ZeroPage OMS

구글에 질서를 부여하는 검색엔진 맛보기

최유경

검색 엔진이란?

- 검색 엔진

웹에서 정보를 수집해 검색 결과를 제공하는 프로그램

- 검색 시스템

검색엔진을 기반으로 구축된 시스템

- 검색 서비스

검색 시스템을 활용해 검색 결과를 서비스로 제공하는 것

구글 검색시 첫 페이지에 제시되는 10개의 결과

첫 페이지를
92% 확률로
사용자가 클릭

30조에서 중요한
10개 고르는 시간
:평균 0.5초

$$\frac{10}{30,000,000,000,000}$$

크롤링(Crawling)

- 웹사이트 등의 정보 자원을 자동화된 방법으로 수집, 분류, 저장하는 것.

1. 크롤러 프로그램은,
2. 주어진 인터넷 주소에 접근해서,
3. 관련된 다른 주소를 찾아내고,
4. ‘색인(index)’이라고 하는 웹페이지 목록을 만들어서,
5. 데이터베이스에 저장

과거의 검색엔진(단어 검색)

- 1996 주로 검색어와 웹페이지 본문 내용을 비교하는 방식
  ->웹페이지 소유자들이 인기 검색어를 자기 웹페이지에 몰래 심어놓음

- 웹페이지의 바탕화면과 같은 색의 글자로 인기 검색어를 많이 작성.
- 웹페이지의 소스 코드에 인기 검색어를 많이 작성.

-> 그 검색어를 검색했을 때 위의 웹페이지가 상위 순위로 뜸. ㅠㅠ

페이지랭크(PageRank)

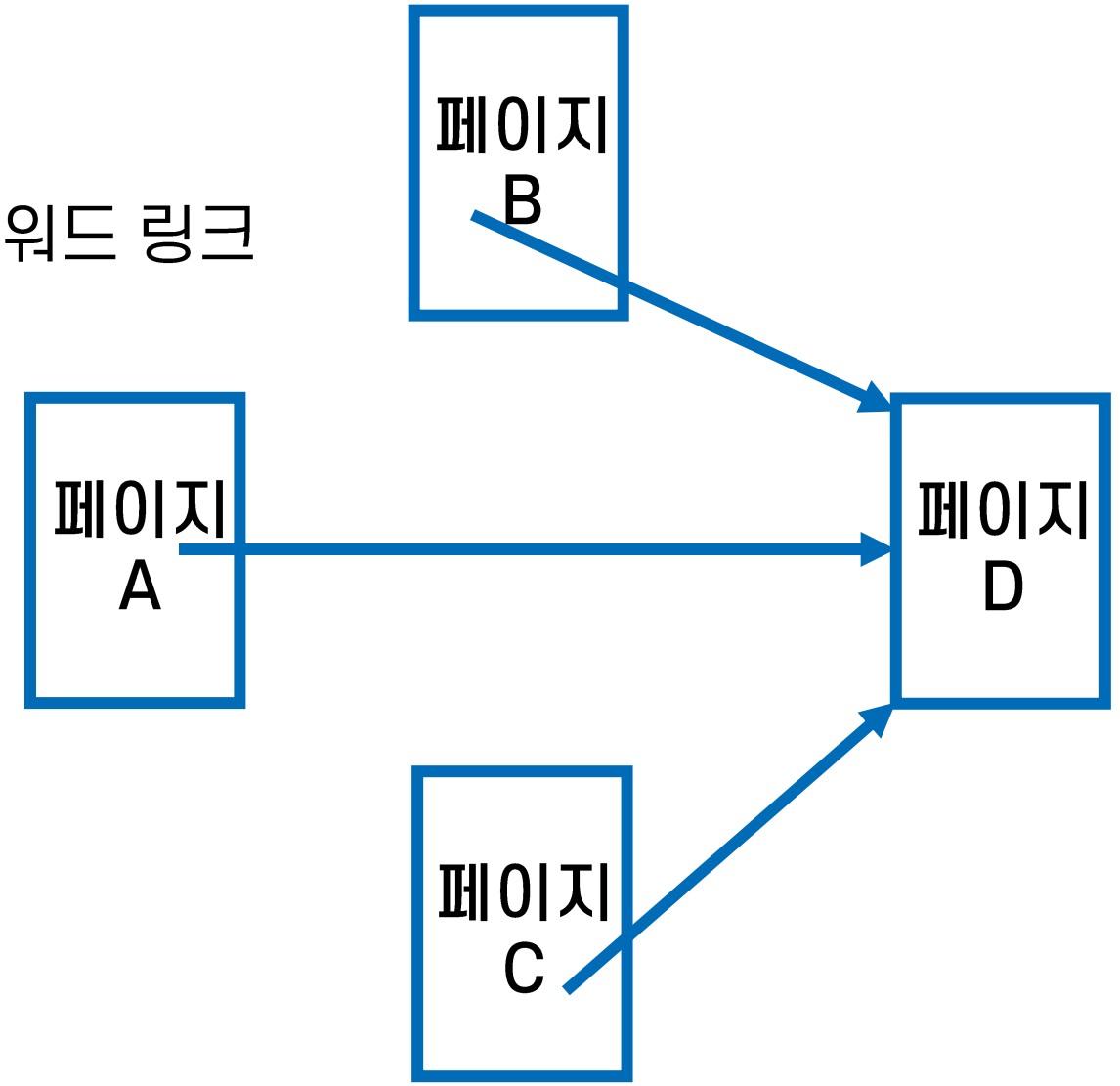
- 구글 창립자인 래리 페이지와 세르게이 브린에 의해 개발
- 웹 그래프를 기반으로 웹 페이지의 순위를 매겨 랭킹하는 알고리즘
  - 내 블로그를 조회수 0인 블로그가 링크하면?
  - 내 블로그를 파워블로그가 링크하면?

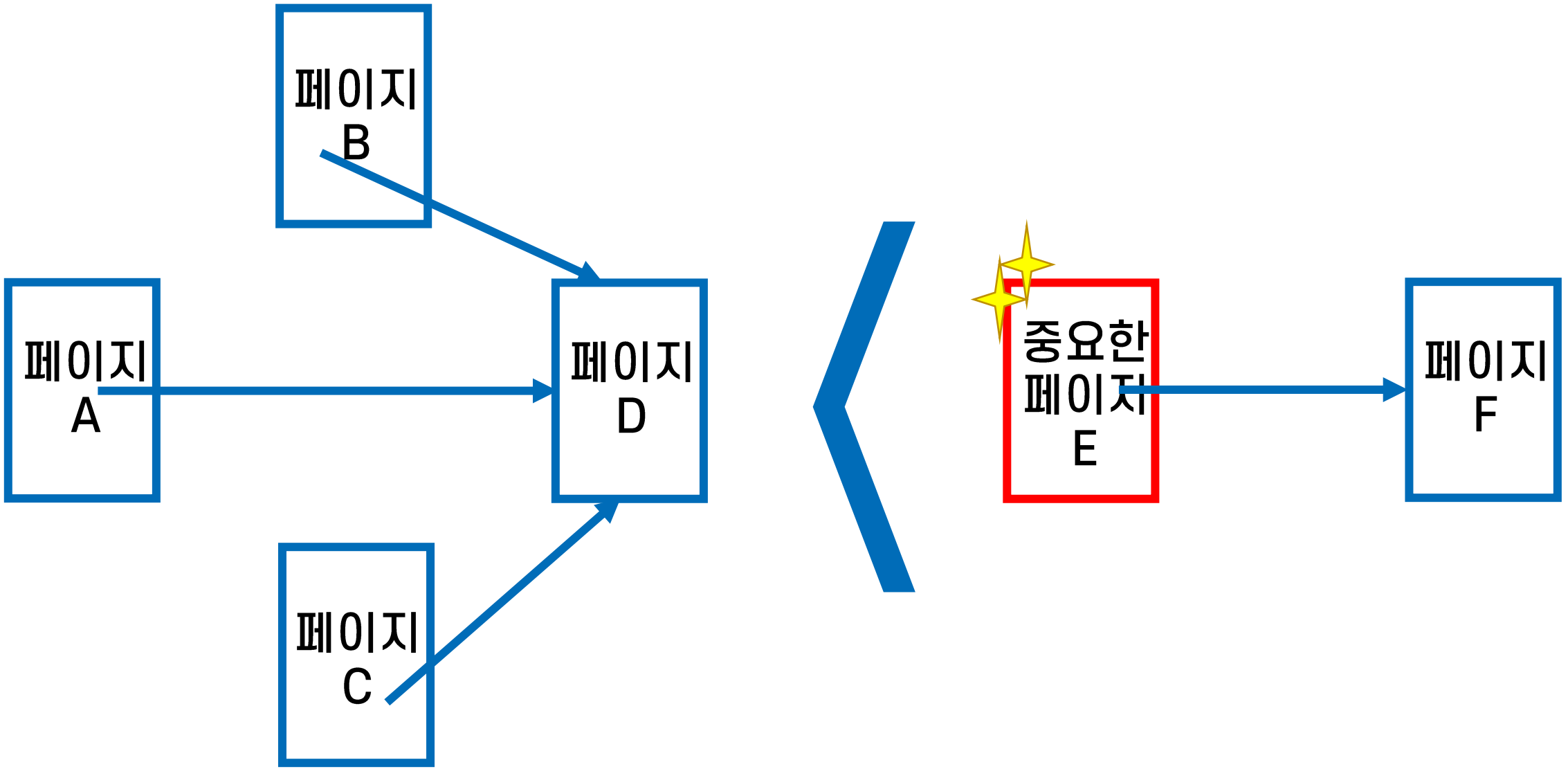
-> 여러 블로그가 링크할수록, 파워블로그가 링크할수록
나의 블로그 중요도 점수가 많이 올라감

백 링크(Backlink)

- 웹페이지에서 밖으로 나가는 포워드 링크
- 그 페이지를 가리키는 백링크

- A, B, C는 D의 백링크




페이지
B
페이지
A
페이지
D
페이지
C
중요한
페이지
E
페이지
F

목표 : 많은 백링크와 높은 랭크값 백링크를 찾는 웹페이지 찾기

페이지랭크의 정의

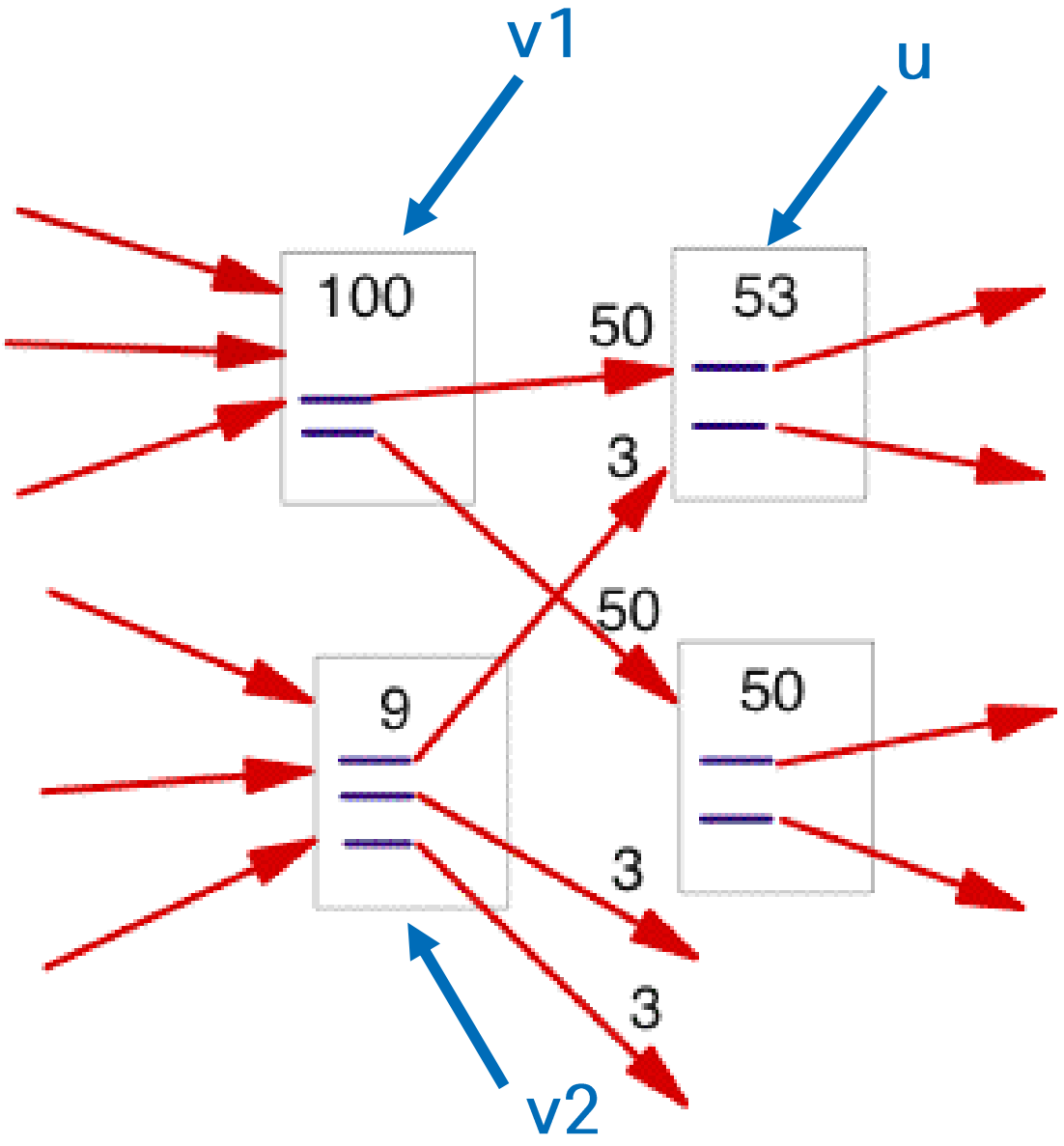
- 단순 페이지랭크

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

- u : 웹 페이지
- Nu : u 페이지로부터 나가는 링크의 개수
- Bu : u 페이지를 가리키는 페이지의 집합
- 전체 웹 페이지의 랭크 총합을 일정하게 하기 위한 c

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

- R(u) = R(v1)/Nv1 + R(v2)/Nv2
  = (100/2) + (9/3)
  = 53

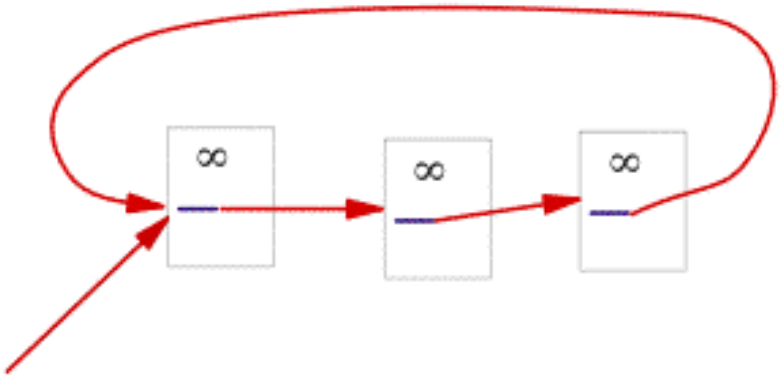


식 뜯어보기

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

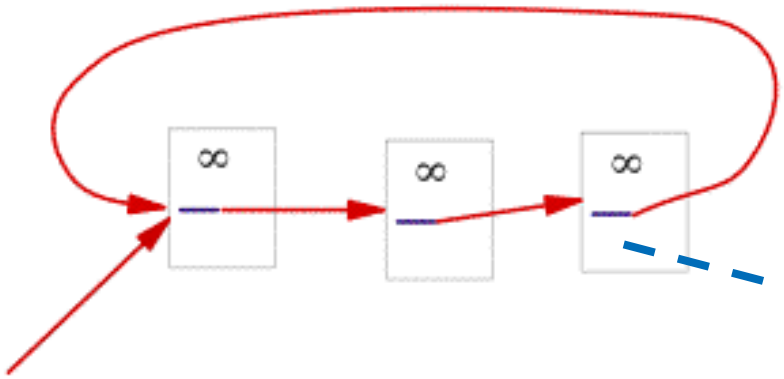
문제점 : 랭크 싱크(rank sink)

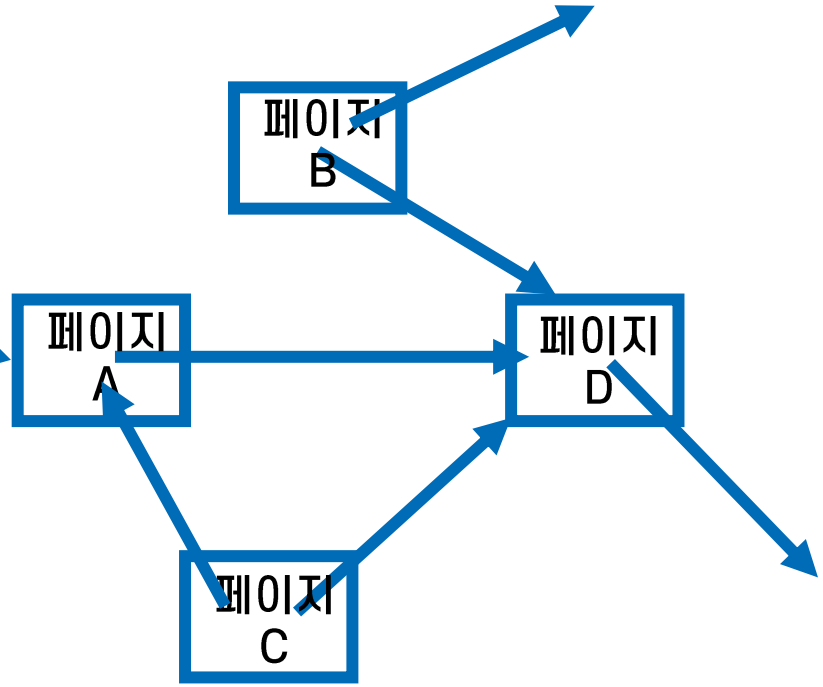
- 두 페이지가 서로 가리키고 있으며 다른 페이지로 연결되어 있지 않은 경우
- 반복연산이 진행되면서 그 루프에서는 랭크가 계속 축적될 뿐 외부로 전혀 분산하지 못한다.



->랜덤 서퍼 모델로 해결

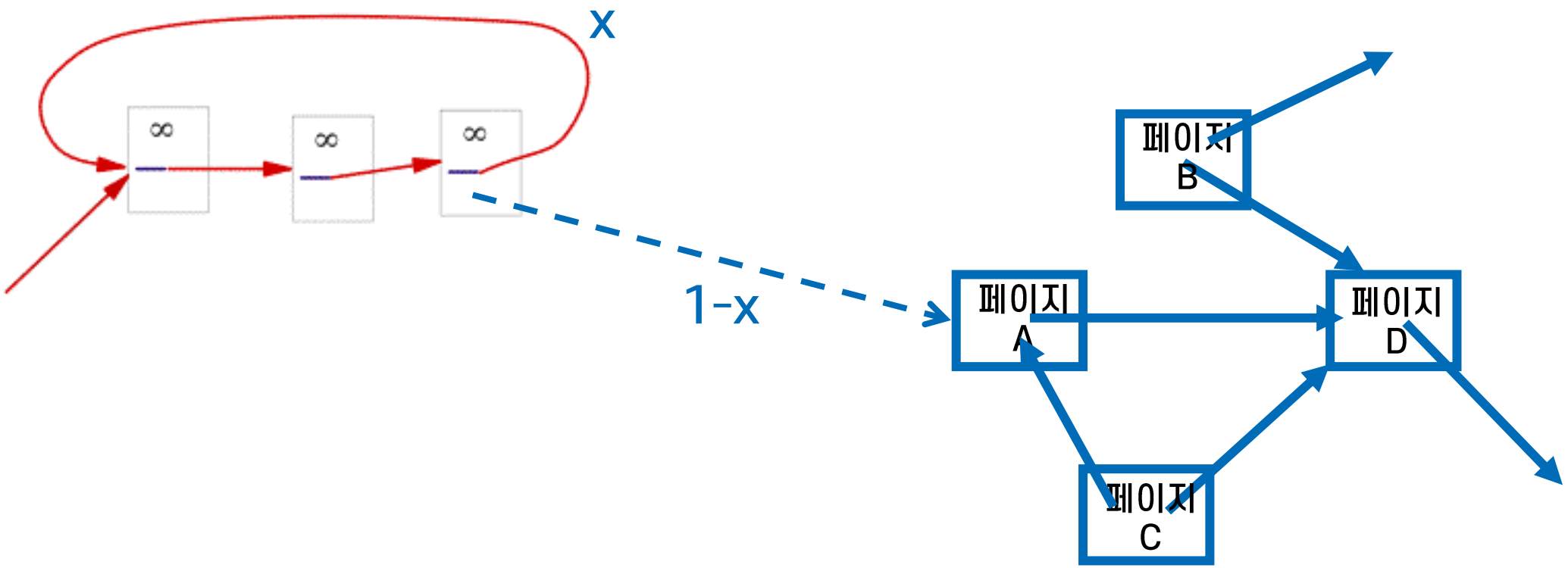
랜덤 서퍼 모델(Random surfer model)





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

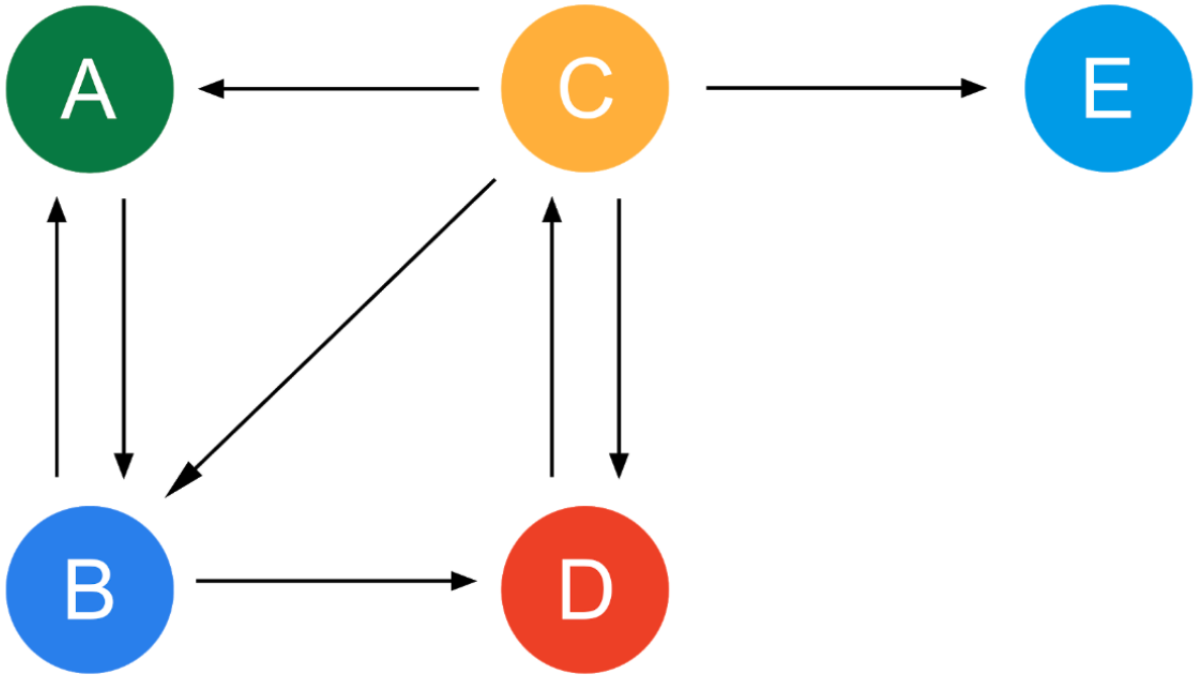
랜덤 서퍼 모델(Random surfer model)



- 무작위 페이지로 점프하는 특정 확률이 더해짐

문제점 : Dangling link

- E는 외부로 나가는 링크가 없다.
- 논문에서는 제외하고 계산



페이지랭크 구현하기 전에…

- 크롤러를 이용해 데이터베이스에 각 웹페이지의 정보를 저장
  - URL을 유일한 정수로 바꾸기
  - 링크를 정수 ID를 이용해서 데이타베이스에 저장하기

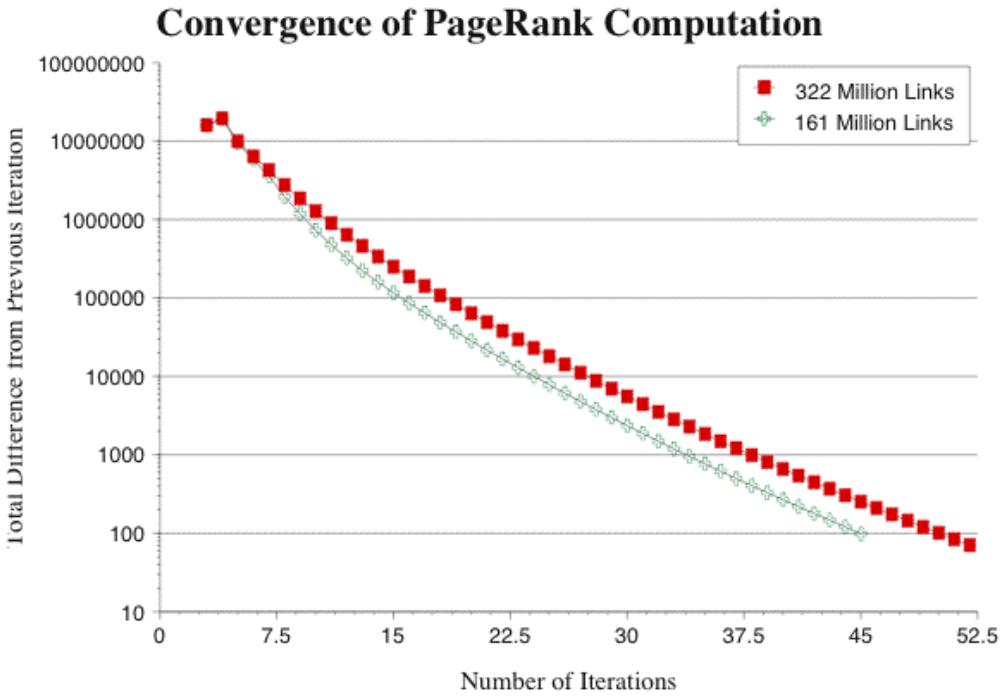
페이지랭크 구현하기

1. ID를 이용해서 링크 구조를 정렬
2. 링크 데이타베이스에서 댕글링 링크를 제거
3. 랭크 값을 초기화
4. 가중치가 수렴할 때까지 반복계산(각 페이지의 가중치에 메모리를 할당)
5. 다시 댕글링 링크를 추가하고, 랭킹을 재연산

페이지랭크의 수렴

- 받은 초기값으로 페이지랭크를 수렴할 때까지 계산
- 초기값과 수렴한 페이지랭크 계산 결과는 관련없고, 초기값과 속도와는 관련 있음

- 속도면에서 큰 장점



페이지랭크(Pagerank)

- 백링크를 피어리뷰의 역할로 이용한다.
- 충분히 시간이 흐른 뒤에 랜덤 워크가 그 노드에 있을 확률

- 제목: 구글에 질서를 부여하는 검색엔진 맛보기
- The PageRank Citation Ranking : Bringing Order to the Web

검색엔진 최적화(SEO)

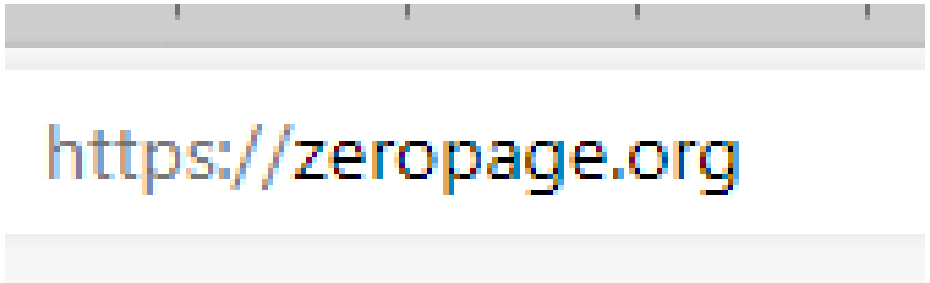
- 여러분의 웹사이트를 검색엔진 검색 결과 첫페이지로 이전시키는 작업
- 타깃고객에게 웹페이지의 콘텐츠를 효과적으로 상위 노출하기 위한 전략적이고 기술적인 작업

- 좋은 글을 올리는게 가장 좋지만… 검색결과에서 광고를 클릭하는 비율 3%
- 현대 디지털 마케팅의 핵심

SEO의 준비

- 페이지의 로딩 속도 올리기
- URL의 길이가 너무 길지 않게 하기
- 타 사이트에서 적당히 링크받기
- 보안에 신경쓰기

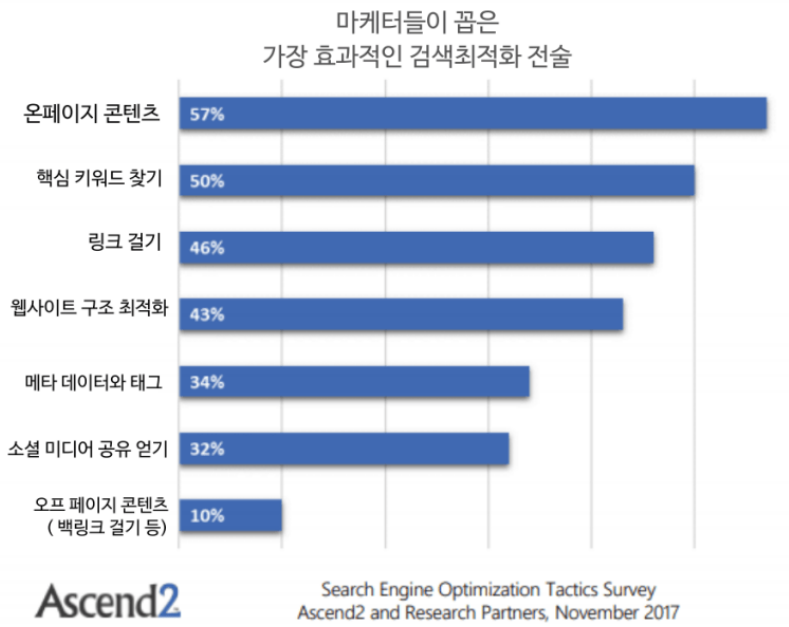
- 제로페이지의 웹페이지는?



SEO의 핵심-콘텐츠와 키워드

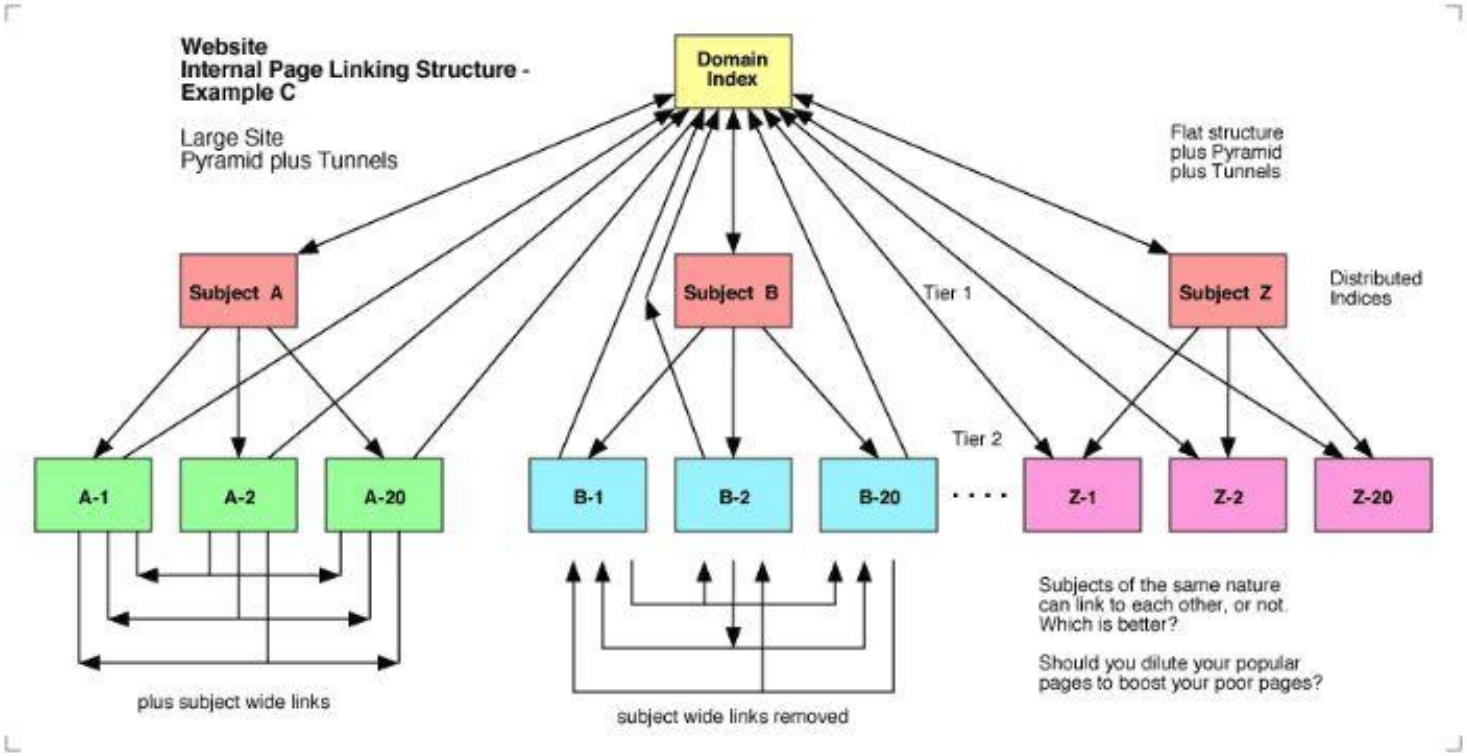
- 자세한 키워드 설정(대회->동아리 대회->동아리 알고리즘 대회)
- 제목 또는 인터넷 주소에 키워드(https://zeropage.org/동아리-알고리즘-대회)
- 부제목에도 키워드
- 이미지의 파일명에도 키워드

콘텐츠가 검색엔진최적화에 가장 중요하다



SEO의 핵심-링크걸기

- 핵심 콘텐츠를 선정하고 연관성을 가진 콘텐츠로 링크 걸기



감사합니다.

논문 제목: The PageRank Citation Ranking : Bringing Order to the Web