4. Informatie zoeken

Auteur	Bètapartners
Team	Wikiwijs Maken Auteurs
Laatst gewijzigd	25 november 2014
Licentie	CC Naamsvermelding-GelijkDelen 3.0 Nederland licentie
Webadres	https://maken.wikiwijs.nl/45922/

Inhoudsopgave

4 Informatie zoeken

De hoeveelheid informatie op internet is indrukwekkend en neemt nog steeds exponentieel toe. Google heeft berekend dat er rond juli 2008 1 triljoen unieke URL's http://nl.wikipedia.org/wiki/URL waren. Cijfers uit midden 2009 ramen de hoeveelheid data op internet op 487 biljoen gigabyte. Als al die informatie zou worden uitgeprint en ingebonden, zou dat resulteren in tien stapels boeken die van de aarde tot Pluto reiken. Doordat steeds meer mensen internet gebruiken en data genereren, neemt deze stapel boeken sneller toe dan een space shuttle kan bijhouden.
De centrale vraag in dit hoofdstuk is hoe je in deze stapel informatie toch nog iets kunt vinden en welke technieken er zijn ontwikkeld om je daarbij te helpen. De verschillende aspecten worden behandeld in de volgende paragrafen:

4a. Zoeksystemen
4b. Uitgelicht: Google (hoe werkt een zoekmachine?)
4c. Zoeken op internet: tips en tricks
4d. Het verborgen internet

Download voor je verder gaat met het hoofdstuk nu eerst de opdrachten:

https://maken.wikiwijs.nl/userfiles/5/5deb7d6448543310dc8f76cdfa6d19fd.doc

Het icoontje  geeft aan wanneer je een opdracht moet maken.
Vul de antwoorden en je naam + klas in in het Word document en upload aan het einde van het hoofdstuk de antwoorden in de Postbus.

4a Zoeksystemen

Om op internet informatie te kunnen vinden is een zoekinstrumentarium ontwikkeld dat beter bekend is onder de naam zoekmachine. Kort gezegd struinen zoekrobots het hele internet af en vullen een enorme database met data. Om te kunnen reageren op een zoekopdracht beschikt de zoekmachine over de letterlijke weergave van miljarden webpagina's. Afhankelijk van hun functie zijn zoekmachines zijn te verdelen in *algemene* en *specifieke* zoekmachines.

Algemene zoekmachines zoeken het hele internet af. Bekende algemene en internationale zoekmachines zijn Google, Yahoo, search.msn.com, search.aol.com, askjeeves en Lycos. Voorbeelden van algemene zoekmachines die alleen binnen het Nederlandse domein zoeken zijn altavista.nl en ilse.nl.

Een bijzondere vorm van een algemene zoekmachine is de metazoekmachine. Hiermee kun je in verschillende zoekmachines tegelijkertijd zoeken. Het resultaat wordt als een lijst gepresenteerd aan de gebruiker. Voorbeelden van metazoekmachines zijn http://ixquick.com/ned/ en http://www.webcrawler.com.

Daarnaast zijn er specifieke zoekmachines die ook wel verticale zoekmachines worden genoemd. Deze zoekmachines richten zich op een bepaald specialisatiegebied en zijn ontworpen om op die gebieden betere prestaties te leveren dan de algemene zoekmachines. Een voorbeeld is Google Scholar dat wetenschappelijke artikelen doorzoekt of jaap.nl waarmee je naar koop- en huurhuizen kunt zoeken.

Een andere vorm van een specifiek zoeksysteem laat je zoeken binnen een bepaalde website. Het verschil met zoeksystemen als Google Scholar en jaap.nl is dat deze zoekmachines zoeken in een statische hoeveelheid gegevens. Voorbeelden daarvan zijn te vinden op bijvoorbeeld telefoongids.nl of ikea.nl. De laatste website biedt als extraatje een zogeheten avatar ('Anna') aan wie je vragen kan stellen. Bij de meeste zoekmachines kun je zoeken op een trefwoord en krijg je pagina's terug die die zoekterm bevatten. Zoeken met een avatar werkt anders. Stel dat je intypt dat je honger hebt, dan zal de avatar je een antwoord geven dat te maken heeft met het restaurant. Vertel je de avatar dat je dorst hebt, dan vertelt ze dat ze je geen suggestie kan doen. Dat komt omdat van tevoren is bedacht welke woorden in de consumentenvragen moeten matchen met welke producten en diensten. In dit geval is 'honger' wel gelinkt aan restaurant, maar 'dorst' niet.
Een goedwerkende vraag- en antwoord machine ontwikkelen kost veel tijd en wordt nog niet op grote schaal gebruikt.

Naast zoekmachines kun je ook gebruik maken van internetgidsen om informatie te vinden. Toen de eerste zoekmachines nog niet zo goed werkten als nu, waren de internetgidsen heel geschikt om snel relevante informatie te vinden over een bepaald onderwerp. Internetgidsen worden handmatig gemaakt, waarbij de makers zelf pagina's beoordelen op hun relevantie en kwaliteit. Het grote nadeel is dat internetgidsen voortdurend onderhouden moeten worden omdat pagina's verdwijnen en er ieder moment nieuwe informatie beschikbaar komt. Daarnaast is de selectie van de pagina's natuurlijk subjectief. Een bekend voorbeeld van een Nederlandse internetgids is http://www.startpagina.nl. Internationale internetgidsen zijn directory.google.com en directory.yahoo.com

Maak opdracht 4-1 en 4-2 en bekijk ook 4-3 (extra stof).

4b Uitgelicht: Google (Hoe werken zoekmachines?)

In deze paragraaf leer je hoe zoekmachines werken aan de hand van de zoekmachine Google. Bekijk eerst het onderstaande filmpje van Het Klokhuis over hoe Google werkt.

//www.youtube.com/embed/bLsIU_2Jes4

Kort samengevat bestaat Google dus uit de volgende onderdelen:

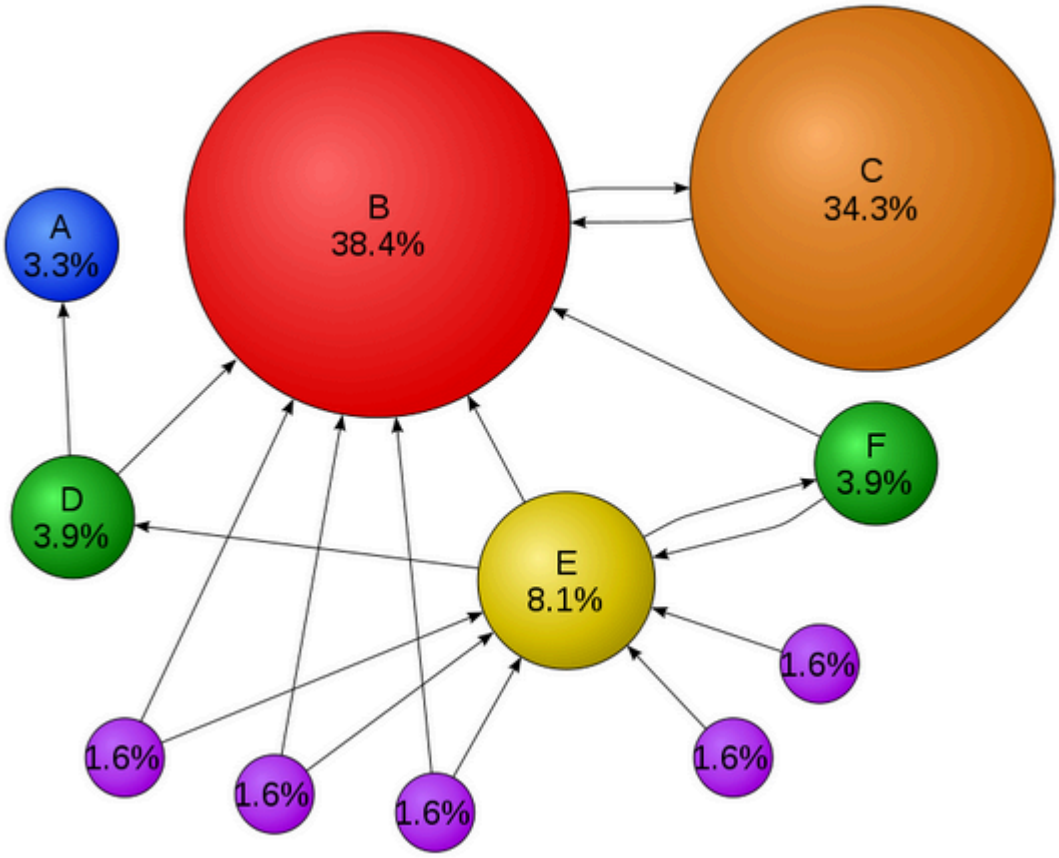
- Spider: haalt webpagina's op van het internet en extraheert de tekst.
- Indexer: bekijkt de teksten van de verschillende pagina's en geeft elk woord een score.
- Database: slaat alle lijsten met woorden op.
- Zoekmachine: kijkt welke websites uit de database de woorden bevatten die het beste bij een zoekterm passen.
- Webserver: zorgt ervoor dat de communicatie van Google met een gebruiker. Hij presenteert de resultaten die de zoekmachine vindt aan de gebruiker.

De zoekmachine in actie:

In deze animatie zie je alle onderdelen van Google in actie. De animatie bestaat uit vier knoppen.

Een zoekmachine geeft niet zomaar alle resultaten terug in een willekeurige volgorde. Om de gebruiker goed van dienst te zijn, gebruiken alle algemene zoekmachines bepaalde technieken om de beste pagina's bovenaan de zoekresultaten te laten verschijnen. Verschillende factoren kunnen daarbij worden meegewogen zoals bijvoorbeeld:

- **metatags;** dit zijn een soort sleutelwoorden die in de html van een pagina kunnen worden geplaatst. De maker van een pagina kan deze metatags zelf toevoegen. Tegenwoordig worden deze tags bijna niet meer gebruikt voor het meewegen omdat mensen door deze tags makkelijk hun positie op de ranglijst kunnen verbeteren en zelfs misleidende tags kunnen toevoegen.
- **datum laatste wijziging;** als een pagina al vier jaar niet mer gewijzigd is, kan het zijn dat informatie verouderd is. Dat kan een reden zijn om een pagina een lagere positie te geven in de zoekresultaten.
- **bezoekersaantallen;** als een pagina weinig wordt bezocht, kan het zijn dat deze weinig relevante of interessante informatie bevat. Ook dat kan een reden zijn van een lagere ranking.
- **de inhoud van andere pagina's op een website;** stel dat het woord 'vis' slechts eenmaal op een pagina voorkomt en op de andere pagina's van je website niet voorkomt, dan kan het zijn dat deze pagina niet over vissen gaat en dus minder interessant is voor iemand die informatie zoekt over vissen.
- **de inhoud van de websites die naar jou linken;** als in deze andere websites wel vaak het woord 'vis' voorkomt, dan is onze pagina over vis misschien toch wel relevanter dan op basis van het aantal woorden 'vis' verwacht kan worden.
- **de populariteit van sites die naar jou linken;** als veelbezochte websites naar jouw pagina doorlinken, dan kan dat iets zeggen over de kwaliteit van jouw pagina. Dat kan een reden zijn om die pagina een hogere positie te geven in de zoekresultaten.
- **het aantal pagina's dat naar een pagina linkt en waar zelf ook veel naar gelinkt wordt.** Dit is een van de belangrijkste parameters waarop de zoekresultaten van Google worden geordend, en heet *pageranking*. In onderstaand figuur kun je zien hoe het basaal werkt:



Bron: wikipedia. (http://nl.wikipedia.org/wiki/PageRank)

Stel dat het bovenstaande figuur het hele internet zou zijn, dan is de kans dat een willekeurige bezoeker pagina B bezoekt 38,4%. Dat komt omdat veel pagina's naar deze pagina linken. De kans dat iemand uitkomt op een van de paarse websites is 1,6% omdat niemand naar deze pagina's linkt. Je zou verwachten dat websites A en C een even groot percentage zouden moeten krijgen, omdat er telkens maar een website is die naar ze linkt. Echter; er wordt veel gelinkt naar website B waardoor deze een zekere autoriteit krijgt. Als deze naar een andere website linkt (C) weegt dat zwaarder dan de link van D naar A. Bovendien heeft C maar een link: die naar B. Daarmee wordt gesuggereerd dat er kennelijk een sterke relatie bestaat tussen A en C.

De formule voor de Google pagerank ziet er zo uit:

- PR(A) = (1 - d) + d * {(PR(T1) / C(T1) + ... + PR(Tn) / C(Tn))}
  d is de dampingfactor: de waarschijnlijkheid dat een gebruiker een pagina verlaat voor een andere pagina (standaard d=0.85)
  T1,T2,...Tn: citaties; dit zijn de pagina's die naar pagina A verwijzen
  C(x): dit zijn het aantal uitgaande links van pagina x
  PR(x) is de pagerank van pagina x

  Bron: http://nl.wikipedia.org/wiki/PageRank

Vrij vertaald staat in deze formule: hoe meer pagina's met een hoge pagerank en weinig links naar je linken, hoe hoger je pagerank wordt. Op de universiteit leer je precies wat die formule betekent en hoe het allemaal precies werkt.

Google's methode om de meest relevante websites bovenaan te plaatsen, werkt goed en is waarschijnlijk een grote reden van de populariteit van deze zoekmachine. Er zit ook een nadeel aan. Als jij op je website geen enkele link hebt, en niemand linkt naar jou, dan ben je onvindbaar. Daarnaast zou je kunnen stellen dat Google met deze methode een ijsberg creeert waarvan alleen het topje dat op de eerste pagina's verschijnt, druk bezocht wordt. Lager gerankte pagina's met even relevante informatie

krijgen minder attentie en dus minder links, terwijl er relatief steeds meer gelinkt zal worden naar pagina's die hoog eindigen en veel aandacht krijgen.

Lees eerst opdracht 4-4 door, bekijk de documentaire "Google: achter het scherm" van VPRO's Tegenlicht (50 minuten!) en maak dan opdracht 4-4.

"Google: achter het scherm":

https://maken.wikiwijs.nl/userfiles/8f4d7fc990e4b640d36dc6b6fab19b0a.swf

De zoekmachine in actie

1: Websites zoeken

- De spider zoekt het Internet af naar alle websites die hij kan vinden

2: De gevonden websites indexeren

- De spider extraheert alle woorden die hij op een website heeft gevonden en stuurt de teksten naar de indexer.
- De indexer telt hoe vaak een woord op een site voorkomt en slaat deze informatie op in de database.

3: De zoekpagina van Google opvragen

- De gebruiker vraagt de webpagina van Google om een zoekterm in te voeren

4: Een zoekterm invoeren

- De gevraagde zoektermen worden door de webserver doorgegeven aan de search engine
- De search engine bekijkt en de zoektermen en bepaalt op een geheime manier welke sites hij uit de database moet halen en naar de gebruiker moet sturen.

4c Zoeken op internet: tips en tricks

Globaal kun je op verschillende manieren zoeken op internet:

- Trefwoord: bijvoorbeeld 'boomgaard' op een algemene zoekmachine als Yahoo.
- Op onderverdeelde categorieen (zoeken op onderwerp): http://dir.yahoo.com/
- Geografisch: http://maps.google.com

Bedenk allereerst wat je wil gaan zoeken. Als je resultaten niet goed zijn, voeg dan zoektermen toe of maak ze meer of minder specifiek. Tenslotte kun je je zoekopdracht verfijnen door alleen in een specifieke bron te zoeken, bijvoorbeeld door site:marktplaats.nl aan je zoekopdracht toe te voegen.

Het volgende filmpje gaat over manieren om op internet te zoeken:

http://dotsub.com/media/f779c51c-8732-4df8-9836-b5b2df3a4fe4/embed/dut

Maak opdracht 4-5.

Een manier om de resultaten van je zoekopdracht te verfijnen is het gebruik van zogenaamde booleaanse [http://nl.wikipedia.org/wiki/George_Boole] operatoren. Hoe de operatoren eruit zien, is per zoekmachine verschillend maar de basisfuncties zijn altijd hetzelfde:

De operatoren AND, OR en NOT geven een relatie aan tussen zoektermen:
AND geeft aan dat termen tegelijkertijd in een pagina moeten voorkomen, bijvoorbeeld *bank* AND *hypotheek* of *kever* AND *auto*.
NOT geeft aan dat bepaalde termen zeker niet in een pagina moeten voorkomen, bijvoorbeeld *jaguar* NOT *auto*.
OR kan worden gebruikt wanneer het voorkomen van een van beide termen voldoende is, bijvoorbeeld *Clinton* OR*Obama* als je iets wil weten over de laatste twee democratische presidenten in de VS.

Het is ook mogelijk om deze operatoren te combineren, bijvoorbeeld:
bank AND *hypotheek* OR *lening* NOT *DSB*

Als je een specifieke combinatie van woorden zoekt, kun je ook gebruik maken van 'aanhalingstekens'. Typ je eigen naam maar eens in met en zonder deze aanhalingstekens en bekijk het verschil in het aantal resultaten.

Daarnaast hebben zoekmachines doorgaans nog extra functionaliteiten. Hieronder staat een gedetailleerde uitleg van een Google-pagina.

https://support.google.com/websearch/answer/136861?hl=en

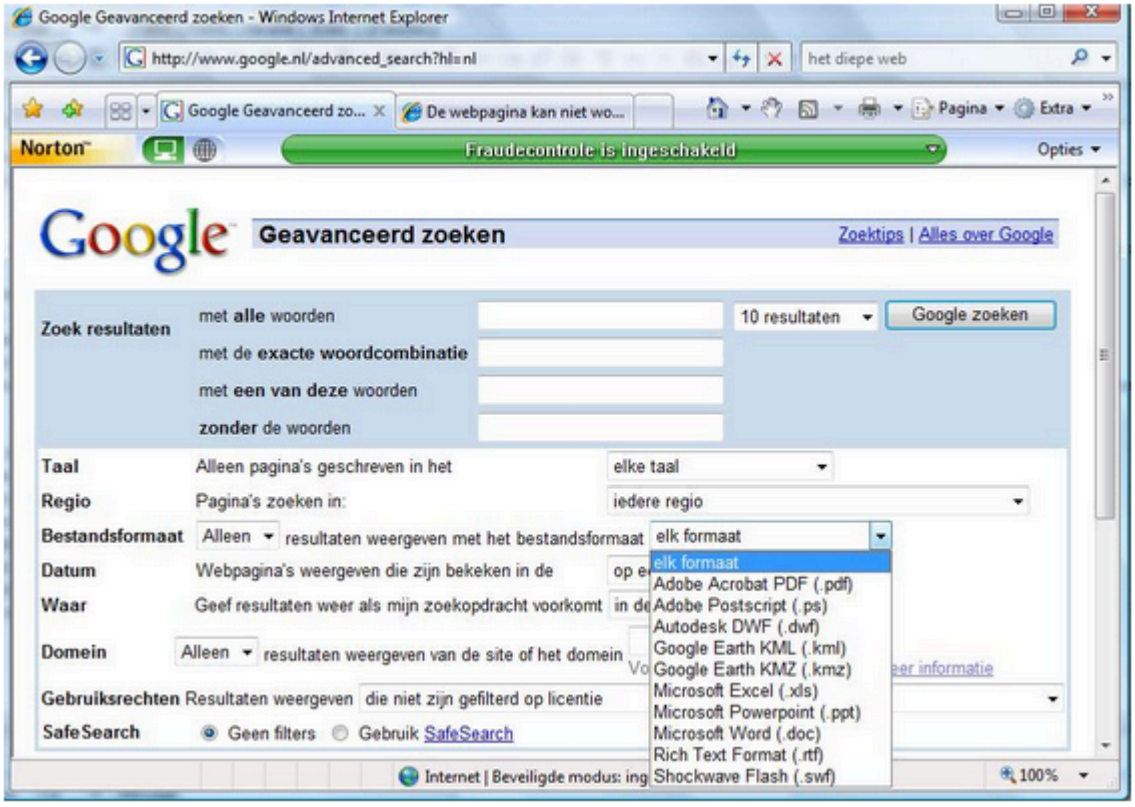
Maak opdracht 4-6.

4d Het verborgen internet

Er staat heel veel informatie op het internet dat zoekmachines wel kunnen vinden. Maar de verborgen informatie is ongeveer 400 keer zo groot. Dat komt doordat zoekmachines niet alle pagina's van een website indexeren. Daarnaast kan een website informatie bevatten die niet voor zoekmachines toegankelijk is:

- de informatie staat in een database, bijvoorbeeld telefoonnummers op telefoongids.nl
- de informatie staat op een website in de vorm van bijvoorbeeld een pdf of excelbestand
- de informatie is alleen beschikbaar nadat je hebt ingelogd.
- de informatie staat op een pagina die van het internet verdwenen is. (code 404, dode links)

Verborgen informatie: databases en documenten
Het is mogelijk om op zoek te gaan naar deze verborgen informatie. In dat geval is een goede strategie om eerst in een zoekmachine een aantal zoektermen in te vullen die relevante pagina's zullen opleveren. Nu kun je in die gevonden websites specifiek gaan zoeken door middel van de zoekfunctie op die pagina. Als de pagina geen zoekfunctie heeft, kun je gebruik maken van 'geavanceerd zoeken'. Als je zoekt met Google kun je bijvoorbeeld een domein opgeven. Jaguars op de website oldtimersnederland.nl kun je dan vinden door het volgende in te typen: *jaguar site:http://www.oldtimernederland.nl/*. Wil je geen informatie vinden van pagina's die eindigen op .com dan kun je dat ook aangeven met *site: -.com*



Ook is het mogelijk om op websites te zoeken naar bepaalde documenten; het documenttype kun je selecteren in het geavanceerd zoeken menu van Google. Het is dus ook mogelijk om binnen een website op trefwoord naar bepaalde documenten te zoeken.

Maak opdracht 4-7.

Upload nu de opdrachten van hoofdstuk 4 in de Postbus.

Over dit lesmateriaal

Colofon

Auteurs	Bètapartners
Team	Wikiwijs Maken Auteurs
Laatst gewijzigd	25 november 2014 om 20:24
Licentie	De Nederlandse Creative Commons 3.0 licentie waarbij de gebruiker het werk mag kopiëren, verspreiden en doorgeven en afgeleide werken mag maken onder de voorwaarden: Naamsvermelding en Gelijk Delen, zie http://creativecommons.org/licenses/by-sa/3.0/nl/. Meer informatie over de CC Naamsvermelding-GelijkDelen 3.0 Nederland licentie licentie.

Aanvullende informatie over dit lesmateriaal

Van dit lesmateriaal is de volgende aanvullende informatie beschikbaar:

Leerniveaus	HAVO 4, VWO 6, HAVO 5, VWO 5
Leerinhoud en doelen	Informatica
Eindgebruiker	leerling/student
Trefwoorden	a1 wetenschap en technologie, a2 maatschappij, e-klassen rearrangeerbaar

Bronnen

https://maken.wikiwijs.nl/userfiles/8f4d7fc990e4b640d36dc6b6fab19b0a.swf

//dotsub.com/media/f779c51c-8732-4df8-9836-b5b2df3a4fe4/embed/dut