Achter de schermen van Google

In deze opdracht geeft Jan Brandts, medewerker aan de Universiteit van Amsterdam, een wiskundig getinte uitleg over de werking van de zoekmachine Google. Google rangschikt zijn zoekresultaten aan de hand van het belang van de gevonden bladzijdes, genaamd de PageRank. Met behulp van Goozzles (Google puzzles) wordt uitgelegd hoe deze PageRank in 1998 door Larry Page en Sergey Brin werd bepaald. Veel plezier!

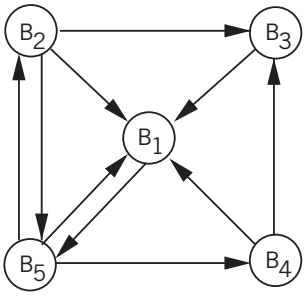
Het woord Googol werd in 1920 bedacht door de 9-jarige Milton Sirotta voor het getal 10100, een één met honderd nullen.

Als je Google gebruikt om te zoeken naar een bepaalde term, krijgt elke gevonden bladzijde een rapportcijfer, de PageRank. Deze rapportcijfers worden gebruikt om de bladzijdes in volgorde van belangrijkheid weer te geven. Geen kleine klus, want internet heeft op dit moment meer dan tien miljard pagina's. Het kost maar liefst drie dagen om alle tien miljard rapportcijfers uit te rekenen met behulp van grote hoeveelheden aan elkaar geschakelde supercomputers. Dat is erg duur, en daarom doet Google dit maar één keer per maand.
De rangschikking van Google stemt vaak verrassend goed overeen met wat mensen echt belangrijk vinden. In deze opdracht proberen we uit te leggen hoe Google dat doet.

Google wordt soms aangeklaagd door bedrijven die vinden dat hun PageRank te laag is. Ze verdedigen zich dan door te stellen dat de PageRank slechts hun mening voorstelt.

1. Een wiskundig kabouterraadsel

Stel, er zijn vijf kabouters die we even B_1, B_2, B_3, B_4 en B_5 noemen. Ieder van deze kabouters heeft vrienden voor wie hij alles over heeft. In onderstaand plaatje kan je zien wie een vriend van wie is: de pijl van B_2 naar B_1 betekent dat B_1 door B_2 als vriend wordt beschouwd. De kabouters wijzen dus in figuur 1 met pijlen hun vrienden aan.



Figuur 1. Niet-wederzijdse vriendschapsrelaties.

Behalve kabouters is er ook een goede fee. Zij heeft enorme hoeveelheden knikkers, die ze graag aan de kabouters wil uitdelen. Er is echter een probleem: zodra een kabouter knikkers krijgt, zal hij ze eerlijk verdelen over zijn vrienden. Zo zitten kabouters nu eenmaal in elkaar.

De fee besluit daarom elke kabouter een zodanig aantal knikkers te geven (minstens één), dat nadat iedere kabouter zijn knikkers heeft verdeeld over zijn vrienden, ze allemaal weer net zoveel knikkers hadden als ervóór. Maar hoe kan ze dit doen? Omdat dit niet even snel uit te leggen is, komen we er later op terug.

PageRank wordt ook te koop aangeboden. Bedrijven zorgen dan tegen betaling voor links van hoog genoteerde pagina's naar de jouwe.

2. Het verband tussen raadsels en PageRank

Het world wide web kun je opvatten als een vriendennetwerk van kabouters. Alleen zijn de kabouters vervangen door webpagina's met pijlen ertussen. Een pijl van B_1 naar B_2 geeft dan aan dat bladzijde B_1 een hyperlink heeft waarmee je naar B_2 kunt surfen.

Bij de kabouters komt de hoeveelheid knikkers waarmee het raadsel is opgelost, overeen met de populariteit van de kabouters. Een kabouter heeft veel knikkers als:

- hij vrienden heeft met veel knikkers,
- die vrienden niet veel andere vrienden hebben.

Dit komt overeen met wat Page en Brin in hun PageRank-model tot uitdrukking willen laten komen, namelijk, een webbladzijde is belangrijk als:

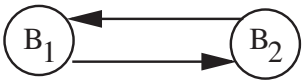
- er naar verwezen wordt door belangrijke bladzijdes,
- die bladzijdes niet naar veel andere bladzijdes verwijzen.

Je bent dus goed af als Koningin Beatrix op haar webbladzijde naar de jouwe verwijst, maar al een stuk minder goed als ze ook naar al haar andere onderdanen blijkt te verwijzen!

Hierna volgen enkele eenvoudige opgaven om mee te oefenen.

Opgave 1

Bepaal de zgn. evenwichtsverdeling die, na het verdelen van de knikkers volgens de pijlen, dezelfde knikkerverdeling oplevert.

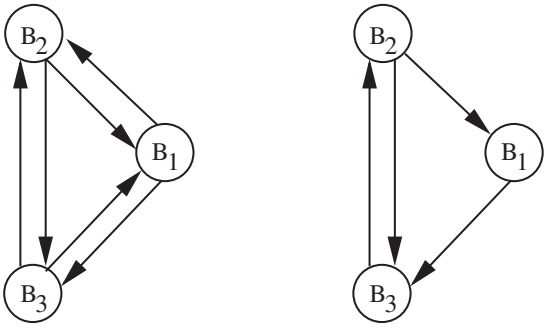


Figuur 2. Eenvoudigste versie van het raadsel.

Controleer dat als je een oplossing hebt gevonden, dit niet de enige oplossing is: als je alle kabouters tweemaal zoveel knikkers had gegeven, had dit ook gewerkt. Sterker nog, ieder veelvoud van een oplossing is weer een oplossing!

Opgave 2

Bepaal de evenwichtsverdeling van elk van de volgende twee netwerken:



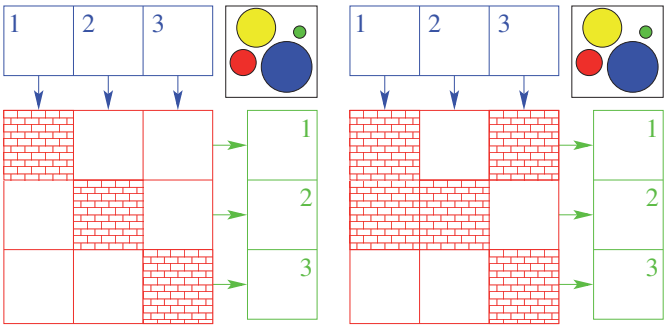
Figuur 3. Een eenvoudig en een wat moeilijker raadsel.

De raadsels uit figuur 3 kun je ook weergeven zoals in figuur 4. Hierbij is het idee als volgt.

- Plaats knikkers in de bovenste drie vakjes,
- Verdeel ze eerlijk over de lege vakjes er verticaal onder,
- Verplaats ze horizontaal naar de rechter drie vakjes.

Als de hoeveelheden knikkers in de rechtervakjes nu hetzelfde zijn als waarmee je bovenin begon, heb je de gevraagde evenwichtsverdeling gevonden.

Raadsels zoals in figuur 4 noemen we Goozzles (Google puzzles).



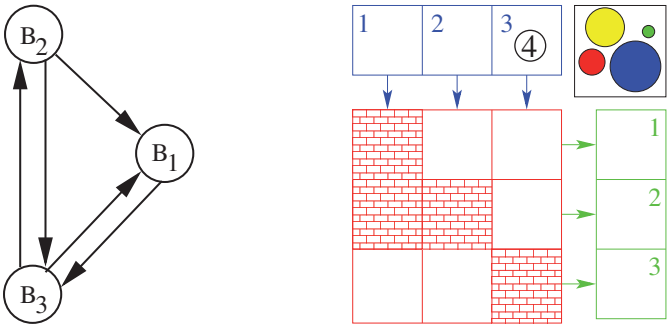
Figuur 4. Figuur 3 in een ander jasje.

Opgave 3

Los opgave 2 nog een keer op, maar nu als Goozzle (figuur 4).

Opgave 4

Los de volgende Goozzle op. Om je op weg te helpen hebben we één van de drie gezochte getallen al ingevuld!



Figuur 5. Deels ingevulde Goozzle.

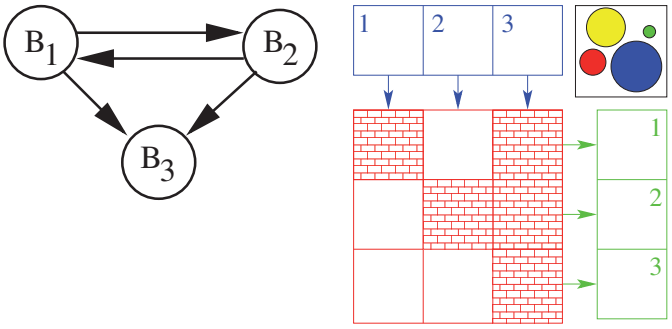
Sommige mensen willen liever niet geholpen worden bij het puzzelen, en hadden de 4 in bovenstaande opgave liever niet kado gekregen. Toch is het niet echt een kado: we hadden immers al gezien dat veelvouden van een oplossing ook weer oplossingen zijn. Er is dan ook een veelvoud dat inderdaad een 4 op die positie heeft. Als een oplossing tot gebroken knikkers - breuken dus - leidt, dan is dat niet erg: na afloop vermenigvuldigen we alles weer met een getal zodanig dat alle breuken verdwijnen.

Opgave 5

a. Ga nu terug naar figuur 1. Teken de Goozzle voor dit vriendennetwerk en los hem op.
b. B_1 is de kabouter waar de meeste pijlen naartoe wijzen. Beredeneer de resultaten voor B_1 en B_5.

3. Onvolkomenheden in het model

De voorgaande Goozzles hebben allemaal een oplossing. Dit is echter niet altijd het geval, zoals uit het volgende voorbeeld blijkt.



Figuur 6. Onoplosbare Goozzle.

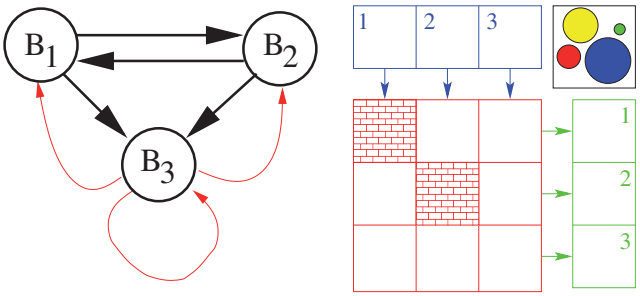
In dit voorbeeld krijgt kabouter B_3 knikkers van B_1 en B_2, maar zelf geeft hij niets weg. Hij zal dus altijd meer bezitten dan ervoor, tenzij B_1 en B_2 geen knikkers hadden. Maar, de fee gaf iedere kabouter minstens één knikker! Er is dus geen oplossing van deze Goozzle.

Ongeveer viervijfde van de documenten in het World Wide Web bevat geen hyperlinks. Denk hierbij aan jpg-, gif-, en pdf-bestanden. Dergelijke documenten heten dangling nodes.

Page en Brin argumenteren dat als een surfer in een dangling node aankomt, hij bij gebrek aan hyperlinks een willekeurig nieuw webadres in de browserbalk zal intikken. In ons plaatje komt dit erop neer dat je vanuit een dangling node pijlen trekt naar ieder van de ander bladzijden, inclusief de dangling node zelf, ondanks dat deze links dus eigenlijk geen van alle echt bestaan. Het reizen naar een andere webpagina zonder daarbij een hyperlink te volgen wordt *teleportatie* genoemd.

Opgave 6

Los de volgende Goozzle op:



Figuur 7. Goozzle uit figuur 6 inclusief teleportatie.

Reden om ook een pijl te trekken naar de dangling node zelf, is dat ieder van de drie bladzijden B_1, B_2, B_3 evenveel profiteert van de PageRank die B_3 uiteindelijk krijgt. Geen van de bladzijden wordt dus bevoordeeld in deze behandeling van dangling nodes.

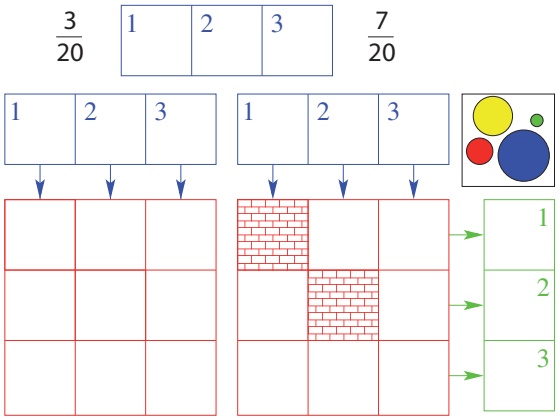
Met deze aanpak van dangling nodes zijn nog niet alle problemen de wereld uit. Bijvoorbeeld, het world wide web zou kunnen bestaan uit meerdere groepen van bladzijdes die onderling geen links hebben. Bladzijdes uit verschillende groepen kunnen dan niet eerlijk met elkaar worden vergeleken, omdat de hoeveelheid knikkers binnen iedere groep met een willekeurig getal vermenigvuldigd kan worden.

4. Het uiteindelijke basismodel

Brin en Page merkten terecht op, dat niet alleen als een surfer in een dangling node aankomt, hij een nieuw webadres in de browserbalk kan intikken. Hij doet dit ook op andere momenten. Vraag is wanneer, en hoe vaak.

De α-factor.
Een surfer zal een deel α (met $0 \le \alpha \le 1$) van de tijd hyperlinks volgen, en een deel $(1-\alpha)$ een nieuw adres in de browserbalk intikken.

Het uiteindelijke model bestaat uit een combinatie van het model met de oplossing voor dangling nodes, en het volledige teleportatie-model. We hebben al gezien dat in dit laatste model iedere bladzijde als het ware een link heeft naar iedere andere bladzijde, inclusief zichzelf. Hoe wordt nu deze combinatie in de praktijk gemaakt? Door te stellen dat een deel α van de knikkers via de pijlen moet lopen van het oorspronkelijke model, en het resterende deel volgens de pijlen van het volledige teleportatie-model.



Figuur 8. Goozzle met teleportatie en α-factor.

Naar aanleiding van wat wiskundig speurwerk valt te beredeneren dat Google de volgende waarde hanteert:

$$\alpha = 0.85 = \frac{17}{20}$$

In de teleportatiematrix - links in figuur 8 - wordt dus 3/20-e deel van de knikkers gelijkelijk over alle bladzijden verdeeld. De resterende knikkers (17/20-e deel) worden – via de rechtermatrix in figuur 8 - verdeeld op basis van de aanwezige hyperlinks (inclusief dangling nodes).

Natuurlijk is het tot zover beschreven model nog lang niet de hele waarheid. Google heeft ongetwijfeld nog heel veel kleine en grote slimmigheden om efficiënt om te gaan met zoekopdrachten en pageranking. Deze zijn, begrijpelijkerwijs, voor het grootste deel geheim.

Als je meer wilt weten over dit onderwerp, dan kun je meedoen aan de 'Google PageRank'-webklas. Deze klas is geschikt voor leerlingen uit 5/6 vwo. De lessen - via internet - worden begeleid door Jan Brandts. Via kleine stapjes kom je uiteindelijk heel veel te weten over de wiskunde achter Google. Speciale voorkennis is niet vereist. Wel moet je bereid zijn om er gedurende vier weken (in het voorjaar) wat extra tijd en energie in te steken.

De eerstvolgende 'Google PageRank'-webklas start in het voorjaar 2009. Aanmelden kan via http://www.studeren.uva.nl/webklassen. Hier vindt je ook informatie over allerlei andere (wiskunde)webklassen die de Universiteit van Amsterdam verzorgt.