

PageRank-algoritmen

Gruppe 23 – markusbs & brageei

Hva er PageRank?

- Grunnlaget for Googles søkemotor
- Ordspill på “Web page” og “Larry Page”
- En metode for å rangere nettsider
- Bygger på idéen om at viktige sider lenkes til av mange andre sider

Hvordan fungerer PageRank?

- Hver side starter med en lik verdi (f.eks. PageRank = 1)
- **Distribusjon av PageRank:** PageRank-verdien spres fra en side til alle sidene den lenker til. Dette innebærer at en side fordeler sin PageRank verdi likt mellom alle utgående lenker.
- **Samlet PageRank:** Sider som mottar mange lenker fra andre sider med høy PageRank, får en høyere samlet PageRank, siden hver lenke bidrar til å øke deres rangering.

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda) \cdot \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

- $PR(u)$: PageRank-verdien for siden u , som måler dens viktighet.
- λ/N : En grunnverdi for alle sider, der:
 - λ (lambda): dempningsfaktoren, som justerer hvor mye tilfeldig "hopp" vektet.
 - N : totalt antall sider.
- $(1-\lambda)$: Justerer vekten fra lenker (Dempningsfaktor).
- \sum : Summen av PageRank-verdier fra sider v som lenker til u , delt på L_v , antall utgående lenker L_v fra hver av disse sidene.

PageRank Formel

Viktigheten av Dempningsfaktoren

- 1 - d sannsynlighet for å besøke en tilfeldig nettside

Tenk deg at du surfer på nettet. Dempningsfaktoren sier noe om hvor ofte du klikker på en lenke på siden du er på, sammenlignet med hvor ofte du bare skriver inn en ny side tilfeldig.

- Typisk verdi for d er 0.85 = 15% sannsynlighet

Hvis dempningsfaktoren er 0.85, betyr det at du i 85 % av tilfellene følger en lenke, mens i de resterende 15 % "hopper" du til en tilfeldig side.

- Hindrer at algoritmen "låses" i mindre nettverk

Dempningsfaktoren hindrer at PageRank bare går rundt i små, sammenkoblede nettverk. Ved å legge til sjansen for å "hoppe" til andre sider, får vi en mer balansert og realistisk rangering av nettsidene.

Eksempel på Bruk

1. Scenario:

- Vi har tre nettsider: A, B, og C.
- Side A har høy viktighet (score) og lenker kun til B.
- Side B får dermed mye verdi fra A og lenker videre til flere sider, inkludert C.

2. Hvordan PageRank flyter gjennom nettverket:

- Side A: En viktig side med høy verdi som overfører nesten hele sin PageRank til B.
- Side B: Får mye verdi fra A og fordeler denne verdien videre til sine lenkemottakere, inkludert C.
- Side C: Mottar en liten mengde PageRank fra B, noe som betyr at C er indirekte påvirket av A sin verdi gjennom B.

3. Resultat:

- A gir B en høyere PageRank.
- B sprer verdien sin over flere sider, slik at C får en liten del av verdien.
- Dette illustrerer hvordan PageRank synker gjennom lenkene og hvordan sider med mange utgående lenker sprer sin verdi tynnere.

Fordeler og Begrensninger

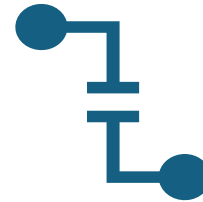


Fordeler:

Relevansbasert rangering: Vurderer antall og kvaliteten på lenker som peker til en nettside.

Skalerbar: Algoritmen kan effektivt håndtere store mengder data, millarder av nettsider

Autoritative kilder: Sider som lenker til andre sider fungerer som stemmer for hvor autoritativ og pålitelig en side er, noe som hjelper med å finne kvalitetskilder.



Begrensninger:

Link spam: Kan manipuleres av kunstige lenkestrukturer (f.eks. Lenkefarmer)

Statisk: Favoriserer eldre nettsider over nyere mer relevante sider

Mangel på innholdsvurdering: PageRank vurderer ikke innholdet på siden direkte, men bare lenkestrukturen.

PageRank's Betydning i dag

- **PageRank i moderne søkemotorer:** PageRank er fortsatt en del av algoritmene som brukes i dagens søkemotorer, men den utgjør bare én av mange faktorer. Søkemotorer som Google har utviklet seg enormt siden PageRank først ble introdusert, og i dag tar de hensyn til hundrevis av signaler (som relevans, brukeratferd, innholdskvalitet og aktualitet) i tillegg til PageRank.
- **Tilpasninger og forbedringer:** Siden lanseringen har PageRank blitt tilpasset for å møte utfordringer med manipulerende lenkestrukturer, som lenkefarmer. Moderne algoritmer har også justert hvordan dempningsfaktoren brukes og legger større vekt på andre kvalitetsfaktorer for å unngå at bare lenker avgjør rangeringen.
- **Grunnlaget for andre rangeringsalgoritmer:** PageRank har inspirert en rekke andre rangeringsalgoritmer, både innen søk og i andre områder. Den grunnleggende ideen om at nettverk kan brukes til å evaluere viktigheten av noder (her nettsider) har blitt brukt i alt fra sosiale nettverk til akademisk siteringsanalyse.
- **PageRanks arv i søketeknologi:** Selv om PageRank nå bare er en del av søkealgoritmer, har den lagt grunnlaget for hvordan vi tenker på rangering på internett. Ideen om at noen sider er viktigere basert på lenker fra andre relevante sider har endret måten søkemotorer fungerer, og den historiske betydningen av PageRank gjør den fortsatt relevant i dagens algoritmer.



Takk for
oppmerksomheten!