Discussion Paper

Aandachtspunten bij de waarneming en de verwerking van internetgegevens

De standpunten in dit document zijn die van de auteur(s) en komen niet noodzakelijk overeen met het beleid van Centraal Bureau voor de Statistiek

2016 | 07

Léon Willenborg

Content

Summary

Internetdata vormen een interessante bron met informatie over artikelen en hun prijzen. Echter het geautomatiseerd waarnemen en verwerken van internetinformatie is niet triviaal. In dit stuk wordt ingegaan op de mogelijkheden die internetdata bieden, de beperkingen die er voor gelden en enkele problemen waar men tegen aan loopt (of zou kunnen lopen) als men deze data wil gebruiken voor de berekening van de CPI. Het CBS wil internetdata voorlopig alleen gebruiken voor schoenen en kleding. Scannerdata worden voor andere artikelgroepen gebruikt. In dit stuk wordt daarom ook ingegaan op de vergelijking van beide typen data, de voor- en de nadelen die ze kennen voor gebruik bij de CPI.

1. Inleiding[1]

Internet vormt een rijke bron van data die met vrucht gebruikt kan worden bij het maken van de CPI. Het voordeel van deze data is dat ze publiek toegankelijk is en gratis beschikbaar. In sommige gevallen is de kwaliteit van de gegevens goed, omdat er alle reden is voor web shops om de informatie die daar staat over hun artikelen correct te vermelden. Het gaat dan om artikelinformatie en prijzen - aanbodprijzen wel te verstaan.

Nadelen hebben deze data echter ook. Om te beginnen: het CBS heeft er geen zeggenschap over. Web shops bepalen zelf welke informatie ze via hun sites aanbieden, in welke vorm ze die presenteren, of ze op een bepaald moment de inhoud of de vorm ervan wensen te veranderen, wanneer ze dat wensen te doen. etc. En via het internet zijn alleen aanbodprijzen waar te nemen en geen transactieprijzen. Nu is het nog zo dat in veel gevallen één prijs wordt gerekend voor een artikel, op één moment. Maar het zou heel wel kunnen zijn dat in de toekomst de prijs die iemand voor een artikel moet bepalen ook af hangt van de consument: lijkt hij/zij interesse voor een artikel te hebben (blijkend uit surfgedrag)? Of heeft de consument eerder iets gekocht bij deze web shop? Welke andere producten koopt hij/zij, naast dit product bij de web shop? Afhankelijk van het antwoord kan dan een prijs worden bepaald die een klant dient te betalen. In dat geval heeft een artikel op één moment geen vaste prijs meer, maar vele. En deze prijs hangt niet allen van het artikel zelf maar ook van enkele kenmerken van de consument. Maar hij/zij zou daarnaast ook nog van de schaarste op het moment van aankoop kunnen afhangen. Denk aan prijzen van vliegtickets die oplopen naarmate meer geboekt is op een bepaalde vlucht, met last minute aanbiedingen die dan weer tamelijk goedkoop zijn omdat ze bedoeld zijn als lokkertjes om de laatste plaatsen in een vliegtuig bezet te krijgen. Dit principe zou men ook kunnen toepassen bij stoelen in een theater. Of bij gewilde producten die bij opbod worden verkocht (zoals op Marktplaats of bij e-Bay en soortgelijke sites). Deze wijze van verkopen is op dit moment nog niet wijd verbreid, maar zou in de toekomst wel eens vaker kunnen worden toegepast of zelfs gemeengoed kunnen worden. Internet leent er zich uitstekend voor.

Hierboven is één manier beschreven om internetdata te gebruiken, namelijk als bron van prijs- én artikelinformatie. Tot nu heeft het CBS vooral op deze manier, via Internetrobots, informatie verzameld op het internet. Bij web shops wordt daarbij dagelijks prijsinformatie verzameld samen met informatie over de desbetreffende artike-

[1] De auteur dankt Olav ten Bosch voor zijn review van een eerdere versie van dit stuk. Ook Robert Griffioen, Guido van den Heuvel, Quan Le en Jeroen Pannekoek hebben commentaar geleverd op eerdere versies van dit stuk. Ook hen is de auteur daarvoor dank verschuldigd.

len (meta-informatie). Het gaat daarbij om grote hoeveelheden data, die automatisch worden verzameld door diverse internetrobots, ieder toegesneden op een specifieke web winkel.

Daarnaast zijn andere tools gemaakt, semi-automatisch werkend, waarmee de prijzen op andere sites gemonitord kunnen worden. Hier gaat het om data, die minder massaal zijn. Sites die hier worden waargenomen zijn bijvoorbeeld die van bioscopen en van autorijscholen. Ze verzamelen en monitoren prijsgegevens, signaleren verschillen met de vorige raadpleging van een site. Een (consumptie)analist gaat vervolgens na of het om prijsverschillen gaat (waar de interesse naar uit gaat) of om veranderingen van de website die niets met prijsveranderingen te maken hebben, maar alleen met de opmaak van de site.

In beide boven beschreven gevallen is de prijsinformatie het hoofddoel. Deze wijze van dataverzameling is, in het eerste geval, bedoeld als alternatief voor de reguliere waarneming dan wel, in het tweede geval, ter vervanging van de (informele) inspectie van bepaalde sites via internetbrowsers door consumptieanalisten. Dit is een erg interactieve, bijna handmatige activiteit om prijsdata te verzamelen, waarbij de analisten zelf verschillen met vorige raadplegingen van deze sites in de gaten moeten houden.

Een derde gebruik van internet is als bron van gedetailleerde artikelinformatie, niet prijzen maar artikelomschrijvingen en kenmerken. Prijsinformatie met summiere artikelinformatie komt uit een andere bron, namelijk scannerdata. Voor prijsinformatie zijn scannerdata veel aantrekkelijker omdat het hier transactieprijzen betreft. Maar scannerdata bevatten soms te weinig details over de verkochte producten. Dat is geen probleem als aanvullende meta-informatie uit een andere bron verkregen gekoppeld kan worden. Zo'n bron zou soms internet kunnen zijn. Met speciale internetrobots dient de metadata verzameld te worden. Het verzamelen van aanvullende meta-informatie kan veel minder frequent geschieden dan het verzamelen van prijsinformatie. Immers de meta-informatie van een product verandert niet. Nieuwe producten worden weliswaar geregeld, maar mondjesmaat, op de markt gebracht.

De prijzen-verzamelende robots zijn nuttig in geval een winkel geen scannerdata kan of wil leveren. In ieder geval zijn winkels waar zowel scannerdata als internetdata te krijgen zijn van belang om het verband tussen beide soorten gegevens te zoeken. De volgende vragen zijn dan interessant: Zijn de prijzen van producten afhankelijk van het verkoopkanaal (fysieke winkel of internet)? Worden andere producten aangeboden of verkocht in de fysieke winkel dan via het internet? Worden er verschillende services aangeboden voor aankopen via de fysieke winkel of via het internet?

Het doel van het dit rapport is om zaken die om nader onderzoek vragen bij het gebruik van internetdata op een rij te zetten. Het gaat dan om het verzamelen en bewerken van gegevens die daar te vinden zijn. Het stuk zou kunnen dienen als bron voor enkele projecten op dit terrein.

Het stuk is verder als volgt opgebouwd. In paragraaf 2 wordt nader ingegaan op internetdata en hun eigenschappen, vooral in vergelijking met scanner data. In dit stuk zijn we vooral geïnteresseerd in problemen die kunnen spelen bij internetdata en het verzamelen en verwerken daarvan. In paragraaf 3 wordt kort ingegaan op wat voor internetdata kenmerkend is, hoe ze waargenomen en verwerkt worden. Paragraaf 4 vormt de kern van dit stuk. Hier worden allerlei specifieke onderwerpen besproken die te maken hebben met internetdata, van het monitoren van internetrobots (om te zien of ze naar behoren functioneren) tot het mogelijk gebruik van verzamelsites. In paragraaf 5, tot slot, worden enkele conclusies gepresenteerd.

2. Internetdata

Internetdata lijken in een bepaald opzicht op secundaire databronnen, vaak registers. Het zijn namelijk ook data sets die gemaakt zijn door andere instanties en voor andere dan statistische doelen, meestal administratieve. Internetdata zijn echter anders dan registerdata omdat ze vluchtiger zijn (ze worden frequent bijgewerkt), ze zijn minder gestructureerd (webpagina's in plaats van bestanden met records), ze zijn van wisselende kwaliteit (bij sites van webwinkels is de kwaliteit goed. Daar zijn voldoende prikkels voor. Bij zoeksites is de kwaliteit vaak minder goed, waarschijnlijk omdat ze zijn samengesteld met classificatiesoftware die niet goed is ingeregeld), ze zijn heterogeen samengesteld (zo komen numerieke waarden, gestructureerde tekst, vrije tekst, plaatjes, tabellen, etc. samen allemaal voor op webpagina's).

Een ander verschil met registerdata is dat internetdata in zekere zin fluïde zijn: wat er aan data op een site staat is niet constant. De inhoud wijzigt voortdurend. Door onderhoud op zo'n site kan de inhoud binnen enkele uren behoorlijk veranderen. Afhankelijk van het moment waarop een web robot wordt geactiveerd kan men veel of weinig 'oogsten', gezien de vulling van de site. Registerdata daarentegen zijn statisch en het eindresultaat van een traag verlopend administratief proces. Internetdata daarentegen zijn 'heet van de naald'.

Het gebruik van deze data door het CBS bij het maken van statistieken is daarom ook niet zonder risico. Een website kan haar data anders aanbieden, of zelfs uit de lucht gaan voor een poos, of zelfs permanent. Maar aangezien deze websites belangrijk zijn voor de verkoop van producten is er ook een zekere dwang om ze in de lucht te houden en om ze te vullen met data van goede kwaliteit: de omschrijvingen van producten en hun prijzen moet betrouwbaar zijn, anders lopen klanten weg en kopen hun spullen bij de concurrentie.

Maar ondanks de genoemde negatieve aspecten, zijn er ook voordelen verbonden aan het gebruik van internetdata. Vooralsnog kwalificeren we dergelijke als een rijke, interessante bron, maar één waar wel met verstand dient te gebruiken, zich ten volle bewust van de risico's ten aanzien van de daar beschikbare informatie. Overigens

hoeft dat niet beperkt te blijven tot prijsgegevens. Op internet zijn ook veel omschrijvingen van artikelen te vinden - metadata. Dit soort data kan heel nuttig zijn bij het verrijken van scannerdata. Denk bijvoorbeeld aan scannerdata voor elektronische apparatuur. Op internet zijn vaak gedetailleerde lijsten te vinden met specificaties voor dit soort producten te vinden. Bij scannerdata is doorgaan alleen het type-nummer gegeven en niet de bijbehorende specificaties.

Bij internetdata loopt de dataverzameling en de dataverwerking naadloos in elkaar over. Het is niet goed mogelijk een duidelijke scheidslijn te trekken tussen beide. Daarom worden in dit stuk beide onderwerpen besproken, zonder verder onderscheid te maken tussen dataverzameling en dataverwerking. De verzameling van gegevens bij webwinkels gebeurt overigens met behulp van internetrobots, ook wel web scrapers genoemd.

2.1 Webinfo

2.1.1 Overzicht

In deze paragraaf geven we een globaal overzicht van het soort websites waar het CBS mogelijk interessante informatie vandaan zou kunnen halen. Het om de volgende vier typen sites: webwinkels, informatiesites van fysieke winkels, verzamelsites en productinformatie sites.

2.1.2 Webwinkels

De web scrapers die tot nu toe zijn ingezet op het CBS hebben informatie verzameld van webwinkels, dus sites waar men zowel informatie over artikelen aantreft (meta-informatie in de vorm van tekst en foto's) als de mogelijkheid heeft om kleding te kopen.

Deze sites zijn het meest aantrekkelijk om prijs- en meta-informatie te verzamelen. De kwaliteit van beide soorten data is goed, de informatie wordt goed bijgewerkt en de informatie is volledig (omvat alle producten die men verkoopt). Een webwinkel heeft ook alle reden om deze informatie goed te hebben. Daar hangt immers de verkoop van hun producten van af.

2.1.3 Informatiesites van fysieke winkels

Daarnaast komt het voor dat er winkels zijn die informatie over hun artikelen op een website hebben geplaats, maar waarbij het niet mogelijk is om ook via deze website te kopen. Daarvoor moet men naar een fysieke winkel.

Primark is een voorbeeld een fysieke winkel met een informatiesite. Klanten kunnen zich dan in hun eigen omgeving oriënteren op de producten die worden aangeboden, en keuzes maken. Maar voor de feitelijke verkoop moeten ze dan naar één van de vestigingen van Primark.[2] De informatie die op de site staat is (veel) meer dan in een folder van een tuincentrum of iets dergelijks. Er staat informatie over vrij veel artikelen, in de vorm zoals men dat bij een webwinkel zou verwachten, zij het met minder tekst (alleen een korte omschrijving van de artikelen), maar wel met plaatjes van alle artikelen. Onduidelijk is welk deel van de collectie hier vermeld is. Hoewel Primark (op dit moment) geen webwinkel is, lijkt het erop dat zijn informatiesite met een internetrobot waar te nemen is.

Men mag verwachten dat de informatie op de informatiesite van Primark van goede kwaliteit is, maar misschien minder dan bij een webwinkel. De feitelijke verkoop vindt immers in fysieke winkels plaats. Voor een belangrijk deel wellicht aan klanten die zich helemaal niet eerst op de website hebben georiënteerd op het beschikbare aanbod. Het moet nagegaan worden over welk deel van de kleding informatie op de informatiesite te vinden is. Is het representatief voor wat in de winkel te krijgen is? Bevat het alle gangbare artikelen? Als dat zo is, lijkt de site een geschikte databron.

Omdat Primark een fysieke winkel is, is regionale waarneming (in principe) mogelijk[3]. Maar deze vorm van waarneming is duur en beperkt qua hoeveelheid te verzamelen data (frequentie van de waarneming en de hoeveelheid waar te nemen artikelen). Om die redenen is regionale waarneming van Primark niet aantrekkelijk en geen optie. Prijswaarnemers alleen maar vanwege Primark in dienst houden is ongewenst.

De kledinggegevens bij Primark zijn beschikbaar in een vorm die sterk lijkt op die van webwinkels. Wibra heeft inmiddels ook een informatiesite. Per kledingartikel is er een korte omschrijving (‘Dames T-shirt’), de prijs (‘€9,99’), wat informatie over samenstelling en beschikbare maten (‘65% polyester, 35% viscose, maten S t/m XXL’) en een foto. De beschikbare tekst-informatie is summier en vaak niet uniek voor een kledingstuk. De foto’s zijn belangrijke informatiedragers voor deze site. Daar is ook nog een folder te vinden; waarschijnlijk nog een overblijfsel van vroeger. Overigens is de kledingcollectie bij Wibra beperkt tot dames- en kinderkleding, in het goedkope segment. Onduidelijk is ook welk deel van de collectie die in de (fysieke) winkels te koop is, vermeld wordt op die site. Nader onderzoek zou nodig zijn om na te gaan of het zinvol is om een site als die van Wibra te gebruiken. Mogelijk is de informatie die er te vinden is te karig. En mogelijk is Wibra niet interessant genoeg voor de CPI.

2.1.4 Verzamelsites

[2] De vraag is hoe lang dit nog zo het geval zal zijn. Bij een andere goedkope kledingwinkel (Zeeman) kan wel via internet besteld worden, en is verzending gratis bij bestellingen vanaf 50€. Voor kleinere bestellingen betaalt men voor de verzendkosten (4,95€ zag ik als bedrag bij enkele producten).

[3] Uiteraard zou Primark daar dan wel toestemming voor moeten geven.

Er zijn echter ook nog verzamelsites die als toegangspoorten tot websites te gebruiken zijn[4]. Een zoektocht naar een geschikt kledingstuk kan starten op zo'n verzamelsite. Kopen van artikelen kan alleen bij de webwinkels waar men uiteindelijk terecht komt.

De verzamelsites bevatten informatie over artikelen (meta-informatie, prijzen en plaatjes), maar er is een kans dat artikelen niet goed ingedeeld zijn. Dit komt waarschijnlijk omdat de indeling met behulp van programmatuur is gemaakt die niet goed alle relevante gevallen onderscheidt. Zie Paragraaf 4.9 voor meer informatie over verzamelsites en de problemen die ze opleveren als ze als databron zouden worden gebruikt.

2.1.5 Productinformatiesites

Productinformatiesites sites bevatten (soms uitgebreide) informatie over producten. Het gaat dan bijvoorbeeld om elektronische en andere apparatuur, die in diverse winkels wordt verkocht. Zeer uitgebreide specs zijn vaak beschikbaar voor dit soort apparaten (smartphones, PC's, laptops, tablets, audiovisuele apparatuur, witgoed, e.d.). Deze sites zijn interessant vanwege de metadata (tekst en foto's) die ze bevatten, niet vanwege eventuele prijsinformatie. Deze informatie kan gebruikt worden als aanvulling op scannerdata, die vaak wat karig is ten aanzien van de beschikbare metadata. Anders dan de sites met prijsinformatie hoeven deze sites veel minder frequent waargenomen te worden. Immers zó vaak komen er geen nieuwe producten bij.

2.2 Gevolgen voor de verwerking

Uit het bovenstaande volgt dat men de verschillende typen informatiebronnen (webwinkels, informatiesites en verzamelsites) ieder op hun merites moet beoordelen. Dit heeft ook gevolgen voor de waarneming en de verwerking van de gegevens van dergelijke sites. Puur kijkend naar de kwaliteit van de informatie en de volledigheid ervan (wat de beschreven artikelen betreft) kan men concluderen dat van deze drie sites van webwinkels de beste, de meest volledige en meest actuele informatie leveren. Ze hebben daar immers alle belang bij. Dat betekent dat men de data van deze sites voor de verwerking niet speciaal hoeft te controleren op prijs e.d. Die informatie is doorgaans goed. Als er al een keer een prijs niet goed gespecificeerd is, zal die waarschijnlijk spoedig gecorrigeerd worden. Te midden van de grote hoeveelheden data die men verzamelt bij dergelijke sites zal zo'n incidenteel foutje geen enkele

[4] Voorbeelden zijn fashionchick.nl, kleding.nl en winkelstraat.nl.

invloed hebben op de prijsontwikkeling bij de desbetreffende website, laat staan de CPI voor kleding. Uiteraard heeft een webwinkel er belang bij om alle artikelen die men wil verkopen ook op de site te vermelden, en dus ook om deze informatie actueel te houden.

De informatie op informatiesites is in principe vrijblijvender dan die op de site van een webwinkel, omdat ze bedoeld zijn om te informeren en niet als verkoopinstrumenten. Potentiële klanten kunnen een beeld krijgen van het actuele aanbod. Zoals een folder van een supermarkt, drogisterij, tuincentrum, etc. inzicht geeft in een selectie van het aanbod. Bijvoorbeeld artikelen die in de aanbieding zijn, of die als lokkertjes worden gebruikt om mensen de winkel in te krijgen. Als dat zo is, is zo'n informatiesite wellicht niet geschikt als bron voor de CPI, omdat de beschreven deelcollectie niet representatief is voor het totale aanbod.[5] Maar als de verzamelsite een goed inzicht geeft in de gangbare artikelen van de winkel, dan is zo'n verzamelsite wel bruikbaar. Wat in een specifiek geval aan de hand is moet nader worden uitgezocht, bijvoorbeeld door navraag bij de desbetreffende winkel.

Informatie op verzamelsites kan niet zonder meer verwerkt worden, vanwege de kans op 'vervuiling', dus van artikelen die niet thuis horen bij de producten waar men in geïnteresseerd is, maar die daar te onrechte vermeld zijn, zoals lipstick of sneakers bij jacks. Deze vervuiling moet eruit gefilterd worden, en wel softwarematig. Dat zou op basis van tekst- of foto-informatie moeten gebeuren. Of dat mogelijk is, en wat in de praktijk het beste werkt, moet nader uitgezocht worden.

3. Internetrobots

3.1 Twee typen internetrobots voor de prijswaarneming

Het CBS gebruikt al enige jaren internetrobots om prijsinformatie van web shops te verzamelen. Dit type verzamelt geheel automatisch bulk informatie van websites. Er is nog een ander type internetrobot in gebruik. Dit werkt semi-automatisch en wordt gebruikt voor specifieke websites (bijvoorbeeld van bioscopen en autorijscholen) waar kleine hoeveelheden prijsgegevens te vinden zijn die bovendien niet sterk muteren. In dit geval is de robot een hulpmiddel voor de monitoring van specifieke websites. Hij wordt gebruikt om een analist snel inzicht te geven in wat er op de website is veranderd, indien dat het geval is. Voorheen werd dit werk gedaan door analisten die de benodigde informatie zelf van de desbetreffende websites haalden. Hiermee moesten ze dan beoordelen of er prijzen waren veranderd of niet. Een saai en foutgevoelig werk, waarvoor de inzet van zo'n robot te verkiezen is.

[5] Maar mogelijk weer wel als vooral de gangbare producten zijn vermeld.

We noemen dit interactieve type robot meer voor de volledigheid, niet omdat we er in dit stuk uitgebreid bij stil willen staan. Hier zijn we meer geïnteresseerd in web robots van het eerste type, die bulk data van het internet verzamelen. Tot nu toe zijn die gebruikt om prijs- én artikelinformatie te vergaren voor een aantal kledingwinkels. Deze informatie is dan weer gebruikt om rechtstreeks prijsindexcijfers mee te berekenen. Deze aanpak heeft echter bezwaren. Men zou verfijnder te werk kunnen gaan en niet alle data gebruiken maar een steekpoef hiervan. Door die in de tijd te volgen en te verversen heeft men een panel, op basis waarvan men prijsindices berekent. Dan heeft men beter onder controle welke producten men meeneemt. Deze aanpak lijkt meer op die welke in de regionale waarneming wordt gebruikt, behalve dat de omvang van de steekproef (veel) groter is.

3.2 Internetdata versus scannerdata

In vergelijking met scannerdata zijn internetdata (in zekere zin) inferieur, in die zin dat het hier om aanbodprijzen gaat en niet om verkoopprijzen. Men kent dan immers de prijs niet die voor een artikel daadwerkelijk is betaald. Bovendien weet men niet hoeveel van ieder artikel verkocht is in een bepaalde maand. Daar staat tegenover dat een winkel niet altijd scannerdata wil leveren aan het CBS. Bij internetdata speelt dat niet omdat ze publiek zijn.[6] Aangezien het echter om grote hoeveelheden data gaat die regelmatig[7] dienen te worden verzameld, is automatische verzameling en verwerking van deze gegevens geboden. Het 'scrapen' van websites is een andere manier van gebruik van de websites dan waarvoor ze primair bedoeld zijn, namelijk visuele inspectie door potentiële klanten. Er wordt veel meer informatie verzameld en men dient te zorgen dat de site niet wordt overbelast. Verder dient men terughoudend te zijn met het downloaden van gegevens van websites om te voorkomen dat de indruk wordt gewekt dat het om een hackeraanval gaat, en de kans bestaat dat de site 'op slot gaat' voor de robot.

Hierboven is gesteld dat in het algemeen scannerdata te verkiezen zijn boven internetdata. Maar in de praktijk kunnen scannerdata ook tegenvallen en kennen ze ook hun problemen. Dat begint al met de beschikbaarstelling. Een winkel kan weigeren deze data te leveren aan het CBS. Het CBS kan niet eisen dat men deze data levert. Als men een winkel al zover weet te krijgen om scannerdata te leveren dan kan het zijn dat de beschikbare meta-informatie met betrekking tot de artikelen te summier

[6] Overigens is maar de vraag of dat betekent dat het CBS hier met webrobots dagelijks data mag verzamelen. Juridisch gezien is dit een grijs gebied. In ieder geval is het netjes de desbetreffende webwinkel op de hoogte te stellen, en om toestemming te vragen. Anders zou de webwinkel actie kunnen ondernemen zodat de robot de toegang tot de site wordt ontzegd. Een goed contact met een webwinkel waarvan de site 'afgegraasd' wordt door een CBS-robot heeft sowieso voordelen. Men hoort dan immers eerder van voorgenomen (ingrijpende) wijzigingen op de site en wordt er niet onverwacht mee geconfronteerd.

[7] Op het moment nog dagelijks. Maar het is denkbaar dat die frequentie in de toekomst verandert. Prijzen van artikelen veranderen niet zo vaak dat dagelijkse waarneming noodzakelijk is. Verder komen er ook niet dagelijks veel nieuwe artikelen bij. Omdat artikelen, na hun introductie, vaak minstens één seizoen mee gaan, zouden ze ook niet gemist worden bij een minder frequente waarneming. Zie ook paragraaf 4.5.

is, voor het linken van soortgelijke artikelen of het gebruik in modellen (voor regressie of stratificatie). Mogelijk heeft de winkel zelf de data geaggregeerd waarbij essentiële informatie is verdwenen, bijvoorbeeld de locatie of het type winkel.[8] Daarnaast kunnen er problemen zijn met 'retouren', artikelen die door klanten (gratis) retour worden gestuurd. In de kledingbranche gebeurt dit op grote schaal. Voor de verkoop van kleding via internet is de optie dat men gekochte producten kan terugsturen van levensbelang, omdat men geen pasmogelijkheden heeft als in fysieke kledingwinkels.[9] Dit is niet zozeer een manco van de scannerdata zelf dan wel van het waarnemen en verwerken van de verkoopinformatie in de CPI, waarbij op maandbasis cijfers worden gemaakt en er geen revisies mogelijk zijn van eerder gepubliceerde prijsindexcijfers. In feite zijn retouren annuleringen van verkopen, die gecorrigeerd moeten worden. Het probleem is alleen dat zo'n correctie vaak pas kan worden uitgevoerd als de maand waarin de (zogenaamde) koop heeft plaatsgevonden al is afgesloten. Verder is maar de vraag hoeveel metadata een geleverd scannerdatabestand bevat. Ook dat kan tegenvallen.

3.3 Waarneemstrategie

In vergelijking met de regionale waarneming met behulp van prijswaarnemers gaat het bij internet bulk data, om grote hoeveelheden data die maandelijks van een webwinkel worden vergaard. Internetrobots verzamelen dagelijks grote hoeveelheden gegevens van artikelen (kenmerken en prijzen) en slaan die op. Het is echter niet gezegd dat al deze data ook moeten worden gebruikt bij het berekenen van de CPI. Het is meer een voorraad waaruit men kan putten bij het maken van de CPI. Het is voordelig de selectie niet door de webrobots te laten maken bij de dataverzameling, maar later bij de verwerking. Niet alleen is dat eenvoudiger waarnemen, maar het is ook aantrekkelijk om alles te hebben en hier vervolgens uit te putten. Dan kan men ook zien wat men niet gebruikt bij de berekeningen voor de CPI.

3.4 Verwerking

De verwerking van de verzamelde webgegevens betreft, onder andere: de selectie van artikelen, als de website meerdere soorten artikelen verkoopt dan waar men specifiek geïnteresseerd is, bijvoorbeeld ook schoenen en accessoires (tassen, riemen, e.d.); de interpretatie van de kenmerken die artikelen beschrijven (via automatisch coderen); de classificatie van artikelen (bijvoorbeeld volgens de Klcl voor kleding, een interne classificatie); de berekening van gemiddelde prijzen, berekening van prijsindices. Voor dit laatste kan men regressiemodellen gebruiken of op stratificatie gebaseerde methoden.

[8] AH bijvoorbeeld kent verschillende typen winkels, ieder met een eigen assortiment.
[9] Webwinkels hebben hier doorgaans ook geen problemen mee, omdat mensen die veel retoursturen vaak ook goede klanten zijn.

Bij dit geautomatiseerde verwerken van de internetgegevens kan van alles mis gaan, zowel bij de dataverzameling als de dataverwerking. Bij de dataverwerking bestaat de kans dat de website zodanig verandert dat de robot vastloopt. Of, subtieler, dat hij correct lijkt te werken, maar dat bij nadere inspectie toch niet gedaan blijkt te hebben. Bij het verwerken van de artikelbeschrijvingen (kenmerken) heeft men soms te maken met vaste teksten en soms met open tekst (omschrijvingen). Vaste teksten verwerken is relatief eenvoudig; open teksten verwerken daarentegen is lastiger en vergt kennis van textmining technieken. Hier heeft men te maken met een open vocabulaire dat branche-gericht is en vaak zelfs winkelspecifiek, het gebruik van, afkortingen, van synoniemen, van hypo- en hyperoniemen, etc. Indien de bedoeling is om zinnen te 'begrijpen' is het zaak de grammaticale structuur van de zin te bepalen door middel van syntaxanalyse. [10] Dat is een stuk lastiger.

De bevindingen tot nu toe met internetdata zijn dat fouten in prijzen of omschrijvingen van artikelen relatief zeldzaam zijn. Het is ook in het belang van een webwinkel om zowel prijs als de artikelomschrijving van een artikel correct, volledig en up-to-date op de site te hebben. Probleem daarbij is wel dat plaatjes de beschrijvingen vaak aanvullen. Een plaatje kan informatiever zijn voor een persoon dan duizend woorden. Dat is wat anders bij de geautomatiseerde verwerking ervan.

De bedoeling van dit stuk is om op een aantal zaken te wijzen die van invloed zijn op de efficiënte verwerking van internetdata. Omdat de waarneming van internetdata van vitaal belang is voor de kwaliteit van de uiteindelijk verkregen data, wordt die hier ook meegenomen. Het betreft verschillende aspecten die een rol spelen hierbij, zoals monitoring door internet robots ten aanzien van de kwaliteit van de verzamelde data, de te volgen waarneemstrategie[11], data editing, textmining, en automatisch coderen.

4. Onderwerpen

Hieronder worden enkele kwesties besproken die op dit moment spelen, of die kunnen gaan spelen bij het verzamelen van informatie op websites. In afzonderlijke deelparagrafen worden deze aan de orde gesteld. De volgorde van de aan de orde gestelde kwesties is niet willekeurig. Hoe eerder een kwestie aan bod komt hoe urgenter die is. Problemen die op het einde aan bod komen spelen om dit moment nog geen rol, maar mogelijk wel in de toekomst.

[10] Op een nog basaler niveau kan men problemen hebben met de karakterset waarin omschrijvingen zijn gespecificeerd. Dit kan problemen opleveren met letters met leestekens (é, è, ê ,ö, etc.). Maar dit soort computertechnische 'problemen' zijn relatief eenvoudig te ondervangen.

[11] Hoeveel data heeft men nodig? Hoe moeten die data worden verzameld: via slepen of zoeken? Hoe vaak moet men waarnemen?

4.1 Monitoring van internetrobots

Omdat websites kunnen veranderen, kunnen internetrobots vast lopen, of niet alle data op de site verzamelen. Indien dat gebeurt dient zo snel mogelijk te worden ingegrepen om te voorkomen dat de robot al te veel data van de desbetreffende site verliest. Bij dagelijkse waarneming is er wel enige uitval mogelijk, omdat de data veel redundantie bevatten (de prijzen van de meeste artikelen zijn op opeenvolgende dagen meestal hetzelfde. Bij ieder incident mag men verwachten dat de web scraper robuuster wordt. Maar een volgend incident kan nooit worden uitgesloten. En men kan ook niet anticiperen waar het volgende probleem zich zal voordoen. Dat heeft niets te maken met de competentie van de robotontwikkelaars, maar met het grote aantal mogelijkheden waar men veranderingen kan aanbrengen. Op al deze mogelijkheden anticiperen is niet mogelijk. Maar ook ernaar streven is niet gewenst, omdat men dan voor een hele hoop potentiële varianten de internetrobot robuust maakt, terwijl de meeste zich wellicht nooit zullen voordoen. Men mag wel hopen dat de web scrapers steeds robuuster worden voor allerlei veranderingen. Zekerheid hierover kan nooit worden verkregen. Onverwachte fouten en storingen kunnen zich altijd voordien. Hierop is niet te anticiperen.[12]

Om te voorkomen dat een web scraper gedurende een langere tijd disfunctioneert en inferieure of geen data waarneemt, is het van belang het gedrag van zo'n web scraper te monitoren. Dit dient zo te gebeuren dat tijdig kan worden ontdekt dat er iets mis is. Deze monitoring dient in essentie softwarematig te gebeuren. De vraag is dan: welke informatie gebruikt men als indicatoren voor het goed functioneren van een web scraper? Waarschijnlijk is vergelijking recent verzamelde data van belang. Als die goed zijn heeft men een ijkpunt. Maar welke indicatoren geven aan dat een website correct en volledig is waargenomen?[13] Uiteraard kan van een gegeven sec (een prijs of ander kenmerk van een item) niet zeggen of het correct is. Dat kan alleen als er meerdere gegevens zijn, bijvoorbeeld van eenzelfde artikel in andere winkels, of van eenzelfde winkel in een andere maand (niet te ver verwijderd van elkaar in de tijd). Men kan dan nog steeds niets zeggen over de correctheid van de gegevens, hooguit over de plausibiliteit.[14] Dat laatste zegt iets over hoe gegevens 'bij elkaar' passen. Daar zit vaak nog de nodige speling in.

[12] En als men het wel probeert te doen loopt men het risico veel gevallen te beschrijven die in theorie mis zouden kunnen gaan, maar die in die praktijk nooit blijken voor te komen. Dan doet men dus een hoop werk voor niets.

[13] Op dit moment worden gebruikt: aantal artikelen, aantal bezochte pagina's, aantal menu items. Verder wordt gecheckt of er kolommen zijn die leeg zijn of overal dezelfde waarde bevatten.

[14] De correctheid van een gegeven – stemt het overeen met de werkelijkheid – is problematisch. In strikte zin zelfs nooit vast te stellen. In de praktijk te bewerkelijk of te duur om vast te stellen. In de statistiekpraktijk kijkt men daarom liever naar onderlinge consistentie in plaats van naar correctheid. De lijkt de enige pragmatische manier. Maar het veronderstelt wel dat er vergelijkingsmateriaal is. Van data die met geen andere data te vergelijken zijn kan men niet zeggen hoe plausibel ze zijn.

4.2 Meta-informatie: tekst en foto's

De beschikbare meta-informatie op internet kan men ruwweg in twee stukken opdelen: tekstuele informatie en visuele informatie (plaatjes in de vorm van digitale foto's). In eerste instantie hebben we ons gericht op de tekstuele informatie. Deze is het gemakkelijkst toegankelijk en vormt een rijke bron. Momenteel wordt onderzocht of de visuele informatie nog interessante aanvulling biedt op de tekstinformatie.

4.2.1 Tekst

De informatie over de afzonderlijke artikelen is de informatie waar het CBS in geïnteresseerd is. Tot voor kort was dit uitsluitend tekstuele informatie (artikelomschrijvingen, ook wel aangeduid als meta-informatie, en prijzen). Plaatjes zijn recent in de scope betrokken als mogelijke bron van productinformatie. De tekstuele informatie is weer ruwweg in twee soorten te verdelen: gestructureerde informatie (variabelen en scores) en ongestructureerde informatie (min of meer vrije beschrijvingen van producten).

De meta-informatie over de artikelen is van belang in de verwerking van de productgegevens. Dit kan zijn om producten te kunnen classificeren. Hierbij komen gelijksoortige producten in eenzelfde klasse terecht, van een (vaste) classificatie.[15] Of men kan de informatie gebruik in hedonische modellen, die prijzen van producten relateren aan sommige van hun kenmerken.

In eerste instantie is alleen gekeken naar de gestructureerde informatie, omdat die gemakkelijker te verwerken is. Hiermee zijn al vrij snel eerste resultaten (prijsindices) verkregen op basis van groepsgemiddelden. Deze zagen er heel behoorlijk uit. De prijsindices gaven seizoenspatronen te zien. De gebruikte methode neemt automatisch nieuwe en verdwijnende producten mee.[16]

Vervolgens is onderzoek gestart om na te gaan of the ongestructureerde tekstuele informatie nog belangrijke informatie bevat die niet in de gestructureerde informatie aanwezig is. Het is een stuk lastiger om die informatie zoveel mogelijk geautomatiseerd te verwerken. Hiervoor is kennis nodig van gebieden als textmining, feature extraction, machine learning en computational linguistics. Het onderzoek heeft al wat eerste inzichten opgeleverd, maar is nog niet voltooid. Indien extra informatie 'win-

[15] Voor intern gebruik is een aparte kledingclassificatie gemaakt.

[16] Dit is verantwoordelijk voor het seizoenpatroon in de prijsindices voor de desbetreffende webwinkel. Als de berekening uitsluitend op identieke producten was gebaseerd had dit waarschijnlijke een dalende prijsindex te zien gegeven. Die kan zeker niet goed zijn. Maar de gebruikte methode heeft ook de eigenschap prijsveranderingen te genereren die louter gebaseerd zijn op aanbodveranderingen, zonder dat individuele artikelen van prijs zijn veranderd. Of dit effect ook speelt is niet onderzocht. Hoe (on)wenselijk dat is, is de vraag. Het gaat om aanbodcijfers niet om verkoopgegevens, en dus ook niet om veranderingen in verkochte aantallen. De vraag is of een veranderde samenstelling van het aanbod ook niet een verandering in de prijsindex mag laten zien.

baar' blijkt te zijn, moet nog worden nagegaan hoe deze extra informatie de resultaten (prijsindices) beïnvloedt, en of er betere prijsindices mee kunnen worden berekend.

De tekstuele informatie is van goede kwaliteit als het om webwinkels gaat.[17] Daar heeft zo'n winkel alle belang bij. Dit betreft zowel de prijzen als de artikelkenmerken. Die informatie moet immers kloppen en voor personen (potentiële klanten) begrijpelijk zijn.

4.2.2 Digitale foto's

Op dit moment wordt uitsluitend tekstuele informatie (in de vorm van kenmerken om omschrijvingen) en numerieke (prijzen) informatie verzameld van internet sites en verwerkt. Maar een deel van de informatie over artikelen is aanwezig in de vorm van plaatjes (digitale foto's). Die zijn weliswaar voor mensen gemakkelijk toegankelijk. Maar het is lastiger om hier softwarematig kenmerken uit af te leiden (feature extraction). Het kan zijn dat foto's informatie bevatten die niet in de omschrijvingen van de desbetreffende producten zijn genoemd. Als men in staat zou zijn dergelijke kenmerken uit zo'n foto af te leiden dan vormen de foto's een mogelijk interessante bron van productkenmerken.[18]

Ook zouden foto's gebruikt kunnen worden om 'verontreinigingen' bij verzamelsites op te sporen en te verwijderen. (Zie paragraaf 2.1.4) Bijvoorbeeld zouden de lipsticks of sneakers te midden van de jacks herkend kunnen worden en verwijderd. Wellicht is het gemakkelijker dit soort verontreinigingen te herkennen op basis foto's dan op basis van beschrijvingen. Onderzoek zal dat moeten uitwijzen.[19]

Een ander gebruik van digitale foto's is om er gelijke artikelen mee te vinden, bij eenzelfde webwinkel (in verschillende maanden) of bij verschillende webwinkels (in dezelfde of in dicht bij elkaar liggende maanden) als eenzelfde product op meer plaatsen wordt verkocht. Foto's fungeren dan als een soort sleutel. Men zou bijvoorbeeld de kleurverdeling over de pixels kunnen gebruiken, of een hieruit afgeleid getal[20]. Het zou ook mooi zijn als deze methode zou werken voor een (niet al te kleine) rechthoekige uitsnedes van foto's, of na bepaalde bewerkingen van de foto's zoals contrastvergroting, dithering, soft focus, etc.

[17] Althans uitschieters ziet men heel weinig. In de grote hoeveelheden data maakt een enkele uitschieter ook niet veel uit. Men hoeft daarom niet speciaal te controleren op uitschieters. Dat was bij de regionale waarneming anders. Daar kan één afwijkende waarde een grote invloed hebben op een (elementaire) prijsindex. Niet echt op de CPI, omdat het slechts één bijdrage is te midden van vele, die vast niet allemaal fout zijn.

[18] Waarbij nadere analyse moet uitwijzen of deze extra informatie heel verschillende uitkomsten oplevert.

[19] Hier spelen weer twee soorten fouten, van de eerste en van de tweede soort. Bij de ene worden verontreinigingen niet herkend en blijven ze dus ten onrechte achter. In het andere geval wordt iets als verontreiniging herkent, maar is dat feitelijk niet.

[20] Een soort hash code, die bijna altijd verschillende waarden oplevert voor verschillende kleurverdelingen, die op hun beurt al snel verschillend zijn voor verschillende foto's.

4.3 Kenmerken van artikelen, relaunches en koppelen

Met kenmerken van artikelen gaat het niet alleen om prijzen, maar ook attributen van artikelen (merk, materiaal, maakwijze, etc. bij kleding). En ook web-id's van artikelen. Die laatste definiëren een artikel ondubbelzinnig. Ze kunnen worden gebruikt om artikelen in verschillende (niet al te ver uit elkaar liggende) maanden aan elkaar te linken. Als de artikelpopulatie niet zou veranderen dan was dit de ideale manier om artikelen 'aan elkaar te koppelen'. Voor ieder artikel zou men het prijsverloop kunnen vaststellen. Met behulp van proxy-gewichten over de omzet zou men vrij eenvoudig een prijsindex kunnen schatten. Helaas is de werkelijkheid wat gecompliceerder. Artikelen komen en gaan.

Sommige artikelen zijn totaal nieuw in de zin dat er geen echte voorganger is aan te wijzen. Immers ooit was er een eerste radio, telefoon, typemachine, televisie, PC, laptop, smartphone, etc. Deze totaal nieuwe producten vormen aparte klassen in een productclassificatie. Als ze worden gesignaleerd moet de gebruikte classificatie er op worden aangepast. Maar dit soort ingrijpende innovaties is relatief zeldzaam.

De meeste nieuwe producten betreffen echter artikelen die een relatief marginale verbetering zijn van een bestaand product. En vaak is het dat niet eens het geval, en gaat het om een bestaand product dat in een nieuwe verpakking wordt aangeboden ('oude wijn in nieuwe zakken'). Zo'n product wordt beschouwd als een 'relaunch'. De bedoeling hiervan is om de verkoop te stimuleren ('er is iets nieuws om aan te prijzen') en meer te verdienen, door een de relaunch tegen een hogere prijs te verkopen. Als SKU's[21] bekend zijn is het eenvoudig om relaunches te koppelen aan hun voorgangers. Als dat niet zo is kan geprobeerd worden te koppelen op (secundaire) kenmerken. Dat koppelen dient overigens automatisch te gebeuren omdat het om grote hoeveelheden gaat. Het koppelen van dit soort data is een voorbeeld van een verwerkingsslag die plaats vindt als de data zijn waargenomen en niet tijdens de waarneming.

Deze toepassing is een uitdaging om de meta-informatie gestructureerd op te slaan zodat hiermee de koppelingen redelijk eenvoudig kunnen worden uitgevoerd. De bedoeling is deze informatie te gebruiken om relaunches statistisch te koppelen aan soortgelijke, bestaande producten. Dit levert dan langere prijsketens op die als basis kunnen dienen voor prijsindexberekeningen. Prijsstijgingen ten gevolge van relaunches blijven dan niet verborgen, maar worden in de prijsindexberekeningen meegenomen. Een ander gebruik van deze informatie is in hedonische modellen. De desbetreffende achtergrondkenmerken kunnen in hedonische regressies worden gebruikt. Overigens moet nader onderzoek bij iedere toepassing uitwijzen hoe bruikbaar deze detailinformatie feitelijk is, zowel bij het statistisch koppelen van artikelen als bij het gebruik in hedonische modellen.

[21] Stock keeping units (SKU's). Producten met eenzelfde SKU kunnen als 'soortgelijk' worden beschouwd..

4.4 Proxy-gewichten en zelfwegendheid

Internetdata zijn aanboddata en geen omzetdata. Bij internetdata heeft men geen omzetgegevens. Maar die zouden uit een andere bron verkregen kunnen worden. Misschien niet de gegevens over de verkopen voor de sites waarin men in is geïnteresseerd, maar voor de hele winkel (inclusief de verkopen in fysieke winkels), of van en soortgelijke winkel, of van de hele branche.

Daarnaast bestaat de mogelijkheid dat omzet toch gereflecteerd wordt in het assortiment. Dat kan het geval zijn als de diversiteit van een groep artikelen verband houdt met de omzet: hoe meer omzet hoe meer diverse de artikelgroep is samengesteld. Als men dan de prijzen waarneemt van de artikelen op de site neemt men automatisch gewichten mee die samenhangen met de omzet.

De vraag is dan vervolgens, hoe men dan maandprijzen voor artikelgroepen op websites moet berekenen. Alle waargenomen prijzen in één maand op een hoop gooien en deze dan gaan middelen? Maar dan zouden heel populaire artikelen een te laag gewicht krijgen. Men kan ook ieder artikel één keer meetellen in een maandportie. Dat is iets ingewikkelder omdat men dan artikelen in één maand moet 'ontdubbelen'. Dat vergt kennis van web-id's of andere identificatoren van artikelen waarmee men ze met elkaar in verband kan brengen.

4.5 Optimale waarneming

Optimaliteit van de waarneming kan op twee manieren worden beschouwd: optimaal voor het CBS en optimaal voor de berichtgever. Het eerste aspect heeft te maken met het maken van prijsstatistieken. De bedoeling is om hier de informatie vandaan te halen die nodig is voor het berekenen van prijsindices. Het tweede aspect betreft met de hinder die de webwinkel ondervindt als het CBS hun website 'afgraast'. Door data effectief en efficiënt te verzamelen krijgt het CBS wat het nodig heeft aan data, zonder dat de website onnodig belast wordt. Dat is gemakkelijkere gezegd dan gedaan. Het betekent dat een doorwrocht plan ('een waarneemstrategie') moet worden opgesteld om de benodigde data te vergaren en niet meer dan deze. De website moet niet onnodig belast worden. De waarnemingsstrategie moet ook rekening houden met uitval van de website, waarbij daarna nog een aantal keren gepoogd wordt de gewenste data te verzamelen.

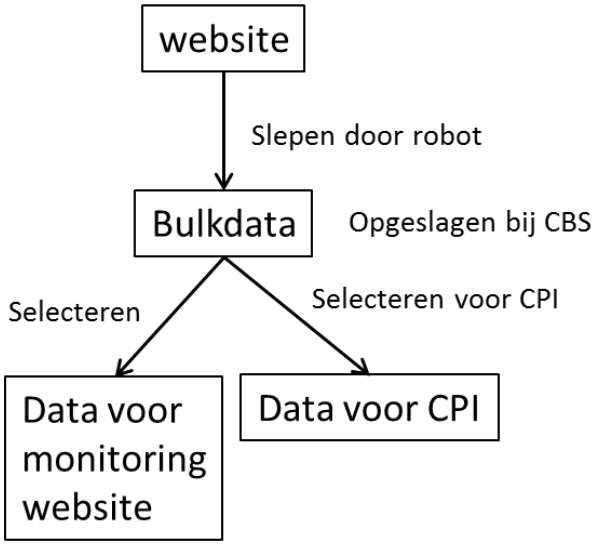
4.5.1 Data verzameld door internetrobots

Bij de huidige internetrobots die voor de CPI worden gebruikt, wordt als het ware een sleepnet gebruikt: alle beschikbare data worden meegenomen. In Figuur 1 is deze wijze van dataverzameling schematisch weergegeven. Achteraf wordt het (voor de CPI) bruikbare materiaal eruit gehaald, ontdubbeld, etc. Dit levert in een keer veel data op.

Al het verzamelde datamateriaal van de websites wordt opgeslagen, en kan ook dienen ter analyse. Men kan hiermee bijvoorbeeld nagaan of er nog nieuwe producten op de markt zijn gekomen.

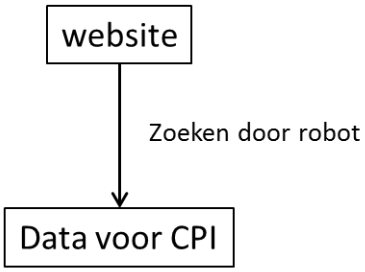
Omdat prijzen van artikelen niet dagelijks veranderen, en artikelen langere tijd bestaan (minstens een seizoen) kan men ook overwegen niet dagelijks waar te nemen, maar enkele keren per maand. De dagen dat men waarneemt wordt dan wél de sleepnetmethode gebruikt.

Een alternatieve waarnemingsmethode zou zijn om gericht waar te nemen, dat wil zeggen, op zoek te gaan naar prijsinformatie van specifieke producten. Dit wordt met 'zoeken' bedoeld. Zie Figuur 2, waar dit schematisch is aangegeven. Dit zoeken is van belang als men een mandje van producten heeft aangewezen waarvan men gedurende enige tijd de prijsontwikkeling wil volgen. Dit is te vergelijken met de werkwijze bij de reguliere waarneming. Alleen kan in het geval van internetdata het mandje groter zijn en kan er frequenter worden waargenomen. Indien een artikel ontbreekt dient een vervangend, soortgelijk artikel gevonden te worden.



Figuur 1. Data van een website en het gebruik ervan

Gezien de hoeveelheid te verwerken data dient dit automatisch te gebeuren. Met behulp van statistische koppelen kan men relaunches koppelen aan hun (mogelijke) voorgangers.



Figuur 2. Gericht zoeken van data op een website

De bedoeling van dit zoeken is met name om de website minder te belasten, dus om de responsdruk op die site te verminderen. Dat is vooral van belang voor de winkel, en heft voor het CBS weinig meerwaarde. Afhankelijk van de precieze inrichting kan het zelfs nadelig zijn, in die zin dat men minder goed in staat is om ontwikkelingen op de website te volgen. Een ander nadeel is dat deze wijze van data verzamelen complexer is dan de data met de sleepnetmethode te vergaren, en hier later uit te selecteren.

4.5.2 Efficiënte dataverzameling

Het lezen van een website kost vrij veel tijd. Vermeden moet worden dat alsof er een DOS-aanval[22] door hackers wordt uitgevoerd om de site 'plat te krijgen'. De 'leestijd' nodig voor web scrapers is een reden om doelmatig te verzamelen, niet meer dan nodig. Ook kan het reden zijn om de belasting van een website te verdelen over de tijd. Dat betekent dat men, als de sleepnetmethode wordt toegepast, de verzameling van de data over een maand wordt verdeeld. Op één dag de dameskleding, een week later de herenkleding en weer een week later de kinderkleding. Bij het waarnemen van een website is het bovendien van belang dag en tijd handig te kiezen. Bij voorkeur niet op tijdstippen dat de site wordt bijgewerkt of wanneer men veel geïnteresseerden en potentiële klanten op de site kan verwachten.

4.6 Welke data voor prijsindexberekeningen?

Los van het verzamelen van prijsinformatie van internet, is er de vraag van welke producten men eigenlijk het beste kan gebruiken voor de indexberekeningen in de CPI. Het is verleidelijk in eerste instantie om alle waargenomen prijzen daarvoor te gebruiken, dat wil zeggen van alle artikelen en/of van alle waargenomen dagen in een maand. Dat draagt echter het risico in zich dat men te veel ruis mee neemt (in de vorm van producten die atypisch zijn: producten die te weinig worden verkocht of die fungeren als lokkertjes of die uit goedkope restpartijen afkomstig zijn, etc.) of dat men verkeerde gewichten gebruikt voor de prijzen. Als men van ieder product alle waargenomen prijzen in die maand mee neemt[23] dan weegt men dus met het aantal keren dat een product is waargenomen door de internetrobot in die maand. Dat kan betekenen dat producten die niet erg veel verkocht worden een groot gewicht krijgen en producten die veel uitverkocht (en dus veel gevraagd) een laag gewicht; precies tegengesteld aan wat men zou willen. Men kan dus ook overwegen om van ieder product waar minstens één prijs van is waargenomen in één maand ook maar één prijs mee te nemen voor die maand.

[22] DoS attack = Denial-of-Service attack, is een poging om een machine of netwerk onbereikbaar te maken voor de beoogde doelgroep van gebruikers. Dit kan door een vloed aan verzoeken te versturen die de service doen crashen of geen ruimte geven voor andere activiteiten.

[23] Dit is zoals de zogenaamde groepenmethode die bij de desbetreffende webwinkel is toegepast.

Om dergelijke complicaties te vermijden kan men waarschijnlijk beter meer gecontroleerd te werk gaan, net als bij de regionale waarneming maar dan met een (veel) groter mandje. Men kan besluiten afzonderlijk producten weg te laten, of zelfs hele productgroepen, namelijk als die niet typisch zijn of zelfs afwijkend voor het kledingassortiment van de desbetreffende winkel. Dit gebeurt momenteel ook. Een interessant vraag is welk deel van de kleding men moet volgen om een goede prijsindex te kunnen berekenen. Enkele jaren geleden zijn twee scenario's onderzocht waarbij slechts een deel van alle kledingartikelen van de desbetreffende webwinkel is meegenomen in de berekening. Deze scenario's waren: top 80% per doelgroep (dames, heren of kinderen); top 50% per doelgroep.

Een andere kwestie betreft welke data in de tijd men nu precies moet gebruik om maandelijks goede prijsschattingen te maken. Omdat er veel redundantie is bij dagelijkse waarneming, zou men wel met data van minder dagen toekunnen. Daar is door de auteur en een collega al eerder naar gekeken. Toen is gekeken naar drie scenario's om met data van minder dagen te gebruiken (eerste drie weken van de maand, eerste drie vaste werkdagen van de maand, één dag per maand). Deze en soortgelijke scenario's zouden kunnen worden herhaald bij data van andere webwinkels. Niet dat dit onmiddellijk moet betekenen dat men per se ook minder data moet gebruiken voor de schattingen. Maar op deze manier krijgt men een beter gevoel voor de veranderlijkheid van de prijsschattingen en de redundantie in de data.

4.7 Internetrobots voor metadata

De internetrobots die tot nu toe zijn ingezet zijn vooral bedoeld voor het verzamelen van prijsinformatie. De desbetreffende websites fungeren aldus als bron van prijsinformatie en kunnen zo rechtstreeks gebruikt worden voor het berekenen van prijsindices.

Een andere mogelijk gebruik van internet is echter als bron van meta-informatie van producten. Als scannerdata onverhoopt te weinig metadata bevatten, dan zou internet de ontbrekende productinformatie kunnen leveren. Denk bijvoorbeeld aan elektronische apparatuur. Op basis van merk en typenummer kan men op internet vaak (zeer) gedetailleerde meta-informatie vinden Dit vergt dan echter wel speciale internetrobots, die nog ontwikkeld moeten worden.

Een aanloopprobleem zou kunnen zijn dat men artikelen aantreft waarvan men op informatiesites als tweakers geen omschrijvingen meer vindt omdat het desbetreffende artikel te oud is en niet meer verkocht wordt. Maar dit probleem zal na de aanloop vanzelf verdwijnen. Van artikelen die niet meer courant zijn heeft men eerder toch informatie verzameld en is deze dus gewoon beschikbaar.

4.8 Mogelijke problemen met prijsinformatie

4.8.1 Personalisatie

Een risico bij de waarneming van aanbodprijzen is dat prijzen (op één moment) voor sommige artikelen afhankelijk zijn van de persoon die op de website van de webwinkel in kwestie kijkt: of het een nieuwkomer betreft, iemand die al eerder gekocht heeft bij de webwinkel of zelfs een goede klant. Voor ieder type klant zou men een prijs kunnen bepalen, die eventueel ook nog beïnvloed wordt wat de klant op dat moment nog meer bestelt.

Als het fenomeen van gepersonaliseerde prijzen een grote vlucht neemt betekent dit dat internetdata als direct te gebruiken 'prijzenbron' ongeschikt wordt. Immers de prijs die een klant uiteindelijk betaalt is niet bekend via de website; het is een transactieprijs. Als de afwijking tussen getoonde en betaalde prijzen niet al te groot is, dan is het een beperkt probleem, en geven de getoonde prijzen een aardig beeld van de betaalde prijzen. Als die discrepantie groter is dan is er wel een probleem.

Modelmatig corrigeren kan alleen als men over data beschikt met betrekking tot gevraagde én betaalde prijzen, zeg op steekproefbasis.[24] Men zou men met behulp van modellen gemiddeld betaalde prijzen kunnen proberen te schatten. Maar het blijft een lapmiddel, gesteld dat men dat al zou kunnen toepassen. Anders kan men eigenlijk alleen nog maar zijn toevlucht nemen tot transactiedata. Maar die moeten er dan wel zijn (in de vorm van scannerdata). En als ze voorhanden zijn, hoeft men geen internetdata te gebruiken.

Dat prijzen steeds meer gepersonaliseerd worden is niet door een web scraper te bepalen. Dit vereist menselijke kennis, bijvoorbeeld die van een consumptieanalist. Dat vereist wel dat men ontwikkelingen bij webwinkels nauwkeurig volgt. Op internet kan men verwachten dat ontwikkelingen veel sneller gaan dan vroeger bij fysieke winkels.

4.8.2 Maatwerkkleding

Een nieuwe trend zou kunnen worden dat steeds meer betaalbare maatwerkkleding beschikbaar komt, in plaats van confectiekleding. Met internet lijkt dit een reële optie. Een klant specificeert via een sjabloon op de website van een webwinkel voor kleding (of bij die van ene kleermaker)[25] welk kledingstuk hij/zij wil hebben, welke stof gebruikt moet worden, welke zijn/haar maten zijn, welke versieringen moeten worden

[24] Als die integraal bekend zijn iedere maand dan kan met net zo goed alleen de transactieprijzen gebruiken.

[25] Die hoeft zich niet eens in Nederland te bevinden. Dit levert een nieuw aspect op voor wat de afbakening van waar te nemen webwinkels betreft. Tot nu toe nemen we alleen webwinkels in Nederland waar. Maar wat te doen als men in toenemende mate producten bestelt in het buitenland. Voor fysieke winkels is de beperking tot Nederland een natuurlijke afbakening. Voor webwinkels geldt dat eigenlijk niet.

aangebracht, etc. en geeft vervolgens opdracht om het te maken. Het kledingstuk wordt vervolgens gemaakt (ergens ter wereld) en vervolgens naar hem/haar opgestuurd. Van kleding die op deze manier gemaakt wordt ziet men op internet de prijs niet, omdat het om een uniek product gaat voor een specifieke klant. Prijzen en beschrijvingen van geproduceerde kledingstukken zal bij zo'n bedrijf moeten worden opgevraagd. Dat is een probleem als zo'n bedrijf in het buitenland zit. Maar mogelijk worden zulke diensten aangeboden via grotere kledingwinkels in Nederland, waar men ook ongewenste kleding aan kan retourneren.

Mocht het maken van betaalde maatkleding een hoge vlucht nemen, dan betekent dit waarschijnlijk dat een deel van de uitgaven aan kleding niet meer zichtbaar wordt. In ieder geval zullen de prijzen van deze maatwerkkleding niet op het web te vinden zijn.

4.8.3 Onzichtbare prijzen

Een ander probleem dat zich in dit verband kan voordoen is dat de prijs van een artikel niet getoond wordt voor kijkers op de site. Op de Amazon.com site heeft de auteur een artikel gezien (een werktuig) waarvan de prijs pas getoond werd zodra het artikel in het mandje was geplaatst. Het is onduidelijk of dit bij Nederlandse webwinkels ook voorkomt en of het een probleem is of kan worden voor bepaalde webwinkels. Als het slechts bij een enkel artikel voorkomt is het geen probleem en kan men dergelijke artikelen buiten de waarneming laten.

4.8.4 Mannelijke en vrouwelijke artikelen

In een recent artikel in de Washington post[26] blijkt dat er voor nogal wat artikelen (in de VS) waar een mannelijke of vrouwelijke versie van bestaat (bv stepjes voor kinderen in verschillende kleuren, rood voor jongetjes en roze voor meisjes) er (soms aanzienlijke) prijsverschillen zijn. Dit voorbeeld geeft aan dat het (soms) zaak is onderscheid te maken tussen beoogde doelgroepen van een artikel, die in dit geval samenhangt met het geslacht van de beoogde gebruiker.

Of dit soort fenomenen ook in Nederland voorkomen is de auteur niet bekend. Dat zou nader onderzocht kunnen worden. Mocht het voorkomen, dan is het van belang op het onderscheidende kenmerk te stratificeren. In het genoemde voorbeeld (ontleend aan het Washington Post artikel) is dit 'geslacht'. Maar het zou ook 'kleur' kunnen zijn, of een ander kenmerk waar dit verschil mee wordt aangeduid. Normaliter zou dit onderscheid tussen artikelen geen probleem hoeven zijn, als de onderscheidende informatie maar mee wordt genomen.

[26] "Why you should always buy the men's version of almost anything" van Danielle Paquette in de Washington Post van 22 December 2015. (Met dank aan Edwin de Jonge die de auteur heeft geattendeerd op dit artikel.)

4.9 Verzamelsites

Verzamelsites zijn websites waar wordt doorverwezen naar websites. Ze zijn ideaal om producten te zoeken in één branche (bv kleding) over vele winkels heen. Ze worden op dit moment niet als databron gebruikt, maar ze zijn daar mogelijk geschikt voor.

Het is zinvol verzamelsites en webwinkels te onderscheiden. Verzamelsites zijn een soort wegwijzers naar webwinkels; zij verkopen zelf niets. De informatie op de sites van webwinkels lijkt van betere kwaliteit te zijn. Dat is begrijpelijk: men heeft er alle belang bij dat de artikelen juist zijn omschreven, gegroepeerd en geprijsd. Correcte informatie van dit soort informatie is belangrijk voor de vindbaarheid van artikelen, de beslissing om aan te kopen en de aankoop zelf.

Bij verzamelsites is het geen probleem dat een paar foutjes aanwezig zijn, in de zin van verkeerd geclassificeerde artikelen. Dit is te beschouwen als verontreiniging van de data, ten gevolge van verkeerde (automatische) classificatie van goederen. Zo ziet men wel lipstick of een sneaker te midden van jacks, omdat het naam van sommige van deze artikelen de string 'jack' bevat. Voor een klant is dit soort 'verontreiniging' niet echt een probleem, hooguit oogt de website daarmee wat slordig. Een klant ziet onmiddellijk aan de naam of de foto dat het om een afwijkend product gaat dat niet thuis hoort in een bepaalde groep artikelen en skipt dat vervolgens.

Te veel producten die niet in een categorie thuishoren is dus geen probleem. Te weinig wel. Als in het voorbeeld de lipstick met het de string 'jack' erin nu te midden van de jacks staat en niet meer bij de lipstick, dan is er een flinke kans dat het product niet gevonden wordt door een klant. Gemiste artikelen in een groep ten gevolge van misclassificaties is dus wel een probleem voor een klant (en de winkelier, die omzet mis loopt).[27]

In het geval van internetrobots is dit soort vervuiling in zoverre een probleem dat de data in principe extra bewerkt moet worden, met het doel om de misclassificaties op de verzamelsite zoveel mogelijk te elimineren of te corrigeren. Dus in het gegeven voorbeeld zou de lipstick, sneaker of andere artikelen die er niet thuishoren) eruit gefilterd moeten worden. Maar dan moet wel eerst bekend zijn dat er vervuiling in de data zit, en welke precies.[28] Vervuilende artikelen zou op basis van beschikbare

[27] Dit is te vergelijken met fouten van de eerste en de tweede soort in de statistiek: Bij het toetsen van hypothesen: ten onrechte verworpen of ten onrechte aangenomen. Of bij het koppelen van records: ten onrechte gekoppeld, of ten onrechte niet gekoppeld.

[28] Deze kan soms eenvoudig gedetecteerd worden, namelijk als de prijzen van deze artikelen flink afwijken van die van de rest, en ze dus als uitbijters te beschouwen zijn. Men zou dan uitbijterdetectiemethoden kunnen gebruiken om deze fouten op te sporen. De artikelen die dan niet gedetecteerd worden vallen niet uit de toon bij de artikelen die in de groep thuis horen en veroorzaken weinig verstoring van de gemiddelde waarden, maar eventueel wel van varianties. Maar deze vervuilende artikelen behandelen als uitbijters is eigenlijk niet goed. Beter is het om de 'vervuiling' eruit te filteren.

meta-informatie verwijderd moeten worden of op basis van digitale foto's. Dat is in paragraaf 4.2.2 al aan de orde gesteld.

In het beste geval is de vervuiling klein en kan men de data verwerken alsof er geen sprake is van vervuiling. Dit levert slechts een kleine fout op, maar bespaart extra werk.

Of verzamelsites, ondanks hun vervuilde data, uiteindelijk aantrekkelijker zijn als databron is zonder nader onderzoek niet te zeggen. Ze zouden aantrekkelijk kunnen omdat ze maar één internetrobot vergen om de site te ontsluiten, als tenminste data op de verzamelsite zelf bruikbaar zijn, en doorlinken naar de onderliggende webwinkel niet nodig is.

De 'vervuiling' op verzamelsites zou softwarematig opgespoord en verwijderd moeten worden. Mogelijk kan dit op basis van fotoherkenning geschieden. Of dat mogelijk is moet nog nader onderzocht worden.

5. Conclusies

Enige jaren geleden is het CBS begonnen met het verzamelen van prijsinformatie van internet. Er zijn robots gemaakt om bulkinformatie – prijzen zowel als meta-informatie van artikelen - te verzamelen van sites van enkele webwinkels. In eerste instantie was het idee om zoveel mogelijk van het verzamelde materiaal te gebruiken, waarbij een simpele manier is gebruikt om prijsindices te berekenen. Dat alles om snel wat resultaten te kunnen berekenen. De resultaten zagen er goed uit, vergeleken met die gebaseerd op de traditionele methode (de regionale waarneming). Seizoenpatronen bijvoorbeeld waren in deze resultaten duidelijk te zien.

Maar het is duidelijk dat opnieuw moet worden gekeken naar de wijze van dataverzameling én het datagebruik. Mogelijk kan de data efficiënter worden verzameld, en is ook minder data nodig. Dit is op verschillende manieren te bereiken. Door van minder artikelen prijzen te verzamelen. Door de robot op minder dagen prijzen te laten verzamelen. Het idee hierbij is dat dan nog steeds de bulk- of sleepnetmethode wordt gebruikt.

Een andere optie is om gericht te zoeken naar de prijzen van specifieke artikelen. In het laatste geval zou dat kunnen betekenen dat men andere internetrobots moet schrijven. En ook dat een waarneemstrategie moet worden ontwikkeld. Of men kan nog steeds de sleepnetmethode te gebruiken om de data te verzamelen, maar dat deze niet in zijn geheel wordt gebruikt voor CPI-berekeningen. Hiervoor worden selecties gemaakt, van dagen en/of artikelen. Voor de prijsindexberekeningen en de monitoring van de informatie op de desbetreffende websites is dat prima, maar voor de vermindering van de responsdruk heeft die werkwijze geen betekenis. Welke de

beste[29] waarneemstrategie is (voor CBS én webwinkels) dient nog te worden uitgezocht.

Ook in een andere zin is het van belang de verzamelde internetdata beter te benutten. Er is productinformatie aanwezig in omschrijvingen en ook in foto's. Aan het uitbaten van de eerste soort data wordt al een tijdje gewerkt. Het verwerken van standaard omschrijvingen is eenvoudig. Lastiger is het gebruik van min of meer vrije tekst die soms in productomschrijvingen wordt gebruikt. Het ontsluiten daarvan is niet triviaal. Verder zit een deel van de informatie over producten in digitale foto's. Voor mensen is het niet moeilijk dit soort informatie te begrijpen. Voor computers is dit lastiger, maar niet onmogelijk.

Een interessante mogelijkheid, die nader onderzocht moet worden, is of digitale foto's gebruikt kunnen worden om 'vervuiling' te herkennen zodat die uit een door een robot verzamelde dataset gefilterd kan worden. Te denken valt hierbij aan verzamelsites die naast kleding ook andere producten bevatten, die echter niet altijd goed geclassificeerd hoeven te zijn.

In het geautomatiseerd verwerken van tekst- en foto-informatie op grote schaal zal de nodige aandacht besteed moeten worden. Daarbij speelt meteen het probleem, net als bij de robots die bulkinformatie verzamelen, hoe te monitoren dat de software naar behoren werkt, de goede data verzamelt en ook álle data verzamelt die het geacht wordt te verzamelen. Als dat mis gaat bij internetrobots betekent dat dat men van een bepaalde periode geen goede (bijvoorbeeld onvolledige) informatie heeft verzameld. En dat is een niet te repareren verlies als men daar pas na een tijdje achter komt. Als het verwerken van teksten of foto's niet goed gaat is dat minder rampzalig, als men tenminste beschikt over volledig bronmateriaal.

Voor het verrijken van scannerdata kan het internet ook een interessante bron zijn. Voor sommige producten (bv elektronica) is vrij gedetailleerde informatie te vinden op internet, die bovendien gemakkelijk te koppelen is met de producten in scannerdata (op basis van EAN-code of apparaat-code (merk, type apparaat, code/typeaanduiding volgens de fabrikant) Hiervoor moeten nog wel aparte internet robots worden gebouwd. De informatie die zij verzamelen verandert niet zo heel snel, en om die reden hoeven zij niet zo frequent te draaien als de robots die bulkinformatie verzamelen. In tegenstelling tot deze robots moeten zij gericht op zoek naar informatie, meta-informatie van producten.

Het is onduidelijk hoe de websites zich zullen ontwikkelen. Wordt de prijsinformatie steeds meer gepersonaliseerd, waarbij de prijs voor een (standaard) product afhangt van de klant? Indien dat het geval is, wordt de prijswaarneming via internet steeds minder aantrekkelijk, althans voor die segmenten waar ze wordt toegepast. Dat geldt ook als het kopen van betaalbare maatwerkkleding een hogere vlucht gaat nemen.

[29] Ook wat 'beste' precies is bij dit soort data, waar de webwinkel zeggenschap over heeft, ook wat de presentatie ervan op de website betreft. Die presentatie kan zo maar veranderen.

Websites zijn in deze gevallen niet bruikbaar meer als databron omdat ze domweg geen prijzen meer bevatten van kleding, hooguit van de kosten om bepaalde kleding te maken (prijs van de stofsoort per strekkende meter, prijs van bepaalde versieringen of van een bepaalde maakwijze, etc.) Als de bedrijven die dit soort kleding maken in het buitenland zitten is er een extra probleem voor het CBS ten aanzien van de waarneming.

Los van ontwikkelingen met betrekking tot personalisatie kan internet een uitstekende bron zijn voor aanvullende artikelinformatie in de vorm van scannerdata. Deze data blijven in principe het meest aantrekkelijk omdat ze ook transactieprijzen betreffen en hoeveelheden bevatten. Maar een winkel moet dit soort data wel willen leveren. En indien nodig, moeten ze zodanig zijn dat ze verrijkt kunnen worden met internetinformatie die er aan gelinkt kan worden door middel van exacte koppeling.

Naast personalisatie van prijzen kunnen er nog andere problemen zijn met prijzen op internet. Die kunnen ontbreken bij een eerste inspectie en pas getoond worden als een klant ze geselecteerd heeft en interesse lijkt te tonen. Of dit fenomeen vaak voorkomt (of zal komen) is de vraag. Maar indien het gebeurt moet de waarneming wellicht ook anders. Dan moeten artikelen eerst geselecteerd worden en de prijzen geregistreerd als ze in een mandje zitten. Maar gekocht worden ze zeker niet door het CBS. Het is voorstelbaar dat een webwinkel dat soort gedrag niet prettig vindt van een robot.

Tot slot zijn er nog informatiesites en verzamelsites die in potentie databronnen zijn voor prijsinformatie. Of dergelijke sites ook werkelijk aantrekkelijk zijn als bron moet nader worden uitgezocht. Ieder type site heeft zijn eigen (mogelijke) problemen. Voor informatiesites zou moeten worden nagegaan of ze representatief zijn voor de artikelen die de bijbehorende winkel verkoopt op ene bepaald moment, en hoe actueel de informatie is? Voor de verzamelsites gelden vragen als: welke bedrijven zitten op zo'n verzamelsite? Met andere woorden, zijn deze bedrijven representatief? Of kan men zich inkopen? Wat is de kwaliteit van de data op zo'n site? Hoeveel misgeclassificeerde artikelen zijn er gemiddeld? Hoe gemakkelijk zijn die misclassificaties op te sporen en te elimineren? Hoe stabiel zijn de sites? Dit soort vragen dient te worden beantwoord alvorens men kan besluiten om informatie- of verzamelsites te gebruiken als databron.

Verklaring van tekens

Niets (blanco)	Een cijfer kan op logische gronden niet voorkomen
.	Het cijfer is onbekend, onvoldoende betrouwbaar of geheim
*	Voorlopig cijfer
**	Nader voorlopige cijfer
2014–2015	2014 tot en met 2015
2014/2015	Het gemiddelde over de jaren 2014 tot en met 2015
2014/'15	Oogstjaar, boekjaar, schooljaar enz., beginnend in 2014 en eindigend in 2015
2012/'13–2014/'15	Oogstjaar, boekjaar, enz., 2012/'13 tot en met 2014/'15

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Vormgeving: Centraal Bureau voor de Statistiek, Studio BCO
Ontwerp: Edenspiekermann

Inlichtingen
Tel. 088 570 70 70, fax 070 337 59 94
Via contactformulier: www.cbsl.nl/infoservice