Hulptroepen bij de ontsluiting van av-materiaal

Beeld en Geluid gastheer van twee CATCH projecten

[Inleiding]
CATCH (Continuous Access To Cultural Heritage) is een nationaal onderzoeksprogramma van de NWO-gebieden exacte- en geesteswetenschappen. Binnen CATCH worden methoden en technieken ontwikkeld waarmee erfgoedbeheerders hun digitale collecties beter toegankelijk kunnen maken. Informaticaonderzoekers en erfgoedinstellingen als het Rijksmuseum, de Koninklijke Bibliotheek, het Nationaal Archief, Naturalis en het Nederlands Instituut voor Beeld en Geluid werken in CATCH nauw samen. De onderzoekers verrichten het grootste deel van hun onderzoek dan ook binnen de instelling, die optreedt als projectbegeleider, eindgebruiker, en leverancier van data en content. Beeld en Geluid is betrokken bij twee CATCH deelprojecten: CHOICE (CHarting the informatiOn landscape employIng ContExt Information) en MUNCH (MUltimedia aNalysis for Cultural Heritage).

De Beeld en Geluid context

Meer materiaal sneller en beter beschikbaar kunnen maken voor gebruikers. Dit is de kern van Beeld en Geluid's motivatie voor deelname aan CATCH. Onder invloed van de digitalisering van het omroepproductieproces is het collectiebeleid van Beeld en Geluid ingrijpend gewijzigd: er wordt niet langer geselecteerd. Hierdoor stroomt inmiddels het complete Nederlandse programma-aanbod van de publieke omroepen het archief binnen. Als gevolg daarvan is het aantal te catalogiseren radio- en televisieprogramma's meer dan verdriedubbeld, naar ruim 32.000 uur op jaarbasis. Daarnaast wordt een groot deel van de bestaande collecties opnieuw beschreven, in het kader van het project Beelden voor de Toekomst. Technieken die het ontsluitingsproces kunnen ondersteunen en versnellen worden bij Beeld en Geluid dan ook enthousiast omarmd.

Het beschrijven wordt nu voor het overgrote deel handmatig gedaan. Ondanks het nieuwe geavanceerde catalogussysteem IMMIX en de (deels) geautomatiseerde metadataworkflow vanuit de productieomgeving (zie eerdere artikelen hierover in IP 2006-04 en IP 2007-06) is dit nog altijd een dure en tijdrovende bezigheid. De meeste programma's kunnen alleen toegankelijk worden gemaakt op formele gegevens en een vaak summiere inhoudsbeschrijving. Willen gebruikers details, dan zullen ze de programma's in hun geheel moeten afspelen. Om de enorme hoeveelheid materiaal duurzaam te kunnen ontsluiten (denk ook aan nieuwe aanwas zoals internetproducties en user generated content) streeft Beeld en Geluid naar een situatie waarin uiteindelijk 80% van de beschrijvingen (semi) automatisch wordt gegenereerd.

Het CHOICE-project

Aan CHOICE wordt gewerkt door onderzoekers van het Max Planck instituut, de Vrije Universiteit en het Telematica-instituut. Het belangrijkste doel van dit project is om automatisch ontleende thesaurustermen voor te stellen aan documentalisten tijdens het catalogiseerproces. Deze suggesties worden met behulp van Natural Language Processing (NLP) en Semantische webtechnieken afgeleid uit contextuele tekstbronnen over radio of televisieprogramma's zoals websites, online televisiegidsen, recensies en kijkcijferonderzoek. De methode is niet gebaseerd op training en heeft dus geen vooraf geannoteerde datasets nodig. Er wordt gebruik gemaakt van de impliciete en expliciete kennis die vastligt in de GTAA, de Gemeenschappelijke Thesaurus Audiovisuele Archieven die bij Beeld en Geluid in gebruik is.

De termen binnen de GTAA zijn verdeeld over zes lijsten ofwel 'assen': Onderwerp, Genre, Maker, Persoon, Geografische Naam en Naam. Het projectteam heeft deze assen

omgezet naar een internationale webstandaard. Er werden links aangebracht *tussen* de verschillende assen en en verrijkte de onderwerpsas met synoniemen, enkelvoudvormen en Engelse vertalingen. Vervolgens is een corpus samengesteld: 258 afleveringen van televisieprogramma's als Andere Tijden, Beeldenstorm, de Donderdag Documentaire, de Nieuwe Wereld en Dokument, met een daarbij behorend aantal van 364 contextdocumenten. Voor dit (documentaire) genre werd gekozen omdat de verwachting is dat de CHOICE methode hier het meeste profijt oplevert: er is veel aanbod, de vraag naar snelle, gedetailleerde ontsluiting is hoog en er zijn meerdere actuele websites voorhanden die relatief veel tekst bevatten.

Het eigenlijke proces van het analyseren en suggereren van trefwoorden verloopt volgens een denkbeeldige 'annotatiepijplijn'. Eerste stap is het inzetten van Algoritme no. 1. Dit algoritme detecteert die segmenten van de contextdocumenten, waarin mogelijk trefwoorden aanwezig zouden kunnen zijn. Deze termen worden gelinkt aan de eigenlijke GTAA termen. Algoritme no. 2 gebruikt vervolgens de relaties tussen de gevonden trefwoorden om semantische clusters te maken, deze te wegen en de meest centrale termen als suggestie te presenteren aan de documentalist, die op basis daarvan de juiste thesaurusterm kiest. Experimenten hebben inmiddels uitgewezen dat ongeveer de helft van de eerste 10 suggesties een goede is. Deze 10 bestrijken 65% van de trefwoorden die documentalisten hadden toegekend.

De automatische methode levert per document veel meer trefwoorden op dan de handmatige. Elk afzonderlijk beschrijven deze 'automatische' trefwoorden het programma daarbij minder precies. De overlap tussen deze sets levert echter een interessante nieuwe manier van zoeken op, nl. naar de inhoudelijke overeenkomst tussen (groepen) documenten. Immers: 7 automatisch ontleende, overlappende en weinig precieze trefwoorden pikken een andere inhoudelijke overeenkomst op dan dat ene, heel precieze en specifieke trefwoord dat door een documentalist wordt toegekend.

Het CHOICE project wil uiteindelijk een raamwerk creëren voor het trefwoordsuggestie-systeem: de CHOICE Documentalist Support omgeving. Hiertoe wordt momenteel gebouwd aan prototypes voor een database voor contextdocumenten en voor een geïntegreerde zoekomgeving voor metadata, teksten, semantische annotaties en AV-materiaal. Vervolgonderzoek gaat zich ook bezighouden met de beoordeling van de suggesties. Zitten er aperte fouten in? Worden er termen totaal gemist? Ook de (relatieve) waarde van de gesuggereerde termen voor de verschillende gebruikerstypen (i.c. documentalist, professionele omroepklant, algemeen publiek) wordt nader onderzocht.

Het MUNCH-project

MUNCH - met als technische partners de Universiteit van Amsterdam en de Vrije Universiteit - richt zich op de automatische analyse van digitale bewegende beelden. Er is gekozen voor een multimodale aanpak: aangestuurd door een ontologie worden de beeldeigenschappen van shots geanalyseerd *samen met* de (geschreven en gesproken) taal die dat beeld beschrijft. Het project bestrijkt daarmee niet één maar drie researchgebieden.

Als eerste onderdeel heeft het projectteam een multimediathesaurus gebouwd. Deze bestaat uit een set high-level detectoren in een onderlinge structuur. Met deze detectoren kunnen automatisch 'beeldconcepten' worden opgespoord, zoals 'opstijgende vliegtuigen', 'George Bush sr.', 'de Amerikaanse vlag', een explosie, een boot of een persoon. De detectoren worden getraind door een algemene automatische methode te combineren met handmatig geannoteerde beelden uit het nieuws- en actualiteitendomein. Er zijn er inmiddels ruim 400.

Het MUNCH team heeft dit lexicon verrijkt met semantische beschrijvingen en structuur, verkregen vanuit het online woordbestand WordNet. Men heeft ook onderzocht in hoeverre de relaties (en niet alleen de losse termen) binnen de GTAA kunnen worden gebruikt. De Beeld en Geluid thesaurus werd nog eens uitgebreid met nieuwe, afgeleide verbanden. Het testen van de uitgebreide GTAA wees uit dat er een gunstig effect uitgaat van de integratie met de semantiek van andere, bredere thesauri. Voor het ontologie-werk werd voor een deel samengewerkt met het CHOICE team.

Om het zoeken in transcripten van het gesproken woord te verbeteren experimenteert MUNCH met het gebruik van taaltechnologie. Men werkt daarbij aan het probleem van de temporele 'mismatch' tussen het noemen van objecten in spraak en hun verschijning in het beeld. Er is een oplossing gevonden door de spraak die hoort bij shots vóór en na het shot met het betreffende object, mee te nemen. Deze shots worden gewogen: hoe verder weg van het 'hoofdshot', hoe lager het gewicht. Het betrekken van de spraak in de naburige shots verbeterde het zoekresultaat aanzienlijk.

De ontologie-, spraak- en visueelgebaseerde methodes van MUNCH zijn geëvalueerd met een zoekopdracht. Een van de uitkomsten was dat de resultaten van de drie strategieën vergelijkbaar waren. Interessanter was dat werd aangetoond dat er winst geboekt kan worden door de methoden te combineren.

Het MUNCH team gaat zich nu richten op het verder verbeteren van de zoekresultaten door het combineren van heterogene informatiebronnen zoals metadata van documentalisten, ondertitels, automatische annotaties, contextdocumenten en nog meer externe thesauri. Om de resultaten te kunnen toetsen aan gebruikersbehoeften heeft MUNCH ook het initiatief genomen tot de inrichting van een logsysteem voor het vastleggen van het gebruikersgedrag op de Beeld en Geluid website. Een dergelijke omvangrijke databron moet het projectteam gedetailleerd inzicht verschaffen in de gebruikerstypen, het zoekgedrag, de zoekvragen en de relaties tussen deze elementen. MUNCH zal deze data analyseren om in de tweede helft van het project specifieke usecases te kunnen opzetten.

De verwachtingen

Het resultaat van het CHOICE-project wordt binnen Beeld en Geluid prettig snel zichtbaar: de implementatie van een 'Documentalist Support omgeving' is nu al terecht gekomen op de ICT roadmap. Door de omzetting van de eigen GTAA naar een internationale webstandaard kan straks gemakkelijker worden gezocht in andere culturele collecties. De CHOICE methode creëert ook nieuwe opties: zoeken naar en in contextbronnen en zoeken naar groepen vergelijkbare documenten.
Vanuit het perspectief van de documentalist zullen de trefwoordsuggesties bepaald (nog) niet perfect zijn. Niettemin zal met behulp van deze 'tips' sneller en meer consistent kunnen worden gekozen voor de juiste thesaurusterm.

MUNCH heeft een horizon die – waar het gaat om de omzetting in implementeerbare technologie – wat verder weg ligt. Dit type geavanceerde research, de koppeling van beeldanalyse, taal- en spraaktechnologie en het redeneren met ontologieën, vertegenwoordigt een relatief nieuwe onderzoeksrichting. Een televisieprogramma kan binnen de MUNCH strategie dan ook voor het eerst automatisch worden benaderd als een geïntegreerd product van beeld, gesproken woord en geassocieerde teksten en metadata. De kennis en informatie die in al deze bronnen ligt opgesloten kan straks binnen één en dezelfde zoekactie worden aangeboord. Deze benadering komt tegemoet aan de complexe, gelaagde semantiek van een videoproductie. Het is dan ook de verwachting dat het multimodale werk van MUNCH uiteindelijk relatief complete zoekresultaten zal opleveren.

Voor beide projecten geldt dat ze belangrijk bijdragen aan de ontwikkeling van technieken voor automatische en semi-automatische ontsluiting van *Nederlands* audiovisueel materiaal. Resultaten worden geïntegreerd en gecombineerd met de uitkomsten van andere CATCH-projecten en komen aldus beschikbaar voor andere erfgoedinstellingen.

Annemieke de Jong is projectleider van CHOICE en MUNCH namens Beeld en Geluid