Centraal Bureau voor de Statistiek

Programma Impact ICT

Onderzoeksrapport nr. 8

Medegefinancierd door het Ministerie van Economische Zaken en Prima

Internetwaarneming van prijzen voor het maken van statistiek bij het CBS. Stand van zaken

Nico Heerschap en Olav ten Bosch

Aan: *Stuurgroep Impact ICT*
Datum: *14 maart 2013*

INTERNETWAARNEMING VAN PRIJZEN VOOR HET MAKEN VAN STATISTIEK BIJ HET CBS. STAND VAN ZAKEN

Samenvatting:

Deze notitie geeft een beeld van het tot op heden uitgevoerde onderzoek bij het CBS naar de mogelijkheden om internetwaarneming in te zetten voor prijzen van producten en diensten. Hoewel men positief is over de mogelijkheden om internetwaarneming in te zetten bij prijzen, bevindt onderzoek zich nog in een verkennende fase. Pilots richten zich op de prijswaarneming in de domeinen vliegtickets, benzine, bioscopen, rijscholen en kleding.

Trefwoorden:
Internetwaarneming, internetrobots, prijzen, prijzen hulprobot

Inhoud

1. Inleiding en achtergrond

Steeds meer activiteiten uit de “fysieke” wereld hebben zich op één of andere wijze ook naar het internet verplaatst. Daarmee is internet een belangrijkere bron van informatie geworden, ook voor het maken van statistiek.

Om gegevens van het internet te halen wordt gebruikt gemaakt van zogenaamde internetrobots of -crawlers. Een internetrobot is een programma dat het internet afzoekt naar de gewenste gegevens en deze opslaat. Deze gegevens vormen vervolgens de basis, mogelijk in combinatie met andere bronnen, voor het maken van statistiek. Het kan daarbij gaan om informatie van één of enkele websites of het doorzoeken van een deel van het internet.

Het eerste gebied waarop het CBS heeft geëxperimenteerd met de inzet van internetrobots is het domein van prijzen geweest. Daarbij ging het, in 2009, om prijzen van vliegreizen en benzine van onbemande tankstations. Omdat de resultaten van dit experiment bemoedigend waren is dit vanaf 2011 uitgebreid met onder meer prijzen van de domeinen bioscopen, rijscholen en kleding.

Het Programma Impact ICT is binnen het CBS begin 2011 opgezet om het gebruik van Internet als nieuwe databron (*IaD*), waaronder de inzet van internetrobots en smartphones, te stimuleren en te ondersteunen. Het Programma is mede gefinancierd door EZ en had vooral een aanjagersfunctie.

Het terrein van internetwaarneming van prijzen is niet specifiek in het Programma Impact ICT opgenomen [zie 4], omdat de financiering van deze werkzaamheden uit andere bronnen afkomstig is en omdat de werkzaamheden al waren begonnen voordat het Programma Impact ICT van start was gegaan. In het Programma Impact ICT is het terrein van internetwaarneming van prijzen echter wel als alternatief voorgesteld mochten de werkzaamheden op het terrein van internetrobots en de woningmarkt, in 2012 geen voortgang vinden. Dit laatste is niet het geval geweest. In het tweede kwartaal van 2012 is besloten om het internetrobots project Woningmarkt, dat gestart was in 2011, te continueren [zie voor de belangrijkste resultaten 5].

Dat betekent dat de belangrijkste bijdrage van het Programma Impact ICT aan het project internetwaarneming van prijzen is geweest de opzet van een noodzakelijke infrastructuur voor de inzet van internetrobots in het algemeen. Deze infrastructuur is een belangrijke randvoorwaarde geweest voor het überhaupt kunnen inzetten van internetrobots.

Vanuit het Programma ICT is afgesproken dat kort zal worden gerapporteerd over de werkzaamheden en voortgang van het terrein van internetwaarneming van prijzen. Er is namelijk veel belangstelling voor dit terrein van beleidsmakers en andere statistiek gebruikers, zeker niet in de laatste plaats door de aandacht die gegenereerd is door het Billion Prices Project van het Massachusetts Institute of Technology (http://bpp.mit.edu/). In dit project worden op basis van internetwaarneming dagelijkse inflatiecijfers gemaakt van een groot aantal landen.

2. Doel van notitie

Doel van deze notitie is dan ook om kort aan te geven welke werkzaamheden tot op heden binnen het CBS zijn uitgevoerd op het terrein van internetwaarneming van prijzen.

Deze rapportage is bijna geheel gebaseerd op de volgende notities:

- Hoekstra, R., O. ten Bosch en F. Harteveld, maart 2012, *Automated data collection from web services for official statistics: first experience*, gepubliceerd in The Journal of the International Association for Official Statistics (IAOS), 28, 2012, pag. 99-111;
- Bosch, O. ten, D. Windmeijer en R. Griffioen, mei 2012, *Internetwaarneming bioscopen en autorijscholen in de CPI*, intern CBS-rapport.
- Heuvel, G. van den, 16 november 2012, *Kleding en internetrobots*, intern CBS-rapport.

3. Uitgevoerde werkzaamheden en stand van zaken

3.1 Vliegtickets en benzine

In 2009 is bij het CBS gestart met een experiment om prijzen waar te nemen van vliegtickets en benzine van onbemande tankstations. Doel van de pilot was vooral om ervaring op te doen met het waarnemen van prijzen op het internet voor het maken van statistiek. De belangrijkste bevindingen van dat onderzoek zijn [zie 1]:

- dat het technisch goed mogelijk is om prijzen met behulp van internetrobots van internet te halen en op basis daarvan statistieken te maken;
- dat voor het ontwikkelen en onderhouden van internetrobots aparte vaardigheden nodig zijn en dat de te gebruiken software afwijkt van de standaardprogrammatuur;
- dat, als mogelijk, gestreefd moet worden naar zo generiek mogelijke tools of onderdelen. Dit wordt belangrijker naarmate het aantal websites waarop moet worden waargenomen toeneemt;
- dat er altijd sprake is van een keuze van het in-huis ontwikkelen van internetrobots of het uitbesteden van deze techniek aan een derde partij;
- dat de beste mogelijkheden, ten opzichte van de traditionele manier van waarnemen, liggen bij gebieden waarbij sprake is van het waarnemen van een grote hoeveelheid data op een beperkt aantal sites. Is er slechts sprake van een zeer beperkt aantal waarnemingen, zoals bij het incidenteel waarnemen van prijzen van vliegtickets, dan is het lastig een positieve business case te formuleren voor het bouwen en onderhouden van een gespecialiseerde internetrobot hiervoor;
- dat het waarnemen via internet van prijzen kan leiden tot een herevaluatie van de gebruikte methoden om inflatiecijfers en de CPI samen te stellen. Zo wordt in het geval van prijzen nu maandelijks op één specifiek tijdstip waargenomen. Op basis van die ene doorsnede in de tijd wordt het inflatiecijfer ge-

maakt. Bij het waarnemen met internetrobots kan 24 uur per dag en 7 dagen per week worden waargenomen. Hierdoor kan een beeld ontstaan van prijzen van producten en diensten per dag, per week of per maand waarbij de prijzen van het product of dienst over die periode heen (sterk) kunnen fluctueren. Dat is bijvoorbeeld het geval bij vliegtickets. Dit roept methodologische en theoretische vragen op hoe inflatiecijfers anders kunnen worden samengesteld dan de huidige methode op basis van enkele waarnemingen op één tijdstip per maand. Of hoe cijfers sneller en frequenter kunnen worden gepubliceerd;

- dat er op zich geen sprake is van juridische blokkades voor het maken van statistiek op basis van de op internet verzamelde prijzen. Er is ruimte voor statistiek en onderzoek als geen "substantieel deel" van een website wordt gespiderd[1].

Ondanks dat de conclusies van de pilot positief waren heeft men de volgende stappen van analyse naar verdere productie (nog) niet gezet. Daarbij hebben twee redenen een rol gespeeld. Ten eerste omdat de capaciteit heeft ontbroken om verdere analyses op de data uit te voeren en om de nieuwe methodologische en theoretische vragen te beantwoorden, maar belangrijker omdat één van de conclusies was dat het - door het kleine aantal waarnemingen bij de vliegtickets - in dit geval goedkoper was om handmatig de cijfers van het internet te halen dan de inzet en het onderhoud van de internetrobots.

Door de bemoedigende resultaten heeft het project wel een impuls gegeven om vanaf 2011 het aantal domeinen waarop onderzoek wordt gedaan naar internetwaarneming van prijzen op het internet uit te breiden. Dit betreffen de domeinen bioscopen en rijscholen en daarnaast kleding

3.2 Bioscopen en rijscholen

Doel van de pilot is om te onderzoeken in hoeverre het mogelijk is om robotgestuurde waarneming in te zetten bij het waarnemen van prijzen op internet bij bioscopen en rijscholen in plaats van de huidige telefonische waarneming. De bestaande methodologie blijft gehandhaafd. Baten van de pilot worden vooral gezocht in lastendrukvermindering en operationele voordelen, maar ook in de mogelijkheid om meer prijzen te kunnen waarnemen dan nu het geval is. Het gaat hier om een situatie waarbij weinig informatie moet worden verzameld van een relatief groot aantal websites (ca. 50).

De belangrijkste bevindingen van de pilot tot nu toe zijn [zie 2]:

- dat het geheel automatiseren met internetrobots van een groot aantal websites, inclusief het waarnemen, extraheren en vergelijken van de prijzen, en de opzet van een achterliggende infrastructuur geen goede optie lijkt. Gezien het aantal sites zou het beheer en configuratie van deze robots veel werk met zich mee kunnen brengen;

[1] Het begrip "substantieel" is een rekbaar begrip.

- dat als alternatief is gezocht naar mogelijkheden om de internetwaarneming te ondersteunen met meer generieke tools;
- dat daarvoor is gekeken naar een aantal scenario's en dat uiteindelijk wordt voorgesteld om een *prijzen hulprobot* te ontwikkelen en bij het waarneemproces in te zetten. Deze hulprobot ondersteunt de medewerker in die zin dat een relevant deel van de webpagina wordt vergeleken met een eerder opgeslagen versie. Als die aan elkaar gelijk zijn betekent dat dat er geen wijzigingen zijn en de eerder gedane waarneming van de prijs van kracht blijft. Is er wel iets gewijzigd dan vraagt dat om nadere analyse. Belangrijk hierbij is dat uit een analyse van de huidige waarneming blijkt dat er maar een relatief klein deel van de prijzen maand op maand verandert. Daarmee wordt de waarde van de inzet van zo'n hulprobot vergroot. Slechts een klein deel van de waarnemingen vraagt dan om nadere analyse;
- dat er een nulversie van de hulprobot is gebouwd en naar tevredenheid van de gebruikers is getest. Er wordt gestreefd naar een zo generiek mogelijk opzet van de hulprobot, zodat deze in een latere fase ook voor andere domeinen kan worden ingezet.

3.3 Kleding

Doel van deze pilot is om prijzen van kleding waar te nemen met een aantal kenmerken, zoals artikelnaam, artikelcode (een unieke id van het artikel) en - als afgeprijsd - de oorspronkelijke prijs. Gekozen is om te starten met de websites van twee aanbieders. Een belangrijk eerste doel is om te kijken welke dingen men zoal in de data tegenkomt.

De belangrijkste bevindingen van deze pilot tot nu toe zijn [zie o.a. 3]:

- dat het technisch goed te doen is om de prijzen en kenmerken van kleding van de website van de aanbieders te halen;
- dat de meeste waar te nemen kenmerken van kleding over de tijd heen goed stabiel blijven en ook goed te interpreteren zijn. Voor sommige kenmerken geldt dat minder. Zo is bij afgeprijsde artikelen niet altijd duidelijk hoe de oorspronkelijke prijs moet worden geïnterpreteerd, bijvoorbeeld als daadwerkelijke prijs, of gemiddelde prijs of als een adviesprijs.
- dat de frequentie van waarnemen een punt van discussie is. Een hoge frequentie van waarnemen, bijvoorbeeld één keer per week, levert beter inzicht in in de levensloop van artikelen. Dat geldt bijvoorbeeld voor artikelen, die maar een paar dagen op de site staan. Anderzijds levert het echter een grote hoeveelheid data op die minder makkelijk zijn te verwerken en te interpreteren. Minder vaak waarnemen, bijvoorbeeld één keer per maand, levert minder data op die gemakkelijker is te verwerken en te interpreteren, maar daardoor kunnen bepaalde ontwikkelingen worden gemist.
- dat sommige artikelen een soort knipperlichtpatroon vertonen. Ze staan voor een paar dagen op de site, verdwijnen dan om er een paar dagen later weer op te duiken.

- dat voortgang wordt gemaakt met methodologisch onderzoek. Zo wordt gekeken hoe tijdelijke uitval van robots methodologisch kan worden opgevangen bijvoorbeeld door extrapolatie van eerder waargenomen prijzen. Als uiteindelijk bij meer sites wordt waargenomen neemt de robuustheid ook toe, omdat de waarneming op de ene site de tijdelijk weggevallen waarneming op een andere site kan opvangen. Daarnaast wordt onderzoek uitgevoerd naar de mogelijkheden om artikelen te classificeren.
- dat het missen van volume data bij internetwaarneming mogelijk kan worden opgelost door – evenals bij de huidige traditionele waarneming – gebruik te maken van andere beschikbare bronnen. Het voordeel van scannerdata boven internetwaarneming maar zeker ook boven de traditionele wijze van waarnemen is dat daar wel sprake is van prijs- en volumegegevens.
- dat ieder domein van prijzen zo zijn eigen merkwaardigheden kent. Hetgeen betekent dat domeinkennis van groot belang is en blijft.
- dat het gewenst is om een lange reeks aan data te verkrijgen, minimaal 2 jaar, om valide conclusies te kunnen trekken over de mogelijkheden om internetwaarneming in te zetten bij prijzen.

4. Juridische aspecten

Op zich lijken juridische aspecten hier een minder belangrijke rol te spelen dan bij waarneming van andere informatie op het internet omdat het niet gaat om persoonsinformatie[2]. Om zo open mogelijk te zijn naar de buitenwereld is er wel sprake geweest van het opnemen van deze werkzaamheden in een melding aan de Functionaris Gegevensbescherming van het CBS en het uitvoeren van een Privacy Impact Assessment. Ook is informatie op de site van het CBS gezet over de wijze waarop internetwaarneming wordt uitgevoerd.

Meer dan pure juridische aspecten moet het CBS hier bedacht zijn op imagorisico's. Voorkomen dient te worden dat een website eigenaar het bezoek van statistische robots negatief uitlegt, terwijl het juist bedoeld is om de lastendruk te verminderen. Of dat het waarnemen via internet door het CBS negatief in de pers komt. Uitgangspunt bij de werkzaamheden is dat er wordt gestreefd naar een zo minimaal mogelijke belasting van de te spideren websites en een zo groot mogelijke openheid naar buiten toe. Het CBS staat open voor de suggesties van de website eigenaren.

5. Conclusies en voortzetting van het onderzoek

5.1 Conclusies

De inzet van internetrobots voor prijswaarneming ten behoeve van het maken van statistieken, zoals inflatiecijfers en de CPI, staat nog in de kinderschoenen. De uitgevoerde pilots bevinden zich nog in de fase waarbij gezocht wordt naar de beste technische oplossingen, onderzoek naar wat de data precies voorstelt, hoe daarvan

[2] Als het gaat om eenmanszaken is dit minder evident.

statistiek kan worden gemaakt en eerste stappen op het terrein van de ondersteunende methodologie.

Technisch is veel en ook steeds meer mogelijk. Een uitdaging is nog wel om te zoeken naar een efficiënte methode om waarneming te laten plaatsvinden van minimale hoeveelheden informatie verspreid over veel (verschillende) websites. Het waarnemen van veel informatie op één of enkele sites is prima te doen. Het wordt echter moeilijker naarmate het aantal te spideren sites zich uitbreidt, ook als het gaat om een beperkte hoeveelheid informatie. Het aantal robots moet navenant worden uitgebreid, waardoor het onderhoud en beheer van de robots sterk kan toenemen.

Zoals ook bij andere internetrobots projecten is geconstateerd, gaat het echter niet zozeer om de waarneming an sich, maar juist om de stappen daarna, namelijk het interpreteren en analyseren van de verzamelde data en het uiteindelijk maken van kwalitatief goede statistieken. Vooral bij het domein kleding worden de eerste stappen gezet op het terrein van de benodigde methodologie. Men is optimistisch over de mogelijkheden om op basis van internetwaarneming ook uiteindelijk cijfers te kunnen publiceren. De verwachting is wel dat dit in combinatie zal gaan met of als aanvulling op de traditionele wijze van waarnemen en het gebruik van scannerdata.

De pilot waarbij een hulprobot wordt gebouwd wijkt hier enigszins vanaf. Daarbij wordt de waarneming ondersteund vanuit het idee dat als het betreffende deel van de website niet is gewijzigd ook de waarneming en dus de prijs niet is gewijzigd. Er is in feite geen sprake van een inhoudelijke interpretatie of directe vergelijking van de data. Dat maakt de werkwijze een stuk gemakkelijker.

De inzet van internetrobots kan *in potentie* de volgende voordelen opleveren:

- nieuwe en snellere statistieken;
- meer detail;
- minder lastendruk;
- operationele voordelen.

Onderzoek bij het waarnemen van prijzen op het internet bij het CBS richt zich op dit moment vooral op lastendrukvermindering en operationele voordelen. In de iets verdere toekomst kan men ook denken aan onderzoek naar de mogelijke effecten van internetwaarneming op bijvoorbeeld bestaande methodologie of voor het samenstellen van onzekere, maar wel veel snellere statistieken[3]. Dat vraagt wel om een andere manier van statistiekmaken (statistische cultuur). Bij internetwaarneming is de immense hoeveelheid data, ook in de vorm waarin het wordt gepresenteerd, een gegeven en het startpunt van het maken van statistieken en niet de vooraf bedachte en opgezette vragenlijst van de onderzoeker.

[3] Bijvoorbeeld alleen gebaseerd op de prijzen van producten zoals die op internet worden aangeboden.

5.2 Voortzetting van het onderzoek in 2013 e.v.

Het onderzoek naar het waarnemen van prijzen op internet zal vooralsnog worden voortgezet op basis van de eigen resources van het CBS. Om daarin stappen te kunnen maken is het verstandig:

1. Om kennis zo veel mogelijk te bundelen en samenwerking met externen aan te gaan. Wat dit laatste betreft is het goed te melden dat samenwerking is gezocht en gevonden met statistische bureaus van andere Europese landen.
2. Onderzoek op basis van internetrobots bij het CBS, ook als het gaat om andere terreinen, zoals de woningmarkt [zie voor resultaten 5], vacatures [zie voor resultaten 6] e.d., te bundelen in een Internetdata servicecenter. Hiermee kan niet alleen de kennis en ervaringen worden gedeeld, maar ook het onderhoud en de wijze waarop internetrobots van het CBS worden ingezet op het internet.

6. Referenties

1. Hoekstra, R., O. ten Bosch en F. Harteveld, 2012, *Automated data collection from web services form official statistics: first experience*, gepubliceerd in The Journal of International Association for Official Statistics (IAOS)), 28, 2012, pag. 99-111;
2. Bosch, O. ten, D. Windmeijer en R. Griffioen, 2012, *Internetwaarneming Bioscopen en autorijscholen in de CPI*, mei, 2012, intern CBS-rapport
3. Heuvel, G van den, 2012, *Kleding en internetrobots*, 16 november 2012, intern CBS-rapport
4. Heerschap, N., 2011, *Programma Impact ICT 2012*, juni 2011.
5. Bosch, O. ten, en D. Windmeijer, 2013, *Eindrapportage onderzoek inzet internetwaarneming bij de woningmarkt*, 2011 en 2012, februari 2013.
6. Heerschap, N. en anderen, 2013, *Eindrapportage onderzoek inzet internetwaarneming bij de vacaturestatistieken 2011 en 2012*, maart 2013.