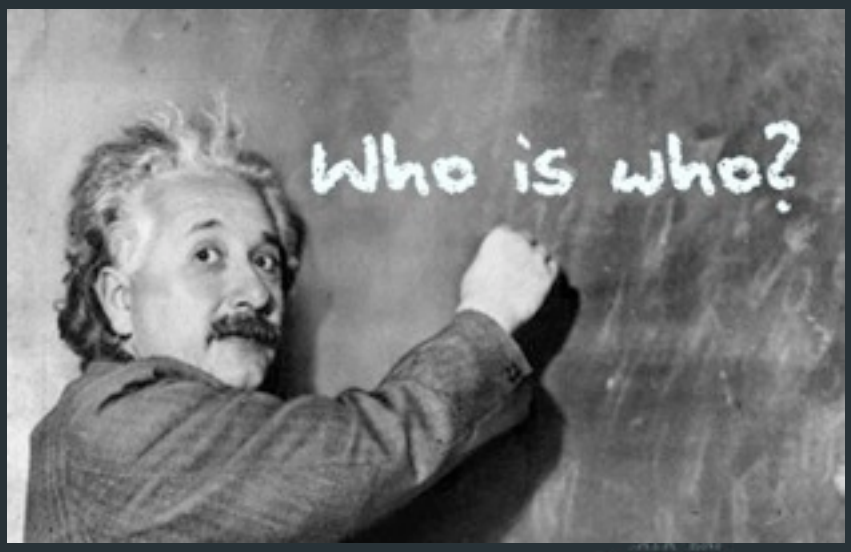
Mini Whitepaper

KENNIS IS DE SLEUTEL TOT IDENTIFICATIE



Human Inference

Kennis is de sleutel tot idenficatie

Steeds meer bedrijven erkennen het belang van goede verwerking van relatiegegevens. Of het nu gaat om ontdubbeling van bestanden, online zoeken, invoercontrole of het samenvoegen van doublures: het is de bedoeling relaties al dan niet als gelijken te identificeren. Om een betrouwbare en zinvolle uitspraak over de mate van overeenkomst te kunnen doen, is toegevoegde kennis vereist. Kennis over de relatiegegevens en de omgeving en de cultuur waarin zij voorkomen. Deze kennis moet op een consistente en intelligente manier worden verzameld en vastgelegd en vervolgens in de identificatiemethode worden geïncorporeerd. Dat is feitelijk de enige juiste manier van verwerking van relatiegegevens.

Traditionele methoden

Het gebruik van identificatiemethoden als matchcodes, trefwoorden, postcode-huisnummer en relatienummers, vaak in combinatie met andere gegevens zoals een geboortedatum, levert in vele gevallen niet het gewenste resultaat. De gevonden overeenkomsten zijn in veel gevallen onbetrouwbaar en niet specifiek genoeg (mismatches en missed matches). De matchcode BAAGEN73 levert in 2013 veel 40-jarige kandidaten uit Gendringen en Genemuiden op, maar houdt geen rekening met de fonologische overeenkomst tussen mevrouw Van Balen en mevrouw Baalen. Een groot gevaar bij het gebruik van trefwoorden is, dat degene die het trefwoord kiest, vaak iemand anders is, dan degene die de relatie later opzoekt. Hoeveel mensen geven niet bij het zoeken naar relaties uit ‘Bourtange’ het trefwoord ‘BOE’ op? Een combinatie van postcode en huisnummer zegt wel iets over een locatie, maar niets over de mensen die er wonen of hebben gewoond. Daarnaast is het in Nederland ook nog eens zo dat er meerdere straten bij een bepaalde postcode kunnen horen. Het gaat hierbij om zo’n drieduizend postcodes. De combinatie met het huisnummer is dan dus niet meer uniek. Over de foutgevoeligheid van relatienummers is al veel geschreven. Maar bedenk ook dat de gegevens van elke Nederlander in gemiddeld negenhonderd bestanden zijn opgeslagen. Hiermee krijgt iedere inwoner zoveel unieke nummers toegekend, dat de klantvriendelijkheid ver te zoeken is.

Conclusie: het zoeken met traditionele methoden kent vele nadelen. Er worden te veel niet relevante relaties gevonden en/of de werkelijk gezochte relaties worden gemist.

Traditionele zoekmethoden zijn onvolledig

Mathematische vergelijking

Wanneer men de vergelijking van relaties baseert op zuiver mathematische methoden, gebeurt dit op basis van een overeenkomst in het aantal en de volgorde van bepaalde letters. Er wordt echter geen rekening gehouden met de betekenis van die reeksen van letters. Ook het gebruik van afkortingen en acroniemen blijft bij deze manier van vergelijken buiten beschouwing. Zo lijkt bij een dergelijke methode de letterreeks ‘allr’ meer op

de familienaam ‘Aller’ dan op de afkorting van het bijvoeglijke bedrijfswoord ‘allround’. Ook bestaat er weinig overeenkomst tussen de bedrijfsnaam ‘Eerste Nederlandse Taxi- en Automobielmaatschappij’ en het hiervan afgeleide acroniem ‘ENTAMij’.

Dit gebrek aan kennis over de betekenis van de relatiegegevens zorgt bij puur mathematische vergelijkingsmethoden voor veel overkill en underkill (onterecht gevonden matches en onterecht niet gevonden matches).

Betekenis

Bij het analyseren van relatiegegevens kunnen we o.a. de volgende betekeniscategorieën onderscheiden:

- titulatuur: Firma Bakker, Gebroeders De Boer, Mevrouw Versteegh
- titels: Drs. Philip van Meerdingen, Carel baron Sloet tot Oldenhave, Willem Wanders MSc
- voornamen: David Bertelink, Kim Kaasjager, Marie-Louise van Houwelingen
- voorletters: H.A.F.M.O. van Mierlo, M. Boogerd
- voor- en tussenvoegsels: Jan van der Graaf, Theo Rutten meergenaamd Roethof
- familienamen: Sophie Beer, Annelies van Aakster Bussen
- toevoegingen: J. Holsboer & Co. IT-services, Gerard Hamming Hzn., Walter Delleman jr.
- beroepsaanduidingen: B. Vink cardioloog, Ton de Vos informatieanalist
- geografische aanduidingen: Tilburgse Betonmortelfabr. BEMOTI, Drankenhandel Vd Spek Arnhem
- rechtsvormen: Human Inference Enterprise BV, Jansen & Tilanus GmbH
- bedrijfswoorden: Arnhemse Steenfabriek, Oudman & Partners Planologische Consultancy BV
- bedrijfseigennamen: MarktSelect BV, Zuivelfabriek Campina, Kledingherstelbedrijf Van der Naald
- rangtelwoorden en tijdsaanduidingen: Eerste Twentse Stoomspinnerij Anno 1907

Bij het vastleggen van de betekenissen moet men ook rekening houden met de ambiguïteit van verschillende onderdelen van de tenaamstelling. Zo is het item ‘art’ zowel voornaam, familienaam als (afgekort) bedrijfswoord <zie figuur 1>.

Een item dat wordt vastgelegd in de kennis moet dus in al zijn betekenissen worden vastgelegd. Alleen op deze wijze kan eenduidige interpretatie plaatsvinden en worden er geen appels met peren vergeleken.

Daarnaast speelt ook de omgeving waarin een item zich bevindt een rol. In het laatste voorbeeld van figuur 1 zien we bijvoorbeeld dat het item ‘art’ wordt gevolgd door een punt en een voorzetsel. Contextanalyse leert dat we in dergelijke gevallen zeer waarschijnlijk met een bedrijfswoord te maken hebben.

De heer Art de Vries

Mevrouw J. Art-de Vries

Art Gallery de Vries

Art. voor Kunstnijverheid De Vries BV

Fig. 1: Betekenisambiguïteit van het item ‘art’

Bij het vergelijken van relatiegegevens zijn de context en het verschil in betekenis van groot belang voor de mate van overeenkomst van de records.

Regels

Naast het vastleggen van kennis in bepaalde betekeniscategorieën, is de notatiewijze van de verschillende onderdelen van de tenaamstelling van groot belang. Hierbij hebben we te maken met geschreven en ongeschreven landspecifieke regels. Zaken als meervoudsvorming, afkortingen, acroniemvorming, synoniemen,

de vorming van samenstellingen en bijvoeglijke afleidingen zijn hierbij belangrijke factoren. Wanneer een tenaamstelling in verschillende databestanden voorkomt, dan is de kans groot dat de notatie zal verschillen. Enkele voorbeelden?

'Int. Transp. Ond.' is gelijk aan 'Internationale Transportonderneming' en niet aan 'Intern Transplantieonderwijs'.

'ENTAMij' is het afgeleide acroniem van 'Eerste Nederlandse Automobiel- en Taximaatschappij'.

De afgekorte string 'arnh' in de tenaamstelling 'Arnh. Zuivelcoöperatie Van Otten & Zonen' zal taalkundig gezien een bijvoeglijke geografische afleiding moeten zijn: 'Arnhemse Zuivelcoöperatie Van Otten & Zonen'.

Het gaat dus ook om het verzamelen en intelligent vastleggen van kennis over de kennis.

Figuur 2 en 3 illustreren hoe het ontbreken van kennis en regels de mate van overeenkomst kan beïnvloeden.

B.V. Sloten Amsterdam Bedrijfsauto's <-->

B. v. Amsterdam Autobedrijven Sloten

Internationaal Transport Idema <-->

Int. Transp. I. de Man

A.B.H. Dekker-de Goey <-->

H.A.B. de Goey-Dekkers

Fig. 2: Onterecht hoge overeenkomst van de relatieparen door het ontbreken van kennis en regels

Gar. Van Sloten A'dam BV <-->

Amsterdams Automob. Bedr. Sloten VOF

Cebuco <-->

Centraal Bureau voor Courantenpubliciteit

Esveha <-->

SVH

Fig. 3: Onterecht lage overeenkomst van de relatieparen door het ontbreken van kennis en regels

Het zal duidelijk zijn, dat het gebruik van kennis en regels over deze kennis in alle gevallen een beter identificatieresultaat tot gevolg heeft.

Alternatieven?

Er zijn uiteraard identificatiemethoden, waarbij de gebruiker zelf kennis kan toevoegen. Dit heeft echter een aantal evidente nadelen.

Het vastleggen en onderhouden van kennis is, zoals uit het voorgaande al blijkt, een dynamisch en complex proces, waarbij deskundigheid is vereist. Wanneer de gebruiker zelf kennis invoert om een "kennisloze" identificatiemethode te optimaliseren, levert hij uiteindelijk een grotere tijdsinspanning met een kwalitatief minder resultaat. Het toevoegen van kennis alleen is immers niet genoeg. Ook contextanalyse en kennis van taalkundige regels zijn nodig voor een goede identificatie.

Op deze manier maakt het bedrijf uiteindelijk hogere kosten dan aanvankelijk voorzien. Het zelf toevoegen, onderhouden en nabewerken van kennis komt de gebruiker in kwestie letterlijk duur te staan.

Feitelijk is er geen alternatief. Een goede identificatiemethode maakt gebruik van geïncorporeerde kennis

over relatiegegevens en de omgeving en de cultuur waarin zij voorkomen. Ook in uw organisatie.

Over Human Inference

Human Inference helpt al meer dan 25 jaar overheid en bedrijfsleven om beter met hun klanten om te gaan, door hen alle pijn rondom klantgegevens en informatiekwaliteit uit handen te nemen.
Zo kan de **Belastingdienst** vooraf uw juiste gegevens invullen.
Centerparcs stuurt u een persoonlijk aanbod, waardoor zij 20% meer rendement op hun marketing halen.
I**NG** kon pijnloos samengaan met de Postbank.
Nutricia realiseerde in 3 maanden de basis voor nog gezondere marketingcampagnes.
En **Aegon, ING Lease, SNS property Finance** en vele anderen voorkomen miljoenen aan fraude, ieder kwartaal opnieuw.