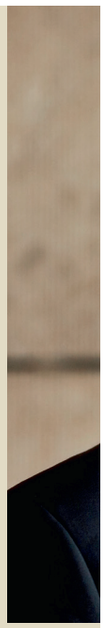
Zoeken in de toekomst

Actuele ontwikkelingen op het gebied van 'zoeken en vinden' kwamen op 26 maart jl. aan bod tijdens de VOGIN-IP-lezing 2015, een dag met lezingen en workshops in de Openbare Bibliotheek Amsterdam. Acht sprekers presenteerden hun bevindingen uit het brede informatieveld.

Door: Andrea Langendoen, redacteur IP en werkzaam bij de KB, en Carola van der Drift, werkzaam bij East Asian Library van de Universiteit Leiden

Foto's: Eef Evers



Bob de Graaf

Is *intelligence* informatie?

Zijn *intelligence* en informatie hetzelfde en welke betekenis heeft dat onderscheid in deze tijd van informatieoverload? Over die vraag boog Bob de Graaf, hoogleraar inlichtingen- en veiligheidsstudies bij zowel de Universiteit Utrecht als de Nederlandse Defensieacademie in Breda, zich in zijn lezing. Zijn stelling is dat *intelligence* informatie is die 1. tijdig is, 2. een hoge relevantie heeft en 3. een indicatie van betrouwbaarheid heeft. Stel dat je bijvoorbeeld als trainer van Ajax in de kleedkamer van Feyenoord belandt en daar de opstelling voor de komende wedstrijd aantreft. Die informatie voldoet aan alle voorwaarden, en is daarmee *intelligence*. Maar als je het pas na de wedstrijd aantreft is zij bijvoorbeeld niet meer tijdig en is daarmee tamelijk nutteloze informatie geworden.
De Graaf schetste hoe we op dit moment door alle technologische en maatschappelijke veranderingen aan de vooravond van een nieuwe intelligence-revolutie staan. Waar vroeger door inlichtingendiensten gericht naar informatie werd gezocht (bijvoorbeeld ten tijde van de koude oorlog om spionnen op te

Greg Notess

Hoe maak je een martini?

Wie een paar jaar geleden op internet zocht naar het recept van een perfecte martini werd via zijn zoekopdracht geleid naar pagina's met recepten. Maar wie tegenwoordig die vraag stelt krijgt meteen het recept te zien. Dat is volgens keynote speaker Greg Notess – schrijver, blogger en bibliothecaris – de grote transitie die het web de afgelopen jaren heeft doorgemaakt: die van trefwoord naar de 'knowledge graph' (of de kenniskaart, zoals Google het op de Nederlandstalige versie noemt), het informatieve blok in je scherm.

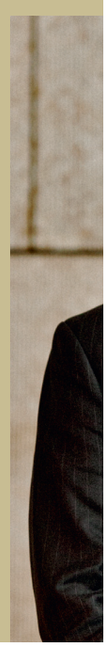
Hartstikke mooi natuurlijk, zo'n snelle hap informatie, maar zoals Notess aangaf: een beetje informatieprofessional wil natuurlijk graag weten waar die informatie vandaan komt (heel vaak Wikipedia, zo blijkt) en hoe up-to-date die is. De gemiddelde gebruiker, waar Google natuurlijk uiteindelijk op mikt, zal zich daar niet zo druk over maken, maar uit onderzoeken die Notess daarnaar uitvoerde, bleek wel dat je je in heel wat bochten moet wringen om meer te weten te komen over welke informatie wordt gebruikt. Want zoals Notess kon concluderen: het kost soms maanden voordat een aanpassing in een bron (bijvoorbeeld Wikipedia) ook zichtbaar is in de knowledge graph. Voor je martini niet zo'n ramp, maar voor meer tijdsgebonden informatie wil je natuurlijk graag dat zij actueel is en klopt.

> Het kost, aldus Notess, soms maanden voordat een aanpassing in een bron als Wikipedia ook zichtbaar is in de knowledge graph

Erik Elgersma

Informatie maakt het verschil

Wie 's avonds na het eten een Monatoetje oplepelt, denk er vast niet over na dat aan het totstandkomen van zo'n puddinkje (kokos met mangosaus deze maand) een heel informatietraject voorafgaat. Erik Elgersma, global Director Business Insight van Friesland Campina, gaf een kijkje in de keuken van het internationale melkbedrijf waar we-



sporen) wordt er nu vooral op basis van heel veel data gezocht naar anomalieën ('afwijkingen') en patronen. Een gevaarlijke ontwikkeling, aldus De Graaf, omdat het verhaal eromheen er niet meer toe doet. Gevolg daarvan zijn veel *false positives*: toevallige overeenkomsten en misinterpretatie. Denk bijvoorbeeld aan de journaliste die ten tijde van de aanslag bij de 'Boston-marathon' googlede op 'rugzak' en 'snelkookpan' – en bezoek kreeg van politiemannen.

Leveren deze 'sleepnetmethodes' dus wel echt die beloofde extra veiligheid op? De vraag is niet alleen wat er door het verzamelen van big data is voorkomen, maar ook wat er door is veroorzaakt. Het alternatief is dus om terug te keren naar de narratieve analyse, bijvoorbeeld in social media, en zo weer meer tijd te nemen voor het verhaal achter de patronen.

Inlichtingendiensten zoeken nu vooral op basis van heel veel data naar anomalieën ('afwijkingen') en patronen, zegt De Graaf

De Winters pikantste stelling is dat de overheid onze grootste hacker van informatie is

Brenno de Winter

'Thanks for all the fish'

Brenno de Winter, onderzoeksjournalist, kwam met een aantal huiveringwekkende verhalen. Of waren het tips om je eigen omgeving te hacken? Mensen zijn zich vaak nauwelijks bewust hoeveel informatie ze (bewust en onbewust) vrijgeven. Vaak is dat niet zo'n probleem, maar als die informatie in handen komt van slimme mensen met slechte bedoelingen dan kan dat ineens heel gevaarlijk worden. Kijk bijvoorbeeld naar de vrolijke Facebook-pagina van een gezin dat net verhuisd is: de hele familie voor de nieuwe woning (het huisnummer is zichtbaar) en de foto van de blinkend nieuwe huissleutels ernaast. Sleutels die je via een site als keysduplicated.com ('Copy your house keys with your phone') heel eenvoudig kunt laten namaken. Voila!

We worden omgeven door informatie, maar er is bijzonder weinig controle op. Keer op keer komen bedrijven met zogenaamd handige toepassingen, met soms een grote impact op onze privacy, maar vaak is er nauwelijks discussie over en veel te weinig toezicht. De Winter heeft zelf veel onderzoek gedaan hoe makkelijk het is om een paar stappen verder te gaan en persoonsgegevens te vinden (soms via sites met gehackte informatie als *pastebin*) en op die manier zelf AIVD'tje te gaan spelen of zelfs iemands digitale leven over te nemen.

De Winters boodschap is dan ook dat de overheid die informatiestromen veel meer zou moeten sturen, in plaats van die zelf voor allerlei soms twijfelachtige doeleinden toe te passen (De Winters pikantste stelling is zelfs dat de overheid onze grootste hacker van informatie is).

reldwijd jaarlijks 11,5 miljard liter melk wordt verwerkt (en daarnaast nog een heleboel pakken Appelsientje). Hij liet zien hoe informatie over de markt (*market intelligence*) een cruciale rol speelt bij de strategische besluiten van het bedrijf.

Dagelijks is een heel team aan informatiespecialisten bezig met het vergaren van informatie. Door het bijhouden en analyseren van relevante nieuwsberichten (iedere dag worden er aan het kennismanagementsysteem ca. 500 toegevoegd) proberen zij grip te krijgen op de wereld om hen heen. Waar nodig wordt in projecten uitvoeriger ingegaan op een bepaald onderwerp – om de kaasmarkt in Uruguay te analyseren bijvoorbeeld. Daarbij wordt gebruik gemaakt van methoden die ook bij bijvoorbeeld militaire inlichtingendiensten worden toegepast, zoals HUMINT (*human intelligence*) en OSINT (*open source intelligence*). Het is aan de informatiespecialisten om deze informatie te analyseren en op betrouwbaarheid te toetsen.

Op basis van al deze bronnen worden adviezen uitgebracht die de directie helpen bij het bepalen welke producten in welk gebied kunnen worden uitgezet.

Overigens, besluit Elgersma, maakt die feitelijke informatie ongeveer 20 procent van het besluit uit. De uiteindelijke besluitvorming wordt voor 80 procent bepaald door intuïtie van de besluitnemer.

We maken ook gebruik van methoden die bij bijvoorbeeld militaire inlichtingendiensten worden toegepast, aldus Elgersma

Toon Steenbakkers

De Douane monitort het internet

Ja, ook de Nederlandse Douane zoekt en vindt via internet. Sterker nog, het is een dagtaak. Op het internet staan miljoenen antwoorden; het gaat erom de juiste vraag te stellen. Aldus Toon Steenbakkers, informatiespecialist bij de Nederlandse Douane. Hij noemde een rits aan programma's en methodes die de Douane inzet om het internet te monitoren. Met Coosto houden ze social media in de gaten en door het nieuws op ODIN Daily te volgen kan de Douane rekening houden met bepaalde 'risico's', zoals producten met gevaarlijke ingrediënten die op de markt zijn.

Steenbakkers stipte heel kort verschillende methodes aan die ze toepassen. Duidelijk wordt dat de Douane zelf van alles ontwikkelt en er sprake is van veel innovatie. Zo is met behulp van Website Watcher-software een tool gebouwd om het internet te monitoren.

Bij het werk van de Douane speelt het strenge privacybeleid een grote rol. Want ook al werkt de Douane met informatie die iedereen op internet kan vinden, het betekent nog niet dat zij hier ook eenvoudigweg mee aan de slag kan gaan. Informatie die gebruikers op Facebook zetten mag niet zomaar worden doorzocht. Ook mag de Douane geen nepaccounts maken en vriendschapsverzoeken uitsturen. Medewerkers mogen alleen officieel, uit naam van de Douane, online handelen. Juristen onderzoeken continu wat wel en niet mag en waar de grens ligt, om te waken dat die grens niet wordt overschreden.

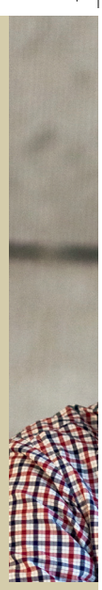
> Bij het zoekwerk van de Douane speelt het strenge privacybeleid een grote rol, benadrukt Steenbakkers

Peter Mika

Semantische mark-up met schema.org

Bij zijn introductie van de semantische mark-up van Schema.org legde Peter Mika, informaticus bij Yahoo! Labs in Barcelona, een link met het vak van zijn toehoorders, allen informatieprofessionals. Die maken het met behulp van ordening en metadata makkelijker om informatie te vinden. En dat is ook het doel van Schema.org, een samenwerkingsverband van de zoekmachines Bing, Google, Yahoo! en Yandex Search. Samen hebben de zoekmachines een schema opgesteld om data te structureren. Met dit schema kan HTML-code van bekende tags worden voorzien, zodat het voor zoekmachines eenvoudiger wordt deze aan internet gerelateerde codes te doorzoeken.

Semantische mark-up – oftewel codering in de webpagina – functioneert op vele manieren hetzelfde als metadata: het zorgt voor informatie *over* de informatie. Met semantische mark-up kun je eigenschappen toekennen aan elementen op webpa-



Thomas Mensink

Beelden classificeren zonder voorbeelden

Thomas Mensink, onderzoeker in *Computer Vision en Machine Learning* bij de Universiteit van Amsterdam, opende zijn lezing met twee vragen: wat is een axolotl en wat is een aye-aye? Het publiek geeft vrijwel moeiteloos de antwoorden. De ene is een soort salamander, de ander een vingerdier. Waarna Mensink grapt dat hij duidelijk een ander publiek dan gewoonlijk voor zich heeft.

De vragen zijn een opstapje naar het onderwerp van Mensinks lezing: kunnen we computers afbeeldingen laten classificeren? En kunnen die dat niet alleen sneller, maar ook beter dan mensen doen?

Het classificeren van afbeeldingen is voor mensen niet zo eenvoudig als het klinkt. Het vergt training en kennis van de klasses (termen waaronder je iets kunt indelen). Een getrainde professional (in dit geval Mensink zelf) kan afbeeldingen classificeren met een foutmarge van iets meer dan vijf procent. Hetzelfde geldt voor computers – ook zij moeten getraind worden. Als er traindata (een testset aan data) beschikbaar is, ligt de foutmarge van bijvoorbeeld GoogLeNet (een project van Google om de computer te trainen voor het classificeren van afbeeldingen) dicht bij die van mensen: 6,8 procent.

gina's en zo de informatie structureren. Daardoor kunnen zoekmachines specifieke gegevens uit pagina's trekken en deze direct in de zoekresultaten laten zien. Denk bijvoorbeeld aan het adres van een bedrijf of aan de sterbeoordeling van een restaurant. Het zijn gegevens die we tegenwoordig meteen te zien krijgen na het invoeren van een zoekopdracht, zonder verder te hoeven klikken naar de website zelf. Op dit moment heeft zo'n vijftien procent van alle webpagina's semantische mark-up.

Maar, benadrukt Mika, het resultaat van semantische mark-up hangt volledig af van de vaardigheden van de informatieaanbieder. Die is verantwoordelijk voor het correct gebruik van semantische mark-up. De zoekmachines tonen slechts resultaten.

Mika: Het resultaat van semantische mark-up hangt af van de vaardigheden van de informatieaanbieder

Vossen: Computers moeten niet alleen teksten analyseren, maar ook zelf beslissingen nemen

Piek Vossen

Een nieuwe kijk op het nieuws

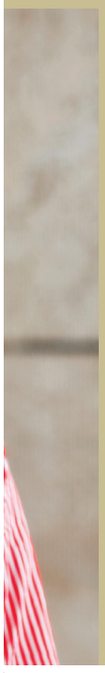
De dagelijkse stroom aan nieuwsberichten – wereldwijd circa 1,5 miljoen berichten, afkomstig van zo'n 30.000 bronnen – is onmogelijk bij te houden. Kunnen computers hulp bieden? Hier komt het NewsReader-project van keynote speaker Piek Vossen, hoogleraar Computationele Lexicologie bij de VU, om de hoek kijken. Het doel van NewsReader is om een computer het nieuws bij te laten houden. De grote vraag is alleen: wat willen we dat de computer precies meet?

Het antwoord van het NewsReader-project hierop is: niet het volume aan nieuws, maar de specifieke gebeurtenissen in hun context. Door de computer nieuwsteksten te laten analyseren en hier het 'wat, waar, wanneer en wie' uit te halen, wordt gepoogd een compleet overzicht te maken van het nieuws. De computer moet dus niet alleen teksten analyseren, maar ook zelf beslissingen nemen. Eenzelfde gebeurtenis wordt door verschillende bronnen op verschillende manieren verwoord, maar toch beschrijven die bronnen dezelfde gebeurtenis. Personen of organisaties, acties en momenten kunnen op oneindig veel verschillende manieren benoemd worden, maar beschrijven toch hetzelfde. De computer moet hierin dus steeds een onderscheid weten te maken.

Door middel van het markeren van elementen uit een tekst met RDF (semantische mark-up, vergelijkbaar met wat schema.org doet) kunnen de elementen uit elkaar gehouden worden en kunnen verbanden gelegd worden. Dankzij RDF kan de computer de tekst verder interpreteren. Wie doet wat, waar, wanneer, waarom, en met wie, voor wie, tegen wie, et cetera.

Dit is natuurlijk niet genoeg. Om nuttig te zijn moet de informatie ook voor mensen interpreteerbaar zijn. Het programma SynerScope kan de data visualiseren en de bijbehorende nieuwsteksten tonen.

Helemaal accuraat is het systeem nog niet. Maar als deze onjuistheden voorkomen kunnen worden, zou NewsReader weleens geschiedenis kunnen gaan schrijven. Letterlijk. <



Wat moeten we ons voorstellen bij het classificeren van afbeeldingen? Aan de hand van specifieke elementen in een afbeelding, zoals kleur, patronen en bepaalde vormen, kan een computer een classificatiesysteem leren en aan de hand daarvan vaststellen in welke klasse een afbeelding hoort. Zwart, wit en gestreept duidt op een zebra, maar dit had ook een streepjescode kunnen zijn. Gelukkig dat de computer dus ook vormen in een afbeelding kan lezen.

Maar niet elk concept in elke afbeelding is de computer aangeleerd. Daarom is het belangrijk dat de computer ook zonder voorbeelden afbeeldingen kan classificeren. Aan de hand van context kan een computer voorspellen waar een afbeelding ingedeeld moet worden. Hoe vaak komt een klasse samen met andere klasses voor? Op basis hiervan kan een computer inschatten wat het onderwerp van een afbeelding is, zonder dat de computer dit onderwerp ooit eerder heeft gezien – en dit is waar het Mensink om gaat. Deze contextuele informatie kan ook van internet gehaald worden. Flickr-tags – termen die door gebruikers van Flickr aan hun foto's zijn toegekend – blijken bijvoorbeeld goed te werken. Zo vindt zelfs een stukje crowdsourcing plaats.

Mensink toont in zijn lezing aan dat computers al heel wat kunnen met afbeeldingen zonder dat ze ze ooit 'gezien hebben'. Nu is het zaak dit te perfectioneren.

Mensink toont aan dat computers al heel wat kunnen met afbeeldingen zonder ze ooit te hebben 'gezien'