Ika Eegdeman, en Hanneke Smulders

Topaas

Information retrieval in parlementaire context

Het informatiedomein van de Staten-Generaal is niet afgebakend, maar divers van aard en alle beleidsterreinen van alle ministeries omvattend. Een zo veelomvattend domein stelt bijzondere eisen aan de functionaliteit van het vrije-zoek-systeem. Zo is bijvoorbeeld het kunnen opvangen van verschillende betekenissen van hetzelfde begrip in verschillende contexten van groot belang. In het nieuwe informatiesysteem TOPAAS moet deze optie gerealiseerd worden.

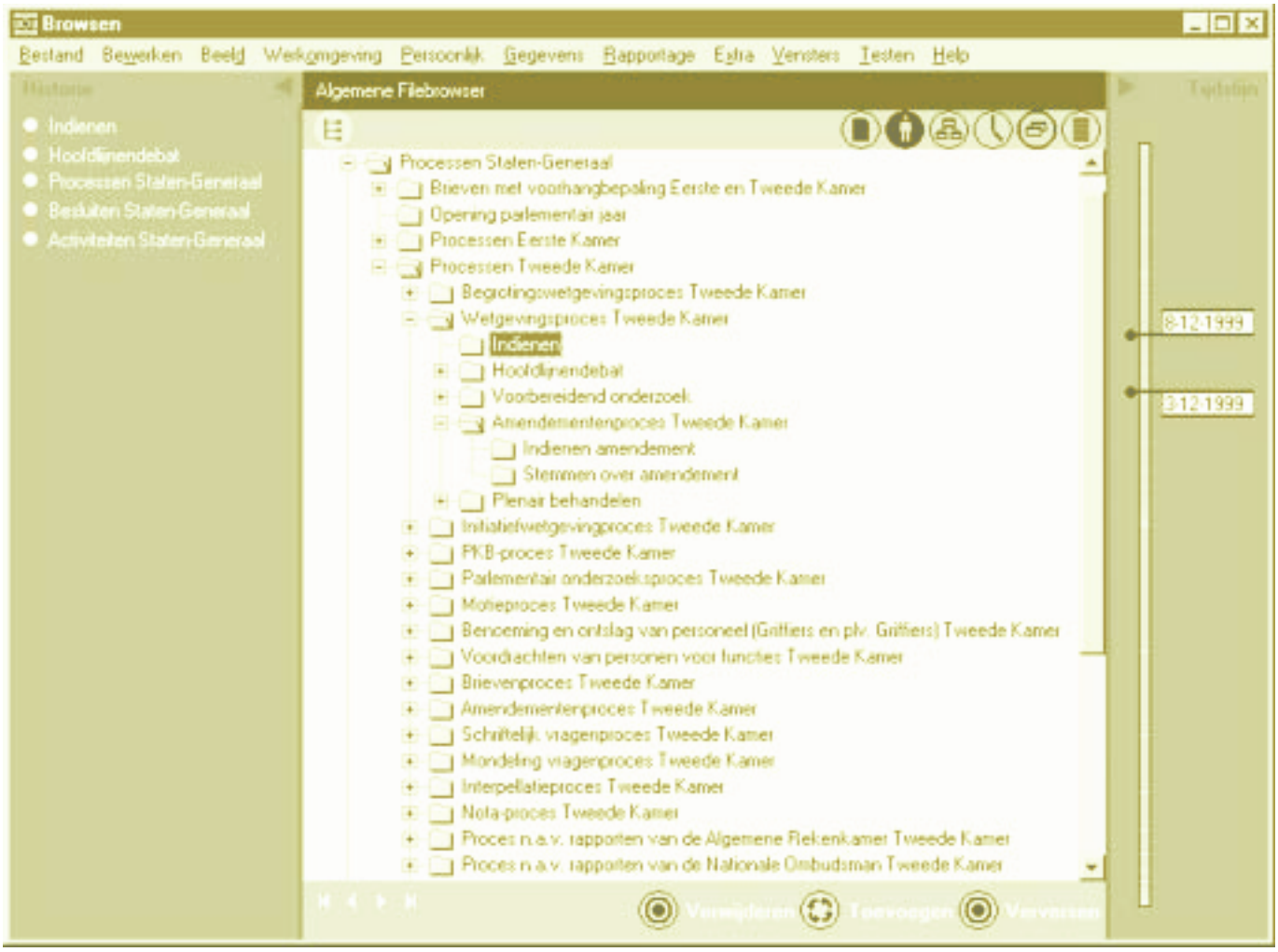
De ontsluitingsmethoden en -technieken van de huidige informatiesystemen binnen de Staten-Generaal omvatten behalve formele ontsluiting ook trefwoorden uit de Parlementsthesaurus en diverse soorten abstracts. In TOPAAS (zie kader op p. 24) zullen deze technieken worden aangevuld met:
- integrale full-text indexering van parlementaire documenten vanaf 1995;
- een in het systeem geïntegreerde thesaurus;
- technieken die het voor de eindgebruiker gemakkelijker maken om de meest relevante documenten terug te vinden;
- informatiedossiers;
- actieve persoonlijke en groepsinteresseprofielen.

Het ligt in de verwachting dat deze nieuwe gereedschappen in de toekomst verder zullen worden uitgebreid met onder meer automatisch gegenereerde samenvattingen, autoclasseren, multilinguïstisch zoeken en grafische vormen van ontsluiting.

Context van het onderwerp

In maatwerkwerkomgevingen zullen informatiespecialisten en eindgebruikers uitgebreide mogelijkheden krijgen om gestructureerd te zoeken. Het is echter noodzakelijk dat ook associatief en vrije-tekst-zoeken, in combinatie met gestructureerd zoeken, op gedegen wijze in het nieuwe informatiesysteem wordt ondersteund. Ook de Staten-Generaal heeft de eis geformuleerd dat het nieuwe retrievalsysteem concept-zoeken moest ondersteunen.
Een voorbeeld van een informatievraag vanuit de context van de Staten-Generaal illustreert deze eis: 'Ik zoek alle moties van Bolkestein over de hervorming van de Europese Unie'. In deze vraag zijn een aantal metadata te herkennen, namelijk: *Bolkestein* als *indiener* en *motie* als *soort document*. *Hervorming Europese Unie* is in deze vraag al wat lastiger te herkennen. Moeten we dit onderwerp terugvinden door het gebruik van één of meer thesaurustermen? Moeten we de termen maar gewoon inbrengen als zoekgegevens en hopen dat ze voorkomen in de documenten?
In het genoemde voorbeeld zijn we als eindgebruiker het meest gebaat bij een vrije-tekst-zoeksysteem dat de zoekvraag als volgt aanpakt:
- alle moties van Bolkestein selecteert en
- vervolgens deze moties doorzoekt op het voorkomen van de morfologische varianten van hervorming Europese Unie (Europese, Europees, European, Euro) en
- ook op het *concept* 'hervorming van de Europese Unie' laat zoeken.

Op die manier selecteert het zoeksysteem documenten waarin termen als Agenda 2000, Fischlerplan, uitbreiding Oost-Europa, landbouwhervorming, eigen bijdrage EU en financiering Europese Unie voorkomen, zonder dat deze termen ook daadwerkelijk in de zoekvraag zijn meegegeven. Hierbij is ook van belang dat de vraagexpansie plaatsvindt op samengestelde begrippen en niet alleen op de delen daarvan.
Voor de vrije-tekst-zoekfunctionaliteit van TOPAAS is zeer uitgebreid onderzoek gedaan naar de complexe zoekprogramma's die meegedaan hebben aan de TREC-onderzoeken. Enkele van de door ons opgedane ervaringen willen wij hier kort weergeven. De onderzochte producten realiseren op verschillende manieren dat na het intikken van een zoekvraag naar meer gezocht en gekeken wordt dan alleen de letterlijk ingetikte letters. Het ene zoekprogramma leunt hiervoor volledig op een handgemaakte thesaurus, het andere maakt gebruik van handgemaakte semantische netwerken en weer andere leggen tijdens het indexeren in een eigen kennisnetwerk of in vectoren in een eigen index verbanden tussen woorden vast.
In de laatste twee gevallen vindt het concept-zoeken met



name plaats op basis van de eigen collectie, met andere woorden: op basis van de terminologie die ontleend wordt aan het geïndexeerde materiaal. Dit in tegenstelling tot de eerstgenoemde systemen waarbij de extra kennis in het systeem wordt verkregen door het systeem van tevoren te verrijken. Hierbij kan worden gebruikgemaakt van thesauri, woordenlijsten, semantische netwerken, enzovoort.
Dat het informatiedomein van de Staten-Generaal zeer vele vakgebieden omvat, is niet bevorderlijk voor een goede precisie. Ons is gebleken dat het handmatig toevoegen van domeinkennis aan het semantisch netwerk wel degelijk een goede precisie mogelijk maakt.

Het zoeken op begrippen met een uitvoerige thesaurus is zeer bevorderlijk voor een acceptabele recall en precision. Als het zoeksysteem de thesaurus of het semantische netwerk zelf uitbreidt op basis van de inhoud van het informatiesysteem, dan levert een verzameling titelbeschrijvingen niet altijd bevredigende resultaten op. In het algemeen speelt de omvang van zowel de dataverzameling als de thesaurus een grote rol.
Indien het uitbreiden van de thesaurus of het semantische netwerk louter mensenwerk is, lijkt het ons onwerkbaar voor een universeel domein. Het aankopen van algemene Nederlandse concepten combineren met het toevoegen van eigen domeinkennis aan de thesaurus lijkt ons wel een goede weg ter verbetering van recall en precisie. Een voorbeeld van in te brengen kennis is de beide betekenissen van het fenomeen *koppeling,* namelijk die van koppeling van lonen aan uitkeringen én de koppeling van gegevensbestanden in het kader van het vreemdelingenbeleid. Het zoveel mogelijk gescheiden houden van kennis en data voorkomt dat na iedere aanvulling van kennis het herindexeren van data nodig is.

Context van de gebruiker

Gebruikersvriendelijkheid en gebruikersondersteuning tijdens een zoeksessie staan hoog op het 'verlanglijstje'. Het ligt immers in de verwachting dat, zodra Kamerleden vanaf hun eigen werkplek of vanuit huis kunnen gaan zoeken in TOPAAS, zij dat ook steeds meer zelf zullen gaan doen. Ook niet-vakmatig betrokken burgers die op de hoogte willen zijn van hetgeen plaatsvindt in het Nederlandse parlement moeten op eenvoudige wijze gewenste informatie tijdig, volledig, geïntegreerd en consistent ter beschikking kunnen hebben. Hierbij mag het parlementaire vakjargon geen belemmering vormen.
Wij willen er dan ook van af dat de gebruiker de juiste zoektermen móét gebruiken. Dat geldt voor zoeksystemen waar je verplicht bent te zoeken met de gecontroleerde vocabulaires waarmee de documenten handmatig zijn ontsloten. Maar vooral ook voor zoeksystemen die zich louter baseren op woordfrequentie. In het eerste geval kan een uitgebreid systeem van verwijzingen misschien nog uitkomst bieden, in het tweede geval ontbreekt dergelijke ondersteuning. Nog afgezien van alle mogelijke relevante

documenten die je zou missen omdat jouw zoekwoorden er niet letterlijk in voor komen.
Wij denken niet dat, in die gevallen dat de gebruiker niet exact weet wat hij zoekt, hij gebaat is bij een louter Booleaanse recall. Niet alleen documenten waarin de zoekwoorden in de onbedoelde betekenis voorkomen, maar ook documenten waarin de zoekwoorden in de bedoelde betekenis voorkomen kunnen als ruis ervaren worden. Een vraag stel je meestal met een zekere bedoeling. Het zoeksysteem zou die bedoeling en uit het zoekgedrag af te leiden bijzondere interesses moeten kunnen verwerken in de relevantie rangorde, bijvoorbeeld via de inzet van een bijzonder gebruikersprofiel.

Feedback

Vrije-tekst-zoeksystemen leveren al snel veel ruis op. Ons inziens is het niet erg dat er, zeker na de eerste vraagstelling, ruis is. Het zoeksysteem dient echter wel iets dat echt bruikbaar is te presenteren bij de eerste dertig titels. Daar kan de vraagsteller dan mee verder. Dit moet zo verwerkt worden dat er steeds meer relevante items bij de eerste dertig komen. Intelligente relevance ranking is dus onontbeerlijk. De ordening dient dan ook gebaseerd te zijn op veel meer criteria dan alleen de frequentie, denk aan de woordvolgorde in de zoekvraag of het aantal verschillende zoekwoorden dat in een document wordt gevonden. Het frequentiecriterium mag zeker niet de doorslag geven.

Ontevredenheid over het eerste zoekresultaat is bij vrije-tekst-zoeken trouwens niet echt ter zake, aangezien het vinden van een bruikbaar zoekresultaat in grote hoeveelheden volledige teksten doorgaans pas kan lukken in een zoeksessie van meer dan één vraag-en-antwoord. Wij zien het zoekproces als een iteratief proces, waarbij het van belang is dat de gebruiker wordt gestimuleerd en ondersteund in het 'definiëren' van zijn informatiebehoefte. Het streven is om het interactieproces, dat in de traditionele situatie plaatsvindt tussen eindgebruiker en informatiespecialist, na te bootsen.

Belangrijke middelen om tijdens het interactieproces te komen tot verbetering van recall en precision zijn naar ons idee de feedback vanuit het semantische netwerk of de thesaurus én de feedback door de gebruiker.
Feedback van de gebruiker kan *actief,* door aan te geven dat je vondst 26 heel bruikbaar vindt of dat het derde zoekwoord echt moet voorkomen in de tekst, maar ook *passief,* indien de zoekmachine informatie over de context van de gebruiker gebruikt.

De gebruiker moet weliswaar handig zijn in het zoeken, actief hulp bieden kan toch geen kwaad, denken wij. De zoekmachine kan bijvoorbeeld, het liefst in dialoogvorm, met hulp van het semantisch netwerk of de thesaurus actief feedback geven op hetgeen gevonden is.
Een 'conceptual grouping' tool zou de verschillende deelaspecten of betekenissen van de gestelde vraag moeten kunnen presenteren en daarmee de gebruiker helpen te komen tot een betere precisie. Door bij de gebruiker 'terug te komen' met de mogelijke betekenissen c.q. contexten van een zoekterm kan de precisie aanmerkelijk worden verhoogd. Tevens helpt de zoekmachine de gebruiker met een dergelijke dialoog om op een relatief eenvoudige wijze te vinden wat hij nodig heeft, ook al begon die gebruiker met het stellen van een voor het betrokken domein veel te ruime en ambigue vraag.

TOPAAS

TOPAAS is de voorgestelde naam voor het nieuwe informatiesysteem van de Staten-Generaal, Kabinet der Koningin en het ministerie van Algemene Zaken (Rijksvoorlichtingsdienst). Doelstelling van het project is het realiseren van een toekomstvaste infrastructuur voor de informatievoorziening ter ondersteuning van het primaire proces: (mede)wetgeving en controle van de regering.
Het systeem behelst enerzijds een vervanging en integratie van de huidige documentaire informatiesystemen Stairs en Globit Dis. Stairs is een retrievalsysteem van IBM waarin de bekende Parac-bestanden met bibliografische beschrijvingen van diverse parlementaire en andere typen documenten zijn opgenomen. Globit Dis is het huidige postregistratiesysteem. In TOPAAS zullen documenten in verschillende formaten elektronisch (Word, pdf of html) ter beschikking worden gesteld. Dit maakt het vrije-tekst-zoeken mogelijk, alsmede het leggen van directe links tussen documenten.
Anderzijds biedt het systeem een uitbreiding van functionaliteit via de component procesbeheer- en ontsluiting, die de parlementaire procesgang beheert en toegankelijk maakt. Procesinformatie wordt geïntegreerd ontsloten met de bovengenoemde documentverzamelingen en de bijbehorende metagegevens. Zo biedt TOPAAS ook de functionaliteit voor het plannen van vergaderingen en het opstellen van plenaire of commissieagenda's. Door de genoemde integrale ontsluiting kunnen bijvoorbeeld in een agenda directe links worden opgenomen naar de bijbehorende full-text documenten.

TOPAAS zal de mogelijkheid bieden om op een eenvoudige manier inzicht te verkrijgen in besluitvorming en de voortgang van bijvoorbeeld een wetgevingsproces. De daarbij betrokken personen & organisaties, geplande vergaderingen en documenten zullen, ook voor extern geïnteresseerden, snel via de zogenaamde Filebrowser (vergelijkbaar met de Windows-Verkenner) opvraagbaar zijn.
Het nieuwe informatiesysteem zal bestaan uit zeer gestructureerde gegevens, zoals de samenstelling van parlement, senaat en commissies, adresgegevens, vergaderdata, soorten bijeenkomsten, besluitvormingstappen, de stappen van diverse parlementaire processen en formele titelbeschrijvingen van brieven, boeken en parlementaire documenten. Daarnaast omvat TOPAAS een omvangrijke hoeveelheid ongestructureerde informatie, met name in abstracts, dossiertoelichtingen en in de volledige teksten van parlementaire documenten. Momenteel is de bouw van het nieuwe informatiesysteem in volle gang. De verwachting is dat het in 2000 operationeel zal zijn.

Voorbeeld: de gebruiker vraagt naar 'lekken'. De zoekmachine vraagt vervolgens: bedoelt u leeuw.krant (67 procent) of rijksrecherche (17 procent) of uitlekken (17 procent)? De percentages geven aan hoe de deelaspecten verdeeld zijn over het zoekresultaat.
Het als zodanig herkennen van een ambigue begrip en het uit de zoekvraag afleiden van de gewenste context blijken niet zonder meer beschikbaar bij de zoeksystemen die nu op de markt zijn. Via technische omwegen, maar vooral ook met feedback van de gebruiker, is hier toch iets aan te doen.

Een feedback-mogelijkheid die voor TOPAAS zeer bruikbaar zou zijn, is die waarbij de gebruiker vrij zoekt naar een persoonsnaam, bijvoorbeeld Bolkestein. De zoekmachine zou dan terug moeten komen met de tegenvraag: in welke hoedanigheid wilt u documenten met betrekking op Bolkestein terugvinden: als indiener, als auteur, als geïnterviewde?
Ook voor deze vorm van gebruikersondersteuning bij het vrije-tekst-zoeken kan een thesaurus, samen met de nodige attributen, ons inziens een belangrijke rol spelen. Net als bijvoorbeeld een periodethesaurus het mogelijk kan maken om te zoeken naar 'bodemvervuiling in de jaren '80' of naar 'de houding van D'66, PvdA en VVD t.o.v. Schiphol tijdens de drie kabinetten Lubbers'.

Serendipity

Zoals eerder aangegeven zal in TOPAAS het gestructureerd zoeken borg staan voor die situaties waarin de gebruiker weet wat hij zoekt en een compleet antwoord wenst. Een functie van ongestructureerd zoeken die wij ook onderkennen, is de mogelijkheid om per toeval iets te ontdekken en verbanden te leggen waar men zelf niet op gekomen was: 'serendipity'.
Een vraag als 'Ik wil alles over vluchtelingen van het afgelopen jaar', zou als resultaat een enorme hoeveelheid documenten kunnen opleveren, waar een gebruiker niet direct iets mee kan. Deze vorm van zoeken zou de zoekmachine moeten ondersteunen door het suggereren van inhoudelijk gerelateerde concepten, al dan niet via een visualisatietool (denk aan de *Aquabrowser* van Medialab of *Thinkmap* van Plumb Design). Deze suggesties zullen de eindgebruiker op ideeën moeten brengen die zijn zoektocht vergemakkelijken of de precisie zullen verhogen. Uiteraard vereist het een geavanceerd semantisch netwerk om de gebruiker niet alleen de diverse invalshoeken van het fenomeen vluchtelingen te kunnen tonen, bijvoorbeeld: sociale voorzieningen, koppeling van gegevensbestanden, overheidsuitgaven, IND, Europese wetgeving, vreemdelingenwet en vluchtelingenverdrag, maar ook van élk ander fenomeen.

Wij zijn van mening dat, in een tijdperk waarin de verzamelingen een zodanige omvang krijgen dat een handmatig ontsluitingssysteem niet meer haalbaar is, we ons toch zullen moeten richten op automatische ontsluitingsmogelijkheden. De complexe analyse- en probabilistische software die actief kan zijn tijdens zoeksessies, kan ook actief zijn tijdens het indexeren in het kader van autoclassificatie.*

Tot slot: het ideale text retrieval-systeem bestaat helaas niet. Dat bevestigde professor Keith van Rijsbergen ook tijdens de IP-Lezing op 7 oktober 1999, toen hij meldde dat binnen de meeste text-retrieval-systemen maximaal veertig procent van alle relevante artikelen daadwerkelijk wordt gevonden. Onze algemene indruk is wel dat het combineren van zoveel mogelijk verschillende indexeer- en retrievaltechnieken het gehele retrievalsysteem krachtiger maakt.

Noot

* Zie ook Spitters, M. en J. van Gent, 'Adjust: automatische thesauriële ontsluiting van grote hoeveelheden krantenartikelen'. In: *Informatie Professional,* nr. 10 1999, p. 29-31.

Dit artikel is mede geschreven naar aanleiding van het artikel 'Vrije tekst en gecontroleerde vocabulaires' van Gerhard Riesthuis in Informatie Professional nr. 10/1999. Zie ingezonden brief van de auteurs in Informatie Professional nr. 11/1999.

I. Eegdeman is medewerker bij Projectbureau Integratie Bestanden, Tweede Kamer (I.Eegdeman@tk.parlement.nl).
J.A.M. Smulders is consultant bij Infomare (infomare@wxs.nl).