Het nut van data mining

Dr. Wojtek Kowalczyk, Vrije Universiteit in Amsterdam
(wojtek@cs.vu.nl)

Dr. Wojtek Kowalczyk

Dr. Wojtek Kowalczyk is werkzaam aan de Vrije Universiteit, waar hij onderzoek doet naar het minen van Data Streams, en les geeft in Data Mining Techniques. Hij is trekker in het Diana-project (Data Interception and ANAlysis), en heeft een aantal prijzen gewonnen op het terrein van patroonherkenning en machine learning.

Waar je ook heen gaat, wat je ook doet, je laat je visitekaartje achter… Echt!

Je staat op, doet je PC aan en… je internetprovider weet dat je online bent. Vervolgens wordt elke klik in je internetbrowser in meerdere computers opgeslagen. Elke email die je stuurt of ontvangt wordt bij je internetprovider geregistreerd. Elk document, plaatje, liedje of filmpje dat je van het internet plukt kan onderschept en geanalyseerd worden. Heb je een mobieltje? Is die aan? Dan wordt je locatie constant gemonitord met een nauwkeurigheid van enkele honderden meters. Bel je een vriendin? Uiteraard worden dan haar telefoonnummer, haar locatie en de duur van het gesprek ook geregistreerd. Heb je je bonuskaart of air-miles kaart gebruikt? Oeps, nu weten ze precies wat je hebt gekocht, voor hoeveel en hoe laat dit voorval zich geschiedde. Ben je de gelukkige eigenaar van een auto? Kijk eens om je heen: door hoeveel camera's word je eigenlijk bekeken? En wat te denken van al die sensoren die veilig in het wegdek opgeborgen zitten? Ga je misschien met het openbaar vervoer? Wees gerust! De nieuwe OV-chipkaart zal genoeg gegevens over je bij elkaar sprokkelen: elke sensor die jouw kaart kan "zien" zal een kort berichtje naar de centrale sturen: "kaart X gezien om Y uur door sensor Z"… Op deze centrale kan dan direct "X" worden gekoppeld aan de eigenaar van de kaart, aan jou dus. Ja, het is waar: elke persoon in Nederland staat geregistreerd in meer dan 1000 databanken! Denk aan de gegevens die ze van je hebben bij je bank, de winkels die je regelmatig bezoekt, je videotheek, je school, universiteit, enzovoort, enzovoort. Het "Big brother is watching you"-tijdperk lijkt snel in zicht te komen; misschien wel sneller dan menigeen zou wensen. Wat weten 'ze' eigenlijk allemaal van je? Een enge gedachte om zo even bij stil te staan, nietwaar? Maar er is ook een positieve kant aan deze onuitputtelijke stroom van informatie die je in je alledaagse leven produceert: deze kan met behulp van krachtige algoritmen worden geanalyseerd om je zo te helpen betere keuzes te maken, je te verdedigen tegen mensen met kwade bedoelingen, enzovoort. Oftewel: om je leven beter te maken!

In dit artikel zullen we drie toepassingen van data mining bekijken: fraudedetectie, dynamische verkeersregeling, en *recommender systems*.[1]

Fraudedetectie

Creditcards maken ons het leven gemakkelijker. We gebruiken ze om er mee te betalen in het restaurant, vluchten en hotels mee te boeken, van alles te kopen op internet en wat nog meer niet. Ze zijn bijzonder gemakkelijk in de omgang: in sommige gevallen is het kaartnummer en de verloopdatum al voldoende om een transactie te doen. Het is dus niet verassend dat er gemakkelijk mee valt te frauderen. Wist je bijvoorbeeld dat er alleen al in Groot-Brittannië elke 9 seconden wordt gefraudeerd met een creditcard? En dat daar elke dag ongeveer 2 miljoen euro verloren gaat door creditcardfraude? En dat bijna één op de drie creditcardgebruikers wel eens door creditcardfraude is getroffen? Maar toch worden creditcards als veiliger dan contant geld beschouwd. Gemiddeld zijn slechts 0.1% van alle transacties frauduleus – dat is één fraudegeval per duizend normale transacties.

De helden die achter deze lage misdaadcijfers zitten zijn 'slimme' systemen die permanent alle transacties (in real-time) bekijken en deze automatisch blokkeren als ze onraad ruiken. Ze kunnen honderden transacties per seconde behandelen; ze leren automatisch nieuwe fraudepatronen herkennen gebaseerd op transacties die in het verleden frauduleus bleken te zijn en ze produceren kwalitatief zeer goede fraudewaarschuwingen. Sommige bedrijven, zoals Fair Isaac (*www.fairisaac.com*), claimen dat hun fraudedetectie-oplossingen de fraudegevallen met meer dan de helft kunnen reduceren. Je kunt je dus wel voorstellen hoeveel geld je kunt besparen met zulke systemen. En dat niet alleen in Groot-Brittannië.

Het principe achter deze slimme systemen is verbazingwekkend simpel. Voor elke kaart houden ze een vector bij met daarin getallen die enkele typische karakteristieken representeren van het gebruik van de kaart: bijvoorbeeld het gemiddeld aantal transacties per week, de kenmerkende tijd en plaats van transacties en de gemiddelde transactiebedragen. Deze vector, die soms profiel of signatuur wordt genoemd, wordt continu vergeleken met de meest recente transacties: voldoen ze aan het profiel of niet? Als de verschillen tussen de transactie en het profiel van de kaart te groot zijn, kan de kaart worden geblokkeerd of kan de bank worden gewaarschuwd. Het is duidelijk dat het profiel zich moet aanpassen in het geval dat de kaarthouder zijn gedrag geleidelijk aanpast. Hoe kunnen we bepalen hoe groot een mismatch tussen twee transacties is? Dat kan bijvoorbeeld met behulp van neurale netwerken die getraind zijn op honderden miljoenen frauduleuze

1. Zie voor meer informatie het artikel over recommender systems van Rory Sie. (Red.)

2. Bovenstaande getallen zijn gebaseerd op gegevens gepubliceerd op www.apacs.org.uk

en niet-frauduleuze transacties.

Hoewel het idee om profielen te gebruiken voor fraudedetectie erg simpel en aantrekkelijk blijkt, is het vinden van goede profielvariabelen een ware kunst. Vanzelfsprekend worden de belangrijke details van moderne fraudedetectiesystemen geheim gehouden. Meer informatie over het bouwen van profielen (in de context van telecomfraude) is te vinden in artikel [1].

Dynamische Verkeersregeling

Niemand vindt het leuk om zijn tijd in de file te verkwisten. In 2005 was de totale 'filedruk' in Nederland meer dan 10 miljoen kilometerminuten, ofwel gemiddeld een permanente file van ruim 19 km het hele jaar door, dag en nacht. Natuurlijk zijn er plannen om de situatie te verbeteren door meer asfalt te leggen en rekeningrijden in te voeren. Maar er is nog een mogelijkheid: slim gebruik maken van de vele gigabytes aan verkeersdata!

Laten we ons een systeem indenken dat in real-time de data van alle auto's inzamelt: hun locatie en hun bestemming. Verder verzamelt het systeem alle beschikbare informatie over de huidige situatie op de weg: wegwerkzaamheden, ongelukken op de weg, het weer en de files. Met al deze gegevens kan het systeem voor elke auto een optimale snelheid en route geven, gebruikmakend van bepaalde statistische filemodellen (het ontstaan van een file is een 'toevallige' gebeurtenis, waarbij de kans een functie van zowel de wegcapaciteit als de drukte op de weg is). Om een voorbeeld te geven: een auto die om zeven uur 's ochtends uit Den Bosch zou vertrekken met als bestemming Amsterdam, zou kunnen worden opgedragen 92 km/u te rijden om de kans op een opstopping bij Breukelen om half negen te minimaliseren. Deze 'lokale plannen' zouden elk moment kunnen worden aangepast aan de actuele situatie. Op deze manier zou de doorstroming worden geoptimaliseerd en de totale verspilde tijd in de file aanzienlijk worden geminimaliseerd.

Vandaag de dag lijkt zo'n dynamische, datagedreven oplossing misschien meer op een science-fictionverhaal dan op een werkelijke oplossing. Toch zijn er recentelijk vele pogingen gedaan om een globaal verkeersoptimalisatiesysteem te bouwen. Zo bestaat er het REACT-project (*Realizing Enhanced Safety and Efficiency in European Road Transport*, *www.react-project.org*), gesponsord door de Europese Unie, wat als doel heeft om de efficiëntie van ons weggebruik te verbeteren en het aantal verkeersslachtoffers tot een minimum te beperken. Een mooi overzicht van de meest recente initiatieven voor een dynamisch verkeerscontrolesysteem is te vinden in artikel [2].

Recommender systems

Zou je een geweldige DVD willen zien? Mag ik er eentje aanbevelen? Klinkt te mooi om waar te zijn. Ten eerste zou ik, om je een goed advies te kunnen geven, moeten weten wat voor soort films jij leuk vindt en wat voor soort niet. Ten tweede zou ik moeten weten welke films je al gezien hebt. En bovenal zou ik, ten derde, alle bestaande films al moeten kennen om jou degene die je waarschijnlijk het leukst zou vinden te kunnen aanbevelen. Het moge duidelijk zijn: geen mens is hiertoe in staat. Maar… in de VS is er een DVD-verhuurbedrijf, *Netflix*, die werkelijk voortreffelijke aanbevelingen kan doen. Hoe doen ze dat? Wel, ze hebben 5 miljoen actieve klanten die niet alleen DVD's huren, maar Netflix vervolgens ook vertellen of ze hem leuk vonden of niet – dit doen ze door de film op een schaal van 1 tot 5 in te delen. Op deze manier verzamelt Netflix elke dag 2 miljoen nieuwe ratings. Sinds 1997 hebben ze 1,4 miljard ratings verzameld. Bovendien houdt het bedrijf een database bij met karakteristieken van de ongeveer 70.000 films die ze bezitten; waaronder het genre, hoofdrolspelers, regisseur en de premièredatum. Vervolgens analyseren ze hun data elke dag met een stel krachtige data mining-algoritmen om hun klanten elke dag zo goed mogelijke aanbevelingen te kunnen doen – één miljard per dag. Hoe nauwkeuriger hun aanbevelingen zijn, hoe tevredener klanten ze hebben en hoe meer DVD's ze verhuren. Hoe meer feedback ze krijgen, des te betere aanbevelingen kunnen ze doen. Zo simpel is het. Een echte *perpetuum mobile*. Het hele ratingsysteem heet *Cinematch.*[3]

Hoewel de aanbevelingen die Netflix maakt al zeer goed zijn, wil het bedrijf natuurlijk graag weten of hun algoritmen nog verbeterd kunnen worden. In plaats van hulp in te schakelen van data mining-experts, heeft Netflix simpelweg 100 miljoen ratings op internet gezet en $1.000.000 uitgeloofd voor de eerste persoon (of team) wat erin slaagt hun oorspronkelijke ratingsysteem met minstens 10% (in nauwkeurigheid) te overtreffen. Aan deze oproep is massaal gehoor gegeven. Binnen vier maanden (de competitie is in oktober 2006 van start gegaan) hebben zich meer dan 15.000 teams aangemeld. Het beste resultaat tot dusver (daterend van 31 januari 2007) is een verbetering van 6.75%, dus de hoofdprijs wacht nog steeds op een winnaar! Misschien iets voor jou? Ga naar *www.netflixprize.com* en check it out! ⌀

• *[1] Cahill, M. H., Lambert, D., Pinheiro, J. C. and Sun, D. X. (2000). Detecting fraud in the real world, Technical report, Bell Labs, Lucent Technologies. http://citeseer.ist.psu.edu/cahill00detecting.html*
• *[2] Traffic Planing and Logistic. The Special Theme of the ERCIM News 68, January, 2007. http://ercim-news.ercim.org/images/stories/EN68/EN68.pdf*
• *[3] Linden, G., Smith, B. and York, J. (2003). Amazon.com Recommendations. Item-to-Item Collaborative Filtering. http://www.win.tue.nl/~laroyo/2L340/resources/Amazon-Recommendations.pdf*

3. Je kunt meer te weten komen over Cinematch door naar http://blog.recommenders06.com/?p=35 te surfen en te kijken naar de presentatie van Jim Bennet – de hoofdfiguur achter het hele systeem. Ook kun je een blik werpen op een korte klassieker over het aanbevelingssysteem van Amazon.com, zie artikel [3].