문서간의 유사도를 이용한 개선된 PageRank 알고리즘

이경희°, 김민구, 박승규
아주대학교 정보통신전문대학원
lkh°@ceai.ajou.ac.kr {minkoo, sparky}@ajou.ac.kr

Improved PageRank Algorithm Using Similarity Information of Documents

Kyunghee Lee MinKoo Kim Seungkyu Park
Graduate School of Information and Communication Ajou University

요 약

웹에서의 검색 방법에는 크게 Text-Based 기법과 Link-Based 기법이 있다. 본 논문은 그 중에서 Link-Based 기법의 하나인 PageRank 알고리즘에 대해 연구 하고자 한다. 이 PageRank 알고리즘은 각 페이지의 중요성을 수치로 계산하는 방법이다. 하지만 이 알고리즘에서는 페이지에서 페이지로 링크를 따라갈 확률의 값을 일정하게 주어서 모든 페이지의 값을 획일적으로 계산하였기 때문에 각페이지의 검색 효율성에 문제가 있다고 판단하여 , 이를 해결하고자 본 논문은 페이지사이의 유사도를 측정하여 유사도에 따라 링크를 따라가는 확률 값인 Damping factor값을 다르게 부여하여 검색의 효율성을 높였다. 이를 위하여 두 가지 방법의 실험을 통하여 구현, 증명하였다.

1. 서 론

검색엔진에서 과거에는 주로 Text-Based 검색방법을 사용하였지만, 웹의 성장이 거듭되면서 링크를 이용한 Link-Based 검색방법의 이용이 늘어나고 있는 추세이다. Link-Based 검색방법은 Text-Based 검색방법의 단점을 많이 보완하여 발전하여왔고, 많은 알고리즘이 존재한다. 그 중 대표적인 알고리즘인 PageRank 알고리즘의 방법이 있다. 이 알고리즘은 페이지에서 링크된 페이지의 importance를 수치화시켜 페이지간의 관계를 보여주는 것이다 . 그러나 페이지에서 페이지로 링크를 따라갈 확률을 일정하게 부여하여, 페이지마다 가지고 있는 링크가 서로 다름에도 불구하고, 획일적인 Damping factor 값을 적용시켜 검색의 효율이 떨어지는 점이 발생하였다. 이에 본 논지에서는 이런 점을 개선하고자 페이지에서 링크된 페이지와의 유사도를 이용하여 서로 다른 Damping factor 값을 적용하여 검색의 효율성을 개선시키는 알고리즘을 설계와 구현을 통하여 제시하고자 한다. 앞으로 본 논문의 2장은 PageRank 알고리즘에 대해 논하고,3장에서는 PageRank 알고리즘의 문제를 해결한 유사도를 이용한 PageRank에 대해 살펴본다. 그리고 4장에서는 유사도를 이용한 PageRank 알고리즘의 구현에 대해 알아보고,5장에서 그 실험의 결과를 분석하고 6장에서 결론을 맺고자 한다.[1)]

2. PageRank 알고리즘의 이론

2.1 웹의 링크 구조

웹의 링크 구조는 Junhoo Cho 박사가 아래와 같은 이론으로 자세하게 설명하였다. 현재의 웹은 매우 복잡한 구조로 연결되어있다. 모든 웹 페이지들은 forward link와 backlink를 가지고 있다. 웹 페이지가 가지고 있는forward links 나 backlink의 많은 개수들을 가지고 있는데, 예를 들면 네스케프 홈페이지의 backlink는 62,804개이다. 이점을 감안하여 PageRank 알고리즘은 important 한 페이지들은 많은 다른 웹 페이지에서 link를 한다는 점을 착안한 것이다. 어떤 웹 페이지의 backlink의 갯수가 많을수록 그 페이지의 중요성은 높은 것이다. [1]

2.2 PageRank 알고리즘의 구성이론

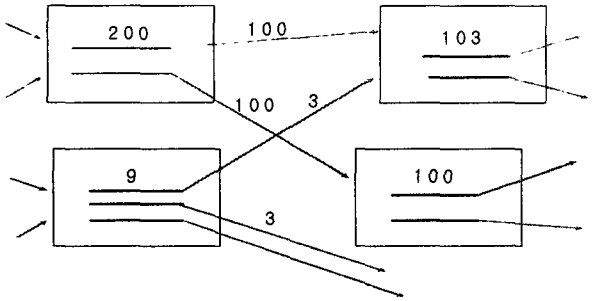
PageRank는 웹 페이지들끼리 관계되는 중요성을 측정하기 위하여 쓰는 방법이다. 이 방법은 웹의 graph에 기반한 모든 웹 페이지들에 대한 ranking을 계산하기 위한 방법이다. 웹에서 모든 페이지끼리의 링크는 거대한 그래프 구조로 이루어져 있다고 말할 수 있다. PageRank알고리즘은 그래프 구조를 이용한다. PageRank의 graph를 보면 node와 node사이에 link로 이루어져 있으며, 여기에 방향이 주어지는 directed graph를 이용하는 것이다. PageRank 알고리즘의 공식을 보면 다음과 같다

$$\mathbf{R(u)} = \mathbf{c} \sum_{v \in B_u} \frac{R(v)}{N_v} \qquad [1]$$

위의 공식은 PageRank 의 간단한 알고리즘 공식을 보여준 것이다. u라는 페이지의 PageRank를 계산하려면 위의 공식을 적용하면 된다. V는 u페이지의 backlink들이고 Nv는

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

v웹페이지의 총 링크 수이다. u페이지의 backlink를 합한 것이 바로 u의 PageRank 가 되는 것이다. 여러 페이지들이 어떤 한 페이지를 링크하는 것은 링크된 페이지가 그만큼 중요한 페이지라는 것이다. 중요한 페이지일수록 PageRank 의 값이 높게 되는 것이다. 어떤 페이지의 PageRank의 값이 높다는 것은 그만큼 다른 페이지에서 link을 많이 한다는 것이다. 이처럼 PageRank 알고리즘은 각 페이지의 중요성을 수치화 시키는 알고리즘이라 할 수 있다. 위의 공식을 그대로 웹 페이지에 적용시킨 과정을 도식화 해보면 아래 그림과 같다.

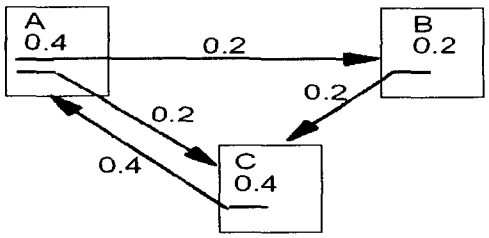


[그림1 웹에서의 PageRank 계산방법]

위의 그림은 웹의 한 단면에서 PageRank를 계산하는 방법을 보여준 것이다. 어떤 페이지의 PageRank 가 200이 되고 그 페이지에는 두개의 링크가 존재한다면, 하나의 링크 값이 100으로 링크된 페이지에 전달 된다. 여러 개의 페이지에서 링크된 웹 페이지는 바로 이 backlink의 값들의 합이 된다.

2.3 PageRank 알고리즘의 반복 문제

위의 공식에서 간단하게 PageRank 알고리즘을 계산하였다. 여기에 한가지 문제가 존재한다. 서로 다른 두 개의 웹 페이지가 다른 페이지로의 링크 없이 서로가 링크 한다고 가정하면, 두 개의 웹 페이지 사이에는 loop(반복)의 문제가 존재 할 것이다. 아래 그림과 같은 loop 가 존재한다.[1]



[그림2 PageRank알고리즘에서 발생하는 loop 문제]

이 loop form을 rank sink 문제라고도 한다. 이 rank sink 문제를 해결하려면 각 페이지의 PageRank 를 계산하는 요소를 변화시키면 된다. 즉 기본적인 공식에서 사이클을 따라가지 않고 다른 페이지로 가게 될 확률을 집어넣으면 되는 것이다. 문제를 해결한 공식은 아래와 같다.

$$\mathbf{R(u)= c \sum_{v\in B_u} \frac{R(v)}{N_v} + \frac{(1-c)}{N}} \qquad [2]$$

2.4 PageRank 알고리즘의 계산

PageRank 알고리즘에서 웹 페이지의 초기값은 웹 그래프에서의 node의 개수가 된다. 웹에서의 node란 바로 페이지를 말하는 것이다. c 는 Damping factor로써 link된 페이지를 가게 될 확률 값이다. 그리고 반복적인 문제의 해결 공식으로 링크를 따라가지 않고, Random 하게 다른 웹 페이지로 가게 될 확률 값이 바로 1-c가된다. 시작페이지부터 더 이상 링크가 존재하지 않을 때 까지 알고리즘은 반복적으로 작동된다. 효율적인 PageRank 알고리즘 계산방법은 아래와 같다.

∀Source[s] = 1/N
while (residual > τ) {
  $\forall_d$Dest[d] = 0
  while (not Links.eof()) {
    Links.read(source,n,$dest_1$,$dest_2$,$dest_3$...$dest_n$)
    for j=1....n
      Dest[$dest_j$]=Dest[$dest_j$] + Source[source]/n
  }
  $\forall_d$Dest[d] = c ×Dest[d] + $\frac{1-c}{N}$
  residual = ||source - Dest||
  Source = Dest
} [2]

3. PageRank Algorithm using Similraity

3.1 PageRank Algorithm 의 전반적인 문제점

PageRank Algorithm 은 각 페이지마다 importance를 구하는 알고리즘이다. PageRank알고리즘 공식에서 보면 각 페이지에서 Damping factor로 링크를 따라가는 확률을 0.85로 계산하였고, random하게 다른 페이지로 가게 될 확률을 0.15의 값으로 획일하게 계산하였다. 그러나 페이지끼리의 관계 중요성을 고려하지 않고 획일적인 확률 값을 0.85로 구현하였다. 본 논지에서는 각 페이지의 PageRank 알고리즘을 이용하여 중요성을 수치화 할 때 유사도를 이용하여 좀더 정확한 검색 효율성을 얻고자 한다. 즉 어떤 페이지에서 다른 페이지로 링크를 할 때 두 페이지 사이에 유사도를 측정하여 유사도가 높으면 damping factor를 높게 주고, 유사도가 낮으면 damping factor 값을 낮게 주는 알고리즘을 적용하여 검색의 정확성을 높이고자 한다.

3.2 PageRank I Algorithm Using Similarity의 이론

이 알고리즘의 전체적인 작동방법은 우선은 순수한 PageRank 알고리즘에서 Damping factor값만 변형시킨 것이다. 어떤 한 웹 페이지에서 다른 페이지로 가는 확률의 Damping factor 값을 적절하게 조절하여 각 페이지의 중요성을 계산하는 방법이다. 순수한 PageRank 알고리즘에서는 이 Damping factor를 일정하게 0.85의 값으로 부여 하였다. 유사도를 이용한 PageRank 알고리즘에서는 각 페이지와 페이지사이의 유사정보를 측정하여, 만일 어떤 A

라는 페이지와 그 페이지가 링크하는 페이지의 유사도가 높으면 보통 link를 따라갈 확률이 높은 것이고, 유사도가 낮으면 link를 따라가는 확률보다 random 하게 다른 페이지로 갈 확률이 높다.

$$R(u)= c \sum_{v \in B_u} \frac{R(v)}{N_v} + \frac{(1-c)}{N} \qquad [2]$$

유사도에 따라서 Damping factor의 값인 c를 다르게 부여하였다. 유사도가 0.4 이상이면 Damping factor의 값을 0.9를 적용시키고, 유사도가 0.2 이상이면 Damping factor의 값을 0.85로 적용하였다.

3.3 PageRank II Algorithm Using Similarity의 이론

PageRank II Algorithm은 유사도를 이용한 PageRank I과는 비슷하지만 유사도에 따라 Damping factor의 범위를 다르게 부여하였다. 이렇게 Damping factor의 범위 차이를 많이 주는 이유는 페이지와 관계있는 링크를 따라갈 확률 값을 높게 주어 검색의 정확성을 높이고자 함에 있다. 유사도가 높은 페이지끼리는 link을 따라갈 확률 값을 높게 부여하였고, 유사도가 낮은 페이지들은 link를 따라갈 확률 값을 낮게 부여 하였다. 공식은 아래와 같다.

c = 0.5 * sim + 0.70 ①

c = 2 * sim + 0.40 ②

만일 유사도인 sim이 0.2보다 크면 ①번 공식을 적용하고 유사도가 0.2보다 작으면 ②번 공식을 적용한다.

4. PageRank Algorithm Using Similarity의 구현

4.1 알고리즘 구현의 실험 환경

실험 환경으로는 울트라 6.0 서버환경을 사용 하였으며, 알고리즘의 구현을 위해 C language 을 사용하였다. 실험 데이터로는 외국의 성능 평가용 데이터 집합인 WTX 문서 중에서 5개의 Directory를 임의로 추출하여 실험 했으며,하나의 Directory 에는 평균적으로 15,000개의 문서를 포함하고 있다. 이 문서들을 실험집합으로 사용하였다.

5. PageRank Algorithm Using Similarity 의 결과분석

[표1 WTX001디렉토리의 결과]

	Query NO	Original PageRank	PageRank I Using Sim	PageRank II Using Sim
WTX 001	4	0.083333333	0.0666667	0.07142857
	10	0.000896861	0.0010941	0.00145985
	60	1	1	1
	95	0.001666667	0.0013423	0.00095551
	128	0.5	1	0.41476870

위의 표는 첫 번째 Directory WTX001 에 대한 실험결과이다. 이중 5개의 query에 대한 Original PageRank 와 유사도를 이용한 PageRank 알고리즘의 결과를 비교 한 것이다.4번 query 와 95번 Query 에 대해서는 결과를 비교해보면 성능이 조금 떨어졌고, 나머지 query에 대해서는 현저히 증가하였다. Damping factor의 범위를 적게 적용한 유사도를 이용한 PageRank I 보다는 Damping factor의 범위를 크게 부여한 유사도를 이용한 PageRank II 알고리즘이 현저히 향상된 것을 알 수 있다.

[표2 WTX003 디렉토리의 결과]

	Query NO	Original PageRank	PageRank I Using Sim	PageRank II Using Sim
WTX 003	27	0.5	0.5	1
	49	0.003584229	0.00374531	0.00392156
	86	0.002873563	0.00806451	0.00729927
	89	0.000327761	0.00314465	0.00352112
	122	0.000327654	0.00035026	0.00037355

위의 표에서 보면 Original PageRank 알고리즘보다 유사도를 이용한 PageRank 알고리즘의 성능이 훨씬 좋은 것을 알 수 있다. 유사도를 이용한 PageRank I 보다는 다양한 Damping factor 값을 준 PageRank II의 성능이 평균 97%로 향상되었다. 다섯 개의 Directory 실험값의 평균값으로 나타내 보면 Original PageRank 알고리즘보다 유사도를 이용한 PageRank I 알고리즘이 평균 25.8% 향상되었고, 유사도를 이용한 PageRank II 알고리즘은 원래의 PageRank 알고리즘보다 44.2% 향상된 것을 알 수 있다. 5개의Directory 중에서 모든 값이 향상된 것이 아니고, 그 중 몇 개의 query는 Original PageRank 보다 결과가 좋아지지 않은 것도 있지만 전반적인 값들이 향상된 것 을 알 수 있다.

6. 결론

웹의 사용량이 증가하면서 웹에서의 검색방법도 다양하게 발전되고 있다. 그 중 웹 문서의 특징인 link을 이용한 검색 알고리즘 연구도 다양하게 발전하여 왔다.

그 중 google 사이트에서 사용한 PageRank 알고리즘방법을 사용하여 PageRank 의 문제점을 조금 수정시켜 검색의 효율성을 높이고자 유사도를 이용한 PageRank 알고리즘을 구현하여 결과를 분석하였다. 분석한 결과 Original PageRank 알고리즘 보다 페이지끼리의 유사도를 이용하여 Damping factor를 다르게 부여한 PageRank I 과 PageRank II의 성능 이 훨씬 좋아졌다. Damping factor의 값을 0.85~0.9 사이의 값을 부여한 PageRank I 알고리즘은 Original PageRank 알고리즘보다 평균 25.8% 가 향상된 것을 볼 수 있고, Damping factor의 값을 유사도에 따라 0.4~0.95까지 다양하게 적용한 PageRank II 알고리즘은 Original PageRank 알고리즘보다 44.2%의 성능이 향상된 것으로 나타났다. 하지만 이 실험 결과는 full-text 로 한 것이 아니고, 5개의 Directory를 임의로 추출하여 실험을 한 것이기 때문에 full-text 보다 좁은 범위에서 실험 평가하였다. 향후 완전한 full-text 에서 실험 연구되어야 하고, 모든 값들이 향상된 것이 아니기 때문에 향후 이 부분에 대해서도 좀더 많은 연구가 필요하다 할 것이다.

참고 문헌

[1] Junghoo Cho " The PageRank Citation Ranking : Brining order to Web" (1998)

[2]Taher H.Haveliwala " Efficient Compution of PageRank " (1999)

[3] Jon M.Kleinberg " Authritative Sources in a Hyperlinked environment" (1998)