FINAAL RAPPORT
INFO-NS
“Intelligente” exploitatietools voor niet gestructureerde informatie ten behoeve van de noden van de Federale Politie

Versie 1.2: 28 november 2005

Auteurs:

Jan De Beer
Marie-Francine Moens
Interdisciplinair Centrum voor Recht en Informatica
K.U.Leuven

Nishant Kumar
Jan Vanthienen
Onderzoekscentrum in Beleidsinformatica
K.U.Leuven

1 Inleiding

Het *INFO-NS project* of *"Intelligente" exploitatietools voor niet gestructureerde informatie ten behoeve van de noden van de Federale Politie* betreft een evaluatie van commerciële zoekmachines en text mining tools voor het doorzoeken en verwerken van de tekstuele bronnen van de Federale Politie van België.

De snelle evolutie van de technische mogelijkheden (transport, communicatie, financiële spitstechnologie,...) biedt criminelen streeds meer mogelijkheden om hun misdrijven met meer complexiteit, diversiteit en flexibiliteit te plegen. De klassieke politiestrategie (*crime fighting*), waarbij onveiligheid wordt aangepakt met toezicht, interventie, en onderzoek, wordt aangevuld met een informatiegestuurde politiezorg (*intelligence led policing*).

De geïntegreerde politiediensten van België (Lokale en Federale Politie) beschikken over enorme hoeveelheden van niet gestructureerde informatie. Zij hebben een grote nood inzake de exploitatie en analyse van deze informatie. Onder niet gestructureerde informatie wordt bedoeld alle digitale informatie die op een niet gestructureerde manier is opgeslagen in databanken: 1) niet gestructureerde teksten waaronder de oorspronkelijke tekst van PVs (processen-verbaal), informatierapporten, interne nota's, maar ook open bronnen zoals persartikels, rapporten, wetgeving; 2) niet gestructureerd beeld- en geluidsmateriaal zoals foto's van sporen en plaatsen, video-fragmenten en geluidsopnamen van manifestaties, enz.

Het INFO-NS-project is gegroeid binnen de Federale Politie omwille van een absolute noodzaak aan een adequaat systeem voor:

- de selectie van de meest pertinente documenten;
- de identificatie van de meest gezochte items;
- de identificatie van relaties tussen items;
- de structurering en categorisering van deze informatie en hun opslag in databanken voor verder gebruik.

De gebruikers van een dergelijk intelligent systeem, op het niveau van de geïntegreerde politie zijn:

- het geheel van de diensten die voor de managementondersteuning zorgen en die betrokken zijn bij het verwerken van informatie (gestructureerd of niet) met het oog op analyse en met als doel het uitstippelen van beleidslijnen;
- de operationele diensten die gestructureerde en niet gestructureerde informatie gebruiken in het kader van onderzoeken en voor het zoeken naar bijkomende aanwijzingen/informatie.

De door de onderzoeksploeg uit te voeren opdracht in het INFO-NS project is als volgt gedefinieerd:

- Een analyse van de behoeften van de geïntegreerde politie en de verwachtingen van de verschillende diensten inzake de exploitatie van niet gestructureerde informatie.
- Een benchmarking van de op de markt beschikbare (commerciële) producten na voorafgaande opstelling van testcases en evaluatiecriteria, in samenspraak met de betrokken politiediensten.
- Het opstellen van een overzicht van de randvoorwaarden die moeten vervuld zijn om een integratie van de commerciële producten en optimale exploitatieresultaten mogelijk te maken.

Gezien de huidige technologische beperkingen in de exploitatie van beeld en geluid is het onderzoek beperkt tot de evaluatie van de commerciële tools die tekst doorzoeken en informatie eruit ontginnen.

De onderzoeksploeg van het INFO-NS-project bestaat uit twee junior onderzoekers: Jan De Beer en Nishant Kumar, verbonden aan de K.U.Leuven, respectievelijk aan het Interdisciplinair Centrum voor Recht en Informatica (ICRI), en het Onderzoekscentrum in Beleidsinformatica (LIRIS). Ze werden respectievelijk begeleid door Marie-Francine Moens (hoofddocent aan de K.U.Leuven en coördinator van INFO-NS) en Jan Vanthienen (gewoon hoogleraar aan de K.U.Leuven). De opdrachtgever van het project is het Federaal Wetenschapsbeleid (programma AGORA), en de projectbeheerder is Mevrouw Lieve Van Daele.[1] In alle fasen van het project is er nauw samengewerkt met de volgende personen van de Federale Politie: de Heer Kris d'Hoore (DGS/DSB), de Heer Guy Dahmen en de Heer Marc Duforez (DGS/DST), Mevrouw Martine Pattyn en de Heer Paul Wouters (CGC/ASA), en de overige leden van de gebruikersgroep, waaronder de Heer Pascal Fleron (CIA Marche-en-Famenne), de Heer Marc Vandewalle en de Heer André Deblaere (DGJ/DJO), de Heer Fons Schoonaerts (DGJ/DJF), de Heer Jean Marc Lietaer en de Heer Guy Verberren (DGJ/DJF/FCCU). Het begeleidingscomité van dit project bestaat uit bovengenoemde personen, alsook Mevrouw Diane Reynders en Mevrouw Fabienne Polain (Service Politique criminelle), Mevrouw Laura Szabo (SPF Intérieur), Mevrouw Annick Castiaux (FUNDP), de Heer Sven Forster (FEDICT), de Heer Christophe Onraet, de Heer Johan Truyens, en de Heer Alain Hugelier (SPF Défense), Mevrouw Gaël Kermarrec (Federaal Coördinatiecomité), de Heer Wim Geens (CPPL), de Heer Yannic Hulot (Pol Fed OCDEFO), de heer Marc Borry (Pol Fed CDC), de Heer Marc Vandendriessche (Pol Fed DSB) en de Heer Anatole Wojciechowski (Pol Fed CGC). De onderzoeksploeg dankt de input en inzet van deze personen.

Het INFO-NS-project liep over een periode van 13 maanden (van 1 oktober 2004 tot en met 31 oktober 2005). De onderzoeksploeg heeft de genoemde doelstellingen van het INFO-NS-project behaald. Dit finaal rapport geeft een overzicht van de voornaamste onderzoeksresultaten. Het is als volgt georganiseerd. Een eerste sectie beschrijft de resultaten van een literatuurstudie over de mogelijkheden van de huidige tools voor het exploiteren van ongestructureerde, tekstuele informatie, en het gebruik van deze tools door politiediensten wereldwijd. Een tweede deel beschrijft de analyse van de gebruikers- en systeembehoeften van de Federale Politie van België. Vervolgens beschrijven we de evaluatiecriteria die de onderzoeksploeg heeft

[1] http://www.belspo.be/agora

opgesteld. Sectie 5 beschrijft de criteria en de resultaten van een eerste selectie van commerciële producten. Een belangrijk deel van dit rapport bevat de evaluatie van de resultaten van de geselecteerde tools voor wat betreft het doorzoeken van documenten, het classificeren, het extraheren van informatie en het visualiseren van de resultaten, opgenomen in sectie 6. Tenslotte worden de randvoorwaarden besproken die moeten voldaan zijn indien men wil overgaan tot de implementatie van elk van de geselecteerde tools.

2 Literatuurstudie

Het objectief van de literatuurstudie is om verslag uit te brengen van de mogelijkheden om niet gestructureerde informatie en gestructureerde informatie te verwerken in het kader van politionele werkzaamheden. Er werd een overzicht gemaakt van de beschikbare technologieën voor de ontsluiting, de extractie en classificatie van tekstuele informatie. Daarnaast werden beslissingsondersteunende systemen bestudeerd die worden gebruikt voor de visualisatie en exploitatie van gestructureerde informatie die beschikbaar is in politiediensten. Beslissingsondersteunende technologieën worden vooral toegepast in het domein van strategische en tactische misdrijfanalyse. De besproken systemen omvatten geografische informatiesystemen, statistische technieken, online analytische verwerking, graph mining waaronder ook sociale netwerkanalyse valt, en automatische kennisextractie uit databanken (data mining).

We verduidelijkten en illustreerden de technologieën met een aantal buitenlandse praktijkvoorbeelden. Als voornaamste voorbeeld beschouwden we het pilootproject COPLINK dat momenteel operationeel is in een aantal politiediensten in de staat Arizona (USA). De architectuur van het COPLINK-systeem heeft drie lagen (*three-tier*). *Connect* is het platform van gegevensintegratie en -ontsluiting. Dit platform wordt gekarakteriseerd door haar eenvoud in gebruik; een bewuste keuze, gezien de heterogene gebruikersgroep van politieagenten met hoofdzakelijk uniforme, eenvoudige informatienoden. *Detect* is ontwikkeld specifiek voor speurders en misdrijfanalisten. Zo biedt *Detect* geavanceerde zoekfuncties en de mogelijkheid tot constructie van conceptnetwerken. *Collaboration* is het platform ter ondersteuning van de onderlinge communicatie en samenwerking door een groep van agenten of speurders. De studie van het COPLINK-systeem gaf ons een goed inzicht in de vereisten van de gebruikers en van het systeem zelf, en leverde ons een aantal criteria voor de evaluatie van intelligente exploitatietools voor politionele informatie.

Daarnaast werden nog enkele andere systemen onderzocht waaronder CrimeStat, STAC, CLEAR, FLINTS, OVER, en DataDetective. Het viel ons op dat een grondige evaluatie van de systemen met betrekking tot nauwkeurigheid, schaalbaarheid en performantie zeer dikwijls ontbreekt. Er wordt ook weinig aandacht besteed aan niettemin cruciale aspecten zoals beveiliging, privacy, rechtsgeldigheid van de afgeleide informatie, en de aanpasbaarheid van deze tools in het licht van de integratie van velerlei, heterogene bronnen.

Tabel 2.1 geeft een overzicht van de huidige technologie voor het exploiteren van politionele informatie. Voor meer details over de resultaten van dit onderzoek verwijzen we naar het analyseverslag – hoofdstukken 2 en 3.

Technologie †	Toepassingen
Enabling technologies – op ongestructureerde informatie	
Informatieontsluiting	Zoeken op sleutelwoorden, entiteiten (personen, groepen, organisaties, gebeurtenissen, voertuigen, plaatsen, drugs,...), metadata, fuzzy (fonetisch, morfologisch) en meertalig op zowel interne als externe (multimediale) documenten
Informatie-extractie	Automatische generatie (augmentatie) van gestructureerde informatie
NERC	Extractie van namen van entiteiten (personen, groepen, organisaties, drugs,...)
Tekstcodering	Schematisering van feitenmateriaal (inzichtelijke voorstelling in tijd en ruimte) en kwalitatieve analyses; frequentieanalyse van woorden of woordcombinaties, gerelateerde termen of concepten (MDS, LSI), topic maps
Classificatie	Documentenbeheer, ook voor zoekresultaten
Zoekagenten (monitoring)	Geïnformeerd blijven van mogelijks nieuwe feiten over de behandelde of gelijkaardige onderzoeken (signaalfunctie), voorzien in verbeterde onderlinge samenwerking en coördinatie
Informatie-exploitatie – op gestructureerde informatie	
Gestructureerde querytalen	Parametrisch zoeken naar entiteiten en feiten, met navigatie tussen de weergegeven informatie
Beslissingsondersteunende systemen	
Visualisatie & exploratie	Verkenning van de informatie, schematisering, trendanalyse, beeldvorming
Geografische informatiesystemen	Misdaadmapping
Statistisch (ruimtelijk/temporeel)	Misdaad hot spot analyse, analyse seriemisdaden (*journey-to-crime* en *correlated walk* analyse), generatie van veiligheidsstatistieken o.a. voor beeldvorming
Online analytische verwerking	Trendanalyse, beeldvorming
Graph mining	
Linkanalyse, conceptenruimte	Analyse van de correlatie tussen entiteiten, constructie van geattribueerde relationele grafen; misdaadnetwerken, communicatienetwerken, transactienetwerken, conceptnetwerken, topic maps,...
Gerelaxeerde subgraafisomorfisme	Zoeken naar verdachte patronen (*target patterns*) in geobserveerde activiteiten en ander feitenmateriaal
Sociale netwerkanalyse	Karakterisering van misdaad- en andere netwerken; topologie, sleutelfiguren (bendeleiders, tussenpersonen, informanten, enz.) aan de hand van SNA metrieken
Blokmodellering	Identificatie van groeperingen en spilfiguren, hiërarchisch
Data mining	Zoeken naar patronen, trends, en relaties voor beschrijving en voorspelling
Clustering	Profilering van delinquenten, hot spot analyse
Classificatie, record linkage, patroonherkenning	Beschrijving (karakterisering) van types daders en hun misdrijven aan de hand van modus operandi, persoonlijke en andere kenmerken, dadervoorspelling, auteurherkenning, biometrieken (identificatie), valse identiteitsdetectie
Regressie	Beschrijving en voorspelling misdaadprevalentie en andere statistieken over tijd en/of ruimte
Associatieregels	Ontdekken van patronen in modus operandi, recidiven
Kennisgebaseerde systemen	Gebruik van logisch programmeren voor het clusteren van seriemisdaden en het toekennen van nieuwe misdrijven aan de gevormde clusters
Gevalsgebaseerd redeneren	Ophalen van antecedenten of gerelateerde feiten

† Het gebruikte classificatieschema voor technologie is slechts ter overzicht; veel van deze technologieën zijn immers deels overlappend of steunen op elkaar.

Tabel 2.1

3 Analyse van de vereisten

In deze fase van het project is een analyse van de **behoeften van de gebruikers** opgesteld. Voor een contextuele duiding van een domein dat voor de onderzoeksgroep onvertrouwd was, werd vooreerst de politiehervorming in België bestudeerd. Met het verkregen inzicht in de structuur en de organisatie van de eengemaakte politie werden de potentiële gebruikers bepaald en ingedeeld in de profielen: operator, beheerder, onderzoeker, operationele en strategische analist. Daarnaast werden de gegevensinfrastructuur, inclusief de informatiebronnen en informatiestromen nauwkeurig onderzocht. Tabel 3.1 geeft een overzicht van de informatiebronnen opgedeeld naar oorsprong en structurering.

Binnen de verschillende departementen en diensten van de politie werd navraag gedaan naar de wenselijkheid van informaticatools die een ondersteuning kunnen bieden voor het ophalen, doorzoeken, beheren, structureren, visualiseren, en extraheren van relevante informatie uit (elektronisch beschikbare) tekstuele documenten. Daarnaast werden ook de noden van de exploitatie van gestructureerde informatie geanalyseerd. De behoeften werden vastgelegd middels vragenlijsten en gesprekken met gezagvoerende personen in een aantal diensten van de Federale Politie. De resultaten van dit onderzoek zijn uitgewerkt als functionele noden in Tabel

3.2. Deze studie gaf ons een goed inzicht in de doelstellingen van INFO-NS en in de absolute noodzaak om informatie te zoeken in ongestructureerde bronnen.

	Gestructureerd	Ongestructureerd
Intern	• ANG • INDEX • NEMESIS • Interventiegegevens • Expertendatabanken • ...	• Processen-verbaal (aanvankelijk, navolgend, synthese) • Kantschriften, gerechtelijke dossiers • Informatierapporten (RIR, RAR) • Onderzoeksfiches (DOS) • Analyserapporten • Interne nota's • Interne documentatie (PolDoc, IntraDoc) • Seiningsbladen (RIB) • Video, geluid, foto's • Vingerafdrukken • ...
Extern	• Kruispuntbanken (KBO, KSZ) • Rijksregister (RR) • DIV • CWR • SIDIS • NIS • Euro DB • Expertendatabanken (gerecht, overheidsdiensten) • ...	• Media (kranten, tv, teletekst, nieuwssites) • Internet • Belgisch Staatsblad • Wetgeving (richtlijnen, reglementen) • Rapporten domeinexperten, universiteiten • Boeken, literatuur • ...

Tabel 3.1

Configuratie

Indexeringsproces

Het beheer van het indexeringsproces, waarbij een centrale *index* (inventaris) wordt aangemaakt en onderhouden van alle documenten waarvoor exploitatie wenselijk is.

Vindplaatsen

Instellen van de vindplaatsen van de documentendepots die geïndexeerd dienen te worden. Onder *documentendepot* wordt begrepen de fileservers en/of bestandensystemen met documenten die mee worden geïndexeerd.

Connectiviteit

Instellen van de toegang tot de documentendepots (connectiviteit- en toegangsprotocol).

Planning

Controle over de uitvoeringstijden van het indexeringsproces (hernieuwingsfrequenties).

Beveiliging

Instellen van de toegangsrechten, waaronder autorisatie voor het beheer van de tool, afscherming van de indexstructuren van de tool, en controle over het uitvoeren van de functionaliteiten op de documenten door de verschillende gebruikers.

Indexering

Crawling

Autonoom scannen van de documentendepots voor de aanmaak en het onderhouden van de centrale index.

Toevoegen

Indexeren van documenten in de depots die nog niet in de index zijn opgenomen.

Verwijderen

Verwijderen van de index van documenten die niet meer voorkomen in de depots (zoals na afloop van de bewaartermijn).

Hernieuwen

Hernieuwen van de index van gewijzigde documenten.

Documentformaten

De indexering van verschillende documentformaten, zowel qua bestandstype als qua inhoudstaal.

Type

De ondersteuning van verscheidene bestandstypes, in hoofdzaak .PDF, .DOC, .XLS, .PPT, en .HTML (webpagina's).

Taal

De ondersteuning van verscheidene inhoudstalen, in hoofdzaak beperkt tot de officiële landstalen (Nederlands, Frans, Duits), en het Engels voor wat betreft open bronnen.

Metadata

Toekennen van metadata aan de geïndexeerde documenten (zie ZM), extern of via hergebruik van bestaande gegevens.

Extern

Toekenning van metadata via de tool.

Hergebruik

Hergebruik van bestaande databanken met metadata.

Indelen collectie

Automatische indeling van de collectie volgens de inhoud van de documenten. Men onderscheidt clustering van classificatie.

Clustering

Het groeperen van documenten naar overeenkomst in inhoud, mogelijks hiërarchisch en in overlappende clusters.

Classificatie

Zoals clustering, maar met meertalige labelling van elke cluster (*categorie* of *klasse* genaamd). [a]

Zoeken

Zoeken op metadata

Filteren van de collectie volgens een aantal criteria met betrekking tot een of meerdere documentattributen met vastgestelde semantiek (*metadata*). [b] Als nuttige metadata beschouwen we:

Id
Een identificator die elk geïndexeerd document uniek aanduidt. Dit laat toe om gericht te zoeken naar een welbepaald document, bijvoorbeeld op basis van een afgedrukte resultaatlijst (zie AEP).

Url
De URL (webadres) van documenten gepubliceerd op het intranet of internet.

Titel
Een bevattelijke titel van ieder document.

Type
Het type van informatiebron.

Taal
De taal waarin het document werd opgesteld.

Herkomst
De herkomst van het document; de politiezone, centrale dienst, intranet of internet.

Tijdstip feit
Het tijdstip van het beschreven feit.

Tijdstip publicatie
Het tijdstip van publicatie (indexatie) van het document.

Als nuttige zoekfilters beschouwen we:

Exacte tekstovereenkomst
Het selecteren van documenten waarvan het beschouwde, tekstuele metadatum exact overeenstemt met de opgegeven zoekstring.
Toepasbaar op: id, url, titel

Partiële tekstovereenkomst
Het selecteren van documenten waarvan het beschouwde, tekstuele metadatum partieel overeenstemt met de opgegeven zoekstring.
Toepasbaar op: id, url, titel

Nominale overeenkomst
Het selecteren van documenten waarvan het beschouwde, nominale metadatum is opgenomen in een opgegeven verzameling van gewenste waarden.
Toepasbaar op: type, taal, herkomst document

Tijdsovereenkomst
Het selecteren van documenten waarvan het beschouwde, tijdsaanduidend metadatum voor, tussen, na of gelijk is aan de opgegeven tijdsaanduiding.
Toepasbaar op: tijdstip feit, tijdstip publicatie

Combinatie
Een combinatie van bovenstaande zoekfilters op metadata.
Toepasbaar op: alle metadata

Zoeken in vrije tekst

Filteren van de collectie volgens de inhoud van het document, met ondersteuning van

Zoektermen

Het opgeven van een combinatie van een of meerdere zoektermen; losstaand, in nabijheid (*proximity search*), of als geheel (*phrase search*).

Meertaligheid

Het terugvinden van documenten in een taal onafhankelijk van de taal waarin de zoekopdracht werd gesteld. [a]

Partiële overeenkomst

Het terugvinden van documenten op basis van partiële overeenkomst tussen de zoektermen en de inhoudstermen. Dit moet toelaten om verschillende schrijfwijzen, spellingsfouten, typ- of OCR fouten, afkortingen,... enigszins op te vangen.

Gerelateerde termen

Het terugvinden van documenten met inhoudstermen die semantisch nauw verbonden zijn aan de opgegeven zoektermen (niet beperkt tot synoniemen). [c]

Zoeken naar entiteiten

Filteren van de collectie volgens de vermelding van een of meerdere entiteiten, met ondersteuning van

Entiteitstypes

Zoeken naar verschillende types van entiteiten, waaronder: personen, groeperingen, organisaties, voertuigen, plaatsen, tijdstippen, producten (drugs, hormonen, wapens,...), geldbedragen, terrorismefinanciering,...

Zoeknamen

Het opgeven van een combinatie van een of meerdere namen van entiteiten; losstaand of in nabijheid (*proximity search*).

Meertaligheid

Het terugvinden van vermeldingen in een taal onafhankelijk van de taal waarin de zoeknamen werden geformuleerd. [a]

Voorbeeld: Zoeken naar "Oostende" geeft documenten met vermelding "Oostende", "Ostende", "Ostend", en diens meer.

Partiële overeenkomst

Het terugvinden van documenten op basis van partiële overeenkomst tussen de zoeknamen en de entiteitsvermeldingen. Dit moet toelaten om verschillende schrijfwijzen, typ- of OCR fouten, verkorte (afkortingen en initialen) of verlengde vormen (familienamen), woordordes,... enigszins op te vangen. Een basisvereiste is het fonetisch zoeken op namen; overeenkomst in klank. [d] Zoeken op partiële overeenkomst kan worden aan- of uitgeschakeld.

Zoeken volgens indeling

Filteren van de collectie volgens de indeling uit II.

Categorie selectie

Filteren door selectie van één of meerdere weergegeven categorieën.

Categorie opzoeking

Filteren door opzoeking van één of meerdere categorieën volgens (meertalig) label.

Cluster selectie

Filteren door selectie van één of meerdere weergegeven clusters.

Zoeken volgens voorbeeld

Filteren van de collectie volgens inhoudelijke gelijkheid met een opgegeven voorbeelddocument, met ondersteuning van

Meertaligheid

Het terugvinden van gelijkaardige documenten in een taal onafhankelijk van de taal van het voorbeelddocument. [a]

Voorbeeld: Zoeken naar antecedenten of gerelateerde feiten (verbandlegging).

Monitoring

Passieve vorm van documentbevraging, waarbij op basis van gebruikersprofielen en/of voorgaande zoekopdrachten, nieuwe relevante informatie automatisch ter kennis wordt gesteld aan de geïnteresseerde gebruikers.

Interactie
Geassisteerde formulering zoekopdracht Het automatisch samenstellen van de zoekopdracht door navigatie en selectie binnen een *thesaurus* of *topic map*.
Verfijning zoekopdracht Progressieve verfijning van de zoekopdracht op basis van voorgaande resultaten. **Samengestelde zoekopdrachten** Opeenvolgende (multimodale) zoekopdrachten met progressieve uitdunning van het resultaat, optioneel met aanduiding van en navigatie doorheen de weergegeven zoekgeschiedenis. **Relevantiefeedback** Manuele aanduiding van relevante en irrelevante zoekresultaten, opdat de tool een beter onderscheid kan maken.
Herhaalde zoekopdracht De mogelijkheid tot het bewaren van de zoekopdracht voor een eventueel zelfde bevraging op een later tijdstip.
Weergave zoekresultaat Verkorte weergave van de geïndexeerde documenten die door de tool relevant werden bevonden in het licht van de uitgevoerde zoekopdracht. **Rangschikking volgens relevantie** Uitgezonderd van het louter zoeken op metadata en volgens indeling, het geordend weergeven van de resulterende documenten volgens afnemende mate van relevantie. **Weergave relevantie** Aanduiding van de relevantie voor ieder resulterend document. **Weergave metadata** Aanduiding van de metadata voor ieder resulterend document (zie ZM). **Weergave synopsis** Aanduiding van de verkorte inhoud voor ieder resulterend document. **Consultatielink** Aanduiding van een verwijzing ter consultatie van ieder brondocument, mogelijks gerealiseerd via één der metadata aanduidingen (id, titel of URL bijvoorbeeld).
Exporteren zoekresultaat Exporteren van het zoekresultaat. **Printen** Afdrukken van het zoekresultaat. **Naar bestand** Opslaan van het zoekresultaat naar een interpreteerbaar (courant ondersteund of open) bestandsformaat.
Herschikken zoekresultaat Ordenen van het zoekresultaat volgens metadata (zie ZM).

Indelen zoekresultaat

Automatische indeling van de documenten die door de tool relevant werden bevonden in het licht van de uitgevoerde zoekopdracht, volgens de inhoud van de documenten. Men onderscheidt clustering van classificatie (zie II).

Verkennen indeling

Navigatie doorheen de indeling uit II of AI, met een overzicht van de geïndexeerde documenten ressorterend in de bezochte clusters (categorieën).

Consultatie document

Consultatie van ieder brondocument. [e]

Weergave inhoud

Weergave van de inhoud van het brondocument.

Markeren relevante inhoud

Aanduiding van de passages binnen het brondocument die overeenkomen met de zoekopdracht.

Voorbeeld: Aanduiding zoektermen en zoeknamen.

Printen inhoud

Afdrukken van de inhoud van het brondocument.

Kopiëren inhoud

Kopiëren van een selectie van de inhoud van het brondocument.

Kwalitatieve analyse

Ontdekken relaties

De identificatie van relaties in de teksten van een uitgekozen verzameling van documenten.

Termen

De identificatie van geassocieerde termen; termen die vaak samen voorkomen in documenten, indien mogelijk met ondersteuning van meertaligheid.

Concepten

De identificatie van concepten; verzamelingen van geassocieerde termen, indien mogelijk met ondersteuning van meertaligheid.

Entiteiten

De identificatie van geassocieerde entiteiten, indien mogelijk met ondersteuning van meertaligheid.

Tekstcodering

Het annoteren van teksten ter structurering van de informatie die erin vervat zit.

Geassisteerde annotatie

De geassisteerde annotatie van teksten (cfr. NVIVO [f] en i2 TextChart [g]).

Entiteitherkenning

De automatische herkenning en classificatie van vermeldingen van namen van entiteiten, indien mogelijk met ondersteuning van meertaligheid.

Exploitatie analyse

De exploitatie van de analyses zoals uitgevoerd in takenpakketen KR en KC.

Termrelaties

De visualisatie van geassocieerde termen, of de mogelijkheid tot het exporteren naar een interpreteerbaar (courant ondersteund of open) bestandsformaat.

Concepten

De visualisatie van concepten of de mogelijkheid tot het exporteren naar een interpreteerbaar bestandsformaat.

Entiteitsrelaties

De visualisatie van geassocieerde entiteiten (*sociale netwerken*, cfr. UCINET en PAJEK [h]) of de mogelijkheid tot het exporteren naar een interpreteerbaar bestandsformaat.

Tekstannotaties

De visualisatie van tekstannotaties of de mogelijkheid tot het exporteren naar een interpreteerbaar bestandsformaat.

Schemacreatie

Het maken van grafische voorstellingen van de inhoud van teksten (feitenmateriaal) op basis van de geannoteerde teksten uit KC (cfr. i2 Analyst Notebook [g]).

[a] Ondersteuning voor het Nederlands, Frans, Duits en Engels is een basisvereiste.

[b] Wanneer men zou opteren voor het gebruik van codes en keuzelijsten is het wel aangewezen om meertalige labels in natuurlijke taal te voorzien. [a]

[c] We vermelden het bestaan van sleutelwoordlijsten (thesauri) voor misdaadfenomenen, die mogelijks door de tool kunnen aangewend worden.

[d] Naamvariaties zijn ruimer dan door fonetisch zoeken kan gecompenseerd worden, zo bijvoorbeeld het gebruik van initialen, verkorte vormen, bijnamen (aliassen), roepnamen, verlenging familienaam door huwelijk, enz.

[e] Uitgezonderd het downloaden van de brondocumenten vanuit de tool, dat wegens beveiligingsredenen niet kan worden toegestaan.

[f] Meer informatie over NVIVO op www.qsrinternational.com.

[g] Meer informatie over i2 op www.i2.com.

[h] Meer informatie over UCINET en PAJEK op www.byeday.net/sna/software.html.

Tabel 3.2

In deze fase van het onderzoek werden tevens de **technische noden** vastgelegd. Deze omvatten:

1. Interoperabiliteit: De mogelijkheid tot het gebruik van de tool door meerdere, gedistribueerde client-systemen, die in verbinding staan met de tool (mogelijks onrechtstreeks) via netwerkprotocollen.

Voor de mogelijkheid tot interoperabiliteit met toekomstige systemen en een flexibele inpassing in de huidige (toekomstige) architectuur, wordt de voorziening van een publieke API (*Application Programming Interface*) gevraagd, die programmatorische toegang verschaft tot de aangeboden functionaliteiten van de tool.

2. Systeemarchitectuur: De servermachine voor de tool wordt mogelijks een HP-machine. Wat vaststaat is dat deze RedHat Enterprise Linux V3 -- Advanced server -- Upgrade 4 als besturingssysteem zal hebben. Voorts werd de benodigde opslagruimte voor de documentencollectie DOCMAN, inclusief indexstructuren geschat op ongeveer 1.8 TB.

3. Systeembronnen: Deze vereiste betreft het beperkt gebruik van systeembronnen, waaronder processorcapaciteit, werkgeheugen, opslagruimte, enz. Harde eisen kunnen a priori niet gesteld worden, tenzij wat betreft de maximaal toegelaten netwerkbandbreedte die, zoals voor alle andere toepassingen, werd vastgesteld op 9.6 kilobit per seconde.

4. Beveiliging: Wat betreft de afdwingbaarheid en de preciese uitwerking van de toelatingsvoorwaarden tot het consulteren en het uitvoeren van de opgetekende functionaliteiten op de documenten, zijn er nog geen concrete specificaties. Wat echter vaststaat is dat de tool het in voegen zijnde beveiligingsprotocol dient te implementeren, en dat moet worden voorzien in het loggen van alle transacties.

5. Documentformaten: De ondersteuning van verscheidene documentformaten, in hoofdzaak PDF, DOC, XLS, PPT, en webpagina's (HTML). Als versies van PDF noteren we 1.2 en 1.3 en wat betreft de overige formaten uit de Microsoft Office suite, alle versies gaande van 6.0. De te ondersteunen inhoudstalen zijn in eerste instantie beperkt tot de officiële landstalen (Nederlands, Frans, Duits), en het Engels voor wat betreft open bronnen.

6. Voldoening: Aandacht voor een aantal subjectieve parameters, waaronder up-to-dateheid van de zoekresultaten (gezien de eindige tijd nodig voor een tour doorheen de collectie door het crawlingproces), de gebruikersvriendelijkheid van de tool (eenvoud in gebruik, intuïtieve gebruikersinterface, snelle responsietijd), eenvoud in installatie en onderhoud, duidelijke en volledige handleidingen en referentiemateriaal, goede ondersteuning, etc. Eigen indrukken hiervan door installatie, studie, gebruik, en evaluatie van de tools zullen worden opgetekend naast de objectieve, cijfermatige testresultaten, en mede worden gepubliceerd in het benchmarkingverslag.

Voor meer details over de resultaten van dit onderzoek verwijzen we naar het analyseverslag, hoofdstukken 4 en 5.

4 Evaluatiecriteria

In deze fase van het project zijn de evaluatiecriteria vastgelegd waaraan de commerciële softwarepakketten voor het doorzoeken van tekst en het ontginnen van informatie uit tekst tijdens de benchmarkingfase zullen worden onderworpen, indachtig de opgetekende prioriteiten aangaande de functionaliteiten uit de analyse van de vereisten, en een aantal gangbare criteria uit de wetenschappelijke literatuur.

Concreet hebben we drie categorieën van criteria onderscheiden, die we respectievelijk aanduiden met de termen conformiteits-, kwalitatieve, en technische criteria. Voor elk van deze groepen werkten we eerst een algemeen, generisch evaluatiemodel uit als formeel en herbruikbaar raamwerk, dat we vervolgens specificeerden zodat het evaluatiemodel voldoet aan de vereisten van het project INFO-NS. We erkennen dat een aantal van deze criteria en evaluatiemodellen is uitgewerkt ter volledigheid van deze studie, en niet is toegepast in de latere benchmarkingfase omwille van een gebrek aan tijd gekoppeld met een duidelijke prioriteitsbepaling van de gewenste functionaliteiten, en het vervallen van de relevantie/noodzaak van enkele evaluaties omwille van de soms erg gelijkaardige, prille resultaten behaald met de tools, op basis waarvan reeds voldoende inzichten werden verkregen.

Het vastleggen van objectieve en adequate evaluatiecriteria was beslist geen triviale taak. Vooreerst is het begrip relevantie als de kwaliteit of de bruikbaarheid van de geproduceerde resultaten in hoge mate subjectief en contextafhankelijk (gebruiker, taak en finaliteit). Bijgevolg berusten evaluaties veelal op menselijke interactie en beoordeling, waardoor de voorgestelde criteria zuinig dienen om te springen met benodigde tijd en moeite. Tenslotte spelen een groot aantal van diverse en moeilijk vergelijkbare (combineerbare) factoren en criteria, ook niet-kwalitatieve, een rol bij de uiteindelijke appreciatie van een bepaalde tool.

1. Conformiteitscriteria

We gingen na in hoeverre elk van de voorgeselecteerde tools beantwoordt aan de functionele noden van de geïdentificeerde gebruikersprofielen. In sectie 5 zullen we zien dat we deze aftoetsing konden gebruiken voor het uitvoeren van de voorselectie zelf, waarbij we enkel de tools overhouden die de moeite lonen om diepgaande, kwalitatief inhoudelijke testen mee te gaan uitvoeren.

De ingrediënten voor onze werkmethode bestaan uit (a) Tabel 3.2 met individuele, uitgesplitste functionaliteiten, (b) de prioriteiten die per profiel aan elk van deze functionaliteiten werden toegekend, en (c) de voorziening ofte de mate van ondersteuning van de tools voor elk van deze functionaliteiten. Voorts bakenden we een beperkt aantal *use cases* af. Elke use case vertegenwoordigt een logische groepering van meerdere, samenhangende functionaliteiten. Zo onderscheiden we bijvoorbeeld het zoeken op metadata, het zoeken in vrije tekst, het ontdekken van gerelateerde entiteiten, enz. als afzonderlijke use cases.

Vervolgens stelden we voor elk van de use cases een boomstructuur op (zie figuur 2.1 in het evaluatiecriteriaverslag voor een uitgewerkt voorbeeld), die wordt opgebouwd door de individuele functionaliteiten waaruit de use case bestaat, op een logische

manier te integreren. Aan elk van de knooppunten uit deze structuur kunnen we een objectieffunctie hechten die, gegeven een particuliere tool en een gebruikersprofiel, een score berekent van deze tool voor de voorgestelde functionaliteit in dit knooppunt, met inachtneming van de prioriteiten van het beschouwde profiel. De berekening van deze scores verloopt gradueel, startend bij de knooppunten aan de uiteinden van de boom, wiens scores systematisch worden gecombineerd (additief, multiplicatief, of via andere operatoren) naar knooppunten hogerop in de boom, om finaal uit te komen op een globale score voor de tool op deze use case volgens het beschouwde profiel. Wanneer men deze berekening herhaalt voor alle gevallen, bekomt men als resultaat een tabel die per tool, per use case, en per profiel een conformiteits- of overeenkomstscore geeft.

Naast het optekenen van de deelresultaten, ondermeer voor de belangrijkste functionele componenten (takken) in de boomstructuur, en de samenstelling van de scoretabel, kunnen nog twee gecomprimeerde tabellen worden gemaakt. Door het samenstellen via de prioriteiten kan men immers komen tot (a) een globale score van elke tool voor elk profiel, en (b) een globale score van elke tool voor elke use case. Bij interpretatie van al deze cijfergegevens zijn het niet zozeer de absolute, dan wel de relatieve scores die een beoordeling mogelijk maken.

2. Kwalitatieve crtiteria

Tools kunnen dan wel volledige ondersteuning bieden voor bepaalde use cases, wat vooral van belang is, is uiteraard de kwaliteit waarmee ze deze taken volbrengen. In de mate van het mogelijke hebben we voor elk van de use cases aanvaardbare en uitvoerbare kwalitatieve criteria opgesteld. Aangezien deze voor sommige use cases geheel analoog verlopen en bij wijze van illustratie, beperkten we ons tot een beschrijving van de gevolgde methodologie voor de use cases ``het zoeken in vrije tekst'', ``het zoeken volgens indeling'', en ``de automatische herkenning van entiteiten in documenten''.

Zoeken in vrije tekst

Voor het zoeken in vrije tekst gingen we uit van (a) een collectie van ter beschikking gestelde documenten, (b) een collectie van typische zoekopdrachten zoals geponeerd door een gebruiker van de tool, mogelijks ingezameld per profiel. We laten deze zoekopdrachten elk afzonderlijk los op ieder van de tools, en vergelijken de gerangschikte lijst met documenten (geordend volgens relevantie) zoals die door elk van de tools werd geproduceerd als antwoord op onze zoekvraag .

Om de evaluatie praktisch niet te overladen, beschouwden we voorts enkel die documenten waarvoor de tools onderling het meest oneens zijn. Dit zijn documenten met erg verscheiden posities en/of opnames in de rangschikkingen. Voor elk van deze documenten lieten we de gebruiker de werkelijke relevantie beoordelen na consultatie van het document. Als eerste criterium stelden we dan de afwijking tussen deze werkelijke relevantiewaarden en de door een tool berekende relevantiewaarden, die ons ofwel gegeven zijn door de tool zelf, of die we konden afleiden (schatten) aan de hand van de rangschikking. In geval de verschillende tools het toch vrijwel eens zijn over de rangschikking van de documenten volgens relevantie, is het toch nuttig de

rangschikking van één tool te controleren. Immers alle tools kunnen een gelijkaardige, maar foute rankschikking genereren.

Als tweede criterium maten we de mate waarin minder relevante documenten zich mengen in de resultaatlijst, en de gebruiker hinderen in diens zoektocht naar relevante informatie. We definieerden de evaluatiemaatstaf *rpref*. Deze maatstaf is een adaptatie van de recent voorgestelde, binaire preferentiemaatstaf *bpref* naar reële (fuzzy) relevantiewaarden en is eveneens enkel toepasbaar voor geordende resultaatlijsten. *bpref* is een waardevolle maatstaf voor het relatief vergelijk van informatieontsluitingstools, die in tegenstelling tot de meer courante maatstaven vrij robuust is (consistent blijft) wanneer er wordt voorzien in slechts een beperkte set van relevantiemetingen. Gezien het opzet van onze selectieprocedure is deze maatstaf dus uiterst relevant. We onderstrepen dat aan de noodzakelijke en voldoende voorwaarde voor een correct gebruik van deze maatstaf in ons model is voldaan. De voorwaarde stelt immers dat de kans op aanwezigheid van een document in de lijst van relevante documenten onafhankelijk is van het al dan niet relevant zijn van dit document.

De filosofie achter beide preferentiemaatstaven (*rpref* en *pbref*) berust op een meting van de mate waarin irrelevante documenten zich mengen in de resultaatlijst, dit is, gerangschikt worden voor en tussen relevante documenten. Een grote inmenging zorgt er immers voor dat de gebruiker mogelijks heel wat tijd verliest met het consulteren van weinig relevante documenten.

Zoeken volgens indeling

Inzake het zoeken volgens indeling bestuderen we in welke mate de clustering (hiërarchische groepering) van de gehele documentencollectie (of als alternatief de geclusterde resultaatlijst van een zoekopdracht) een hulp kan betekenen bij het zoeken naar informatie. Die hulp bestaat er immers in van a priori, dit is, voor het loslaten van de zoekopdracht, aan te geven tot welke clusters men zijn/haar zoekopdracht wenst te beperken. Dit principe maakt het immers mogelijk om in sommige gevallen de precisie van de geproduceerde resultaten gevoelig te verhogen, zonder daarbij al te veel relevante documenten a priori uit te sluiten.

Teneinde dit te evalueren hernamen we het evaluatiescenario voor het zoeken in vrije tekst, waarvan we de gegevens en manuele relevantiebepalingen handig kunnen overnemen. Als extra gegeven laten we de gebruiker voorafgaand aan de uitvoering van een zoekopdracht voor elke tool afzonderlijk aangeven welke clusters van deze tool als relevant worden gedacht voor de zoekopdracht, dit is, tot dewelke men de zoekopdracht zou beperken indien men effectief kon gebruikmaken van de clustering. Eens dit is opgetekend, nemen we als criterium de werkelijke relevantie van deze geselecteerde clusters, alsook de irrelevantie van de niet geselecteerde clusters, die we berekenen (schatten) op basis van de gekende en voor waar genomen relevantiebepalingen van de gebruiker op de beperkte selectie van documenten (cfr. het zoeken in vrije tekst).

Entiteitsherkenning

Voor de automatische herkenning van entiteiten in documenten gebruikten we klassieke maatstaven uit de literatuur omtrent classificatietechnieken, met name

precision (precisie) en *recall* (volledigheid). Concreet gaan we uit van (a) een collectie van ter beschikking gestelde documenten, en (b) een collectie van beschouwde entiteitstypes (personen, groeperingen, voertuigen, enz.). We laten de tools alle vermeldingen van deze entiteitstypes aanduiden in de documenten, en beschouwen ter evaluatie - net zoals bij het zoeken in vrije tekst - enkel die vermeldingen waarvoor de tools onderling het meest oneens zijn. Hier geldt ook dezelfde opmerking als hoger.

Door manuele classificatie van deze beperkte verzameling van vermeldingen, meten we voor elk entiteitstype afzonderlijk de precisie van een tool als het percentage van correct geklasseerde vermeldingen ten aanzien van diegene die de tool heeft aangeduid, en de recall als het percentage van aangeduide en correct geklasseerde vermeldingen ten aanzien van alle vermeldingen van het beschouwde entiteitstype. Tenslotte kunnen we beide maatstaven combineren tot de klassieke *F-measure*, eventueel uitgemiddeld over alle entiteitstypes.

3. Technische criteria

Naast de functionele ondersteuning en de kwaliteit van uitvoering, beschouwden we tenslotte nog een laatste, erg heterogene groep van technische criteria. Deze criteria doelen enerzijds op het efficiënt gebruik van systeembronnen (waaronder geheugen, opslagruimte, netwerkbandbreedte) met bijzondere aandacht voor de schaalbaarheid naar grotere, realistische documentencollecties toe, en anderzijds op eerder subjectieve beoordelingen van de software, waaronder de gebruikersvriendelijkheid, de systeem/gebruikersinteractie, de kwaliteit van documentatie, en dergelijke meer.

Voor een gedetailleerd overzicht van de evaluatiecriteria verwijzen we naar het evaluatiecriteriaverslag (verslag 2).

Deze fase van het project resulteerde reeds in twee wetenschappelijke publicaties.

5 Eerste selectie van commerciële producten

Er werden een groot aantal bedrijven gecontacteerd die een product aanbieden voor de ontsluiting van informatie uit tekst (zie Tabel 5.1).

1. Autonomy Systems Ltd *Autonomy*

2. SPSS Ltd. *Clementine*

3. SAS Ltd. *SAS Text Miner*

4. Hummingbird Ltd. *Hummingbird KM*

5. Hyperwave Information Management, Inc

6. ClearForest Corp.

7. Verity software

8. Inxight Software *Inxight SmartDiscovery and Inxight VizServer*

9. Convera

10. Dolphin Search

11. i2 Ltd.

12. NetMap Ltd.

13. Entrieva Inc.

14 . Kofax *Mohomine*

15. Coplink

16. Ask Jeeves, Inc

17. Google

18. SRA International, Inc

19.IBM Belgium Luxembourg SA/NV

20. Megaputer-*Text Analyst*

21. Oracle *Oracle Text & Oracle Enterprise Search (2 producten)*

22. Metacarta

23. Sentient Information Systems Ltd. *Data Detective*

Tabel 5.1

Een eerste filtering van de tools is gebaseerd op productinformatie verkregen via verschillende kanalen zoals informatiebrochures en materiaal beschikbaar via het Internet die het product beschrijven, en gelijkaardige studies van de NASA en USA Homeland Security diensten. Daarnaast werden er vergaderingen belegd met de technische experts en vertegenwoordigers van deze bedrijven, die werden ondervraagd over het product. Deze vergaderingen gingen dikwijls gepaard met een demonstratie van het product door vertegenwoordigers van het bedrijf dat het product aanbiedt.

Enkel deze producten werden geselecteerd waarvan de functionaliteit beantwoordt aan de door de Federale Politie opgestelde vereisten. Deze eerste selectie resulteerde in de volgende lijst van producten:

1. Autonomy
2. Clementine
3. SAS Text Miner
4. Hummingbird
5. Inxight
6. Convera
7. Hyperwave
8. Entrieva
9. Verity
10. Oracle Text & Oracle ES

Door het niet verkrijgen van de software werden Autonomy en Entrieva geschrapt van deze lijst, terwijl Fast en Temis in de plaats werden gesteld tijdens verder marktonderzoek.

Voor een gedetailleerd overzicht van de selectieprocedure verwijzen we naar het verslag met de voorselectie van de tools (verslag 3).

6 Evaluatie van de geselecteerde producten

In deze fase van het project INFO-NS werden de testcases en testscenario's vastgelegd in samenspraak met de Federale Politie en werd de software van de geselecteerde tools voor het doorzoeken van tekst en het extraheren van informatie uit tekst getest en geëvalueerd.

Er werd aangegeven wat de aandachtspunten en de criteria zijn bij de evaluatie van de voorgeselecteerde tools, ingebed in een aantal concrete testcases voor de vermelde use cases uit het protocolakkoord, en een aantal algemene evaluatiepunten die los staan van enige use case, of betrekking hebben op alle use cases. Hierbij werd gebruik gemaakt van de hogerop beschreven conformiteits-, kwalitatieve- en technische criteria.

Testcases zijn opgesteld voor het zoeken in vrije tekst en op metadata, het classificeren van tekst, de extractie van entiteiten en het linken of het vinden van relaties tussen entiteiten. Er werd voor gekozen om de testcases omtrent `visualisatie' en `het opslaan van gevonden informatie' (zoals voorzien in het protocolakkoord) niet als afzonderlijke testcases te behandelen, maar deze aspecten voor elk van de te evalueren use cases nader te specificeren en op te nemen als onderdeel van hun evaluatie. De technische aspecten van een tool werden ook geëvalueerd.

We geven hier enkel een korte samenvatting van de evaluatieresultaten. Wegens contractverplichtingen met de betrokken leveranciers van de tools kunnen we in dit finaal rapport enkel een algemene appreciatie geven van de tools, zonder een betrokken tool bij naam te vermelden. We verwijzen echter naar een uitgebreid benchmarking rapport dat aan de Federale Politie werd overhandigd samen met de gedetailleerde registratie van alle benchmarking resultaten. Elk tool is geëvalueerd op basis van een zeer gedetailleerde lijst van uit te voeren testen.

Op basis van het behoeftenonderzoek bij de Belgische politie inzake exploitatietools voor ongestructureerde informatie, kwamen we tot een lijst van elf voorgeselecteerde producten die geschikt werden bevonden voor nadere evaluatie. Door de aard van de functionele noden en de aangeboden ondersteuning door de producten, kunnen we de geëvalueerde producten indelen in een beperkt aantal van gerelateerde families: informatie-ontsluiting (documentontsluiting), informatiebeheer (documentbeheer), informatie-extractie, en data mining (text mining).

Ondanks de grote overeenkomsten in architectuur, bediening, werkwijzen, en achterliggende technieken, kwamen toch een aantal duidelijke verschillen aan het licht tijdens de evaluaties. Zo werd voor de toepassing **documentontsluiting** bevonden dat de expressiviteit van de zoektaal (voortbouwend op het eenvoudigweg zoeken naar sleutelwoorden), de eenvoud in het gebruik, de voorziening van een duidelijke en makkelijk hanteerbare zoekinterface, en de kwaliteit van de geproduceerde zoekresultaten voorname beoordelingscriteria zijn waar tools vaak erg onderscheiden op scoren. Het mag dan ook opmerkelijk genoemd worden dat er één enkele tool is in de testsuite die al deze aspecten succesvol weet te combineren in een enkele toepassing die consistent de beste resultaten wist neer te zetten.

De **integratie van (reeds bestaande) gestructureerde informatie**, zoals ondermeer nuttig bij documentontsluiting, wordt in elke geteste tool erg gelijkaardig en voldoende ondersteund. Concreet wordt voorzien in connectoren naar de meest gebruikelijke (externe) opslagsystemen voor dergelijke gegevens. Het resultaat van deze connectoren is een overname van de gestructureerde informatie in de door de tool aangelegde indexstructuren, onder de vorm van bevraagbare metadata bij de geïndexeerde documenten. Zo kan bijvoorbeeld de datum van opmaak, het pv nummer, de politiezone, het type misdrijf, enz. apart worden opgenomen naast de tekstuele inhoud, en aldusdanig bevraagbaar worden gemaakt.

De **automatische classificatie** van documenten vindt men terug als standaard uitbreiding bij alle geteste documentontsluitingstools, en als standaard toepassing in alle geteste text mining tools.

In documentontsluitingstools wordt doorgaans gekozen voor het manueel opstellen van een regelgebaseerde taxonomie in een grafische interface, typisch door een domeinexpert. De expressiviteit van de regelgebaseerde classificatietaal ondersteunt minimaal de classificatie op basis van letterlijk voorkomende woorden of woordcombinaties in de tekst, en wordt in een aantal tools uitgebreid tot maximaal de volledige zoektaal. Dit betekent dat men alle documenten die beantwoorden aan een zelf gedefinieerde zoekvraag automatisch kan laten onderbrengen in een welbepaalde klasse.

Text mining tools daarentegen richten zich voor classificatie op de toepassing van standaard machinelerentechnieken. Deze technieken werken ofwel op basis van voorgeclassificeerde voorbeelddocumenten (gesuperviseerde technieken), of door het automatisch ontdekken van groepen van inhoudelijk gerelateerde documenten (ongesuperviseerde clusteringstechnieken). Een aantal tools maken een hybride toepassing mogelijk, met de automatische generatie van doorgaans editeerbare classificatieregels.

Bij de evaluatie van gesuperviseerde technieken voor een taxonomie met slechts drie klassen, zijn de resultaten op één product na (82% precisie en recall) eerder

teleurstellend te noemen. Als verklaring wordt gegeven dat het onderscheidingsvermogen van de gebruikte technieken en instellingen voor de documenten uit de aangelegde politionele corpus te beperkt is voor een echt succesvolle toepassing in een reële omgeving. Een meer doordachte keuze dan de standaard gekozen instellingen van de tools, maar vooral een verbeterde voorstelling van de documenten dan louter een verzameling van woorden, kan hierin verandering brengen.

Een aantal documentontsluitingstools wagen zich aan de **automatische extractie van entiteitsnamen** uit tekst (waaronder namen van personen, organisaties, locaties, geldbedragen, enz.). Deze toepassing wordt als uitbreiding geleverd bij alle geteste text mining tools, en vormt de basis voor informatie-extractietools. Daar waar de eerste familie van tools zich in hoofdzaak baseert op editeerbare en uitbreidbare woordenboeken, gaan informatie-extractietools doorgaans een stap verder met de voorziening van manueel editeerbare en uitbreidbare regelsets. Een regel kan hierbij de vorm aannemen van een reguliere expressie (bijvoorbeeld, een geldbedrag is een serie van decimale cijfers of groeperingstekens, gevolgd door één der valutasymbolen). Soms zijn ook complexere regels mogelijk op basis van taalkundige analyses van de tekst, zoals uitgevoerd door de tool. Hierdoor kan ook de context van de entiteitsvermelding mee worden beschouwd. Zo kunnen bijvoorbeeld bepaalde voorzetsels wijzen op een plaatsnaam eerder dan een persoonsnaam.

Beide technieken zijn echter entiteitstype- en taalspecifiek, en de mate van ondersteuning en omvattendheid van de standaardinstallatie vormt dan ook een belangrijk beoordelingspunt. Andere voorname punten zijn de voorziening van een overzichtelijke en navigeerbare gebruikersinterface, afdoende mogelijkheden voor het exporteren of uitwisselen van de resultaten naar andere tools (de garantie op interoperabiliteit), en uiteraard de kwaliteit van de extractie.

De kwaliteit van extractie is meetbaar in termen van *precisie* (het percentage aan geëxtraheerde entiteiten dat correct is), *recall* (het percentage aan entiteiten dat correct werd geëxtraheerd), en een gecombineerde *F-waarde*. De precisie van alle tools op de meest voorname entiteitstypes (personen, organisaties, locaties, enz.) in een kleine, meertalige selectie aan documenten is doorgaans erg hoog (tot 97%). De recall daarentegen is meestal ondermaats (< 50%). Deze resultaten vermoeden het gebruik van weinig omvattende woordenboeken en/of regelsets.

Het leggen **van verbanden tussen entiteiten** is een logische verderzetting van het extraheren van entiteiten, maar overstijgt de capaciteiten van alle geteste documentontsluitingstools. Text mining tools bieden visualisatiemogelijkheden die eenvoudigweg kunnen aangeven welke entiteiten samen voorkomen binnen eenzelfde tekstfragment. Een aantal informatie-extractietools gaan een stap verder met de mogelijkheid tot het manueel opstellen van regelgebaseerde extractiesjablonen (*templates*). Een sjabloon gebruikt de taalkundige en structurele annotaties van de tekst (zoals afgeleid door de tool) om het voorkomen van een interessant feit aan te duiden. Als eenvoudig voorbeeld, het voorkomen van een snelheidsaanduiding en een nummerplaat of wagenmerk in eenzelfde zin kan duiden op de vermelding van een snelheidsovertreding. Meer geavanceerde functionaliteiten, waaronder het automatisch herkennen van aliassen of identiteit, wordt voorlopig niet ondersteund. Dit beïnvloedt de directe bruikbaarheid en de kwaliteit van de resultaten in ongekende

mate negatief. Een kwalitatieve evaluatie werd niet nagestreefd. Dit vereist een afzonderlijke, diepgaander studie.

Voor een toepassingsoverschrijdende beoordeling van de geteste producten op basis van de gepubliceerde resultaten, dient men een nauwgezette afweging te maken met de prioriteiten, behoeften, en vereisten van de uiteindelijke aankoper, beheerder, en eindgebruiker van het product. Ons onderzoek kan door de grote diversiteit van deze afwegingen geen eenduidige beoordeling geven, wel kan het een hulp en referentie zijn bij het maken van een dergelijke beoordeling.

Voor meer details over de resultaten van deze evaulatie verwijzen we naar het Benchmarkingverslag (verslag 5) en bijhorende testresultaten.

7 Randvoorwaarden

In een laatste fase van het project hebben we de randvoorwaarden (basiscondities en aanbevelingen) geformuleerd voor de implementatie van de bestudeerde exploitatieproducten binnen de huidige of toekomstige informatie-architectuur van de eengemaakte politie.

De randvoorwaarden worden ondergebracht onder de hoofdingen productontplooiing, productinpasbaarheid, beveiliging, informatiebeheer, en marktpositie van de leveranciers. Niettegenstaande vloeien deze randvoorwaarden voort uit de opgedane ervaringen met de voorgeselecteerde en geteste producten, en de kennis verkregen uit de studie van de huidige informatie-architectuur binnen de politiediensten.

We bespraken de marktpositie van de producenten/leveranciers van de voorgeselecteerde producten, als antwoord op de randvoorwaarde die stelt dat er voldoende garanties moeten zijn op de stabiliteit van het bedrijf (gunstig toekomstperspectief), en de mogelijkheid tot lokale *on-site* ondersteuning.

De beveiliging van de toegang tot de informatie is fundamenteel bij de implementatie van een exploitatieproduct; het kan geenszins de bedoeling zijn dat de bestaande beveiligingsmechanismes, zoals van kracht in de ANG, worden omzeild door de invoering van een dergelijk product. Beveiliging vormt een basisvoorwaarde, die zowel op centraal als decentraal niveau moet worden voorzien, in gelijke hoedanigheid. We hebben het huidige beveiligingsprotocol geschetst en de rol van het product daarin.

We bespraken de voornaamste aanbevelingen of vereisten met betrekking tot de cliënt- en servercomponent van het product, opdat deze (eenvoudiger) zouden kunnen worden geïntegreerd in de huidige of toekomstige operationele omgeving. In onze studie gingen we uit van twee niveaus van potentiële, operationele omgevingen voor de inpassing van het beoogde exploitatieproduct; centraal en decentraal.

Voorwaarden met betrekking tot het licentieschema, de kostprijs van het product, en de ondersteuningsgaranties van de producent/leverancier, werden buiten beschouwing gelaten. Deze en andere bepalingen vormen immers het onderwerp van het lastenboek dat door de dienst telematica bij openbare aanbesteding zal worden opgesteld.

Voor meer details verwijzen we naar het verslag over de randvoorwaarden (verslag 6).

8 Conclusies

De onderzoeksploeg heeft de doelstellingen van INFO-NS-project behaald. De resultaten kunnen nu gebruikt worden in de selectie en implementatie van een intelligent exploitatietool voor het doorzoeken van de documenten van de Federale Politie en het ontginnen van informatie in deze documenten.

9 Lijst van onderzoeksrapporten

De auteurs van de onderstaande lijst gepubliceerde rapporten betreffen Jan De Beer en prof. Marie-Francine Moens (ICRI/LIIR), Nishant Kumar en prof. Jan Vanthienen (ECON/LIRIS).

1. *INFO-NS Analyseverslag*, uitgegeven op 1 april 2005, 104p.

2. *INFO-NS Evaluatiecriteria*, uitgegeven op 2 mei 2005, 59p.

3. *INFO-NS Benchmarking van de tools, betreffende een voorselectie van de tools*, uitgegeven op 2 mei 2005, 40p.

4. *INFO-NS Testcases*, uitgegeven op 19 juli 2005, 24p.

5. *INFO-NS Benchmarkingverslag*, uitgegeven op 24 oktober 2005, 120p.

6. *INFO-NS Randvoorwaarden*, uitgegeven op 24 oktober 2005, 21p.

7. *INFO-NS Finaal rapport*, uitgegeven op 19 oktober 2005, 25p.

10 Lijst van publicaties

Kumar, N., De Beer, J., Vanthienen, J. & Moens, M.-F. (2005). Evaluation of Intelligent Exploitation Tools for Non-structured Police Information. In *Proceedings of the ICAIL 2005 Workshop on Data Mining, Information Extraction and Evidentiary Reasoning for Law Enforcement and Counter-terrorism.*

Kumar, N., De Beer, J. , Vanthienen, J. & Moens, M.-F. (2005). Evaluation of Intelligent Information Retrieval Tools for Unstructured Police Case Reports. In *Proceedings Conferentie Informatiewetenschap 2005*.