

Jan Scholtes is bijzonder hoogleraar 'Text Mining' bij de Universiteit van Maastricht en directeur van Zylab. Bij het einde van de regeerperiode van Barack Obama mag zijn bedrijf ook nu weer de zoeksoftware leveren voor het doorzoekbaar maken van de presidentiële mailtjes. Hij is ook een van de sprekers op de VOGIN-IP-lezing op 9 maart. Twee redenen om hem een paar vragen te stellen.

Heb je vanuit wetenschappelijk invalshoek suggesties wat nog meer nodig is dan een recht-toe-recht-aan zoekmachine om in die miljoenen mailtjes van Obama de informatie met een beetje behoorlijke recall en precisie terugvindbaar te maken?

'De tijd van recht-toe-recht-aan zoeken is wel een beetje voorbij. Er is veel te veel data. Het maakt dan niet meer uit hoe goed je query is, je krijgt altijd te veel hits (een precisieprobleem). Daarnaast weet je eigenlijk nooit of je iets gemist hebt (een recall-probleem). Een betere oplossing is om data tevoren te analyseren, bekijken, classificeren en dan zoveel mogelijk te sorteren, zodat de gebruiker beter het overzicht kan bewaren. Dit kun je doen met technieken uit de tekst-mining. (Semi)-gestructureerde informatie zoekt nu eenmaal makkelijker. Alleen kunnen we die structuur niet handmatig toevoegen, daar hebben we slimme algoritmes voor nodig.'

Maakt het daarbij nog veel verschil of het gaat om mailtjes, om tweets of om longreads op bijvoorbeeld Medium (waar Obama ook regelmatig publiceerde)?

'Nee, niet echt. De lengte van documenten is wel in zekere mate van invloed op de uitkomst van relevance ranking-algoritmes, maar met de juiste normalisatietechnieken zijn veel van die problemen te voorkomen. Ook kun je chat logs of tweets binnen een bepaalde tijdsperiode als een virtueel document beschouwen, zodat je helemaal geen last hebt van dat soort problemen.'

Hoewel intussen door de werkelijkheid achterhaald, speelde onlangs de vraag hoe de FBI in één week 650.000 opgedoken mailtjes van Hillary Clinton kon doorzoeken. Daarbij kwamen ook beweringen langs dat dat wel in één dag had gekund.

'Alles kan met de juiste tools en voldoende hardware, maar je moet natuurlijk wel weten wat

Q&A



Jan Scholtes

je ongeveer zoekt. Weet je dat niet, dan zul je de resultaten van de automatische analyse – overigens nog maar een fractie van dat oorspronkelijke aantal – wel handmatig moeten analyseren.'

Verwacht je dat hoeveelheden e-mail – zoals van opvolgende Amerikaanse presidenten – in hetzelfde tempo blijven groeien of verwacht je dat verdere groei meer gaat zitten in andere nieuwe communicatiemediën of kanalen? En hoe moet je die dan archiveren en doorzoekbaar maken?

'Ja, de exponentiële groei is duidelijk herkenbaar in de grootte van de archieven van Clinton (1 miljoen), Bush (110 miljoen) en Obama (1 miljard). Informatie in Slack en andere repositories kun je ook prima analyseren. Met betrekking tot audio zal ik in tijdens de VOGIN-IP-lezing laten zien hoe je daarin kan zoeken met fonetische technieken zonder dat je een en ander eerst hoeft uit te typen.'

Zou je bedrijven en organisaties die misschien met iets kleinere hoeveelheden te maken hebben, hetzelfde aanraden als voor de Obama-mail?

'Ja, want de problemen die organisaties als het Witte Huis nu hebben, hebben gewone stervelingen over een aantal jaren. De continue exponentiële groei van de hoeveelheid informatie leidt op de langere termijn onherroepelijk tot vergelijkbare problemen.'

Op 9 maart bij de VOGIN-IP-lezing zal Jan Scholtes op de meeste van deze punten verder ingaan.