Een Theorie voor het Bestuderen van Information Retrieval Modellen

T.W.C. Huibers, B. van Linder en P.D. Bruza

UU-CS-1994-37
September 1994

Utrecht University

Department of Computer Science

Padualaan 14, P.O. Box 80.089,
3508 TB Utrecht, The Netherlands,
Tel. : ... + 31 - 30 - 531454

ISSN: 0924-3275

Een Theorie voor het Bestuderen van Information Retrieval Modellen

T.W.C. Huibers[*†] *B. van Linder*[†] *P.D. Bruza*[§]

Samenvatting

In dit artikel wordt een theoretisch raamwerk voor het bestuderen van information retrieval (IR) modellen gepresenteerd. Deze studie richt zich met name op de wijze waarop modellen besluiten dat een informatie item *omtrent* een ander informatie item is. Het raamwerk vindt zijn oorsprong in de *Situation Theory*. Zogenaamde *infons* en *profons* stellen elementaire informatie-dragers voor. Deze kunnen bewerkt worden door middel van fusie operatoren. Middels deze operatoren kunnen relaties tussen informatie-dragers worden vastgelegd. Een verzameling infons vormt een zogenaamde *situatie* waarmee informatie voorkomend in objecten, zoals documenten, gemodelleerd kan worden. Een willekeurig information retrieval model kan afgebeeld worden in dit raamwerk. Afhankelijk van het soort model zijn hiervoor speciale functies gedefiniëerd. Binnen het theoretisch raamwerk definiëren wij een verzameling postulaten, die gebruikt kunnen worden om de omtrentheid relaties geassocieerd met information retrieval modellen, te beschrijven. Aan de hand van deze postulaten zijn wij in staat kwalitatieve uitspraken te doen over de verschillende omtrentheid-relaties die door de verschillende information retrieval modellen geïnduceerd worden. Ook is het mogelijk kwalitatieve uitspraken te doen over kwantitatieve grootheden als recall en precision. Aan de hand van het boolse retrieval model tonen wij de toepasbaarheid van ons theoretische raamwerk in de praktijk van de information retrieval.

1 Is een information retrieval theorie nodig?

Het terugvinden van opgeslagen relevante informatie wordt door de explosieve groei van het informatie-aanbod steeds moeilijker. Informatie-opslag wordt immers steeds goedkoper en daardoor omvangrijker. Een groeiende hoeveelheid (dure) informatie

*Correspondentie te richten aan deze auteur. Email: theo@cs.ruu.nl.
†Vakgroep Informatica, Universiteit Utrecht, Postbus 80.089, 3508 TB Utrecht.
§School of Information Systems, Queensland University of Technology, Brisbane, Australië.

verdwijnt hierdoor ongebruikt (of ongelezen) om de simpele reden dat er geen middel voorhanden is om deze informatie effectief terug te vinden. Deze problematiek wordt ook wel het *"information retrieval vraagstuk"* genoemd en kan als volgt gekarakteriseerd worden:

> "Op welke wijze kan men relevante informatie onderscheiden van niet relevante informatie met betrekking tot een zekere informatie-behoefte"

Systemen die dit vraagstuk automatisch proberen op te lossen worden information retrieval (IR) systemen genoemd. Dergelijke systemen worden ontwikkeld aan de hand van een bepaald model. Zo'n model probeert een oplossing aan te dragen voor het vraagstuk. Vele onderzoeksgebieden hielden zich sinds de jaren vijftig bezig met dit probleem. Nadat van Rijsbergen in 1989 in "Towards an Information Logic" ([Rij89]) een logische benadering voorstelde, grepen ook logici de kans om voor het information retrieval vraagstuk modellen te ontwikkelen. Op dit moment bestaan er vele modellen waarvan de probabilistische, vector space en logische modellen de meest bekende zijn.

Nieuwe modellen zijn meestal gebaseerd op de drie bovenstaande "originele" klassen van modellen, zo zijn het Hidden Markov model ([MS94]) en Index Expression Belief Networks ([BG94]) extensies van de Probabilistische variant, en zijn Imaging ([CR94]), Situation Theory ([LR92, Rij93]) en Terminologic Logic ([Oun94, Seb94]) invullingen van de logische mogelijkheid. Alhoewel verscheidene auteurs aangeven dat de optimalisatie in bepaalde varianten -meestal die van anderen- marginaal is en significante verbeteringen niet meer mogelijk zijn, kunnen zij geen formeel bewijs geven om hun veronderstelling te onderbouwen.

Over deze problematiek gaat dit artikel. Welk model is beter dan een ander model? Waarom zijn bepaalde IR modellen niet meer significant te verbeteren? Wat zijn de theorieën achter de retrieval mechanisme? Deze problematiek hebben wij eerder behandeld ([HB94], [BH94]). Dit artikel kan gezien worden als een uitbreiding op het formeel systeem geïntroduceerd in [HB94].

Gerelateerd aan deze problematiek is dat binnen de information retrieval bepaalde veronderstellingen heersen die tot op heden niet formeel bewezen zijn. Een van deze veronderstellingen zullen we in dit artikel met onze theorie nader bestuderen. Wij hebben deze veronderstelling het *vermoeden van IR-Utopia* genoemd:

> **Vermoeden (IR-Utopia)** Als men het onderliggende informatiedomein van een information retrieval systeem uitbreidt, levert dit systeem altijd een beter retrieval resultaat op.

Deze veronderstelling is gebaseerd op de intuïtieve gedachte dat hoe *meer informatie* er aanwezig is, des te *beter* zijn de conclusies die getrokken worden uit deze

informatie. Als er systemen zijn waarvoor dit vermoeden waar is dan hoeven we de explosieve groei van het informatie-aanbod niet langer meer als een information retrieval probleem te beschouwen.

Tot op heden werden information retrieval systemen in het bijzonder vergeleken aan de hand van hun *recall* en *precision* waarden. Dit zijn statistische waarden die aangeven hoe doortastend en accuraat een bepaald IR systeem is (zie 3.1). Wij onderkennen zeker het grote nut van deze statistische waarden. Echter om strikte uitspraken te doen omtrent eigenschappen van het ene model ten opzichte van een ander model, zouden we toch over meer formele vergelijkingsmiddelen moeten kunnen beschikken. Ook voor het bewijzen van stellingen over het gedrag van IR systemen zijn statistische testen niet afdoende, en lijkt er behoefte te zijn aan een meer logische karakterisering van deze systemen.

Wij stellen een algemeen raamwerk voor waarin de theoretische fundamenten van verscheidene IR systemen uitgedrukt kunnen worden. In dit logisch raamwerk formuleren wij een aantal postulaten, die min of meer wenselijke eigenschappen van IR modellen uitdrukken. Aan de hand van de postulaten waaraan een IR model voldoet kunnen deze systemen vergeleken worden, en wordt een theoretische bestudering mogelijk gemaakt. Daarnaast blijkt het mogelijk te zijn bepaalde formele uitspraken te doen over de recall en precision waarden die voor een bepaald IR model gelden.

2 Situation Theory in de context van information retrieval

Information retrieval houdt zich bezig met de vraag *"op welke wijze is het mogelijk om juist die documenten te vinden die zo relevant mogelijk zijn met betrekking tot een zekere informatiebehoefte"*. In 1971 introduceerde Cooper een objectieve notie van relevantie genaamd *logisch relevant* ([Coo71]. Deze notie plaatst het begrip relevantie in een logische context, en ontrekt het aan subjectieve interpretaties. Bij logische relevantie gaat het erom of men van een informatie-item via een of andere logische afleiding een ander informatie-item kan afleiden. Als dit het geval is, dan is het eerste item relevant met betrekking tot het tweede.

Deze aanpak nemen we als uitgangspunt. Om echter het beladen woord relevantie te vermijden gebruiken we de term *omtrentheid* (om aan te duiden dat een informatie-item omtrent een ander informatie-item is). Ons doel is deze omtrentheid-relatie te bestuderen in een algemene informatie-theorie. Dergelijke informatie-theorieën zijn onder andere te vinden in de Situation Theory ([BE87, BE90, Dev91]). Wij zijn ervan overtuigd dat Situation Theory een voldoende krachtige informatie-theorie is voor het bestuderen van de omtrentheid relatie tussen documenten of tussen documenten en een vraagstelling. Lalmas & Van Rijsbergen hebben recen-

telijk de voordelen van het gebruik van deze Situation Theory in IR behandeld ([LR92, LR93, Rij93]). Onze aanpak verschilt in die van hen in dat wij Situation Theory niet als onderliggend model gebruiken om IR te bedrijven, maar als een middel om op theoretische wijze de eigenschappen van reeds gedefiniëerde IR systemen te bestuderen.

2.1 Infons: de atomaire informatiedragers

Informatie speelt in information retrieval een centrale rol. Het eerste punt van aandacht in onze theorie is op welke wijze informatie bruikbaar en doorzichtig gemodelleerd kan worden. Een aanknopingspunt kan gevonden worden in het relationele indexering-werk van Farradane ([Far80a, Far80b]). In Farradane's onderzoek wordt informatie gemodelleerd met behulp van een afgebakende verzameling van relatie-typen tussen een onderliggende verzameling van termen. Dit concept is nauw verbonden met de fundamentele informatiedragers van Situation Theory, de *infons* ([BE87, BE90, Dev91]):

Definitie 2.1 (Infons) Een infon is een structuur $\langle\langle \mathrm{R}, a_1, \ldots, a_n; i\rangle\rangle$ die de informatie representeert dat de relatie R geldt (als $i = 1$) of niet geldt (als $i = 0$) tussen de objecten $a_1, \ldots, a_n$.

De waarde i wordt de polariteit van de infon genoemd. Zoals de naam reeds suggereert wordt deze waarde gebruikt om positieve ($i = 1$) of negatieve informatie ($i = 0$) te modelleren.

In relationele indexering zijn de objecten termen, die al dan niet gerelateerd zijn aan elkaar. Een concreet voorbeeld van een infon dat bruikbaar zou zijn voor IR is het volgende, dat gebaseerd is op de relationele indexering aanpak. Stel je voor dat een document alleen maar bestaat uit de tekst "Een schrijver schreef een boek". De informatie die door dit document is vertegenwoordigd kan gemodelleerd worden door de infon: $\langle\langle \mathrm{F}, \texttt{schrijver}, \texttt{boek}; 1\rangle\rangle$. De letter F beschrijft een functionele relatie tussen de termen `schrijver` en `boek`. Dergelijke relaties kunnen gebruikt worden om contextuele informatie af te leiden voor IR. Om redenen van praktische aard is het echter nog niet mogelijk om dit idee te gebruiken in de huidige generatie IR systemen. In huidige systemen wordt de informatie in de documenten voor een gedeelte gemodelleerd met een verzameling van termen, *keywoorden*, genaamd. Deze worden op automatische wijze uit het document gedistilleerd. Dit brengt met zich mee dat de relaties die tussen de keywoorden gelden, niet langer meer voorhanden zijn (zoals de relatie `schrijver,boek` in het bovenstaand voorbeeld). In essentie proberen IR systemen af te leiden welke relaties *er hadden kunnen zijn* om te bepalen of een document omtrent een bepaalde informatiebehoefte is.

Bezien van uit het oogpunt van de Situation Theory resulteert het gebruik van keywoorden om informatie te modelleren in primitieve infons. Deze keywoord-achtige

infons kunnen gezien worden als niet verder te splitsen "sub-informatie". Een analogie in de chemie is die van protonen ten opzichte van atomen. Daarom zullen wij deze specifieke infons *profons* noemen. Profons vormen een sub-klasse van infons gebaseerd op een relatie I. Deze relatie I drukt een ongespecificeerde unaire relatie uit, die staat voor het feit dat het indexproces alle kennis van relaties, waar dat keywoord onderdeel van was, verwijderd heeft. Als zodanig kan Ibeschouwd worden als een existentiële kwantor. Als een document d geïndexeerd wordt met het keywoord `schrijver`, kan men dit zien als het feit dat de informatie gedragen door het begrip `schrijver` aanwezig is in het document d. De expressie $\langle\langle$I,`schrijver`; 1$\rangle\rangle$ zal gebruikt worden om de corresponderende profon te beschrijven.

Het is interessant om op te merken dat het huidige IR onderzoek zich richt op positieve infons, voor wat betreft de indexering. Bij weten van de auteurs bestaan er geen indexsystemen die 'negatieve' keywoorden, dat wil zeggen, profons met polariteit nul, opleveren.

Tot op dit moment hebben we infons en profons laten zien als een mechanisme om keywoorden te beschrijven. De vraag blijft op welke wijze we documenten kunnen modelleren. Zoals eerder gezegd zijn documenten informatiedragers en in deze hoedanigheid kunnen zij gemodelleerd worden als een verzameling infons. Deze verzameling wordt een *situatie* genoemd, in analogie met de definitie van abstracte situaties zoals door Devlin ([Dev91]) gebruikt in de Situation Theory, en is een abstracte representatie van de informatie die aanwezig is in het document.

2.2 Informatie fusie

Een van de prettige eigenschappen van informatie is dat men het kan bewerken. Zo kunnen bijvoorbeeld twee stukjes informatie aan elkaar *'geplakt'* worden om zodoende een nieuw stuk informatie te krijgen. Op het niveau van de infons kan deze samenstelling bereikt worden door speciale operatoren.

Neem bijvoorbeeld de keywoorden `water` en `vervuiling`. Deze kunnen samengesmolten worden om de term `water vervuiling` te krijgen. Dit is een voorbeeld van *informatie fusie.* Het is opvallend dat de informatie in de term `water vervuiling` precies de combinatie vormt van de informatie aanwezig in `water` en `vervuiling`. Fusie wordt op het niveau van infons gemodelleerd door middel van de operator $\odot$: $\langle\langle$I,`water`; 1$\rangle\rangle\odot\langle\langle$I,`vervuiling`; 1$\rangle\rangle$. Deze operator representeert verschillende mogelijke bindingen: zo zijn er in Farradane's indexeringssysteem negen verschillende relaties elk met hun eigen karakteristieken aanwezig, in boolse retrieval zijn er drie (de **or**, de **and** en de **not**).

Op welke manier worden informatie-items in IR samengesteld? Wij onderscheiden hier drie verschillende klassen van bindingen: de sterke binding met verschillende relaties, de zwakke binding en de keuze mogelijkheid.

Allereerst beschouwen we de *sterke binding*. Deze wordt gebruikt om bijvoorbeeld de term `water vervuiling` uit de keywoorden `water` en `vervuiling` te krijgen. Voor een sterke bindingsrelatie R wordt de gecombineerde infon $\langle\langle$I,`water`; 1$\rangle\rangle$ $\odot_R$ $\langle\langle$I,`vervuiling`; 1$\rangle\rangle$ geschreven als $\langle\langle$R,$\langle\langle$I,`water`; 1$\rangle\rangle$, $\langle\langle$I,`vervuiling`; 1$\rangle\rangle$; 1$\rangle\rangle$. Bij voorkeur zullen we dit infon weergeven als $\langle\langle$R,`water`,`vervuiling`; 1$\rangle\rangle$. Hierbij is de relatie R een element uit een gegeven verzameling relaties. Zo is het mogelijk dat R een element is van een verzameling connectieven rechtstreeks verkregen uit de karakterisering van het informatie-item zoals voorgesteld in [Bru93]. In zo'n geval kan de informatie `vervuiling in rivieren` vertaald worden naar de infon $\langle\langle$In,`vervuiling`,`rivieren`; 1$\rangle\rangle$. Dergelijke relaties kunnen een goed middel zijn om context-gevoelige informatie te modelleren. Het is ook mogelijk om twee infons van verschillende typen die met elkaar in relatie staan, samen te voegen. Bijvoorbeeld het samenvoegen van infon $\langle\langle$R,`water`,`vervuiling`; 1$\rangle\rangle$ met de infon $\langle\langle$I,`rivier`; 1$\rangle\rangle$ met een relatie R' resulteert in $\langle\langle$R',$\langle\langle$R,`water`,`vervuiling`; 1$\rangle\rangle$, `rivier`; 1$\rangle\rangle$.

De *zwakke binding* hebben we nodig om aan te geven dat twee infons samen een informatie-item karakteriseren zonder dat er een relatie tussen beiden bestaat. Deze operatie zullen we met $\odot_\wedge$ aangeven. Stel dat een document zowel over `water` als over `vuur` gaat zonder dat er direct verband tussen deze twee keywoorden bestaat. Dan levert een zwakke binding $\langle\langle$I,`water`; 1$\rangle\rangle$ $\odot_\wedge$ $\langle\langle$I,`vuur`; 1$\rangle\rangle$ de situatie {$\langle\langle$I,`water`; 1$\rangle\rangle$, $\langle\langle$I,`vuur`; 1$\rangle\rangle$} op. Deze situatie beschrijft nu dat er informatie over zowel `water` als `vuur` bestaat, maar dat deze profons niet noodzakelijkerwijze aan elkaar gerelateerd zijn.

Tenslotte hebben we nog de *keuze mogelijkheid*. Deze keuze mogelijkheid wordt in de IR eigenlijk alleen maar gebruikt om de informatie-behoefte en niet om documenten te karakteriseren. Net zoals bij de negatieve infons zijn er weinig IR modellen die gebruik maken van een keuze mogelijkheid. Een mate van onzekerheid wordt, wederom uit praktische overwegingen, niet gebruikt bij het indexeren. Echter bijvoorbeeld in het geval van ambiguïteit zou een keuze mogelijkheid wenselijk kunnen zijn. Beschouw bijvoorbeeld een document bestaande uit de tekst "`Alexander heeft een mooi slot gekocht`". Een geavanceerd IR model met thesaurus zou dergelijke informatie misschien willen modelleren met de infon $\langle\langle$Kopen,`Alexander`,`Fietsslot`; 1$\rangle\rangle$ *of* $\langle\langle$Kopen,`Alexander`,`Kasteel`; 1$\rangle\rangle$. Het opsporen en kunnen weergeven van dergelijk ambiguïteit zorgt in het algemeen voor een betere retrieval (dit is aangetoond in [San94], echter met behulp van recall en precision waarden en niet op formele wijze). Wij zullen deze operatie aangeven met behulp van een $\odot_\vee$.

Beginnend met profons (de oorspronkelijke keywoorden) kunnen we situaties opbouwen die exact die informatie modelleren die in het document aanwezig is. Hiervoor moeten alle infons die informatie bezitten die in het document voorkomt op de juiste wijze samengevoegd worden.

Nu we ieder informatie-item kunnen modelleren komen we terug op het informatie

retrieval omtrentheid: wanneer is een situatie omtrent een andere situatie? Deze omtrentheid formaliseren we in de volgende paragraaf.

2.3 De omtrentheid relatie tussen situaties

In de literatuur kan men verschillende omschrijvingen vinden van hetgeen wij de omtrentheid relatie noemen: *"topically related"* ([Coo71]), *"about"* ([Mar77]), *"likely to contain information about"* ([Rij92]), en *"correspondent to"* ([Nie92]). Formeel representeren wij de omtrentheid relatie door middel van het symbool $\mapsto$: intuïtief betekent $S \mapsto T$ dat situatie S omtrent situatie T is, en $S \not\mapsto T$ betekent dat S niet omtrent T is.

Conform de werkelijkheid is het vaak niet meteen duidelijk of een situatie S omtrent een andere situatie T is. Deze omtrentheid moet worden afgeleid. Dergelijke min of meer logische afleidingen spelen een belangrijke rol in zowel de IR als in de Situation Theory ([BE90]). Zo kan bijvoorbeeld van de infon $\langle\langle$R,`water`,`vervuiling`; 1$\rangle\rangle$ intuitïef de infon $\langle\langle$I,`vervuiling`; 1$\rangle\rangle$ afgeleid worden. Dit soort afleidingen wordt gebruikt om strikte afleidingmechanismen van IR modellen te modelleren. Vele takken van Situation Theory beperken hun aandacht tot strikte afleidingen. Om echter met de onzekerheden die een rol spelen in de IR om te kunnen gaan, zijn we genoodzaakt plausibele afleidingen te onderzoeken. Een voorbeeld van een plausibele afleiding is die waarin de infon $\langle\langle$R,`lucht`,`vervuiling`; 1$\rangle\rangle$ plausibel wordt afgeleid uit de infon $\langle\langle$I,`vervuiling`; 1$\rangle\rangle$.

In het algemeen kan een strikte omtrentheid relatie beschreven worden door middel van een effectief berekenbare verzameling van axioma's en regels. Dit is bijvoorbeeld het geval voor afleidbaarheid in de klassieke propositie logica, maar ook voor afleidbaarheid in verscheidene modale logica's ([HC84]).

Definitie 2.2 (Afleidingssysteem) Een afleidingssysteem $\mathcal{A}$ is een paar van de vorm (Ax, Rule), met Ax een verzameling axioma's en Rule een verzameling regels van de vorm $R(S_1, .., S_k, S_{k+1})$. Hier zijn $S_1, \ldots, S_k$ de premissen van de regel, en S_{k+1} de conclusie.

Opmerking 1 Merk op dat we geen uitspraak doen over *hoe* precies de afleidingsrelatie bepaald wordt door het afleidingssysteem. Het is bijvoorbeeld mogelijk dat dit analoog aan de klassiek logische afleiding gebeurt: een situatie is afleidbaar uit een andere situatie als er een rijtje van "tussenliggende" situaties bestaat, die ofwel een axioma zijn ofwel uit vorige situaties ontstaan zijn middels toepassing van een regel. Men kan echter ook aan andere afleidingsrelatie denken. Zo is het bijvoorbeeld mogelijk een default theorie (cf. [Rei80]) als een afleidingssysteem te zien. In dat geval wordt de afleidingsrelatie gedefinïeerd door het al dan niet element van een extensie zijn; deze afleidingsrelatie is in het algemeen niet op de wijze van

de klassiek logische afleidingsrelatie te schrijven. Dit geldt in het algemeen voor niet-monotone afleidingsrelaties.

In de volgende sectie zullen wij door middel van postulaten een aantal axioma's en regels beschrijven. De mate waarin de omtrentheid relatie geassocïeerd met een bepaald IR systeem aan deze postulaten voldoet, geeft een kwalitatieve indicatie van de aard van dit systeem. Deze laatste eigenschap geldt in het bijzonder met betrekking tot recall en precision.

3 Postulaten voor information retrieval

De doelstelling van een IR systeem is, gegeven een informatie-behoefte, enerzijds zoveel mogelijk relevante documenten, en anderzijds zo weinig mogelijk irrelevante documenten op te leveren. Beschouw een verzameling documenten $\mathcal{D}$. Neem aan dat er van ieder document een document-karakterisering $\chi(d)$ bestaat. De karakterisatering van een document is een zo goed mogelijke benadering van de inhoud van een document. Het information retrieval proces begint met een vraagstelling van de gebruiker. De vraagstelling bestaat uit een aantal keywoorden die komen uit de verzameling $\chi(\mathcal{D})$. De keywoorden in de vraagstelling vormen de karakterisering van de informatie-behoefte van de gebruiker.

Documenten bevatten informatie en kunnen als zodanig gemodelleerd worden als een set infons, of met andere woorden, als een situatie. Stel dat een situatie S_d correspondeert met het document d. Zoals bij documenten is er bij deze situatie een andere situatie $S_{\chi(d)}$ die overeenkomt met de karakterisering $\chi(d)$. In werkelijkheid is $S_{\chi(d)}$ een ruwe benadering van S_d.

Een vraagstelling kan gezien worden als een verzoek om bepaalde informatie en kan daarom ook gerepresenteerd worden als een situatie. Ofwel alle aspecten die een rol spelen in information retrieval kunnen afgebeeld worden in Situation Theory. Dit geldt zowel voor documenten en karakteriseringen van documenten, evenzo voor de vraagstellingen. De vraag blijft nu, op welke wijze we in IR *het beter zijn dan andere modellen* kunnen plaatsen in het raamwerk gebaseerd op Situation Theory.

In dit raamwerk is *omtrentheid* een relatie tussen situaties. Daardoor wordt omtrentheid gezien als een fundamentele relatie met betrekking tot informatie. Dit verschilt van andere aanpakken die informatie omtrentheid uitdrukken in termen van informatie die bevat is in andere informatie.

In deze paragraaf wordt een aantal postulaten gepresenteerd als een afleidingssysteem. Het doel van deze postulaten is een vergelijking tussen verschillende modellen mogelijk te maken, aan de hand van het al dan niet voldoen aan bepaalde postulaten. Als zodanig kunnen deze postulaten vergeleken worden met de postulaten voor

niet-monotoon redeneren zoals gepresenteerd in [KLM90], en die voor belief revision zoals gedefiniëerd in [Gär88].

Om het definiëren van bepaalde postulaten, in het bijzonder het consistentie- en CWA postulaat, mogelijk te maken, is het noodzakelijk de beschikking te hebben over het *complement* van een situatie. Intuïtief kan het complement van een samengesteld informatie-item gezien worden als dat informatie-item dat precies de complementaire informatie bevat.

Definitie 3.1 (Het complement van situaties) Het complement van situaties S, genoteerd als $\overline{S}$, wordt inductief gedefiniëerd naar de opbouw van infons.

- $\overline{\langle\langle \mathrm{I},\ldots;\ \mathrm{i}\rangle\rangle} = \langle\langle \mathrm{I},\ldots;\ \text{1-i}\rangle\rangle$
- $\overline{\langle\langle \mathrm{R},\ldots;\ \mathrm{i}\rangle\rangle} = \langle\langle \mathrm{R},\ldots;\ \text{1-i}\rangle\rangle$ voor sterke bindingen R.
- $\overline{I_1 \odot_\wedge I_2} = \overline{I_1} \odot_\vee \overline{I_2}$ waarbij I_1 en I_2 willekeurige infons zijn.
- $\overline{I_1 \odot_\vee I_2} = \overline{I_1} \odot_\wedge \overline{I_2}$ waarbij I_1 en I_2 willekeurige infons zijn.

Beschouwen we de situatie $\{\langle\langle \mathrm{I},\texttt{water};\ 1\rangle\rangle, \langle\langle \mathrm{I},\texttt{vuur};\ 0\rangle\rangle \odot_\vee \langle\langle \mathrm{I},\texttt{vlammen};\ 0\rangle\rangle\}$. De bijbehorende infon is $\langle\langle \mathrm{I},\texttt{water};\ 1\rangle\rangle \odot_\wedge (\langle\langle \mathrm{I},\texttt{vuur};\ 0\rangle\rangle \odot_\vee \langle\langle \mathrm{I},\texttt{vlammen};\ 0\rangle\rangle)$. Het complement hiervan is de infon $\langle\langle \mathrm{I},\texttt{water};\ 0\rangle\rangle \odot_\vee (\langle\langle \mathrm{I},\texttt{vuur};\ 1\rangle\rangle \odot_\wedge \langle\langle \mathrm{I},\texttt{vlammen};\ 1\rangle\rangle)$.

In feite zijn de volgende postulaten *postulaat-schema's*: in deze postulaten wordt de $\odot$ gebruikt als een parameter voor de verzameling operatoren $\odot_r$ met $r \in \mathrm{R}$ of als de zwakke binding $\odot_\wedge$ of als mogelijke keuze-binding $\odot_\vee$. Dus indien bijvoorbeeld de verzameling R uit één element bestaat, dan representeert ieder postulaat-schema drie verschillende postulaten.

Mogelijke axioma's

Het axioma van Reflexiviteit:

Reflexiviteit
$S \vdash\!\!\!\mapsto S$

Dit axioma stelt dat iedere situatie omtrent zichzelf is. Dit is waarschijnlijk het axioma dat het meest algemeen geaccepteerd is: bij ons weten bestaan er geen met modellen van IR systemen geassociëerde omtrentheid-relaties die niet reflexief zijn.

De *extensie*-axioma's zeggen iets over de aard van de binding. Is deze binding "sterk" (in het geval van *Linkse Extensie*) of is deze "zwak" (in het geval van *Rechtse Extensie*) met betrekking tot de omtrentheid relatie. Met andere woorden, deze axioma's doen een uitspraak over het resultaat van het combineren van infons: is de gecombineerde infon met betrekking tot de omtrentheid-relatie sterker of zwakker dan de afzonderlijke infons.

Extensie

links	*rechts*
$S \odot T \mapsto S$	$S \mapsto S \odot T$

Het axioma van *Commutativiteit* wordt op de gebruikelijke wijze gedefiniëerd. Dit axioma geeft aan dat situaties met betrekking tot hun binding verwisselbaar zijn. Merk op dat dit postulaat per definitie geldt voor de $\odot_V$ binding als een direct gevolg van de eigenschappen van verzamelingen.

Commutativiteit

$$S \odot T \mapsto T \odot S$$

Basis afleidingsregels

De regel van *Symmetrie* wordt ook op de gebruikelijke wijze gedefiniëerd. Deze regel stelt dat er geen enkel verschil bestaat tussen concluderen dat situatie S omtrent situatie T is en concluderen dat situatie T omtrent situatie S is.

Symmetrie

$$\frac{S \mapsto T}{T \mapsto S}$$

In enkele retrieval systemen, bijvoorbeeld het boolse retrieval systeem, wordt symmetrie uitgesloten door striktheid van de omtrentheid-relatie. Symmetrie komt bijvoorbeeld wel voor in systemen waarbij de matching functie gebaseerd is op een overlappingswaarde, zoals vector space retrieval en coordination level matching.

Het postulaat van *Transitiviteit* stelt dat indien een situatie S omtrent een situatie T is, en situatie T tegelijkertijd omtrent een situatie U is, ook situatie S omtrent U is.

Transitiviteit

$$\frac{S \mapsto T \qquad T \mapsto U}{S \mapsto U}$$

In tegenstelling tot de regel van Symmetrie komt Transitiviteit juist niet voor in modellen waar sprake is van een overlap matching functie.

Consistentie van situaties, in de Situation Theory ook wel coherentie genaamd ([Dev91]), stelt dat het niet mogelijk is dat een situatie S zowel omtrent een situatie T als omtrent het complement $\overline{T}$ is. Hoewel consistentie een zeer gebruikelijke eis is in modellen die een uitbreiding van de klassieke propositie logica vormen, is het niet direct duidelijk of het voldoen aan dit postulaat een wenselijke eigenschap is.

Consistentie

$$\frac{S \vdash\!\!\mapsto T}{S \not\vdash\!\!\mapsto \overline{T}}$$

De *Closed World Assumption (CWA)* is een bekend postulaat uit de theorie van database-systemen en die van niet-monotone redeneersystemen ([Rei78]). Intuïtief stelt de CWA dat alleen positieve feiten opgeslagen worden: ontbreken van een positief feit wordt als een indicatie van het waar zijn van de negatie van het feit beschouwt. In onze theorie stelt het CWA postulaat dat indien een situatie S niet omtrent een situatie T is, S omtrent het complement van T is.

CWA

$$\frac{S \not\vdash\!\!\mapsto T}{S \vdash\!\!\mapsto \overline{T}}$$

De regels van *Linkse en Rechtse Equivalentie* drukken uit dat indien twee situaties omtrent elkaar zijn, en dus min of meer "equivalent" genoemd mogen worden, zich "equivalent" gedragen onder de omtrentheid-relaties. Dat wil zeggen dat voor iedere situatie U zodanig dat S omtrent U is, T ook omtrent U is, en ieder situatie U die omtrent S is, is ook omtrent T.

Equivalentie

links

$$\frac{S \vdash\!\!\mapsto T \qquad T \vdash\!\!\mapsto S \qquad U \vdash\!\!\mapsto S}{U \vdash\!\!\mapsto T}$$

rechts

$$\frac{S \vdash\!\!\mapsto T \qquad T \vdash\!\!\mapsto S \qquad S \vdash\!\!\mapsto U}{T \vdash\!\!\mapsto U}$$

De regels van Equivalentie geven in feite de *extensionaliteit* van een omtrentheid-relatie aan. Het voldoen aan dit "extensionaliteits" postulaat wordt niet noodzakelijk wenselijk geacht voor IR modellen die met een zo "intensioneel" concept als informatie werken.

Fusie afleidingsregels

In vele modellen vragen we ons af hoe de omtrentheid-eigenschap bewaard blijft onder fusie. Is het zo dat de omtrentheid-relatie zich monotoon gedraagt onder het samenstellen van infons? De postulaten die een antwoord op deze vraag formuleren vallen in twee categorieën uiteen. Bij de postulaten van *Zwakke Monotone Fusie* leggen we eisen op aan beide situaties die we combineren, terwijl we bij de postulaten van *Sterke Monotone Fusie* geen eis opleggen aan de "toegevoegde" situatie.

Monotone Fusie

Zwakke				Sterke	
links		*rechts*		*links*	*rechts*
$S \mid\!\!\mapsto U \quad T \mid\!\!\mapsto U$		$S \mid\!\!\mapsto T \quad S \mid\!\!\mapsto U$		$S \mid\!\!\mapsto T$	$S \mid\!\!\mapsto T$
$S \odot T \mid\!\!\mapsto U$		$S \mid\!\!\mapsto T \odot U$		$S \odot U \mid\!\!\mapsto T$	$S \mid\!\!\mapsto T \odot U$

Merk op dat de regels van Sterke Monotone Fusie echt sterker zijn dan die van Zwakke Monotone Fusie: het voldoen aan de regels van de eerste groep impliceert automatisch dat er ook voldaan wordt aan de regels van de tweede groep. Het is echter mogelijk dat bepaalde omtrentheid-relaties wel aan de regels van Zwakke Monotone Fusie, maar niet aan die van Sterke Monotone Fusie voldoen.

De volgende vier postulaten zijn de negatieve versies van Sterke en Zwakke Monotone Fusie. Wat kunnen we concluderen als we eenmaal weten dat een situatie S niet omtrent een situatie U is? Geldt dat dan ook voor samenstelling van situaties waarin S voorkomt? De antwoorden op deze vraag worden uitgedrukt middels de regels van *Zwakke en Sterke Negatieve Fusie.*

Negatieve Fusie

Zwakke				Sterke	
links		*rechts*		*links*	*rechts*
$S \not\mid\!\!\mapsto U \quad T \not\mid\!\!\mapsto U$		$S \not\mid\!\!\mapsto T \quad S \not\mid\!\!\mapsto U$		$S \not\mid\!\!\mapsto U$	$S \not\mid\!\!\mapsto T$
$S \odot T \not\mid\!\!\mapsto U$		$S \not\mid\!\!\mapsto T \odot U$		$S \odot T \not\mid\!\!\mapsto U$	$S \not\mid\!\!\mapsto T \odot U$

Het laatste postulaat dat wij hier definiëren is dat van *Wederzijdse Fusie.* Dit postulaat is een soort combinatie van voorgaande fusie regels.

Wederzijdse Fusie

$$\frac{S \mid\!\!\mapsto T \quad U \mid\!\!\mapsto V}{S \odot U \mid\!\!\mapsto T \odot V}$$

De hierboven gedefiniëerde postulaten worden door ons gebruikt om tot beschrijvingen van omtrentheid-relaties te komen. Merk op dat wij geen uitspraken doen over compleetheid dan wel minimaliteit van deze verzameling postulaten. Het is zowel mogelijk dat bepaalde postulaten overbodig zijn, als dat er eigenschappen van een omtrentheid-relatie zijn die door geen van deze postulaten afdoende beschreven worden.

3.1 Een kwalitatieve beschouwing van recall en precision

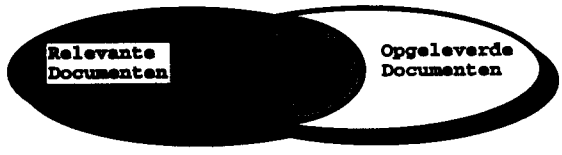
Zoals reeds in de introductie is verteld, is het thema van dit artikel tot een kwalitatief vergelijk van de met IR modellen geassocieerde omtrentheid-relatie te komen.

Een interessante mogelijkheid die geboden wordt door het in de voorgaande secties geïntroduceerde formeel model behelst een kwalitatief onderzoek van recall en precision. Hier zullen we met name bekijken wat voor een bepaald IR model de relatie is tussen het voldoen aan bepaalde postulaten, en de recall en precision waarden die voor dat model gelden.

Zowel recall als precision zijn gedefiniëerd als een relatie tussen de documenten die relevant zijn met betrekking tot een bepaalde vraagstelling, en de documenten die het IR systeem oplevert voor diezelfde vraagstelling.

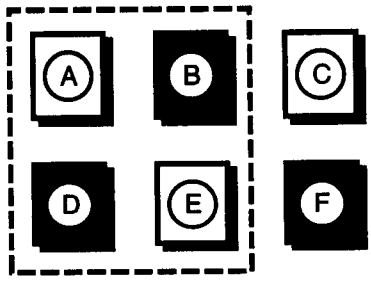
$$\text{Recall} = \frac{|\ \text{Relevant en Opgeleverd}\ |}{|\ \text{Relevant}\ |}$$

$$\text{Precision} = \frac{|\ \text{Relevant en Opgeleverd}\ |}{|\ \text{Opgeleverd}\ |}$$



Merk op dat beide waarden zich tussen 0 en 1 bevinden. Merk ook op dat er meestal een bepaalde trade-off bestaat tussen beide waarden: men zou naïef kunnen trachten de recall waarde van een systeem te verhogen door het aantal opgeleverde documenten te verhogen, maar in het algemeen leidt dit dan weer tot een verlaging van de precision waarde.

Voorbeeld 3.1 (Recall en precision)
De figuur aan de rechterzijde geeft een retrieval situatie weer, de grijze blokjes stellen relevante informatie-items met betrekking tot een zekere informatie behoefte voor, de witte blokjes stellen irrelevante informatie-items voor. De blokjes in het grote gestippelde vierkant werden door een bepaald systeem als relevant beschouwd (in termen van het raamwerk zijn dit de situaties die omtrent S_q gaan). De recall waarde is hier $\frac{2}{3}$ en de precision waarde is $\frac{2}{4}$.



Als we de omtrentheid relatie van een bepaald IR model (geheel) gekarakteriseerd hebben door middel van een (deelverzameling) van het afleidingssysteem gedefiniëerd in de vorige sectie, kunnen we trachten uitspraken te bewijzen als "toevoeging c.q. weglating van deze regel zou recall/precision positief/negatief beinvloeden". Om dit soort uitspraken formeel te maken, introduceren we wat extra terminologie.

Definitie 3.2 (Monotoniciteit) Zij $\mathcal{A} = (\text{Ax}, \text{Rule})$ een afleidingssysteem.

- $\mid\!\!\sim_{\mathcal{A}}$ is *monotoon* dan en slechts dan als voor alle situaties S en T geldt:

$$S \mid\!\!\sim_{\mathcal{A}} T \Rightarrow \forall S'(S \subseteq S' \Rightarrow S' \mid\!\!\sim_{\mathcal{A}} T)$$

- $\mid\!\!\sim_{\mathcal{A}}$ is *monotoon in zijn axioma's* dan en slechts dan als voor alle situaties S en T, en voor alle afleidingssystemen $\mathcal{A}' = (\mathrm{Ax}', \mathrm{Rule})$ met $\mathrm{Ax} \subseteq \mathrm{Ax}'$, geldt:

$$S \mid\!\!\sim_{\mathcal{A}} T \Rightarrow S \mid\!\!\sim_{\mathcal{A}'} T$$

- $\mid\!\!\sim_{\mathcal{A}}$ is *monotoon in zijn regels* dan en slechts dan als voor alle situaties S en T, en voor alle afleidingssystemen $\mathcal{A}' = (\mathrm{Ax}, \mathrm{Rule}')$ met $\mathrm{Rule} \subseteq \mathrm{Rule}'$, geldt:

$$S \mid\!\!\sim_{\mathcal{A}} T \Rightarrow S \mid\!\!\sim_{\mathcal{A}'} T$$

Opmerking 2 Merk op dat een afleidingssysteem wiens afleidingsrelatie op dezelfde wijze gedefiniëerd is als de klassiek logische, vanzelf monotoon is, zowel in zijn axioma's als in zijn regels. Dit geldt echter niet voor willekeurige afleidingssystemen. Beschouw bijvoorbeeld een afleidingssysteem $(\emptyset, \mathrm{CWA})$. Op een query $\langle\langle$I,`water`; 1$\rangle\rangle$ zal dit afleidingssysteem vanuit de verzameling $\{\langle\langle$`rivieren`; 1$\rangle\rangle$ niet $\langle\langle$I,`water`; 1$\rangle\rangle$ af kunnen leiden (maar wel $\langle\langle$I,`water`; 0$\rangle\rangle$). Breiden we het afleidingssysteem uit met het axioma $\langle\langle$I,`rivieren`; 1$\rangle\rangle \mid\!\!\sim \langle\langle$I,`water`; 1$\rangle\rangle$, dan zal $\langle\langle$I,`water`; 0$\rangle\rangle$ niet meer afgeleid worden.

Als een eerste aanzet tot een volledige inductieve theorie van IR modellen, geven wij hier drie stellingen die uitspraken doen over de consequenties op de recall waarde die verandering van het IR model tot gevolg kan hebben.

Stelling 3.1 *Als een IR model volledig beschreven wordt door een afleidingssysteem $\mathcal{A}$, en $\mid\!\!\sim_{\mathcal{A}}$ is monotoon, dan zal door het uitbreiden van de karakterisering van de documenten de recall van het model gelijk blijven of hoger worden.*

Merk op dat stelling 3.1 overeenkomt met het postulaat van Links Sterke Monotone Fusie. Immers voor situaties S en S' betekent $S \subseteq S'$ niets anders dan dat $S' = S \odot_{\wedge} U$ voor zekere situatie U. Het postulaat van Linkse Sterke Monotone Fusie stelt dan dat uit $S \mid\!\!\sim T$ geconcludeerd kan worden dat $S' = S \odot_{\wedge} U \mid\!\!\sim T$, wat overeenkomt met monotonie van de afleidingsrelatie $\mid\!\!\sim$. Ofwel, iedere omtrentheidrelatie die voldoet aan het postulaat van Linkse Sterke Monotone Fusie is monotoon.

Stelling 3.2 *Als een IR model volledig beschreven wordt door een afleidingssysteem $\mathcal{A} = (\mathrm{Ax}, \mathrm{Rule})$, en $\mid\!\!\sim_{\mathcal{A}}$ is monotoon in zijn regels, dan zal ieder IR model dat volledig beschreven wordt door een afleidingssysteem $\mathcal{A}' = (\mathrm{Ax}, \mathrm{Rule}')$ met* $\mathrm{Rule} \subseteq \mathrm{Rule}'$ *een gelijkblijvende of hogere recall waarde hebben.*

Stelling 3.3 *Als een IR model volledig beschreven wordt door een afleidingssysteem* $\mathcal{A} = (\mathrm{Ax}, \mathrm{Rule})$, *en* $\vdash\!\!\!\rightarrow_{\mathcal{A}}$ *is monotoon in zijn axioma's, dan zal ieder IR model dat volledig beschreven wordt door een afleidingssysteem* $\mathcal{A}' = (\mathrm{Ax}', \mathrm{Rule})$ *met* $\mathrm{Ax} \subseteq \mathrm{Ax}'$ *een gelijkblijvende of hogere recall waarde hebben.*

Het intuïtive idee achter stellingen 3.2 en 3.3 is dat in monotone afleidingssystemen het aantal afgeleide situaties toeneemt door het toevoegen van regels c.q. axioma's. Immers, alle tevoren afgeleide situaties zijn nog steeds afleidbaar, en het is mogelijk dat nieuwe situaties afgeleid kunnen worden middels de toegevoegde regels. Beschouwen we voorbeeld 3.1. Stel dat de omtrentheid relatie die de informatie-items A, B, D en E heeft afgeleid monotoon is in zijn regels (bij monotonie in axioma's kan een geheel analoge redenering gehouden worden). Bij toevoeging van regels aan de omtrentheid relatie zou het zo kunnen zijn dat informatie-item F nu ook opgeleverd zou worden (en mogelijk ook item C maar dat heeft geen invloed op de recall waarde). Omdat de eerder gevonden items nog steeds opgeleverd worden (dit is precies de monotonie van de omtrentheid relatie) zou het vinden van informatie-item F tot gevolg hebben dat de recall waarde stijgt van $\frac{2}{3}$ tot 1.

Een nadere beschouwing van voorbeeld 3.1 maakt ook duidelijk dat we geen strikte uitspraken over precision waarden kunnen doen. Toevoegen van regels en/of axioma's zou voor omtrentheid relaties die monotoon zijn in hun regels/axioma's de precision kunnen vergroten, maar dit is niet noodzakelijk zo. Immers, door het toevoegen van regels zou informatie-item F gevonden kunnen worden, daarmee de precision waarde verhogend tot $\frac{3}{5}$, maar het is ook mogelijk dat alleen item C als extra gevonden zou worden. In dit laatste geval zou de precision waarde terugvallen tot $\frac{2}{5}$.

Met deze stellingen hebben we een eerste gereedschap om kwalitatieve uitspraken te doen over de verschillende recall waarden die verschillende systemen hebben. Beschouwen we bijvoorbeeld het Coordinate Information Fields retrieval model en het Strict Coordinate Information Fields retrieval model ([BH94]). De verzameling regels van het afleidingsysteem dat Coordinate retrieval volledig beschrijft is een uitbreiding van de verzameling regels van het Strict Coordinate retrieval model met de regel van Symmetrie (cf. [BH94]). Omdat in beide systemen de afleiding gedefiniëerd is als in de klassieke propositie-logica, en er dus sprake is van monotoniciteit in de regels, kunnen we met behulp van stelling 3.2 concluderen dat de recall waarde behorende bij een Coordinate retrieval model minstens zo goed is als die behorende bij een Strict Coordinate retrieval model. Merk op dat we deze uitspraak kunnen doen zonder ook maar één test daadwerkelijk uit te voeren.

Stelling 3.1 bevestigt tenminste een gedeelte van het eerder geformuleerde IR-Utopia vermoeden. Beschouwen we immers de omtrentheid-relatie geassocieerd met een model van een bepaald IR systeem. Als deze relatie monotoon is, kunnen we concluderen dat uitbreiding van het onderliggende informatie-domein mogelijk tot een verhoging van de recall waarde leidt, maar in ieder geval geen verlaging van deze

waarde tot gevolg zal hebben. Ofwel, voor bepaalde IR systemen geldt het IR-Utopia vermoeden met betrekking tot de recall waarde.

4 De theorie in de praktijk

4.1 Omtrentheid in boolse retrieval systemen

In een model van boolse retrieval systemen bestaat de karakterisering uit een set van termen die komen uit een set $\mathcal{T}$. Door middel van de functie *sit* worden termen vertaald naar situaties. Stel $t \in \mathcal{T}$, dan $sit(t) = \{\langle\langle \text{I},t;\ 1\rangle\rangle\}$. De situaties horende bij een document karakterisering ($S_{\chi(d)}$), bestaat uit een fusie van $\mid \chi(d) \mid$ infons $S_{\chi(d)} = \{sit(t) \mid t \in \chi(d)\}$.

Voorbeeld 4.1 (Julius Caesar) Bekijk de volgende twee informatie-dragers: Informatie-drager d_1 bestaat uit de informatie dat "*C*aesar *m*ag *B*rutus" en informatie-drager d_2 "*A*ntonius *h*aat *B*rutus". Met de karakteriseringstaal zoals reeds beschreven kunnen de informatie-dragers als volgt vertaald worden:

$$\begin{aligned}
\mathcal{T} &= \{C, M, B, A, H\} \\
\chi(d_1) &= \{C, M, B\} \\
\chi(d_2) &= \{A, H, B\} \\
S_{\chi(d_1)} &= sit(C) \odot_\wedge sit(M) \odot_\wedge sit(B) \\
S_{\chi(d_1)} &= \{\langle\langle \text{I,C};\ 1\rangle\rangle, \langle\langle \text{I,M};\ 1\rangle\rangle, \langle\langle \text{I,B};\ 1\rangle\rangle\} \\
S_{\chi(d_2)} &= \{\langle\langle \text{I,A};\ 1\rangle\rangle, \langle\langle \text{I,H};\ 1\rangle\rangle, \langle\langle \text{I,B};\ 1\rangle\rangle\}
\end{aligned}$$

In boolse retrieval wordt de vraagstelling gespecificeerd door middel van een formule. Deze formules worden geconstrueerd met de set $\mathcal{T}$ en de logische connectieven $\wedge, \vee$ en $\neg$. We hebben reeds gezien dat $\wedge$ afgebeeld kan worden naar $\odot_\wedge$. De logische operator $\vee$ wordt vertaald naar de keuze-fusie ($\odot_\vee$ operator).

Een formule kan een negatie bevatten, om bijvoorbeeld uit te drukken dat gebruiker documenten wil dit niet gaan over een zeker keywoord t. Met betrekking tot deze negatie is het belangrijk op te merken dat boolse retrieval aan de *Closed World Assumption* voldoet ([Rij86]). Op het niveau van profons wordt negatie als volgt gemodelleerd: $sit(\neg t) = \{\langle\langle \text{I},t;\ 0\rangle\rangle\}$.

De vertaling van boolse formules naar situaties, gaat als volgt:

$$\begin{aligned}
sit(t) &= \{\langle\langle \text{I},t;\ 1\rangle\rangle\} \\
sit(\neg t) &= \{\langle\langle \text{I},t;\ 0\rangle\rangle\} \\
sit(\psi_1 \vee \psi_2 \vee \ldots \vee \psi_n) &= sit(\psi_1) \odot_\vee sit(\psi_2) \odot_\vee \ldots \odot_\vee sit(\psi_n) \\
sit(\psi_1 \wedge \psi_2 \wedge \ldots \wedge \psi_n) &= sit(\psi_1) \odot_\wedge sit(\psi_2) \odot_\wedge \ldots \odot_\wedge sit(\psi_n)
\end{aligned}$$

De functie *sit* is een bijectieve functie. Daardoor kan de inverse functie sit^{-1} gedefiniëerd worden die, gegeven een situatie een daarmee corresponderende formule als resultaat opleveren.

Voorbeeld 4.2 (Caesar) Zij een vraagstelling $(C \wedge \neg A)$ gegeven, die de informatiebehoefte *"ik wil alle informatie-items die over Caesar gaan en waarin Antonius niet voorkomt"* voorstelt. Na gebruik gemaakt hebben van de sit-functie, is het resultaat de situatie $\{\langle\langle \text{I,C; 1}\rangle\rangle, \langle\langle \text{I,A; 0}\rangle\rangle\}$.

In boolse retrieval, wordt omtrentheid geformaliseerd via een afleidingssysteem dat bestaat uit dat van de klassieke propositie logica uitgebreid met de CWA regel (zie [Bru93] voor meer details).

Definitie 4.1 (Boolse Situatie Omtrentheid) Het afleidingssysteem behorend bij de omtrentheid-relatie geassocieerd met een boolse retrieval model wordt als volgt gedefiniëerd. De verzameling van axioma's Ax wordt gegeven door Ax = {Reflexiviteit, $\odot_\vee$-Rechtse Extensie, $\odot_\wedge$-Links Extensie, $\odot_\vee$ en $\odot_\wedge$-Commutativiteit}. De verzameling regels Rule wordt gegeven door {CWA, Equivalentie, $\odot_\vee$-Rechtse Sterke Monotone Fusie, $\odot_\wedge$-Rechtse Zwakke Monotone Fusie, $\odot_\vee$-Wederzijdse Fusie}.

Met het *Julius Caesar* voorbeeld, het boolse retrieval model en de gedefiniëerde $\mathcal{B}$, kunnen we bewijzen dat een situatie *omtrent* een andere situatie is.

Voorbeeld 4.3 (Caesar)

$$\begin{aligned} S_q &= \langle\langle \text{I,C; 1}\rangle\rangle \odot_\wedge \langle\langle \text{I,A; 0}\rangle\rangle \\ S_{\chi(d_1)} &= \{\langle\langle \text{I,C; 1}\rangle\rangle, \langle\langle \text{I,M; 1}\rangle\rangle, \langle\langle \text{I,B; 1}\rangle\rangle\} \end{aligned}$$

Als volgt kan men laten zien dat $S_{\chi(d_1)} \vdash\!\!\!\rightarrow S_q$:

$$\begin{array}{ll} & \{\langle\langle \text{I,C; 1}\rangle\rangle, \langle\langle \text{I,M; 1}\rangle\rangle, \langle\langle \text{I,B; 1}\rangle\rangle\} \\ \vdash\!\!\!\rightarrow & \{\langle\langle \text{I,C; 1}\rangle\rangle\} \qquad \odot_\wedge\textit{-Linkse Extensie} \\ \not\vdash\!\!\!\rightarrow & \{\langle\langle \text{I,A; 1}\rangle\rangle\} \qquad \textit{Via inspectie} \\ \vdash\!\!\!\rightarrow & \{\langle\langle \text{I,A; 0}\rangle\rangle\} \qquad \textit{Closed World Assumption} \\ \vdash\!\!\!\rightarrow & \{\langle\langle \text{I,C; 1}\rangle\rangle \odot_\wedge \langle\langle \text{I,A;0}\rangle\rangle\} \qquad \odot_\wedge\textit{-Rechtse Zwakke Monotone Fusie} \end{array}$$

Opmerking 3 De omtrentheid-relatie in het boolse retrieval model (zoals hierboven beschreven) is niet monotoon. Dit is een gevolg van het feit dat CWA tot de regels van het afleidingssysteem behoort dat deze omtrentheid-relatie beschrijft. Immers, het voldoen aan het CWA postulaat is een bekende indicatie van het niet-monotoon zijn van een relatie. Als een gevolg hiervan, geldt stelling 3.1 niet voor

boolse retrieval modellen: het uitbreiden van de karakterisering van de documenten, levert niet noodzakelijk een hogere recall waarde op. Men zou een verlaging van de recall waarde binnen het boolse retrieval model als gevolg van een uitbreiding van de karakterisering van documenten als een teken van de "slechtheid" van de uitbreiding kunnen beschouwen.

5 Conclusies

In dit artikel hebben we een raamwerk gepresenteerd dat gebruikt kan worden voor het bestuderen van information retrieval modellen. Dit raamwerk bestaat uit een zeer algemene en flexibele taal, gebaseerd op concepten uit de Situation Theory, die gebruikt wordt om verschillende IR modellen te formaliseren. In deze taal hebben we een aantal postulaten gedefiniëerd, om daarmee de omtrentheid-relaties die door verschillende IR systemen geïnduceerd worden, te beschrijven. Uitgaande van deze taal en de beschrijving door middel van het voldoen aan postulaten, hebben we drie algemene stellingen geformuleerd die uitspraken doen over de relatie tussen de postulaten die een systeem beschrijven en de recall waarde van het IR systeem.

Toekomstig onderzoek zal zich met name richten op andere algemene stellingen, in het bijzonder stellingen die uitspraken doen over precision. Ook het doen van kwalitatieve uitspraken over het vervangen van regels en axioma's in de beschrijving van een IR systeem ("bij het vervangen van deze regel door die regel verandert de recall op deze wijze") is een onderwerp van verder onderzoek. Verder is het belangrijk eventuele compleetheid en minimaliteit van de hier gegeven verzameling postulaten te bestuderen. Een algemeen punt van onderzoek is tenslotte de expressiviteit van het Situation Theory-achtig raamwerk: is het mogelijk om in dit raamwerk willekeurige informatiesystemen, van hypertext tot en met databases, te modelleren, en in hoeverre kunnen er formele uitspraken gedaan worden over complexe systemen.

Dankwoord

Wij bedanken Thomas Arts en John-Jules Meyer voor hun opbouwende kritiek.

Referenties

[BE87] J. Barwise and J. Etchemendy. *The Liar, An Essay on Truth and Circularity.* Oxford University Press, 1987.

[BE90] J. Barwise and J. Etchemendy. Information, infons, and inference. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications, Volume I*, chapter 2. CSLI, 1990.

[BG94] P.D. Bruza and L.C. van der Gaag. Index expression belief networks for information disclosure. *International Journal of Expert Systems*, 7(2):107–138, 1994.

[BH94] P.D. Bruza and T.W.C. Huibers. Investigating aboutness axioms using information fields. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–121. ACM, Springer-Verlag, July 1994.

[Bru93] P.D. Bruza. *Stratified Information Disclosure, a Synthesis between Hypermedia and Information Retrieval.* PhD thesis, University of Nijmegen, March 1993.

[Coo71] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[CR94] F. Crestani and C.J. Rijsbergen. Information retrieval by imaging. In *Proceedings of the BCS 16th Information Retrieval Colloquium.* British Computer Society, March 1994. (to appear).

[Dev91] K. Devlin. *Logic and Information.* Cambridge University Press, 1991.

[Far80a] J. Farradane. Relational indexing, part I. *Journal of Computer Science 1*, pages 267–276, 1980.

[Far80b] J. Farradane. Relational indexing, part II. *Journal of Computer Science 1*, pages 313–324, 1980.

[Gär88] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* The MIT Press, Cambridge, Massachusetts and London, England, 1988.

[HB94] T.W.C. Huibers and P.D. Bruza. Situations, a general framework for studying information retrieval. In *Proceedings of the BCS 16th Information Retrieval Colloquium.* British Computer Society, March 1994. (to appear).

[HC84] G.E. Hughes and M.J. Cresswell. *A Companion to Modal Logic.* Methuen, London, 1984.

[KLM90] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[LR92] M. Lalmas and C.J. van Rijsbergen. A logical model of information retrieval based on situation theory. In *Proceedings of the BCS 14th Information Retrieval Colloquium*, pages 1–13. British Computer Society, Springer-Verlag, April 1992.

[LR93] M. Lalmas and C.J. van Rijsbergen. A model of an information retrieval system based on situation theory and dempster-shafer theory of evidence. In V.S. Alagar, S. Berger, and F. Dong, editors, *Incompleteness and Uncertainty in Information Systems*, 1993.

[Mar77] M.E. Maron. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, pages 38–43, January 1977.

[MS94] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–327. ACM, Springer-Verlag, July 1994.

[Nie92] J. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. In N. Belkin, P. Ingwersen, and A.M. Pejtersen, editors, *Proceedings of the Fifteenth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–151. ACM, ACM Press, June 1992.

[Oun94] I. Ounis. *Logique Terminologique pour la correspondance entre Graphes de Concepts dans le cadre d'un Système de Recherche d'Informations.* PhD thesis, Université Joseph Fourier, June 1994.

[Rei78] R. Reiter. On closed-world data bases. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 55–76, New York, 1978. Plenum Press.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Rij86] C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, Vol. 29(6):481–485, 1986.

[Rij89] C.J. van Rijsbergen. Towards an information logic. In N.J. Belkin and C.J. van Rijsbergen, editors, *Proceedings of the Twelfth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 77–86. ACM, ACM Press, June 1989.

[Rij92] C.J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, Vol. 35(3):291–298, 1992.

[Rij93] C.J. van Rijsbergen. Two essays in information retrieval. Departmental Research Report IR-93-3, University of Glasgow, November 1993.

[San94] M. Sanderson. Word sense disambiguation and information retrieval. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. ACM, Springer-Verlag, July 1994.

[Seb94] F. Sebastiani. A probabilistic terminological logic for modelling information retrieval. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual Internatiol ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–130. ACM, Springer-Verlag, July 1994.