

Gereedschap voor media-mining: automatische classificatie en clustering

Franciska de Jong (UT/TNO)

Al snel na de introductie van het begrip *data-mining*, ergens in de jaren '90, werd duidelijk dat 'knowledge discovery', zoals data-mining ook wel wordt genoemd, niet alleen 'verborgen' gegevens uit klassieke databases kon vergaren. Via text-mining, audio-mining en media-mining is het begrip inmiddels van toepassing verklaard op alle soorten digitale data, en is het woord *reality-mining* gevallen (meer dan 3000 hits op Google). En reality, wat valt daar eigenlijk *niet* onder? Ook zonder een antwoord op die filosofische vraag is evident dat de hoeveelheid en de diversiteit van de soorten kennis die met mining-technieken hanteerbaar kunnen worden gemaakt, enorm is. In dit artikel zal worden ingegaan op de vraag welke technieken er ingezet kunnen worden om een succes te maken van informatie- en kennisextractie voor onder meer multimedia-archieven.

Wat is media-mining?

Audiovisuele collecties zijn van oudsher ontsloten geweest op basis van metadata die beschikbaar waren op het moment van archivering: titel, maker, datum van opname, dataum van uitzending, thema, hoofdpersoon, etc. Over de inhoud van audio en video kon je alleen meer te weten komen door een bestand af te spelen of de makers te benaderen. Onze A/V archieven waren heel lang schatkamers zonder sleutel. Met de opkomst van technologie voor digitale opname en opslag wordt het steeds interessanter om hulpmiddelen voor ontsluiting van zwak-gestructureerde datacollecties toe te gaan passen op audio-visueel materiaal en de informatie plus de structuur die erin schuilgaat expliciet te maken. Twee technieken zullen hier worden besproken.

- clustering: het sorteren en bundelen van materiaal op basis van inhoud, te vergelijken met dossiervorming
- classificatie: het toekennen van labels aan documenten en/of de gevonden 'clusters'

Doel: de zoekmachinegenerator

Door de toenemende omvang en heterogeniteit van digitale media-archieven en multimedia-collecties is de Google-variant voor zoektechnologie niet langer toereikend. Het is tijd voor is ontGoogling. Met Google zoek je altijd op dezelfde manier, voornamelijk in tekst, maar eventueel ook in ondertitels bij plaatjes. Soms leidt dat tot tevredenheid, maar heel vaak moet een gebruiker zich door grote aantallen irrelevante hits heenploegen. De *precisie* van full-text zoekmachines, hoe geavanceerd en snel ook, laat vaak te wensen over. De hoeveelheid bronnen en bestandsformaten, de structuur in het web (verbanden tussen collecties) en de diversiteit in gebruikersgedrag neemt nog alsmaar toe. Wat we eigenlijk nodig hebben is een zoekomgeving die zich automatisch aanpast. Aan de kenmerken van de gebruikers (talenkennis, achtergrondkennis), aan de context van waaruit hij zoekt (locatie, tijdstip), aan de collecties waarin de gebruikers geïnteresseerd is en aan de beschikbare bandbreedte. Niet een zoekmachine, maar een

omgeving die automatisch de instellingen van de zoektechnologie optimaliseert: een zoekmachinegenerator.

Lange termijn

Onder meer door de opkomst van probabilistische zoektechnologie zal het in de toekomst mogelijk worden om zoekresultaten voor tekst, geluid en beeld en te integreren. Verder zullen collecties en archieven gekoppeld kunnen worden zonder dat aanpassing van de locale ontsluitingspraktijk nodig is. We zullen niet alleen X-language (cross-language) kunnen zoeken, maar ook X-medium, X-collection, X-domain. Dat zal bijdragen aan flexibele gebruikersomgevingen aan de ene kant, en versterkte autonomie in het beheer van archieven aan de andere kant.

Wat nu al kan

Met media-mining kan de inhoud van media-archieven voorgestructureerd worden. Documenten over hetzelfde onderwerp kunnen met clustering bij elkaar worden geplaatst. Alleen clusters die qua onderwerp aansluiten bij de zoekterm worden dan doorzocht en geordend op relevantie. Dat draagt bij aan de precisie. Bij het aanleggen van een clusterstructuur kan natuurlijk gebruik worden gemaakt van bestaande metadata, maar de term 'mining' is vooral van toepassing als niet of nauwelijks gebruik gemaakt kan worden van vooraf beschikbare metadata. Een cruciale techniek is machine learning: zonder voorkennis 'leert' het systeem de semantische gelijkensissen tussen documenten. Hoe dat leren verloopt is het gemakkelijkst uit te leggen. voor tekstueel materiaal. Op basis van de woorden en woordcombinaties die in een tekst voorkomen wordt als het ware een vingerafdruk gemaakt van een document: een statistisch profiel waarin de frequentie is verwerkt waarmee woorden in elkaars nabijheid voorkomen en de frequentie van die woorden in het algemeen. Documenten die qua onderwerp overeenkomen zullen vergelijkbare woordcombinaties bevatten en kunnen op basis daarvan worden geclusterd. Ook kan worden bepaald of een cluster een hoge nieuwswaarde heeft door het profiel te vergelijken met bestaande profielen.

Voor informatie in andere modaliteiten dan tekst zijn varianten van deze technologie van toepassingen. Voor video-materiaal bijvoorbeeld kan een profiel worden opgebouwd met behulp van beeldkenmerken en met behulp van clustering kan materiaal met vergelijkbare kenmerken bij elkaar worden gevoegd. Technologie voor beeldanalyse kan tot nu toe echter maar een beperkt aantal beeldkenmerken automatisch vaststellen. Een alternatief is daarom om al het 'talige' materiaal zoals ondertitels of spraak te benutten, en daarop clustering toe te passen.

Bij classificatie gaat het erom om aan een document of aan een cluster van documenten een onderwerpslabel toe te kennen. Ook dat is een techniek waarbij machine learning kan worden ingezet. Een collectie van documenten die handmatig zijn voorzien van een label uit een relevante thesaurus of ontologie (termenlijsten) fungeert als leer- of trainingsset. Zo'n systeem leert daaruit welke typerende woordcombinaties optreden bij documenten over een bepaald onderwerp. Kostbare kennis die in gestolde vorm aanwezig is in een handmatig geannoteerd archief kan op die manier worden hergebruikt.

Volgende stap: hiërarchische clusters

Zoals gezegd kan bij het aanmaken van clusters aan de hand van de frequenties van woordcombinaties of andere kenmerken worden bepaald welke documenten bij elkaar horen. De overeenkomst in profiel wordt uitgedrukt in een getal en experimenteel wordt er een grenswaarde vastgesteld voor documenten die aan eenzelfde cluster kunnen worden toegewezen. Maar ‘bij elkaar horen’ of ‘op elkaar lijken’, dat zijn relatieve begrippen. Documenten kunnen meer of minder op elkaar lijken. Ook kun je in een cluster van documenten natuurlijk vaak subgroepen aanwijzen. Het ligt dus voor de hand om bij clustering op zoek te gaan naar hiërarchische structuur. Hiërarchisch clusteren kan grofweg op twee manieren gebeuren: door de collectie in eerste instantie op te vatten als één cluster en vervolgens in stappen op te splitsen in deelclusters (top-down), of door te beginnen bij individuele documenten en die op meerdere niveaus te groeperen (bottom-up). In alle gevallen is het de kunst om een goede grenswaarde in te stellen voor het al dan niet combineren of splitsen van clusters.

De keuze tussen dit soort alternatieven wordt vaak bepaald door allerlei praktische bijkomstigheden: schaalbaarheid is niet voor alle varianten van de techniek gegarandeerd, terwijl clustering juist bij grote aantallen documenten toegevoegde waarde heeft. Een ander kritisch punt is soms de gewenste voorspelbaarheid. Door allerlei statistische bewerkingen kan de uitkomst van met name top-down clustering voor grote collecties per keer fluctueren. Er zijn toepassingsscenario's waarbij dat ongewenst is. En verder geldt vaak de eis om de technieken niet alleen toe te kunnen passen op gefixeerde collecties, maar ook op zich uitbreidende collecties, zoals in het geval van een nieuwsarchief. Daarvoor wil je niet elke dag opnieuw met clusteren beginnen. Neem als voorbeeld een dossier over veeziekten. Na de eerste berichten over het uitbreken van een ziekte kunnen er regelmatig nieuwe gevallen gemeld worden die dan aan een bestaand dossier moeten worden toegevoegd. Ook kan de aanleiding om een dossier te splitsen pas na verloop van tijd ontstaan.

Combinaties met andere technieken

Er zijn verschillende ontsluitingstechnieken die voldoende uitgerijpt zijn om met clustering en classificatie gecombineerd te kunnen worden: spraakherkenning (voor het genereren van transcripties van gesproken audio), taaltechnologie (termextractie, taalidentificatie, analyse van de opbouw van woorden), indexering (bepaling van de meest onderscheidende termen) en presentatie (chronologie, headline-extractie, samenvatting). Spraakherkenning die bedoeld is voor ontsluiting van archieven wijkt in een aantal opzichten af van de vormen van spraakherkenning waarmee veel mensen al vertrouwd zijn, zoals dicteertechnologie en dialoogsysteem voor telefonische inlichtingen. Het gaat bij media-ontsluiting om het afleiden van transcripties, ofwel: tekstuele weergaven, die gebruikt kunnen worden bij het zoeken naar relevante passages: Googlen in spraak, ook wel aangeduid als audio-mining. Dat is op zich al een interessante ontwikkeling (zie bijvoorbeeld <http://www.willemfrederikhernans.nl/>), maar voor een heteroog media-archief, waarin behalve spraakmateriaal ook gerelateerde teksten of videomateriaal en/of archiefstukken zijn opgenomen, kan de combinatie met clustering uiterst krachtig uitpakken. Je kunt daarmee automatisch multimedia-dossiers opbouwen, waarbinnen *cross-media* gezocht en gebladerd kan worden. Om niet alleen de documenten, maar ook het dossier zelf van metadata te voorzien kan het interessant zijn om technieken toe te passen als automatische samenvatting en extractie van eigennamen en headlines.

Overzicht van toepassingsdomeinen

De structurering van ruwe content heeft een breed scala van toepassingen. Een aantal voorbeelden:

- media-industrie: omroeparchieven, krantenarchieven, nieuwsdiensten
- e-culture: interviews, oral-history-archieven (getuigenissen, locatiebeschrijvingen, etc.)
- content-management: ontsluiting van opnames van vergaderingen (raadsvergaderingen, professionele teams, bestuursvergaderingen, teleconferenties)
- security: ontsluiting van spraak en/of beeld, ofwel 'reality recordings'
- industrie- en transportsector: patronen in incidentrapportages
- medische sector: medische documentatie (tekst, spraak en/of beeld), al dan niet in combinatie met meetgegevens (numerieke data, sensordata)
- telecomsector: gebruikersgedrag (doorklikpatronen voor webpagina's, interactiepatronen tussen gebruikers van mobiele telefoons, al dan niet in combinatie met de inhoud van de communicatie)

Conclusie

De technieken en toepassingen die hier zijn beschreven zijn voor een deel al gerealiseerd, en voor een deel toekomstmuziek. Een van de mogelijke obstakels voor introductie is de koppeling aan bestaande workflow en processen. Een naadloze integratie met bestaande tools voor content-management vereist samenwerking van diverse partijen. Verder hebben zowel clustering als classificatie baat bij training. De statistische modellen leveren betere resultaten op als ze getraind worden met voorbeeldmateriaal dat al handmatig geclusterd en/of geclassificeerd is. Dat is een arbeidsintensief proces waarvoor niet altijd de middelen beschikbaar zijn. Veelbelovend daarentegen is dat door het gebruik van statistische benaderingen de analyse van tekst en beeldmateriaal steeds meer naar elkaar toegroeien. Op termijn kunnen beeldkenmerken en tekstuele features vermoedelijk in vergelijkbare modellen worden verwerkt, en kunnen beeld en tekst niet allen op dezelfde manier, maar ook geïntegreerd worden doorzocht.

Op veel terreinen moet overigens nog een begin gemaakt worden met experimenteren, maar er komt almaar meer materiaal in digitale vorm beschikbaar dat impliciete informatie en kennis bevat die met behulp van *mining* boven water kan komen. In combinatie met profielen voor gebruikers en gebruikscontext kan dat allemaal worden benut om het idee van een zoekmachinegenerator te realiseren.

Thema's voor in aparte kaders

A. Lopende onderzoeksprogramma's

Er zijn verschillende onderzoeksprogramma's waarbinnen classificatie- en clusteringtechnieken voor taal-, spraak- en beelddata worden onderzocht en ontwikkeld. Het is een onderzoeksterrein waaraan zowel ICT-groepen als domeinspecialisten werken.

- bsik-programma's

- MultimediaN (gericht op content uit het nieuwsdomein en op e-culture; <http://www.multimedien.nl/>)

- BioRange (gericht op content uit het bio-medische domein; <http://www.nbic.nl/NL/biorange.html/>)
- NWO
 - CATCH (Continuous Access to Cultural Heritage: <http://www.nwo.nl/catch/>)
 - IMIX (Interactieve Multimodale Informatie-eXtractie: <http://www.nwo.nl.imix/>)
- SENTER
 - Waterland (content-ontsluiting ten behoeve van archivering bij de publieke omroepen)
- EU-IST
 - M4 (MultiModal Meeting Manager; ontsluiting van opnames van vergaderingen <http://www.m4project.org/>)
 - AMI (Augmented Multiparty Interaction: tools voor browsing van multi-modale registraties van vergaderingen; <http://www.amiproject.org/>)

B. Public domain toolkits

Voor sommige vormen van clustering zijn toolkits voor machine learning beschikbaar die gratis kunnen worden gedownload. Geïnteresseerde lezers kunnen onder meer terecht op: <http://www.cs.waikato.ac.nz/~ml/>

C. Internationale competitie

Onderzoek naar clusteringstechnologie voor media-archivering krijgt in internationaal verband onder meer aandacht in de vorm van benchmark-evaluaties voor topic-detectie en topic-tracking (TDT). Deelnemende onderzoeksgroepen voeren ieder dezelfde taken uit op dezelfde dataset uit om de prestatie van hun systemen onderling te kunnen vergelijken. De TDT-evaluatie wordt georganiseerd door het Amerikaanse National Institute for Standards and Technology (<http://www.nist.gov/speech/tests/tdt/>). Er zijn vergelijkbare evaluaties voor onder meer tekstretrieval, zoektechnologie voor biomedische informatie (genomics) en videoretrieval.

D. Personalialia:

Franciska de Jong is hoogleraar taaltechnologie bij de onderzoeksgroep Human Media Interaction van de faculteit Electrotechniek, Wiskunde, en Informatica van de Universiteit Twente. Daarnaast is zij verbonden aan TNO. Voor meer details: <http://hmi.ewi.utwente.nl/~fdejong/>