Geschatte grootte van het geïndexeerde World Wide Web

Maurice de Kunder

Universiteit van Tilburg
Tilburg, Noord-Brabant
maurice@dekunder.nl

Versie: 1.4

Inhoudsopgave

1. Inleiding

De omvang van het World Wide Web neemt met iedere seconde toe. Elke toetsaanslag die ik hier maak gaat vergezeld met één, maar hoogstwaarschijnlijk veel meer nieuwe webpagina's op het World Wide Web. In 2001 groeide het World Wide Web met 7,5 miljoen webpagina's per dag [21]. Dit geeft al aan hoe enorm snel de grootte van het World Wide Web toeneemt. De zoekmachine Yahoo! claimt dat ze 19,2 miljard webpagina's [19] geïndexeerd heeft anno augustus 2005, Google volgt met 8 miljard pagina's [17] en MSN Search heeft volgens eigen zeggen 5 miljard pagina's [18] geïndexeerd. De cijfers zijn afkomstig van de zoekmachines zelf, maar in hoeverre zijn deze correct en betrouwbaar? Om daar achter te komen worden in deze scriptie zes metingen uitgevoerd om de grootte van de index (in aantal webpagina's) van de bovenstaande vier zoekmachines te schatten, op basis waarvan de totale grootte van het geindexeerde World Wide Web in termen van het aantal webpagina's geschat zal worden.

Door het dynamische karakter van het World Wide Web is het niet mogelijk om exact vast te stellen wat de grootte hiervan is. Dat neemt niet weg dat het mogelijk is om een schatting van de grootte van het geïndexeerde World Wide Web te maken. In deze scriptie wordt onder het geïndexeerde World Wide Web (ook wel Surface Web [22] genoemd) de webpagina's verstaan die door de grootste zoekmachines op dit moment (Google, MSN Search, Yahoo! en Ask) opgenomen zijn.

Om de grootte van het geïndexeerde World Wide Web te bepalen zijn in het verleden verschillende methodes gebruikt die kort beschreven worden in sectie 2.6. De grootte van het geïndexeerde World Wide Web zal in deze scriptie mede door gebruikmaking van Zipf's law [23] bepaald worden. In de methodologie zal Zipf's law verder naar voren komen. Daarnaast worden zes corpora gebruikt waarmee de metingen uitgevoerd zijn. Een corpus is een verzameling tekstuele documenten, bijvoorbeeld krantenartikelen of webpagina's. In zes metingen worden woorden uit de verschillende corpora logaritmisch geëxtraheerd, van hoogfrequente naar laagfrequente woorden. Daarna worden deze woorden gebruikt om queries op de zoekmachines Google, MSN Search, Yahoo! en Ask uit te voeren en wordt het totaal aantal geretourneerde resultaten opgevangen. De corpora dienen als basis om verhoudingsgetallen van woorden te bepalen aan de hand van documentfrequenties. Deze verhoudingsgetallen worden uiteindelijk vermenigvuldigd met de geretourneerde resultaten die door de 4 zoekmachines aangeleverd worden. Hierdoor wordt het mogelijk om een schatting van het aantal webpagina's per zoekmachine te bepalen. Om de totale grootte van het geïndexeerde World Wide Web te schatten wordt een methode beschreven die de overlap tussen zoekmachines schat, en die overlap in mindering brengt op de som van de geschatte groottes van de indexen van de zoekmachines.

2. Theoretisch kader

Sectie 2.1 geeft een korte beschrijving van enkele begrippen. De opvolgende secties 2.2 en 2.3 beschrijven het Surface World Wide Web, Deep World Wide Web en web crawlers. Daarna zal in sectie 2.4 beschreven worden waaruit het World Wide Web bestaat. In sectie 2.5 wordt een aantal factoren besproken die naar verwachting de metingen zullen beïnvloeden, en die deels (maar niet volledig) kunnen worden ingeschat. De laatste sectie 2.6 beschrijft eerdere benaderingen om de grootte van het World Wide Web vast te stellen.

2.1 Basisbegrippen

In deze scriptie komen enkele begrippen veelvuldig voor, in het bijzonder de begrippen hyperlink, webpagina, url, browser, zoekmachine, query en PHP-Script. In deze sectie worden deze begrippen kort beschreven.

Een hyperlink (ook wel link genoemd) is een verwijzing / koppeling van een webpagina naar een andere webpagina. Een hyperlink bestaat uit een url en een beschrijving waar op geklikt kan worden. Een url geeft de exacte locatie van een webpagina aan. Een voorbeeld van een url is: http://www.mijnwebsites.nl/mijnwebpagina.html. Een url verwijst in de meeste gevallen naar een webpagina op het World Wide Web. Soms ook naar ftp-servers en andere diensten waar in deze scriptie niet verder op ingegaan wordt. Een webpagina is een document die door een browser geopend en bekeken kan worden. Een webpagina is opgebouwd uit HTML-code. Een of meerdere webpagina's vormen een website. De grootte van het World Wide Web wordt in deze scriptie geschat op aantal webpagina's en niet aantal websites. Om een webpagina te bekijken gebruik je een browser. Een browser is een programma waarmee je webpagina's kan bekijken. De bekendste browsers op dit moment zijn Internet Explorer, Firefox en Opera.

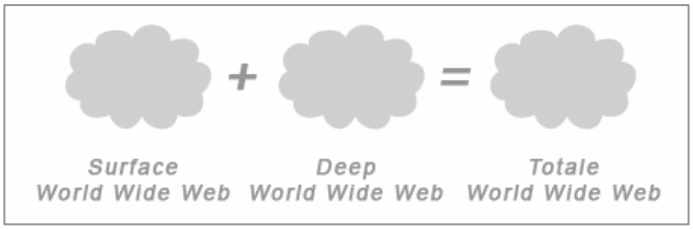
Alle zoekmachines op het World Wide Web maken gebruik van queries. Met een query kan een selectie op een groter geheel gemaakt worden. Je kan een zoekmachine bijvoorbeeld de opdracht geven om alle webpagina's die het woord "scriptie" bevatten op te halen. Je 'maakt' de query door eerst de term "scriptie" in het zoekvenster van de zoekmachine in te voeren waarna je op de knop "zoeken" klikt. Nadat op de knop "zoeken" geklikt is wordt een query richting de zoekmachine gestuurd, waarna de gebruiker de resultaten met het woord "scriptie" terug krijgt.

In de metingen (sectie 3 en 4) wordt alleen gebruik gemaakt van 1-woord-queries. Het is onmogelijk om met meerwoordsqueries (bijvoorbeeld "hans anders") schattingen van de grootte van de zoekmachines te maken. Ten eerste omdat de documentfrequenties in de corpora niet

zomaar opgeteld mogen worden (dit heeft te maken met Booleaanse logica (bv. keuzes tussen AND en OR)). Daarnaast is het onduidelijk of de deelnemende zoekmachines in de metingen dezelfde morfologische regels toepassen (zie sectie 2.5.4). Queries kunnen ook gebruikt worden in vrijwel alle programmeertalen, onder andere ook in de programmeertaal PHP. (Queries worden binnen een programmeertaal vaak SQL-statements genoemd.) In deze scriptie worden queries veelvuldig gebruikt om bijvoorbeeld de overlap tussen de zoekmachines te schatten. Ook worden queries gebruikt om de groottes van de vier deelnemende zoekmachines en het geïndexeerde World Wide Web te schatten. Door gebruikmaking van PHP-scripts kunnen meerdere queries automatisch achter elkaar naar een zoekmachine verzonden worden waarna het resultaat van de queries wordt opgeslagen in een database. Alle schattingen zijn automatisch door PHP-scripts uitgevoerd. In bijlage II staat de url waar alle PHP-scripts gedownload kunnen worden.

2.2 *Surface World Wide Web vs. Deep World Wide Web*

Het is ondoenlijk om elke webpagina van het totale World Wide Web in kaart brengen. Het is zelfs onmogelijk, omdat niet iedere webpagina zichzelf registreert bij een of meerdere zoekmachines of directories. Daarnaast zijn veel webpagina's niet van buitenaf te bereiken doordat er geen webpagina's zijn die via een hyperlink naar deze webpagina's verwijzen. Alle webpagina's die opgenomen zijn door zoekmachines (of directories) vormen het Surface World Wide Web, vanaf nu verkort tot Surface Web.



Figuur 1: Totale World Wide Web

Wanneer een webpagina (die geregistreerd is bij een zoekmachine) naar een webpagina verwijst waarvan het bestaan bij die desbetreffende zoekmachine onbekend is, dan zal deze onbekende webpagina vroeg of laat ook in een zoekmachine opgenomen worden. Deze aardige eigenschap zorgt ervoor dat zelfs wanneer een gebruiker zijn of haar webpagina's niet aanmeldt bij een zoekmachine, deze toch opgenomen zal worden in een zoekmachine als er een link van buitenaf bestaat naar deze webpagina.

Wanneer een webpagina echter niet geregistreerd is bij een zoekmachine en wanneer er geen enkele andere webpagina is die naar deze webpagina verwijst, dan zal deze altijd onzichtbaar blijven voor alle zoekmachines. Daarnaast maken steeds meer websites gebruik van dynamisch gegenereerde webpagina's. Zoekmogelijkheden op een website zijn daarvan een voorbeeld; de informatie wordt pas getoond nadat een zoekterm door een gebruiker opgegeven is. De getoonde webpagina's bestaan in dit soort situaties vaak niet als permanente HTML-pagina's,

maar worden dynamisch gegenereerd door een applicatie en een achterliggende database. Deze dynamisch gegenereerde content zorgt ervoor dat een groot gedeelte van het World Wide Web onzichtbaar is en blijft en niet opgenomen wordt in een of meerdere zoekmachines. Deze onzichtbare webpagina's vormen het Deep World Wide Web, vanaf nu verkort tot Deep Web.

Doordat zoekmachines nooit in staat zullen zijn om alle webpagina's van het Deep Web te indexeren kan alleen een grove schatting gemaakt worden om de grootte van het totale World Wide Web vast te stellen. In een poging om de grootte van het totale World Wide Web te bepalen heeft Bergman in 2001 een onderzoek uitgevoerd naar het Surface Web & Deep Web [21]. Hij kwam tot de conclusie dat het Deep Web 400 tot 550 maal groter was dan het Surface Web, wat op dat moment 1 miljard webpagina's groot was.

Dit onderzoek zal zich niet focussen op de grootte van het Deep Web, maar dit onderzoek zal een schatting maken hoe groot het Surface Web is. Het Surface Web is naar mijn mening een stuk interessanter en informatiever omdat alle webpagina's van het Surface Web opgenomen zijn bij verschillende zoekmachines waardoor gebruikers deze berg van informatie kunnen gebruiken en doorzoeken. Dit in tegenstelling met het Deep Web. De grootte van het Surface Web representeert het aantal webpagina's dat daadwerkelijk vindbaar en doorzoekbaar is. In deze scriptie zal de definitie van het Surface Web als volgt luiden: Het Surface Web bevat alle webpagina's waarvan het bestaan bekend is bij één van de vier grootste zoekmachines van dit moment, namelijk; Google, Yahoo!, Msn Search en Ask[1]. Deze zoekmachines hebben namelijk op het moment van schrijven de grootste databases [25] [28].

2.3 Zoekmachines en web crawlers

In maart 2006 heeft searchenginewatch.com het percentage zoekopdrachten naar verschillende zoekmachines van internetters thuis, op het werk en op universiteiten gemonitord [27]. Hierbij zijn 1,5 miljoen Engelssprekende mensen (waarvan 1 miljoen in de VS woonachtig) een maand lang gevolgd. Het gebruik van de zoekmachines was daarbij als volgt verdeeld: Google (42,7%), Yahoo! (28,0%) en Msn Search (13,2%), gevolgd door de zoekmachines van Aol (7,6%), Ask (5,9%) en overige (2,6%). Omdat Aol gebruik maakt van Google's databases is deze niet meegenomen in de metingen (sectie 3 en 4) maar heeft Ask deze plek ingenomen.

Om het Surface Web op te bouwen zijn zogenaamde web crawlers [14] nodig. Een web crawler, hierna crawler genoemd, zijn geautomatiseerde programma's die websites aflopen en uitlezen die zij tegenkomen. Crawlers worden gebruikt door zoekmachines om een database

[1] De Nederlandse zoekmachine Ilse.nl is niet meegenomen in de zes metingen omdat deze geen aantal webpagina's retourneert. Het is zodoende onmogelijk om de methodiek toe te passen.

(gegevensbank) op te bouwen. Er bestaan zeer veel zoekmachines die elk eigen crawlers hebben, maar er zijn maar enkele grote crawlers die het overgrote deel van het Surface Web in pacht hebben. Een crawler is als het ware de gegevensleverancier voor een zoekmachine.

De vier grootste crawlers zijn Googlebot (Google) [11], Inktomi / Yahoo Slurp (Yahoo!) [10], Msnbot (Msn Search) [9], Teoma (Ask) [7]. Daarnaast zijn er nog vele andere kleinere crawlers (200+) [12] [13]. De zoekmachines Alltheweb en Altavista maken gebruik van Yahoo's Inktomi gegevens [8], Lycos maakt gebruik van de database van Ask/Teoma, HotBot maakt gebruik van verschillende databases, waaronder Teoma en Googlebot en Aol maakt gebruik van Google's database.

Naast deze wirwar van databases zijn er ook nog meta-zoekmachines. Deze meta-zoekmachines verzamelen zoekresultaten van verschillende grote zoekmachines, ordenen de zoekresultaten en gegeven de zoekresultaten uiteindelijk weer. Zij beheren zelf dus niet de gegevens maar gebruiken bestaande data van onder meer Google, Yahoo!, Msn en Ask.

In dit onderzoek worden enkel de vier grootste zoekmachines van dit moment gebruikt in de verschillende metingen. Aangezien het daarnaast onmogelijk is om alle zoekmachines te analyseren is gekozen voor een opzet met de vier grootste zoekmachines. De overige zoekmachines maken daarnaast ook vaak gebruik van gegevensbanken van de vier grote zoekmachines.

Vanaf nu wordt met World Wide Web het Surface Web bedoeld tenzij anders aangegeven.

2.4 Bestandstypen en indexering

De grootte van het geïndexeerde World Wide Web is van vele factoren afhankelijk. Welke bestandstypen worden meegenomen om de grootte van het World Wide Web te bepalen, wanneer is een geretourneerde hit[2] een website of document en hoe gaan de verschillende zoekmachines om met het aantal resultaten die zij terug geven. In de begindagen van het World Wide Web werden hoofdzakelijk html-documenten geïndexeerd door de toen nog heersende zoekmachines Lycos en Altavista. Tegenwoordig worden naast webpagina's ook nog volop andere documenttypen meegenomen in het indexeerproces van Google, Msn, Yahoo! en Ask.

Google indexeert inmiddels de volgende bestandsformaten [1] [2]

- HyperText Markup Language (html)

[2] Een resultaat (bijvoorbeeld een webpagina of document) dat door een zoekmachine gevonden is.

- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)
- Lotus WordPro (lwp)
- MacWrite (mw)
- Microsoft Excel (xls)
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks, wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Shockwave Flash (swf)
- Text / ASCII (ans, txt)
- Meer dan 1 miljard Usenet berichten

Msn, Yahoo! en Ask hebben een soortgelijke lijst van bestandstypen. Dit impliceert dan ook dat de resultaten die door een zoekmachine geretourneerd worden niet allen 'echte' webpagina's (HTML-bestanden) zijn. Een deel van de resultaten bestaat uit een ander bestandsformaat. In de metingen is geen onderscheid gemaakt tussen webpagina's of andere typen bestanden. Het bestandstype maakt in feite niets uit; wanneer immers het zoeken naar informatie leidt tot het vinden van de benodigde informatie zal het er niet toe doen of deze informatie verkregen is op een webpagina of in een ander document.

In deze alinea zal kort ingegaan worden op het indexeren van documenten. Het indexeerproces omhelst het ophalen van webpagina's of andere documenten door crawlers, waarna deze verwerkt worden om uiteindelijk in een database van de desbetreffende zoekmachine terecht te komen. Zoekmachines waaronder Google, Msn, Yahoo! en Ask hebben, zoals eerder besproken meerdere crawlers om webpagina's uit te lezen. Bij het verwerken van webpagina's en andere documenten wordt alleen tekstuele informatie geëxtraheerd. Opmaakcodes en andere aanwezige informatie worden weg gestript. Hierna wordt de 'platte tekst', die uit het verwerken ontstaan is, opgeslagen in een of meerdere enorme databases van de verschillende zoekmachines. Alle databases gebruiken als opslagmethode een 'Inverted Index'. Een Inverted Index maakt het mogelijk om snel door een database te zoeken. Daarnaast werkt een Inverted Index ruimte besparend, omdat niet alle teksten volledig opgeslagen worden maar alleen de locaties van de woorden in de index tabel. Een Inverted Index is een indexstructuur waarbij woorden in één lijst opgeslagen worden en waarbij de ID's van woorden verwijzen naar hun oorspronkelijke locatie in één of meerdere documenten [29]. Sommige eigenschappen van een webdocument worden daarnaast gebruikt om o.a. de zoekresultaten te ordenen door prioriteit

toe te kennen aan bepaalde woorden. Voor het bepalen van de grootte van de zoekmachines en het World Wide Web heeft de manier van opslag en ordening geen effect.

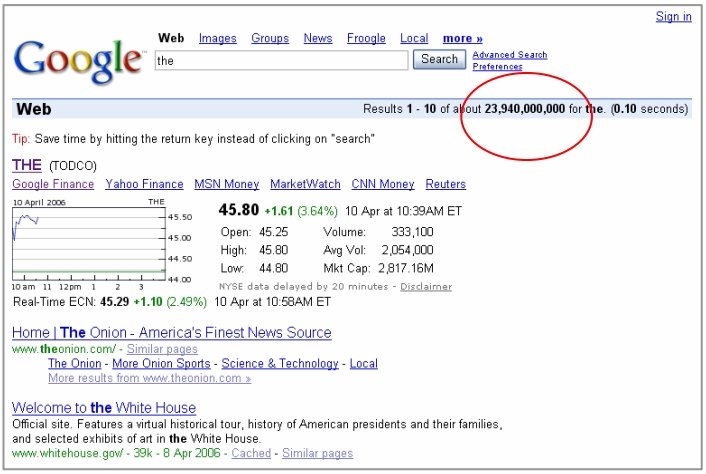
Naast webpagina's en documenten bieden de verschillende zoekmachines ook een mogelijkheid om te zoeken naar afbeeldingen. Deze optie maakt gebruik van de bestaande tekstuele database van Google, Msn, Yahoo! of Ask. De optie om op afbeeldingen te zoeken filtert op bestandskenmerken van afbeeldingen binnen de bestaande tekstuele database. Om deze stelling verder te verifiëren heb ik enkele steekproeven uitgevoerd.

Als eerste zijn keywords ingevoerd bij de vier zoekmachines (afbeeldingen optie geselecteerd). Hierna zijn de url's van de geretourneerde afbeeldingen gebruikt om in het standaard zoekgedeelte van de zoekmachine in te voeren (het tekstuele gedeelte van de zoekmachine). In alle gevallen (> 100) bleek de url ook te bestaan in dit gedeelte. Hierdoor kan aangenomen worden dat de resultaten, wanneer men zoekt in het tekstuele gedeelte van een zoekmachine, url's bevat van de afbeeldingenmogelijkheid van de verschillende zoekmachines en vice versa.

2.5 Bias

2.5.1 Betrouwbaarheid aantal geretourneerde resultaten

De grootste bias in dit onderzoek is de aanname dat het aantal resultaten die door de verschillende zoekmachines geretourneerd worden waar zijn (figuur 2). De rode omkadering geeft het aantal resultaten weer dat Google retourneert voor het woord "the". De geretourneerde aantallen worden in de metingen gebruikt om de groottes van alle vier de zoekmachines te schatten. In hoeverre zijn de geretourneerde aantallen eigenlijk juist? De zoekmachinemarkt wordt steeds groter en daardoor belangrijker en interessanter voor adverteerders. Hierdoor zullen zoekmachines zoals Google, Altavista, Yahoo! en Ask er alles aan doen om te zorgen dat zij de meest relevante resultaten als eerste weergeven (high precision). Minstens zo belangrijk is het aantal zoekresultaten dat door een zoekmachine geretourneerd wordt. Hoe groter het aantal zoekresultaten, des te groter een zoekmachine in de ogen van adverteerders en gebruikers lijkt (high recall). Adverteerders en gebruikers zijn daardoor



Figuur 2: Aantal geretourneerde resultaten

eventueel geneigd om bij zo'n grote zoekmachine eerder advertenties te plaatsen of deze zoekmachine in de toekomst te gaan / blijven gebruiken.

Er bestaat echter grote twijfel over de betrouwbaarheid van het aantal zoekresultaten die door een zoekmachine geretourneerd worden [3]. Jean Véronis vermeldt in zijn onderzoek een grote stijging van het aantal geretourneerde zoekresultaten voor Engelse en Franse woorden in de maand maart 2005 ten opzichte van februari 2005. Twee hypotheses kunnen deze stijging volgens Véronis verklaren.

1. De webpagina's die eerst alleen als url in de supplementaire index[3] werden vermeld zijn nu volledig geïndexeerd, hierdoor is het aandeel van de hoofdindex aanzienlijk gestegen.
2. Het indexaandeel is nog hetzelfde, maar de extrapolatieformules worden kunstmatig verhoogd (getuned), waardoor uiteindelijk de hoofd- en supplementaire index onzichtbaar wordt.

Evenals Véronis hoop ik dat hypothese 1 de juiste is. Er is meer onderzoek nodig om een van deze hypotheses te bevestigen. Deze bias kan dus tot op heden moeilijk opgeheven worden aangezien moeilijk inzicht te krijgen is in het percentage waarmee het World Wide Web in een bepaalde tijd daadwerkelijk groeit en of deze groei (gedeeltelijk) veroorzaakt wordt door Google, Altavista, Yahoo! en Ask zelf. Zij zouden immers de resultaten kunnen beïnvloeden door kunstmatig het aantal resultaten in een periode te laten stijgen.

Wouter Mettrop [26] schrijft in zijn onderzoek 'Internet Search Engines – fluctuations in document accessibility' dat fluctuaties veroorzaakt worden doordat zoekmachines gebruik maken van verschillende indexen, performance afhankelijke algoritmen en het verwijderen van identieke webpagina's uit de indexen. Hij heeft een minder argwanende kijk op de fluctuaties van zoekmachine resultaten dan Jean Véronis. In sectie 4 wordt verder ingegaan op het fluctuatiegedrag van de vier zoekmachines.

2.5.2 Afrondingen van geretourneerde aantallen

Drie van de vier zoekmachines geven, wanneer er meer dan 1000 resultaten gevonden zijn, een afgerond aantal terug. De zoekmachines die afgeronde getallen terug geven zijn Google, Yahoo!, Ask. Msn Search rondt bij geen enkele grootte af.

Google en Yahoo! ronden als volgt af, namelijk:

[3] De supplementaire index is een aanvulling op de hoofdindex, deze index bevat alleen maar de url's van webpagina's, de volledige content van deze webpagina's zijn nog niet daadwerkelijk gecrawld en opgenomen in de hoofdindex

Tot en met 1.000 worden niet afgerond	
1.000 tot en met 10.000 word afgerond op 10	1.234 > 1.230
10.000 tot en met 100.000 word afgerond op 100	12.345 > 12.300
100.000 tot en met 1 miljoen word afgerond op 1.000	123.456 > 123.000
1 miljoen tot en met 10 miljoen word afgerond op 10.000	1.234.567 > 1.230.000
10 miljoen tot en met 100 miljoen word afgerond op 100.000	12.345.678 > 12.300.000
100 miljoen tot en met 1 miljard word afgerond op 1.000.000	123.456.789 > 123.000.000
Miljarden worden afgerond op 10 miljoenen	1.234.567.890 > 1.230.000.000

Voor Ask gelden de volgende afrondingen:

Tot en met 1.000 worden niet afgerond	
1.000 tot en met 10.000 word afgerond op 10	1.234 > 1.230
10.000 tot en met 1 miljoen word afgerond op 100	12.345 > 12.300
1 miljoen tot en met 10 miljoen word afgerond op 1.000	1.234.567 > 1.235.000
Meer dan 10 miljoen worden afgerond op 10.000	12.345.678 > 12.350.000

Deze afrondingen zorgen ervoor dat de te schatte grootte van het World Wide Web in de metingen minder zuiver zijn voor de 3 zoekmachines Google, Yahoo! en Ask.

2.5.3 Dode links

Web crawlers indexeren webpagina's met een bepaalde regelmaat. Soms keren zij iedere dag terug naar een webpagina om de database van de desbetreffende zoekmachine up-to-date te houden, maar meestal worden webpagina's een keer in de week, in de maand of nog minder vaak opnieuw bezocht en geïndexeerd. Doordat de tijdsspanne tussen het eerste bezoek en het volgende bezoek van een crawler erg groot kan zijn bestaat de kans dat sommige webpagina's niet meer bestaan terwijl deze wel opgenomen zijn de database van een zoekmachine. Men spreekt in zo'n geval van een dode link. Het percentage van dode links verschilt per zoekmachine. Uit onderzoek van Lawrence en Giles [24] blijkt dat zoekmachines tussen de 1,6% tot 5,3% aan dode links bevatten. Het onderzoek van Lawrence en Giles vond plaats in 1997, in die tijd speelden de zoekmachines Google, Yahoo, Msn Search en Ask nog geen grote rol of zij bestonden nog niet. De zoekmachines Hotbot, Altavista, Northern Light, Excite, Infoseek en Lycos waren toen de grootste spelers. Hierdoor is het onmogelijk om de dode link percentages uit het onderzoek van Lawrence en Giles te gebruiken in deze scriptie, vandaar dat in deze scriptie de dode link percentages opnieuw geschat zijn.

Om het percentage aan dode links voor de vier zoekmachines Google, Msn Search, Yahoo! en Ask te schatten, zijn 82.753 url's uit de overlap-steekproeven 1 en 2 gebruikt (zie sectie 3.2). Via een PHP-script zijn de url's gecontroleerd of deze vindbaar waren op het World Wide Web. Doordat webservers tijdelijk niet bereikbaar kunnen zijn, is het PHP-script drie maal herhaald, verspreid over één week. Na de controle bleek dat Yahoo! 1,95% aan dode links bevatte, gevolgd door Msn

Search 2,52%, Google 2,72% en Ask eindigde met maar liefst 4,59%. Deze percentages zijn in mindering gebracht op de geschatte grootte per zoekmachine, zie metingen sectie 4.1 t/m 4.6. Bijlage I toont de return-codes welke onder de definitie 'url niet gevonden / dode link' valt.

	Url's bezocht	Url's niet gevonden / dode links	Percentage
Ask	21030	966	4,59%
Google	21063	573	2,72%
Msn Search	19550	493	2,52%
Yahoo!	21110	412	1,95%

Tabel 1: Dode link percentages

2.5.4 Morfologische regels

Morphology rules of in het Nederlands "morfologische regels" kunnen ervoor zorgen dat zoekmachines eenzelfde query (bijvoorbeeld het woord "bloemen") anders interpreteren.
Zo kan zoekmachine X resultaten tonen voor het woord "bloemen" maar ook voor het woord "bloem" of "bloementuin", terwijl zoekmachine Y alleen resultaten toont voor het exacte woord "bloemen". Een oplossing voor deze bias is door alle resultaten (url's) die een zoekmachine retourneert voor een bepaald woord, te controleren op aanwezigheid van dat woord. Helaas is deze methode tegenwoordig onmogelijk. Laagfrequente woorden leveren bij de kleinste zoekmachines al snel meer dan 1.000 resultaten op, hoogfrequente woorden zelfs miljoenen of miljarden. Daarnaast laten de meeste zoekmachines niet toe om verder dan de eerste 1000 resultaten te bladeren, waardoor niet alle url's kunnen worden gecontroleerd.

2.6 Eerdere benaderingen

In deze sectie worden drie eerdere onderzoeken naar de grootte van het World Wide Web kritisch belicht.

2.6.1 Lawrence en Giles: Searching the World Wide Web

Lawrence en Giles hebben in 1998 de grootte van het geïndexeerde web geschat door de overlap tussen zoekmachines vast te stellen. Om de overlap vast te stellen hebben Lawrence en Giles 575 queries naar zes zoekmachines verzonden die in 1998 populair waren (HotBot, Altavista, Northern Light, Excite, Infoseek en Lycos). Alle url's die door de zoekmachines geretourneerd werden zijn, na het verzenden van een query, bezocht. Hierbij werden voor alle geretourneerde url's de desbetreffende webpagina's opgehaald. Daarna werd gecontroleerd of deze webpagina's ook daadwerkelijk één of meerdere woorden van de desbetreffende query bevatten. Volgens Lawrence en Giles worden er namelijk webpagina's door zoekmachines geretourneerd die niet

meer bestaan of veranderd zijn en daardoor niet meer het desbetreffende woord(en) bevatten. Daarnaast geven zij aan dat zoekmachines verschillende "morphology rules" gebruiken, zie sectie 2.5.4. Om deze bias te voorkomen worden, zoals zojuist beschreven, alle url's gecontroleerd en het aantal geretourneerde resultaten naar beneden bijgesteld. De geschatte grootte van het geïndexeerde World Wide Web bedroeg in Augustus 1997 volgens Lawrence en Giles minstens 320 miljoen webpagina's.

Kritiek

Lawrence en Giles hebben in het experiment gebruik gemaakt van 575 queries. Wanneer een zoekmachine minder dan 600 url's voor een query retourneerde werd deze gebruikt in het experiment. Deze afbakening zorgt ervoor dat queries met frequent voorkomende woorden niet meegenomen worden in het experiment, waardoor de grootte van het World Wide Web geschat wordt op basis van laagfrequente woorden. Tegenwoordig is deze methode voor de meeste woorden of woordcombinaties onuitvoerbaar. Een zoekmachine retourneert al snel voor een zeldzaam woord meer dan 1.000 webpagina's; hierdoor wordt het onmogelijk om alle webpagina's te controleren. Ten eerste is het praktisch onuitvoerbaar om voor elk woord alle url's na te speuren, waardoor in sommige situaties miljoenen, zoniet miljarden webpagina's nagelopen moeten worden. Ten tweede kunnen de meeste zoekmachines niet verder bladeren dan 1.000 resultaten / url's.

2.6.2 Gulli en Signorini: The Indexable Web is More than 11.5 billion pages

De grootte van het geïndexeerde World Wide Web is in deze studie geschat door de grootte van, en overlap tussen, de zoekmachines te schatten. Daarbij maken Gulli en Signorini (2005) gebruik van een methodiek die door Bharat en Broder [34] is voorgesteld. Om de overlap en grootte van de zoekmachines vast te stellen zijn twee procedures van Bharat en Broder (1998) gebruikt. In de eerste procedure worden willekeurig url's uit de index van de deelnemende zoekmachines gesampled (sampling procedure). In de tweede procedure worden de gesamplede url's bij de deelnemende zoekmachines gecontroleerd waarbij gekeken wordt welke url's in de index staan (checking procedure).

Sampling procedure

Om willekeurig url's uit een zoekmachine-database te selecteren word door Bharat en Broder (1998) voorgesteld om een set van enkelvoudige en gecombineerde woordqueries te verzenden naar een zoekmachine. Hierna werd uit de eerste 100 resultaten één url willekeurig geselecteerd. Om een corpus met woorden op te bouwen is door Gulli en Signorini (2005) de Dmoz directory gebruikt, anders dan bij Bharat en Broder (1998), die de Yahoo! directory gebruikt hebben. De Dmoz directory en Yahoo! directory zijn beide grote verzamelingen gecategoriseerde url's met

een korte omschrijving. Deze directories worden door mensen bijgehouden, zij dragen zorg dat de url's inclusief een korte omschrijving up-to-date zijn en in de juiste categorie vermeld staan.

De woorden van de Dmoz directory werden gesorteerd op voorkomen en gesplitst in blokken van 20 woorden. Binnen elk blok werd voor elke zoekmachine één woord geselecteerd om uiteindelijk 438.141 queries met één term te vormen. Deze queries werden verzonden naar Google, Msn Search, Yahoo! en Ask.

Checking procedure

Bharat en Broder gebruikten een query gebaseerde controleprocedure. Van elke url werden de meest voorkomende woorden verzonden naar een zoekmachine, waarna gecontroleerd werd of de zoekmachine de daarbij behorende url weergaf in de lijst met resultaten. Gulli en Signorini (2005) gebruikten echter een vereenvoudigde vorm voor de bestaanscontrole van een url bij een zoekmachine. Alle zoekmachines die door Gulli en Signorini (2005) gebruikt zijn boden een mogelijkheid om het bestaan van een url direct te controleren. Dit vereiste wel dat de url's genormaliseerd waren; sommige url's bevatten namelijk 'vreemde' karakters die niet door het controlemechanisme van zoekmachines herkend worden.

Om een schatting van de grootte van het geïndexeerde World Wide Web te maken gebruikten Gulli en Signorini (2005) de volgende methode. Voor elke zoekmachine Z werd het gemiddelde van de drie andere zoekmachines geschat die ook webpagina's van zoekmachine Z geïndexeerd hadden. Uiteindelijk was hierdoor de overlap tussen de zoekmachines bekend en de relatieve en absolute groottes van elke zoekmachine. De grootte van het geïndexeerde World Wide Web werd in april 2005 op tenminste 11,5 miljard webpagina's geschat. De 11,5 miljard webpagina's zijn het gemiddelde van de geschatte relatieve en absolute groottes voor alle zoekmachines gezamenlijk.

Kritiek

Het opdelen in blokken van 20 woorden is nergens op gebaseerd. Waarom niet 50 of 100? Wellicht hadden de woorden geselecteerd kunnen worden op basis van Zipf's law (zie sectie 3.3). De vereenvoudigde methode van Gulli en Signorini om url's te controleren kan eenvoudig toegepast worden en geeft direct resultaat voor genormaliseerde url's. Het normaliseren van url's verstoort echter wel de willekeurige selectie van url's. Door het normaliseren worden lange url's en url's zonder standaard karakters uitgesloten van deelname. Deze selectieve uitsluiting kan ervoor zorgen dat de overlap percentages onnauwkeurig zijn. Juist dit soort url's kunnen de overlap tussen de zoekmachines verkleinen. In deze scriptie zullen url's niet genormaliseerd worden, maar worden url's met speciale karakters hergecodeerd zodat de zoekmachine de url's juist interpreteert.

2.6.3 Rhodenize en Trudel: Estimating the size and content of the World Wide Web

Anders dan de vorige twee onderzoeken gebruiken Rhodenize en Trudel (2003) 'quadrats' om de grootte van het World Wide Web te schatten. Deze techniek is afkomstig uit de biologie en wordt daar gebruikt om schattingen van populaties te maken. Om de grootte van het World Wide Web te schatten genereren zij willekeurig IP-adressen verdeeld over A-, B- en C-klasse netwerkadressen. Bij een A-klasse netwerk staan de eerste 8 bits van het IP-adres vast, de laatste 24 geven het lokale netwerk aan. Bij een B-klasse netwerk staan de eerste 16 bits van het IP-adres vast, de laatste 16 geven het lokale netwerk aan. Bij een C-klasse netwerk staan de eerste 24 bits van het IP-adres vast, de laatste 8 geven het lokale netwerk aan. Wanneer een IP-adres begint met 1 t/m 126, dan heb je te maken met een A-klasse netwerk met maximaal 16 miljoen hosts[4], een IP-adres van een B-klasse netwerk begint met 128 t/m 191 en kan maximaal 65.534 hosts bevatten. Een C-klasse netwerk heeft IP-adressen beginnend met 192 t/m 223 [30], deze klasse kan maximaal 254 hosts bevatten. Niet alle 8 bits (0 t/m 255) worden gebruikt, de overige nummers zijn gereserveerd voor o.a. test doeleinden.
Elk IP-adres wordt in het onderzoek van Rhodenize en Trudel met een Perl-script op poort 80 bezocht (standaard poortnummer voor een webserver). Er zijn 197.100 verschillende IP-adressen bezocht (65.700 IP-adressen per netwerk klasse). Wanneer het Perl-script een response van de server op een van de 197.100 IP-adressen ontving, is deze aangemerkt als webserver. Van de 65.700 IP-adressen uit Klasse A waren 166 (0,253%) webservers, van de 65.700 IP-adressen uit Klasse B waren 307 (0,467%) webservers en van de 65.700 IP-adressen uit Klasse C waren 992 (1,510%) webservers. Om de populatiegrootte vast te stellen zijn het aantal gevonden webservers vermenigvuldigd met het totaal aantal quadrats (mogelijke combinaties) voor elk van de drie netwerkklassen. In tabel 2 worden de geschatte populatiegroottes weergegeven.

Klasse A Netwerk	Klasse B Netwerk	Klasse C Netwerk	Totaal
5.298.740	5.012.280	8.041.867	18.352.887

Tabel 2: Geschatte aantal webservers per netwerk klasse

Het totale aantal webservers dat via de quadrats methode geëxtrapoleerd werd kwam uit op 18,50 miljoen webservers in de periode 16 februari t/m 15 maart 2003.

Kritiek

Rhodenize en Trudel maken een schatting van het aantal webservers, maar dit zegt niets over het aantal webpagina's of websites. Op één webserver met één IP-adres kunnen immers meerdere websites draaien. Daarnaast controleren zij alleen poort 80, maar potentieel kan een webserver eveneens op een van de andere 64.535 beschikbare poortnummers draaien.

[4] Computers, routers en andere randapparatuur

3. Methodologie

In dit onderdeel worden de verschillende stappen uitgewerkt die benodigd waren om de metingen gestalte te geven. Sectie 3.1 beschrijft het opbouwen en analyseren van de corpora. Sectie 3.2 beschrijft het belang van het bepalen van de overlap tussen de zoekmachines. Daarnaast wordt in deze sectie beschreven hoe de overlap tussen zoekmachines geschat kan worden, en worden resultaten gegeven van deze overlap-schattingsmethode op de gebruikte zoekmachines. De laatste sectie 3.3 geeft een beschrijving van Zipf's law met daarbij een grafische weergave van de zes corpora.

3.1 Opbouwen en analyseren van de corpora

Voor het uitvoeren van de metingen zijn zes corpora opgebouwd, geanalyseerd en uiteindelijk is het extract hiervan (woordfrequenties + documentenfrequenties) gebruikt in de metingen. Een corpus is een verzameling artikelen of webpagina's terug gebracht tot een ASCII-codering (platte tekst).

Om de grootte van het World Wide Web te bepalen kan een corpus uitkomst bieden. Het idee daarachter is als volgt: Wanneer van een bepaald woord bekend is wat het aandeel van dat woord binnen het corpus is, kan dit getal gebruikt worden samen met het aantal gevonden webpagina's dat bijvoorbeeld Google terug geeft voor datzelfde woord. Dit aandeel word ook wel documentfrequentie genoemd.

Een vaak voorkomend engels woord is het woord *the.* Stel nu dat Google voor het woord *the* 10.000.000.000 resultaten (aantal webpagina's / documentfrequentie) gevonden heeft en dat binnen een corpus het woord *the* in 25.000 van de in het totaal 50.000 artikelen of webpagina's voorkomt (50% van de artikelen of webpagina's bevat het woord *the*). Dan kan de volgende, zeer eenvoudige, rekensom gemaakt worden.

$$\frac{\text{documentfrequentie zoekmachine}}{\text{documentfrequentie corpus}} * \text{Tot. aantal docu. in corpus} = \text{Geschat aantal webpagina's}$$

$$\frac{10.000.000.000}{25.000} * 50.000 = 20.000.000.000$$

De grootte van het door Google geïndexeerde web zou dan volgens deze schatting 20 miljard webpagina's zijn. Dit voorbeeld rekent nu met één woord en één zoekmachine. Wanneer deze schatting meerdere keren voor verschillende woorden en met verschillende zoekmachines gemaakt wordt dan wordt het mogelijk om een gemiddelde grootte van het World Wide Web te schatten.

De geschatte grootte van het World Wide Web is afhankelijk van het corpus. In corpus X kan het woord *the* 50% voorkomen terwijl in corpus Y het woord *the* 60% voorkomt, dit zorgt er dan ook voor dat de geschatte grootte van het World Wide Web verschilt per corpus.

In deze scriptie wordt gebruik gemaakt van zes verschillende corpora, namelijk;

- 2 Krantencorpora (Nederlands + Engels)
- 2 Dmoz corpora (Nederlands + verschillende talen (merendeel Engels))
- 2 Wikipedia corpora (Nederlands + Engels)

Iedere corpora heeft een eigen 'vingerafdruk', zo zal een krantencorpus grotere artikelen en hierdoor gemiddeld meer woorden bevatten dan een webpagina. Bij webpagina's kan de context over meerdere webpagina's verdeeld worden, terwijl een krantenartikel de gehele context bevat. Hierdoor zal het aantal woorden op een webpagina vaak lager liggen dan een krantencorpus. Zo zijn er nog een aantal andere verschillen tussen de corpora. De verschillen tussen de corpora zullen mogelijk leiden tot verschillen in de geschatte grootte van het World Wide Web.

Genereren van woord- en documentfrequenties

Een belangrijke stap in het proces is het genereren van de woord- en documentfrequenties van alle artikelen en webpagina's uit de zes corpora. Woord- en documentfrequenties van de Kranten-corpora (Engels + Nederlands), Dmoz corpora (Engels + Nederlands) en de Wikipedia corpora (Engels + Nederlands) worden met behulp van een PHP-Script automatisch geteld.

In figuur 3 staat een gedeelte van de woord- ('wf' in figuur) en documentfrequenties ('df' in figuur) uit de Engelse krantencorpus weergegeven, oftewel een frequentietabel. Deze lijst varieert per corpus op lengte, woord- en documentfrequentie. Corpus X heeft daarom ook altijd een andere 'vingerafdruk' dan corpus Y.

In het eerste veld staat het woord ('keyword 'in figuur) vermeldt, gevolgd door de woordfrequentie, deze geeft

	Keyword	wf	df
	wives	3899	3137
	also	433208	228722
	wrote	49216	32939
	scholarly	1600	1431
	articles	11665	8503
	on	2369166	381139
	jonathan	7834	6692
	swift	4530	3158
	richard	44267	36108

Figuur 3: Frequentietabel Engelse krantencorpus

aan hoe vaak het desbetreffende woord (bijvoorbeeld ‘wives’) voorkomt in het corpus. In dit geval komt het woord ‘wives’ 3899 maal voor in het corpus van 398.247 Engelse krantenartikelen. De documentfrequentie geeft aan in hoeveel documenten een woord voorkomt. In het Engelse krantencorpus met 398.247 artikelen komt het woord ‘wives’ in 3.137 documenten voor. De kans dat het woord ‘wives’ voorkomt in de geanalyseerde artikelen uit het Engelse krantencorpus is dus: 3137/398.247*100 = 0,79%. Veel groter is de kans dat het woord ‘also‘ (57,43%) en ‘on’ (95,70%) voorkomt. Met deze gegevens kan dus bepaald worden wat het aandeel van een woord in een corpus is.

Ieder corpus heeft een soortgelijke frequentietabel zoals deze gedeeltelijk afgebeeld staat in figuur 3. Deze frequentietabel vormt de basis voor de daadwerkelijke metingen. De documentfrequentie zal onder andere gebruikt worden om de geschatte grootte van het World Wide Web te bepalen. Hierbij zal het percentage van de documentfrequenties per woord doorberekenend worden naar een schatting van de totale grootte van het World Wide Web in termen van aantallen webpagina’s.

Restricties bij het verwerken

Bij het vergaren van alle woord- en documentfrequenties voor het Dmoz corpus zijn er enkele restricties opgelegd bij het verwerken van de webpagina’s. De Wikipedia en krantencorpora bestonden uit enkele grote databases met teksten, bij het verwerken golden dezelfde restricties als voor het Dmoz corpus (wanneer van toepassing). De restricties die werden gehanteerd zijn als volgt:

1) Websites met zogenaamde frames zijn uitgesloten. Een webpagina met frames bestaat uit twee of meerdere pagina’s. Vaak is er één voor het menu gereserveerd en één voor de content. Hierbij ontstaat het probleem dat onduidelijk is welke webpagina het menu is of welke de content bevat, vandaar dat deze webpagina’s niet meegenomen zijn. Veelal worden dit soort pagina’s eveneens niet meegenomen door zoekrobots (crawlers die door zoekmachines zoals Google, Msn Search, Yahoo! en Ask gebruikt worden).
2) Er geldt een maximale connectie time-out van 5 seconden, als binnen deze tijd niet door de server gereageerd is wordt de webpagina van deze server uitgesloten van deelname.
3) Server redirects worden tot maximaal 2 niveaus diep gevolgd.
4) Client redirects worden niet gevolgd, de web crawlers van Google, MSN Search, Yahoo! en Ask doen dit in de meeste gevallen ook niet, browsers daarentegen wel.
5) Websites die het teken ‘ð’ bevatten worden niet geïndexeerd en verwerkt, dit om problemen te voorkomen met tekensets die niet het westerse alfabet gebruiken. Zodoende worden woorden in het Chinees, Russisch, Japans etc. niet meegenomen.

6) Daarnaast worden websites die geen of minder dan 10 unieke woorden bevatten uitgesloten van deelname. Hierbij gaat het in de meeste gevallen om webpagina's die niet meer bestaan, die in Flash of m.b.v. afbeeldingen gemaakt zijn. Uit Flashcontent en afbeeldingen kunnen namelijk geen woorden afgeleid worden. Daarnaast zijn sommige van deze webpagina's bedoeld als intro en hebben daardoor erg weinig woorden op de begin pagina.
7) Webpagina's groter dan 500 kb worden niet verwerkt, dit omdat grotere bestanden een te grote belasting vormen voor de server.

De zes corpora

In tabel 3 staan enkele gegevens omtrent de zes corpora weergegeven. Deze gegevens zijn verkregen naar verwerking van ieder van de corpora.

Corpora	*Taal*	Aantal artikelen of webpagina's	Aantal woorden	Unieke woorden	Gemiddeld aantal woorden per artikel of webpagina
Kranten	*ENG*	398.247	308.771.369	1.542.878	775
	NL	163.227	70.677.208	787.794	433
Dmoz	*VT*	531.624 [5]	254.094.395	4.395.017	478
	NL	35.383 [6]	13.880.707	593.994	392
Wikipedia	*ENG*	1.036.437 [7]	274.131.085	2.744.183	264
	NL	163.445 [8]	36.457.793	956.302	223

NL = Nederlands / ENG = Engels / VT= Verschillende talen

Tabel 3: Samenvatting corporagegevens

3.1.1 Krantencorpora

Engels

De eerste corpus bevat krantenartikelen uit The New York Times tussen 1999 en 2002. Dit corpus was in eerste instantie nog niet bruikbaar omdat het alleen hele artikelen bevatte. Om het corpus bruikbaar te maken is het omgezet naar een tabel zoals op pagina 17 beschreven is. De tabel bevat 1.542.878 unieke woorden.

Nederlands

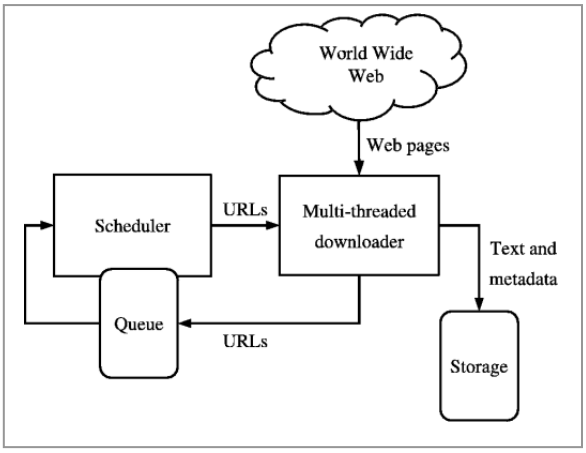
De tweede corpus bevat krantenartikelen uit Nederlandse kranten (De Gelderlander 1997 en 1998 + ANDA). ANDA is de naam van een (niet meer bestaande) gezamenlijke redactie van regionale dagbladen van het Wegener-concern. Deze redactie verzorgde buitenland-, economie-, en achtergrondartikelen die alle regionale kranten konden overnemen. Deze corpus is eveneens verwerkt en omgezet; De relevante gegevens zijn in tabel 3 terug te vinden. Na het verwerken ontstond een tabel ter grootte van 787.794 unieke woorden.

[5] Bruikbaar van 761.817 webpagina's
[6] Bruikbaar van 66.572 webpagina's
[7] Bruikbaar van 1.294.634 artikelen
[8] Bruikbaar van 215.026 artikelen

3.1.2 Dmoz corpora

Verschillende talen

Het Dmoz corpus bevat woorden van 531.624 willekeurig geselecteerde webpagina's uit de Dmoz.org directory. De Dmoz.org directory bevat meer dan 4,4 miljoen webpagina's in verschillende talen. Door middel van een PHP-script zijn van de 4,4 miljoen websites 761.817 webpagina's willekeurig geselecteerd en bezocht. Uiteindelijk waren 69,8% webpagina's bruikbaar; de overige 30,2% webpagina's bestonden niet meer of voldeden niet aan de op de pagina 18 opgestelde restricties.



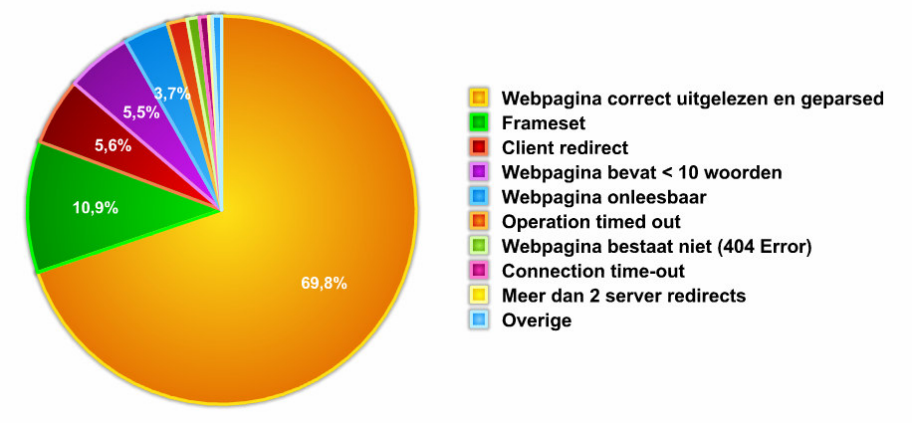
Figuur 4: Werking van de crawler
Overgenomen uit Wikipedia

De twee krantencorpora konden direct verwerkt worden, maar voor het Dmoz corpus moesten 761.817 webpagina's bezocht worden. Het ophalen van de enorme hoeveelheid verschillende webpagina's verliep met een snelheid van ongeveer 3100 tot 3200 webpagina's per uur. Het ophalen van de webpagina's gebeurde met behulp van een crawler. Deze crawler is geschreven in de programmeertaal PHP. De werking van de crawler is in figuur 4 weergegeven.

Het duurde meer dan 10 dagen om de webpagina's te crawlen en te verwerken. Het crawlen en verwerken heeft in de periode van 21 februari t/m 8 maart 2006 plaatsgevonden.

Het Dmoz corpus kan gezien worden als representatie van webpagina's op het World Wide Web. De vingerafdruk van deze corpus zal het meest overeenstemmen met de vingerafdruk van het World Wide Web. Hoogstwaarschijnlijk zal met behulp van deze grote steekproef een betrouwbare schatting van het World Wide Web bepaald kunnen worden.

De woorden van de 531.624 bruikbare webpagina's zijn verwerkt naar een frequentiemodel zoals in figuur 3 op pagina 17. In totaal zijn 254.101.756 woorden geteld. Veel woorden komen meerdere keren op verschillende webpagina's voor waardoor uiteindelijk van 4.395.025 unieke woorden de woord- en documentfrequentie bekend was.



	Aantal	Percentage
Webpagina correct uitgelezen en verwerkt	531.624	69,8%
Frameset	83.357	10,9%
Client redirect	42.466	5,6%
Webpagina bevat < 10 woorden	41.593	5,5%
Webpagina onleesbaar	28.385	3,7%
Operation timed out	12.545	1,6%
Webpagina bestaat niet (404 Error)	7.655	1,0%
Connection time-out	5.865	0,8%
Meer dan 2 server redirects	2.218	0,3%
Overige	6.109	0,8%
Totaal	761.817	100%

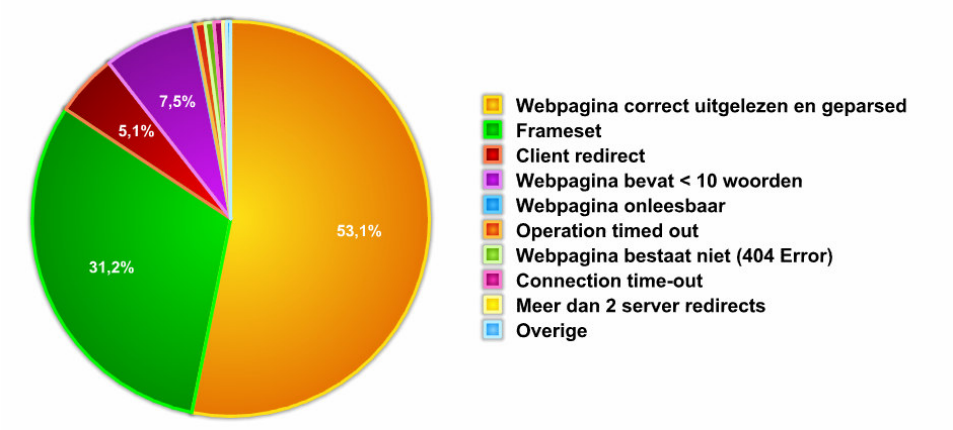
Tabel 4: Verwerking Dmoz corpus (Verschillende talen)

Zoals tabel 4 laat zien zijn 30,2% van de webpagina's niet gebruikt in de verwerking en opname in het corpus. 10,9% van de webpagina's bevatte een frameset, zo'n webpagina bevat niet de daadwerkelijke tekstuele inhoud van een webpagina. Een frameset zorgt alleen voor de weergave van meerdere webpagina's. Omdat door een frameset onduidelijk wordt welke van deze webpagina's de tekstuele inhoud bevat, zijn framesets uitgesloten van opname in het corpus. Daarnaast werden 5,6% webpagina's doorgestuurd naar andere webpagina's, deze webpagina's bevatten dus geen tekst, vandaar dat deze ook uitgesloten zijn van opname. 5,5% van de webpagina's bevatte minder dan 10 woorden. 3,7% van de webpagina's bevatte vreemde tekens. De overige items in Tabel 4 voldeden anderszins niet aan de op pagina 18 opgestelde restricties bij het verwerken.

Nederlands

De vierde corpus bevat woorden van 35.383 Nederlandse webpagina's. De Nederlandse webpagina's zijn, evenals de Engelse webpagina's, verkregen via de Dmoz.org directory. Van de

4,4 miljoen webpagina's in de Dmoz.org directory zijn alle 66.572 Nederlandse webpagina's, die de extensie '.nl' bevatte, bezocht met dezelfde crawler die ook voor het Dmoz corpus (verschillende talen) gebruikt is. Van de 66.572 webpagina's welke bezocht zijn door de crawler was uiteindelijk maar 53,1% bruikbaar. De overige 46,9% webpagina's bevatte framesets die niet verwerkt konden worden, bevatten minder dan 10 woorden of voldeden anderszins niet aan de op pagina 18 opgestelde restricties.



	Aantal	Percentage
Webpagina correct uitgelezen en verwerkt	35.383	53,1%
Frameset	20.778	31,2%
Client redirect	3.401	5,1%
Webpagina bevat < 10 woorden	4.987	7,5%
Webpagina onleesbaar	75	0,1%
Operation timed out	558	0,8%
Webpagina bestaat niet (404 Error)	479	0,7%
Connection time-out	477	0,7%
Meer dan 2 server redirects	204	0,3%
Overige	230	0,3%
Totaal	66.572	100%

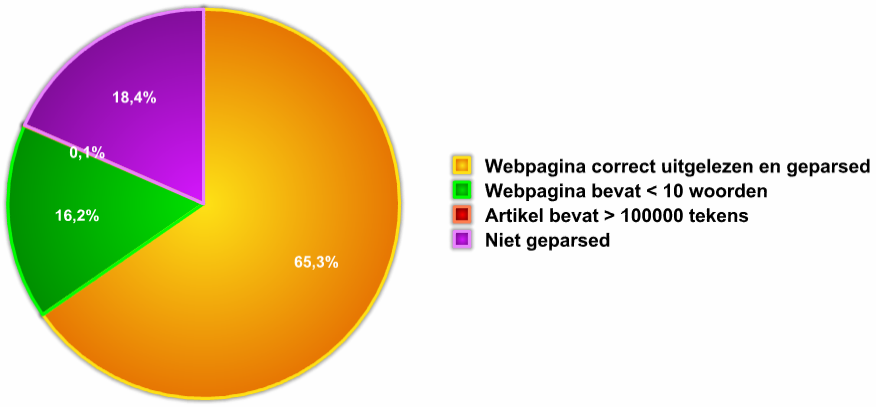
Tabel 5: Verwerking Dmoz corpus (Nederlands)

De woorden van de 35.383 bruikbare webpagina's zijn verwerkt naar een frequentiemodel zoals in figuur 3 op pagina 17, waarbij het totaal aantal woorden uitkwam op 13.880.707. Veel woorden komen meerdere keren op verschillende webpagina's voor waardoor uiteindelijk 593.994 unieke woorden overbleven waarvan de woord- en documentfrequentie bekend was.

3.1.3 Wikipedia corpora

Engels

De vijfde corpus bevat woorden uit 1.036.437 Engelse Wikipedia artikelen. Deze artikelen zijn verwerkt waarna een lijst met 2.744.680 unieke woorden ontstond waarvan de woord- en documentfrequenties bekend was. Om het corpus van Wikipedia op te bouwen is als eerste via de website van Wikipedia [15] een zogenaamde dumpfile gedownload. Deze dumpfile wordt in het xml-formaat aangeleverd en is voor iedereen beschikbaar onder de GNU Free Documentation License. Nadat de dumpfile gedownload was, is deze geïmporteerd in een MySQL-database met behulp van de tool xml2sql [16]. Hierna zijn de woord- en documentfrequentie van alle voorkomende woorden in de artikelen met behulp van een PHP-script geteld.



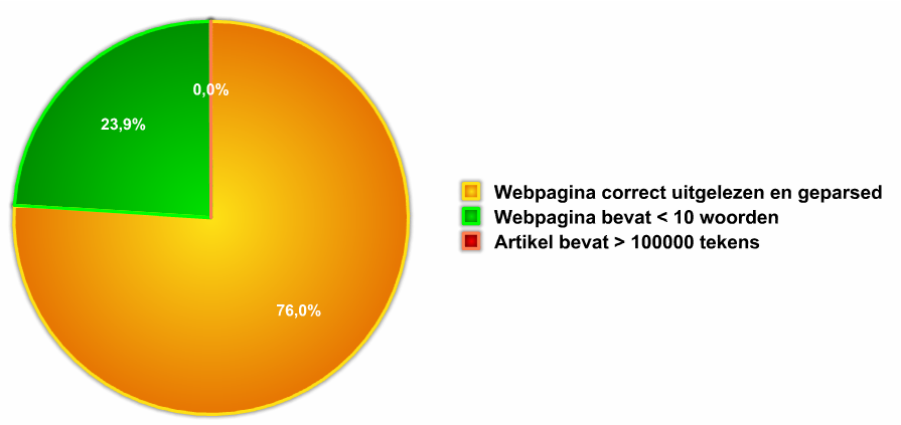
	Aantal	Percentage
Artikelen correct uitgelezen en verwerkt	1.036.437	65,3%
Artikel bevat < 10 woorden	257.098	16,2%
Artikel bevat > 100000 tekens	1.099	0,1%
Subtotaal	1.294.634	81,6%
Niet verwerkt	291.375	18,4%
Totaal	1.586.009	100%

Tabel 6: Verwerking Wikipedia corpus (Engels)

In totaal waren er 1.586.009 artikelen, maar een gedeelte (291.375 artikelen) zijn niet verwerkt vanwege tijdgebrek. 257.098 artikelen zijn overgeslagen omdat deze minder dan 10 woorden bevatten; dit waren meestal synoniemen die verwezen naar andere artikelen en dus niet daadwerkelijk het artikel zelf. 1099 Artikelen waren te groot om te verwerken.

Nederlands

De zesde corpus bevat woorden uit 163.445 Nederlandse Wikipedia artikelen. Ook van deze corpus zijn de woord- en documentfrequenties berekend. 51.581 artikelen zijn uitgesloten van deelname omdat deze kleiner waren dan 10 woorden; het ging dan meestal om doorverwijspagina's. Wikipedia bevat bijvoorbeeld voor het woord "zeeland" meerdere artikelen. Omdat het woord "zeeland" dus meerdere betekenissen heeft, gebruikt Wikipedia doorverwijspagina's waarin de verschillende mogelijke betekenissen staan vermeld. Uiteindelijk ontstond een lijst met 956.365 unieke woorden waarvan de woord- en documentfrequentie bekend was.



	Aantal	Percentage
Artikelen correct uitgelezen en verwerkt	163.445	76,01%
Artikel bevat < 10 woorden	51.495	23,95%
Artikel bevat > 100000 tekens	86	0,04%
	215.026	100%

Tabel 7: Verwerking Wikipedia corpus (Nederlands)

3.2 Overlap tussen de zoekmachines bepalen

Om de grootte van het World Wide Web te bepalen met behulp van de vier grootste zoekmachines, is het noodzakelijk om de overlap tussen de zoekmachines vast te stellen. Aangezien de deelnemende zoekmachines een behoorlijke omvang hebben is de kans reëel dat zoekmachines voor een groot gedeelte dezelfde url's geïndexeerd hebben. Door middel van drie steekproeven met 31.153 url's, 51.600 url's en 52.096 url's is paarsgewijs en kruislings gecontroleerd hoe groot de overlap tussen de zoekmachines is, zowel voor de Nederlandse corpora als voor de Engelse corpora. Een PHP-script heeft deze taak uitgevoerd.

3.2.1 Steekproef 1 - Zipf-woordselectie - 31.153 url's

De 31.153 url's zijn verkregen via de volgende methode. Van ieder corpus zijn logaritmisch log(1.07) woorden (van hoogfrequente naar laagfrequente woorden) geselecteerd. Log(2) levert te weinig woorden op, waardoor er minder url's voor deze steekproef beschikbaar zouden zijn.

Log(1.07) resulteerde in een lijst met 784 woorden voor de zes corpora gezamenlijk. Deze 784 woorden zijn verzonden naar elk van de vier zoekmachines, waarbij de top 10 resultaten van url's opgeslagen werden. Zo ontstond een lijst met 4 * 784 * 10 = 31.360 url's.

ID	Keyword	Positie	Zoekmachine	Url	Taal	Google	Yahoo	Msn	Ask	Datum
1	the	1	Google	http://www.theonion.com/	ENG	1	1	1	1	2006-04-19 17:35:21
2	the	2	Google	http://www.whitehouse.gov/	ENG	1	1	1	1	2006-04-26 16:36:49
3	the	3	Google	http://www.weather.com/	ENG	1	1	1	1	2006-04-28 12:50:39
4	the	4	Google	http://www.economist.com/	ENG	1	1	1	1	2006-05-02 11:15:03
5	the	5	Google	http://www.guardian.co.uk/	ENG	1	1	1	1	2006-05-06 22:58:50
6	the	6	Google	http://www.alltheweb.com/	ENG	1	1	1	1	2006-05-08 14:01:10
7	the	7	Google	http://www.gimp.org/	ENG	1	1	1	1	2006-05-09 10:21:04
8	the	8	Google	http://www.telegraph.co.uk/	ENG	1	1	1	1	2006-05-10 12:13:48
9	the	9	Google	http://www.theregister.co.uk/	ENG	1	1	1	1	2006-05-10 20:01:08
10	the	10	Google	http://www.independent.co.uk/	ENG	1	1	1	1	2006-05-11 14:30:43
11	of	1	Google	http://www.loc.gov/index.html	ENG	1	1	1	1	2006-04-19 17:35:22
12	of	2	Google	http://www.bankofamerica.com/	ENG	1	1	1	1	2006-04-26 16:36:54
13	of	3	Google	http://www.house.gov/	ENG	1	1	1	1	2006-04-28 12:50:43
14	of	4	Google	http://www.yourdictionary.com/	ENG	1	1	1	1	2006-05-02 11:15:07
15	of	5	Google	http://www.ed.gov/	ENG	1	1	1	1	2006-05-06 22:58:53
16	of	6	Google	http://www.nih.gov/	ENG	1	1	1	1	2006-05-08 14:01:14
17	of	7	Google	http://www.usdoj.gov/	ENG	1	1	1	1	2006-05-09 10:21:06
18	of	8	Google	http://www.usda.gov/	ENG	1	1	1	1	2006-05-10 12:13:48
19	of	9	Google	http://www.fbi.gov/	ENG	1	1	1	1	2006-05-10 20:01:08
20	of	10	Google	http://chronicle.com/	ENG	1	1	1	1	2006-05-11 14:30:43
21	and	1	Google	http://www.nasa.gov/	ENG	1	1	1	1	2006-04-19 17:35:25
22	and	2	Google	http://www.barnesandnoble.com/	ENG	1	1	1	1	2006-04-26 16:36:57

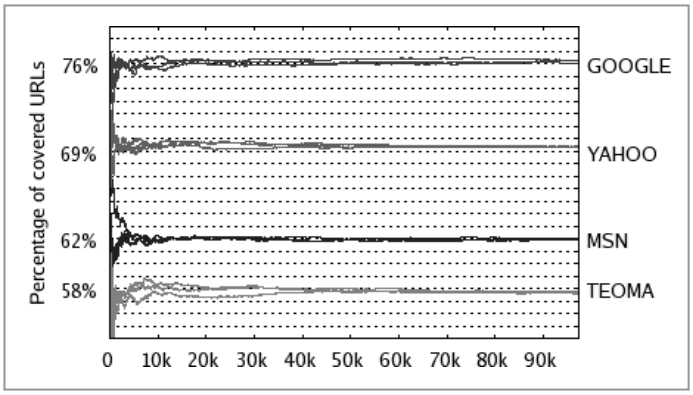
Figuur 5: Gedeeltelijke weergave van de overlapdatabase

Vooral de zoekmachine Msn Search gaf niet altijd een lijst met 10 resultaten terug, dit kwam doordat er gesponsorde links (dit zijn eveneens url's) tussen de resultaten vermeld stonden. Deze gesponsorde links zijn niet meegenomen in de uiteindelijke lijst. Opdrachtgevers van gesponsorde links betalen namelijk om bovenaan in de lijst te komen staan; zou men niet betalen dan zou men ook niet bovenaan de webpagina verschijnen. Daarnaast zijn het geen gecrawlde webpagina's, het is dus goed mogelijk dat de adverteerder alleen bij Google adverteert en niet bij Msn Search, Yahoo! of Ask. Doordat de gesponsorde links niet meegenomen zijn is de lijst met url's iets naar beneden bijgesteld tot 31.153 url's. Doordat alleen de eerste 10 url's van de geretourneerde lijst gecontroleerd zijn zal de overlap groter zijn dan in werkelijkheid. Zoekmachines trachten namelijk met allerlei technieken, bijvoorbeeld de PageRank methode van Google [33], om de belangrijkste

webpagina's bovenaan in de lijst te plaatsen. De andere drie zoekmachines gebruiken waarschijnlijk soortgelijke technieken. Deze technieken behoren tot de best bewaarde bedrijfsgeheimen van de verschillende zoekmachines. Aannemende dat de 'ranking'-methoden van de verschillende zoekmachines ongeveer hetzelfde werken, mag verwacht worden dat de kans op overlap groter is naarmate url's hoger in de lijst staan. Het is belangrijk om vast te stellen dat er op deze manier een ondergrens-schatting van de overlap wordt bepaald.

Wanneer willekeurig url's geselecteerd zouden worden uit alle url's die een zoekmachine retourneert, dan kan wel een schatting gegeven worden, maar of deze onder- of overschat is kan dan niet achterhaald worden. In het onderzoek van Gulli en Signorini [25] worden willekeurig één of meerdere url's uit de eerste 100 resultaten geselecteerd. Hierdoor zal de overlap tussen de zoekmachines groter zijn dan in werkelijkheid. In deze scriptie worden alleen de eerste 10 url's meegenomen, hierdoor wordt nog meer de bovengrens van de overlap benaderd.

Gulli en Signorini hebben van 486.752 url's de bekendheid gecontroleerd bij vier zoekmachines. Na 20.000 tot 40.000 url's stabiliseerde de dekkingsgraad voor alle zoekmachines, zie figuur 6. Aan de hand van de figuur is af te lezen dat de schommeling vanaf 20.000 url's maximaal 1% naar boven of beneden is. De aanname die zij maken is dat de percentages van de dekkingsgraad voor alle zoekmachines na 20.000 url's geen significante veranderingen vertonen. Helaas gaat deze aanname niet op voor de dekkingsgraad in steekproef 1 en 2 van deze scriptie. De mogelijke verklaring hiervoor word gegeven in de volgende sectie 3.2.2, Steekproef 2.



Figuur 6:
De dekkingsgraad voor elk van de vier zoekmachines wordt nauwkeuriger bepaald naarmate er meer url's gecontroleerd zijn.
Overgenomen uit Gulli en Signorini, 2005

3.2.2 Steekproef 2 - Willekeurige woordselectie - 51.600 url's

Deze steekproef lijkt sterk op steekproef 1 maar verschilt op enkele punten. In steekproef 1 worden uit ieder corpus *logaritmisch* woorden geselecteerd. In deze steekproef worden *willekeurig* 225 woorden uit ieder corpus geselecteerd met een documentfrequentie > 10. Wanneer de documentfrequentie van een woord binnen een corpus 10 of meer is, dan kan verwacht worden dat elke zoekmachine ook minstens 10 of meer url's voor een woord kan vinden, de corpora zijn immers vele malen kleiner dan de vier deelnemende zoekmachines.

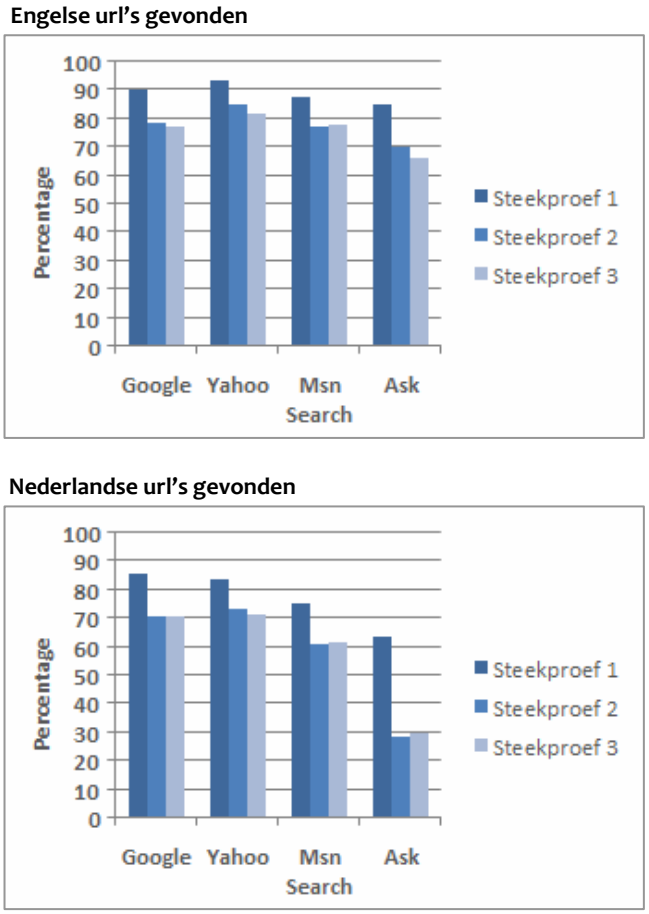
Om de lijst met 51.600 url's te verkrijgen zijn 225 willekeurige woorden per corpus naar de vier deelnemende zoekmachines verzonden. Van de lijst met resultaten die de zoekmachines retourneerden zijn de eerste 10 url's opgeslagen. Uiteindelijk leverde deze actie 54.000 url's op (225 woorden * 6 corpora * 4 zoekmachines * 10 url's). Soms werden er niet 10, maar 9 of 8 url's door een zoekmachine getoond. De overige url's waren gesponsorde links, evenals in steekproef 1. Wanneer de url's gesponsorde links bevatte zijn de url's voor dat desbetreffende woord niet meegenomen in de steekproef. Hierdoor bleven er uiteindelijk 51.600 url's over waarmee de steekproef is uitgevoerd.

3.2.3 Steekproef 3 - Willekeurige woordselectie - 52.096 url's

Steekproef 3 is identiek aan steekproef 2, alleen zijn opnieuw willekeurige woorden verzonden naar de vier zoekmachines waarbij telkens de eerste 10 url's in een database opgeslagen zijn om uiteindelijk een lijst met 52.096 url's te vormen.

3.2.4 Dekkingsgraad url's van andere zoekmachines

In figuur 7 word de dekkingsgraad van de drie steekproeven getoond. Opvallend is het duidelijke verschil tussen steekproef 1 en steekproef 2 en 3. In steekproef 2 en 3 zijn beduidend minder url's door de vier zoekmachines gevonden dan in steekproef 1. De reden van dit verschil is hoogstwaarschijnlijk toe te schrijven aan de verschillende methode van woordselectie.



Figuur 7: Dekkingsgraad zoekmachines (Visueel)

Doordat in steekproef 1 veel hoogfrequente woorden geselecteerd zijn (logaritmische selectie), is de kans ook groter dat zoekmachines url's retourneren van bekende webpagina's die deze hoogfrequente woorden bevatten. Deze bekende webpagina's zullen eerder in meerdere zoekmachines opgenomen zijn dan minder bekende webpagina's. De 'ranking'-methode van een zoekmachine zorgt ervoor dat de belangrijkste/bekende webpagina's bovenaan komen staan. Wanneer een hoogfrequent woord, bijvoorbeeld "microsoft" ingevoerd wordt, dan zullen één of meerdere

webpagina's van www.microsoft.com/nl hoogstwaarschijnlijk bij de eerste 10 resultaten vermeld staan. Dit komt omdat de Page ranking van deze webpagina's beduidend hoger liggen dan een andere webpagina waar misschien nog wel veel vaker het woord "microsoft" in voorkomt. De vier zoekmachines die in deze scriptie onderzocht worden gebruiken 'ranking'-algoritmes om belangrijke en bekende webpagina's bovenaan de lijst met resultaten te plaatsen. Hierdoor is de kans groot dat frequent voorkomende woorden ook url's opleveren welke bij alle zoekmachines bekend zijn.

Engelse url's gevonden

	Google	**Yahoo!**	**Msn Search**	**Ask**
Steekproef 1	90,00%	93,73%	87,88%	84,97%
Steekproef 2	78,89%	85,02%	77,06%	69,99%
Steekproef 3	77,08%	81,59%	78,03%	66,29%

Tabel 8a: Dekkingsgraad steekproef 1,2 en 3

Nederlandse url's gevonden

	Google	**Yahoo!**	**Msn Search**	**Ask**
Steekproef 1	85,59%	83,75%	75,62%	63,52%
Steekproef 2	70,54%	73,54%	61,36%	28,73%
Steekproef 3	70,71%	71,35%	61,50%	29,67%

Tabel 8b: Dekkingsgraad steekproef 1,2 en 3

In steekproef 2 en 3 zijn 225 woorden willekeurig per corpus geselecteerd met een documentfrequentie die groter is dan 10. Het Dmoz corpus bevatte bijvoorbeeld 329.375 woorden met een documentfrequentie > 10. Uit deze selectie van 329.375 woorden zijn 225 woorden willekeurig geselecteerd. Een klein deel van de 329.375 woorden zijn vaak voorkomend, een veel groter deel is dat niet (Zipf's law) (zie sectie 3.3). Om deze reden bevatten steekproef 2 en 3 weinig hoogfrequente woorden. Hoe minder frequent een woord voorkomt, des te groter is de kans dat een relatief onbekende webpagina bovenaan in de lijst (bij de eerste 10 resultaten) komt te staan. Voor hoogfrequente woorden heeft een webpagina een hogere ranking nodig dan voor laagfrequente woorden om bij de eerste 10 resultaten te eindigen. Daarnaast is de kans dat een minder bekende / onbekende webpagina voorkomt in meerdere zoekmachines kleiner dan voor een bekende webpagina. De voorgaande punten zouden mogelijkerwijs de beduidend lagere dekkingsgraad in steekproef 2 en 3 ten opzichte van steekproef 1 kunnen verklaren.

Tabel 8a en 8b tonen de herkenningspercentages van url's van de verschillende zoekmachines. Ask scoort het minst in alle steekproeven, zowel voor Engelse als voor Nederlandse url's. Van alle Nederlandse url's herkent Ask in steekproef 2 en 3 maar 28,73% en 29,67%. Deze percentages liggen een stuk lager in vergelijking met de andere zoekmachines. Hetzelfde geldt in mindere mate ook voor Engelse url's. Uit deze cijfers mag je dan ook afleiden dat Ask de minste url's herkent en daardoor ook de minste webpagina's indexeert. Deze zoekmachine kan dan ook gekwalificeerd worden als de kleinste van de onderzochte zoekmachines. De achterstand van Msn Search op Yahoo! en Google is minder groot dan voor Ask. Voor Engelse url's komt het herkenningspercentage in de buurt van Google, met Nederlandse url's heeft de zoekmachine beduidend meer moeite.

Erg opvallend is de hoge url herkenning door Yahoo!. De metingen in sectie 4 laten namelijk zien dat Google voor Engelse en Nederlandse url's de grootste zoekmachine is, althans dat suggereren de documentfrequenties die Google retourneert (zie ook bijlage 3). De documentfrequenties liggen bij Google voor vrijwel ieder woord hoger dan voor de overige zoekmachines, waardoor je zou verwachten dat Google ook de meeste url's zou herkennen, maar in alle drie de paarsgewijze overlap-steekproeven herkende Yahoo! de meeste Engelse url's en in 2 van de 3 steekproeven de meeste Nederlandse url's.

Een mogelijke verklaring zou kunnen zijn dat in deze drie steekproeven telkens de eerste 10 url's gebruikt zijn. De eerste 10 url's die een zoekmachine terug geeft zijn over het algemeen bekende / regelmatig voorkomende url's op het web. Dit komt onder andere doordat er veel naar deze url's verwezen wordt vanuit andere webpagina's. De kans dat een bekende url in een zoekmachine opgenomen wordt is allicht groter dan voor een minder bekende url. Wanneer url's uit de top 100 of top 1000 geselecteerd zouden worden zullen er meer url's tussen zitten die minder bekend zijn en niet bij alle zoekmachines geregistreerd staan. Het is mogelijk dat Google meer url's bevat die minder bekend zijn dan Yahoo!. Dit zou een mogelijke verklaring zijn dat de dekkingsgraad van Yahoo! groter is dan Google, terwijl dit feitelijk niet het geval is. In een vervolgonderzoek zou opnieuw de overlap bepaald moeten worden, alleen dan met een groter bereik, bijvoorbeeld willekeurig 10 url's uit de eerste 500 resultaten. Een nadeel van deze methode is dat de geschatte overlap lager zal zijn dan werkelijkheid het geval is. Een andere mogelijke verklaring is dat Google de aantallen kunstmatig hoog houdt. Google doet zich hierdoor groter voor dan dat ze in werkelijkheid is (zie sectie 2.5.1).

3.2.5 Paarsgewijze overlap

De tabellen 9 + 10 op de volgende pagina geven paarsgewijs de niet overlappende url's per zoekmachine paar weer. Met andere woorden; het aantal url's (percentage) van zoekmachine X die niet voorkomen in zoekmachine Y^1 ,Y^2, Y^3 en vice versa. De percentages in steekproef 1 liggen beduidend lager dan in steekproef 2 + 3, dit vanwege de andere benadering van woordselectie, zie sectie 3.2.4. Steekproef 2 en 3 zijn methodisch identiek uitgevoerd. De percentages van steekproef 2 en 3 liggen, zoals verwacht, erg dicht bij elkaar. Alleen de paarsgewijze overlap van Msn Search met Google, Yahoo! en Ask wijken erg af, het is onduidelijk waarom dit het geval is.

Voorbeeld: Yahoo! bevat 6,21% van Google's url's niet (steekproef 1).
Msn Search! bevat 9,18% van Google's url's niet (steekproef 1).
Ask bevat 13,92% van Google's url's niet (steekproef 1).

Percentages niet overlappende Engelse url's

Url van	Url niet in	Steekproef 1 *n = 16.546*	Steekproef 2 *n = 25.590*	Steekproef 3 *n = 26.016*
Google	**Y**	6,21%	19,13%	18,81%
	M	9,18%	21,40%	18,96%
	A	13,92%	33,53%	34,06%
Yahoo!	**G**	11,42%	24,64%	24,63%
	M	13,45%	24,44%	23,92%
	A	14,31%	29,97%	31,63%
Msn	**G**	8,73%	16,24%	22,49%
Search	**Y**	5,03%	10,69%	20,10%
	A	13,07%	26%	35,57%
Ask	**G**	8,24%	21,85%	21,60%
	Y	5,77%	14,53%	16,46%
	M	10,94%	23,01%	23,03%

Tabel 9: Paarsgewijze overlap niet herkende Engelse url's

Percentage niet overlappende Nederlandse url's

Url van	Url niet in	Steekproef 1 *n = 14.607*	Steekproef 2 *n = 26.010*	Steekproef 3 *n = 26.080*
Google	**Y**	17,62%	30,23%	32,87%
	M	20,51%	34,04%	32,99%
	A	34,53%	72,63%	71,80%
Yahoo!	**G**	15,86%	31,38%	31,69%
	M	23,95%	40,16%	40,82%
	A	28,88%	68,65%	67,26%
Msn	**G**	10,89%	22,93%	22,44%
Search	**Y**	13,91%	23,02%	25,92%
	A	37,26%	72,66%	72,08%
Ask	**G**	12,13%	33,51%	33,18%
	Y	12,55%	25,82%	26,94%
	M	21,17%	41,72%	41,68%

Tabel 10: Paarsgewijze overlap niet herkende Nederlandse url's

3.2.6 Kruislingse overlap

Overlap kan op verschillende manieren berekend worden. Een paarsgewijze overlapschatting is alleen niet toereikend om de grootte van het geïndexeerde WWW te bepalen. Door vast te stellen hoeveel procent van de webpagina's kruislings bij elkaar in de database zitten kan wel de grootte van het geïndexeerde WWW geschat worden. Het grootste probleem is dat er meerdere combinaties zijn om de kruislingse overlap te schatten. Een webpagina kan bijvoorbeeld opgenomen zijn in Google en Ask, maar niet in Yahoo! en Msn Search. Een andere webpagina is misschien alleen opgenomen door Yahoo! en niet door de overige zoekmachines. De kruislingse overlap kan op verschillende manieren berekend worden. De meest logische stap is om te

sorteren op grootte; beginnend met de grootste zoekmachine. De reden voor deze strategie is dat het geïndexeerde World Wide Web minstens zo groot is als de grootste zoekmachine. Op 1 juni 2006 zijn de gemiddelde documentfrequenties per zoekmachine uit 154 queries bepaald. Google retourneerde gemiddeld de meeste documentfrequenties gevolgd door Yahoo!, Msn Search en Ask. Om de kruislingse overlap te schatten zijn dezelfde url's gebruikt als de paarsgewijze overlap. Tabel 11 a toont de percentages voor de steekproeven 1 t/m 3.

Y=1 G = 0	=	Url's wel in Yahoo!, niet in Google
M=1 G = 0 Y=0	=	Url's wel in Msn Search, niet in Google, niet in Yahoo!
A=1 G = 0 Y=0 M = 0	=	Url's wel in Ask, niet in Google, niet in Yahoo!, niet in Msn Search

Google, Yahoo!, Msn Search, Ask

	Y=1 G=0			M=1 G=0 Y=0			A=1 G=0 Y=0 M=0		
Steekproef	1	2	3	1	2	3	1	2	3
Dmoz corpus (VT) *n = 8500*	5,60%	12,22%	12,88%	0,78%	1,95%	2,22%	0,35%	2,19%	1,62%
Krantencorpus (ENG) *n = 8440*	6,03%	12,19%	13,64%	0,47%	1,65%	2,82%	0,47%	2,07%	2,43%
Wiki corpus (ENG) *n = 8650*	6,39%	11,39%	11,48%	0,77%	1,64%	2,70%	0,53%	1,56%	1,38%
Dmoz corpus (NL) *n = 8750*	5,10%	11,94%	11,76%	1,11%	2,65%	2,27%	0,74%	2,81%	2,45%
Krantencorpus (NL) *n = 8580*	7,38%	17,27%	17,06%	1,69%	3,15%	3,97%	1,64%	4,88%	3,87%
Wiki corpus (NL) *n = 8680*	7,60%	15,13%	15,96%	1,92%	3,25%	3,65%	1,75%	4,75%	4,54%

Tabel 11 a: Kruislingse overlap niet herkende Nederlandse en Engelse url's

In sectie 3.2.4 kwam naar voren dat Yahoo! de grootste dekkingsgraad had. Van de 6 metingen herkende Yahoo! 5 maal de meeste url's. Dit suggereert dat Yahoo! de grootste zoekmachine zou kunnen zijn. Vandaar dat Google en Yahoo! omgewisseld zijn en de volgorde in de onderstaande tabel als volgt is: Yahoo!, Google, Msn Search en Ask.

Yahoo!, Google, Msn Search, Ask

	G=1 Y=0			M=1 G=0 Y=0			A=1 G=0 Y=0 M=0		
Steekproef	1	2	3	1	2	3	1	2	3
Dmoz corpus (VT) *n = 8500*	3,30%	8,33%	9,51%	0,78%	1,95%	2,22%	0,35%	2,19%	1,62%
Krantencorpus (ENG) *n = 8440*	2,74%	7,75%	8,83%	0,47%	1,65%	2,82%	0,47%	2,07%	2,43%
Wiki corpus (ENG) *n = 8650*	3,29%	6,27%	9,68%	0,77%	1,64%	2,70%	0,53%	1,56%	1,38%
Dmoz corpus (NL) *n = 8750*	6,01%	10,24%	12,14%	1,11%	2,65%	2,27%	0,74%	2,81%	2,45%
Krantencorpus (NL) *n = 8580*	9,14%	14,35%	16,30%	1,69%	3,15%	3,97%	1,64%	4,88%	3,87%
Wiki corpus (NL) *n = 8680*	8,78%	12,96%	14,82%	1,92%	3,25%	3,65%	1,75%	4,75%	4,54%

Tabel 11 b: Kruislingse overlap niet herkende Nederlandse en Engelse url's

3.3 Zipf's law

Veel natuurlijk voorkomende zaken zoals de grootte van een stad, inkomen en woordfrequentie zijn verdeeld volgens een Zipf-distributie. Deze logaritmische verdeling impliceert dat grote voorkomens uiterst zeldzaam zijn, terwijl kleine voorkomens vaak voorkomen. Deze regelmatigheid of 'wet' wordt ook wel Zipf's law en soms Pareto law genoemd [31] [32]. Van origine stelt Zipf's law dat de log(frequentie-ranking) van woorden binnen een tekst of corpus een lineair verband oplevert. In een corpus van natuurlijke talen is de frequentie van een woord grofweg omgekeerd evenredig aan zijn positie in een frequentietabel. Zo zal het frequentste woord ongeveer tweemaal zo vaak voorkomen als het tweede frequentste woord, dat tweemaal zo vaak voorkomt als het vierde frequentste woord [23]. In alle metingen wordt dit natuurlijke verband gebruikt om een representatieve selectie van woorden uit elk van de zes corpora te verkrijgen. Door de woorden logaritmisch uit een corpus te extraheren blijft de natuurlijke woordverdeling intact. Van ieder corpus zijn alle $n^{log(1,6)ste}$ woorden geselecteerd waardoor, afhankelijk van de grootte van het corpus, een lijst van circa 25 tot 30 woorden ontstond. De logaritmische extractie van woorden zorgt ervoor dat vaak voorkomende woorden binnen een corpus ook zwaarder meewegen in de metingen dan minder vaak voorkomende woorden. Hieronder staat simplistisch de Zipf-woordselectie uitgewerkt voor de zes verschillende corpora. Voor het rekengemak is $n^{log(2)}$ gebruikt (voorbeeld 1), voor de daadwerkelijk Zipf-woordselectie in de metingen is $n^{log(1,6)}$ gebruikt (voorbeeld 2). De reden dat in de metingen $n^{log(1,6)}$ gebruikt is komt doordat met $n^{log(1,6)}$ meer woorden geselecteerd kunnen worden dan met $n^{log(2)}$. Een getal lager dan 1,6 zou dubbele posities opleveren.

3.3.1 Zipf-woordselectie

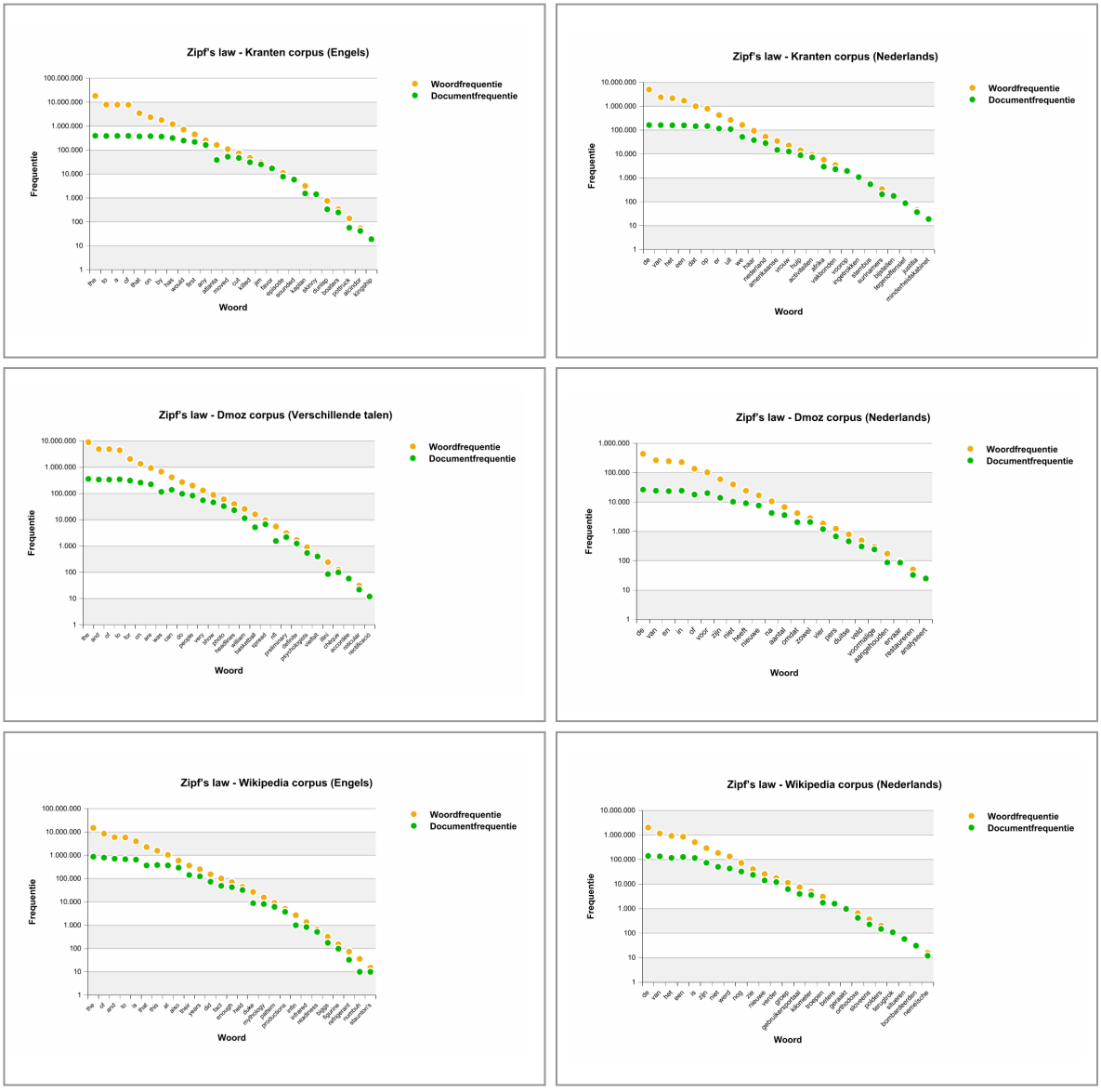
Als eerste wordt een lijst met woorden, document- en woordfrequenties opgebouwd uit een corpus X. Neem aan dat van 10.000 woorden (komende uit corpus X) de document- en woordfrequenties geïnventariseerd zijn. Deze 10.000 woorden worden gesorteerd op documentfrequentie, van frequent voorkomende woorden naar minder frequent voorkomende woorden. Uiteindelijk worden alle $n^{log(2)ste}$ woorden (uit de gesorteerde lijst met 10.000 woorden) geselecteerd welke uiteindelijk de volgende posities opleveren:

- Voorbeeld 1 – log(2)
  1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192.
- Voorbeeld 2 – log(1,6)
  1, 2, 3, 4, 7, 10, 17, 27, 43, 69, 110, 176, 281, 450, 721, 1153, 1845, 2951, 4722, 7556.

Met andere woorden; nadat de lijst met 10.000 woorden gesorteerd is van vaak voorkomende woorden naar minder vaak voorkomende woorden wordt het eerste, twee, vierde, achtste etc. woord geselecteerd (voorbeeld 1) en in een nieuwe lijst geplaatst met de daarbij behorende woord- en documentfrequentie. Voor elke corporus is een Zipf-woordenlijst gegenereerd. In de metingen worden deze lijsten gebruikt om de grootte van de zoekmachines te schatten (zie sectie 4).

3.3.2 Zipf-verdeling per corpus

Hieronder wordt de Zipf-verdeling per corpus getoond. De Y-as geeft logaritmisch de woord- en documentfrequenties weer. Op de X-as staan alle woorden uit de desbetreffende corpus waarbij telkens het log(1.6) ^{ste} woord weergegeven is. Elke gele stip representeert de woordfrequentie en elke groene stip de documentfrequentie van een woord.



Figuur 8: Woord- en documentfrequenties – Zipf-verdeling

De zes grafieken in figuur 8 laten duidelijk de Zipf-verdeling voor elke corpus zien. De woordfrequentie loopt in een vrijwel rechte lijn naar rechts af. De documentfrequentie buigt bij frequent voorkomende woorden af. Hoogfrequente woorden komen vaak meerdere keren in een artikel of webpagina voor, daardoor loopt de woordfrequentie in een rechte lijn op de logaritmische schaal terwijl de documentfrequentie afloopt bij hoogfrequente woorden. Laagfrequente woorden komen vaker éénmaal in een document voor dan hoogfrequente woorden, waardoor de woord- en documentfrequenties dicht bij elkaar liggen.

4. Resultaten

In deze sectie wordt de geschatte grootte van het Geïndexeerde World Wide Web bepaald in zes onafhankelijke metingen. Daarbij worden de zes corpora, geschatte kruislingse overlap, gegenereerde Zipf-woordenlijsten en het percentages dode links uit de voorgaande secties gebruikt.

De zes metingen vonden tussen 1 juni 2006 en 26 juni 2006 plaats. Elke meting maakt gebruik van een gegenereerde Zipf-woordenlijst. Deze lijsten bestaan uit 23 tot 28 woorden, afhankelijk van de grootte van het corpus.

De woorden uit de zes woordenlijsten zijn iedere ochtend om 8:00 uur verzonden naar de zoekmachines Google, Msn Search, Yahoo! en Ask. Dit resulteerde in 154 queries per dag. De documentfrequenties die de zoekmachines retourneerden zijn opgeslagen in een MySQL-database. Een PHP-script zorgde ervoor dat dit gehele proces automatisch verliep.

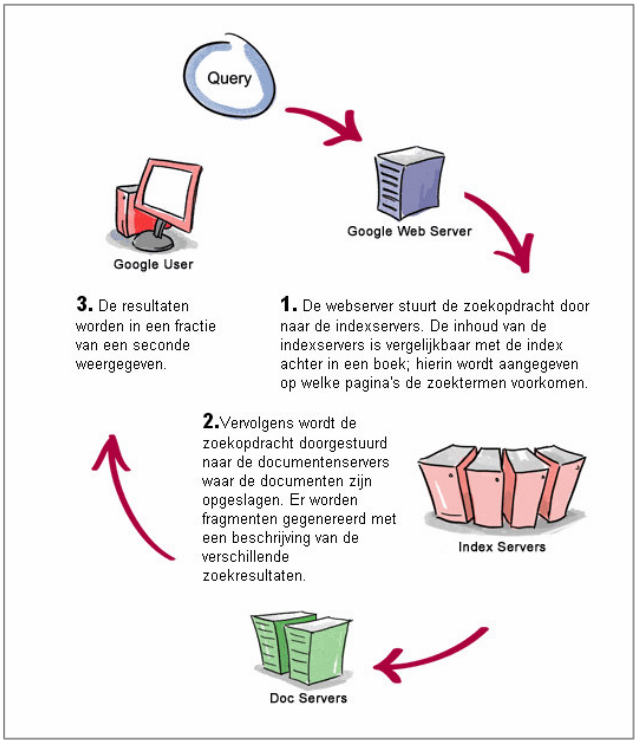
Alle zoekmachines boden de mogelijkheid om te filteren op taal. Deze mogelijkheid is gebruikt om het Nederlandse World Wide Web te bepalen. De filter zorgt ervoor om binnen een taaldomein te zoeken. Zo wordt het mogelijk om documentfrequenties voor Nederlandse woorden te bemachtigen. Alle zoekmachines boden de opties “pagina’s uit Nederland” en “pagina’s in het Nederlands”. De keuze viel hierbij op “pagina’s in het Nederlands”, omdat de corpora een taaldomein afbakenen en geen land.

Om een ondergrens-schatting van het totale World Wide Web te maken zijn geen filters opgelegd voor zowel de woorden van de Engelse corpora als de verschillende talen corpora. Dit heeft als nadeel dat een deel van het World Wide Web niet in de schatting zal worden meegenomen. Om dit overige deel te schatten zal voor iedere taal een corpus gegenereerd moeten worden. Het opleggen van geen filter heeft ook een groot voordeel; buitenlandse talen gebruiken ook Engelse woorden in een tekst, url of titel van een webpagina. Hierdoor worden een groot aantal niet

Engelse webpagina's mee genomen. Er worden dus meer dan alleen Engelse webpagina's meegenomen in de metingen.

Nadat het PHP-script gereed was met het uitlezen van de 154 documentfrequenties, werd met hetzelfde PHP-script de grootte per dag / per zoekmachine / per corpus geschat. Van ieder woord is de documentfrequentie in het corpus (die in bijlage III weergegeven wordt) gedeeld door de documentfrequentie van de zoekmachine om uiteindelijk vermenigvuldigd te worden met het totaal aantal woorden in het corpus (sectie 3.1). Door deze methode toe te passen wordt via ieder woord de grootte van de desbetreffende zoekmachine geschat. Zoekmachines bevatten ook dode links (sectie 2.5.3). Het geschatte percentage aan dode links is op elk van de zoekmachines in mindering gebracht.

In de periode dat de zes metingen plaatsvonden kwam al snel naar voren dat alle zoekmachines fluctuaties in de documentfrequenties vertoonden. Vooral Google schommelde met het aantal resultaten. Zo leverde het woord "the" bij Google op 1 juni 2006 17,56 miljard resultaten op, terwijl op 2 juni 2006 ineens 23,23 miljard resultaten geretourneerd werden. In 2001 heeft Wouter Mettrop [26] onderzoek verricht naar het fluctuatiegedrag van zoekmachines. Daarin kwam naar voren dat er niet één maar verschillende oorzaken aan dit fluctuatiegedrag ten grondslag liggen. Een van de verklaringen is dat zoekmachines verschillende indexen gebruiken. Voor een gebruiker lijkt het alsof een zoekmachine altijd dezelfde index raadpleegt, maar in werkelijkheid worden verschillende indexen geraadpleegd. Dit impliceert volgens Wouter Mettrop dat iedere index ("Index Servers" in afbeelding 9) regelmatig gerepliceerd moeten worden. Repliceren houdt in dat de wijzigingen die in de verschillende indexen gemaakt zijn onderling gesynchroniseerd worden. Na het repliceren zouden alle indexen identiek aan elkaar moeten zijn (meerdere replica's), ware het niet dat het repliceren zoveel tijd kost dat onderwijl het repliceren alweer nieuwe gegevens worden weggeschreven naar de verschillende indexen. Hierdoor zullen de indexen nooit 100% aan elkaar gelijk zijn.



Figuur 9: Levensloop Google-zoekopdracht
Overgenomen van Google.nl

Een andere verklaring volgens Mettrop is het feit dat verschillende delen van het World Wide Web verschillende indexen raadplegen. Technische tekortkomingen kunnen er dan voor zorgen dat een index tijdelijk niet bereikbaar is waardoor een zoekmachine onvolledige zoekresultaten toont. Fluctuaties kunnen daarnaast ook veroorzaakt worden door performance afhankelijke algoritmes, aldus Mettrop. Hoe meer queries afgehandeld worden, hoe minder deze algoritmes worden geactiveerd en hoe meer kans op fluctuaties. De laatste mogelijkheid die door Mettrop genoemd wordt zijn fluctuaties die veroorzaakt worden door het verwijderen van dubbele webpagina's. Zoekmachines met dubbele webpagina's maken een zoekmachine namelijk minder efficiënt. Jean Véronis [3] stelt daarentegen grote vraagtekens bij de sterke groei en fluctuaties van enkele zoekmachines, zie daarvoor sectie 2.5.1. Hoe dan ook; fluctuaties zijn een feit en aan de orde van de dag zoals in de zes metingen duidelijk zal blijken.

De zes metingen (2 per corpus) op de volgende pagina's zijn verdeeld in twee secties; de eerste sectie "Geschatte grootte per zoekmachine" en een tweede sectie "Geschatte grootte van het geïndexeerde World Wide Web". In de volgende alinea's volgt een korte beschrijving van deze twee secties.

De eerste sectie toont per corpus een grafiek met daarin gewogen (micro-average) en ongewogen (macro-average) geschatte groottes van de deelnemende zoekmachines. De ongewogen grootte per dag is geschat door de formule in sectie 3.1 toe te passen voor elk woord in de gegenereerde Zipf-woordenlijst. Daarna is de som van deze geschatte groottes gedeeld door het aantal woorden in de gegenereerde Zipf-woordenlijst.

De gewogen grootte per dag verschilt op één punt met de ongewogen grootte; de documentfrequentie van een woord bepaalt het aandeel in de geschatte grootte. Hoogfrequente woorden zullen hierdoor zwaarder meewegen dan laagfrequente woorden. De formule voor gewogen groottes voor een enkel woord is als volgt:

$$\frac{\text{documentfrequentie zoekmachine}}{\text{documentfrequentie corpus}} * \text{Tot. aantal docu. in corpus} * \text{documentfrequentie corpus}$$

Deze formule levert *per woord* de gewogen geschatte grootte van een zoekmachine op. Om voor elke zoekmachine de gewogen geschatte grootte te bepalen is de som van alle gewogen geschatte groottes per woord gedeeld door de som van de documentfrequenties van de gegenereerde Zipf-woordenlijst. Gewogen schattingen zorgen voor een stabieler beeld in de resultaten, het onderdrukt uitschieters naar boven en naar beneden. Met name laagfrequente woorden in een corpus zorgen voor grote schommelingen wanneer de geschatte grootte van een zoekmachine bepaald wordt. In een corpus kan het woord "hottentottententententoonstelling"

bijvoorbeeld oververtegenwoordigd zijn wanneer dit woord in het corpus 5 maal voorkomt en in een zoekmachine 25 maal, waardoor de geschatte grootte van het geïndexeerde World Wide Web vele malen groter uitvalt voor dit woord dan voor de andere woorden in het corpus. Om deze gevallen te dempen worden de geschatte groottes gewogen.

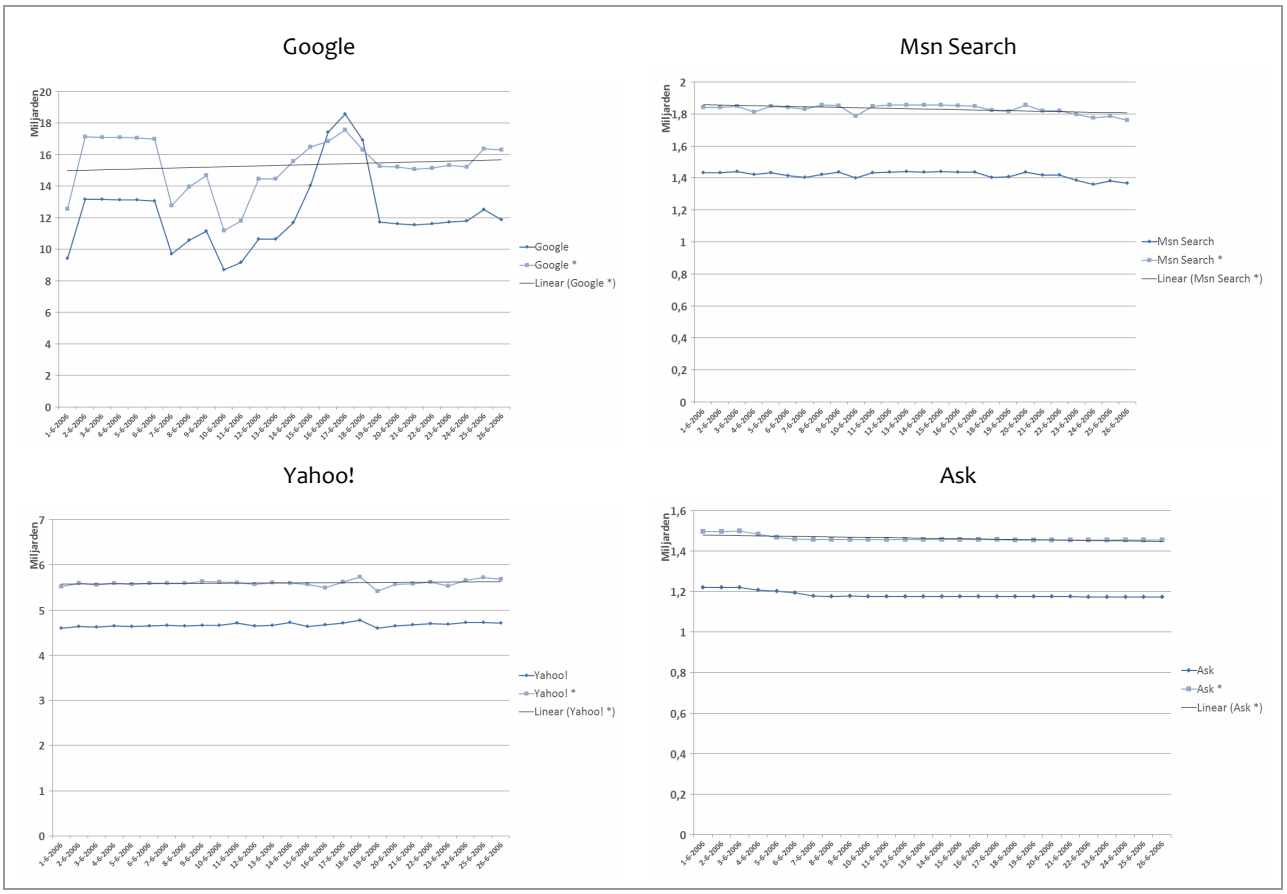
Ook tonen de grafieken in deze sectie een trendlijn, waarbij gebruik gemaakt is van lineaire regressie. Lineaire regressie is verkozen boven alternatieven (waaronder exponentiele regressie) omdat het aantal meetwaarden klein is, en de meetperiode kort (een maand). Ook al groeit het web licht exponentieel, de korte-termijngroei kan het beste benaderd worden met een lineaire regressie.

De tweede sectie "Geschatte grootte van het geïndexeerde World Wide Web" geeft zes ondergrens-schattingen van de grootte van het geïndexeerde World Wide Web. Om deze groottes te schatten zijn de geschatte groottes per zoekmachine gebruikt en is de kruislingse overlap tussen de zoekmachines bepaald (zie sectie 3.2.6). De grafiek in deze sectie gebruikt verschillende afkortingen, zij hebben de volgende betekenis; het eerste deel (OL1, OL2 of OL3) geeft de steekproef aan (OL staat voor OverLap). Het tweede gedeelte geeft de volgorde aan. GYMA betekent bijvoorbeeld Google, Yahoo!, Msn Search en Ask. De reden dat twee verschillende volgordes gebruikt zijn is omdat Google in de volgende secties ontzettend fluctueert in grootte, waardoor het een onbetrouwbare factor kan zijn om de grootte van het geïndexeerde World Wide Web te bepalen. Vandaar dat de grootte ook bepaald wordt door Yahoo! als uitgangspunt te nemen (volgorde YGMA).

4.1 Metingen I t/m VI afzonderlijk

4.1.1 Krantencorpus (Engels) - Geschatte grootte per zoekmachine

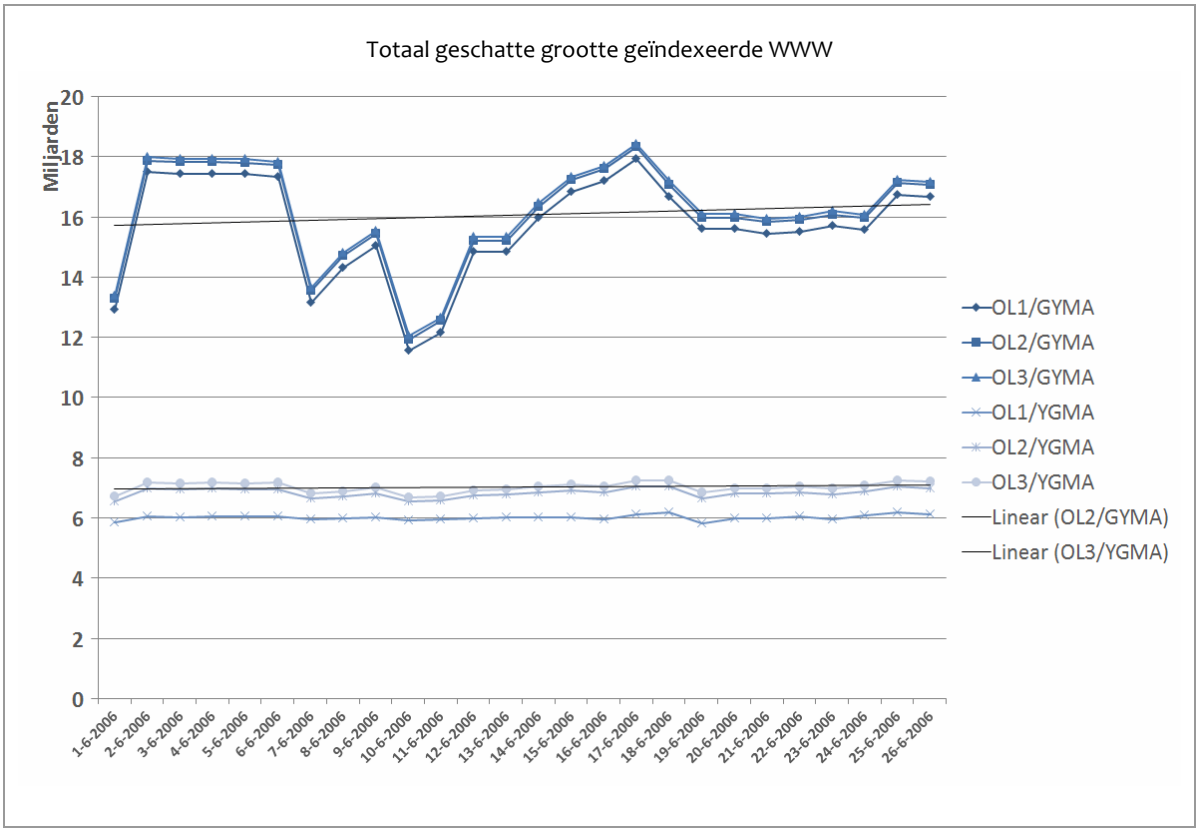
Deze meting is uitgevoerd met het krantencorpus (Engels). Wat direct opvalt zijn de grote fluctuaties bij Google in de maand juni 2006. De gewogen metingen bevinden zich tussen de 11 en 18 miljard webpagina's. Daarnaast zijn ook redelijke fluctuaties bij Yahoo! zichtbaar, deze fluctuaties liggen tussen de 5,5 en 5,8 miljard. De overige twee zoekmachines fluctueren in mindere mate (< 0,1 miljard). Ask levert de meest constante waarden in deze meting. Tijdens de testfase[9] in de maanden maart, april en mei 2006 werden ook al fluctuaties opgemerkt, ook Wouter Mettrop [26] en Jean Véronis [3] hebben dit fluctuatiegedrag in 2001 en 2005 vastgesteld (zie sectie 2.5.1 & sectie 4).



Figuur 10: Meting 1 – Engels - Geschatte grootte per zoekmachine.
*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**
Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

[9] De fase waarin de PHP-scripts voor deze metingen ontwikkeld werden

4.1.2 Krantencorpus (Engels) - Geschatte grootte van het geïndexeerde World Wide Web



Figuur 11: Meting 1 - Engels - Geschatte grootte World Wide Web (gewogen)
OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3
GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask
Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

De geschatte grootte van het geïndexeerde World Wide Web wordt sterk bepaald door de volgorde waarin de zoekmachines gesorteerd zijn. In sectie 3.2.6 is de kruislingse overlap bepaald voor de volgorde 'Google, Yahoo!, Msn Search en Ask' (GYMA) en 'Yahoo!, Google, Msn Search en Ask' (YGMA). In figuur 11 is duidelijk het verschil te zien. De geschatte grootte van het geïndexeerde World Wide Web ligt voor de volgorde GYMA tussen de 11,6 en 18,5 miljard webpagina's. Voor de volgorde YGMA ligt de grootte tussen de 5,9 en 7,3 miljard. De rekenvoorbeelden in tabel 12a en 12b op de volgende pagina laten zien hoe de geschatte grootte van het geïndexeerde World Wide Web op 1 juni 2006 bepaald is.

In figuur 11 zijn niet alle trendlijnen weergegeven. Voor de volgordes GYMA en YGMA is enkel een trendlijn weergegeven voor de steekproef met de minste overlap (steekproef 2 of 3). De reden hiervoor is dat steekproef 1 een (te) grote overlap heeft gemeten, steekproef 2 en 3 geven nog steeds een overschatting van de overlap maar deze is minder groot dan in steekproef 1. De volgende metingen bevatten ook alleen trendlijnen voor de steekproef met de minste overlap.

Rekenvoorbeelden:

1) OL1/GYMA: (Kruislingse steekproef 1 / Volgorde = 'Google, Yahoo!, Msn Search en Ask')

Google	100% van 12.571.685.633	=	12.571.685.633	
Yahoo!	6,03% van 5.518.310.091	=	332.754.098	Yahoo! heeft 6,03% url's die Google niet heeft
Msn Search	0,47% van 1.844.695.275	=	8.670.068	Msn Search heeft 0,47% url's die Google + Yahoo! niet hebben
Ask	0,47% van 1.497.526.899	=	7.038.376	Ask heeft 0,47% url's die Google + Yahoo! + Msn Search niet hebben
			12.920.148.175	**Totaal geschatte grootte**

Tabel 12a: Schatting World Wide Web (GYMA) – 1 juni 2006

2) OL1/YGMA: (Kruislingse steekproef 1 / Volgorde = 'Yahoo!, Google, Msn Search en Ask')

Yahoo!	100% van 5.518.310.091	=	5.518.310.091	
Google	2,74% van 12.571.685.633	=	344.464.186	Google heeft 2,74% url's die Yahoo! niet heeft
Msn Search	0,47% van 1.844.695.275	=	8.670.068	Msn Search heeft 0,47% url's die Yahoo! + Google niet hebben
Ask	0,47% van 1.497.526.899	=	7.038.376	Ask heeft 0,47% url's die Yahoo! + Google + Msn Search niet hebben
			5.878.482.721	**Totaal geschatte grootte**

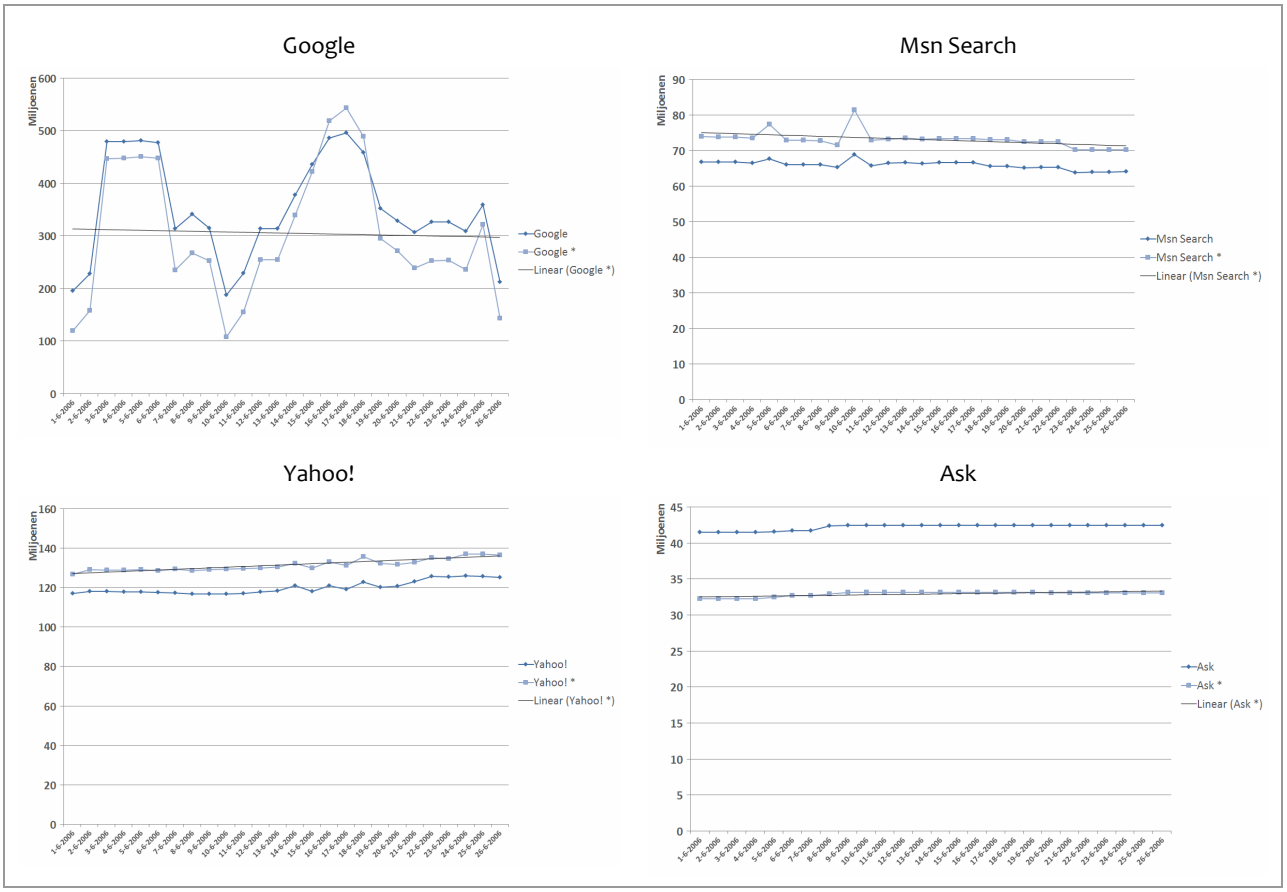
Tabel 12b: Schatting World Wide Web (YGMA) – 1 juni 2006

In het rekenvoorbeeld (tabel 12a en b) wordt de grootte van het geïndexeerde World Wide Web met kruislingse overlap steekproef 1 geschat, in figuur 11 worden ook de geschatte grootte voor steekproef 2 en 3 getoond.

De rekenvoorbeelden laten duidelijk zien dat de overlap nog maar relatief weinig invloed heeft op de geschatte grootte van het World Wide Web. De overlap is (zoals in sectie 3.2 besproken) groter dan in werkelijkheid het geval zal zijn, hierdoor kan een ondergrens-schatting van het World Wide Web gegeven worden. De 'werkelijke' overlap zal een stuk kleiner zijn, waardoor de grootte van het World Wide Web groter zal zijn dan dat hier geschat is.

4.1.3 Krantencorpus (Nederlands) - Geschatte grootte per zoekmachine

In deze meting is gebruik gemaakt van het krantencorpus (Nederlands). Google fluctueert qua grootte in deze corpus heftiger dan in het krantencorpus (Engels). De gewogen metingen lopen uiteen van de 110 tot en met maar liefst 544 miljoen webpagina's! De fluctuaties bij Msn Search, Yahoo! en Ask zijn minder hevig dan bij Google en vertonen een relatief constant patroon. Msn Search daalt licht in de periode 1 t/m 26 juni 2006. Yahoo! en Ask vertonen een licht stijgende lijn.

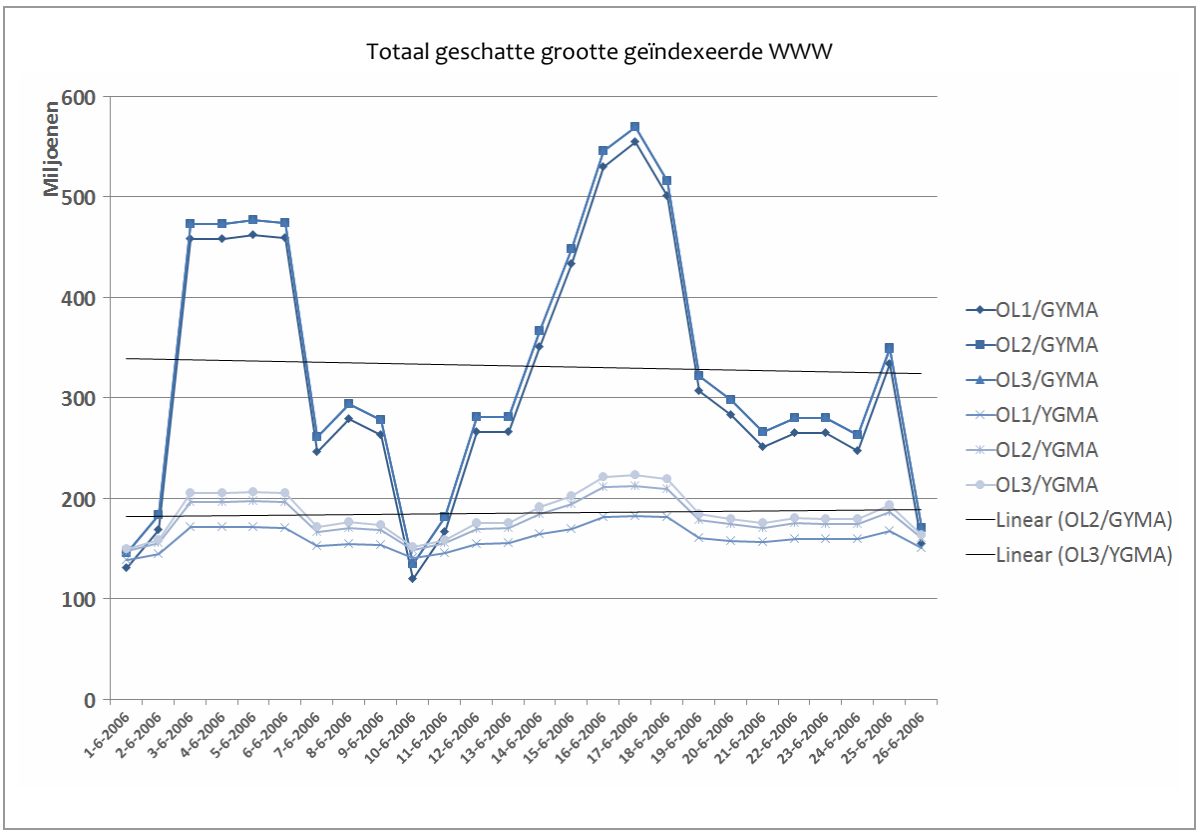


Figuur 12: Meting 2 - Nederlands - Geschatte grootte per zoekmachine

*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**

Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

4.1.4 Krantencorpus (Nederlands) - Geschatte grootte van het geïndexeerde World Wide Web



Figuur 13: Meting 2 - Nederlands - Geschatte grootte World Wide Web (gewogen)
OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3
GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask
Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

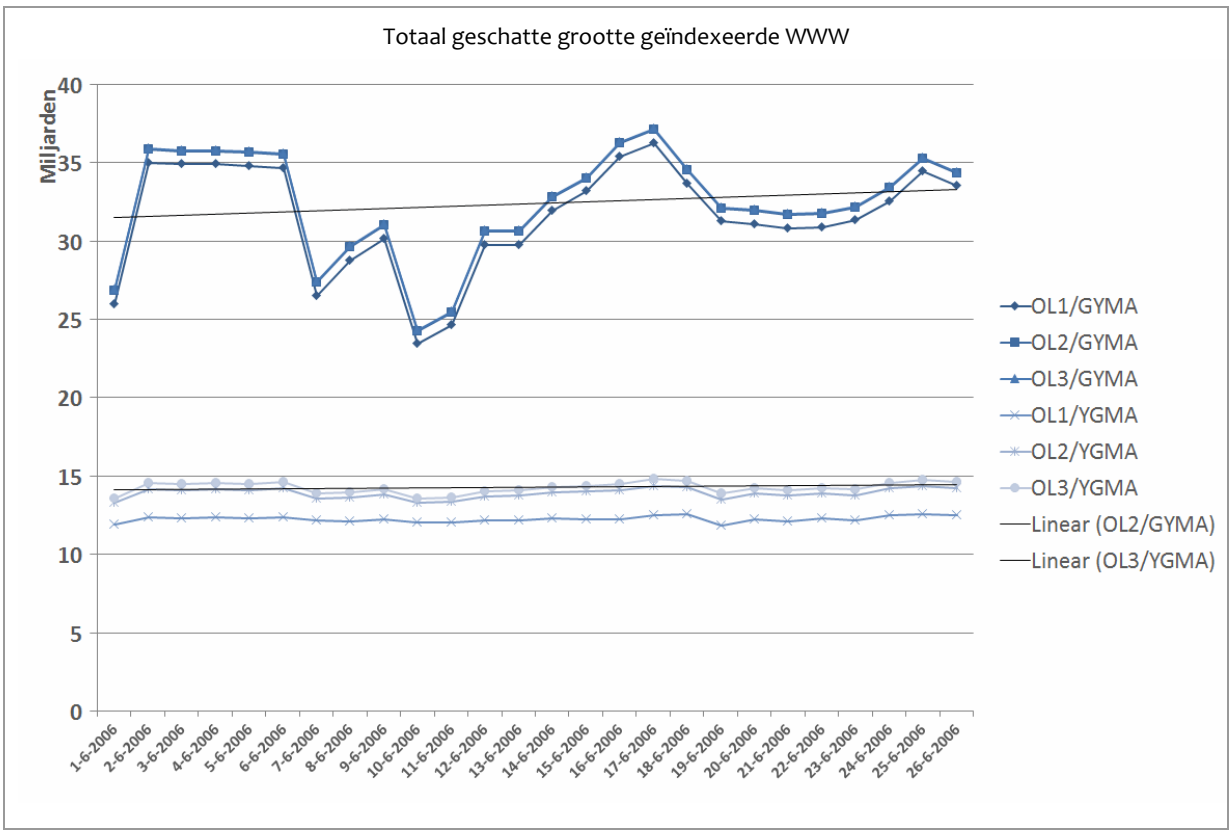
Doordat Google tijdens deze meting erg fluctueerde, fluctueerde de geschatte grootte van het geïndexeerde Nederlandse World Wide Web voor de volgorde GYMA eveneens. Deze schatting is immers grotendeels gebaseerd op de schatte grootte van Google. De grootte van het geïndexeerde Nederlandse World Wide Web liep in de periode 1 t/m 26 juni 2006 uiteen van 131 tot 571 miljoen webpagina's. Omdat de grootte voor de volgorde YGMA sterk rust op de grootte van Yahoo! fluctueerde deze niet zo sterk als voor GYMA. De geschatte grootte ligt voor de volgorde YGMA tussen de 140 en 224 miljoen.

4.1.5 Dmoz corpus (Verschillende talen) - Geschatte grootte per zoekmachine

In deze meting, waarbij gebruik gemaakt is van het Dmoz corpus (Verschillende talen), fluctueerde de geschatte grootte van Google wederom in de periode 1 t/m 26 juni 2006. De geschatte grootte van Msn Search en Ask daalden in deze periode licht. Yahoo! vertoont een lichte stijging. Het Dmoz corpus lijkt, zoals eerder besproken, het meest op het geïndexeerde World Wide Web (zie sectie 3.1.2). De geschatte groottes van de vier onderstaande zoekmachines zijn daardoor ook reëler dan de geschatte groottes met behulp van de Wikipedia en krantencorpora. Deze uitspraak geldt eveneens voor sectie 4.1.6, 4.1.7 en 4.1.8.

Figuur 14: Meting 3 - Verschillende talen - Geschatte grootte per zoekmachine
*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**
Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

4.1.6 Dmoz corpus (Verschillende talen) - Geschatte grootte van het geïndexeerde World Wide Web

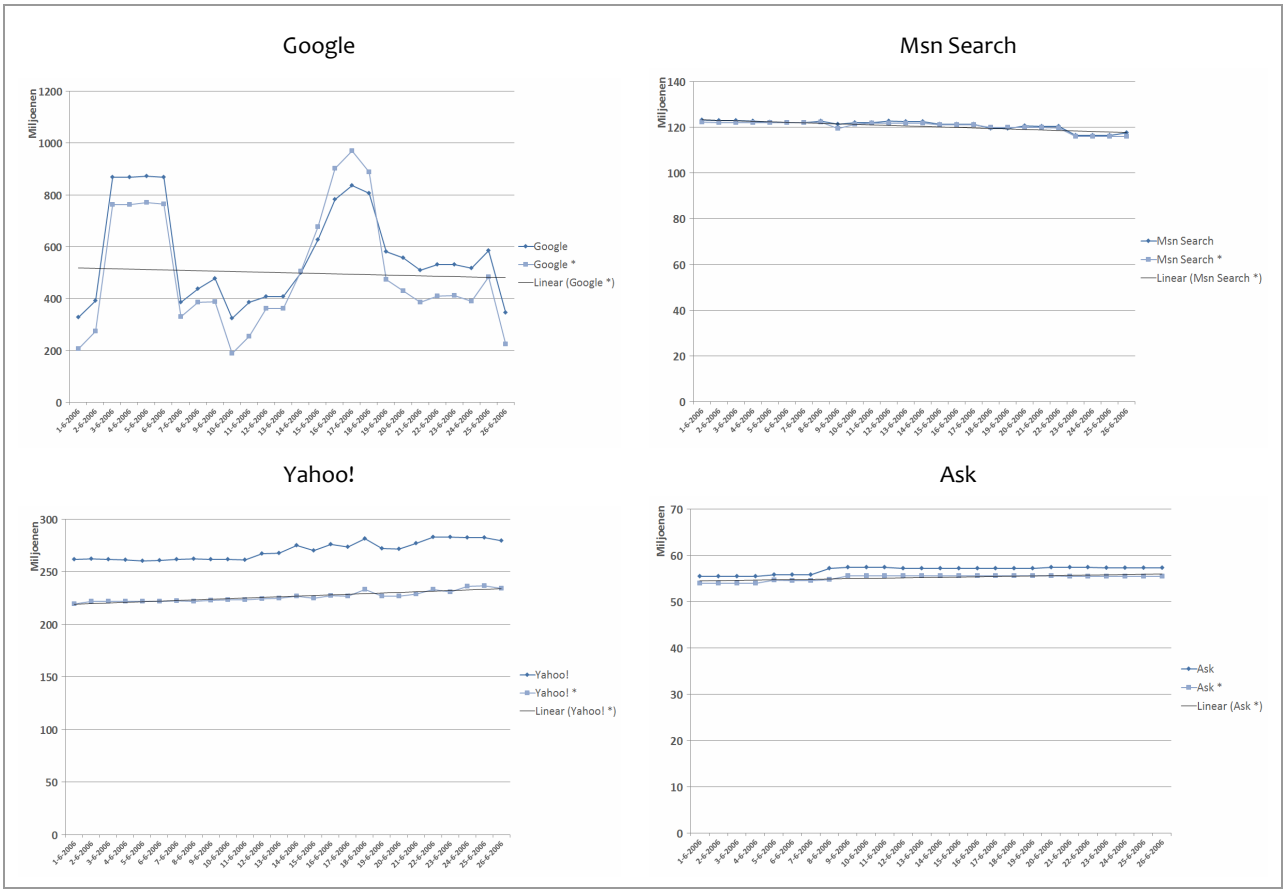


Figuur 15: Meting 3 - Verschillende talen - Geschatte grootte World Wide Web (gewogen)
OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3
GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask
Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

Doordat de geschatte grootte van het geïndexeerde World Wide Web voor de volgorde GYMA erg afhankelijk is van Google, fluctueert deze wederom sterk. De grootte van het geïndexeerde World Wide Web ligt voor de volgorde GYMA tussen de 23,5 en 37,2 miljard webpagina's. Voor de volgorde YGMA ligt de geschatte grootte tussen de 12 en 15 miljard.

4.1.7 Dmoz corpus (Nederlands) - Geschatte grootte per zoekmachine

De meting met het Dmoz corpus (Nederlands) zorgde voor de onderstaande grafieken. De grootste fluctuaties zijn alweer voor rekening van Google. Msn Search daalde in deze periode licht in grootte. Yahoo! en Ask vertonen een lichte stijging.

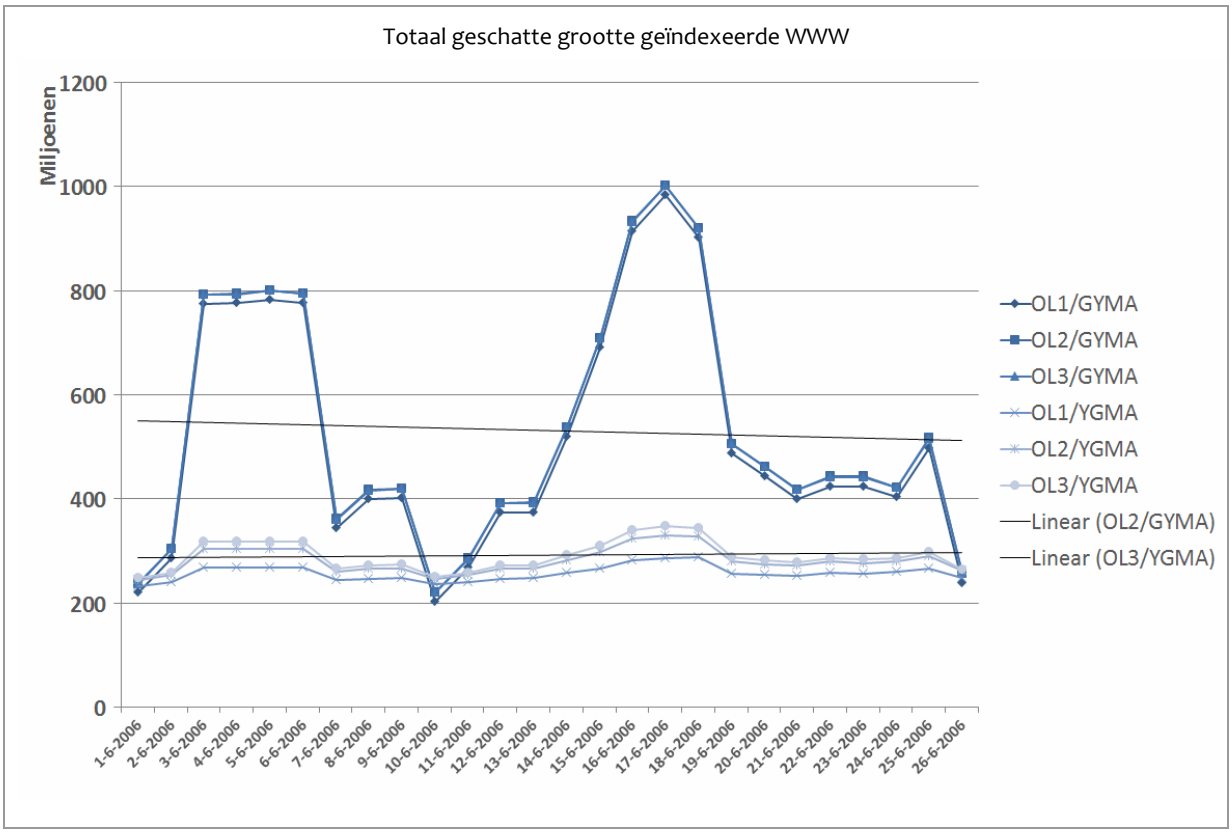


Figuur 16: Meting 4 - Nederlands - Geschatte grootte per zoekmachine

*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**

Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

4.1.8 Dmoz corpus (Nederlands) - Geschatte grootte van het geïndexeerde World Wide Web

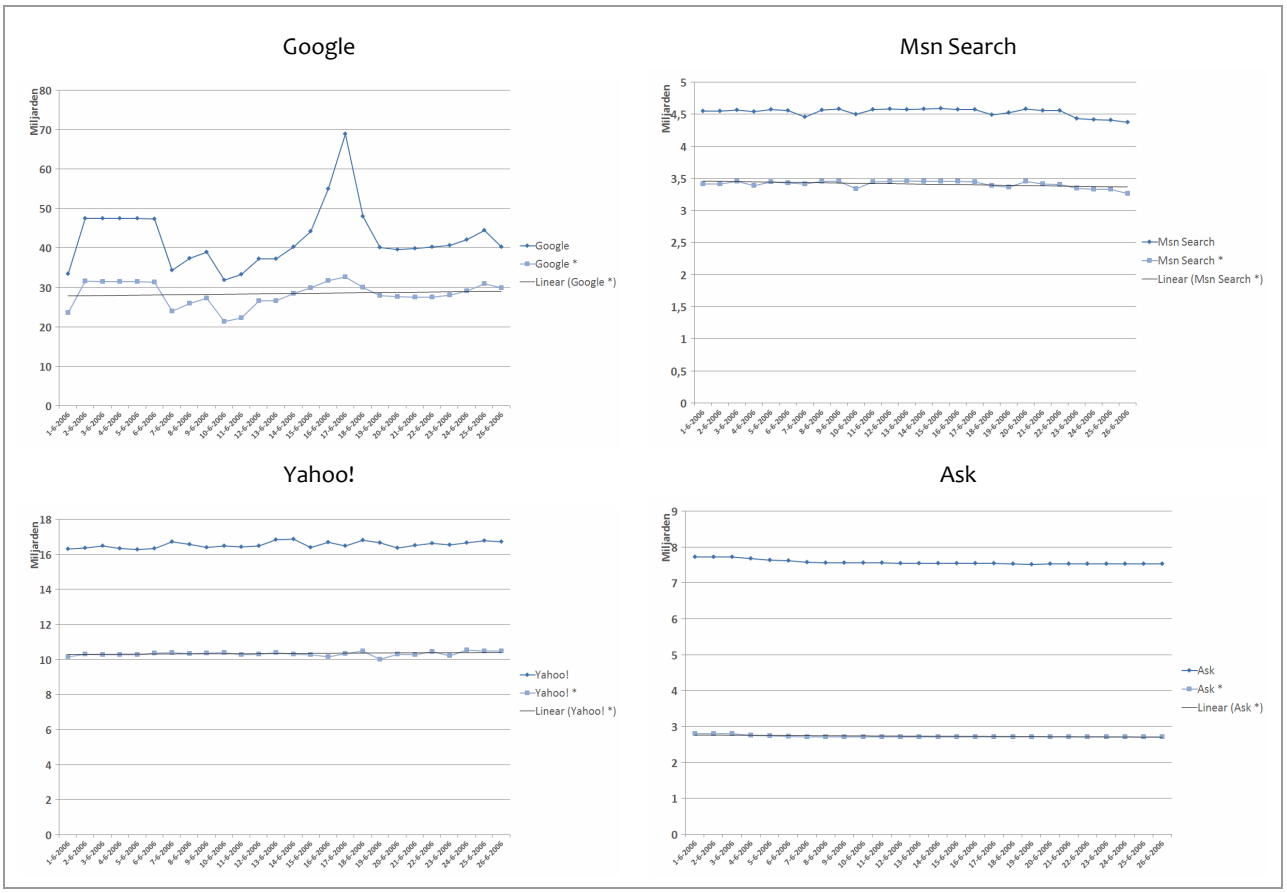


Figuur 17: Meting 4 - Nederlands - Geschatte grootte World Wide Web (gewogen)
OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3
GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask
Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

Ook in deze corpus fluctueert de geschatte grootte, met name voor de volgorde GYMA. De grootte van het geïndexeerde Nederlandse World Wide Web ligt voor de volgorde GYMA tussen de 200 en 1000 miljoen. Voor de volgorde YGMA ligt de geschatte grootte tussen de 246 en 349 miljoen.

4.1.9 Wikipedia corpus (Engels) - Geschatte grootte per zoekmachine

Deze meting heeft gebruik gemaakt van het Wikipedia corpus (Engels). Op 17 juni 2006 werd er een flinke uitschieter gemeten bij de zoekmachine Google, een geschatte grootte van maar liefst 70 miljard webpagina's werd deze dag gemeten. Msn Search en Ask dalen in deze periode zeer licht. Yahoo! vertoont een lichte stijging.

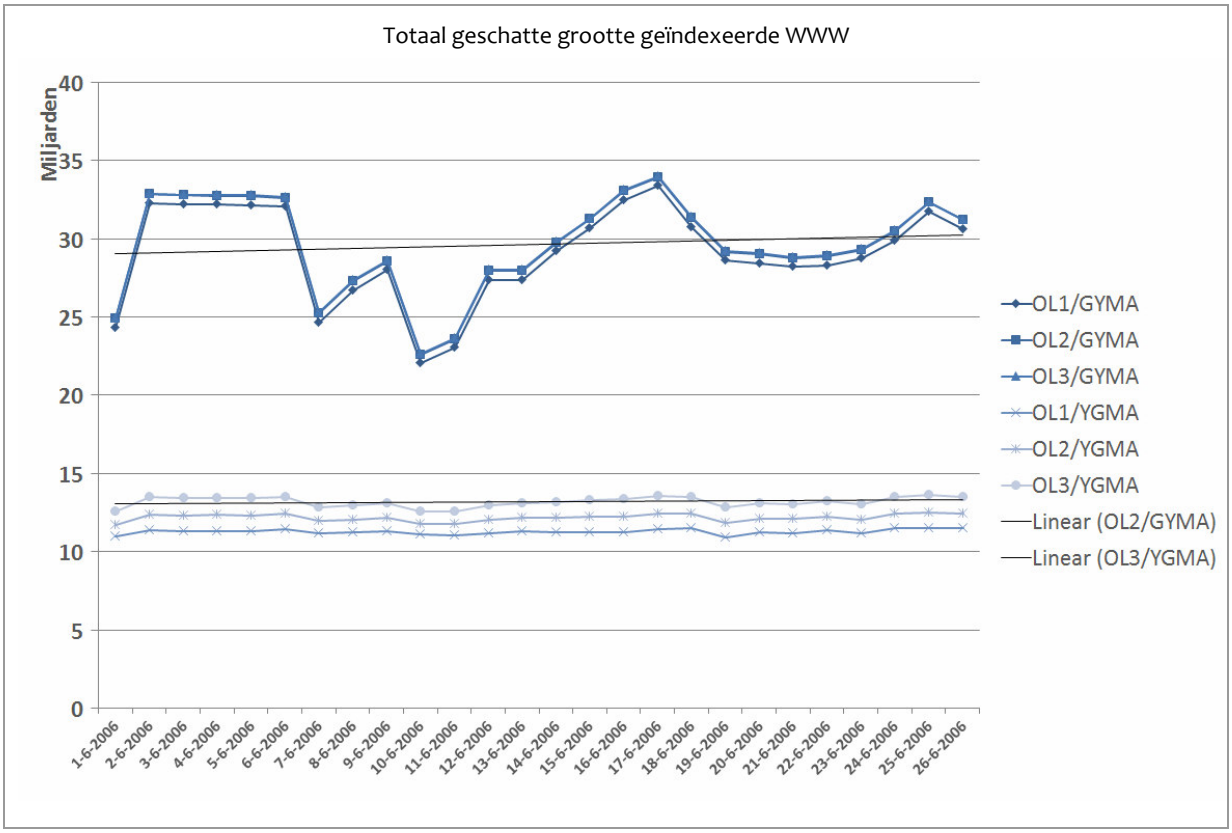


Figuur 18: Meting 5 – Engels - Geschatte grootte per zoekmachine

*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**

Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

4.1.10 Wikipedia corpus (Engels) - Geschatte grootte van het geïndexeerde World Wide Web



Figuur 19: Meting 5 - Engels - Geschatte grootte World Wide Web (gewogen)

OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3

GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask

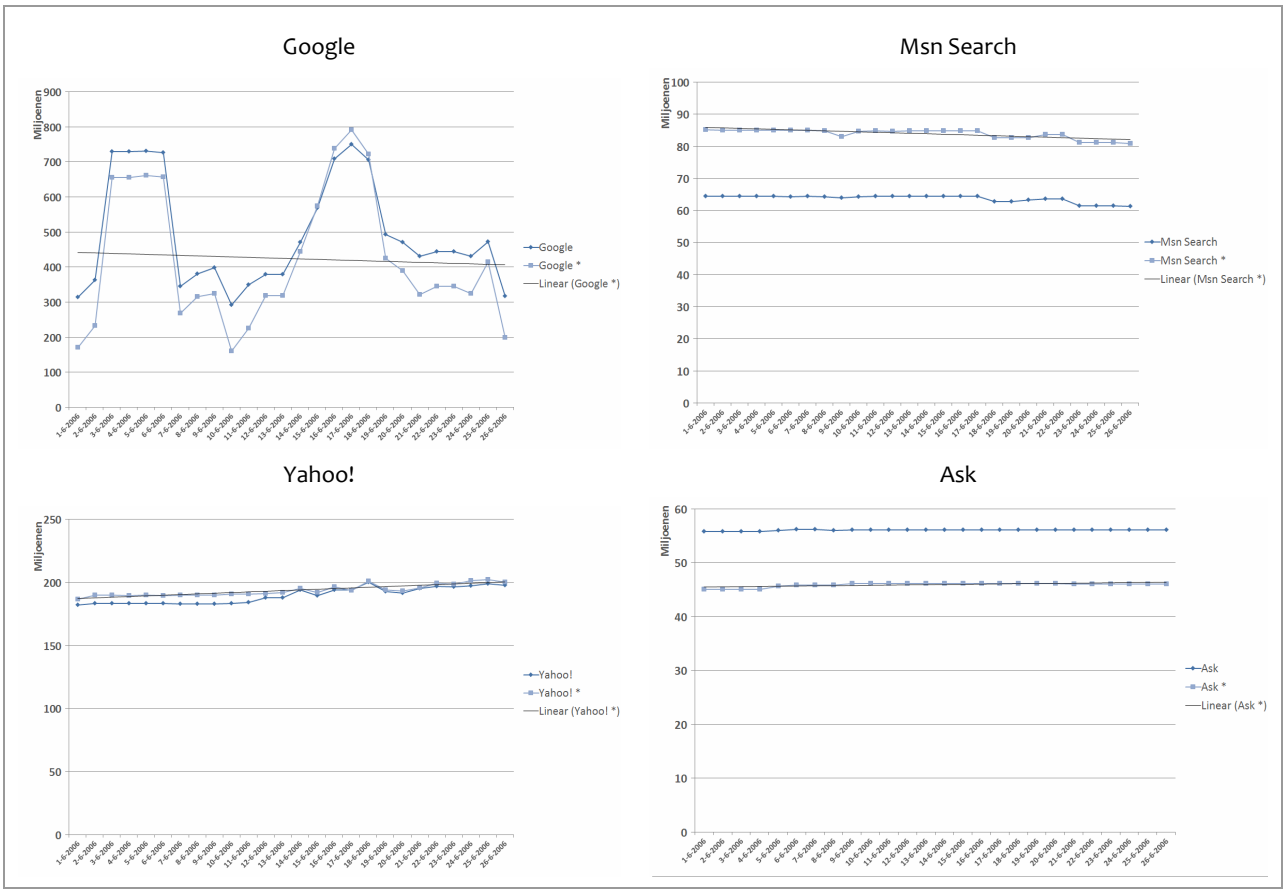
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask

Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

De geschatte grootte van het geïndexeerde Nederlandse World Wide Web voor de volgorde GYMA fluctueert evenals in de andere metingen. De grootte van het geïndexeerde Nederlandse World Wide Web ligt voor de volgorde GYMA tussen de 22,1 en 34 miljard. De YGMA volgorde leverde in de periode 1 t/m 26 juni 2006 een stabiele trend. Voor de volgorde YGMA ligt de geschatte grootte tussen de 11 en 13,6 miljard, afhankelijk van de overlap steekproef.

4.1.11 Wikipedia corpus (Nederlands) - Geschatte grootte per zoekmachine

Net zoals alle andere metingen fluctueert Google fors. Deze meting, die met behulp van het Wikipedia corpus (Nederlands) is uitgevoerd, levert voor Google geschatte groottes op die uiteenlopen van 180 miljoen tot 800 miljoen. Msn Search daalt in deze periode zeer licht en Yahoo! en Ask vertonen een lichte stijging.

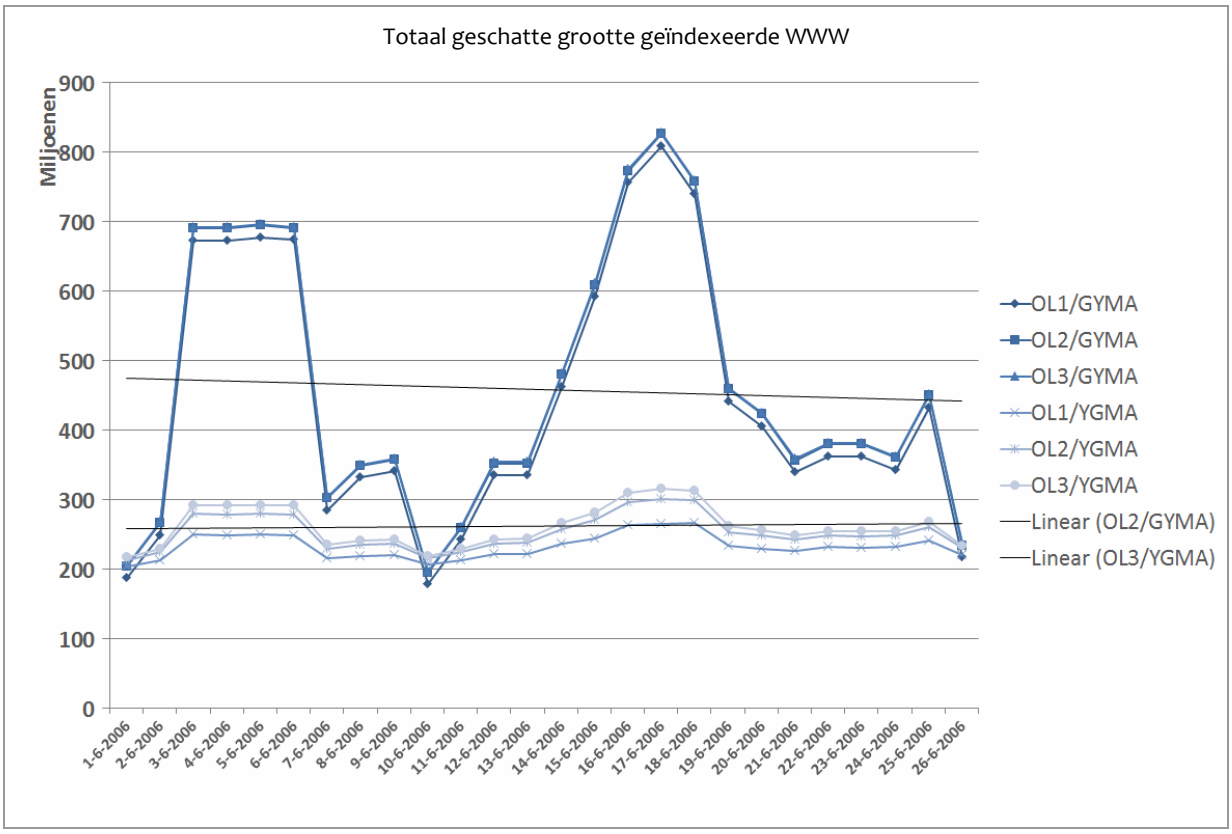


Figuur 20: Meting 6 – Nederlands - Geschatte grootte per zoekmachine

*** = gewogen geschatte grootte, zonder * = ongewogen geschatte grootte**

Linear (zoekmachine *) = Trendlijn zoekmachine voor gewogen geschatte grootte

4.1.12 Wikipedia corpus (Nederlands) - Geschatte grootte van het geïndexeerde World Wide Web



Figuur 21: Meting 6 - Nederlands - Geschatte grootte World Wide Web (gewogen)
OL1 = Overlap steekproef 1, OL2 = Overlap steekproef 2, OL3 = Overlap steekproef 3
GYMA = Gesorteerd op Google, Yahoo!, Msn Search en Ask
YGMA = Gesorteerd op Yahoo!, Google, Msn Search en Ask
Linear (OLx / GYMA of YGMA) = Trendlijn grootste gewogen geschatte grootte voor sortering GYMA of YGMA

Ook in deze corpus fluctueert de geschatte grootte, met name voor de volgorde GYMA. De grootte van het geïndexeerde Nederlandse World Wide Web ligt voor de volgorde GYMA tussen de 178 en 828 miljoen. Voor de volgorde YGMA ligt de geschatte grootte tussen de 197 en 316 miljoen.

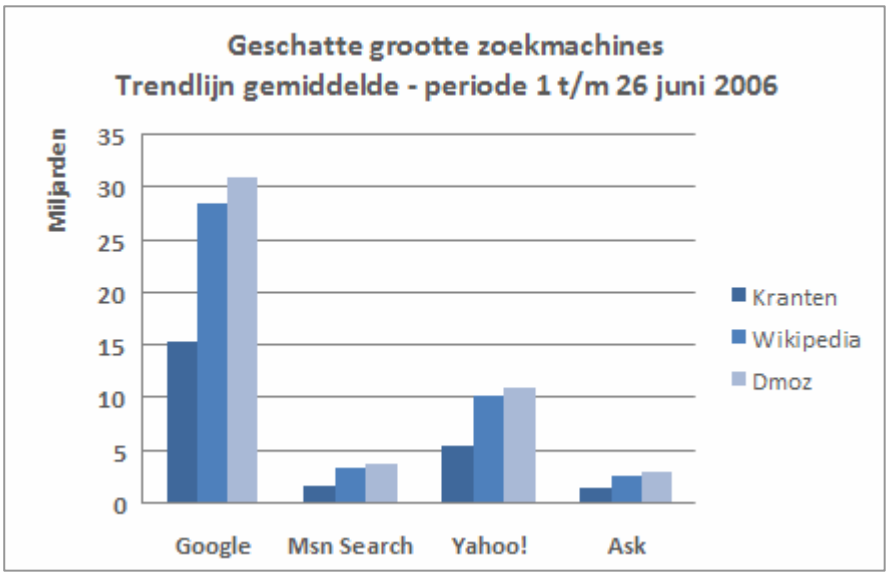
4.2 Overzicht metingen I t/m VI

4.2.1 Geschatte grootte per zoekmachine - Metingen I, III, V (Engels + verschillende talen)

In de vorige secties zijn van ieder corpus afzonderlijk de resultaten weergegeven, in deze sectie worden de trendlijnwaarden van het krantencorpus (Engels), Wikipedia corpus (Engels) en Dmoz corpus (verschillende talen) naast elkaar geplaatst. In de onderstaande tabel 13 en figuur 22 wordt per corpus de gemiddeld geschatte grootte per zoekmachine in miljarden webpagina's weergegeven. Deze gemiddelden zijn gebaseerd op de minimale en maximale trendlijnwaarden uit de zes metingen.

	Google			Msn Search			Yahoo!			Ask		
	min	max	**gem**	min	max	**gem**	min	max	**gem**	min	max	**gem**
Kranten	15	15,7	**15,4**	1,8	1,8	**1,8**	5,6	5,6	**5,6**	1,5	1,4	**1,5**
Wikipedia	28,5	28,5	**28,5**	3,4	3,4	**3,4**	10,3	10,3	**10,3**	2,7	2,7	**2,7**
Dmoz	30	31,8	**30,9**	3,8	3,8	**3,8**	11,2	11,3	**11,3**	3	3	**3**

Tabel 13: Geschatte grootte zoekmachines in miljarden



Figuur 22: Geschatte grootte zoekmachines in miljarden

Figuur 22 toont duidelijk aan het krantencorpus de kleinste grootte per zoekmachine schat. De geschatte grootte met behulp van het Wikipedia en Dmoz corpus liggen dichter bij elkaar. Het krantencorpus verschilt zeer sterk van de andere twee corpora. Dat komt doordat deze corpus anders is opgebouwd. Het krantencorpus bevatte veel lange artikelen, de gemiddelde grootte van een artikel in het krantencorpus bevat 775 woorden, terwijl een artikel uit het Wikipedia corpus gemiddeld 264 woorden bevat en het Dmoz corpus gemiddeld 478 woorden. Daarnaast

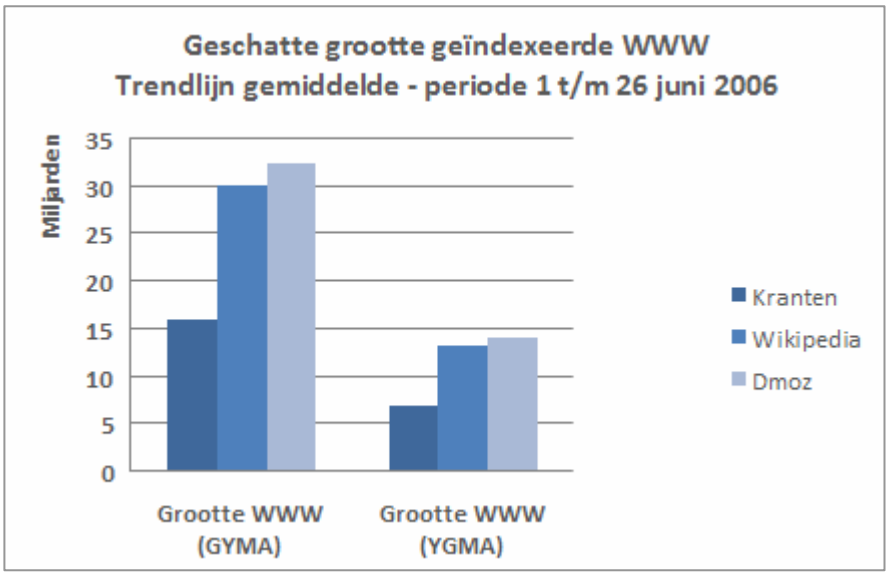
woorden in een krantenartikel vaak hoogfrequente woorden gebruikt. Zo komt bijvoorbeeld het woord 'the' in 99,77% van de Engelse krantenartikelen voor, terwijl dit in het Wikipedia corpus (Engels) 84,07 % en Dmoz corpus (verschillende talen) 67,61 % is (zie bijlage III). Hierdoor zal de geschatte grootte van een zoekmachine met de formule uit sectie 3.1 niet veel groter zijn dan de documentfrequentie die een zoekmachine voor bijvoorbeeld het woord 'the' retourneert. Je kan je dan ook afvragen of het zinvol is om het krantencorpus te gebruiken om de grootte van een zoekmachine te schatten. Het Dmoz corpus geeft naar mijn mening de betrouwbaarste schatting vanwege het feit dat voor deze corpus willekeurig webpagina's bezocht zijn in plaats van artikelen die uit een kranten- of encyclopediedomein komen. Het Dmoz corpus is hierdoor in feite een mini World Wide Web en heeft (zoals in sectie 3.1 besproken) daardoor een soortgelijke 'vingerafdruk' als het geïndexeerde World Wide Web. Het krantencorpus en Wikipedia corpus hebben beide niet deze soortgelijke 'vingerafdruk'.

4.2.2 Geschatte grootte geïndexeerde World Wide Web - Metingen I, III, V (Engels + verschillende talen)

Tabel 14 toont per corpus de geschatte grootte van het geïndexeerde World Wide Web. In deze sectie worden ook het gemiddelde van de minimale en maximale trendlijnen gebruikt om de groottes te schatten.

	Grootte WWW (GYMA)			Grootte WWW (YGMA)		
	min	max	**gem**	min	max	**gem**
Kranten	15,7	16,4	**16,1**	7	7,1	**7,1**
Wikipedia	29,7	30,5	**30,1**	13,2	13,2	**13,2**
Dmoz	32	33	**32,5**	14,2	14,4	**14,3**

Tabel 14: Geschatte grootte geïndexeerde World Wide Web in miljarden



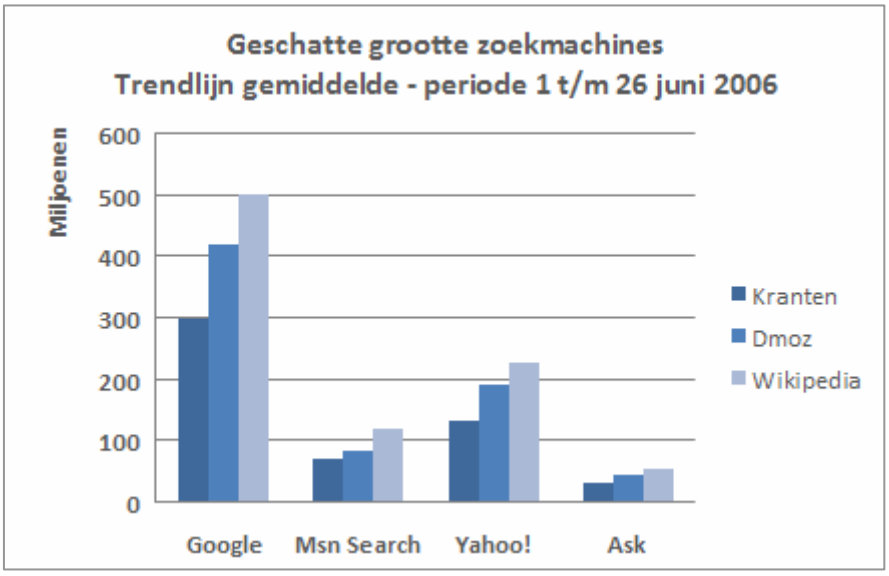
Figuur 23: Geschatte grootte geïndexeerde World Wide Web in miljarden

4.2.3 Geschatte grootte per zoekmachine - Metingen II, IV, VI (Nederlands)

In deze sectie worden de trendlijnwaarden van het krantencorpus (Nederlands), Dmoz corpus (Nederlands) en Wikipedia corpus (Nederlands) getoond. In de onderstaande tabel 15 en figuur 24 wordt per corpus de gemiddeld geschatte grootte per zoekmachine weergegeven. De eenheid van deze groottes is in miljoenen Nederlandse webpagina's. Deze gemiddelden zijn gebaseerd op de minimale en maximale trendlijnwaarden uit de zes metingen. Evenals in sectie 4.2.1 levert het Dmoz corpus de grootste schattingen op.

	Google			Msn Search			Yahoo!			Ask		
	min	max	**gem**	min	max	**gem**	min	max	**gem**	min	max	**gem**
Kranten	300	300	**300**	70	75	**73**	130	140	**135**	32	34	**33**
Wikipedia	400	440	**420**	82	85	**84**	187	200	**194**	45	46	**46**
Dmoz	484	520	**502**	118	123	**121**	220	235	**228**	54	55	**55**

Tabel 15: Geschatte grootte zoekmachines in miljoenen



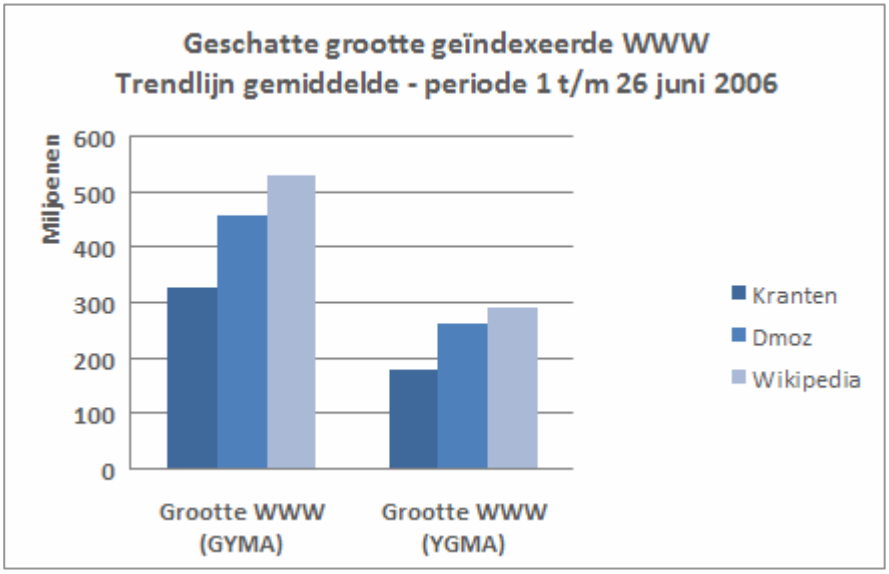
Figuur 24: Geschatte grootte zoekmachines in miljoenen

4.2.4 Geschatte grootte geïndexeerde World Wide Web – Metingen II, IV, VI (Nederlands)

In de onderstaande tabel 16 en figuur 25 worden per corpus de geschatte grootte van het geïndexeerde Nederlandse World Wide Web weergegeven. Wederom is het gemiddelde van de minimale en maximale trendlijnwaarden uit de metingen gebruikt om de geschatte grootte te bepalen.

	Grootte WWW (GYMA)			Grootte WWW (YGMA)		
	min	max	**gem**	min	max	**gem**
Kranten	320	340	**330**	180	180	**180**
Wikipedia	445	470	**458**	264	264	**264**
Dmoz	510	555	**533**	286	296	**291**

Tabel 16: Geschatte grootte geïndexeerde Nederlandse World Wide Web in miljoenen



Figuur 25: Geschatte grootte geïndexeerde Nederlandse World Wide Web in miljoenen

5. Conclusie & Discussie

Om het aantal geschatte webpagina's van het geïndexeerde World Wide Web te bepalen zijn de groottes van de indexen van de vier grootste zoekmachines geschat. Daarnaast is de overlap tussen deze indexen geschat. De grootte van iedere index is geschat via een extrapolatiemethode, die gebruik maakt van een groot tekstcorpus (gebaseerd op verzamelingen van krantenartikelen, encyclopedische artikelen, of informatieve webpagina's), en die de berekende documentfrequentie van woorden uit deze verzamelingen vergelijkt met de opgegeven "hit rate" van de betreffende zoekmachine. De overlap tussen de indexen van de vier zoekmachines is geschat op basis van drie grote steekproeven met vele zoektermen die naar alle zoekmachines werden verzonden. De uiteindelijke schatting van de grootte van het geïndexeerde World Wide Web is gebaseerd op de optelsom van de geëxtrapoleerde groottes van de indexen van de individuele zoekmachines, waarvan de overlap is afgetrokken, en een correctie heeft plaatsgevonden op verwachte "dode links" (niet meer bestaande pagina's).

In deze scriptie wordt een *ondergrens-schatting* van het geïndexeerde World Wide Web bepaald. De extrapolatie van de geschatte zoekmachine indexen is weliswaar geen over- of ondergrens-schatting, maar de geschatte overlap is wel een overschatting. De overlap tussen de zoekmachines wordt hierdoor groter geschat dan dat deze in werkelijkheid is (figuur 22). Door een ondergrens-schatting te maken kan een vrij betrouwbare ondergrens van de geschatte grootte van het geïndexeerde World Wide Web bepaald worden.

Figuur 22: Overlap World Wide Web

Om de grootte van het World Wide Web te schatten ben je daarnaast afhankelijk van interne en externe factoren. Op de interne factoren (zoals de verschillende corpora, overlap en dode links

steekproeven) kan invloed uitgeoefend worden. Deze interne factoren kunnen door de onderzoeker bijgesteld worden. De grootste externe factor waar geen invloed op uitgeoefend kan worden is het fluctuatiegedrag van een zoekmachine. De zoekmachine Google retourneerde soms aantallen die zo erg fluctueerden dat er geen duidelijke trend waarneembaar werd. Hierdoor is het lastig om vast te stellen hoeveel webpagina's Google nu daadwerkelijk bevat.

Daarnaast is het opvallend dat Yahoo! in 5 van de 6 overlap-steekproeven een betere dekkingsgraad heeft, terwijl je zou verwachten dat Google in alle steekproeven de beste dekkingsgraad zou moeten hebben. Bovendien overlapt de database van Yahoo meer met andere zoekmachines dan Google. Dit wekt de indruk dat Google de geretourneerde aantallen kunstmatig ophoogt. Het kan betwist worden of Google inderdaad de grootste zoekmachine is. Als Google niet de grootste index heeft, dan valt die eer te beurt aan Yahoo!. Op grond van deze twijfel is de grootte niet alleen geschat op basis van de volgorde Google, Yahoo!, Msn Search en Ask (GYMA), maar ook op Yahoo!, Google, Msn Search en Ask (YGMA) in sectie 4. De sortering van de zoekmachines heeft grote invloed op de geschatte grootte van het geïndexeerde World Wide Web zoals in sectie 4.2 duidelijk te zien is. De keuze om te sorteren van grootste naar kleinste zoekmachine (in absolute aantallen webpagina's) is gemaakt omdat het geïndexeerde World Wide Web minstens zo groot is als de grootste zoekmachine.

Aangezien de fluctuaties bij Google dermate groot zijn en omdat Yahoo! een betere dekkingsgraad heeft is het onmogelijk om te vertrouwen op de GYMA-schatting van het geïndexeerde World Wide Web. De YGMA-schatting kan daarom als de betrouwbaarste ondergrens opgevat worden. De eindconclusie die hieruit volgt is dat het geïndexeerde World Wide Web tenminste 14.3 miljard webpagina's telt en het geïndexeerde Nederlandse World Wide Web tenminste 291 miljoen webpagina's. Deze groottes zijn geschat met behulp van de Dmoz corpora, die bestaan uit webpagina's, en daarom de beste afspiegeling vormen van teksten zoals ze op het World Wide Web voorkomen (zie sectie 3.1.2 en 4.1.5).

In de periode van 1 t/m 26 juni 2006 steeg de grootte van het geindexeerde World Wide Web met 2%, of 0.28 miljard webpagina's. Wanneer dit verschil door de 26 meetdagen gedeeld wordt betekent dit een toename van 10,8 miljoen webpagina's per dag. Het verschil voor het geindexeerde Nederlandse World Wide Web bedroeg in diezelfde periode 10 miljoen webpagina's. Deelt men dit verschil door de 26 meetdagen dan is het Nederlandse web in deze periode met ongeveer 385.000 webpagina's per dag gegroeid.

In deze scriptie is eerder het Surface Web (geïndexeerde World Wide Web) en Deep Web aan de orde gekomen (zie sectie 2.2). Het Deep Web is volgens Bergman [21] een factor 400 tot 550 groter dan het Surface Web. Door deze factor toe te passen op de betrouwbare YGMA-schatting, telt het Deep World Wide Web een slordige 5,7 tot 7,9 biljoen webpagina's. Het Nederlandse

Deep World Wide Web zou door de vermenigvuldiging 116 tot 160 miljard webpagina's tellen. Deze extrapolatie naar de geschatte Deep World Wide Web grootte is uiteraard volstrekt onbetrouwbaar en kan slechts als een indicatie worden opgevat.

Mogelijk vervolgonderzoek

De methodiek in deze scriptie heeft enkele nadelen. Een factor die kan leiden tot zeer verschillende schattingen is het gebruikte corpus. De resultaten tonen duidelijk aan dat de geschatte grootte per corpus behoorlijk verschilt. De twee meest reële corpora waren het Dmoz corpus (verschillende talen) en het Dmoz corpus (Nederlands). Deze Dmoz corpora komen namelijk het meest met het geïndexeerde World Wide Web overeen, omdat ze zijn opgebouwd uit webpagina's in plaats van krantenartikelen (krantencorpora) en encyclopedische artikelen (Wikipedia corpora).

De Dmoz.org directory bevat echter geen onuitputtelijke lijst url's. Om in een vervolgonderzoek niet afhankelijk te zijn van een 'url leverancier', zoals Dmoz.org, is het raadzaam om de url's via een andere methode te bemachtigen. Door willekeurige woorden (of woord combinaties) in verschillende talen naar meerdere zoekmachines te verzenden en de url's die door de zoekmachines geretourneerd worden op te slaan, kan een onuitputtelijke bron van url's verkregen worden. Deze url's kunnen dan met een crawler bezocht worden waarna ze verwerkt worden in een 'mini geïndexeerd World Wide Web corpus'.

Daarnaast is het zinvol om in toekomstig onderzoek de overlap nauwkeuriger te schatten. Dit kan door een grotere hoeveelheid url's te controleren dan dat in deze scriptie gebeurd is. De afwijking tussen de steekproeven 2 en 3 verschilt namelijk nog enkele procenten, terwijl exact dezelfde methode gebruikt is. Deze afwijking zal kleiner worden naarmate meer url's gecontroleerd worden. Om de overlap te schatten zijn in deze scriptie twee methoden gebruikt. De eerste methode was via Zipf's law en de tweede via willekeurige woordselectie. Willekeurige woordselectie heeft de voorkeur (overlap steekproef 2 en 3) boven Zipf-woordselectie (overlap steekproef 1). Zipf-woordselectie levert, vergeleken met willekeurige woordselectie, een (te) grote overschatting van de overlap. Er worden veel hoogfrequente woorden gebruikt die zorgen voor de (te) grote overschatting (zie sectie 3.2.4). De willekeurige woordselectie levert een kleinere overschatting van de overlap.

6. Bibliografie

[21] Bergman, M. K. (2001). *The Deep Web: Surfacing Hidden Value*.
[24] Lawrence, S., & Giles, C. Lee (1998). *Searching the World Wide Web*. Science, vol. 280,
[25] Gulli, A., & Signorini, A. (2005). *The Indexable Web is More than 11,5 billion pages*.
[26] Mettrop, W. (2001). *Internet search engines – Fluctuations in document accessibility*. Journal of Documentation vol. 57, no. 5, 623-651
[31] Reed, W.J. (2001). *The Pareto, Zipf and other power laws*. Economics Letters. vol. 74, issue 1, 15-19
[34] Bharat & Broder (1998). *A technique for measuring the relative size and overlap of public web search engines. In 7th WWW.*
[1] Google. "Google Corporate Information: Quick Profile", http://www.google.com/corporate/facts.html
[2] Notess, Greg R. (2004). "On The Net - Fiddling with File Types", http://www.infotoday.com/online/mar04/OnTheNet.shtml
[3] Véronis, J. (2005). "Technologies du Langage: Web: Google adjusts its counts", http://aixtal.blogspot.com/2005/03/web-google-adjusts-its-counts.html
[4] Véronis, J. (2005). "Technologies du Langage: Web: Google's missing pages: mystery solved?", http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html
[5] Véronis, J. (2005). "Technologies du Langage: Web: MSN cheating too ?", http://aixtal.blogspot.com/2005/02/web-msn-cheating-too.html
[6] Liberman, M. (2005). "Language Log: More arithmetic problems at Google", http://itre.cis.upenn.edu/~myl/languagelog/archives/001840.html
[7] Ask. "Ask.com Search Engine - Better Web Search", http://search.ask.com/#subject:ask|pg:1
[8] Inktomi. "Inktomi Web Search FAQ", http://support.inktomi.com/Search_Engine/Product_Info/FAQ/searchfaq.html#portals
[9] Msn. "Site Owner Help: About site indexing on MSN Search ", http://search.msn.com/docs/siteowner.aspx?t=SEARCH_WEBMASTER_REF_GettingSiteIndexed.htm
[11] Google. "Webmaster Help Center", http://www.google.com/webmasters/bot.html
[12] Antezeta. "Enhanced Robots.pm for AWStats: Recognize Additional Robots in your Web Analytics Reports - Antezeta", http://www.antezeta.com/awstats-robots.html
[13] Robotstxtorg. "Database of Web Robots, Overview", http://www.robotstxt.org/wc/active/html/index.html
[14] Wikipedia. "Web crawler", http://en.wikipedia.org/wiki/Web_crawler
[15] Wikipedia. "Index of /nlwiki/", http://download.wikimedia.org/nlwiki/
[16] Xml2sql. "Xml2sql", http://xml2sql.sourceforge.net/
[17] Google. "Bedrijfsinformatie van Google: Overzicht van het bedrijf", http://www.google.nl/intl/nl/corporate/
[18] Msn. "Web Search Help: Behind the scenes: How MSN Search works", http://beta.search.msn.com/docs/help.aspx?t=SEARCH_CONC_HowSearchWorks.htm
[19] Yahoo! Search Blog. "Yahoo! Search blog: Our Blog is Growing Up – And So Has Our Index", http://www.ysearchblog.com/archives/000172.html
[20] Wikipedia. "World Wide Web", http://en.wikipedia.org/wiki/World_Wide_Web

[22] Wikipedia. “Deep web”, http://en.wikipedia.org/wiki/Deep_web#Surface_web
[23] Wikipedia. “Zipf’s law”, http://en.wikipedia.org/wiki/Zipf's_law
[27] Sullivan, D. (2006). “comScore Media Metrix Search Engine Ratings”, http://searchenginewatch.com/reports/article.php/2156431
[28] Sullivan, D. (2005). “Search Engine Sizes”, http://searchenginewatch.com/reports/article.php/2156481
[29] Wikipedia. “Inverted index”, http://en.wikipedia.org/wiki/Inverted_index
[30] Wikipedia. “IP-adres”, http://nl.wikipedia.org/wiki/IP-adres
[32] Adamic, L. A. (2002) “Zipf, Power-laws, and Pareto - a ranking tutorial”, http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html
[33] Google. “Redenen om Google te gebruiken”, http://www.google.nl/why_use.html

7. Bijlagen

Bijlage I

Wanneer een webserver een van onderstaande HTTP Header return codes retourneerde is de bijbehorende url geclassificeerd als zijnde dode link:

400, 401, 403, 404, 406, 409, 410, 412
500, 501, 502, 503, 504, Bad gateway

Bijlage II

http://www.dekunder.nl

Bijlage III **Legenda: * = Gewogen T = Geschatte grootte zoekmachine voor betreffende woord**

Kranten Engels (wf = 308771369 - df = 398247) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	the	18412348	5,96%	397321	99,77%	1,76E+10	2,67E+09	7,33E+09	2,14E+09	1,76E+10	2,67E+09	7,35E+09	2,14E+09
2	to	7954106	2,58%	390174	97,97%	1,68E+10	2,66E+09	7,41E+09	2,13E+09	1,71E+10	2,71E+09	7,56E+09	2,17E+09
3	a	7887654	2,55%	391508	98,31%	1,67E+10	2,65E+09	7,92E+09	2E+09	1,7E+10	2,69E+09	8,06E+09	2,03E+09
4	of	7825445	2,53%	394484	99,06%	1,66E+10	2,49E+09	7,1E+09	2,12E+09	1,68E+10	2,51E+09	7,17E+09	2,14E+09
7	that	3499875	1,13%	375806	94,37%	9,47E+09	1,26E+09	3,72E+09	1,07E+09	1E+10	1,34E+09	3,94E+09	1,13E+09
10	on	2369166	0,77%	381139	95,70%	1,36E+10	2E+09	6,01E+09	1,62E+09	1,42E+10	2,09E+09	6,28E+09	1,69E+09
17	by	1798344	0,58%	367495	92,28%	1,32E+10	1,91E+09	5,66E+09	1,61E+09	1,43E+10	2,07E+09	6,13E+09	1,74E+09
27	has	1215620	0,39%	320392	80,45%	6,98E+09	8E+08	2,74E+09	7,36E+08	8,68E+09	9,94E+08	3,41E+09	9,15E+08
43	would	713730	0,23%	249670	62,69%	2,95E+09	3,55E+08	1,53E+09	4,11E+08	4,71E+09	5,66E+08	2,44E+09	6,56E+08
69	first	453538	0,15%	219150	55,03%	5,01E+09	7,02E+08	2E+09	5,78E+08	9,1E+09	1,28E+09	3,63E+09	1,05E+09
110	any	255155	0,08%	163502	41,06%	4,43E+09	6,28E+08	2,13E+09	5,69E+08	1,08E+10	1,53E+09	5,19E+09	1,39E+09
176	atlanta	164181	0,05%	38673	9,71%	3,36E+08	53414771	1,59E+08	32470000	3,46E+09	5,5E+08	1,64E+09	3,34E+08
281	moved	109400	0,04%	52978	13,30%	4,42E+08	46291012	2,03E+08	58340000	3,32E+09	3,48E+08	1,53E+09	4,39E+08
450	cut	73043	0,02%	46799	11,75%	5,45E+08	77800664	3,1E+08	85540000	4,64E+09	6,62E+08	2,64E+09	7,28E+08
721	killed	47130	0,02%	31018	7,79%	2,64E+08	19573221	1,43E+08	36930000	3,39E+09	2,51E+08	1,84E+09	4,74E+08
1153	jim	30941	0,01%	25143	6,31%	4,55E+08	49538157	2,06E+08	48490000	7,21E+09	7,85E+08	3,26E+09	7,68E+08
1845	favor	19099	0,01%	17174	4,31%	1,9E+08	26448187	98700000	21800000	4,41E+09	6,13E+08	2,29E+09	5,06E+08
2951	episode	11171	0%	7788	1,96%	1,55E+08	21456360	86600000	19190000	7,93E+09	1,1E+09	4,43E+09	9,81E+08
4722	sounded	6194	0%	5852	1,47%	33500000	6271173	26100000	7186000	2,28E+09	4,27E+08	1,78E+09	4,89E+08
7556	kaplan	3201	0%	1546	0,39%	54100000	4020622	17100000	3798000	1,39E+10	1,04E+09	4,4E+09	9,78E+08
12089	skinny	1591	0%	1430	0,36%	28100000	11046361	20000000	4122000	7,83E+09	3,08E+09	5,57E+09	1,15E+09
19343	dunlap	762	0%	337	0,08%	8550000	817351	4240000	1466000	1,01E+10	9,66E+08	5,01E+09	1,73E+09
30949	boaters	340	0%	249	0,06%	5480000	951980	4100000	824500	8,76E+09	1,52E+09	6,56E+09	1,32E+09
49518	pottruck	140	0%	57	0,01%	104000	15820	52800	11200	7,27E+08	1,11E+08	3,69E+08	78252042
79228	alcindor	54	0%	42	0,01%	157000	18271	85800	20700	1,49E+09	1,73E+08	8,14E+08	1,96E+08
126765	kingship	20	0%	19	0%	1560000	297264	890000	287700	3,27E+10	6,23E+09	1,87E+10	6,03E+09

Google	Msn Search	Yahoo!	Ask
9,71E+09	1,47E+09	4,69E+09	1,28E+09

Google *	Msn Search *	Yahoo! *	Ask *
1,29E+10	1,89E+09	5,63E+09	1,57E+09

Kranten Nederlands (wf = 70677208 - df = 163227) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	de	5004875	7,08%	163021	99,87%	1,71E+08	93481163	1,58E+08	8365000	1,71E+08	93599290	1,58E+08	8375570
2	van	2396280	3,39%	162116	99,32%	1,23E+08	93462574	1,51E+08	46310000	1,24E+08	94103084	1,52E+08	46627368
3	het	2174207	3,08%	160188	98,14%	1,41E+08	45305529	1,29E+08	37960000	1,44E+08	46165041	1,31E+08	38680157
4	een	1719711	2,43%	158999	97,41%	1,05E+08	46250725	1,3E+08	38060000	1,08E+08	47480595	1,33E+08	39072067
7	dat	1001899	1,42%	147183	90,17%	74700000	32278712	75100000	14680000	82842834	35797323	83286437	16280225
10	op	789310	1,12%	148284	90,85%	1,11E+08	72785203	1,35E+08	38920000	1,22E+08	80119975	1,49E+08	42842079
17	er	427486	0,60%	116544	71,40%	55000000	87043448	70200000	15460000	77030864	1,22E+08	98319394	21652676
27	uit	267076	0,38%	109940	67,35%	54600000	21256498	68800000	14060000	81064164	31559345	1,02E+08	20874765
43	we	166255	0,24%	52510	32,17%	42600000	84936831	37600000	17560000	1,32E+08	2,64E+08	1,17E+08	54585148
69	haar	94221	0,13%	38492	23,58%	25800000	13517645	22700000	4714000	1,09E+08	57322161	96260337	19989922
110	nederland	53581	0,08%	28621	17,53%	46000000	15301750	40800000	12150000	2,62E+08	87266649	2,33E+08	69292060
176	amerikaanse	35195	0,05%	14876	9,11%	6780000	1181949	6490000	1007000	74393591	12968943	71211564	11049314
281	vrouw	22968	0,03%	12757	7,82%	11200000	3545110	11500000	2181000	1,43E+08	45360012	1,47E+08	27906098
450	hulp	14293	0,02%	8944	5,48%	14100000	2159100	8630000	12930000	2,57E+08	39403334	1,57E+08	2,36E+08
721	activiteiten	9397	0,01%	7252	4,44%	15800000	2689458	9750000	12200000	3,56E+08	60533944	2,19E+08	2,75E+08
1153	afrika	5773	0,01%	2970	1,82%	6850000	6607935	4140000	793300	3,76E+08	3,63E+08	2,28E+08	43598646
1845	vakbonden	3476	0%	2328	1,43%	1980000	147048	590000	132700	1,39E+08	10310225	41367668	9304219
2951	voorop	2058	0%	1952	1,20%	1660000	462268	1080000	224200	1,39E+08	38655030	90310020	18747691
4722	ingetrokken	1168	0%	1079	0,66%	1830000	116285	434000	61900	2,77E+08	17591151	65653863	9363996
7556	stembus	640	0%	541	0,33%	638000	27096	169000	27600	1,92E+08	8175229	50989580	8327292
12089	surinamers	344	0%	207	0,13%	506000	43017	114000	12700	3,99E+08	33920463	89893130	10014410
19343	bijstellen	181	0%	176	0,11%	490000	46240	120000	14800	4,54E+08	42884185	1,11E+08	13725907
30949	tegenoffensief	92	0%	88	0,05%	39000	2843	9050	2600	72339239	5273345	16786413	4822616
49518	justitia	45	0%	37	0,02%	170000	16354	53500	6520	7,5E+08	72146334	2,36E+08	28763244
79228	minderheidskabinet	20	0%	19	0,01%	848	509	840	1570	7285079	4372765	7216352	13487705

Google	Msn Search	Yahoo!	Ask
2,02E+08	68564258	1,19E+08	43518005

Google *	Msn Search *	Yahoo! *	Ask *
1,23E+08	75897219	1,29E+08	33865539

Dmoz verschillende talen (wf = 254094395 - df = 531624) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	the	9005508	3,54%	359419	67,61%	1,76E+10	2,67E+09	7,33E+09	2,14E+09	2,6E+10	3,94E+09	1,08E+10	3,16E+09
2	and	4924884	1,94%	341193	64,18%	1,65E+10	2,54E+09	7,23E+09	2,06E+09	2,57E+10	3,96E+09	1,13E+10	3,22E+09
3	of	4915221	1,93%	338596	63,69%	1,66E+10	2,49E+09	7,1E+09	2,12E+09	2,61E+10	3,9E+09	1,11E+10	3,33E+09
4	to	4446931	1,75%	347152	65,30%	1,68E+10	2,66E+09	7,41E+09	2,13E+09	2,57E+10	4,07E+09	1,13E+10	3,26E+09
7	for	2059482	0,81%	310997	58,50%	1,58E+10	2,34E+09	6,76E+09	1,88E+09	2,69E+10	4E+09	1,16E+10	3,22E+09
10	on	1352705	0,53%	260341	48,97%	1,36E+10	2E+09	6,01E+09	1,62E+09	2,78E+10	4,09E+09	1,23E+10	3,3E+09
17	are	937442	0,37%	223867	42,11%	1,05E+10	1,42E+09	4,2E+09	1,19E+09	2,49E+10	3,37E+09	9,97E+09	2,84E+09
27	was	678038	0,27%	115591	21,74%	4,31E+09	8,89E+08	2,51E+09	6,93E+08	1,98E+10	4,09E+09	1,15E+10	3,19E+09
43	can	416578	0,16%	138845	26,12%	7,67E+09	9,06E+08	2,86E+09	8,24E+08	2,94E+10	3,47E+09	1,1E+10	3,15E+09
69	do	272720	0,11%	97707	18,38%	5,73E+09	6,87E+08	2,26E+09	5,46E+08	3,12E+10	3,74E+09	1,23E+10	2,97E+09
110	people	202543	0,08%	83214	15,65%	4,09E+09	3,71E+08	1,57E+09	4,29E+08	2,61E+10	2,37E+09	1E+10	2,74E+09
176	very	130547	0,05%	55382	10,42%	1,91E+09	3,96E+08	1,19E+09	3,41E+08	1,83E+10	3,8E+09	1,14E+10	3,28E+09
281	show	88954	0,04%	46136	8,68%	2,26E+09	5,61E+08	1,14E+09	3E+08	2,6E+10	6,46E+09	1,31E+10	3,45E+09
450	photo	60188	0,02%	33192	6,24%	1,89E+09	5,25E+08	9,95E+08	2,64E+08	3,03E+10	8,4E+09	1,59E+10	4,22E+09
721	headlines	40003	0,02%	23297	4,38%	6,7E+08	17973921	3E+08	67160000	1,53E+10	4,1E+08	6,85E+09	1,53E+09
1153	william	25961	0,01%	11503	2,16%	6,02E+08	58123358	2,27E+08	67650000	2,78E+10	2,69E+09	1,05E+10	3,13E+09
1845	basketball	16172	0,01%	5183	0,97%	2,62E+08	36261691	1,71E+08	46920000	2,69E+10	3,72E+09	1,75E+10	4,81E+09
2951	spread	9658	0%	6731	1,27%	3,12E+08	50564327	1,36E+08	38320000	2,46E+10	3,99E+09	1,07E+10	3,03E+09
4722	nfl	5572	0%	1563	0,29%	1,13E+08	21392471	91100000	26360000	3,84E+10	7,28E+09	3,1E+10	8,97E+09
7556	preliminary	3092	0%	2192	0,41%	1,97E+08	8460214	54500000	15350000	4,78E+10	2,05E+09	1,32E+10	3,72E+09
12089	definite	1698	0%	1248	0,23%	46600000	5858630	23400000	6409000	1,99E+10	2,5E+09	9,97E+09	2,73E+09
19343	psychologists	904	0%	546	0,10%	29600000	2061504	9670000	2476000	2,88E+10	2,01E+09	9,42E+09	2,41E+09
30949	vielfalt	477	0%	404	0,08%	10400000	2900616	9530000	269200	1,37E+10	3,82E+09	1,25E+10	3,54E+08
49518	illini	246	0%	86	0,02%	8400000	727678	3110000	727800	5,19E+10	4,5E+09	1,92E+10	4,5E+09
79228	chèque	126	0%	100	0,02%	4830000	5726893	11100000	48500	2,57E+10	3,04E+10	5,9E+10	2,58E+08
126765	accordée	63	0%	58	0,01%	669000	806598	2500000	49600	6,13E+09	7,39E+09	2,29E+10	4,55E+08
202824	reticular	31	0%	22	0%	1700000	132230	454000	134200	4,11E+10	3,2E+09	1,1E+10	3,24E+09
324519	rectificació	13	0%	12	0%	17	18693	59800	1310	753134	8,28E+08	2,65E+09	58035620

Google	Msn Search	Yahoo!	Ask

Google *	Msn Search *	Yahoo! *	Ask *

2,62E+10	4,8E+09	1,43E+10	3,02E+09

2,61E+10	3,91E+09	1,13E+10	3,19E+09

Dmoz Nederlands (wf = 13880707 - df = 35383) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	de	439839	3,17%	26542	75,01%	1,71E+08	93481163	1,58E+08	8365000	2,28E+08	1,25E+08	2,11E+08	11151337
2	van	266716	1,92%	24433	69,05%	1,23E+08	93462574	1,51E+08	46310000	1,78E+08	1,35E+08	2,19E+08	67064492
3	en	246432	1,78%	23656	66,86%	1,63E+08	93472618	1,53E+08	57730000	2,44E+08	1,4E+08	2,29E+08	86348520
4	in	229302	1,65%	24441	69,08%	1,57E+08	93476418	1,42E+08	33550000	2,27E+08	1,35E+08	2,06E+08	48570011
7	of	136726	0,99%	18055	51,03%	83600000	93276352	92000000	24000000	1,64E+08	1,83E+08	1,8E+08	47033619
10	voor	103431	0,75%	20290	57,34%	1,04E+08	39691297	1,23E+08	37210000	1,81E+08	69216223	2,14E+08	64889178
17	zijn	60320	0,43%	13997	39,56%	74300000	28584669	91000000	28630000	1,88E+08	72259151	2,3E+08	72373744
27	niet	40154	0,29%	10295	29,10%	62900000	23252991	79300000	16590000	2,16E+08	79918463	2,73E+08	57018356
43	heeft	24532	0,18%	9154	25,87%	48900000	17250874	60700000	11600000	1,89E+08	66679886	2,35E+08	44837536
69	nieuwe	17009	0,12%	7672	21,68%	43300000	13952997	47800000	11830000	2E+08	64350742	2,2E+08	54559553
110	na	10631	0,08%	4272	12,07%	40200000	63566867	38500000	6847000	3,33E+08	5,26E+08	3,19E+08	56710534
176	aantal	6761	0,05%	3599	10,17%	29600000	10001645	29200000	15910000	2,91E+08	98329593	2,87E+08	1,56E+08
281	omdat	4199	0,03%	2069	5,85%	17700000	5931202	19300000	4237000	3,03E+08	1,01E+08	3,3E+08	72459048
450	zowel	2834	0,02%	2091	5,91%	13400000	4761145	11600000	2228000	2,27E+08	80566042	1,96E+08	37701255
721	vier	1853	0,01%	1195	3,38%	11000000	3939250	11200000	1857000	3,26E+08	1,17E+08	3,32E+08	54984294
1153	pers	1236	0,01%	679	1,92%	11100000	6037367	5850000	880500	5,78E+08	3,15E+08	3,05E+08	45883257
1845	duitse	801	0,01%	461	1,30%	5290000	1055563	5200000	854500	4,06E+08	81017322	3,99E+08	65585192
2951	veld	498	0%	306	0,86%	3440000	1020180	3010000	729400	3,98E+08	1,18E+08	3,48E+08	84341046
4722	voormalige	300	0%	244	0,69%	2410000	551526	2020000	339500	3,49E+08	79978051	2,93E+08	49231674
7556	aangehouden	176	0%	88	0,25%	1980000	210966	1380000	196600	7,96E+08	84825113	5,55E+08	79048839
12089	ervaar	96	0%	87	0,25%	1230000	171227	386000	51100	5E+08	69638218	1,57E+08	20782429
19343	restaureren	51	0%	33	0,09%	378000	52810	141000	27700	4,05E+08	56623522	1,51E+08	29700276
30949	analyseert	25	0%	25	0,07%	598000	76381	185000	21500	8,46E+08	1,08E+08	2,62E+08	30429380

Google	Msn Search	Yahoo!	Ask
3,38E+08	1,26E+08	2,67E+08	58135662

Google *	Msn Search *	Yahoo! *	Ask *
2,14E+08	1,26E+08	2,24E+08	56631141

Wiki Engels (wf = 274131085 - df = 1036437) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	the	15101722	5,51%	871360	84,07%	1,76E+10	2,67E+09	7,33E+09	2,14E+09	2,09E+10	3,17E+09	8,72E+09	2,54E+09
2	of	8498397	3,10%	797827	76,98%	1,66E+10	2,49E+09	7,1E+09	2,12E+09	2,16E+10	3,23E+09	9,22E+09	2,75E+09
3	and	5964625	2,18%	714756	68,96%	1,65E+10	2,54E+09	7,23E+09	2,06E+09	2,4E+10	3,68E+09	1,05E+10	2,99E+09
4	to	5843826	2,13%	679695	65,58%	1,68E+10	2,66E+09	7,41E+09	2,13E+09	2,56E+10	4,05E+09	1,13E+10	3,25E+09
7	is	4007448	1,46%	650113	62,73%	1,28E+10	1,91E+09	5,36E+09	1,51E+09	2,05E+10	3,04E+09	8,55E+09	2,41E+09
10	that	2272122	0,83%	367388	35,45%	9,47E+09	1,26E+09	3,72E+09	1,07E+09	2,67E+10	3,57E+09	1,05E+10	3,02E+09
17	this	1575527	0,57%	383487	37%	1,24E+10	1,65E+09	5,26E+09	1,51E+09	3,36E+10	4,47E+09	1,42E+10	4,09E+09
27	at	1049962	0,38%	370060	35,71%	1,13E+10	1,59E+09	4,94E+09	1,27E+09	3,16E+10	4,46E+09	1,38E+10	3,54E+09
43	also	592109	0,22%	295318	28,49%	5,4E+09	8,44E+08	2,21E+09	5,78E+08	1,9E+10	2,96E+09	7,76E+09	2,03E+09
69	their	364509	0,13%	142784	13,78%	5,6E+09	5,39E+08	2,05E+09	5,83E+08	4,06E+10	3,91E+09	1,49E+10	4,23E+09
110	years	251743	0,09%	125071	12,07%	2,52E+09	3,35E+08	1,24E+09	3,2E+08	2,09E+10	2,78E+09	1,03E+10	2,65E+09
176	did	153403	0,06%	73706	7,11%	1,2E+09	1,7E+08	8,07E+08	2,19E+08	1,69E+10	2,4E+09	1,13E+10	3,07E+09
281	fact	101336	0,04%	48956	4,72%	1,15E+09	1,02E+08	4,34E+08	1,17E+08	2,43E+10	2,15E+09	9,19E+09	2,47E+09
450	enough	70133	0,03%	42628	4,11%	8,16E+08	1,04E+08	4,9E+08	1,28E+08	1,98E+10	2,53E+09	1,19E+10	3,12E+09
721	held	44785	0,02%	32392	3,13%	9,48E+08	95915676	3,85E+08	1,03E+08	3,03E+10	3,07E+09	1,23E+10	3,31E+09
1153	duke	26861	0,01%	8814	0,85%	1,96E+08	16734904	59200000	14990000	2,3E+10	1,97E+09	6,96E+09	1,76E+09
1845	mythology	15662	0,01%	8142	0,79%	44700000	5005043	21700000	5817000	5,69E+09	6,37E+08	2,76E+09	7,4E+08
2951	pattern	9119	0%	6145	0,59%	2,63E+08	25071096	1,05E+08	31600000	4,44E+10	4,23E+09	1,77E+10	5,33E+09
4722	productions	5055	0%	3768	0,36%	1,33E+08	22894366	76900000	19040000	3,66E+10	6,3E+09	2,12E+10	5,24E+09
7556	infin	2686	0%	995	0,10%	1340000	66383	243000	47300	1,4E+09	69147535	2,53E+08	49269819
12089	infrared	1358	0%	830	0,08%	77000000	4156245	24000000	5634000	9,62E+10	5,19E+09	3E+10	7,04E+09
19343	readiness	669	0%	519	0,05%	65400000	2721043	15000000	3805000	1,31E+11	5,43E+09	3E+10	7,6E+09
30949	biggs	322	0%	176	0,02%	9400000	1012224	3910000	1021000	5,54E+10	5,96E+09	2,3E+10	6,01E+09
49518	figurine	153	0%	98	0,01%	5960000	3028990	7030000	2806000	6,3E+10	3,2E+10	7,43E+10	2,97E+10
79228	refrigerant	74	0%	33	0%	2640000	400308	1800000	568500	8,29E+10	1,26E+10	5,65E+10	1,79E+10
126765	numbuh	36	0%	10	0%	67400	5773	39800	12100	6,99E+09	5,98E+08	4,13E+09	1,25E+09
202824	staunton's	15	0%	10	0%	82900	15054	170000	873900	8,59E+09	1,56E+09	1,76E+10	9,06E+10

Google	Msn Search	Yahoo!	Ask
3,45E+10	4,67E+09	1,66E+10	8,1E+09

Google *	Msn Search *	Yahoo! *	Ask *
2,43E+10	3,51E+09	1,04E+10	2,94E+09

Wiki Nederlands (wf = 36457793 - df = 163445) (2006-06-01)													
log	keyword	W_freq	W_Perc	D_freq	D_Perc	Google	Msn	Yahoo!	Ask	Google T	Msn T	Yahoo! T	Ask T
1	de	2010694	5,52%	140244	85,81%	1,71E+08	93481163	1,58E+08	8365000	1,99E+08	1,09E+08	1,84E+08	9748848
2	van	1157504	3,17%	134792	82,47%	1,23E+08	93462574	1,51E+08	46310000	1,49E+08	1,13E+08	1,83E+08	56154208
3	het	926737	2,54%	117653	71,98%	1,41E+08	45305529	1,29E+08	37960000	1,96E+08	62939000	1,79E+08	52734501
4	een	849966	2,33%	127262	77,86%	1,05E+08	46250725	1,3E+08	38060000	1,35E+08	59400683	1,67E+08	48881180
7	is	511421	1,40%	115738	70,81%	96400000	93472618	1,19E+08	32980000	1,36E+08	1,32E+08	1,68E+08	46574298
10	zijn	292180	0,80%	73448	44,94%	74300000	28584669	91000000	28630000	1,65E+08	63609918	2,03E+08	63710793
17	niet	186655	0,51%	50483	30,89%	62900000	23252991	79300000	16590000	2,04E+08	75284454	2,57E+08	53712191
27	werd	133592	0,37%	43210	26,44%	21000000	6460979	22400000	4534000	79434043	24439128	84729646	17150188
43	nog	73042	0,20%	32416	19,83%	42800000	19351854	58800000	11980000	2,16E+08	97574154	2,96E+08	60404464
69	zie	40786	0,11%	23755	14,53%	25900000	5301852	18100000	3949000	1,78E+08	36479108	1,25E+08	27170882
110	nieuwe	25664	0,07%	14026	8,58%	43300000	13952997	47800000	11830000	5,05E+08	1,63E+08	5,57E+08	1,38E+08
176	verder	17819	0,05%	12069	7,38%	25900000	7545159	27600000	5024000	3,51E+08	1,02E+08	3,74E+08	68037756
281	groep	11386	0,03%	6219	3,80%	12800000	4192071	12100000	3661000	3,36E+08	1,1E+08	3,18E+08	96216778
450	gebruikersportaal	7526	0,02%	3975	2,43%	10100000	32208	283000	170700	4,15E+08	1324336	11636462	7018883
721	kilometer	4949	0,01%	3599	2,20%	5920000	751301	2490000	10450000	2,69E+08	34119587	1,13E+08	4,75E+08
1153	troepen	3069	0,01%	1745	1,07%	1950000	106429	709000	110700	1,83E+08	9968646	66408312	10368689
1845	betere	1852	0,01%	1596	0,98%	5260000	1449264	4390000	892000	5,39E+08	1,48E+08	4,5E+08	91348960
2951	geraakt	1106	0%	966	0,59%	2070000	470181	1640000	258200	3,5E+08	79553554	2,77E+08	43686852
4722	orthodoxe	655	0%	420	0,26%	539000	47940	185000	24700	2,1E+08	18656079	71993631	9612123
7556	sloveens	369	0%	227	0,14%	1600000	14230	67000	8620	1,15E+09	10245913	48241476	6206590
12089	polders	203	0%	148	0,09%	541000	69532	166000	20800	5,97E+08	76788228	1,83E+08	22970649
19343	terugtrok	111	0%	109	0,07%	93800	7740	32200	4910	1,41E+08	11606094	48283752	7362522
30949	situeren	60	0%	58	0,04%	438000	35771	91700	14000	1,23E+09	1,01E+08	2,58E+08	39452241
49518	bombardeerden	32	0%	31	0,02%	24900	1897	3100	1980	1,31E+08	10001780	16344500	10439390
79228	nemeïsche	16	0%	12	0,01%	48	262	203	164	653780	3568549	2764945	2233748

Google	Msn Search	Yahoo!	Ask
3,23E+08	66160301	1,86E+08	58545123

Google *	Msn Search *	Yahoo! *	Ask *
1,76E+08	87416756	1,91E+08	47328821