Archieven Linken met Semantische Zoekmachines

In toenemende mate worden grootschalige archieven toegankelijk gemaakt voor een breed publiek. Prominente voorbeelden worden gegeven door de archieven van landelijke dagbladen, nationale archieven, overheidsarchieven, archieven onder beheer van de Koninklijke Bibliotheek, televisiearchieven zoals beheerd door het Nationaal Instituut voor Beeld en Geluid en, meer algemeen, door archieven van erfgoedinstellingen.

Maarten de Rijke, Krisztian Balog, Marc Bron, Jiyin He, Bouke Huurnink, Valentin Jijkoun, Fons Laan, Edgar Meij, Manos Tsagkias, Andrei Vishneuski, Wouter Weerkamp
Universiteit van Amsterdam

Een archief is geen eiland. Gebeurtenissen beschreven in een nieuwsarchief krijgen een extra dimensie als zij gekoppeld worden aan beeldmateriaal. Historisch televisiemateriaal wint aan betekenis als het gekoppeld wordt aan contemporaine commentaren en nieuwsmateriaal uit de gedrukte pers. En meer specialistische of technisch georiënteerde archieven winnen aan bruikbaarheid als ze gekoppeld zijn aan achtergrondinformatie.

Onderzoek wijst uit dat eindgebruikers er bij gebaat zijn als koppelingen tussen archieven betekenisvol zijn en bijvoorkeur langs semantische lijnen lopen, met een sterke oriëntatie op entiteiten (mensen, locaties, organisaties, artefacten, etc.), op thema's (zoals 'stadsleven,' 'festiviteiten' of 'consumentencultuur') en op gebeurtenissen (zoals 'Praagse lente,' 'Opening van de Kanaaltunnel' of 'Marathon Amsterdam'). Betekenisvolle ontsluiting van archieven komt hiermee neer op zoek- en verkenningstechnologiën rondom entiteiten, thema's en gebeurtenissen plus hun onderlinge relaties.

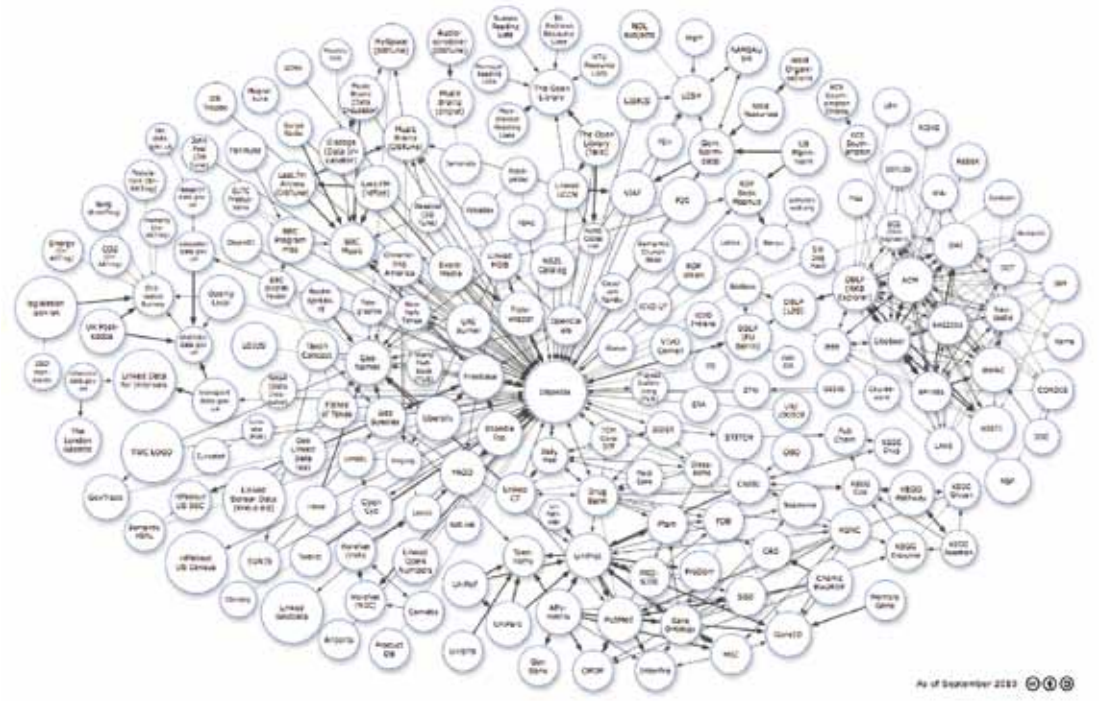
Gezien de omvang van de archieven die nu beschikbaar zijn of komen, zijn handmatige methoden om de gewenste koppelingen te leggen en om entiteiten, thema's en gebeurtenissen te identificeren in archiefobjecten eenvoudigweg niet realistisch. Een belangrijke beweging in onderzoek op het raakvlak van zoekmachinetechnologie en taaltechnologie betreft *semantisch* zoeken, waarbij de gewenste koppelingen tussen archieven langs de genoemde assen automatisch worden gelegd.

Onder de motorkap

Semantische zoekmachines stellen ons in staat om relevante entiteiten, thema's en gebeurtenissen te identificeren in grote hoeveelheden archief- of webdata. Dergelijke zoekmachines bouwen in belangrijke mate op taaltechnologie die erop gericht is om entiteiten thema's, gebeurtenissen en hun relaties in teksten te herkennen. Aan de Universiteit van Amsterdam werken we sinds 2008 aan een gedistribueerde omgeving, genaamd Fietstas, die de vereiste functionaliteit als webservice aanbiedt. Naast laag-niveau functionaliteit zoals tokeniseren en lemmatiseren biedt Fietstas semantische functionaliteiten zoals het herkennen van entiteiten en relaties, het normaliseren van entiteiten en het genereren van 'profielen' van entiteiten. Met name de laatste twee zijn interessant voor het koppelen van archieven. Het normaliseren van entiteiten betreft de volgende taak: welk object in de realiteit vormt de verwijzing van een voorkomen van een naam of beschrijving van een entiteit in een stuk tekst? Zie Figuur 1 voor een illustratie van het soort van documentannotaties dat daarbij door Fietstas gegenereerd wordt.

Parallel aan de genoemde taaltechnologie komen nieuwe algoritmes voor zoekmachi-

Figuur 1: Extractie van entiteiten.



Figuur 2: Verbonden data in de linked open data wolk.

nes op. De standaard informatie-eenheid (‘unit of retrieval’) verschuift van documenten naar scherper gedefinieerde eenheden, zoals entiteiten, netwerken van entiteiten, profielen van entiteiten, of relaties tussen entiteiten. Gebruikers van zoekmachines hebben meer nodig dan een platte lijst van links naar relevante documenten en moderne zoekmachines hebben de ambitie om hierin te voorzien middels *faceted search and exploration* waarin onderwerpsfacetten gecombineerd worden met facetten gebaseerd op entiteiten, thema’s, gebeurtenissen, locatie, tijd, genre, etc.

De twee - taaltechnologie en zoekmachinetechnologie met een nadruk op entiteiten en relaties - komen samen in een breed pallet aan taken. Bijvoorbeeld bij het automatisch genereren van hypertekstlinks, binnen Wikipedia of juist tussen specialistische documenten enerzijds en bronnen van achtergrondkennis andezijds, waarbij zowel op documentniveau als op het niveau van zoekvragen semantische structuur wordt ontdekt en benut.

Recente ontwikkelingen

Een belangrijke uitdaging voor thans vigerende algoritmes voor entity retrieval is dat zij zelden voor mensen interpreteerbare beschrijvingen weten te genereren van gevonden entiteiten of van de relaties tussen entiteiten. Linked Open Data (LOD) is een recente bijdrage van het opkomende semantische web dat de potentie heeft om de gewenste semantische informatie te leveren. De LOD-wolk bevat op dit moment miljoenen concepten uit honderden gestructureerde dataverzamelingen; zie Figuur 2. In lopend werk wordt onderzocht in hoeverre datagestuurde zoekmachine-algoritmes gecombineerd kunnen worden met gegevens uit de LOD-wolk. Voor het koppelen van archieven betekent dit werk een mogelijke verrijking van de te genereren koppelingen, zowel wat betreft hun aantal als wat betreft hun beschrijving en contextualisering.

Een tweede recente ontwikkeling die hier genoemd moet worden is het toepassen van deze koppeling van archieven in een dynamische context, bijvoorbeeld om te achterhalen hoe er in sociale media, zoals blogs, discussiefora en Twitter, gereageerd wordt op een nieuwsbericht. Een interessante uitdaging hierbij is het feit dat het taalgebruik in sociale media creatief is en niet onderworpen aan centrale redactie, waardoor te leggen koppelingen een fluïde karakter krijgen dat in hoge mate temporeel en contextueel bepaald is. Entiteiten spelen hierbij een sleutelrol, maar minstens zo belangrijk zijn de identiteit van deelnemers in online discussies en het gebruik van semantische hulpmiddelen zoals *hashtags*. Hiermee ontstaan koppelingen langs nieuwe - en voor een deel *sociale* - assen.

Conclusie

Met het steeds breder beschikbaar komen van grootschalige online archieven ontstaat

de noodzaak om deze aan elkaar en aan achtergrondkennis te koppelen. Gezien de omvang van dergelijke archieven zijn alleen automatische methoden een optie. Met behulp van moderne semantische zoekmachinetechnologiën zijn inmiddels de eerste stappen gezet om de gewenste koppelingen te genereren - van hoge kwaliteit en op grote schaal.

Dankwoord
Het type onderzoek dat hierboven beschreven wordt, is een onlosmakelijke combinatie van theoretisch, experimenteel en toegepast werk. Zonder data en use cases kan het niet uitgevoerd worden. We danken *NRC Handelsblad*, het *Nederlands Instituut voor Beeld en Geluid* en *Philips Medical Systems* voor discussies en het beschikbaar stellen van documentcollecties, zoekvragen en evaluatiemiddelen.

Het hier beschreven onderzoek wordt mede mogelijk gemaakt dankzij subsidies van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), onder projectnummers 612.066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011, het 7de Kaderprogramma van de EU, onder projectnummer 258191, het ICT Policy Support Programme van de EU, onder projectnummer 250430, het DuOMAn project dat wordt uitgevoerd binnen het STEVIN programma onder projectnummer STE-09-12, en door het Center for Creation, Content and Technology (CCCT).