Predictive Encoding:

glazen bol of black box?

Author(s)
Henseler, Hans

Publication date
2012

Document Version
Final published version

Published in
Automatisering gids

PREDICTIVE CODING: GLAZEN BOL OF BLACK BOX?

Het aantal digitale documenten in bedrijfsomgevingen neemt dermate toe dat men het beoordelen ervan niet kan overlaten aan experts. Predictive coding, zegt Hans Henseler, moet uitkomst bieden. Met behulp van slimme software kan men de relevantie van elektronische documenten voorspellen.

door: HANS HENSELER **beeld:** LEX VAN LIESHOUT / ANP

Digitale informatie in bedrijfsomgevingen neemt explosief toe. Volgens IDC zal de hoeveelheid informatie de komende jaren iedere achttien maanden verdubbelen. Bij een onderzoek naar fraude, omkoping, witwassen, datalekkage enzovoorts, is daardoor steeds meer informatie digitaal beschikbaar. Nu al en zeker in de toekomst is er in feite te veel informatie. Het is onmogelijk en onbetaalbaar om alle informatie door menselijke onderzoekers te laten beoordelen. Uit wetenschappelijke experimenten blijkt bovendien dat mensen helemaal niet zo goed zijn in het beoordelen van grote hoeveelheden documenten. In ieder geval kunnen ze het veel beter als ze daarbij geholpen worden door de computer (zie kader 'TREC Legal track').
Deze ontwikkeling wordt ook wel Technology Assisted Review genoemd en was dit voorjaar het hoofdthema van het symposium e-discovery in Nederland, dat in het voorjaar voor het derde opeenvolgende jaar door de Hogeschool van Amsterdam werd georganiseerd. De ontwikkelingen op dit gebied zijn al enkele jaren aan de gang. Leveranciers van e-discovery-software vallen over elkaar heen om telkens slimme verbeteringen aan te kondigen in hun producten die van oorsprong gebaseerd zijn op het doorzoeken van elektronische documenten aan de hand van steekwoorden. Allerlei technieken die ooit waren bedacht in het kader van kennismanagement en enterprise search verschijnen nu één voor één op het e-discovery-toneel. De businesscase voor e-discovery is echter een stuk beter dan die destijds was voor kennismanagement. Bovendien kunnen hedendaagse servers met 16 CPU cores, 128 Gb RAM en tientallen Tb's aan snelle diskopslag veel data snel verwerken.

ESSENTIËLE FEATURE IN COMMERCIËLE E-DISCOVERY-SOFTWARE

Traditioneel publiceert Gartner in het voorjaar zijn Magic Quadrant voor e-discovery-software. In 2011 noemde Gartner predictive coding alleen bij het onderdeel marketingstrategie als onderscheidend kenmerk. In 2012 wordt het echter expliciet genoemd bij productstrategie als een essentiële feature voor legal reviews. In 2010 is er door het E-Discovery Institute een overzicht gemaakt waarin maar liefst elf e-discovery-producten worden genoemd die over een vorm van predictive coding beschikken. Interessant is dat in datzelfde jaar een van deze leveranciers (Recommind) een patent heeft aangevraagd op zijn predictive-codingtechnologie. Tot grote verbazing van de markt is dit patent in 2011 ook daadwerkelijk verleend. Vooralsnog levert dit vooral veel publiciteit op maar weinig concurrenten denken dat Recommind erin zal slagen om soortgelijke technologie te verbieden. Begin dit jaar hebben op LegalTech andere leveranciers ook predictive-codingtechnologie aangekondigd. Sommigen bieden daarbij overigens ook gelijk slimme dashboards aan en workflow voor het nemen van steekproeven om predictive coding transparanter te maken. LegalTech is met beurzen in New York en Los Angeles al jarenlang het grootste en belangrijkste evenement op het gebied van legal technology.

Voorspellen

Een van de nieuwste technieken is predictive coding die de relevantie van een elektronisch document kan voorspellen (zie kader 'Predictive coding'). Die voorspelling wordt door de computer gedaan aan de hand van een model. Het bijzondere is dat dit model automatisch is gemaakt door een computer aan de hand van voorbeelden die door een expert zijn beoordeeld. De casestudies over deze nieuwe techniek zijn indrukwekkend. Wat te denken van een verzameling van 2 miljoen documenten die onmogelijk door zestig advocaten op tijd gereviewed kon worden. Door middel van predictive coding is met behulp van een senior in tien uur tijd de set van relevante documenten teruggebracht tot 400.000 die wel binnen de termijn gereviewed konden worden en tegen beduidend lagere kosten.
Het probleem bij modellen die automatisch door de computer worden gemaakt, is dat ze voor mensen meestal niet begrijpelijk zijn. Dat geldt zeker voor predictive coding in e-discovery waarbij de features uit duizenden woorden kunnen bestaan. Bewijs dat is geleverd met behulp van zo'n black box is moeilijk te toetsen. De afgelopen jaren werd predictive coding sterk bekritiseerd en slechts enkele leveranciers durfden openlijk te pleiten voor deze nieuwe benadering (zie kader 'Essentiële feature in commerciële e-discovery-software). Daar is begin dit jaar verandering in gekomen nadat nota bene een Amerikaanse rechter, Judge Peck, in een rechterlijke uitspraak het gebruik van predictive coding heeft toegestaan. Haast vanzelfsprekend (in de VS) is men tegen deze uitspraak in beroep gegaan maar inmiddels is de uitspraak door een federale rechtbank bekrachtigd. Inmiddels heeft ook Gartner recentelijk aangegeven dat predictive coding-functionaliteit niet mag ontbreken in met name juridische e-mail-reviews. Daarmee lijkt de doorbraak compleet.
Met deze uitspraken is predictive coding nog niet zomaar van black box omgetoverd in een glazen bol. Wie goed de uitspraak van Judge Peck en verwante artikelen heeft gelezen, weet dat er wel degelijk aanvullende zekerheden worden gevraagd. De rechter vraagt niet om een perfecte oplossing maar wel om een redelijke oplossing waarin transparantie essentieel is. Het ligt voor de hand dat daarmee de nadruk wordt gelegd op de kwaliteit van het hele e-discovery-proces dat doorlopen wordt. Dat gold al voor het verzamelen en verwerken van elektronische documenten met hulp van procedures en software en geldt nu dus ook voor predictive coding. Door middel van een steekproef is relatief eenvoudig vast te stellen of er toch interessante informatie zit in een verzameling documenten die met behulp van predictive coding terzijde is geschoven.
Daarmee komt e-discovery langzaam op bekend terrein van de rechters. De toepassing van statistiek bij de interpretatie van traditioneel forensisch sporenonderzoek is namelijk al jaren in opmars.

Verzamelen

Predictive coding is bij uitstek geschikt om alle relevante documenten te verzamelen, bijvoorbeeld wanneer een partij verplicht wordt om alle informatie over een bepaald onderwerp aan te leveren. Het is dan belangrijk om te kunnen onderbouwen dat alle informatie geëvalueerd is, ook al is een groot deel van de verzamelde documenten nooit door onderzoekers zelf gelezen. Predictive coding is echter minder geschikt in het begin van een onderzoek wanneer er behoefte is aan Early Case Assessment. Bij Early Case Assessment wordt in een vroegtijdig stadium van een onderzoek een eerste verkenning gemaakt van het bewijsmateriaal. Op dat moment volstaat vaak het vinden van enkele relevante bewijsstukken om een beschuldiging te verifiëren en om vast stellen wat de beste onderzoeksstrategie is.
De technieken voor Early Case Assessment worden echter ook steeds slimmer en de computers worden steeds sneller. IBM is er vijftien jaar geleden in geslaagd om met Deep Blue de wereldkampioen schaken te verslaan. Vorig jaar versloeg IBM met Watson Amerika's beste Jeopardy-spelers. Beide prestaties werden vooraf door velen als onhaalbaar beschouwd. Het is niet zozeer de vraag of maar eerder wanneer de opvolger van Watson na het doorlezen van 2 miljoen e-mails en het aanhoren van de beschuldiging, een samenvatting geeft van de meest relevante documenten waartoe de onderzoekers, advocaten en of de rechter zich vervolgens kunnen beperken.

Hans Henseler
(j.henseler@hva.nl) is lector e-discovery in het kennis-centrum Create-IT van de Hogeschool van Amsterdam en partner bij Fox-IT.

2 MILJOEN DOCUMENTEN WERDEN IN TIEN UUR TERUGGEBRACHT TOT 400.000 RELEVANTE DOCUMENTEN

TREC LEGAL TRACK

TREC staat voor Text Retrieval Conference en wordt sinds 1992 jaarlijks georganiseerd met sponsoring van NIST en het Amerikaanse ministerie van Defensie. TREC stimuleert onderzoek naar text retrieval door een grootschalige evaluatie van text-retrievaltechnieken mogelijk te maken in een wetenschappelijke context en in competitieverband. De kracht zit vooral in het beschikbaar maken van grootschalige datasets waarmee deelnemers hun technieken kunnen evalueren. TREC bestaat uit verschillende onderdelen die worden aangeduid als track. Door de jaren heen zijn er steeds verschillende tracks gestart. Zo is TREC Legal track voor het eerst in 2006 georganiseerd en richt zich op text retrieval in juridische context. Met behulp van de testdata uit TREC Legal 2009 is aangetoond dat een technology assisted review niet alleen efficiënter is maar ook superieur aan een review zonder dergelijke ondersteuning. Op dit moment is vooral de Enron-dataset met circa 500.000 e-mails zeer populair bij e-discovery-onderzoeksprojecten. De organisatoren van TREC Legal track hebben dit jaar aangekondigd dat ze een set van 1.000.000 e-mails voorbereiden die afkomstig is uit een failliete onderneming.

Predictive coding

Predictive coding is een proces waarin onderzoekers met behulp van slimme software automatisch de relevantie van documenten voorspellen. De software wordt aan de hand van voorbeelden geleerd wat relevante documenten zijn. Dit kan enkele uren in beslag nemen en bestaat meestal uit meerdere iteraties. In een iteratie voorspelt de software van een aantal, bijvoorbeeld honderd, documenten de relevantie en legt die weer voor aan een expert. Aan de hand van het oordeel van de expert wordt het model aangepast zodat de voorspelling na iedere iteratie beter wordt. De betrouwbaarheid van de voorspelling kan vervolgens bepaald worden met een steekproef uit de niet-relevante documenten.
De slimme software in Predictive coding is gebaseerd op Support Vector Machines (SVM) en Probabilistic Latent Semantic Analyses (PLSA). Documenten worden gepresenteerd als wiskundige vectoren die zijn samengesteld uit vectoren van de woorden die in een document voorkomen. PLSA zorgt ervoor dat de vectoren van woorden die vaak samen voorkomen in een document op elkaar lijken. Een SVM is een bekende techniek uit de patroonherkenning die gebruik maakt van deze vectorrepresentaties. Daarmee wordt een model gemaakt dat voorspelt of een document lijkt op een van de geleerde voorbeelddocumenten.

Voor reacties en nieuwe bijdragen van IT-experts:
Henk Ester,
070 3780397
h.ester@sdu.nl