

PageRank:

Standing on the Shoulders of Giants

(2011)

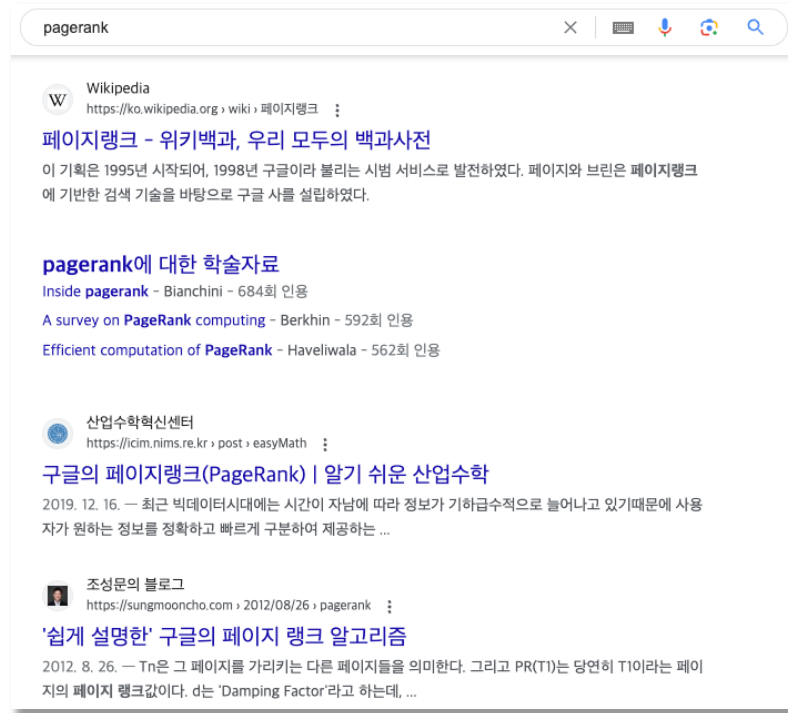
Communications of the ACM

목차

- 연구 배경
- PageRank를 이용한 웹페이지 랭킹
- PageRank vector 계산방법
- “Standing on the Shoulders of Giants”
- HITS
- 결론 및 요약

PageRank의 목적과 활용 예시

- 웹 페이지마다 정량적인 수치로 중요도를 측정
- 중요한 웹 페이지를 검색결과 상단에 위치시킬 수 있음



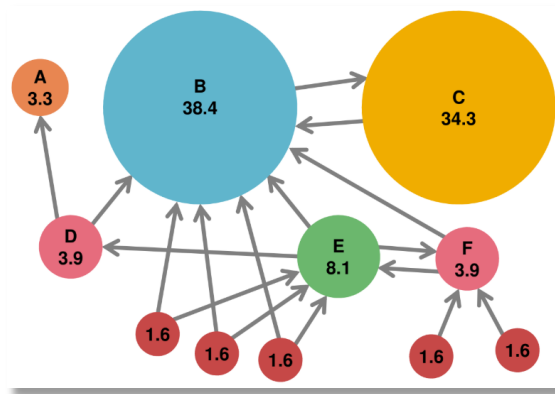
연구 배경

- **새로운 information system 웹 시스템의 등장**
 - 페이지 간에 하이퍼링크로 연결되어 있음
 - 커진 시스템의 규모
- **새로운 정보 검색 방법의 필요성 대두**
 - 하이퍼링크 연결정보를 활용할 수 있는 방법
 - 커진 규모에 맞게 빠르고 효율적인 방법이 필요

제안된 PageRank

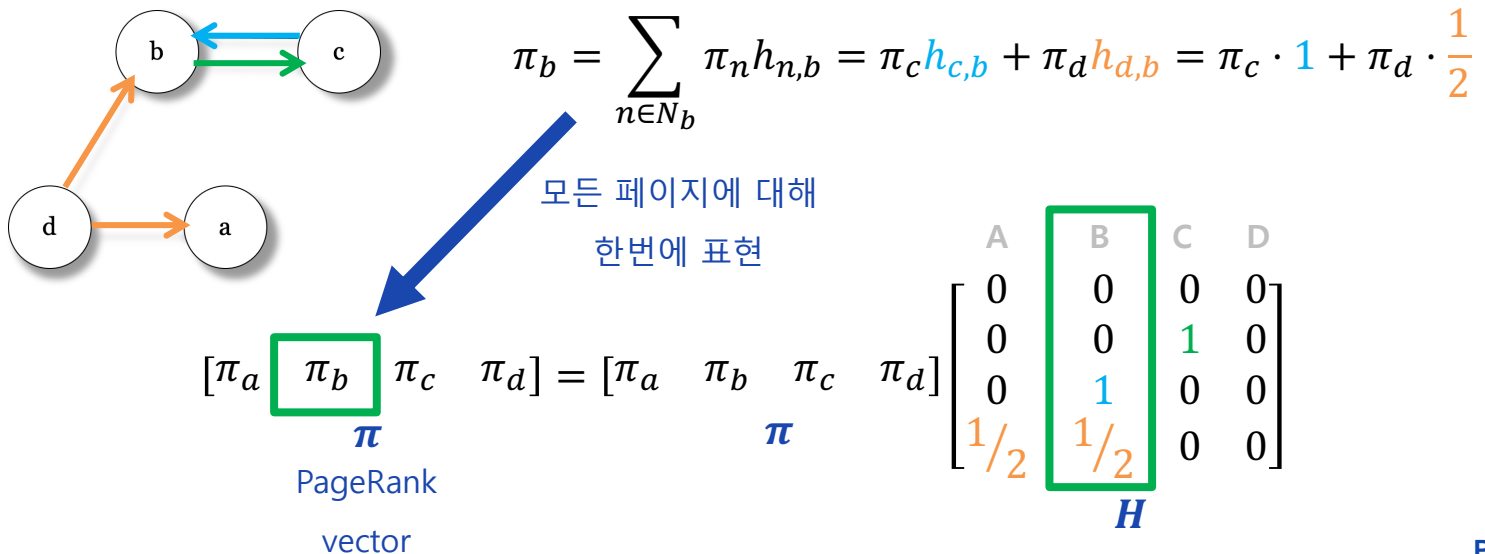
- PageRank

- 효율적이고 연결 정보를 활용하는 웹 페이지 중요도 측정 방법
 - ✓ 중요한 웹 페이지가 가리키는 웹 페이지가 중요할 것 → 순환 이론
 - ✓ 효율적인 계산방법을 사용
- 웹 시스템을 그래프 자료구조로 생각
 - ✓ 노드: 웹 시스템을 구성하고 있는 웹 페이지
 - ✓ 엣지: 웹 페이지 사이에 연결된 하이퍼링크



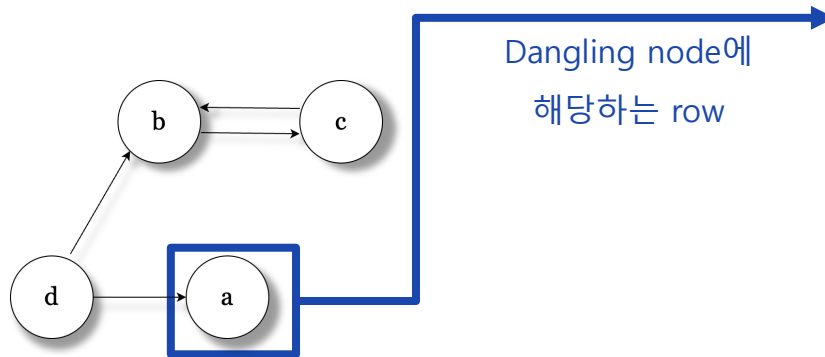
PageRank의 정의

- 가상의 유저인 surfer가 해당 페이지를 방문할 상대적인 빈도수
- 하이퍼링크를 타고 들어오는 페이지의 PageRank value의 weighted sum
 - 가중치 $h_{d,b}$: d에서 b로 이동할 확률, cf. $h_{d,b} = \frac{1}{2}$
 - π_b : 페이지 b의 PageRank value
 - $n \in N_b$: b로 하이퍼링크를 타고 들어오는 페이지 집합, 이에 속하는 페이지 n



Dangling node 문제와 해결방안

- Surfer가 다른 페이지로 나갈 수 있는 하이퍼링크가 없는 경우
 - PageRank vector $\pi = \pi H$
 - 행렬 H 안 특정한 row의 element가 모두 0
- Surfer가 모든 페이지에 임의의 확률로 이동하게 대체
 - PageRank vector $\pi = \pi S$



$$\begin{matrix} & \text{A} & \text{B} & \text{C} & \text{D} \\ \text{A} & 0 & 0 & 0 & 0 \\ \text{B} & 0 & 0 & 1 & 0 \\ \text{C} & 0 & 1 & 0 & 0 \\ \text{D} & 1/2 & 1/2 & 0 & 0 \end{matrix} = H$$

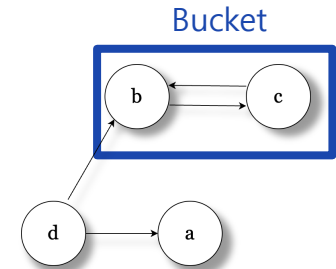


$$\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} = S$$

Bucket 문제와 해결방안

- 오직 cycle을 이루는 하이퍼링크가 존재하는 부분

- PageRank vector $\pi = \pi S$
- Surfer가 bucket 밖으로 나갈 수 없음



- Teleportation 행렬 E 와 더한 google 행렬 G

- Damping factor α : surfer가 하이퍼링크를 타고 페이지를 이동할 확률

$$G = \alpha * \underbrace{\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}}_S + (1 - \alpha) * \underbrace{\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}}_E$$

- 최종적으로 사용하는 google matrix G

- PageRank vector $\pi = \pi G$

세 가지 검증 사항과 첫 번째 검증

- 기존의 수학적 연구 토대를 통해 세 가지 질문에 답을 얻음
 - Q1: 식으로 얻을 수 있는 결과가 과연 존재하는가?
 - Q2: 유일한 결과를 구할 수 있는가?
 - Q3: 효율적인 계산으로 결과를 구할 수 있는가?
- 결과가 존재하는가?
 - Stochastic matrices의 convex combination은 stochastic
 - ✓ Stochastic matrix : 한 row(또는 column)의 elements 합이 1
 - ✓ Convex combination : 각 항에 숫자를 곱하고 결과를 더하는데 숫자의 합이 1
 - $G = \alpha S + (1 - \alpha)E$
 - 합이 1 stochastic
 - G 가 stochastic하면 식 $\pi = \pi G$ 를 만족하는 π 는 적어도 하나 이상 존재

두 번째 검증: 유일한 결과를 얻는가?

- Perron-Frobenius 정리 사용해 증명

- $rx = xA, x > 0$ 를 만족하는 행렬 A가 irreducible하고 음이 아닌 element로 채워져 있다면 유일한 벡터 x 가 존재
 - ✓ 조건1) $rx = xA$ 형태를 만족
 - ✓ 조건2) 행렬 A가 irreducible
 - ✓ 조건3) 행렬 A의 모든 elements가 0 또는 양수
- 식 $\pi = \pi G$ 는 세 조건을 모두 만족
 - ✓ 행렬 G는 stochastic $\rightarrow r=1$
 - ✓ 행렬 G는 strongly connected \rightarrow irreducible
 - ✓ 행렬 G는 stochastic \rightarrow 모든 elements가 $[0,1]$ 사이의 확률값
- PageRank vector π 는 유일

세 번째 검증: 효율적으로 구할 수 있는가?

- 효율성을 위해 행렬 G 가 아닌 행렬 H 를 저장
 - 페이지 수 만큼의 공간만 차지
 - 웹 시스템의 큰 규모를 고려해 적은 저장 비용
- Power method는 행렬 G 를 벡터 π 와 수렴할 때까지 곱함
 - 수렴할 때까지 행렬 G 을 곱하면서 벡터 π 를 업데이트
$$\begin{aligned}\pi(1) &= \pi(0)G \\ \pi(2) &= \pi(1)G \\ &\vdots \\ \pi(k+1) &= \pi(k)G\end{aligned}$$
 - 빠른 속도로 수렴

“Standing on the Shoulders of Giants”

- 선행된 연구들이 쌓아 놓은 지식에 올라서 PageRank가 탄생
 - Giants: Perron, Frobenius, von Mises 등의 PageRank 이전의 연구자들
 - Perron-Frobenius 정리, Power method 같은 연구를 토대로 세가지 검증 사항에 대한 검증을 수행할 수 있었음

Year	Author	Contribution
1906	Markov	Markov theory [19]
1907	Perron	Perron theorem [23]
1912	Frobenius	Perron-Frobenius theorem [6]
1929	von Mises & Pollaczek-Geiringer	Power method [30]
1941	Leontief	Econometric model [17]
1949	Seeley	Sociometric model [28]
1952	Wei	Sport ranking model [31]
1953	Katz	Sociometric model [10]
1965	Hubbell	Sociometric model [9]
1976	Pinski & Narin	Bibliometric model [25]
1998	Kleinberg	HITS [13]
1998	Brin & Page	PageRank [3]




Table 1: PageRank history.

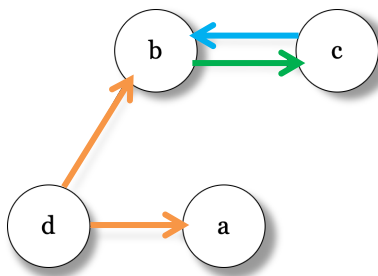
HITS

- 웹 페이지 랭킹 방법
- **Hubs / authorities : 웹 페이지마다 2가지 중요도로 측정**
 - Good authority : good hub가 가리키는 페이지
 - Good hub : good authority를 가리키는 페이지
 - 순환 이론

HITS의 수식적 정의

- Hub scores 와 authority scores

- L : 인접행렬(adjacency matrix)
 - ✓ 그래프의 구조를 표현하기 위한 정방행렬
- $\mathbf{y}^{(k-1)}$: hub vector , $\mathbf{x}^{(k)}$: authority vector
- $\mathbf{x}^{(k)} = L^T \mathbf{y}^{(k-1)}$: authority score는 받고있는 페이지 hub score의 합
- $\mathbf{y}^{(k)} = L \mathbf{x}^{(k)}$: hub score는 가리키는 페이지 authority score의 합



$$\begin{matrix}
 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} auth_a \\ auth_b \\ auth_c \\ auth_d \end{bmatrix} & = & \begin{bmatrix} 0 \\ auth_c \\ auth_b \\ auth_a + auth_b \end{bmatrix} & = & \begin{bmatrix} hub_a \\ hub_b \\ hub_c \\ hub_d \end{bmatrix} \\
 L & \mathbf{x}^{(k)} & & & & \mathbf{y}^{(k)}
 \end{matrix}$$

결론 및 요약

- **결론**

- PageRank는 사용자 수백만명이 직접 구성하고 이용하는 웹에 대한 집단지성 평가방식
- 2011년까지도 PageRank 기반의 방법, 유사한 방법들에 대한 연구가 진행되고 있음

- **요약**

- PageRank는 웹 페이지 랭킹 알고리즘
 - ✓ 페이지간의 연결정보를 활용 / 웹 규모에 맞게 효율적인 방법
- PageRank의 검증과정과 계산 방법
 - ✓ 존재성 / 유일성 / 효율적인 계산 방법
- HITS는 Hubness와 Authority를 사용하는 웹 페이지 랭킹 알고리즘