ILLUSTRATIE GIJS KLUNDER

Waarheidsvinding in 11,5 miljoen documenten

Bewijsmateriaal in teksten

Forensisch onderzoekers ontwikkelen technieken om relevant bewijsmateriaal in documenten te ontdekken. Semantisch zoeken in die teksten houdt rekening met de betekenis en de context van woorden, maar de onderzoekers kijken ook naar wat er juist níet staat. Onder andere de Belastingdienst, de FIOD en de politie zijn nauw betrokken bij deze onderzoeksprojecten. Twee wetenschappers vertellen.

ERIK VAN DER SPEK
is mede-eigenaar van Hendrikx Van der Spek, een communicatiebureau in Bussum. Daarnaast is hij lector bij de masteropleiding Communicatie en Organisatie aan de Universiteit Utrecht.

Zoeken in teksten kan iedereen, maar zoeken naar bewijsmateriaal in miljoenen documenten is een vak apart. Toch is dat een dagelijkse taak voor een groeiend aantal rechercheurs en analisten, bijvoorbeeld bij de Belastingdienst. Zij worden daarbij geholpen door wetenschappers in vakgebieden met namen als Semantic Search, Digital Forensics en E-Discovery. Twee onderzoekers, David Graus en Hans Henseler, lichten een tipje van de sluier op.

Het interview komt op een goed moment: de kranten staan vol over de Panama Papers, die een paar weken eerder zijn gepubliceerd: 11,5 miljoen documenten over 214.500 vennootschappen, in totaal 2620 gigabyte aan gegevens. Wat zouden de onderzoekers daarmee kunnen? ‘Het is in zo’n geval belangrijk een overzicht te krijgen van de informatie’, begint David Graus. ‘Daarin zou semantic search een belangrijke rol kunnen spelen. Je kunt vaststellen welke bedrijven en mensen in de documenten worden genoemd, en je kunt de verbanden daartussen in kaart brengen. Een andere benadering is wat wij ‘outlier detection’ noemen: daarbij kijk je naar patronen die je juist niet verwacht.’

Hans Henseler geeft een andere invalshoek: ‘Stel dat je één geval gevonden hebt, bijvoorbeeld een bedrijf dat een bepaalde dienstverlening aanbiedt die je wilt onderzoeken. Die dienst kun je modelleren. Je kunt dan aan de hand van dat model zoeken aan welke andere klanten ze die dienstverlening aanbieden. Dat kan aan de hand van trefwoorden die karakteristiek zijn voor die dienstverlening, maar dan krijg je veel *false positives*, valse resultaten. Je kunt dat voorkomen door te werken met semantiek en Machine Learning. Dan werk je met voorbeelden die karakteristiek zijn voor de dienstverlening die je onderzoekt. Je traint de computer dan om dat soort voorbeelden te herkennen.’

Semantisch Zoeken

Graus en Henseler zijn betrokken bij een omvangrijk onderzoeksproject dat uitgevoerd wordt aan de Universiteit van Amsterdam, in samenwerking met de Hogeschool van Amsterdam (HvA), onder leiding van professor Maarten de Rijke: Semantic Search in E-Discovery. Bij dit onderzoek, gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), zijn verder het Nederlands Forensisch Instituut (NFI), de Fiscale Inlichtingen- en Opsporingsdienst (FIOD) en Fox-IT, een bedrijf op het gebied van (cyber)security, betrokken om de ideeën in de praktijk te toetsen. Graus is een van de drie promovendi die bijdragen aan dit onderzoek. Henseler is lector E-Discovery aan de HvA en houdt zich bezig met de ontwikkeling van tools om digitaal bewijs te (door)zoeken. Daarnaast is hij medeoprichter en directeur van Tracks Inspector, een bedrijf dat software ontwikkelt voor de politie en voor opsporingsdiensten.

Het onderzoek richt zich op de vraag: hoe kunnen forensisch onderzoekers slimmer zoeken naar bewijs in grote hoeveelheden e-mails en documenten? Met alleen zoeken op woorden mis je veel relevant bewijsmateriaal, zoals Henseler hierboven al aangaf.

De onderzoekers willen daarom technieken ontwikkelen om automatisch relevante facetten in de documenten te ontdekken. Zo richt Graus zich vooral op entity search: het zoeken naar personen, organisaties en locaties waarbij informatie op een semantisch betekenisvolle manier is georganiseerd. ‘Ik heb onder andere gewerkt aan een project waarin we proberen te voorspellen wanneer bepaalde mensen elkaar gaan e-mailen, gegeven het netwerk en de inhoud van de e-mail’, zegt Graus. Semantisch zoeken houdt rekening met de betekenis en de context van woorden. De verwachting is dat onderzoekers daardoor gerichter naar relevante informatie kunnen zoeken.

Zoekmachinetechnologie en chatbots

Binnen het terrein van E-Discovery is een groot aantal aandachtsgebieden te onderscheiden (zie kader Begrippen) en maakt men gebruik van verschillende technologieën. Zoekmachinetechnologie is een belangrijk hulpmiddel, maar dan wel met de nodige aanvullingen, stelt Henseler: ‘Bij Google tik je een woord in en je krijgt tien resultaten, maar je weet niet op welke gronden die selectie tot stand komt. Bij onze algoritmes kunnen we die keuze wél uitleggen. Dat is heel belangrijk in de context van het forensische proces: je moet je resultaten kunnen verantwoorden.’

Een thema dat de laatste tijd sterk in de belangstelling staat, is chatbot-technologie. ‘Google, Microsoft en Facebook investeren allemaal in persoonlijke, digitale assistenten’, zegt Henseler. ‘Neem Siri (van Apple) en Cortana (van Microsoft): dat zijn chatbots waartegen je kunt praten. Ook dat is een voorbeeld van Machine Learning: we beschikken over zoveel data dat we er computers mee kunnen trainen.’ Dat lukt volgens Graus ook op basis van ondertitels van films: ‘Je kunt op internet veel ondertitels downloaden. Als je er maar genoeg van in zo’n model stopt, dan komt er vaak, maar niet altijd, zinnige tekst uit. Als je vraagt ‘What’s the meaning of life?’ produceert de computer soms een diep filosofisch antwoord. De antwoorden zijn grammaticaal correct en semantisch betekenisvol. Je kunt zien dat er nieuwe content wordt gegenereerd die niet in de oorspronkelijke database staat.’

David Graus

Hans Henseler

Daarnaast zijn er nuttige bots die helpen om informatie op een bepaald vakgebied te ontsluiten. ‘Veel van die toepassingen zijn gebaseerd op Watson van IBM, winnaar van de Amerikaanse televisieshow *Jeopardy*’, zegt Henseler. ‘Eén daarvan heeft 2 miljoen patiëntendossiers verwerkt die artsen kan helpen met het vaststellen van een diagnose of suggesties doen voor nader onderzoek. Je hebt ook de virtuele advocaat ROSS, die juridische vragen kan beantwoorden.’

Hoe kunnen forensisch onderzoekers slimmer zoeken naar bewijs in grote hoeveelheden e-mails en documenten? Met alleen zoeken op woorden mis je veel relevant bewijsmateriaal. De onderzoekers willen daarom technieken ontwikkelen om daar automatisch relevante facetten in te ontdekken

Forensische toepassingen

Daarmee komen we op de toepassingen van E-Discovery voor opsporingsdoeleinden. ‘Vooral in de Verenigde Staten is dat al een enorme industrie’, zegt Graus. ‘Zo is er veel software ontwikkeld voor gebruik in een juridische setting. Die wordt bijvoorbeeld gebruikt door advocaten die een bedrijf moeten beschermen en in de data van dat bedrijf op zoek gaan naar een ‘smoking gun’. De andere partij, het Openbaar Ministerie, gaat vanuit het perspectief van de aanklager op zoek naar digitaal bewijs. De software stelt je in staat om te kijken naar communicatiepatronen, samenvattingen te maken van content en onderwerpen en entiteiten in e-mails te identificeren.’

Stel dat de Belastingdienst een computer met een bedrijfsadministratie in beslag neemt. Wat voor type tools heeft die dan tot zijn beschikking? ‘Als het gaat om een financiële administratie maak je gebruik van data science’, zegt Henseler. ‘Iedereen doet belastingaangifte, dus op een gegeven moment weet de Belastingdienst van alle pizzeria’s in Nederland hoe zo’n aangifte er op hoofdlijnen uitziet. Hoeveel verkopen ze, hoeveel kopen ze in, wie hebben ze op de loonlijst staan? Als een bepaalde pizzeria dan heel weinig winst maakt, maar wel heel veel grondstoffen inkoopt, dan valt dat op. Dat kun je markeren en daar kun je inspecteurs op zetten.’

Taalkunde

Hoeveel taalkunde zit er in de technologieën die Graus en Henseler gebruiken? Dat valt tegen, volgens Graus: ‘Patroonherkenning in tekst lijkt heel erg op patroonherkenning in financiële gegevens. Wij zetten tekst om in getallen en gaan er vervolgens patronen in zoeken. Daar komt weinig grammaticale en linguïstische kennis bij kijken. Grammaticale verschijnselen spelen wel een rol, want die zorgen ervoor dat er bepaalde patronen ontstaan. Maar je hoeft die grammatica niet te kennen. We stoppen geen linguïstische kennis in de algoritmes, we proberen die stap overbodig te maken. Wat daarbij helpt, zijn de ontwikkelingen op het gebied van deep learning en neurale netwerken. Een neuraal netwerk kunnen we zonder specifieke grammaticale informatie leren om patronen te herkennen.’

Zoeken naar het onbekende

Analisten en opsporingsambtenaren weten vaak niet waarnaar ze op zoek zijn. Hoe kan de techniek op het gebied van semantic search hen helpen? Henseler beantwoordt deze vraag met een analogie uit de automarkt: ‘Stel dat je zoekt naar een tweedehands auto, maar je weet niet welke. Dan ga je naar Autoscout24 en zie je een aanbod van 2 miljoen auto’s. Maar je krijgt aan de zijkant al meteen een breakdown: diesels en benzineauto’s, gele en blauwe. Dan zeg je bijvoorbeeld: ik wil geen diesel, ik wil geen witte auto en hij mag niet ouder zijn dan 4 jaar. De computer geeft je de mogelijkheden aan, en als onderzoeker bepaal je wat interessant is en wat niet. Bij Autoscout24 zitten al die criteria al netjes in een database, maar in de Panama Papers niet. Het gaat er dus om een dialoog te maken tussen de onderzoeker en de dataset, waarbij de onderzoeker nog steeds de dienst uitmaakt.’ ■

Begrippen

Computational linguistics (Wikipedia): een specialisatie op het grensvlak van taalkunde en informatica waarin de computationele modellering van taalkundige verschijnselen centraal staat;
Digital forensics: een tak van forensische wetenschap gericht op het herstellen en onderzoeken van (beeld- en tekst-) materiaal op digitale gegevensdragers;
E-discovery: het doorzoeken van grote hoeveelheden elektronische data, bijvoorbeeld om bij een juridisch onderzoek bewijsmateriaal te vergaren;
Entity Search: zoeken naar entiteiten (personen, organisaties, locaties) waarbij informatie op een semantisch betekenisvolle manier is georganiseerd;
Machine learning (Wikipedia): een breed onderzoeksveld binnen kunstmatige intelligentie, dat zich bezighoudt met de ontwikkeling van algoritmes en technieken waarmee computers kunnen leren;
Outlier detection: onderzoeksmethode die erop gericht is afwijkingen of afwijkende patronen in data te identificeren.