Robotisering van infor

Deze ontwikkeling wordt niet door iedereen met open

Zoekmachines worden persoonlijke assistenten bij het vinden van informatie: de robotisering van informatiemanagement, concludeert Hans Henseler. Die robotisering beperkt zich niet tot slimme persoonlijke assistenten. Er zullen ook chatbots verschijnen die met mensen communiceren. En autonome systemen die met chatbots informatie verzamelen.

door: HANS HENSELER **beeld:** SHUTTERSTOCK

Zoekmachines worden langzaamaan persoonlijke assistenten die ons helpen met het uitvoeren van taken en vinden van informatie. Researchafdelingen van de grote spelers, zoals Google, IBM, Facebook, Apple en Microsoft, werken hard aan de ontwikkeling van nieuwe technieken waarmee automatisch kennis verkregen wordt uit ongestructureerde informatie. Ze worden daarbij geholpen door steeds snellere computers en de grote hoeveelheid data die door consumenten op sociale media gedeeld en beoordeeld worden.

De toepassingen van deze nieuwe mogelijkheden zijn overal zichtbaar en niet alleen voor consumenten, maar ook voor professionals merkbaar. Neem alleen al de spectaculaire opkomst van data science als tool voor het omzetten van data in waarde, voor inzichten in en antwoorden op veel verschillende gebieden. Het belang van data science wordt onderstreept in de Nationale Wetenschapsagenda waar het is opgenomen bij het onderwerp 'Big data verantwoord gebruiken: zoeken naar patronen in grote databestanden'.

Machineleren

Het vakgebied E-discovery kent inmiddels een lange traditie in het ontwikkelen van praktisch toepasbare methoden waarmee het zoeken naar patronen in grote databestanden (met name tekstueel) verbeterd kan worden. E-discovery-professionals lopen voorop in het toepassen van machineleren om efficiënt en met hoge nauwkeurigheid relevante e-mails en documenten te selecteren uit grote verzamelingen ongestructureerde informatie. Diezelfde tools kunnen ook een uitkomst zijn voor informatieprofessionals die worstelen met een overvloed aan informatie (zie AutomatiseringGids, 11 september 2015).

Uit onderzoek van Blair en Maron in 1985 bleek al dat menselijke reviewers zichzelf sterk overschatten in het goed kunnen beoordelen van documenten. De onderzoekers Cormack (computer science professor, gespecialiseerd in onder meer information retrieval) en Grossman (advocate, gespecialiseerd in E-discovery) toonden in 2011 aan dat menselijke reviewers aanzienlijk beter relevante documenten kunnen vinden met behulp van machineleren dan alleen met boolean search. Uiteindelijk zou dit leiden tot een brede acceptatie van predictive coding, een vorm van technology assisted review (TAR). Deze methode is

TAR 2.0

Een belangrijk nadeel van de eerste generatie technology assisted review-methoden (TAR 1.0) is dat een representatieve verzameling positieve en negatieve voorbeelden nodig is om het proces machineleren te controleren. Voor experts is dit vaak een tijdrovende zaak met name in onderzoeken waarbij relatief weinig relevante documenten en e-mails aanwezig zijn (zogeheten low prevalence). Bovendien vinden senior experts gevoelsmatig dit niet erg zinvol werk omdat veel tijd verloren gaat aan het beoordelen van niet-relevante informatie. Vervolgens is er nog het probleem dat de controleset statisch is, terwijl in de praktijk de expert zijn mening vaak bijstelt. Dat betekent in feite dat het leerproces van begin af aan herhaald moet worden.

Continuous active learning lijkt geen last te hebben van deze problemen. Het model wordt voortdurend tijdens het reviewproces getraind en leert zo gaandeweg onderscheid te maken tussen relevante en niet-relevante documenten en e-mails. Het leerproces kan starten met slechts een paar relevante trefwoorden, een relevante e-mail of een relevant document. Het initiële model wordt gebruikt om documenten te rangschikken zodat de senior expert meer relevante informatie voorgeschoteld krijgt en eerder in het proces een team met minder ervaren collega's kan inschakelen.

matiemanagement

armen ontvangen

Voor reacties en nieuwe bijdragen van IT-experts: Henk Ester, 020 235 6415 h.ester@automatiseringgids.nl

E-DISCOVERY SYMPOSIUM

Hoe zorgen we ervoor dat een 'bonnetje van Teeven' nooit meer zoek raakt? Hoe helpen we de politie om zelf digitaal bewijsmateriaal te onderzoeken? En is robotisering in onze informatiestromen een zegen of een vloek? Op de zevende editie van het jaarlijkse Symposium E-discovery op 21 april in Amsterdam komen deze onderwerpen aan bod, samen met andere actuele onderwerpen rondom het thema 'Robotisering van informatiemanagement'. Voor meer informatie zie: https://ediscoverynl.dmci.hva.nl/events/symposium/symposium-2016.

gebaseerd op machineleren waarmee een computermodel getraind wordt om documenten automatisch te classificeren (zie het artikel Predictive Coding: glazen bol of black box, AutomatiseringGids, 14 september 2012).
In de afgelopen jaren zijn verschillende problemen van predictive coding naar voren gekomen. Deze problemen hebben Cormack en Grossman geïnspireerd tot het verbeteren ervan door middel van continuous active learning. Deze nieuwe methode staat inmiddels bekend als TAR 2.0 en presteert aanzienlijk beter dan de huidige TAR-technieken (zie kader). Het enthousiasme voor TAR 2.0 heeft geleid tot een nieuwe competitie in de Tekst Retrieval Conference (TREC) die de toepasselijke naam Total Recall heeft gekregen. Het doel van een E-discovery-review is namelijk om zoveel mogelijk relevante documenten te vinden in tegenstelling tot bijvoorbeeld het zoeken op internet, waarbij gebruikers meestal op zoek zijn naar één of een paar relevante documenten die hun vraag beantwoorden.
Cormack en Grossman hebben met een team van de University of Waterloo een standaardversie van continuous active learning geïmplementeerd die ter beschikking is gesteld aan deelnemers. Het Intelligent Language Processing Systems-(ILPS)-lab van de UvA en het lectoraat E-discovery van de Hogeschool van Amsterdam leverden één van de vijf teams die aan deze competitie in 2015 hebben meegedaan. Meer details over en resultaten van de deelname van het Amsterdamse team zullen tijdens het zevende jaarlijkse Symposium E-discovery gepresenteerd worden (zie kader).

Zwarte doos

Het meest bijzondere van de Total Recall-competitie zijn de zogenaamde 'sandbox runs'. Dit zijn wedstrijdonderdelen waarbij de teams een virtuele machine moeten inleveren met een verbeterde versie van de standaardimplementatie. De virtuele machine is in feite een zelfstandig werkende computer die als een zwarte doos werkt. De zwarte doos wordt gekoppeld aan een bepaalde collectie en krijgt een aantal woorden of documenten die relevant zijn voor het onderzoek. De zwarte doos krijgt ook toegang tot een expert die documenten op relevantie kan beoordelen. Doel is om zoveel mogelijk relevante documenten te vinden met een zo min mogelijk aantal vragen aan de expert.
Met deze opzet is de bemoeienis van een zoekspecialist volledig uitgesloten. Daarbij is het nu ook mogelijk om in de vaak vijandige E-discovery-praktijk volgens deze methode onafhankelijk en toch betrouwbaar een set relevante documenten te verzamelen. De advocaten waren vooraf erg sceptisch. Maar tot ieders verrassing bleek zelfs het relatief eenvoudige algoritme in de standaardversie goed te presteren. Zo goed zelfs dat de deelnemende teams slechts af en toe beter konden presteren. Inmiddels is besloten om de Total Recall-competitie in 2016 nogmaals te organiseren.
De vraag blijft overigens in hoeverre deze en andere technieken Nederlandse informatieprofessionals kunnen helpen. In de praktijk blijkt namelijk dat vooral natuurlijke taalverwerking voor Engelstalige documenten veel beter is ontwikkeld dan voor Nederlandstalige documenten. Daarnaast hebben de documenten waar informatieprofessionals mee te maken hebben meerdere facetten die beoordeeld moeten worden. Een e-mail of document is niet alleen maar wel of niet relevant, maar kan bijvoorbeeld behoren tot een bepaald dossier. Continous active learning blijkt ook voor een dergelijke classificatietaak geschikt te zijn aldus een publicatie van Cormack en Grossman op SIGIR 2015.
De robotisering van informatiemanagement beperkt zich niet tot slimme persoonlijke assistenten die voor ons informatie zoeken op internet en TAR 2.0. Er zullen ook kunstmatig intelligente assistenten verschijnen die niet alleen met andere computers maar ook met mensen kunnen communiceren. Denk bijvoorbeeld aan een chatrobot op een website of in een programma die eenvoudige supportvragen kan beantwoorden. Weer een stapje verder gaan de autonome systemen die met chatbots informatie verzamelen, zoals in het kader van het Sweetie 2.0 project (zie kader).
Deze ontwikkelingen worden niet door iedereen met open armen ontvangen. Het zijn niet de minsten, zoals Stephen Hawking, Elon Musk en Bill Gates, die geloven in de mogelijkheden van kunstmatige intelligentie, maar die er ook voor waarschuwen als we niet tijdig de risico's onderkennen van volledig autonome systemen. «

CHATROBOTS IN HET SWEETIE 2.0-PROJECT

In het kader van het Sweetie 2.0-project wordt een kunstmatig intelligente chatrobot (chatbot) gebouwd aan de hand van voorbeeldchats uit het eerste Sweetie-project. Met dit project stelde Terre des Hommes in 2013 met groot succes en veel publiciteit de aard en omvang van webcamsex met minderjarige kinderen aan de kaak. De chatbot probeert te achterhalen of een gesprekspartner uit is op betaalde webcamseks. In dat geval heeft de chatbot een strategie om aanvullende informatie te verkrijgen zodat na de chat de betrokken persoon een bericht gestuurd kan worden waarin hij wordt aangesproken op zijn gedrag.
De chatbot in het Sweetie 2.0-systeem kan niet alleen converseren. Het systeem kan op gepaste momenten ook een video van een virtueel minderjarig jongetje of meisje (Sweetie) laten zien om eventuele argwaan over de leeftijd weg te nemen. Ook is het systeem in staat om buiten de chatroom, bijvoorbeeld via e-mail, whatsapp of skype, berichten uit te wisselen die onderdeel zijn van het gesprek en deel kunnen uitmaken van de strategie. Dit systeem wordt met subsidie van de Nationale Postcode Loterij door Terre des Hommes ontwikkeld in samenwerking met forensisch psychologen van de Universiteit Tilburg, Tracks Inspector en juridische specialisten van de Universiteit Leiden.

Hans Henseler is lector E-discovery aan de Hogeschool van Amsterdam en CEO en medeoprichter van Tracks Inspector dat software ontwikkelt voor opsporingsinstanties (j.henseler@hva.nl).