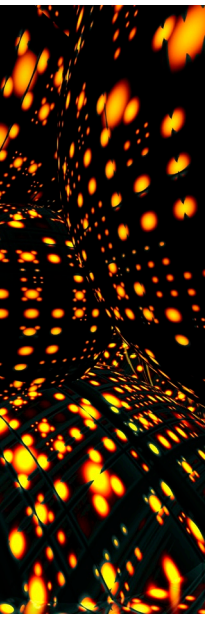

Human Inference en de kunst van het intelligent interpreteren

Het Gouden Record

Hans Lamboo

Al sinds 1986 is het Nederlandse bedrijf Human Inference actief op het gebied van datakwaliteit. Het behoort inmiddels tot de wereldtop als het gaat om gespecialiseerde software die de gehele Data Quality Lifecycle kan ondersteunen. De filosofie achter Human Inference is even simpel als geniaal: het menselijk redeneervermogen vormt het fundament van de suite.

Het in een prachtig pand op een fraaie bosrijke locatie in Arnhem gehuisveste Human Inference onstond in 1986 toen twee medewerkers van de Goudse verzekeringen op zoek waren naar een manier om fraude met claims te ontdekken. Ze begrepen niet dat, hoewel ze toch met grote stelligheid wisten dat een zekere claim hun bureau al eens was gepasseerd, deze door de computers nergens kon worden gevonden, hoezeer ze ook hun best deden het juiste algoritme te vinden.
Principal advisor en taalkundige Holger Wandt heeft nog steeds bewondering voor de benadering die toen is gekozen. "Het is interessant vanuit de taalkundige sfeer, maar het is vooral de manier waarop gegevens worden geïnterpreteerd die uiterst boeiend is. Wat is wat in de database? Een korte maar complexe vraag. Die vraag heeft Human Inference als het ware geïmplementeerd in een machine." Als voorbeeld daarvan schrijft hij 'Arend Tromp' op een stuk papier, iets dat een mens direct herkent als een naam. "Om precies te zijn: een voor- en achternaam," zegt Wandt. "Zet ik tussen beide woorden een &-teken: 'Arend & Tromp' dan interpreteert onze software 'Arend' als familienaam. Maak ik er vervolgens 'Smit, Arend & Tromp' van dan vertellen onze hersens ons dat het om een bedrijfsnaam gaat en dan waarschijnlijk ook nog om een accounts- of advocatenkantoor. Dat is nu precies zoals onze software werkt: net als het menselijk redeneervermogen. Het is de kunst van het intelligent interpreteren."

Lifecycle

De net benoemde CEO Winfried van Holland is verre van een nieuw gezicht bij Human Inference, want hij werkt er al jaren. Hij kent de geschiedenis door en door. "In het begin zonden we een stukje software naar de klant. Die moest de codes in het mainframe hangen en meecompileren. Vervolgens moesten alle settings en thresholds met de hand worden aangegeven. De volgende fase was dat we een product bouwden dat niet meegecompileerd of ingelinkt hoefde worden, maar stand-alone kon werken."
Snel daarna ontstond al de gedachte aan een datakwaliteitlifecycle. Het zoeken naar dubbele records geeft bijvoorbeeld aanmerkelijk betere resultaten als de data eerst zijn gestandaardiseerd. "Zorg dat postcodes, huisnummers, telefoonnummers op een uniforme manier in de databases staan. Zorg dat de namen allemaal met een hoofdletter beginnen en hanteer een vaste volgorde voor voornaam, achternaam, tussenvoegsels, titels enzovoort. Daar hebben we inmiddels verschillende modules voor gebouwd die onderdeel uitmaken van onze suite."
De tweede stap is de afhandeling van een match. Twee of meer records die boven de threshold komen moeten worden vergeleken en zo mogelijk samengevoegd tot één juist record. "Het Gouden Record noemen we dat," zegt Van Holland. "Dat is zeker niet triviaal, want het is niet vaak een record dat het juiste is. Meestal bevat elk record wel één van de juiste gegevens. Maar welke? Het record uit de financiële administratie bevat waarschijnlijk een juist bankrekeningnummer, de marketingdatabase het goede telefoonnummer, het CRM-systeem de correcte postcode. Waarschijnlijk. Je kunt niet zomaar Het Gouden Record samenstellen. En al helemaal geen records wissen, zeker niet in grote ERP-systemen."

De Data Quality Lifecycle (DQL) start meestal met de Inspectiefase. "Je steekt als het ware de thermometer in de databestanden die je hebt en stelt een diagnose over de kwaliteit," zegt Wandt. "Vervolgens kom je in de dynamiek van de DQL terecht. Met dataprofiling kun je vrij gemakkelijk naar voren halen waar inconsistenties in de data zitten, die vrij eenvoudig gewijzigd kunnen worden. Je ontdekt bij profiling ook bijvoorbeeld expliciet verkeerde geboortedatums, bijvoorbeeld omdat 6 procent van de populatie

in de database geboren zou zijn op 1 januari 2001 – oftewel 01-01-01. Dit soort omissies wordt in deze fase herkend. Dan ga je de definitieve cleansing fase in, waarbij je eerst een oplossing definieert om de bestaande databestanden op te schonen door te standaardiseren, valideren, ontdubbelen, mergen en verrijken. Vervolgens definieer je de oplossing om de opgeschoonde databestanden schoon te houden. Daarvoor passen we ons principe van de Data Quality Firewall toe. De nieuwe data komen immers via allerlei kanalen binnen. Na de cleansing ga je de resultaten meten en vergelijken met de nulmeting in de assessment fase. Daarna creëer je een feedback loop naar het bestaande databestand."
De DQ-suite van Human Inference bestaat uit modules voor Profiling, Transformatie, Cleansing, Matching, Merge & Enrich en tenslotte Rapportage. De suite koppelt zo alle fases van de DQL aan elkaar en biedt als het ware een generieke datakwaliteitsstrategie.

Fouttolerantie

Het lijkt zo voor de hand te liggen: slechte data kosten geld. "Toch is het nog steeds geen gemakkelijke zaak om bedrijven te overtuigen van de business case," zegt Van Holland. "Mensen moeten zelf doorhebben dat er iets mis is met hun data. Dan is het sommetje gauw gemaakt."
Hoever kun je gaan met datakwaliteit? Dat begint volgens Wandt met postcode, straat en huisnummer te checken. "Hoewel, er zijn landen – Ierland bijvoorbeeld – waar men niet eens een postcode kent. Maar je kunt vele malen verder gaan dan dergelijke checks. De eerste vraag die je moet stellen: wat is de bedoeling van de acties die je doorvoert om de kwaliteit te verhogen? Wat wil je bereiken? Dat is de *fit-for-use* definitie van Joseph Juran: data zijn van kwaliteit als ze voldoen aan de verwachtingen van de gebruiker en de ontvanger.

> Voor het ontdubbelen van bestanden dien je echter alle data in handen te hebben om te kunnen vergelijken

Met andere woorden: wil je als postorderbedrijf zorgen dat de catalogi die je uitstuurt allemaal op het juiste adres terechtkomen bij de mensen voor wie het bedoeld is? Dan moet je zorgen dat de actualiteit van je adresgegevens hoog is en dat de koppeling tussen naam- en adresgegevens klopt. Maar als je als directeur van een bank of verzekeraar verplicht wordt gesteld om te controleren dat je geen zaken doet met bedrijven of personen die op de een of andere suspectlist staan, dat is van een hele andere orde. De gegevens op die blacklists staan zijn vaak rudimentair, incompleet of verminkt. In de enorme informatiestroom die plaatsvindt bij een financiële instelling moet je toch vast gaan

Holger Wandt: "Wat is de bedoeling van de acties die je doorvoert om de kwaliteit te verhogen?"

stellen of er een potentiële (liaison met een) terrorist in zit. Dat is heel veel werk. Dat vereist een bepaalde manier om naar die data te kijken. De fouttolerantie moet anders worden ingesteld. Daarbij komt dat je geen over- of underkill wilt creëren. Ik bedoel: je zou heel veel personen kunnen vinden die het eventueel zouden kunnen zijn. Maar dan heb je nog eens 20 man aan het werk om uit te zoeken of er inderdaad iemand van de suspectlist tussenzit. Dat wil je ook niet. Je wilt een hoge mate van betrouwbaarheid creëren in je geautomatiseerde vaststelling van mogelijke kandidaten. Een dergelijke actie vereist een heel andere aanpak dan de adresgegevens van het postorderbedrijf. Je kunt heel ver gaan, je kunt veel bereiken. Maar op het moment dat wij als mens twee records bekijken en we moeten vaststellen wat de mate van overeenkomst is en we kunnen niet met zekerheid zeggen of ze nu wel of niet gelijk zijn, dan wil ik eigenlijk dat mijn applicatie dat ook zo doet. Dat de applicatie zegt: 'kijk hier nog eens naar' of om meer informatie vraagt om vast te stellen of het om hetzelfde record gaat of niet."

"Mijn ervaring is dat je met de weg die wij zijn ingeslagen een hele hoge mate van nauwkeurigheid kunt bereiken. Ik weet dat je de 80/20 regel eigenlijk niet meer mag gebruiken, maar die eerste 80 procent kun je met een aantal mathematische methodieken wel bereiken, maar als je zoals wij het determinisme met het probabilisme combineert en eerst vaststelt wat wat is, daarna de appels bij de appels en de peren bij de peren doet, om die

Winfried van Holland: "Het is nog steeds geen gemakkelijke zaak om bedrijven te overtuigen van de business case."

vervolgens op een slimme manier te vergelijken, dan kun je een kwaliteitsniveau halen dat heel erg hoog ligt, in de hoge 90 procent regio."

Namen

De drijfveer voor bedrijven om contact te zoeken met Human Inference verschilt sterk. Van Holland: "Bij de verzekeraars en banken is dat vaak het customer intimacy verhaal. Ze willen weten wie hun klant is. Ze willen een hoge mate van betrouwbaarheid naar hun klanten uitstralen. Dat geldt ook in het hoge segment van de retail. Daaronder is het vaak kostenoptimalisatie. Ook compliancy kan een drijfveer zijn. Fusies en overnames zijn een mooi moment om data op te schonen. Wij hebben daar een speciale aanpak voor; Move & Improve."
Het is zelfs niet altijd eenvoudig om vast te stellen of A. de Vries in de ene dataset dezelfde persoon is als A. Jansen-de Vries, en al helemaal niet als in een derde dataset de persoon A. Jansen voorkomt. "Als je hier als mens naar kijkt, dan constateer je dat de geboortedatum hetzelfde is, het adres is anders, als niet telkens De Vries als meisjesnaam wordt toegevoegd is ook de naam anders," zegt Van Holland. "Dan kan misschien een BSN of voormalig Sofi-nummer uitsluitsel geven, of het e-mailadres of mobiele telefoonnummer, gegevens over social media. Je zoekt naar zoveel mogelijk gegevens voor je profiel om de definitieve vaststelling zo nauwkeurig mogelijk te kunnen doen."

Wandt voegt daaraan toe: "In 2000 is een wet aangenomen waarmee de gebruikersnaam is ingevoerd. Dat houdt in dat als Annemarie de Vries trouwt met Henk Jansen, dan mag Henk Jansen zich ook Henk de Vries, Henk Jansen de Vries, Henk de Vries Jansen noemen – en dan ook nog al dan niet met een koppelstreepje tussen de achternamen. Een van onze grootste klanten, een verzekeraar uit Apeldoorn, vermoedde dat als die gebruikersnaam echt gaat beklijven in de samenleving dat wel eens problemen zou kunnen gaan opleveren in de polisadministratie. Want stel dat de begunstigde in een polis ineens anders in de GBA ingeschreven blijkt te staan, dan heb je een juridisch probleem. Op ons verzoek hebben ze een jaar bijgehouden hoe vaak er een aanvraag binnenkwam voor verandering van familienaam. Dat is in één jaar twee keer gebeurd. We praten dus over hoge uitzonderingen. In landen als Engeland en Duitsland is dat echter geheel anders, daar is het schering en inslag, heel gewoon, net als in de Scandinavische landen."
Human Inference heeft in Nederland diepgaand onderzoek gedaan naar wat werkelijke dubbele achternamen zijn en wat combinatienamen zijn, al dan niet met koppelstreepje. "Dat hebben we goed in kaart gebracht. Met bronnen als het Adelsregister, aangevuld met veel telefonisch onderzoek. We hebben gegevens uitgewisseld met het Meertens Instituut voor Naamkunde en volkstellingsgegevens, en die bestanden tegen elkaar aangehouden om te kijken naar de overeenkomsten en verschillen. Dat bestand hebben we behoorlijk goed voor elkaar; er komen ook niet zoveel nieuwe Nederlandse namen meer bij. Wel uit het buitenland, vanzelfsprekend."

Al die kennis van namen heeft ertoe geleid dat Human Inference een functie aan de suite heeft kunnen toevoegen die naamsuggesties doet. "Dat vind ik echt heel erg mooi," zegt Wandt. "Je typt een voornaam of een achternaam in en vanuit die grote corpus doet het systeem een suggestie van 'zou het ook deze naam kunnen zijn?' Om verschrijvingen tegen te gaan, of om te kijken welke namen op een bepaalde naam lijken. Dat zou je ook heel mooi kunnen gebruiken voor historisch namenonderzoek. Dat doen wij niet, maar daar zou het ook heel erg nuttig voor zijn."
Van Holland heeft nog steeds bewondering voor het basisprincipe van de software. "Als er dubbele records gevonden worden dan zie je de typfouten en andere dingen die verkeerd zijn gegaan. Toch weet onze software dat het om dezelfde records gaat. Dat vind ik nog steeds boeiend om te zien."

De markt

Human Inference is een relatief kleine partij in een hele grote markt. Er komen steeds meer data waarvan de kwaliteit niet altijd even betrouwbaar is. Het toekomstbeeld moet dus voor het bedrijf heel rooskleurig zijn.
Van Holland wijst erop dat er vooral heel veel ongestructureerde data bij komen. "Er worden wel veel data rondom klanten en contactpersonen heen gegenereerd, maar die zitten nog vrij gestructureerd in de databases. Wat we wel zien is dat de

gestructureerde bestanden groter worden door globalisering, het samenvoegen van data. We zijn weliswaar van origine Nederlands, maar in de praktijk zijn we veel meer een Europees bedrijf, met veel klanten in Duitsland, België en Engeland. En we beginnen klanten in de VS te krijgen. De vertrokken CEO Sabine Palinckx heeft vooral de partnerkanalen uitgebouwd en versterkt. We doen dus meer zaken via partners en de cloud. Voorheen werd onze software bijna alleen *on premise* geïnstalleerd, de laatste jaren zien we veel meer SaaS-toepassingen. Een deel van onze producten is daar uitermate geschikt voor. Fire & Forget, noemen we dat: de klant stuurt ons een postcode met huisnummer en krijgt het bijbehorende adres terug. Hij stuurt een naam en krijgt terug of het om een man of een vrouw gaat. Op het moment dat je daar een machine voor neerzet met een website er bovenop, werk je al snel niet meer alleen in Nederland. SaaS biedt ons een grote groeipotentie."

Voor het ontdubbelen van bestanden dien je echter alle data in handen te hebben om te kunnen vergelijken. "Dat moet dus bijna wel on premise. Bovendien is de traditionele enterprisemarkt nog heel erg huiverig om hun data buiten de muren van de onderneming te sturen. Salesforce.com heeft in de marketingdata al heel wat deuren geopend. Maar onze traditionele klant vindt het zelfs al eng om een postcode met huisnummer naar buiten te sturen. Ze vergelijken bestanden het liefst binnen de veilige omgeving van hun bedrijfsmuren. De nieuwe medewerkers hebben de 'Google Experience'. Ze zijn gewend om naar internet te gaan, wat te openen en de functies naar zich toe te halen. Dat zal op den duur ook bij de traditionele bedrijven leiden tot groei in het gebruik van clouddiensten."

> De DQ-suit koppelt alle fases van de DQL aan elkaar en biedt als het ware een generieke datakwaliteitsstrategie

Van Holland ziet ook dat de grote megavendors datakwaliteit als commodity meeleveren. "En zo gaan ze er ook mee om; als commodity. SAP kreeg met de overname van BusinessObjects ook Fuzzy en First Logic in handen. Een gebruiker weet vaak helemaal niet eens dat in zijn BO-omgeving datakwaliteitfuncties zitten, want ze hebben BusinessObjects gekocht voor Business Intelligence. Inmiddels is alles wat met Fuzzy te maken had verdwenen bij SAP. First Logic zit er nog wel in maar dat product kijkt op de traditionele manier naar datakwaliteit: de hele wereld bestaat uit ZIP-codes. De datakwaliteitoplossingen van de grote leveranciers zijn allemaal op de V.S. gericht. Daar heb je in Europa weinig aan. Je reist in ons land 100 kilometer naar het oosten of het zuiden en je komt in een andere cultuur terecht met een andere taal. Vlaamse achternamen worden anders gespeld dan Nederlandse. Fransen draaien vaak voor- en achternaam om, en spellen de achternaam helemaal in kapitalen. Sommige Belgen doen dat ook. In Engeland en Frankrijk staan de huisnummers vóór de straatnaam. Wij vinden dat niet erg, sterker nog, we vinden die verschillende culturen leuk. We hebben software gemaakt die daar allemaal mee kan omgaan."

"Onze speerpunten zijn internationalisatie en gemakkelijk kunnen aanhaken bij grote partijen," aldus Van Holland. "Vroeger was datakwaliteit een op zichzelf staand vraagstuk, nu is dat uitgegroeid tot MDM, CDI, CRM. Vroeger ging het bij datakwaliteit vooral om namen en adresgegevens, nu zitten in de databases ook e-mail adressen, Twitter id's, Facebook pagina's, LinkedIn id's en noem maar op. Sinds 2008 mogen mensen in elke karakterset hun domeinnaam schrijven. Je zult ze in toenemende mate in het Grieks, Chinees, Russisch of Arabisch tegenkomen. De 1,3 miljard mensen in India doen het nog in het Engels. Maar het zal me niet verbazen als de 1,4 miljard Chinezen massaal overstappen op Chinese karakters in de URL's. Ons bedrijf moet meegaan met al die ontwikkelingen, zich aanpassen aan de nieuwe uitdagingen. Ja, er is nog genoeg voor ons te doen."

Hans Lamboo is hoofdredacteur van Database Magazine.