

$$P_{inlink}(D) = P(R \mid D) = C \cdot inlinkCount(D)$$

[수식4] Kraaij : 링크의 개수에 의한 확률 값

*inlinkCount(D)*는 문서 *D*를 링크하고 있는 문서의 개수이고, *C*는 상수 값이다.

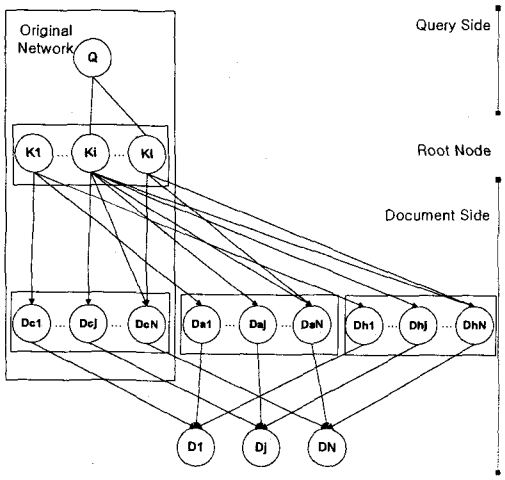
$$P_{URL}(D) = P(EP \mid URLtype(D) = t_i) = \frac{c(EP, t_i)}{c(t_i)}$$

[수식5] Kraaij : URL의 깊이에 의한 확률 값

*URLtype(D)*는 문서 *D*의 URL의 종류이고 $c(EP,T_i)$는 문서의 종류가 EntryPage이고 T_i인 것의 개수, $c(T_i)$는 문서의 종류가 T_i인 것의 개수이다. 이 알고리즘에서는 Language Model에서 질의와 문서간의 관련이 있을 확률 값에 문서의 링크의 개수에 의한 확률값과 URL의 깊이에 의한 확률 값을 적용하였다.

3.1.2 Silva의 알고리즘[6]

Silva의 알고리즘은 Bayesian networks를 검색 시스템에 적용한 belief network model[10]을 내용기반 검색 시스템의 기본 틀로 사용하였다. belief network model을 이용해서 구축한 기본 network에 Kleinberg가 제시한 authority와 hub값을 갖는 node들을 연결하여 network을 확장하였다.



[그림1] 링크정보를 이용한 Bayesian network 확장

이를 기반으로 링크정보에 대한 가중치와 내용정보에 대한 가중치를 결합해서 이용하였다.

3.2 링크정보 검색에 내용정보를 이용한 알고리즘

링크 정보를 이용한 알고리즘에 내용정보를 결합한 알고리즘들은 Kleinberg의 HITS알고리즘을 기반으로 좋은 문서의 척도가 되는 authority값을 계산할 때 문서의 내용정보를 반영하고 있거나 PageRank알고리즘의 PageRank값을 구할 때 내용정보를 반영하고 있다.

3.2.1 Chakrabarti의 알고리즘

Chakrabarti는 Kleinberg의 HITS 알고리즘에서 authority값과 hub값을 구할 때 링크된 문서들 간의 내용정보가 반영되지 않음을 지적하고 있다. Chakrabarti는 내용정보를 반영하는 방법 중 하나로 링크 주변에 있는 앵커텍스트를 이용하고 있다. 앵커 텍스트는 웹 문서에서 다른 문서로 링크가 되어 있는 부분의 주변에 있는 단어들의 집합을 의미한다.

Kleinberg의 HITS 알고리즘에서는 두 문서가 링크 되어있을 때 authority값이나 hub값을 링크의 개수를 이용하여 구하고 있다. 그러나 Chakrabarti는 연결하고 있는 문서의 앵커텍스트에 포함되어있는 단어정보를 이용하고 있다. 문서 p가 q를 링크하고 있을 때 링크의 가중치는 다음과 같이 표현된다.

$$Weight\ \ W(p,q) = 1 + n(t)$$

[수식6] Chakrabarti : 문서 간 링크의 가중치

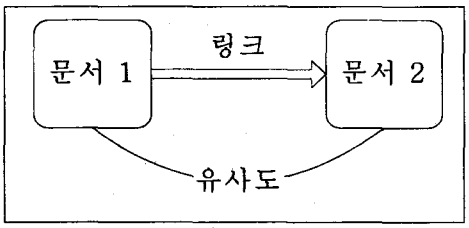
위의 수식에서 *n(t)* 는 문서 p에서 q를 연결하고 있는 링크의 앵커텍스트에 포함된 단어 중에 질의어에 포함된 단어의 개수를 의미한다. 각각 문서에 포함된 모든 링크에 대해 위의 수식과 같이 가중치를 구한 값을 이용하여 Kleignberg의 HITS알고리즘을 수행하였다.

3.2.2 Richardson의 알고리즘

Richardson의 알고리즘은 PageRank알고리즘을 기반으로 하고 있다. PageRank알고리즘은 사용자가 임의로 문서를 열어볼 확률 값을 이용해 문서의 가중치를 계산한다. 그런데, PageRank알고리즘에서는 사용자가 문서를 열어볼 확률을 임의로 정해주기 때문에 실제 사용자가 문서를 열어볼 확률과는 크게 차이가 나게 된다. 이에 대하여 Richardson은 사용자가 문서를 열어볼 확률을 질의어와 관련 있는 문서에 대해서 더 높게 부여하고 있다.

4. 결합 알고리즘의 일반화

내용정보와 링크정보의 결합 알고리즘들은 커다란 두 가지 요소를 이용하고 있다. 하나는 문서와 문서 사이에 존재하는 링크정보이고, 다른 하나는 문서와 문서의 내용정보의 유사도 이다.



[그림 2] 결합 알고리즘의 요소

결합 모델은 다음과 같이 두 가지 형태로 일반화 할 수 있다. 하나는 문서와 문서 또는 문서와 질의어 사이의 유사도를 계산할 때 링크정보를 반영하는 형태이고[수식7], 다른 하나는 문서간의 링크의 가중치를 계산할 때 문서와 문서사이의 유사도를 반영하는 형태이다[수식8].

$$newSim(D1,D2)=sim(D1,\ D2)\cdot linkinfo(D1,\ D2)$$

[수식7] 유사도에 링크정보 반영
('·' 는 결합의 의미를 갖는 연산자)

$$linkWeight(D1,D2)=sim(D1,\ D2)$$

[수식8] 링크 가중치에 유사도를 반영

5. 결론

월드 와이드 웹 환경은 계속해서 많은 문서들이 생성되고 있으며 그에 따른 링크정보도 계속해서 생성되고 있다. 이런 환경에서 검색 시스템은 단순히 문서의 내용정보만을 이용해서는 좋은 결과를 보일 수 없다. 그로 인해서 링크정보를 이용한 다양한 알고리즘이 연구되었다. 그러나 단순히 링크정보만을 이용한 알고리즘들이 한계점에 봉착하였고, 내용정보와 링크정보를 결합한 알고리즘들이 연구되고 있다.

본 논문에서는 내용정보와 링크정보를 결합한 알고리즘들을 분석하고, 그에 따른 일반화된 모델을 제시하였다.

6. 향후 과제

향후과제로는 본 논문에서 분석한 알고리즘들에 대하여 일반화된 실험 데이터를 이용하여 실험해보고 결과를 분석하고자 한다. 또한, 본 논문에서 제시한 일반화된 모델을 이용하여 내용정보와 링크정보를 이용한 개선된 알고리즘을 연구하고 실험해 봄으로써 일반화된 모델의 효용가치를 판단해 보고자 한다.

참고문헌

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

[2] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the ACM-SIAM Symposium on Discrete Algorithms,* pages 668-677, 1998.

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

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

[5] W. Kraaij. The Importance of Prior Probabilities for Entry Page Search. In *Proc of the 25th ACM-SIGIR conference,* 2002.

[6] I. Silva. Link-Based and Content-Based Evidential Information in a Belief Network Model. In *Proc of the 23rd ACM-SIGIR conference,* 2000.

[7] M. Richardson, P Domingos. The Intelligent Surfer : Probabilistic Combination of Link and Content Information in PageRank. *volume 14. MIT Press, Cambridge, MA,* 2002

[8] S, Chkrabarti. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. In *Proc. of the 7th WWW Conference,* 1998.

[9] J.M. Ponte, W.B. Croft. A language modeling approach to information retrieval. In *Proc of the 21st ACM-SIGIR conference,* 1998.

[10] B. Ribeiro and R. Muntz. A belief network model for ir. In *Proc. of the 19th ACM-SIGIR conference,* 1996.