PageRank Algorithm

전북대학교 컴퓨터공학부 정보검색 1분반(이경순 교수)

2019-10-17

컴퓨터공학부 201514740 이동준

Index

전북대학교 컴퓨터공학부 정보검색 1분반(이경순 교수) RankPage Algoritm

What is PageRank?

정의와 배경

정의

월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 **가중치**를 부여하는 방법이다.

이 알고리즘은 서로간에 인용과 참조로 연결된 임의의 묶음에 적용할 수 있다.

연구 배경

문제점 : 정보량은 증가하고 검색결과에 부정확한 정보도 많았다.

근본적인 사항 : 사람의 목적이 모두 다르기에 **" 중요성 "** 을 모두 만족 시켜줄 수 없다.

⇒ 그렇다면?

페이지 내부에 연결된 **링크**를 가지고 중요도를 평가해 보자

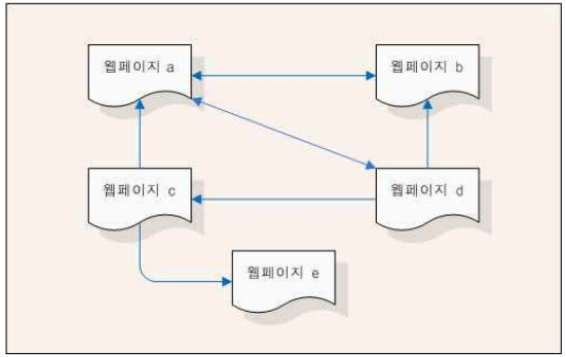
PageRank Algorithm

알고리즘 분석

선형대수학을 이용한 구글 PageRank 분석

확률 행렬화 (마코브 행렬)

1단계: 인접(adjacency) 행렬의 건설



$$A=\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

2단계: 열정규화된 인접행렬 H의 건설

$$H_j = \frac{A_j}{\sum_{k=1}^{n} A_{kj}}, \quad j=1,\cdots,n$$

$$H=\begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad (3)$$

가중치가 '1'이 되게 만들자

3단계: 행렬 H의 열 stochastic화

전제 조건 1 : 모든 원소의 값은 0보다 크거나 같다.
전제 조건 2 : 칼럼 원소의 합은 1이다.

$$S = H + \frac{\mathbf{e}\mathbf{a}^T}{n}$$

$$\sum_{i=1}^{n} H_{ij} = 0 \quad a_j = 1 \qquad \sum_{i=1}^{n} H_{ij} \neq 0 \quad a_j = 0$$

$$S=\begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} & \frac{1}{5} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{5} \end{bmatrix}$$

irreducible인 동시에 stochastic

PageRank Algorithm

알고리즘 분석

선형대수학을 이용한 구글 PageRank 분석

"선형대수학과 구글(Google) 검색엔진" - 페이지랭크 알고리즘 이상구 (성균관대학교)1)

Irreducible : 기약(나누어지지 X)

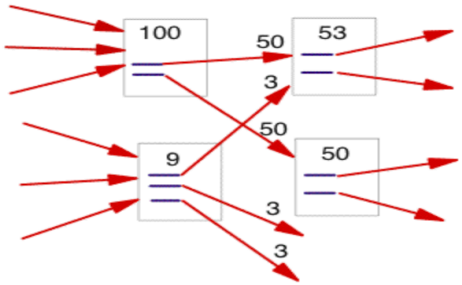


그림 2 단순화된 페이지랭크의 계산

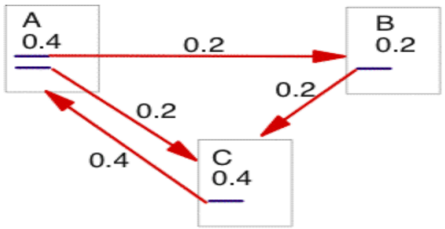


그림 3 정상상태를 이루고 있는 페이지들

4단계: 구글행렬 G의 건설

$$G = mS + (1-m)E \qquad (5)$$

m은 $0 \le m \le 1$이고 $E = \frac{ee^T}{n}$이다.

$$\mathbf{x}_{k+1} = G\mathbf{x}_k$$

단계 5: 거듭제곱법(Power method)

$$\begin{aligned}
\mathbf{x}_k &= G\mathbf{x}_{k-1} \\
&= \left(mS + (1-m)\frac{\mathbf{e}}{n}\mathbf{e}^T\right)\mathbf{x}_{k-1} \\
&= mS\mathbf{x}_{k-1} + (1-m)\frac{\mathbf{e}}{n}\mathbf{e}^T\mathbf{x}_{k-1} \\
&= mS\mathbf{x}_{k-1} + (1-m)\frac{\mathbf{e}}{n} \\
&= m\left(H + \frac{\mathbf{e}\mathbf{a}^T}{n}\right)\mathbf{x}_{k-1} + (1-m)\frac{\mathbf{e}}{n} \\
&= mH\mathbf{x}_{k-1} + \frac{\mathbf{e}}{n}\left(m\mathbf{a}^T\mathbf{x}_{k-1} + (1-m)\right)
\end{aligned}$$

$$G = \begin{bmatrix}
\frac{3}{100} & \frac{22}{25} & \frac{91}{200} & \frac{47}{150} & \frac{1}{5} \\
\frac{91}{200} & \frac{3}{100} & \frac{3}{100} & \frac{47}{150} & \frac{1}{5} \\
\frac{3}{100} & \frac{3}{100} & \frac{3}{100} & \frac{47}{150} & \frac{1}{5} \\
\frac{91}{200} & \frac{3}{100} & \frac{3}{100} & \frac{3}{100} & \frac{1}{5} \\
\frac{3}{100} & \frac{3}{100} & \frac{91}{200} & \frac{3}{100} & \frac{1}{5}
\end{bmatrix}$$

PageRank Algorithm

알고리즘 용도ㄴ

과학 연구 및 학계

연구원의 과학적 영향을 정량화

생물학 - 단백질 네트워크 분석

선거

파키스탄 예시 - Contact youth 농업 그룹에서 모든 유권자를 통해서 투표를 실시

스포츠

NFL, 개인 축구선수 등을 평가하는데 사용함

예측 프로그램

거리에 순위를 지정하여 보행자 및 차량 수를 예측 하는데 이용

기타 검색 알고리즘

HITS(Hyper-Text induced Topic Selection) 알고리즘

Authority : 중요한 정보를 제공하고 있는 페이지

Hub: 중요한 정보를 제공하고 있는 페이지에 링크를 보내고 있는 페이지

PageRank와 차이점

1. 쿼리에 따라 달라집니다. 즉, 링크 분석으로 인한 (허브 및 권한) 점수는 검색어의 영향을 받습니다.
2. 인덱싱 시간이 아니라 쿼리 시간에 실행되며 쿼리 시간 처리와 함께 제공되는 성능에 대한 히트가 발생합니다.
3. 검색 엔진에서는 일반적으로 사용되지 않습니다.
4. 단일 점수가 아닌 문서, 허브 및 권한 당 2 개의 점수를 계산합니다.
5. PageRank에서와 같이 모든 문서가 아니라 '관련된'문서의 작은 서브 세트 ('집중 서브 그래프'또는 기본 세트)에서 처리됩니다.

국내 검색 알고리즘

NAVER

LIBRA	C RANK	D.I.A
2012.4 ~ 2016 댓글, 공감, 스크랩 어뷰징 등 상위 노출을 위한 블로거 꼼수에 대응 허위 포스팅과 퀄리티가 떨어지는 콘텐츠들로 신뢰도를 잃던 시기	2016 ~ C = Creater 블로거가 얼마나 오랫동안 하나의 분야에 집중하여 글을 작성했는지 중요하게 본다. 신규 블로거의 진입 장벽이 너무 높았다.	2018 ~ C Rank + 블로거가 얼마나 한 분야의 전문가 인지 컨텐츠가 좋은지 판단 [통계] – 사용자 '체류 시간' 중요 신규 블로거의 진입장벽을 낮춤 검색 사용자들의 체류시간 + 댓글 + 공감 등 여러 지표를 종합하여 점수를 매김

- 감사합니다 -