Googlen, googlede, gegoocheld? Vrijwel iedereen heeft wel eens met verbazing naar het scherm gestaard wanneer Google het voor elkaar kreeg om binnen een paar tellen het hele web te doorzoeken naar, bijvoorbeeld, de juiste spelling van het werkwoord in het Nederlands. **Jan Brandts** ontsluiert de geheimen van Google, behalve de α...

De PageRank van Google: de grootste matrixberekening ooit

Inleiding

Google vindt in een oogwenk de meest relevante webbladzijden over een bepaald onderwerp. Omdat het web uit zo'n tien miljard bladzijden bestaat, is dit een enorm indrukwekkende prestatie: het lijkt eenvoudiger om een naald in een hooiberg te vinden. De kracht van Google schuilt in de mathematische omschrijving van het begrip belangrijkheid van een webbladzijde: de zogenaamde PageRank. We zullen laten zien welke ideeën de studenten Page en Brin daar over hadden toen ze Google rond 1997 ontwikkelden. Uiteindelijk bedachten ze een wiskundige vergelijking waarvan de PageRank de oplossing is. We zullen deze PageRank-vergelijking afleiden uit een lijstje van begrijpelijke wensen over PageRank, en ons buigen over de vraag hoe Google deze vervolgens uitrekent. Overigens kunnen we deze vraag niet in zijn geheel beantwoorden: een aantal zaken rondom PageRank wordt door Google strikt geheim gehouden. Het Google PageRank-probleem wordt door velen gezien als de grootste (matrix)berekening die ooit is ondernomen.

Beknopte geschiedenis

In nog geen tien jaar tijd is Google niet alleen één van de bekendste zoekmachines op het internet geworden, maar het is tevens een uiterst succesvol bedrijf, een stuk gereedschap om de effectiviteit van webbladzijden te analyseren, een werkwoord in (niet alleen) de Engelse taal (to google something), een wereldwijde in detail inzoomende verrekijker (Google Earth), en voor sommigen zelfs een levenswijze. Het woord 'googol'[1] is in 1920 door de negenjarige Milton Sirotta bedacht toen zijn vader hem vroeg een passende term te bedenken voor het getal 10^{100}.
Het woord Googolplex, waarvan de naam van het hoofdkantoor Googleplex van Google is afgeleid, bedacht hij als naam voor het gigantische getal 10^{googol}. Google werd in 1998 in het leven geroepen door twee studenten van Stanford University in de VS, Larry Page en Sergey Brin. Zij waren bij de eersten die zich realiseerden dat je slim gebruik kunt maken van de structuur van het internet in het proces van informatie vergaren. Informatie vergaren was natuurlijk sinds de geboorte van het World Wide Web in 1989 al hevig veranderd: in plaats van te zoeken middels traditionele gereedschappen zoals kaartenbakken en micro-fiches in bibliotheken en schijven in computers, moest nu ineens het web worden doorzocht voor informatie. Uiteraard leidde dit tot geheel nieuwe uitdagingen. Immers, met zo'n tien miljard bladzijden is het Web enorm groot. Daarnaast verandert zo'n veertig procent binnen een week van inhoud, en drieëntwintig procent zelfs dagelijks, en dus is het Web dynamisch. Bovendien zijn er geen getrainde specialisten die verantwoordelijk zijn voor de inhoud en organisatie zoals bij een bibliotheek. Sterker nog, er zijn zogeheten spammers die er een sport van maken om het Web opzettelijk te desorganiseren. Het belangrijkste verschil met de traditionele informatiebronnen is echter dat het web gelinkt is middels hyperlinks. Toch duurde het nog zo'n tien jaar voordat Page en Brin, maar ook Jon Kleinberg van IBM (ook in 1998) zich realiseerden dat het World Wide Web een schoolvoorbeeld is van wat wiskundigen een gerichte graaf noemen, en als zodanig behandeld zou moeten worden. Omdat er al vele decennia onderzoek wordt gedaan in de zogeheten grafentheorie, konden de resultaten daarvan meteen worden toegepast op het web. Dit leidde uiteindelijk tot de ontwikkeling van Googles zogenaamde PageRank, en Jon Kleinbergs minder bekende HITS (Hypertext Induced Topic Search), wat gebruikt wordt in de zoekmachine Teoma.

Opmerking
Het getal googol lijkt erg groot. Enerzijds is dit ook zo. Immers, het aantal elementaire deeltjes in het heelal wordt geschat op zo'n 10^{70}. Anderzijds, als er zeventig mensen in de rij staan bij een kassa, is het aantal volgordes waarin ze kunnen staan net even meer dan googol.

Het PageRank-model

Het succes van Google berust op de wonderbaarlijk goede resultaten die het door Google ontwikkelde PageRank-model produceert. Zoals gezegd is de PageRank

van een webbladzijde een getal dat de belangrijkheid van die bladzijde aangeeft. Omdat het web uit zo'n tien miljard bladzijden bestaat, zijn er tien miljard PageRank-getallen. Behalve dat we zullen onderzoeken hoe deze getallen zijn gedefinieerd, is het natuurlijk ook interessant om te kijken naar de problemen die ontstaan doordat al deze getallen met regelmaat uitgerekend moeten worden. Omdat het web zo snel van structuur verandert, is het goed mogelijk dat een onbelangrijke bladzijde van vandaag over een maand een belangrijke bladzijde is geworden, en dus zullen die tien miljard getallen met regelmaat moeten worden ververst. Samengevat zullen we ons de volgende vragen stellen:

– Welke criteria worden gebruikt om het belang van webbladzijden te onderscheiden?
– Hoe zetten we deze criteria om in een wiskundige model?
– Hoe berekenen we op efficiënte wijze de resulterende PageRank-getallen?

In de volgende secties zullen we deze vragen proberen te beantwoorden, meestal aan de hand van eenvoudige voorbeelden.

De intuïtie achter het PageRank-model

Het eerste model voor PageRank van Sergey Brin en Larry Page berust op twee hele eenvoudige en logische principes.

> Principe 1: Een webbladzijde is belangrijk als ernaar wordt verwezen door andere belangrijke webbladzijden.
>
> Principe 2: Als een webbladzijde enkel en alleen naar jouw bladzijde verwijst, is dit meer waard dan wanneer deze bladzijde ook naar heel veel ander bladzijden verwijst.

Beide principes klinken heel natuurlijk, maar lijken niet echt tot concrete getallen voor het belang van een bladzijde te kunnen leiden. Immers, om het belang van een gegeven bladzijde te bepalen aan de hand van rekenregels die op deze principes zijn gebaseerd, moet je weten hoe belangrijk de bladzijden zijn die naar jouw bladzijde verwijzen. Dit lijkt in een cirkelredenering te verzanden. Verderop zal echter blijken dat dit geen probleem is.

Een eerste wiskundig model

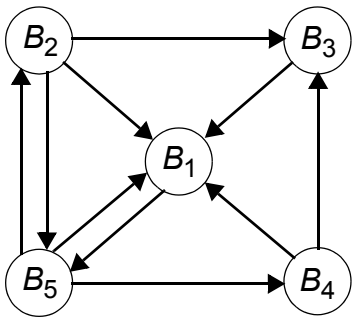

fig. 1 Een World Wide Web van vijf bladzijden, met pijlen als hyperlinks

Laten we een poging doen om bovenstaande twee principes om te zetten in formules die concrete getallen kunnen gaan opleveren. Om de zaken overzichtelijk te houden, nemen we aan dat het World Wide Web uit vijf bladzijden bestaat, die naar elkaar verwijzen zoals weergegeven in figuur 1. Met het oog op de beide principes 1 en 2 lijkt bladzijde B_1 in dit web een belangrijke bladzijde: alle andere bladzijden verwijzen ernaar. Ook B_5 heeft goede papieren: twee bladzijden verwijzen ernaar, waaronder de belangrijke B_1. Maar door principe 2 is B_5 misschien wel belangrijker dan B_1, omdat B_1 alleen naar B_5 verwijst, terwijl B_5 behalve naar B_1 ook nog naar twee andere bladzijden verwijst.

Vraag is natuurlijk hoe de heuristieken verwoord in principes 1 en 2 omgezet kunnen worden in harde wiskunde. Dat deden Page en Brin als volgt. Laat voor iedere gehele j met $1 \le j \le 5$ het symbool P_j staan voor de belangrijkheid van bladzijde B_j. We noemen P_j de PageRank van B_j.

Definitie 3 (PageRank): Veronderstel dat bladzijde B_i naar L_i verschillende bladzijden verwijst, waaronder B_j. Dan draagt B_i een hoeveelheid $\frac{P_i}{L_i}$ bij aan de PageRank van B_j.

In figuur 1 betekent dit dat de PageRank P_1 van bladzijde B_1 wordt berekend middels:

$$P = \frac{P_2}{3} + \frac{P_3}{1} + \frac{P_4}{2} + \frac{P_5}{3} \qquad (1)$$

omdat B_2 naar B_1 verwijst en naar drie bladzijden in totaal, omdat B_3 naar B_1 verwijst en naar één bladzijde in totaal, omdat B_4 naar B_1 verwijst en naar twee bladzijden in totaal, en omdat B_5 naar B_1 verwijst en naar drie bladzijden in totaal. Voor de overige vier bladzijden vinden we op gelijke wijze dat

$$P_2 = \frac{P_5}{3},\ P_3 = \frac{P_2}{3} + \frac{P_4}{2},\ P_4 = \frac{P_5}{3} \text{ en } P_5 = \frac{P_1}{1} + \frac{P_2}{3}. \qquad (2)$$

Het model resulteert dus in vijf lineaire vergelijkingen voor de onbekende getallen $P_1, \ldots, P_5$, die we als volgt kunnen schrijven in matrix-vectorvorm:

$$HP = P, \text{ waarbij } \quad H = \begin{bmatrix} 0 & \frac{1}{3} & 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 1 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix} \text{ en } P = \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{bmatrix}. \qquad (3)$$

Deze vergelijkingen vormen derhalve een speciaal geval van een eigenwaardeprobleem. Er bestaat een oplossing $P \neq 0$ als de matrix H een eigenwaarde gelijk aan één heeft. Middels enig rekenwerk is na te gaan dat dit inderdaad het geval is, en een oplossing P van (3) in gehele getallen is:

$$P_1 = 16, P_2 = 6, P_3 = 5, P_4 = 6, P_5 = 18, \quad (4)$$

ons vermoeden over het belang van B_5 en B_1 bevestigend. Natuurlijk is ieder veelvoud van de gegeven oplossing ook een oplossing, maar voor een rangschikking maakt dit niet uit.

Opmerking
De rijen en de kolommen van H geven duidelijk weer hoe het web in elkaar zit. De tweede rij van H vertelt je bijvoorbeeld dat B_2 niets ontvangt van bladzijden B_1, B_3 en B_4 en de helft van de PageRank van B_5. De vijfde kolom laat zien dat bladzijde B_5 zijn PageRank gelijkelijk verdeelt over bladzijden B_1, B_2 en B_4, enzovoort.
Een belangrijke vraag is of bovenstaand voorbeeld typerend is voor de algemene situatie. Heeft iedere op deze manier verkregen matrix H wel een eigenwaarde één? Antwoord hierop, en aanpassingen van het model die problemen oplossen, geven we hieronder.

Tekortkomingen

Aan de hand van eenvoudige voorbeelden illustreren we nu wat er allemaal fout kan gaan in het zojuist geïntroduceerde PageRank-model. We geven voorbeelden van de volgende problemen:

- Het web is onsamenhangend,
- Een simpele dekpuntiteratie convergeert niet naar de oplossing van de vergelijkingen,
- De PageRank-vergelijkingen hebben alleen de oplossing $P_1 = P_2 = ... = P_n = 0$.

Dit laatste probleem is het meest serieuze en correspondeert met de opmerking dat H geen eigenwaarde gelijk aan één heeft. Gelukkig kan het model worden aangepast zodat al deze problemen worden opgelost.

Het web is onsamenhangend
Een eerste tekortkoming van het model hierboven is dat het geen uitsluitsel geeft over de situatie waarin het web niet samenhangend is. Hiermee bedoelen we dat er twee of meer verschillende groepen van bladzijden zijn die totaal niet naar elkaar verwijzen, zoals in figuur 2. Voor ieder van beide webdelen kunnen onderling de PageRanks worden uitgerekend, maar omdat veelvouden van PageRanks ook weer PageRanks zijn, is het niet duidelijk hoe de PageRanks van de twee groepen met elkaar moeten worden vergeleken.

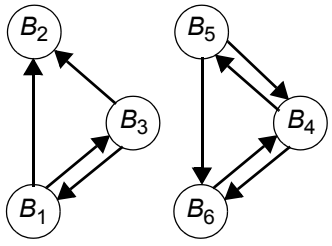

fig. 2 Hoe vergelijk je twee volledig afzonderlijke groepen bladzijden?

Immers, de PageRanks van het linkerdeel kunnen straffeloos met twee worden vermenigvuldigd terwijl die in het rechterdeel gelijk worden gehouden. Of, in meer wiskundige termen, de eigenruimte van de 6×6 matrix H behorende bij de eigenwaarde één is tweedimensionaal.

Opmerking
Een kleine losse cluster van webbladzijden kan onbelangrijk lijken, omdat niemand ernaar verwijst. Desondanks kan een websurfer op één van deze bladzijden aankomen door simpelweg het adres in de adresbalk in te tikken. Behalve het volgen van weblinks is ook het intikken van een adres dus een reële mogelijkheid om ergens aan te komen. We zullen dit later in het model verwerken.

Een eenvoudige dekpuntiteratie convergeert niet
Een tweede tekortkoming is dat de voor de hand liggende dekpuntiteratie

$$P_{k+1} = HP_k, \text{ met gegeven start-vector } P_0, \quad (5)$$

voor de PageRank-vergelijking $HP = P$ niet altijd convergeert naar een oplossing. Een eenvoudig voorbeeld waaruit dit blijkt, is het web met twee bladzijden die naar elkaar verwijzen:

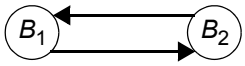

fig. 3 Een web waarvoor de dekpuntiteratie (5) niet convergeert

De PageRank-vergelijkingen voor P_1 en P_2 zijn simpelweg

$$P_1 = P_2 \text{ en } P_2 = P_1. \quad (6)$$

Toch convergeert dekpuntiteratie (5)

$$P_{k+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} P_k, \text{ waarbij } P_0 = \begin{bmatrix} a \\ b \end{bmatrix}, \quad (7)$$

niet als $a \neq b$, omdat in iedere iteratiestap de vectorentries alleen maar worden verwisseld.
Omdat de matrix H in de praktijk afmetingen $n \times n$ heeft met n tegen de tien miljard, is (5) één van de weinige praktische mogelijkheden om de oplossing van $HP = P$ in drie à vier decimalen nauwkeurig te kunnen berekenen. En dan nog duurt het met de huidige supercomputers een dag of drie! Dus, ondanks dat het wiskundig gezien geen vereiste is voor een goed gedefinieerd PageRank-model, is het praktisch erg wenselijk om op een of andere manier te kunnen garanderen dat (5) met een bepaalde snelheid convergeert.

De PageRank-vergelijkingen hebben geen zinvolle oplossing
Een derde en veel serieuzer probleem is, dat het niet is gegarandeerd dat de PageRank-vergelijkingen die worden

opgesteld, ook inderdaad een zinvolle oplossing hebben. Hier is een voorbeeld:

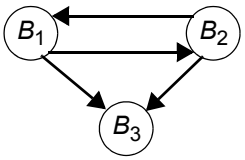

fig. 4 Een web zonder zinvolle PageRank-oplossing

De bijbehorende PageRank-vergelijkingen zijn de volgende

$$P_1 = \frac{1}{2}P_2,\ P_2 = \frac{1}{2}P_1, \text{ en } P_3 = \frac{1}{2}P_1 + \frac{1}{2}P_2 \qquad (8)$$

De eerste twee vergelijkingen laten zien dat

$$P_1 = \frac{1}{2}P_2 = \frac{1}{2}\left(\frac{1}{2}P_1\right) = \frac{1}{4}P_2, \qquad (9)$$

en dus is $P_1 = 0$. Maar dan is ook $P_2 = 0$, en uit de derde vergelijking volgt dat ook $P_3 = 0$.
Kortom, de oplossing $P_1 = P_2 = P_3 = 0$ is de enige oplossing van dit stelsel. Of, met andere woorden, de matrix H behorende bij dit web,

$$H = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \qquad (10)$$

heeft geen eigenwaarde gelijk aan één. Dit is natuurlijk onwenselijk, vooral als het op grotere schaal zou gebeuren.

Opmerking
De oorzaak voor het ontbreken van een zinvolle PageRank in figuur 4 is dat er een bladzijde is waar weliswaar naar wordt verwezen, maar die zelf nergens naar verwijst. Ongeveer tachtig procent van het web bestaat uit dergelijke bladzijden (documenten zoals jpg-, en pdf-files). Dergelijke bladzijden, die in de Engelstalige literatuur dangling nodes heten, kunnen zeer relevante informatie bevatten, dus het zou geen goede oplossing zijn ze gemakshalve maar te negeren.

Reparatie middels teleportatie

De zojuist geformuleerde drie tekortkomingen van het originele PageRank-model kunnen alledrie worden verholpen. We zullen aangeven hoe dit kan worden gedaan. Het gaat echter iets buiten het bestek van deze tekst om ook volledig te bewijzen dat de problemen daadwerkelijk verholpen zijn. We beginnen met het laatst gevonden probleem van de dangling nodes.

Teleportatie vanuit dangling nodes
Een dangling node is een webbladzijde die zelf nergens naar verwijst, zoals een jpg- of pdf-document. Dergelijke bladzijden kunnen er de oorzaak van zijn dat de PageRank-vergelijkingen $HP = P$ alleen de oninteressante oplossing $P = 0$ hebben. Dit is een onvolkomenheid in het oorspronkelijke model, die we als volgt herstellen:

> Principe 4: (dangling nodes) Een webbladzijde die nergens naar verwijst zullen we in het model representeren als een bladzijde die naar alle bladzijden binnen het web, inclusief zichzelf, verwijst.

Dit lijkt de omgekeerde werkelijkheid, maar zo gek is het niet. Immers, je kunt in een dangling node geen link aanklikken. Wat je wel kunt doen om weg te komen, is een nieuw webadres intikken in de navigatiebalk van je webbrowser. Omdat dit ieder adres op het web kan zijn, is het niet onnatuurlijk om vanuit deze bladzijde pijlen te tekenen naar alle andere bladzijden.

> Definitie 5: (teleportatie) De verplaatsing van een dangling node naar een willekeurig ingetypt nieuw webadres wordt in de literatuur aangeduidt met teleportatie.

Opmerkingen
De reden om een dangling node ook naar zichzelf te laten verwijzen, is dat hij er anders nadeel van zou ondervinden dat hij een dangling node is. De voorgestelde aanpassing lijkt in zekere zin eerlijk.

De matrix S met eigenwaarde één
Wiskundig gezien komt een dangling node overeen met een kolom van H waarin alleen maar nullen staan. Principe 4 stelt dat iedere nul in een dergelijke kolom moet worden vervangen door $1/n$, waarbij n het totaal aantal webbladzijden is. De matrix H uit (10) wordt op deze manier een matrix S:

$$S = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \end{bmatrix}, \qquad (11)$$

wat in zijn algemeenheid kan worden opgeschreven als

$$S = H + \frac{1}{n}ae^T, \qquad (12)$$

waarbij $e^T = (1, \ldots, 1)$ en $a_j = 1$ als B_j een dangling node is, en nul als dat niet zo is. Omdat iedere kolom van S, en dus ook iedere rij van de gespiegelde S^T van S, nu optelt tot één, vinden we dat $S^T e = e$, dus S^T heeft een eigenwaarde één. Omdat

$$0 = \det(S^T - \mathrm{I}) = \det(S^T - I^T) = \det(S - I), \qquad (13)$$

gebruikmakend van de regel dat $\det(A) = \det(A^T)$, vinden we dat ook S zelf een eigenwaarde één heeft. Schrijven we nu $\|\cdot\|_\infty$ voor de maximum norm op $\mathbf{R}^n$, dan zien we middels de afschatting

$$\forall w \in \mathbf{R}^n : \left\| S^T w \right\|_\infty \leq \|w\|_\infty \quad (14)$$

die in het bijzonder geldt voor alle eigenvectoren van S^T, dat als $S^T v = \lambda v$, er noodzakelijkerwijs moet gelden dat $|\lambda| \leq 1$. Dus is de eigenwaarde één een zogeheten dominante eigenwaarde van S en S^T .

Onsamenhangendheid en periodiciteit: globale teleportatie

Natuurlijk zal een surfer niet alleen als hij in een dangling node aankomt een nieuw webadres in de navigatiebalk van de webbrowser intikken. Ook op andere bladzijden zal dit soms worden gedaan, simpelweg omdat de interesse in de huidige bladzijde en de links daarvandaan is verdwenen, of omdat het niet snel genoeg leidt tot de gewenste bestemming.

> Principe 6 (α-teleportatie) Een surfer zal een bepaald deel α met $0 < \alpha < 1$ van de tijd niet de links in het web volgen, maar een nieuw adres intikken in de navigatiebalk.

Om dit principe in het model op te nemen, doen we het volgende. Het pure teleporteren kunnen we symboliseren middels de zogenaamde teleportatie-matrix T die gedefinieerd is door

$$T = \frac{1}{n} e e^T = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} . \quad (15)$$

Deze matrix staat model voor het (denkbeeldige) web waarin iedere bladzijde naar iedere bladzijde linkt. Je kan met gelijke kans overal naartoe. Kortom, de ultieme teleportatie.

Echter, het echte web zit zo niet in elkaar. Om toch een deel ter grootte α uit teleportatie te laten bestaan, en de rest van de tijd ter grootte $1 - \alpha$ het model tot nu toe te volgen, is het uiteindelijke model van Google gebaseerd op de matrix G, genaamd de Google-matrix,

$$G = (1 - \alpha)H + \alpha T. \quad (16)$$

Omdat de kolommen nog steeds optellen tot één, geldt het argument uit het vorige hoofdstuk dat G een dominante eigenwaarde één heeft. Bovendien geldt, als $0 < \alpha < 1$ dat het web op kunstmatige wijze samenhangend is geworden.

> Stelling 7. De Google matrix G heeft een ééndimensionale eigenruimte V behorende bij de dominante eigenwaarde één. Iedere $v \in V$ heeft entries met gelijk teken (plus, min, nul), en dus kan de PageRank strikt positief worden gekozen. De overige eigenwaarden van G zijn in absolute waarde kleiner dan α, waardoor de dekpuntiteratie (5) convergeert met asymptotische convergentie $\mathcal{O}(\alpha^k)$.

In de praktijk moet natuurlijk wel een keuze voor α worden gemaakt. Vermoed wordt dat Google $\alpha = 0{,}15$ gebruikt, maar dat is, net als veel andere aspecten van het echte model dat tegenwoordig door Google wordt gebruikt, niet openbaar. Voor bewijzen van bovenstaande uitspraak, veel referenties, en amusante anekdotes over Google, verwijs ik naar onderstaand boek over Google, dat ik van harte aanbeveel.

Jan Brandts,
Universiteit van Amsterdam

Literatuur

Langville, A.N. & C.D. Meyer (2006). *Google's PageRank and Beyond: the science of search engine rankings*. Princeton and Oxford: Princeton University Press.

Noot

[1] Het woord Google ontstond door een spelfout van de investeerders op een cheque aan de oprichters.