Terugblik VOGIN-IP-lezing 2017

Actuele ontwikkelingen op het gebied van 'zoeken en vinden' kwamen op 9 maart jl. aan bod tijdens de VOGIN-IP-lezing 2017, een dag met lezingen en workshops in de Openbare Bibliotheek Amsterdam. Acht sprekers presenteerden hun bevindingen uit heel diverse hoeken van de zoekwereld.

Door: Andrea Langendoen, redacteur IP en teamleider Educatie & Presentatie bij de KB, en Astrid van Wesenbeeck, open science officer en programmaleider Wetenschap voor Iedereen bij de KB

Foto's: Eef Evers

Herbert Van de Sompel

En toen was er niets meer…

Iedereen kent het verschijnsel, je klikt op een hyperlink en je krijgt niks – alleen de deprimerende melding '404 – file not found'. Hyperlinks, zo betoogde keynotespreker Herbert Van de Sompel, zijn zowel de kracht als de achilleshiel van het web. Hij legde uit hoe je als gebruiker al zoekend en klikkend door het web geregeld te maken krijgt met twee verschijnselen: linkrot en contentdrift.

De basis van een hyperlink is eigenlijk heel simpel: pagina a linkt naar pagina b. Zo'n link is statisch en vastgelegd op een bepaald moment in de tijd. Het probleem is alleen dat de pagina waar je naar linkt in de loop van de tijd kan veranderen (contendrift) of zelfs volledig kan verdwijnen (linkrot). Linkrot is daarbij het minst grote probleem: je krijgt gewoon een foutmelding. Zeker in de wetenschap kan contentdrift een risico opleveren omdat je als onderzoeker denkt dat je via een link naar een bepaalde context gaat, maar die kan in de loop van tijd volledig veranderd zijn, zonder dat je dat expliciet merkt. Van de Sompel startte daarom het project Memento, waarbij gebruikers (via webarchieven of versiebeheersystemen) niet alleen de huidige maar ook oudere webversies van een pagina kunnen opvragen. Een soort van tijdreizen op het web dus.

Herbert Van de Sompels project *Memento* is een soort van tijdreizen op het web

Het is onmogelijk (en ook onwenselijk) om te zorgen dat alle content op het web constant blijft. Wel zijn er mogelijkheden om ervoor te zorgen dat beheerde (curated) collecties met bijvoorbeeld overheids-, wetenschappelijke of journalistieke informatie stabiel en integer blijven, bijvoorbeeld door gebruik te maken van permanente links oftewel persistent identifiers (PID's) in plaats van 'gewone' http-adressen.

Daarnaast kan de beheerder die naar andere webpagina's linkt ervoor zorgen dat er bij het linken meerdere opties zijn: niet alleen naar de huidige webpagina maar ook naar snapshots van oudere pagina's (via open webarchieven als het Internet Archive, of programma's als Hiberlink). *(AL)*.

Jan Scholtes

Gaat Artificial Intelligence helpen het zoeken verder te automatiseren?

Als je vandaag de dag op zoek gaat naar informatie, heb je veel goede zoeksoftware tot je beschikking. Sterker nog: de software kan je zelfs helpen als je niet weet waar je naar op zoekt bent. Maar zelfs dat is volgens Jan Scholtes, bijzonder hoogleraar Text Mining en tevens werkzaam bij ZyLab, niet meer voldoende. De tijd van traditioneel zoeken is voorbij: er is simpelweg te veel data, je weet nooit precies wat je krijgt en wat je mist.

Om je verder te kunnen helpen bij je zoekvraag moet je dus gebruik maken van andere technieken, op basis van Artificial Intelligence. Door middel van AI ga je niet op zoek in de informatie, maar wordt de informatie voorgeorganiseerd, geclassificeerd en klaargezet. Zeker voor beroepen waar het zoeken en vinden van informatie een belangrijke rol speelt heeft AI een grote toegevoegde waarde: zowel voor het intelligent zoeken in en het analyseren van de inhoud van documenten, als voor het identificeren, extraheren en classificeren van relevante informatie.

Bij text mining wordt uit grote tekstcorpora relevante informatie geïdentificeerd op verschillende niveaus: entiteiten (be-

Agnes Molnar

Microsoft Graph verbindt kennis

Gemiddeld ontvangen we dagelijks zo'n 63.000 woorden aan nieuwe informatie. Zeg maar een boek per dag. Onmogelijk natuurlijk om al die informatie te verwerken, op te slaan en te doorzoeken. Je hebt dus hulp nodig om daar de relevante informatie uit te kunnen filteren.

Agnes Molnar, managing consultant en CEO van Search Explained, noemt vier patronen waarmee mensen zoeken naar informatie:

> ik weet wat ik zoek en hoe ik dat moet vinden;
> ik weet wat ik zoek maar ik weet niet hoe ik het moet vinden;

grippen als personen, plaatsen, producten) en attributen (de eigenschappen van een bepaalde entiteit, zoals leeftijd van een persoon). Daarnaast kun je ook feiten (de relaties tussen verschillende entiteiten) en gebeurtenissen (relevante activiteiten rondom een entiteit) vinden. Door informatie op deze manier te analyseren kun je computers leren in teksten te zoeken naar patronen (persoon bezoekt plaats) in plaats van naar woorden – en dat ook nog eens op hogere semantische niveaus (vervoegingen van werkwoorden, co-referenties et cetera). Documentclassificatie is een nog doeltreffendere manier om, zeker bij lange teksten, de juiste informatie te vinden. Hiermee kun je documenten automatisch classificeren met als doel het maximaliseren van de recall. Het zorgt ervoor dat relevante documenten automatisch vindbaar zijn zonder te veel afhankelijk te zijn van de zoekvaardigheden van een eindgebruiker. *(AL)*

Door middel van Artificial Intelligence ga je niet op zoek in de informatie, aldus Jan Scholtes, maar wordt de informatie voorgeorganiseerd

Het werk van Bellingcat is vaak niets anders dan een serieuze versie van 'rara waar ben ik', zegt Christiaan Triebert

Christiaan Triebert

Factchecking rock stars

De 'factchecking rock stars', zo worden de onderzoekers van het internationale collectief Bellingcat weleens genoemd. Volgens Christiaan Triebert, onderzoeker bij Bellingcat en Airwars, is het werk van het collectief vaak niets anders dan een serieuze versie van 'rara waar ben ik', daarbij gebruikmakend van open bronnen als journalistieke informatiebronnen, social media, geografische bronnen (wikimapia/digital globe) enzovoort. Geolocatie is de belangrijkste onderzoeksmethode voor Bellingcat.

Het collectief is de afgelopen jaren betrokken geweest bij onder andere onderzoek rondom de MH17, militaire acties in het Midden-Oosten en de verblijfplaats van IS-aanhangers. Triebert toont met drie verschillende casestudies aan hoe hij en zijn collega's op inventieve wijze gebeurtenissen kunnen verifiëren en dateren of personen en locaties kunnen traceren.

Het is fascinerend om te zien hoe de onderzoekers op basis van heel beperkte informatie (satellietbeelden, foto's en soms zelfs alleen geschreven teksten) er toch vaak in slagen om de precieze locatie van een foto te vinden en daarmee voor een onderzoek belangrijke bewijzen te leveren. In een van de voorbeeldcases bleken de onderzoekers op basis van de informatie in een kort filmpje in staat de verblijfplaats van een Tunesische jihadist vast te stellen.

Maar behalve dat dit soort onderzoek heel interessant is, benadrukt Triebert het belang ervan. Zo gebruiken de Verenigde Staten de door Bellingcat gevonden gegevens tegenwoordig om hun verantwoordelijkheid te nemen bij de effecten van hun militaire operaties in bijvoorbeeld het Midden-Oosten. *(AL)*

> ik weet niet wat ik zoek;
> ben ik aan het zoeken?

In onze zoektocht naar kennis worden we geconfronteerd met verschillende uitdagingen. Dat geldt zeker ook binnen organisaties. Er is een overvloed aan informatie. Zoeken naar relevante resultaten is daardoor vaak als het zoeken naar een speld in een hooiberg. Vaak is er ook sprake van 'silo's' binnen een organisatie. Je weet dus niet welke voor jou relevante informatie iemand elders in de organisatie heeft. De vraag is tevens hoe je die informatie deelt binnen een organisatie, over de silo's heen. Daarnaast zijn er vaak verschillende systemen in gebruik binnen een organisatie (technische silo's) en werken mensen met heel verschillende niveaus van zoek- en domeinexpertise.

Net zoals bij Jan Scholtes' lezing is ook Molnars boodschap dat de traditionele manieren van zoeken niet meer toereikend zijn. In een hedendaagse werkomgeving heb je als werknemer tools nodig om de relevante kennis te filteren uit de grote hoeveelheden data. Microsoft Graph (onderdeel van Office 365) is een tool die verbindingen legt tussen jou als gebruiker en de voor jou relevante mensen, kennis en expertise binnen de organisatie. Het systeem leert steeds meer over jou, je werk, je gedrag en selecteert zo de content die het meest dichtbij en relevant is. Hoewel Graph ongetwijfeld een handig systeem is, krijg je als luisteraar toch een beetje de kriebels bij het idee dat deze tool de hele dag bijhoudt wat je doet. *(AL)*

Ook Molnars boodschap is dat de traditionele manieren van zoeken niet meer toereikend zijn

Cees Snoek

'Met deep learning ontlasten we ons eigen brein'

'Er zijn een paar dingen waar computers niet goed in zijn,' antwoordt Cees Snoek op de vraag van een dame uit het publiek of de computer van haar zoontje te zijner tijd misschien zal kunnen aangeven dat het tijd voor hem is om even te rusten. Met andere woorden, kan de computer via de videocamera op basis van gedrag en gezichtsuitdrukking signaleren dat vermoeidheid de energie heeft weggenomen, of dat iemand toe is aan een slok water of aan een paar bewuste diepe ademhalingen.

Snoek, directeur van QUVA Lab (een samenwerking tussen de Universiteit van Amsterdam en de Amerikaanse chipbouwer Qualcomm), denkt dat computers de echt menselijke kwaliteiten niet heel snel zullen kunnen overnemen in de toekomst, maar dat neemt niet weg dat we allemaal zijn verrast als we zien wat al wel mogelijk is met de technologie van deep learning. Met deep learning kan een computer beeld uit een video vertalen naar concepten zoals een boot, een auto, een vliegtuig, maar ook naar gedrag of bewegingen, zoals een kus die een paar seconden duurt of een doelpunt dat wordt gemaakt. De laatste ontwikkelingen in de technologie van deep learning maken het zelfs mogelijk dat de machine beeld kan vertalen naar tekst. Zo kun je een subscript of een ondertiteling creëren bij een video. Concreet betekent dit dat je in zeer rap tempo de concepten of gedragingen waar je naar zoekt, kunt opsporen en dus zelfs van tekst kunt voorzien. Met deep learning ontlasten we ons eigen brein, dat 50 procent van zijn capaciteit nodig heeft bij 'kijken'. Dit kost een computer beduidend minder rekenkracht, maar het blijft dus wel bij objectief registreren. Interpretatie komt er vooralsnog niet te pas. *(AvW)*

> De laatste technologie van deep learning maakt het zelfs mogelijk, aldus Cees Snoek, dat de machine beeld kan vertalen naar tekst

> Het is erg lastig om van 'raw data' chocola te maken, meent Karen Blakeman

Karin Blakeman

Hoe krijg je weer controle over je zoekresultaten?

'You have to work so hard to get Google to behave... get a little bit of control back over Google!' Karin Blakeman, managing director van RBA Information Services, heeft een missie: ons duidelijk maken dat Google manipuleert en dat we niet moeten vergeten om de informatie die we krijgen te beoordelen op betrouwbaarheid. Google wijzigt ongevraagd zoektermen, personaliseert de resultaten, geeft foutieve en ambigue antwoorden, laat vaak na de bron te vermelden en met Google klikken we massaal op advertenties (wat we niet erg vinden zolang we worden geholpen). Google heeft maar één doel en dat is 'to make money' voor moederbedrijf Alphabet. Om de controle terug te krijgen kun je cookies verwijderen, je zoekresultaten depersonaliseren en je zoekcommando's verbeteren, bijvoorbeeld door *intext:* toe te voegen, om te zorgen dat Google de zoekterm daarachter niet laat vervallen, of door *filetype:pdf* toe te voegen als je op zoek bent naar officiële publicaties.

In plaats van veel dingen niet meer te doen, of anders te doen in Google, zijn er gelukkig prima gratis al-

Henk van Ess

Een online speurtocht

Henk van Ess neemt ons mee op een speurtocht: een reis door de tijd, door andere culturen en door spannende virtuele netwerken. Een speurtocht ook waarin je eigen vermogen om doortastend te zijn en om out of the box te denken in combinatie met je kennis over hoe het web is georganiseerd, het succes zullen bepalen.

In opdracht van organisaties zoals Daily Telegraph en CNN gaat Van Ess op zoek naar de identiteit van

ternatieven die je kunt gebruiken, zoals Bing. (Ik zelf slaakte overigens een zachte zucht van verlichting toen ik hoorde dat Google Scholar en Google Web niets met elkaar te maken hebben.)
Blakeman stipt ook het onderwerp 'free big data' aan, een beweging die door overheden wordt gesteund en die tot grote hoeveelheden free data leidt. Ze merkt op dat het erg lastig is om van deze 'raw data' chocola te maken. Ook hier geldt: 'You are on your own!' Gelukkig zijn we als mensen altijd in staat om elkaar te helpen. Karen rondt af met erop te wijzen dat het verstandig is altijd de bron te raadplegen als je een nieuwsbericht op social media onder ogen krijgt. Kennis over de bron geeft je informatie over betrouwbaarheid en je zult geregeld ervaren dat de bron niet betrouwbaar is. *(AvW)*

mensen van wie maar heel weinig bekend is. Denk bijvoorbeeld aan het fragment van Jihadi John, oftewel Mohamed Emwazi. Henk van Ess traint mediaprofessionals, vooral op het gebied van sociale media en datajournalistiek. Hij is geobsedeerd om in data zowel nieuws als verhalen te ontdekken en hij vertelt ons hoe hij het internet heeft afgespeurd naar wegen die naar de identiteit van Jihadi John zouden kunnen leiden.
Sommige wegen liggen ogenschijnlijk voor de hand, zoals het gebruik van (Instant) Google Street View of een online telefoonboek, andere zijn (althans voor mij) minder algemeen, zoals de geocheck waarmee je bestudeert wat er wordt getweet over een event, software waarmee je in rap tempo een filmbestand kunt scannen op een tekst/gesproken woord.
Van Ess maakt een paar dingen duidelijk: zodra je je eenmaal hebt begeven in het netwerk van social media en internet, is het vrijwel onmogelijk om daar een onzichtbare status voor terug te krijgen. Er is meer te vinden als je de juiste zoekcommando's gebruikt (sluit in een zoektocht naar YouTube-content bijvoorbeeld YouTube zelf eens uit door *-site:youtube.com* aan je vraag toe te voegen). En ook de organisatie van het internet an sich is een studie waard. (AvW)

Hoaxes voldoen gelukkig aan een aantal kenmerken, zodat ze systematisch kunnen worden opgespoord met machine learning algoritmes, zegt Srijan Kumar

Srijan Kumar

Hoaxes op Wikipedia

In de Verenigde Staten haalt 62 procent van de volwassenen het nieuws van social media, waarbij vaak wordt doorverwezen naar gratis content. Het gegeven dat op Wikipedia meer dan 20.000 'hoaxes' bekend zijn, staat dan plots in een ander daglicht. Met deze mededeling slaat Srijan Kumar, die een PhD doet in computerwetenschappen aan de Universiteit van Maryland (VS), onbewust een bruggetje naar de presentatie van Karen Blakeman. Van de betrouwbaarheid van Google stappen we zo over naar de betrouwbaarheid van Wikipedia.
Hoaxes zijn artikelen met bewust onjuiste gegevens of helemaal verzonnen broodjeaapverhalen. In geval van Wikipedia betreft het vaak lemma's over fictieve personen of gebeurtenissen. Feitelijk kan iedereen een hoax schrijven op Wikipedia. De vraag is hoe lang het gepubliceerd zal zijn en wie het in de tussentijd voor welk doel gebruikt. Sommige hoaxes overleven jaren lang voordat ze worden ontdekt en gekenmerkt als hoax. En pas met dat kenmerk zijn ze onschadelijk omdat de lezer, zolang ze nog niet verwijderd zijn, dan op zijn minst kan zien wat de status is.

Hoaxes voldoen gelukkig aan een aantal kenmerken zodat ze systematisch kunnen worden opgespoord met machine learning algoritmes. Deze kijken in eerste instantie naar structuur en vorm. Een hoax bevat vaak veel tekst en weinig doorverwijzingen naar het web en naar wikipedia. En als er verwijzingen naar pagina's op Wikipedia zijn, dan is de kans groot dat deze niet onderling met elkaar zijn verbonden. Ook kan gekeken worden naar gedrag van de editor en naar de kenmerken van zijn netwerk. De editor is vaak nieuw op Wikipedia en heeft nog niet eerder gepubliceerd. En een echte hoax-editor laat niet van zich horen als hij ontdekt dat zijn artikel is verwijderd. Uit een experiment van Kumar bleek dat zijn getrainde software 86 procent van de aangeboden hoax-artikelen als zodanig herkende en menselijke beoordelaars maar 66 procent.
Ik weet zeker dat ik na deze lezing de artikelen op Wikipedia van een hoaxscan zal voorzien, en als meerdere mensen dat doen, dan tonen we met deze lezing ook aan dat mediawijsheid een noodzakelijke vaardigheid is voor iedereen die het internet gebruikt als informatiebron. *(AvW)*

Henk van Ess: wanneer iemand zich eenmaal op social media heeft begeven: het is vrijwel onmogelijk om dan nog onzichtbaarheid terug te krijgen

Verder lezen

De presentaties van de lezingen zijn te vinden zijn op vogin-ip-lezing.net/presentaties-2017/