PageRank

국민대학교 소프트웨어학부

1. 도입과 동기

- 웹은 매우 거대하고 이질적
  - 약 1억5천만 페이지 이상이 존재, 매년 2배씩 증가하는 추세
- 웹 페이지의 다양성
- 웹에 익숙지 않은 초보자들을 고려해야 함
- 검색 엔진의 랭킹 기능을 교묘하게 이용하려는 페이지들로부터 비롯되는 문제점이 있음
- 웹 페이지는 평면적이지 않고, 하이퍼텍스트가 존재함

2. Google의 PageRank 설명

- Page A에서 page B로 연결하는 링크 → vote (투표)
- Hyper link가 많은 페이지 → “중요하다”고 평가
- “중요하다”고 평가된 페이지로부터 link된 page → 더욱 중요하게 평가
- 모든 page에 대해 PageRank 계산

3. PageRank algorithm(1)

- Web의 Link 구조
  - Forward/Out link: 페이지 밖으로 나가는 링크
  - Back/In link : 그 페이지를 가리키는 링크
- 일반적으로 In link 많은 페이지가 그렇지 않은 페이지보다 중요함
- 하지만 In link 개수가 페이지의 중요성과 일치하지 않는 경우가 많음
- PageRank : 어떤 페이지의 In link 개수 및 높은 랭크값 In link를 갖는 경우를 포괄하여 페이지의 랭킹 계산

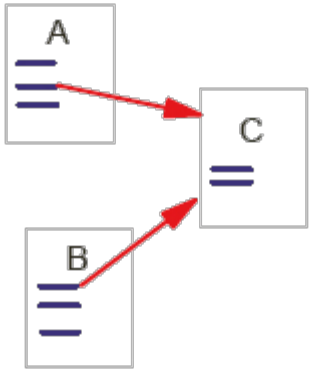


Figure 1: A and B are Backlinks of C

3. PageRank algorithm(2)

- PageRank의 정의
  - Simple ranking R : 앞에서 말한 내용을 기초로 PageRank를 단순화 시킨 버전

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

u : 어떤 페이지

Fu : u가 가리키는 페이지들의 집합

Bu : u를 가리키는 페이지들의 집합

Nu : u로부터 나가는 링크의 개수, Fu의 개수

c : normalization에 사용되는 factor

(전체 웹 페이지의 랭크 총합을 일정하게 하기 위함)

3. PageRank algorithm(3)

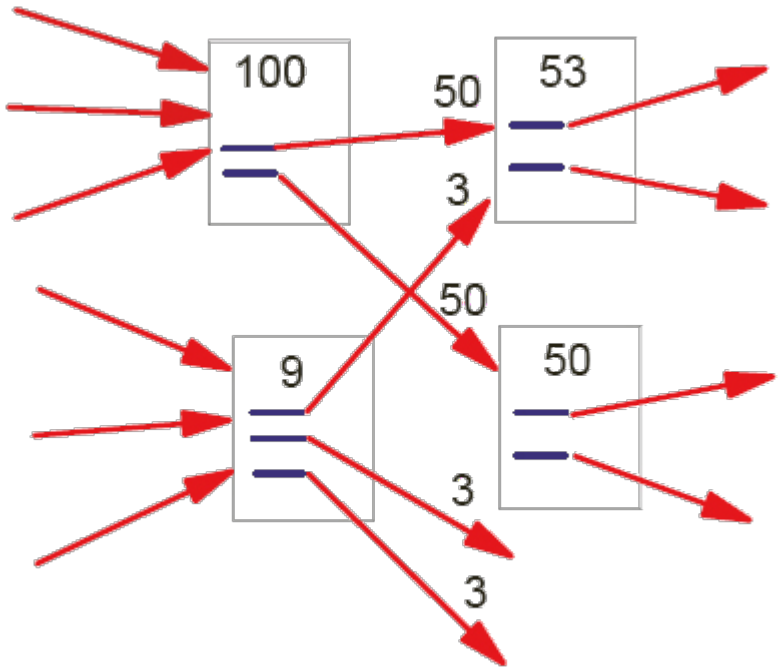


Figure 2: Simplified PageRank Calculation

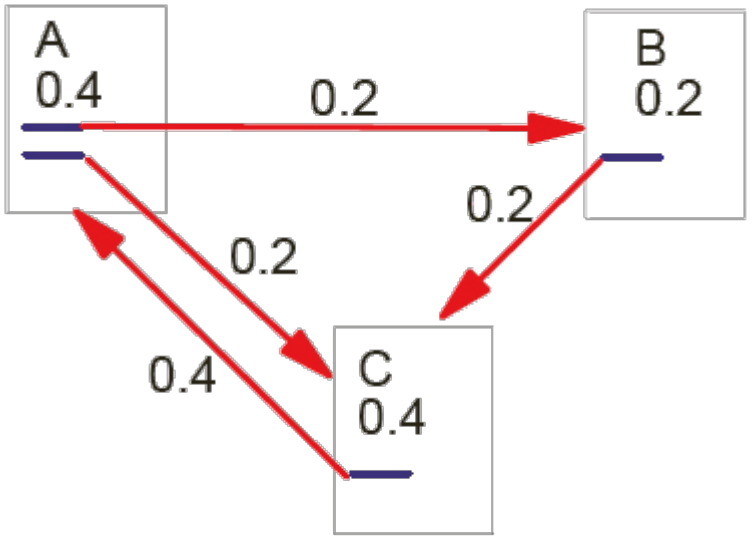


Figure 3: Simplified PageRank Calculation

3. PageRank algorithm(4)

- Rank sink 문제
  - 어떤 페이지들 사이에서 외부로 나가는 링크가 없어 루프 돌며 랭크가 계속 축적되는 것

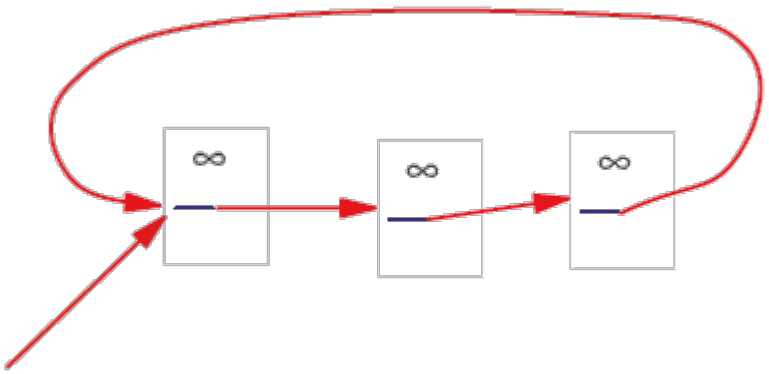


Figure 4: Loop Which Acts as a Rank Sink

3. PageRank algorithm(5)

- Random Surfer Model
  - Random Surfer : 무작위로 일련의 링크들을 무작위로 클릭하는 사람
  - 실제 웹 써퍼가 루프에 빠지면 계속해서 링크를 클릭하지 않고, 다른 페이지로 점프하려 할 것이다. E는 이러한 행동을 나타내는 factor임

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

3. PageRank algorithm(6)

- Dangling Links 문제
  - 외부로 나가는 링크가 없는 페이지를 가리키는 링크
  - 이것의 가중치가 어디로 분산되고 있는지가 불분명하기 때문
  - 다른 페이지의 랭킹에 직접적인 영향을 주지 않기 때문에 모든 PageRank 계산이 다 끝난 뒤에, Dangling link를 첨가하여 계산하여 해결

3. PageRank algorithm(7)

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + .. + \frac{PR(Tn)}{C(Tn)}\right)$$

- A : 어떤 웹 페이지
- T1, T2, … Tn : A를 가리키는 페이지
- d : damping factor(0<d<1), 보통 d = 0.85
- C(A) : A의 Forward link의 갯수

- PR(A)는 단순 반복 알고리즘으로 계산
  그 값은 웹 링크를 normalize해서 행렬로 바꾸었을 때
  주 고유벡터(principal eigenvector)에 해당

4. PageRank 구현

- 각 URL을 고유한 정수(page ID)로 변환하고, 모든 하이퍼링크를 page ID를 이용해서 데이터베이스에 저장

  1. Parent ID를 이용하여 링크 구조를 정렬
  2. Dangling Link 제거 : 반복 작업을 통해 제거
  3. 랭크값 초기화 : 초기값은 결과와 무관. 수렴 속도는 달라짐.
  4. 반복 계산 : 현재 진행 중인 계산의 가중치는 메모리에 저장, 전 단계의 가중치는 디스크에 저장
  5. 가중치들이 수렴하면, dangling link 추가하여 랭킹을 다시 계산(이때 반복횟수는 제거시 반복 횟수와 동일함)

- 2,400만개 페이지의 경우, 워크스테이션에서 약 5시간 소요

5. PageRank의 수렴(1)

- PageRank 계산을 반복적으로 계속하면 모든 페이지 랭크의 평균이 1로 수렴.
- 수렴시키는데 필요한 반복작업 시간은 대략 log n

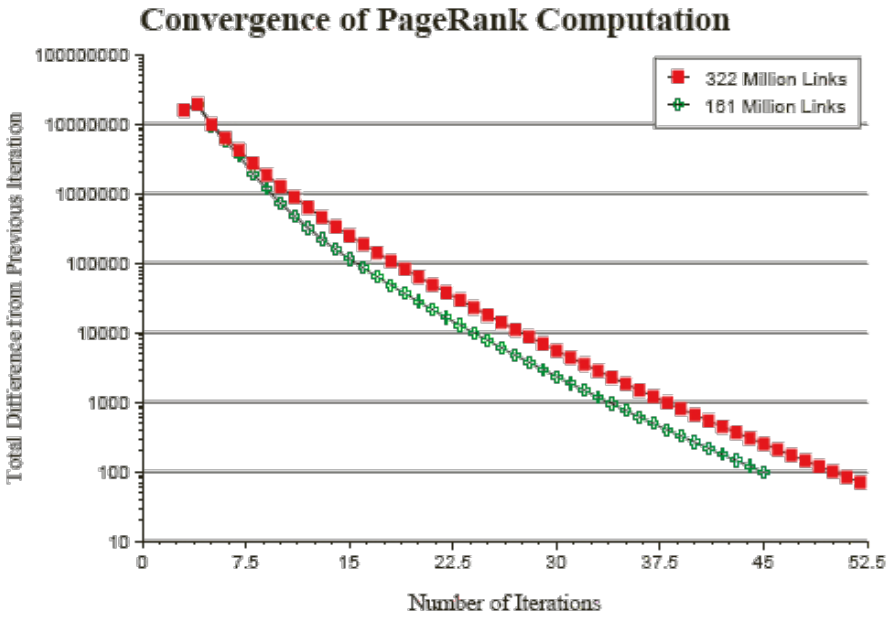
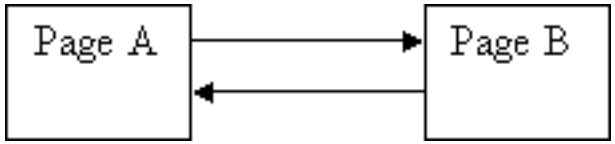


Figure 5: Rates of Convergence for Full Size and Half Size Link Databases

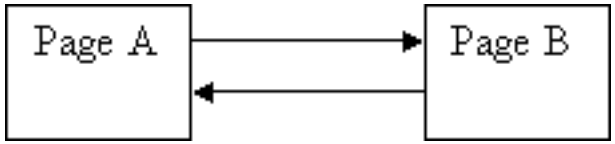
5. PageRank의 수렴(2)



- Guess 1(초기값을 1.0으로 했을 때)
  - d= 0.85
    PR(A)= (1 – d) + d(PR(B)/1)
    PR(B)= (1 – d) + d(PR(A)/1)

    i.e.
    PR(A)= 0.15 + 0.85 * 1 = 1
    PR(B)= 0.15 + 0.85 * 1 = 1

5. PageRank의 수렴(3)



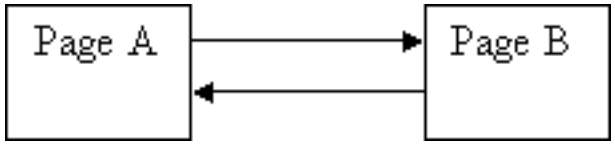
- Guess 2(초기값을 0.0으로 했을 때)
  - PR(A)= 0.15 + 0.85 * 0 = 0.15
    PR(B)= 0.15 + 0.85 * 0.15 = 0.2775

  두번째 반복:
    PR(A)= 0.15 + 0.85 * 0.2775 = 0.385875
    PR(B)= 0.15 + 0.85 * 0.385875 = 0.47799375

  세번째 반복:
    PR(A)= 0.15 + 0.85 * 0.47799375 = 0.5562946875
    PR(B)= 0.15 + 0.85 * 0.5562946875 = 0.622850484375
    …

  PageRank의 평균이 1에 수렴

5. PageRank의 수렴(4)



- Guess 3(초기값을 40.0으로 했을 때)
  - PR(A)= 0.15 + 0.85 * 40 = 34.25
    PR(B)= 0.15 + 0.85 * 0.385875 = 29.1775

두번째 반복:

PR(A)= 0.15 + 0.85 * 29.1775 = 24.950875
PR(B)= 0.15 + 0.85 * 24.950875 = 21.35824375
...

PageRank의 평균이 1에 수렴

6. PageRank를 이용한 검색(1)

- 1600만 페이지 제목만을 이용한 검색 테스트
- University라는 검색어에 대해 타이틀 검색 결과

  좌 : 페이지랭크
  우 : 알타비스타

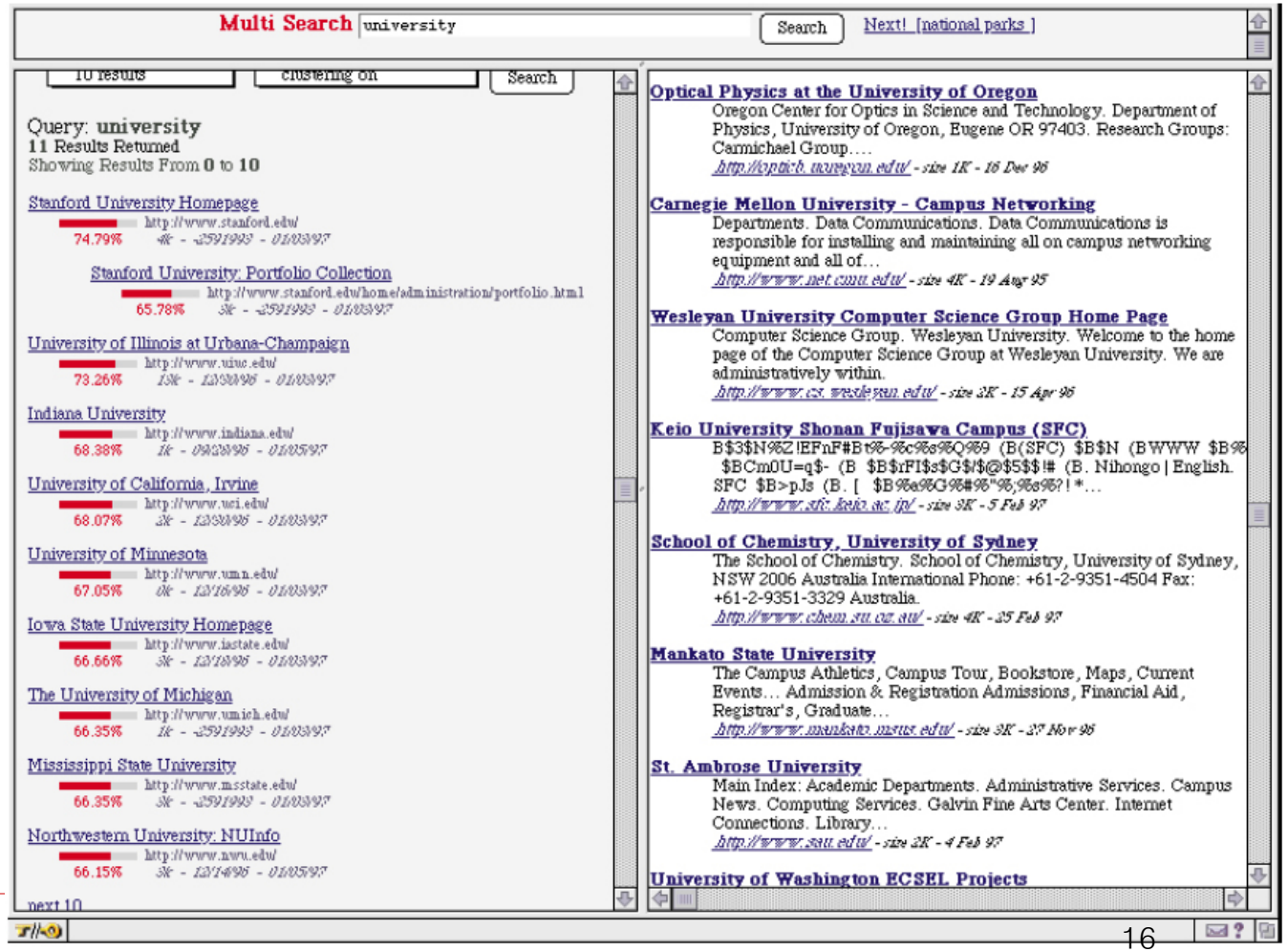


Figure 6: Comparison of Query for "University"

6. PageRank를 이용한 검색(2)

- PageRank에 기반한 Top 15위 페이지 목록

Web Page	PageRank (average is 1.0)
Download Netscape Software	11589.00
http://www.w3.org/	10717.70
Welcome to Netscape	8673.51
Point: It's What You're Searching For	7930.92
Web-Counter Home Page	7254.97
The Blue Ribbon Campaign for Online Free Speech	7010.39
CERN Welcome	6562.49
Yahoo!	6561.80
Welcome to Netscape	6203.47
Wusage 4.1: A Usage Statistics System For Web Servers	5963.27
The World Wide Web Consortium (W3C)	5672.21
Lycos, Inc. Home Page	4683.31
Starting Point	4501.98
Welcome to Magellan!	3866.82
Oracle Corporation	3587.63

Table 1: Top 15 Page Ranks: July 1996