Tekst **Arnout Jaspers** Foto **Shutterstock**

Hoe de zoekmachine zijn gedachten ordent

Google danst

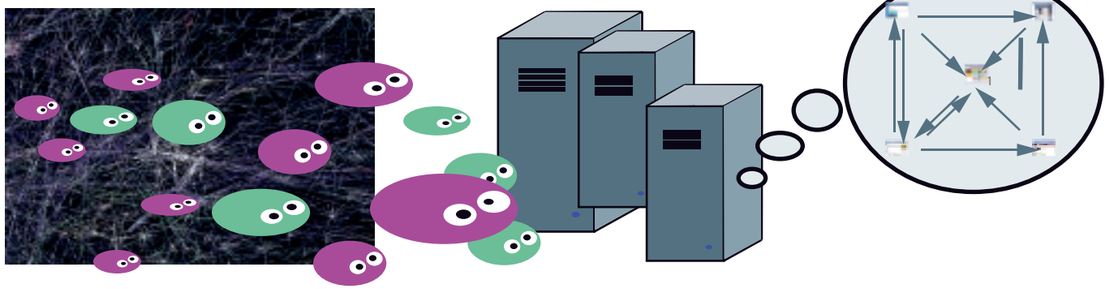
Google viert dit jaar zijn tiende verjaardag. Het begon met twee studenten, Larry Page en Sergey Brin, die lineaire algebra toepasten op grafentheorie. Een blik onder de motorkap bij het Google-algoritme, de gulden standaard voor zoeken op internet.

Jan Brandts, wiskundige aan het Korteweg-De Vries Instituut van de Universiteit van Amsterdam, is expert op het gebied van lineaire algebra en matrixrekening. Dat klinkt specialistisch, en dat is het ook. Toch staat hij regelmatig op een podium of voor een schoolklas met de door hem ontworpen Goozzles, een soort sudoku-variant van het *page rank*-algoritme. Hij vaardigde er zelfs een paar van zijn studenten mee af naar de met bier en wiet doordrenkte ‘universiteit’ op het popfestival Lowlands. Zelden was een onderwerp uit de zuivere wiskunde zo *hot*.
De soms onwerkelijke effectiviteit van Google is grotendeels te danken aan één algoritme, dat de webpagina’s waarin de verlangde zoektermen voorkomen op relevantie sorteert. Maar hoe weet een zoekmachine wat voor de gebruiker het meest relevant is? In feite weet de zoekmachine daar niets van, maar ligt die informatie besloten in de structuur van het internet zelf.
Wiskundig bekeken is het internet een ‘gerichte graaf’, een netwerk van (momenteel) zo’n tien miljard webpagina’s (knooppunten) die al dan niet door links verbonden zijn. ‘Gericht’ slaat op het feit dat een hyperlink tussen twee webpagina’s A en B ofwel van A naar B gaat, ofwel van B naar A, maar nooit allebei, want dat vergt twee aparte links.
Zoekmachines als Google, Yahoo of Baidu (de populairste Chinese zoekmachine) zijn voortdurend bezig met hun webcrawlers het internet af te struinen. Een crawler of spider is een programma dat automatisch via hyperlinks van de ene naar de volgende pagina surft en van elke pagina een korte samenvatting bewaart, bestaand uit het adres, de hyperlinks op die pagina en nog wat trefwoorden. De centrale computers van een zoekmachine slaan al die samenvattingen op en bevatten dus een ‘gestripte’ kopie van het internet, die door hun webcrawlers zoveel mogelijk actueel wordt gehouden. Immers: 40 procent van de webpagina’s verandert wekelijks, 23 procent dagelijks.

Cirkelredenering Waarschijnlijk gebruiken ook andere zoekmachines dan Google tegenwoordig een vorm van het page rank-algoritme om die tien miljard pagina’s in een ranglijst te zetten. De wiskunde achter dit algoritme was onder specialisten allang bekend, maar Page en Brin waren in 1998 de eersten die beseften dat die een geschikt instrument was om informatie op internet te zoeken.
De plaats die een webpagina inneemt in de page rank (weergegeven als een getal tussen 0 en 10, waarbij Google zichzelf altijd bovenaan zet) van een webpagina wordt bepaald door twee principes. Een webpagina is ‘belangrijk’ als andere ‘belangrijke’ webpagina’s ernaar linken, en als een webpagina uitsluitend naar pagina X linkt, is dat voor X meer waard dan als een pagina behalve naar X ook nog naar veel andere pagina’s linkt. Het idee vertoont veel overeenkomsten met menselijk netwerkgedrag: je bent belangrijk als belangrijke mensen jou kennen, en hoe exclusiever de aandacht is die je van een belangrijk persoon krijgt, des te meer baat heb je daar zelf bij. Zelfs het eenrichtingsprincipe geldt: heel veel mensen kennen de uiterst belangrijke persoon Britney Spears terwijl Britney Spears hen niet kent, dus daar worden al die mensen zelf niet belangrijker van.
Dit roept natuurlijk de vraag op: wat is ‘belangrijk’? Op dat punt volgt een cirkelredenering van het soort waar wiskundigen dol op zijn: hoe belangrijk een pagina is, wordt namelijk aangegeven door diens page rank.
Dit systeem lijkt nooit van de grond te komen: hoe kun je ooit de page rank van de eerste pagina bepalen, als die wordt bepaald door de page rank van andere pagina’s, die je evenmin weet?
Toch is het, met wat plausibele aannamen over hoe internetters zich gedragen, mogelijk om je aan je eigen haren uit dit logische moeras omhoog te trekken. In wiskundig opzicht is dit een zogeheten eigenwaardeprobleem van een matrix (zie infographic, pag. 34).
Het berekenen van de page rank van alle webpagina’s vergt het manipuleren van matrices (‘getallenvierkanten’) met 10 miljard maal 10 miljard getallen. Enigszins leesbaar uitgeprint, beslaat zo’n matrix niet een A4’tje, maar een ‘A50’je’: een vel papier waarin je de aardbol kunt verpakken.
Waarschijnlijk – wat geheime diensten uitspoken weten we niet – is dit de grootste matrix-berekening ter wereld, die zelfs op de supercomputers van Google drie dagen duurt. Daarom wordt deze *Google dance* maar eenmaal per maand uitgevoerd.
Als iemand een of meer zoektermen intikt, worden in de gestripte kopie van internet de pagina’s opgezocht die deze termen bevat- ▶

Infographic **Erick Vermeulen**

Hits scoren op Google

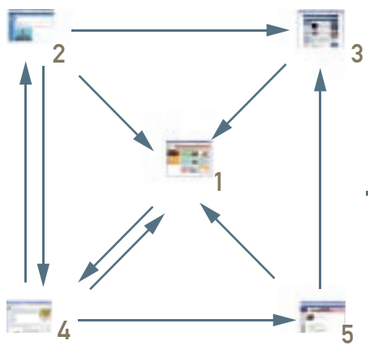


1 Het internet, of beter gezegd het wereldwijde web, bevat naar schatting 10^{10} pagina's.

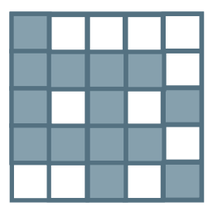
2 Webcrawlers van Google volgen links van pagina naar pagina en brengen zo internet in kaart.

3 Computers slaan een kopie van internet op met samenvattingen van alle gevonden pagina's.

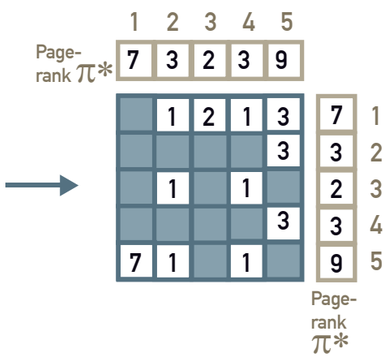
Het page-rank-algoritme



4 Google maakt een ranglijst van alle 10^{10} webpagina's op basis van de structuur van de links. Hierboven staat een internetmodel bestaande uit vijf pagina's. Een pijl geeft een link tussen twee pagina's aan.

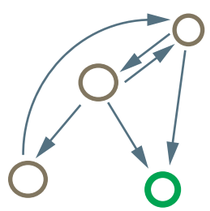


5 De linkstructuur is weer te geven als een zogenoemde S-matrix. Een hokje in bijvoorbeeld kolom 2, rij 3 is open als er een link is van pagina 2 naar pagina 3.

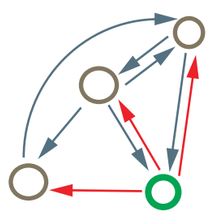


6 De pagina's krijgen hun page-rank (π^*) na oplossen van een puzzel ('Goozzle'). De spelregels: 1) Elke page-rank in de bovenbalk wordt eerlijk verdeeld over de open vakjes in de kolom eronder. 2) Elke page-rank in de zijbalk is de som van de getallen in de rij ernaast.

Teleportatie



7 Een complicatie is, dat de meeste webpagina's geen enkele link bevatten (bijvoorbeeld losse plaatjes en pdf-bestanden). Zo'n 'los eindje' (groen) veroorzaakt in de S-matrix een volledig zwarte kolom, waardoor de puzzel onoplosbaar wordt.



8 De oplossing: doe alsof er vanaf zo'n los eindje links lopen naar alle pagina's (rode pijlen). Die linkstructuur leidt tot de teleportatiematrix E die garandeert dat de puzzel altijd een oplossing heeft.

$$\pi^* = \pi^*(\alpha S + (1-\alpha)E)$$

9 De parameter α bepaalt het relatieve belang van teleportatie versus echte links. Google geeft deze waarde niet vrij, maar Brandts concludeert uit eigen onderzoek dat $\alpha = 17/20$.

Dit is de kern van het algoritme dat Larry Page en Sergey Brin publiceerden. Welke verfijningen er later nog in zijn aangebracht, houdt Google angstvallig geheim.

ten, waarna de gebruiker de links naar deze pagina's in volgorde van hun page rank op z'n scherm krijgt.
Anders dan veel mensen denken, hebben bezoekcijfers dus geen invloed op de Google-ranglijst. Gegeven de dominante positie van deze zoekmachine, werkt het juist omgekeerd: een hoge positie op de ranglijst garandeert hoge bezoekcijfers, omdat zoveel mensen zich bij hun zoektocht naar informatie door die ranglijst laten leiden.
Om met je eigen website hoger op de ranglijst te komen, zul je andere websites die al hoog op de ranglijst staan, moeten overhalen om een link naar jouw site op te nemen. Dat is de economische basis van bedrijven die beweren dat ze je page rank kunnen opkrikken: ze hebben zelf websites met een hoge page rank (of beweren die te hebben), en bieden in feite links te koop aan vanaf die sites. Dit is dubieuze handel. Vanwege het tweede page rank-principe – dat van exclusiviteit – hebben dergelijke links steeds minder effect naarmate zo'n bedrijf meer klanten krijgt. Wie als enige het e-mailadres heeft van Britney Spears, is belangrijk; als iedereen haar mailadres heeft, zijn die mensen opeens een stuk minder belangrijk.
Daar komt nog eens bij dat Google het betalen voor page ranks principieel afwijst. Het heeft zulke bedrijven dan ook al meermalen gestraft door hun page rank terug te zetten naar nul. Daardoor dalen bij de eerstvolgende Google-dans ook de page ranks van alle betalende klanten. "Bedrijven hebben daar wel rechtszaken over gevoerd", zegt Brandts. "Maar Google zegt dan dat de ranglijst niet meer is dan hun eigen mening. En tegen een mening valt juridisch niets te beginnen."

Omslagpunt Het is min of meer toevallig dat de explosieve groei van het *world wide web* gepaard ging met een toename van de geheugen- en rekencapaciteit van de computers die het mogelijk maakt om al die informatie ook efficiënt te doorzoeken. Brandts: "De eerste zoekmachines gaven alleen maar een lijst van webpagina's waarin de ingetikte zoektermen voorkwamen, zonder enige logische rangschikking. In de beginjaren, toen het internet nog klein was, werkte dat nog wel, omdat je niet zo veel hits kreeg. In 1998, het jaar waarin Page en Brin met hun algoritme kwamen, kwam ook wat dat betreft het omslagpunt: het web werd echt te groot voor die methode."
Hoe ineffectief zoekmachines destijds waren, kun je nu nog zien als je bij Google via 'afbeeldingen' zoekt. Brandts: "Web-pagina's vormen een netwerk van links die naar elkaar verwijzen. Voor plaatjes geldt dat niet, dus dan werkt het page rank-principe niet. Je kunt in html (*de programmeertaal waarin webpagina's zijn gecodeerd, red.)* wel trefwoorden aan een plaatje toevoegen die vertellen wat er staat, maar vrijwel niemand doet dat."
Het is voor een computer heel simpel om de tekst van webpagina's te doorzoeken op een of meer zoektermen. Zou het niet mooi zijn als je een zoekmachine aan de hand van een schetsje of voorbeeldfoto naar relevante plaatjes kon laten zoeken? "Momenteel mag je al blij zijn als een iris-scan goed werkt", zegt Brandts. "En daarbij hoeft de computer alleen maar twee heel specifieke plaatjes met elkaar te vergelijken. Zoeken op onderwerp in plaatjes zie ik de eerste tientallen jaren nog niet gebeuren."

Persoonlijk worden Het page-ranken van het hele internet is zo'n gigantische rekenklus dat de vraag opdoemt waarom zoekmachines de procedure niet omkeren: eerst een lijst pagina's op basis van de ingetikte trefwoorden, en die veel kortere lijst daarna pas page-ranken? Dat zou een andere ranking opleveren, omdat de trefwoorden als het ware een hap pagina's uit het internet nemen: links naar andere, niet relevante pagina's worden daarbij doorgesneden. In theorie zou dat een veel persoonlijker resultaat opleveren: iemand die zoekt naar het trefwoord 'zoekmachines' zal niet zo snel uitkomen op de pagina van Britney Spears waar toevallig ook het woord 'zoekmachines' valt, eenvoudigweg omdat er dan wel veel pagina's zijn die naar Britney Spears linken, maar weinig pagina's *over zoekmachines* die naar de site van de zangeres linken.

"Zoeken op onderwerp in plaatjes zie ik de eerste tientallen jaren nog niet gebeuren. Momenteel mag je al blij zijn als een iris-scan goed werkt."

Maar selecties hebben weer hun eigen problemen, vertelt Brandts. "Zo'n op eigen trefwoorden geselecteerde lijst vormt geen samenhangend cluster van webpagina's. De meeste pagina's in zo'n lijst linken helemaal niet naar elkaar, dus heeft het geen zin om daar het page rank-algoritme op los te laten."
Toch is op de individuele gebruiker toegesneden, gepersonaliseerd zoeken wel degelijk de volgende gedroomde stap. Brandts: "In theorie kan het, door de zoekmachine rekening te laten houden met je vorige zoekopdrachten en met de voorkeuren en interesses die je zelf opgeeft. Nu beginnen alle webpagina's nog met hetzelfde 'gewicht' aan de *Google dance*, maar je zou ze allemaal een weegfactor mee kunnen geven. Bijvoorbeeld: als je aangeeft dat sport je interesseert, krijgen alle pagina's waar sporttermen in voorkomen een wat hoger gewicht, wat de uiteindelijke page rank beïnvloedt."
Uiteraard is het in de praktijk niet haalbaar om voor elke gebruiker afzonderlijk een cluster supercomputers drie dagen lang te laten draaien.
Brandts: "Een grove onderverdeling is er nu natuurlijk al, want je kunt kiezen voor webpagina's in een bepaalde taal of uit een bepaald land. Naar verluidt is men al zo ver dat men een stuk of twintig groepen gebruikers onderscheidt. Naarmate de rekensnelheid toeneemt, kun je dat verder verfijnen."
De dominante marktpositie van zoekmachine Google begint inmiddels kritiek op te roepen, maar volgens Brandts is die voor de kwaliteit van de zoekmachines geen probleem: "Wat mij betreft zijn alle andere zoekmachines overbodig. Google wordt gerund door twee mensen die het nog echt leuk vinden om te doen, het zijn eigenlijk academici die dit ook nog wetenschappelijk interessant vinden. Pas als die twee een jaartje of vijftig zijn (*in 2023, red.)* zal dat er wel vanaf gaan." ●