The PageRank Citation Ranking: Bringing Order to the Web

2021-08-15

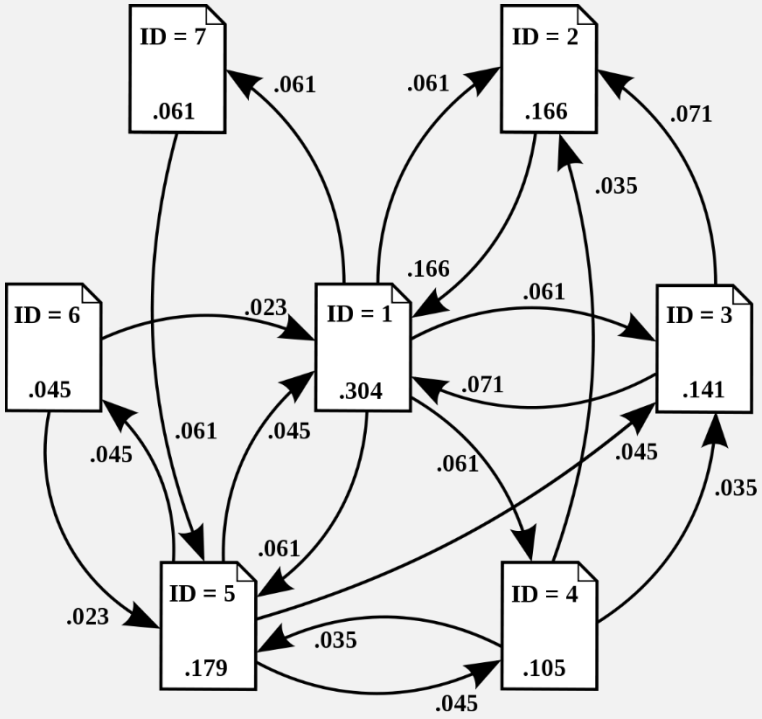
발표자 김찬우

PageRank?

-A method for computing a ranking for every web page based on the graph of the web

-어떤 웹페이지가 더 중요한지 알려주는 척도

-스탠퍼드 대학교에 재학중이던 래리 페이지와 세르게이 브린이 개발, 구글의 시작



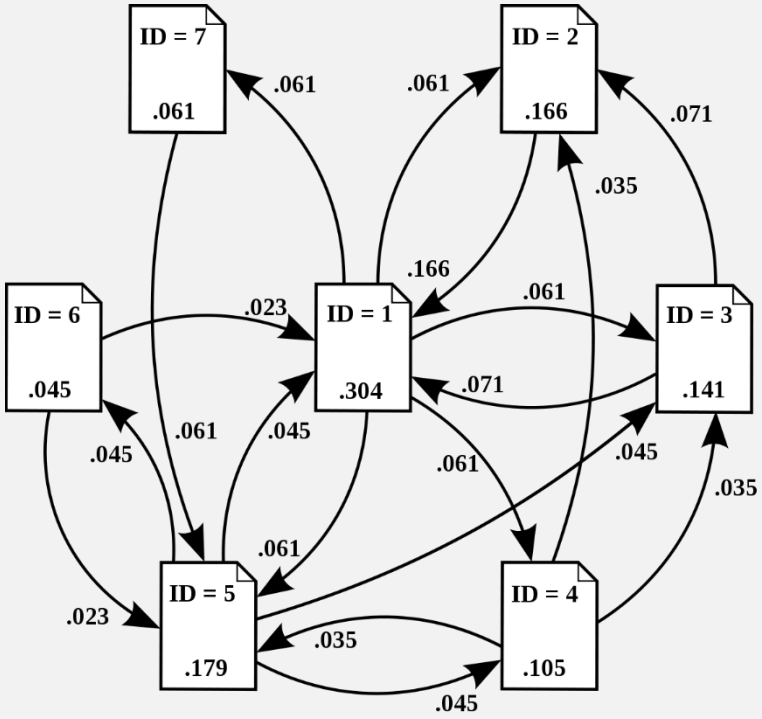
이미지 출처 - 페이지랭크 위키피디아

PageRank 전에는?

-단순한 내용 중심의 검색

-웹페이지가 인용된 횟수만을 가지고 검색에 활용

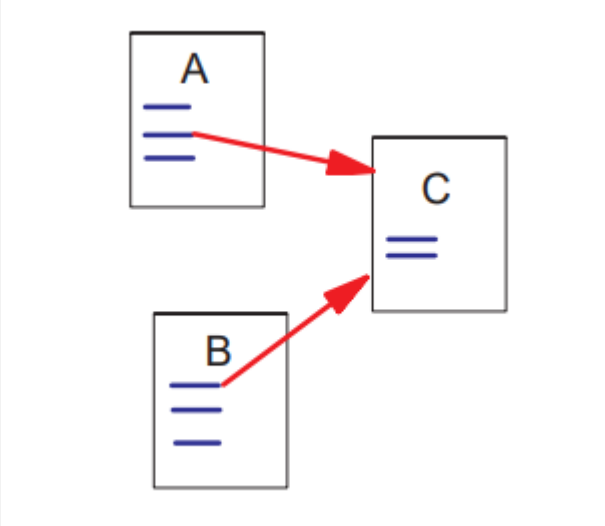
-사용자에게 중요한 웹페이지 순으로 정렬하지 못함



이미지 출처 - 페이지랭크 위키피디아

PageRank의 아이디어

1. 많은 backlink를 가진 웹페이지는 중요한 웹페이지일 것이다.
2. 랭크가 높은 웹페이지의 링크는 랭크가 낮은 웹페이지의 링크보다 영향력이 클 것이다.



이미지 출처 - The PageRank Citation Ranking: Bringing Order to the Web (1998) 이하 본 논문

A와 B는 C의 backlink

C는 A의 forward link

C는 B의 forward link

단순화된 PageRank 수식

$R(u)$: 페이지 u의 랭크
F_u: 페이지 u의 forward link
B_u: 페이지 u의 backlink
N_u: 페이지 u의 forward link 수

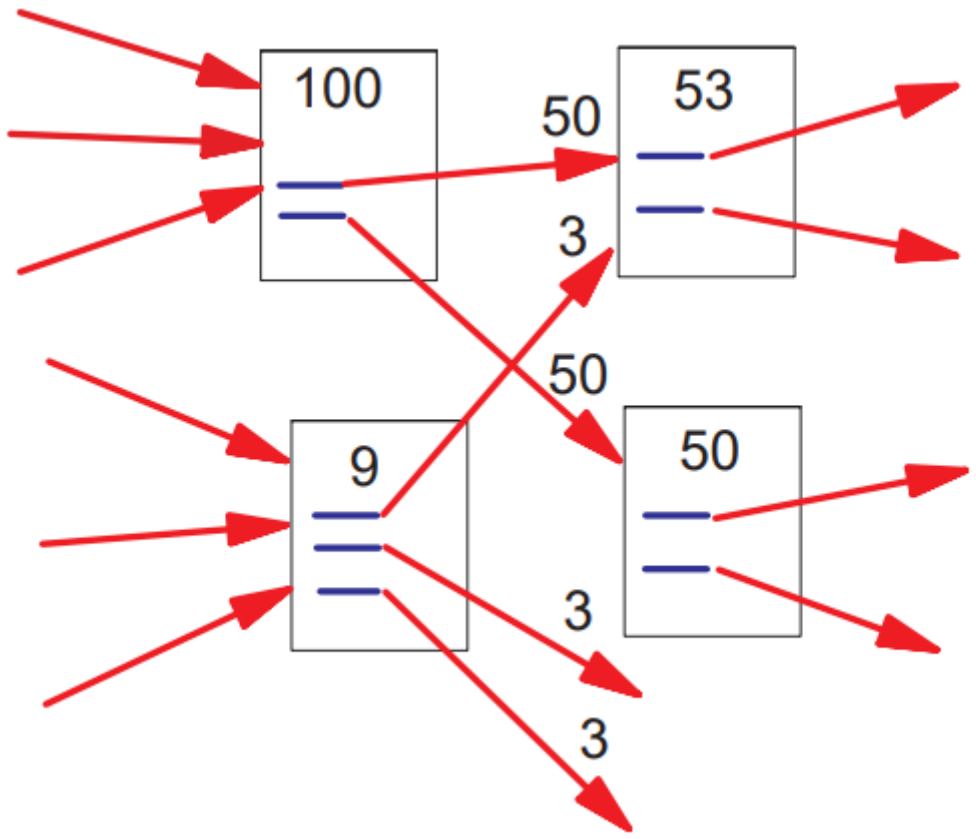
backlink가 얼마나 많이 있는가?

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

backlink가 얼마나 영향력 있는가?

참조를 얼마나 신중하게 했는가?

단순 예시



이미지 출처 - 본 논문

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

모순?

-마치 주주총회 같이 랭크가 더 높은 웹페이지의 링크가 랭크가 낮은 웹페이지보다 다른 웹페이지의 랭크를 결정하는데 더 큰 영향력을 끼침

-웹페이지의 랭크를 결정하는데 웹페이지의 랭크가 영향을 미친다? ?

-〉 순환적인 정의

-계산을 반복하다 보면 페이지랭크가 수렴함

행렬 이용

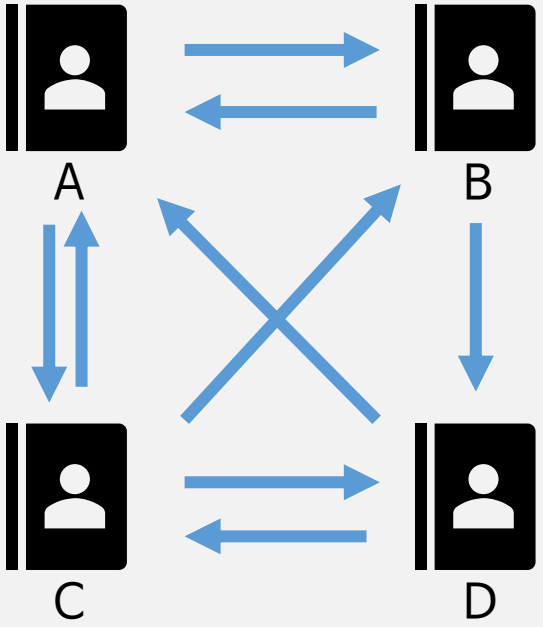
$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{nm} \end{pmatrix}$$

a_{nm}: n번째 페이지에서 m번째 페이지로 넘어갈 확률 즉, $\frac{1}{N_n}$

$$R_{i+1} = cAR_i$$

-i(계산횟수)를 늘려가다 보면 R 행렬이 수렴함

행렬 이용 예시

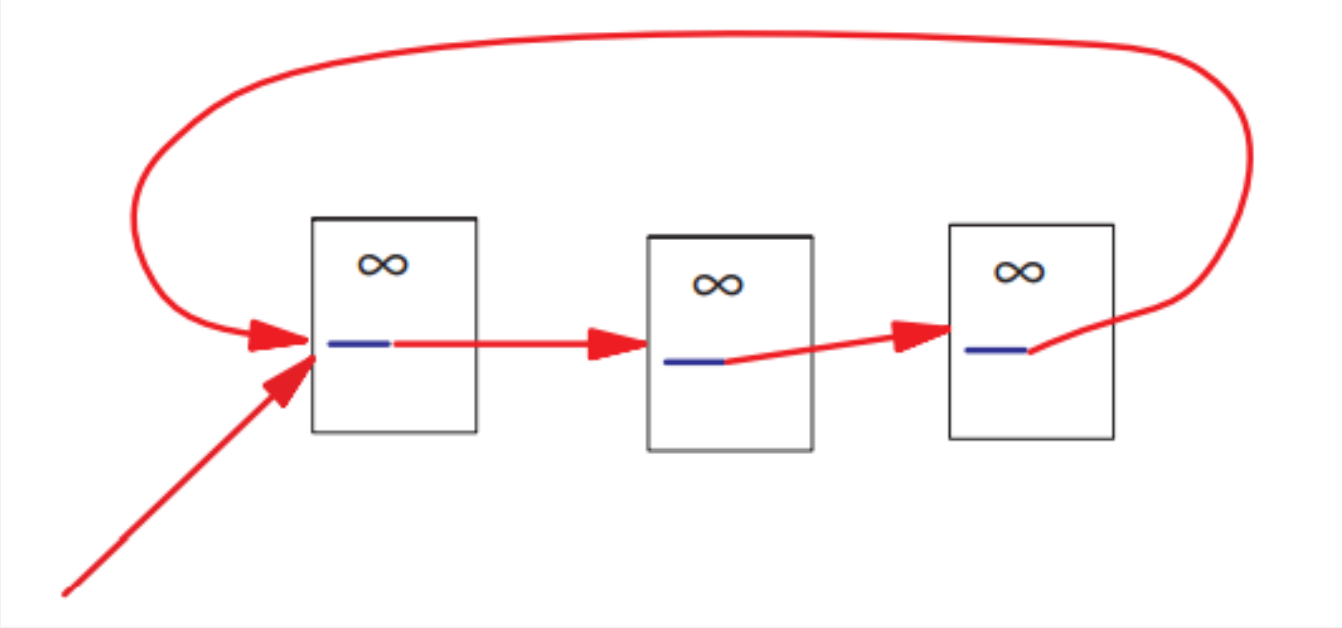


$$R_{i+1} = AR_i$$

-i를 늘려가다 보면 R 행렬이 수렴함

$$A = \begin{array}{c} \text{출발 노드} \\ \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{3} & 0 \end{bmatrix} \end{array}$$

단순화된 모델의 문제점



이미지 출처 - 본 논문

-루프가 생기면 페이지랭크 값이 수렴하지 않는 문제가 발생함.

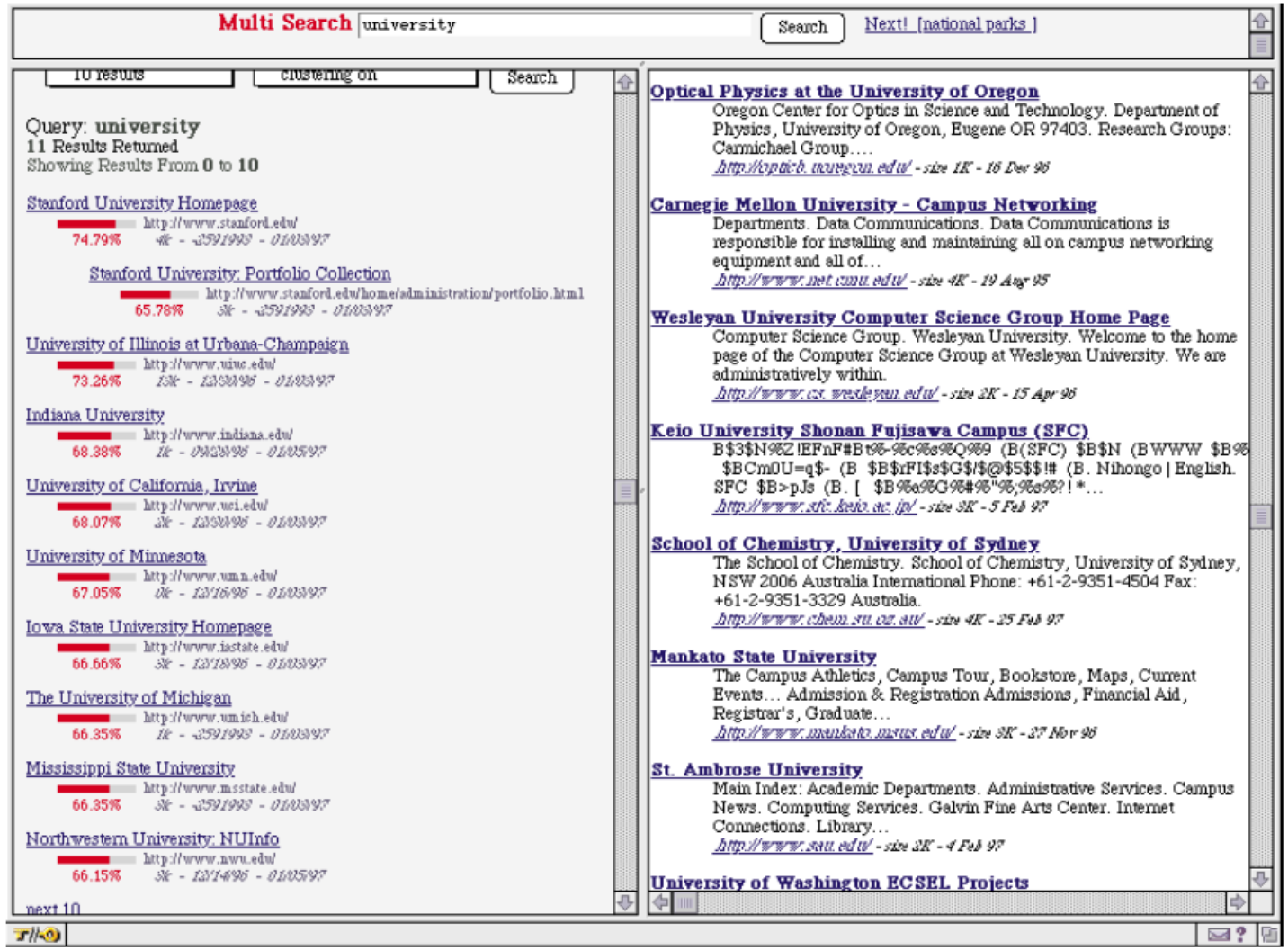
해결책: Random Surfer Model

-인터넷 서핑을 할 때 항상 링크를 타고 들어가는 것은 아니듯이, 링크를 타고 가다가 다른 임의의 웹페이지로 들어갈 수 있도록 만든 모델

$$R_{i+1} = c(AR_i + E)$$

- E 행렬은 이용자 맞춤형 페이지랭크를 만들어냄

검색 엔진 성능 비교



이미지 출처 - 본 논문, 왼쪽이 구글, 오른쪽이 Altavista

-구글이 실제 이용자가 찾을 만한 공식 홈페이지를 보여주는 반면, Altavista는 university가 들어간 페이지를 일관성 없게 보여줌.

-구글은 검색한 단어가 들어있는 페이지를 모두 찾은 후에, 페이지랭크에 따라 정렬함.

참고문헌

-Sergey Brin and Larry Page, The PageRank Citation Ranking: Bringing Order to the Web (1998)

-우쥔, 수학의 아름다움 (p.157-165), 세종서적㈜

감사합니다.