ADVIES ONTWIKKELTRAJECTEN VOOR HET IRN WOPR PROJECT

Dr. Edgar Meij

Prof. Dr. Maarten de Rijke

UNIVERSITEIT VAN AMSTERDAM

Inhoudsopgave

2 Inleiding

Het iRN (Internet Recherche & Onderzoek Netwerk) is een op Internet technologie gebaseerd netwerk, ontwikkeld door de Politie Gelderland-Zuid. Het is beschikbaar voor alle overheidsdiensten uit de Openbare Orde en Veiligheidssector en in- en externe partners, waarmee wordt samengewerkt op het gebied van innovatie en ontwikkeling. Het iRN verzorgt, naast Internet toegang op een forensisch technisch geborgde manier, een netwerk voor samenwerking, kennis delen en ontwikkeling en innovatie op het gebied van intelligent Internet onderzoek. Het iRN is volledig gebaseerd op open source software en open standaarden.

Dit document bevat een verkenning van de mogelijkheden tot onderzoek en implementatie van onderzoeksmodules, uit te voeren door de Universiteit van Amsterdam (UvA) in het kader van het iRN Open Web Observatie project. De modules zijn onderverdeeld aan de hand van een aantal thema's, zoals "zoektechnologie" en "taaltechnologie." Binnen de thema's wordt ingegaan op specifieke onderzoekslijnen waarbij voor ieder item nader wordt beschreven wat het nut is voor het iRN, wat de huidige stand van zaken is en hoeveel werk er nodig is teneinde ieder onderdeel succesvol in te zetten binnen het iRN.

Aan de hand van interviews is een aantal hoofdlijnen en specifieke toepassingen van mogelijke eindgebruikers geëliciteerd. De interviews worden beschreven in het volgende hoofdstuk en zullen worden gebruikt als specifieke use-cases binnen de onderzoekslijnen. De onderzoekslijnen komen daarna aan bod, gevolgd door een aantal concrete toepassingen die aangepakt kunnen worden door een of meerdere onderzoekslijnen te combineren.

3 Interviews

Door middel van interviews met belanghebbenden, betrokkenen en/of eindgebruikers van het iRN is getracht een beeld te vormen van workflows en use-cases die middels gebruik van het iRN opgelost kunnen worden. Als input voor deze interviews is een aantal tools en technieken (ontwikkeld door de UvA) gebruikt; deze zullen terugkomen in Hoofdstuk 3.

De personen die zijn geïnterviewd zijn:

- Eric den Hartogh en Peter van Lierop (Belastingdienst)
- Shanti Morsing en Jennifer van 't Woudt (KLPD)
- Ron Boelsma en Vanessa Dirksen (KLPD)

De vragen die aan bod zijn gekomen in de interviews zijn samen te vatten in de volgende twee kernvragen:

- Stel dat je een webcrawling raamwerk zoals het iRN tot je beschikking zou hebben, wat zou je daarmee kunnen?
- Waar ben je veel tijd mee kwijt in de context van web-gebaseerd onderzoek? Zijn er dingen die nu nog niet of slechts beperkt mogelijk zijn?

Vervolgens is er ingegaan op specifieke tools en technieken die zijn of worden ontwikkeld aan de UvA. Het vervolg van de gesprekken was derhalve opgezet rondom een aantal toepassingen die mogelijk deel uit gaan maken van het iRN raamwerk. Voor iedere toepassing is besproken of en hoe deze ingezet zou kunnen worden in de huidige

Naam	Paragraaf	Projecten	Implementatie	Onderzoek	
Zoekmachines op maat	3.1.1		⊖	⊖	⊕
Meta search/ Data fusie	3.1.2		⊕	⊕	⊕
Adaptief filteren	3.1.3	Peilend.nl	⊕	⊖	⊕
Interactief zoeken	3.1.4		⊕	⊖	⊕
Entiteiten zoeken	3.1.5	Fietstas, EARS, Penta Politica	⊖	⊖	⊕
Entiteiten en hun samenhang	3.1.6	SaHaRa	⊖	⊖	⊕
Metadata	3.1.7	Fietstas, Peilend.nl	⊖	⊖⊖	⊕
Longitudinaal zoeken	3.1.8		⊕	⊖	⊕⊕
Text Mining	3.2.1	Fietstas	⊖	⊖	⊖
Sentiment analyse	3.2.2	Fietstas	⊕	⊖	⊕⊕
Duplicaat detectie	3.2.3	Fietstas	⊖	⊖	⊖
Automatisch samenvatten	3.2.4	Fietstas	⊕	⊖	⊕
Web 3.0	3.2.5		⊕	⊖	⊕
Focused crawling	3.2.6	Ssscrape	⊕	⊖	⊕⊕
Taaldetectie	3.2.7		⊖	⊖	⊖
Machinaal vertalen	3.2.8		⊕⊕	⊕⊕	⊕⊕
Identity tracking	3.3.1	Fietstas, EARS	⊕	⊕	⊕⊕
Persoonsanalyse	3.3.2	EARS	⊖	⊖	⊕⊕
Graaf/Relatie analyse	3.3.3		⊕	⊖	⊕⊕
Concept-, scene-subjectherkenning	3.4.1	Impala	⊕	⊖	⊕⊕
Objectherkenning	3.4.2	Impala	⊕⊕	⊕⊕	⊕⊕
Multimedia linken	3.4.3		⊕⊕	⊕⊕	⊕
Informatie visualisatie	3.4.4	MediaTable	⊕	⊖	⊕⊕

Tabel 1. Overzicht onderzoekslijnen, met bijbehorende paragrafen en relevante projecten binnen de UvA. Middels iconen wordt weergegeven wat de relatieve effort is in termen van implementatie en onderzoek (fundamenteel en adaptief resp. – zie hoofdstuk 3), variërend van zeer weinig ('--'), weinig ('-') en gemiddeld ('+') tot zeer aanzienlijk ('++').

workflow van de geïnterviewden. Ieder van deze werd beschreven en vervolgens werd voor ieder van hen gevraagd:

- Zou dit nuttig zijn binnen je huidige workflow? Kan je een specifiek probleem bedenken dat hiermee opgelost zou kunnen worden? Zo ja, zijn er specifieke dingen waar op gelet moet worden? Zo nee, waarom niet?

Tot slot is er nog ruimte geboden voor een korte brainstorm sessie, aan de hand van de vraag "Aan welke andere zaken moet je denken als je hoort over deze tools en technieken?" Het doel hiervan was het verkrijgen van informatie rondom de ontwikkeling van nieuwe tools en technieken die ondersteuning bij dan wel uitbreiding van kunnen vormen ten aanzien van huidige analyse methoden.

Twee doelgebieden werden door meerdere geïnterviewden genoemd als uiterst interessant: (i) tekstanalyse en (ii) sociale netwerk analyse. Voor beide werd opgemerkt dat het nuttig zou zijn als er relatief simpele analyses voorhanden zouden zijn die gemakkelijk ingebed zouden kunnen worden in complexere workflows (bestaand en/of nieuw). Voorbeelden van dergelijke analyses zijn het achterhalen van de vrienden (van vrienden van) bepaalde Twitteraars of het automatisch identificeren van plaatsnamen in stukken tekst. Aan de hand van de verkregen use-cases wordt concrete invulling gegeven aan de onderzoekslijnen die beschreven worden in het volgende hoofdstuk. Andere, meer complete toepassingen die voortkwamen uit de gesprekken komen terug in hoofdstuk 4.

4 Onderzoekslijnen

In dit hoofdstuk beschrijven we aan de hand van "thema's" verschillende meer specifieke onderzoekslijnen. Deze vormen tezamen de onderzoeks- en ontwikkeltrajecten op macro (thema) of micro (lijnen) niveau. De thema's die we beschrijven zijn *zoektechnologie*, *taaltechnologie*, *sociale netwerk analyse*, *beeld en video analyse* en *machinaal vertalen*. Voor iedere lijn geven we een korte beschrijving, een overzicht van relevante personen en/of projecten binnen de UvA en een indicatie van de benodigde hoeveelheid onderzoeks- en implementatiewerk dat benodigd is om de lijn in te zetten binnen het iRN. In Tabel 1 is een overzicht te vinden van de laatstgenoemde punten. We merken tot slot op dat de mate van benodigd onderzoekswerk beïnvloed wordt door twee factoren: (i) "fundamenteel" onderzoek, dat gebezigd moet worden om basisfunctionaliteit te verkrijgen en (ii) onderzoek dat uitgevoerd moet worden in het kader van taak, domein en data adaptatie; een ander toepassingsdomein legt andere eisen op aan de technologie en is daarmee niet altijd direct toepasbaar. Veelal impliceert een lage implementatie effort een lage mate van fundamenteel onderzoek maar toch nog enige (soms aanzienlijke) vorm van adaptatieonderzoek.

4.1 Zoektechnologie

Het Internet is in de afgelopen decennia een integraal onderdeel geworden van onze samenleving. De manier waarop we informatie tot ons nemen, creëren en uitwisselen is daarmee aanzienlijk veranderd. Ontwikkelingen rondom taal- en zoekmachinetechnologie dragen bij aan, en profteren van, deze verandering: teneinde wijs te kunnen worden uit de alsmaar groeiende hoeveelheden informatie hebben we slimme zoekmachines nodig die op grote schaal teksten kunnen doorzoeken, samenvatten en analyseren.

De algoritmes die deel uitmaken van zoekmachines kunnen ook dienst doen als we onderzoek willen doen naar informatie op het internet. Gezien de immer groeiende schaal zijn handmatige methoden niet langer toereikend en hebben we automatische

technieken nodig. Daarnaast bieden dergelijke technieken ook nieuwe vormen van analysemogelijkheden, die voorheen niet of nauwelijks mogelijk waren. Hieronder beschrijven we specifieke onderzoekslijnen die verband houden met zoektechnologie.

4.1.1 Zoekmachines op maat

In het geval van "normaal" zoeken (ook wel "ad hoc" zoeken genaamd) is er een gebruiker die een zoekvraag ingeeft die typisch bestaat uit een of meerdere sleutelwoorden. De gebruiker bekijkt de resultaten die de zoekmachine teruggeeft en past vervolgens eventueel de zoekvraag aan. Zoekvragen kunnen verschillende vormen aannemen. De meest gangbare variant bestaat uit enkele sleutelwoorden, met of zonder structuur. Een compleet andere variant bestaat bijvoorbeeld uit meerdere documenten. In het laatste geval kan de bijbehorende informatiebehoefte "Vind voor mij zoveel mogelijk documenten die lijken op deze" zijn. Ook kunnen zoekvragen gepaard gaan met metadata, aan de hand waarvan de gebruiker de zoekresultaten wil sturen en/of inperken. Voorbeelden hiervan zijn onder meer een tijdspanne waarin of een bepaalde auteur door wie de documenten verschenen dienen te zijn. Zoeken met behulp van metadata komt aan bod in paragraaf 3.1.7.

Gegeven een aantal tekstdocumenten (zij het web pagina's, blog posts, Tweets, nieuwsartikelen, Marktplaats advertenties, enzovoort) is het mogelijk om een taak- of applicatiespecifieke zoekmachine te maken. Binnen de UvA hebben we aanzienlijke ervaring met het opleveren van dergelijke "zoekmachines op maat." Niet alleen voor tekst documenten, zoals hier beschreven, maar ook voor afbeeldingen en video's. Deze laatste komen in paragraaf 4.4 aan bod.

Voor op tekst gebaseerde zoekmachines gebruiken we in het algemeen Lucene en/of SOLR als basis.[1] Deze open-source zoekmachine biedt legio mogelijkheden om taak- of applicatiespecifieke zoekmachines te maken. Een voorbeeld van een dergelijke zoekmachine zou bijvoorbeeld gebruikt kunnen worden om vragen zoals "Doorzoek alle opgeslagen marktplaats.nl items" of "Welke YouTube gebruikers noemen de term 'jihad' het meest?" te beantwoorden. Ook is het mogelijk kwantitatieve en inhoudelijke vergelijkingen te doen tussen verschillende groepen documenten, zoals "Welk webforum bevat het vaakst de term 'jihad'?"

Zoals blijkt uit de bovenstaande voorbeelden kunnen zoekmachines op maat niet alleen gebruikt worden om specifieke analyses uit te voeren, ook worden ze ingezet als ruggengraat voor andere analyses, zoals hieronder beschreven.

Personen/Projecten: Edgar Meij

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / gemiddeld

4.1.2 Meta search/Data fusie

Het doel van data fusie is het samenbrengen van verschillende bronnen van informatie teneinde een rangschikking van documenten te verkrijgen. Hierbij moet bijvoorbeeld gedacht worden aan de manier waarop Google een rangschikking voor webpagina's genereert. Vroeger werd hiervoor uitsluitend gekeken naar de (relatieve) voorkomens van termen in de zoekvraag in de documenten. Tegenwoordig is het echter gebruikelijk om hier een groot aantal andere factoren bij te betrekken, zoals informatie uit de webgraaf, andere webpagina's, populariteitsinformatie, locatie-informatie en de zoekgeschiedenis van de gebruiker.

---

[1] Zie http://lucene.apache.org.

Een andere toepassing van data fusie is zogenaamde "meta search." Hierbij worden documenten verkregen uit verschillende zoekprocessen samengevoegd in een enkele ranking. Hierbij kan je bijvoorbeeld denken aan het combineren van resultaten van de Google en Yahoo! zoekmachines, maar ook aan het combineren van resultaten van verschillende domeinen. Stel bijvoorbeeld dat we een crawl hebben van alle YouTube video's en ook van alle Marktplaats items. Hoe voegen we de resultaten op een zoekvraag van deze twee samen? Gebruiken we daarvoor een relatief simpele maat (zoals leeftijd of datum) als beslissend argument? Of zijn er slimmere manieren te bedenken?

In zekere zin bouwt deze onderzoeksrichting dus voort op de "zoekmachines op maat" en dan met name op het geval waarin we ofwel meerdere van dergelijke zoekmachines hebben, dan wel waarin we additionele informatie willen betrekken binnen een enkele. Ook is het goed om op te merken dat data fusie inherent taak- en applicatiespecifiek is. Methoden en maten die in het ene geval goed werken kunnen zich heel anders gedragen in een ander geval.

Personen/Projecten: Marc Bron

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: gemiddeld / gemiddeld

4.1.3 Adaptief filteren

Eerder beschreven we al ad hoc zoeken. Binnen de onderzoeksrichting van het adaptief filteren staat de zoekvraag daarentegen vast en is het de bedoeling dat nieuwe documenten automatisch worden gerouteerd ("gefilterd") naar de gebruiker met de zoekvraag. Dit valt te vergelijken met een "RSS feed op maat," waarbij de gebruiker een zoekvraag definieert en op gezette tijden nieuwe resultaten ontvangt die relevant zijn voor de gedefinieerde zoekvraag. De zoekvraag wordt in dit geval veelal aangeduid als "zoek profiel;" een verzameling van sleutelwoorden die het interesse gebied van de gebruiker zo goed mogelijk weergeeft.

Een dergelijk systeem wordt "adaptief" zodra het mogelijk is dat de gebruiker feedback geeft op de teruggegeven resultaten en het systeem daarvan leert. Een simpel voorbeeld hiervan is een ja/nee knop waarmee de gebruiker aan kan geven of een specifiek document al dan niet relevant is. Het systeem leert vervolgens van deze feedback om in de toekomst meer relevante documenten terug te geven. Een bekend voorbeeld van een niet-adaptief filtering systeem is Google Alerts.[2]

Personen/Projecten: Manos Tsagkias, Peilend.nl

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / gemiddeld

4.1.4 Interactief zoeken

Zogenaamde "relevance feedback" maakt gebruik van oordelen van de gebruiker ten aanzien van de relevantie van een resultaat. Een dergelijk oordeel is typisch "wel/niet/enigszins relevant" en kan worden teruggegeven aan de zoekmachine teneinde resultaten in de toekomst te verbeteren (net zoals bij adaptief filteren).

[2] Zie http://www.google.com/alerts.

Een interessante afgeleide hiervan is het bepalen van "meer documenten zoals deze." Hierbij wordt gekeken naar de inhoud van het document en getracht alle andere documenten die daar het meest op lijken op te halen, bijvoorbeeld middels text mining methodes (zie paragraaf 3.2.1). Een variant hierop is een zogenaamde "woordenwolk" per resultaat, die weergeeft wat de meest kenmerkende termen in dat document zijn. Aangezien dat de meest kenmerkende termen zijn, krijgt de gebruiker een idee welke termen gebruikt dienen te worden om meer documenten zoals die terug te krijgen. Ook kan een woordenwolk worden gegenereerd voor de hoogst scorende documenten, aan de hand waarvan de gebruiker geholpen wordt bij het (her)formuleren van een zoekvraag. Iedere term kan dan vervolgens door de gebruiker worden geselecteerd en daarmee aan de query worden toegevoegd.

Deze technieken kunnen ook worden gebruikt om gebruikers een idee te geven van welke documenten andere gebruikers relevant vonden als antwoord op een zoekvraag.

Personen/Projecten: Edgar Meij

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / gemiddeld

4.1.5 Entiteiten zoeken

In sommige informatiebehoeften zijn gebruikers niet slechts op zoek naar documenten, maar naar een specifiek persoon, bedrijf, product, enzovoort. Hier is de gebruiker dus niet op zoek naar documenten, maar naar een specifieke *entiteit* die de zoekvraag beantwoord. Dergelijke informatiebehoeften kunnen niet (direct) beantwoord worden door documenten, aangezien de gebruiker op zoek is naar een of meerdere entiteiten. Bijvoorbeeld als je het antwoord zou willen weten op de vraag "Wie heeft Roland Garros in 2010 gewonnen?", "Hoe heet het meest recente besturingssysteem van Apple?" of "Welke Nederlandse schilders behoren tot de Barok?" Gevallen hiervan die relevant zijn voor het iRN zijn bijvoorbeeld bepaalde personen die actief zijn op een webforum en bepaald taalgebruik bezigen.

Het omgekeerde is ook interessant binnen de context van het iRN, namelijk het geval waarin de entiteiten bekend zijn en je wil weten welke documenten geassocieerd zijn met iedere entiteit. Dit wordt "entity profiling" genoemd en kan worden gebruikt om een tekstueel profiel te verkrijgen van bijvoorbeeld een person of interest (POI). In paragraaf 3.3.3 gaan we hier verder op in.

Om deze informatiebehoeften te kunnen beantwoorden hebben we verscheidene mogelijkheden om entiteiten te herkennen in tekst, die in meer of mindere mate lerend zijn, dat wil zeggen dat feedback van de gebruikers ingezet kan worden om de algoritmes te verbeteren. Met dergelijke technieken is het ook mogelijk om entiteiten te identificeren die voorkomen in een verzameling documenten, zoals diegene die teruggegeven worden aan de hand van een van een zoekvraag.

Personen/Projecten: Krisztian Balog, Marc Bron / Fietstas, EARS, Penta Politica

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / gemiddeld

4.1.6 Entiteiten en hun samenhangen

Dit onderzoeksgebied hangt sterk samen met de vorige. In dit geval kijken we ook naar teksten en "koppelen" we de gevonden entiteiten aan elkaar door te bepalen welke significant veel met elkaar voorkomen. Hierdoor kunnen we bijvoorbeeld bepalen welke

personen een significante associatie met elkaar hebben, of welke plaats vaak genoemd wordt rondom een POI.

Personen/Projecten: Krisztian Balog, Marc Bron /SaHaRa

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / gemiddeld

4.1.7 Metadata

In veel applicaties hebben we tegenwoordig niet alleen tekstuele inhoud voorhanden, maar ook contextuele informatie. Denk aan tijdstippen (waarop bijvoorbeeld een blogpost is gecreëerd), auteursinformatie (door wie bijvoorbeeld de blogpost is geschreven) of aan andere informatie (zoals bijvoorbeeld tags of categorieën). Ook kunnen, op automatische dan wel handmatige wijze, bepaalde stukken van een tekst zijn geannoteerd, bijvoorbeeld om personen, producten of organisaties aan te duiden. Dergelijke, vaak niet-tekstuele informatie wordt in het algemeen aangeduid als *metadata* of *facetten*.

We kunnen metadata op een aantal manieren inzetten, waarvan we er hier twee beschrijven. Allereerst kunnen we de metadata gebruiken in het zoekproces en dan met name als toevoeging aan (of een soort filter op) de zoekvraag. We kunnen bijvoorbeeld de resultaten van een zoekvraag inperken door in te geven dat we alleen geïnteresseerd zijn in blogposts die na een bepaalde datum verschenen zijn. Een alternatief is om uitsluitend metadata te gebruiken. In dat geval zijn we bijvoorbeeld op zoek naar alle blogposts uit een bepaalde rubriek en/of met een bepaalde tag of onderwerp.

Een andere mogelijkheid is om de metadata in te zetten aan de kant van de resultaatpresentatie. Een populaire vorm hiervan is om bijvoorbeeld, gegeven een zoekvraag, de “hits” per metadata veld of waarde (zoals een categorie) weer te geven. De gebruiker kan dan op een categorie klikken om aan te geven dat hij slechts geïnteresseerd in resultaten uit die categorie. Ook kunnen we de metadata gebruiken om inzage te krijgen in hoe woorden worden gebruikt in combinatie met bepaalde metadata of gedurende een tijdsinterval.

Binnen de UvA hebben we ook ervaring met het bouwen van (zelf-)lerende systemen, die *machine learning* toepassen teneinde verschillende zaken kunnen voorspellen. Hierbij moet bijvoorbeeld gedacht worden aan het voorspellen van de populariteit van een nieuwsitem. Gegeven historische informatie (in de vorm van nieuwsitems en bijbehorende commentaren) kan een dergelijk systeem voor een nieuw nieuwsitem met redelijke nauwkeurigheid voorspellen wat de populariteit van dit item gaat worden.

Personen/Projecten: Manos Tsagkias, Wouter Weerkamp, Edgar Meij / Fietstas, Peilend.nl

Benodigd implementatie werk: zeer weinig

Benodigd onderzoek: weinig / gemiddeld

4.1.8 Longitudinaal zoeken

De tijdsdimensie is een bijzondere vorm van metadata, die in veel gevallen inherent aanwezig is (zoals via een publicatie datum of een Last-Modified HTTP header). Indien er dergelijke temporele informatie voorhanden is, is het mogelijk een analyse uit te voeren naar welke termen (of entiteiten) populair zijn op een gegeven moment. Ook is het in dat geval mogelijk om te kijken naar de afgeleide daarvan, namelijk door te kijken naar welke termen of entiteiten populair worden dan wel verdwijnen. Een bekend

voorbeeld van een toepassing die gebruik maakt van temporele informatie is Google Trends.[3] Deze tool laat zien welke termen populair zijn in een bepaalde tijdspanne en kan inzicht verschaffen in (relatief) woordgebruik. In het kader van het iRN is een interessante toepassing van longitudinaal zoeken het verkrijgen van een overzicht van relevante documenten en/of entiteiten over een lange tijdspanne.

Personen/Projecten: Manos Tsagkias, Wouter Weerkamp

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / aanzienlijk

4.2 Taaltechnologie

Een aanzienlijk deel van iedere zoekmachine wordt bepaald door taaltechnologie. Taaltechnologie is een verzamelnaam voor verscheidene onderzoeksgebieden en toepassingen met als gemene deler de analyse van één van de meest complexe informatiemedia: de menselijke taal. Taaltechnologie brengt twee wetenschapsgebieden bij elkaar, namelijk taalkunde en informatie- en communicatietechnologie. Omdat natuurlijke taal in zowel gesproken als geschreven vorm voorkomt, kent taaltechnologie twee gedaantes: spraak- en teksttechnologie. Waar spraaktechnologie gebruikt wordt om menselijke spraak te produceren en/of te begrijpen, wordt tekstuele taaltechnologie met name ingezet voor de analyse van het geschreven woord. Denk aan tekstverwerkers met ingebouwde spelling- en grammatica controle, vertaalmachines die via een ruwe vertaling ook anderstalige webpagina's toegankelijk maken, term suggesties, zoekmachines, enzovoort. In deze paragraaf beschrijven we enkele tekstuele taaltechnologische onderzoeksrichtingen, waaronder classificatie, clustering en sentiment analyse. Hierbij dient opgemerkt te worden dat, zoals hieronder beschreven, taaltechnologie op zich al interessante tools en technieken met zich meebrengt. Deze tools en technieken kunnen echter ook gebruikt worden om stukken tekst te "verrijken" en de output daarvan kan vervolgens ingezet in bijvoorbeeld een zoekmachine.

4.2.1 Text mining

Text mining is een verzamelnaam voor methoden en technieken die als doel hebben kennis te distilleren uit vrije tekst. Voorbeelden van zulke technieken zijn het automatisch samenvatten van tekst (zie paragraaf 3.2.4), het categoriseren van teksten ("tekst classificatie") en het bepalen van op elkaar gelijkende teksten ("tekst clustering"). Ook het automatisch identificeren van entiteiten en het bepalen van relaties daartussen (zoals genoemd in paragraaf 3.1.5 en paragraaf 3.1.6) is een vorm van text mining. Er bestaan verscheidene open-source tools (zoals GATE, WEKA en Lingpipe) waar we meer geavanceerde tools en analysemethoden op baseren.[4]

Bij *tekst classificatie* is het de bedoeling dat nieuwe stukken tekst (denk aan webpagina's, blog posts, tweets, enzovoort) automatisch worden "gelabeld" met een categorie. De verzameling categorieën staat in dit geval dus vast en het is aan het systeem om de toekenning uit te voeren. Het is goed om hier op te merken dat het ophalen van relevante documenten voor een zoekvraag (ad hoc zoeken) opgevat kan worden als een (binarie) classificatie taak. Het systeem bepaalt namelijk voor ieder document of deze wel of niet relevant is gegeven de zoekvraag. Het is dan ook niet verwonderlijk dat vele technieken die toegepast worden bij tekst classificatie elementen gemeenschappelijk hebben met die uit de zoektechnologie. Een voor de hand liggende toepassing van tekst

[3] Zie http://www.google.com/trends.
[4] Zie http://gate.sourceforge.net , http://weka.sourceforge.net en http://alias-i.com/lingpipe.

classificatie binnen het iRN is het bepalen of en in hoeverre documenten gaan over radicalisatie. Een extrapolatie hiervan is het uitvoeren van deze analyse op een verzameling documenten die geassocieerd zijn met een bepaalde persoon, zoals de forum berichten op een webforum. Tekst classificatie is tevens flexibel genoeg om ook andere zogenaamde "features" mee te kunnen nemen, zoals bijvoorbeeld de snelheid van reageren van een bepaald persoon.

In het geval van *tekst clustering* zijn er geen a priori gedefinieerde categorieën, maar is het de bedoeling dat stukken tekst op een zinvolle manier worden gegroepeerd. Op deze manier kunnen bijvoorbeeld documenten worden samengenomen die hetzelfde onderwerp bespreken. Een specifieke toepassing van tekst clustering die relevant is binnen het iRN, is het clusteren van web pagina's (die bijvoorbeeld relevant zijn voor een zoekvraag of die van een bepaald domein komen). Hiermee kan vervolgens op een zinvolle manier bepaald worden welke subset van de documenten bekeken dienen te worden. Ook is het mogelijk ieder cluster automatisch samen te vatten (zie paragraaf 3.2.4) om zo een goed beeld te krijgen van de verschillende "aspecten" van een grote set documenten. Als laatste voorbeeld noemen we een toepassing waarin clusters worden gegenereerd voor een groep webpagina's (bijvoorbeeld alle forum berichten op een webforum) op gezette tijden. Door de clusters op verschillende tijdstippen weer te geven is het mogelijk snel een inhoudelijk beeld te krijgen van de dynamiek van deze pagina's.

Personen/Projecten: Wouter Weerkamp, Jiyin He

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / weinig

4.2.2 Sentiment/Polariteit analyse

Sentiment analyse houdt zich bezig met de classificatie van stukken tekst, waarbij de categorieën gevormd worden door *sentiment*, variërend van polariteit ("positief," "neutraal" of "negatief") tot gevoelens ("blij," "angstig,", "vrolijk," "verdrietig," enzovoort). Een veelgebruikt alternatief om sentiment te bepalen naast tekst classificatie is om een aantal sleutelwoorden aan te wijzen die een bepaald sentiment aanduiden. Als deze woorden voorkomen in een stuk tekst geeft dat een indicatie dat de tekst dat bepaalde sentiment weergeeft.

Een interessante extrapolatie van dit principe kan gebruikt worden om ervaringen rondom een thema of activiteit te voorzien van een sentimentsclassificatie. Zodra we een aantal documenten hebben die dit thema of deze activiteit beschrijven kunnen we sentiment analyse gebruiken om de geassocieerde sentimenten boven water te krijgen. Eenzelfde vorm van analyse kan worden gebruikt om bijvoorbeeld het sentiment rondom een product ("de nieuwe Apple iPod"), een persoon ("Barack Obama"), of ieder ander onderwerp te bepalen. Binnen het kader van het iRN is een dergelijke analyse nuttig om op bijvoorbeeld een grote, geaggregeerde schaal inzicht te krijgen in het sentiment rondom een bepaald onderwerp (zoekvraag) of rondom een bepaalde entiteit zoals een persoon of een plaats.

Personen/Projecten: Wouter Weerkamp, Valentin Jijkoun / Fietstas

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / aanzienlijk

4.2.3 Duplicaat detectie en informatiehergebruik

Een andere manier om te kijken naar taalvergelijkingen, is door te kijken naar de individuele karakters en termen waaruit een tekst bestaat; dergelijke *n-grammen* kunnen ook worden gebruikt om de similariteit tussen twee teksten te bepalen. Een simpele maat om stukken tekst met elkaar te vergelijken is door te tellen hoeveel van de n-grammen overeenkomen en hoeveel er verschillen. De uitkomst van zulke vergelijkingen kunnen vervolgens ingezet worden om te bepalen hoeveel overlap er bestaat tussen twee stukken tekst; een hoge mate van overlap betekent dat de twee nagenoeg identiek zijn.

Deze aanpak wordt bijvoorbeeld ingezet om plagiaat op te sporen of om te bepalen wie de auteur is van een bepaald stuk tekst. Een ander voorbeeld is het bepalen van informatiehergebruik, zoals deze voorkomt bij aanbieders van nieuws. Hier geeft een persbureau nieuwsberichten uit die door nieuwsaanbieders zoals kranten worden opgepakt en aangevuld met additionele informatie; deze maat kan gebruikt worden om te bepalen wat er is toegevoegd, verwijderd of veranderd. In het kader van het iRN kan deze techniek gebruikt worden om de mate van overlap tussen bijvoorbeeld web pagina's te bepalen en inzage te geven in welke gedeelten anders zijn, dan wel de hoeveelheid te bestuderen tekst te verminderen door gekopieerde teksten te herkennen en te verwijderen.

Personen/Projecten: Edgar Meij, Manos Tsagkias, Valentin Jijkoun / Fietstas

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / weinig

4.2.4 Automatisch samenvatten

In paragraaf 3.1.4 noemden we al de mogelijkheid van het genereren van zogenaamde woordenwolken per document of per verzameling van documenten; woordenwolken kunnen dienen als analysemiddel om op een intuïtieve manier de inhoud van een of meerdere documenten weer te geven. Ze geven typisch de meest frequente termen weer, waarbij de grootte van een term proportioneel is aan de frequentie waarmee deze voorkomt. Er bestaan nog andere manieren om op een automatische manier teksten samen te vatten. Deze methoden worden doorgaans onderverdeeld in twee kampen: (i) *extractief* (waarbij sleutelzinnen worden geïdentificeerd) en (ii) *abstractief* (waarbij de belangrijkste elementen worden geparafraseerd in een nieuw stuk tekst). Momenteel is automatisch samenvatten op basis van *extractie* het meest succesvol, aangezien *abstractie* nog steeds een behoorlijke uitdaging vormt.

Een van de meest gangbare methoden om sleutelzinnen uit een (of meerdere) document(en) te extraheren, is door een zoekmodel toe te passen op iedere zin binnen het document; het idee hierachter is om de meest kenmerkende, "centrale" zin of zinnen aan te wijzen. Wat er derhalve wordt gedaan is iedere zin uit het document te beschouwen als een "mini-document" en te vergelijken met iedere andere zin in het document. Op deze wijze kan bepaald worden welke zin het meeste lijkt op alle andere zinnen in het document; deze wordt dan teruggegeven als de zin die de inhoud van het gehele document het beste weergeeft. Binnen het kader van het iRN kan automatisch samenvatten worden ingezet om op een snelle manier stukken tekst inzichtelijk te maken.

Personen/Projecten: Christof Monz / Fietstas

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / gemiddeld

4.2.5 Web 3.0

Tot slot noemen we een aantal zogenaamde Web 3.0 methoden en technieken. Er bestaat nog enige onduidelijkheid omtrent wat deze term precies inhoudt, maar men is het er in het algemeen over eens dat een rijke semantische annotatie structuur deel uitmaakt van het Web 3.0. Hierbij moet gedacht worden aan de rijke, "intelligente" snippets die de Yahoo! en Google zoekmachines in sommige gevallen aanbieden (bijvoorbeeld bij concerten of bioscopen). Ook moet hierbij gedacht worden aan het semantische web en/of de Linked Open Data cloud. Het vakgebied van het semantische web biedt standaarden en technieken aan om op een uniforme manier *concepten* en *relaties* daartussen te definiëren, waarmee het mogelijk moet zijn voor computers om te redeneren. Indien de entiteiten vastgelegd zijn in een dergelijke kennistructuur kan ook de automatische annotaties van stukken tekst met entiteiten (zoals eerder besproken in paragraaf 3.1.5) worden gezien als een vorm van semantische verrijking. In dat geval is het dan ook mogelijk te redeneren met de geëxtraheerde entiteiten, door bijvoorbeeld gebruik te maken van de kennis dat Amsterdam een stad in Nederland is. In het kader van het iRN kunnen dergelijke technieken bijvoorbeeld ingezet worden om op een intelligentere manier documenten terug te vinden.

Naast het herkennen van entiteiten hebben we recentelijk ook een algoritme bedacht waarmee je arbitraire stukken tekst op automatische wijze semantisch kan verrijken door deze te linken aan Wikipedia artikelen. Met deze verrijking kan je bijvoorbeeld gemakkelijk achtergrond informatie opzoeken of kunnen teksten op een semantische manier geclusterd worden.

Personen/Projecten: Edgar Meij

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / gemiddeld

4.2.6 Focused crawling

Focused crawling is een manier van crawlen die rekening houdt met een specifieke informatiebehoefte (eventueel in combinatie met specifieke metadata). Idealiter volgt een dergelijke crawler alleen de links die relevant zijn voor de informatiebehoefte, zoals bijvoorbeeld een POI. Hiertoe wordt begonnen op een aantal zogenaamde *seed* web pagina's en alleen die links gevolgd die voldoende samenhang vertonen met de informatiebehoefte. Hiertoe kan bijvoorbeeld de anchortext gebruikt worden, maar het is ook gebruikelijk om iedere link te volgen (zoals in een normale web crawl) en pagina's die onvoldoende met de informatiebehoefte te maken hebben te schrappen. Een welbekend voorbeeld van focused crawling is de crawlingarchitectuur achter websites als Citeseer.

Personen/Projecten: Ssscrape

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / veel

4.2.7 Taaldetectie

Steeds meer data komt online beschikbaar in verschillende talen die vaak alleen maar met behulp van tolken geïnterpreteerd kunnen worden. Het automatisch herkennen van de specifieke talen is hiervoor een belangrijke eerste stap. Er zijn over de jaren al verschillende taaldetectoren ontwikkeld, maar de meeste van hen zijn met name gericht op Europese talen en talen die door grote groepen van mensen worden gesproken, zoals Arabisch en Chinees. Voor het ontwikkelen van een generieke taaldetector is het nodig om naar grote corpora van al geannoteerde taalbronnen te kijken en hieruit de nodige

statistieken te halen. Daarnaast is het detecteren van documenten met verschillende talen, b.v. Nederlandse documenten met Arabische referenties een open probleem. Ook robuustheid van taaldetectoren om met verschillende coderingen te kunnen werken is een belangrijk punt die nauwelijks aandacht krijgt in de bestaande wetenschappelijke literatuur.

Personen/Projecten: Christof Monz

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / weinig

4.2.8 Machinaal vertalen

Machinaal vertaal systemen vertalen op een automatische wijze documenten in een vreemde taal, zoals Arabisch, naar een bekende taal, zoals Engels. Er is veel werk verricht op dit gebied, waarvan een groot deel is gefinancierd door overheden, met name inlichtingen- en politiediensten. Documenten in vreemde talen zijn vaak belangrijke informatiebronnen in grootschalige onderzoeken met een internationale component, zoals mensen- en wapensmokkel, fraude, en terrorisme. Vaak is de hoeveelheid aan documenten te groot om door tolken te laten vertalen en zijn automatische methoden belangrijk om de werkelijk belangrijke document te identificeren en wellicht door en tolk te laten vertalen.

In het verleden waren machinale vertaalsystemen vaak regelgebaseerd en dus erg taal afhankelijk. Over de laatste tien jaar hebben statistische vertaalsystemen enorme vooruitgang geboekt. Statistische systemen leren automatisch vertaalregels van bestaande vertalingen. Binnen de UvA beschikken wij over een eigen statistisch vertaalsysteem voor verschillende talen, waaronder Arabisch, Perzisch, Pashto, Bulgaars, Duits, Albaans en meer. Alhoewel deze systemen redelijk tot goed werken voor nieuwsbronnen, varieert de vertaalkwaliteit substantieel voor user-generated content zoals web pagina's, blog posts, web fora en e-mail. De uitdaging is om robuustere vertaalsystemen te ontwikkelen die ook binnen deze domeinen toepasbaar zijn.

Personen/Projecten: Christof Monz

Benodigd implementatie werk: aanzienlijk

Benodigd onderzoek: aanzienlijk / aanzienlijk

4.3 Sociale netwerk analyse

Sociale netwerken ontstaan automatisch zodra mensen met elkaar communiceren. Ze kunnen variëren van simpele communicatienetwerken tot netwerken van gelijkgestemden dan wel familie- en vriendennetwerken. Met de komst van het Internet heeft de ontwikkeling van digitale sociale netwerken een enorme vlucht genomen. Niet alleen door vriendensites zoals Hyves en Facebook, maar ook door webfora en (micro)blogging platforms zoals Twitter, waarbij het mogelijk is andere mensen terug te vinden en te volgen. Het doel van sociale netwerkanalyse in de context van het iRN behelst het ontsluiten van sociale processen en daarmee geassocieerde tekstuele en andersoortige uitingen. Hierbij kan bijvoorbeeld gedacht worden aan het bepalen van de meest invloedrijke persoon op Twitter gegeven een zoekvraag, maar ook aan identity tracking & profiling, waarin getracht wordt een zo volledig mogelijk profiel van een POI en zijn/haar contactpersonen te vergaren uit disjuncte bronnen op het Internet, zoals bijvoorbeeld uit web fora en YouTube. Dergelijke informatie kunnen we bijvoorbeeld inzetten om te bepalen of een YouTube gebruiker behoort tot een groep radicaliserende personen of om een forum analyse te ondersteunen, waarbij bijvoorbeeld gekeken

wordt of een bericht, persoon of groep radicaal is/wordt. Idealiter worden gebruikers gevolgd over meerdere media, waarbij gebruikersnamen bijvoorbeeld op (semi-)automatische wijze herleid worden tot een en dezelfde persoon.

4.3.1 Identity tracking

Het doel van identity tracking is het "normaliseren" van personen (d.w.z. "aliassen" in de vorm van gebruikersnamen of daadwerkelijke personen) op verschillende sociale netwerk en andere websites, zodat op (semi-)automatische wijze bepaald wordt dat twee gebruikersnamen te herleiden zijn tot dezelfde persoon. Vervolgens kunnen we al het gevonden materiaal (bijvoorbeeld YouTube video's, YouTube comments, blogposts, Tweets, etc.) bijvoorbeeld samenvoegen tot een profiel van die persoon (zie paragraaf 3.3.2). Ook kunnen we een sociale netwerkanalyse uitvoeren, waarmee het mogelijk is het sociale netwerk van een persoon te analyseren (zie paragraaf 3.3.3).

Personen/Projecten: Wouter Weerkamp / Fietstas, EARS

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: gemiddeld / aanzienlijk

4.3.2 Persoonsanalyse

Persoonsanalyse (of identity profiling) hangt sterk samen met entiteit profiling (zie paragraaf 3.1.5). In dit geval worden de associaties tussen entiteiten en stukken tekst niet uitgevoerd middels informatie extractie methoden (zie paragraaf 3.2.5), maar worden deze verkregen middels de online identiteit van een persoon. In het kader van het iRN kunnen twee gevallen worden onderscheiden: (i) *interne* gebruikers, dat wil zeggen de mensen die gebruik maken van het iRN en (ii) *externe* gebruikers, dat wil zeggen POI's of andere personen op het Internet.

In het interne geval kan gedacht worden aan een social community website zoals Buddypress of Drupal, waarop gebruikers van het iRN zich registreren en onder andere bijhouden wat hun expertise is, wat voor soort onderzoek ze doen en wat hun bevindingen zijn. Ook kunnen bijvoorbeeld competenties, aandachtsgebieden, kennis en expertise worden bijgehouden. Andere gebruikers kunnen deze informatie terugvinden, teneinde bijvoorbeeld hun eigen analyses te ondersteunen of om "experts" te vinden op een bepaald gebied. Zodra dit soort informatie beschikbaar is, kunnen dezelfde soort analyse methoden als die beschreven in paragraaf 3.1.5 worden ingezet om relevante "entiteiten" (in dit geval iRN gebruikers) terug te vinden. Ook maakt een dergelijke opzet het mogelijk op een laagdrempelige manier vondsten te delen met andere iRN gebruikers.

In het externe geval kunnen de methoden en technieken die beschreven zijn in paragraaf 3.1.5 worden ingezet om profielen te genereren van bijvoorbeeld POI's. Net zoals in paragraaf 3.1.4 beschreven, kunnen we deze analyse ook omdraaien en kijken naar welke termen gebruikt kunnen worden om een bepaald persoon terug te vinden. Hiermee kunnen vragen beantwoord worden als "Wat zijn de termen die het sterkst samenhangen met een bepaald persoon?" Maar ook "Wie heeft het over het plegen van een aanslag? Heeft diegene video's op YouTube geplaatst? Welke plaatsten noemt diegene? En wie zijn zijn vrienden?"

Personen/Projecten: Krisztian Balog, Marc Bron / EARS

Benodigd implementatie werk: weinig

Benodigd onderzoek: weinig / aanzienlijk

4.3.3 Graaf/Relatie analyse

Zoals de naam al impliceert, omvatten sociale netwerken niet alleen personen (de knopen in de graaf) maar ook relaties tussen die personen. Dergelijke relaties kunnen worden bepaald door bijvoorbeeld het zogenaamde volgen van een ander persoon op Twitter, door gezamenlijk een commentaar achter te laten op een YouTube video dan wel op een web forum of door iemand te bevrienden op Facebook of Hyves. Dergelijke netwerkstructuren (of *grafen*) kunnen op allerhande manieren worden geanalyseerd alsmede worden ingezet om andere analyses te ondersteunen. Gegeven bijvoorbeeld een lijst met POI's op Twitter kan bepaald worden wie de meest "centrale" persoon is, maar ook welke termen er worden gebruikt binnen die verzameling gebruikers. Dergelijke, afgeleide statistieken (waaronder ook bijvoorbeeld het gedrag van gebruikers op webfora) kunnen ingezet worden in een groter systeem, dat bijvoorbeeld sociaal gedrag koppelt aan bepaald crimineel gedrag. Ook kunnen sociale netwerken gedurende een langere tijd geobserveerd worden, in dat geval kan de dynamiek ervan interessante patronen blootleggen.

Personen/Projecten: Valentin Jijkoun, Wouter Weerkamp

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / aanzienlijk

4.4 Beeld en video analyse

In hun Roadmap beeldtechnologie veiligheidsdomein ("Behoeften en gewenste innovaties voor 'Veilig door innovatie'") identificeren Flight en Hulshof een tweetal technische behoeften als zeer prominent: "Objecten, subjecten volgen + vinden en reconstrueren achteraf" en "Relevante beelden filteren". De UvA heeft werk verricht aan concept-, scene- en persoonsherkenning dat onder de eerste noemer valt en het heeft werk verricht aan multimedia linken en informatie visualisatie dat onder de tweede noemer valt.

4.4.1 Concept-, scene- en subjectherkenning

Concept herkenning houdt zich bezig met het herkennen van visuele concepten in beeldmateriaal. Het woord "concept" wordt doorgaans breed geïnterpreteerd; zo zijn "boot" of "buiten" of "Barack Obama" allemaal voorbeelden van concepten die de moderne, automatische conceptherkenners van de UvA met succes aankunnen. De oplossingen hiervoor aan de UvA zijn gebaseerd op machine leren. Daardoor werken deze op willekeurige domeinen, waarbij de performance mede afhankelijk is van de hoeveelheid trainingsmateriaal die voorhanden is om de algoritmes te informeren.

Personen/Projecten: Arnold Smeulders, Cees Snoek, Marcel Worring / Impala

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / aanzienlijk

4.4.2 Objectherkenning

Voor sommige OOV-taken kan het van belang zijn om een specifiek object in een beeld te herkennen, te denken valt aan een wapen of een kledingstuk. Computergestuurde objectherkenning is uitdagend omdat, anders dan mensen, machines moeite hebben met verschillen in schaal of afstand en met het gedraaid of gedeeltelijk verstopt zijn van een te herkennen object. Veel van de succesvolle methodes voor computergestuurde objectherkenning slagen erin om essentiële visuele kenmerken van objecten te identificeren en op basis van die sleutelkenmerken met aanzienlijke zekerheid te zeggen of een er match is. De technieken die bij de UvA worden toegepast zijn gebaseerd op

machine leren en hun succes is derhalve afhankelijk van de hoeveelheid beschikbaar trainingsmateriaal.

Personen/Projecten: Theo Gevers, Arnold Smeulders / Impala

Benodigd implementatie werk: aanzienlijk

Benodigd onderzoek: aanzienlijk / aanzienlijk

4.4.3 Multimedia linken

Het doel van deze onderzoekslijn is het identificeren van links tussen multimedia items, bijvoorbeeld om verdachten of slachtoffers op te sporen. Hiertoe dienen verschillende multimedia bronnen aan elkaar maar ook aan andere modaliteiten (zoals tekst) gekoppeld te worden. Tevens moet de belangrijkheid en relevantie van mogelijke links automatisch bepaald kunnen worden, aan de hand van de beschikbare (meta)data en eventuele eerdere casussen. De mogelijke links bestaan uit low-level links tussen video fragmenten (als zij bijvoorbeeld dezelfde objecten bevatten), tussen multimedia objecten (bijvoorbeeld op grond van gedeelde metadata) of tussen geïdentificeerde personen.

Personen/Projecten: Marcel Worring, Bouke Huurnink, Marc Bron

Benodigd implementatie werk: aanzienlijk

Benodigd onderzoek: aanzienlijk / gemiddeld

4.4.4 Informatie visualisatie

Het doel van deze onderzoekslijn is het ontsluiten van multimedia collecties. Veel multimedia collecties bevatten alleen metadata zoals “date created” en “file size” en ontberen doorgaans verdere annotaties. Hierdoor is het navigeren door dergelijke collecties op zijn best tijdrovend. Automatische inhoudsanalyse kan metadata genereren in de vorm van zogenaamde *content-based descriptors*. Op dit moment is de nauwkeurigheid van dergelijke technieken op zichzelf onvoldoende om het categoriseren van een collectie te kunnen automatiseren. Visualisatietechnieken ondersteunen gebruikers echter bij het categoriseren van een multimedia collectie, met effectieve tools voor het sorteren, filteren, selecteren en visualiseren van collecties.

Personen/Projecten: Marcel Worring, Ork de Rooij / MediaTable

Benodigd implementatie werk: gemiddeld

Benodigd onderzoek: weinig / aanzienlijk

5 Toepassingen

In het vorige hoofdstuk hebben voor verscheidene onderzoekslijnen aangegeven in welke context of use-case (verkregen uit de interviews) deze interessant is. In dit hoofdstuk gaan we hier dieper op in, door een aantal concrete toepassingen te definiëren (wederom verkregen uit de interviews) die aan te pakken zijn door verscheidene onderzoekslijnen te stapelen en/of combineren. Eerst beschrijven we echter een typering van de methoden en technieken die in de context van het iRN gebruikt kan worden. We onderscheiden twee "typen" van analyse, namelijk (i) reactief en (ii) proactief/samenvattend.

Een reactieve vorm van analyse heeft, zoals de naam al impliceert, een bepaalde input van een gebruiker nodig. De gebruiker heeft bijvoorbeeld een zoekvraag en resultaten

worden vervolgens gecrawled, opgeslagen en weergegeven. In het geval van adaptief filteren kan die ook een doorlopende zoekvraag zijn, waarbij de gebruiker om de zoveel tijd op de hoogte wordt gehouden van nieuwe resultaten. Dit is de meest voor de hand liggende categorie. Andere analyses die hierbij aansluiten zijn, onder meer, sentiment analyse t.a.v. een zoekvraag, het bepalen van een verschil in taalgebruik in de documenten teruggegeven aan de hand van twee zoekvragen en focused crawling, waarbij gegeven een zoekvraag/webpagina alle daaruit en daarnaartoe linkende pagina's gecrawled en samengevat worden.

In het proactieve (of samenvattende) geval, detecteert het systeem "interessante" gevallen, gaat op zoek naar achterliggende, gerelateerde informatie en geeft hier een samenvatting van. Analyses die hierbij aansluiten zijn, onder meer, de volgende.

- Houd bij welke nieuwsberichten meer dan bijzondere aandacht trekken (gemeten aan de hand van bijvoorbeeld het aantal commentaren of views). Haal voor deze berichten gerelateerde items op, zoals hyperlinks in het bericht, hyperlinks in de commentaren, hyperlinks in de profielen van de mensen die reageren, pagina's van (of Tweets over) personen of bedrijven die genoemd worden in het bericht of pagina's die lijken op het nieuwsbericht.
- Kijk welke termen "opkomend" zijn in een bron (zoals Twitter of het nieuws) en haal "omliggende" informatie op. Hetzelfde principe kan toegepast worden op geolocaties die met Tweets geassocieerd zijn.
- Volg bestaande/nieuwe gebruikers op een forum en rapporteer "interessant" gedrag. Bijvoorbeeld over hoeveel nieuwe gebruikers er op een forum zijn bijgekomen en of er daarin een verandering heeft plaatsgevonden. Een ander voorbeeld is het weergeven van gebruikers met "afwijkend" gedrag (zoals bijvoorbeeld radicaliserende personen).

De onderstaande toepassingen zijn concrete voorbeelden die voortgekomen zijn uit de interviews. Ieder van deze dient ter illustratie en kan worden aangepakt middels (combinaties van) onderzoeksrichtingen zoals deze hierboven zijn beschreven.

5.1 Person of interest profiling and retrieval

Als input voor deze toepassing dient een lijst van bekende personen, zoals een lijst met POI's. De verwachte output is een profiel van iedere persoon, waarin beschreven staat wat er van diegene op het internet te vinden is (al dan niet gesplitst per bron en samengevat). Ook dient voor iedere persoon aangegeven te worden met welke andere personen hij/zij in contact staat en waar de berichten uit bestaan (wederom al dan niet gesplitst per bron en samengevat). Een variant hierop vat ook de interactiegeschiedenis samen en kan dienen als input voor een toepassing die bepaalt of een bepaalde persoon crimineel gedrag vertoont.

Teneinde deze toepassing te kunnen realiseren hebben we de volgende "bouwstenen" nodig:

- Meta search (3.1.2) – wordt gebruikt om meerdere internet bronnen te ontsluiten en samen te voegen.
- Entiteiten zoeken (3.1.5) / Persoonsanalyse (3.3.2) – wordt gebruikt om gerelateerde entiteiten te identificeren en beschrijven.
- Entiteiten en hun samenhangen (3.1.6) – wordt gebruikt om de relaties met de gevonden entiteiten te omschrijven.
- Automatisch samenvatten (3.2.4) – wordt (optioneel) gebruikt om de gevonden stukken tekst automatisch in te korten, om zo de eindgebruiker van het iRN te ondersteunen.

- Identity tracking (3.3.1) – wordt gebruikt om verschillende aliassen te disambigueren.
- Graaf/Relatie analyse (3.3.3) – wordt gebruikt de relaties met andere personen op sociale media te identificeren en beschrijven.
- Taaldetectie (3.2.7) en machinaal vertalen (3.2.8) – wordt (optioneel) gebruikt om stukken tekst in een vreemde taal te vertalen.
- Concept-, scene- en subjectherkenning (3.4.1) – wordt (optioneel) gebruikt om een persoon te herkennen in bijvoorbeeld YouTube video's.

5.2 Temporele analyse

Dit kan of een reactieve (waarin de gebruiker een zoekvraag ingeeft) of een proactieve toepassing zijn. Het doel is een profiel over de tijd weergeven van belangrijke gebeurtenissen in het nieuws. Hiertoe worden verschillende bronnen gecombineerd, waaronder Twitter en nieuws websites. Teneinde deze toepassing te kunnen realiseren hebben we de volgende "bouwstenen" nodig:

- Meta search (3.1.2) – wordt gebruikt om meerdere internet bronnen (in dit geval nieuws websites en Twitter) te ontsluiten en samen te voegen.
- Entiteiten zoeken (3.1.5) – wordt gebruikt om plaatsen te identificeren.
- Metadata (3.1.7) – wordt gebruikt om de populariteit van een nieuwsitem te voorspellen.
- Sentiment analyse (3.2.2) – wordt (optioneel) gebruikt om de belangrijkste stemmingen rondom iedere gebeurtenis aan te geven.
- Automatisch samenvatten (3.2.4) – wordt (optioneel) gebruikt om de gevonden stukken tekst automatisch in te korten.

5.3 Adaptief filteren van nieuws

Het doel van deze toepassing is de automatische analyse en aggregatie van lokaal, regionaal, nationaal en (eventueel) internationaal nieuws. Deze toepassing kan dienen als (i) "thermometer" in de samenleving en (ii) beschouwd worden als complementair aan bijvoorbeeld (politie) rapporten. Dit is typisch een reactieve toepassing, waarin de gebruiker een vaste zoekvraag heeft. Zodra er relevant nieuws binnenkomt wordt dit naar de gebruiker met de informatiebehoefte gestuurd. Teneinde deze toepassing te kunnen realiseren hebben we de volgende "bouwstenen" nodig:

- Meta search (3.1.2) – wordt gebruikt om meerdere internet bronnen te ontsluiten en samen te voegen.
- Adaptief filteren (3.1.3) – wordt gebruikt om relevante informatie te identificeren.
- Interactief zoeken (3.1.4) – wordt (optioneel) gebruikt om de gebruiker te helpen een optimaal zoekprofiel te definiëren.
- Text mining (3.2.1) – wordt (optioneel) gebruikt om groepen relevante nieuwsitems te identificeren (clusteren), dan wel om de meest relevante politie rapporten te linken (classificeren).
- Duplicaat detectie (3.2.3) – wordt gebruikt om aan te geven of en hoe verschillende relevante nieuwsitems van elkaar verschillen.
- Automatisch samenvatten (3.2.4) – wordt (optioneel) gebruikt om de gevonden stukken tekst automatisch in te korten.
- Taaldetectie (3.2.7) en machinaal vertalen (3.2.8) – wordt (optioneel) gebruikt om stukken tekst in een vreemde taal te vertalen.

5.4 Mogelijke doelwitten

Het doel van deze proactieve toepassing is om mogelijke doelwitten bepalen uit bijvoorbeeld web fora, Twitter, of andere sociale media. De output van de toepassing is een lijst met mogelijke doelwitten, gebaseerd op een tekstuele analyse van verschillende bronnen. Teneinde deze toepassing te kunnen realiseren hebben we de volgende "bouwstenen" nodig:

- Meta search (3.1.2) – wordt gebruikt om meerdere internet bronnen te ontsluiten en samen te voegen.
- Entiteiten zoeken (3.1.5) – wordt gebruikt om plaatsen te identificeren.
- Text mining (3.2.1) – wordt gebruikt om voor ieder stuk tekst aan te geven of het om een mogelijke aanslag gaat.
- Automatisch samenvatten (3.2.4) – wordt (optioneel) gebruikt om de gevonden stukken tekst automatisch in te korten.
- Identity tracking (3.3.1) – wordt gebruikt om gevonden aliassen te disambigueren.
- Graaf/Relatie analyse (3.3.3) – wordt (optioneel) gebruikt de relaties met andere personen te identificeren en beschrijven.
- Taaldetectie (3.2.7) en machinaal vertalen (3.2.8) – wordt (optioneel) gebruikt om stukken tekst in een vreemde taal te vertalen.

6 Conclusie

In dit document hebben we beschreven wat de mogelijkheden omtrent onderzoek en implementatie zijn, uit te voeren door de UvA in het kader van het iRN Open Web Observatie project. Aan de hand van een viertal onderzoeksthema's hebben we meer specifieke analyse mogelijkheden uitgelicht, waarbij voor ieder nader wordt beschreven wat het nut is voor het iRN, wat de huidige stand van zaken is en hoeveel werk er nodig is teneinde ieder onderdeel succesvol in te zetten. Om een duidelijk beeld te krijgen van use-cases zijn een aantal interviews afgenomen bij de mogelijke eindgebruikers van het iRN. Deze use-cases zijn gebruikt als concrete voorbeelden, zowel bij de onderzoekslijnen als op zichzelf staande toepassingen.