LISA-D en onbetrouwbare kennis

Een fundamenteel onderzoek naar de mogelijkheid om LISA-D te gebruiken voor expertsystemen en information retrieval systemen

R.J. te Kronnie

Katholieke *Universiteit* Nijmegen

Afstudeerscriptie Nr. 318, 6 juli 1994

Afdeling Informatiesystemen
Faculteit der Wiskunde en Informatica

Informatie
Systemen

LISA-D en onbetrouwbare kennis

Een fundamenteel onderzoek naar de mogelijkheid om LISA-D
te gebruiken voor expertsystemen en information retrieval systemen

Afstudeerscriptie 318,

ter verkrijging van de graad doctorandus
aan de Katholieke Universiteit Nijmegen,
Faculteit der Wiskunde en Informatica,
Afdeling Informatiesystemen

door:

R.J. te Kronnie

geboren te Winterswijk, 6 Maart 1968

begeleider(s):

Dr. Th.P. van der Weide
Drs. H.A. Proper

6 juli 1994

Inhoudsopgave

Hoofdstuk 1

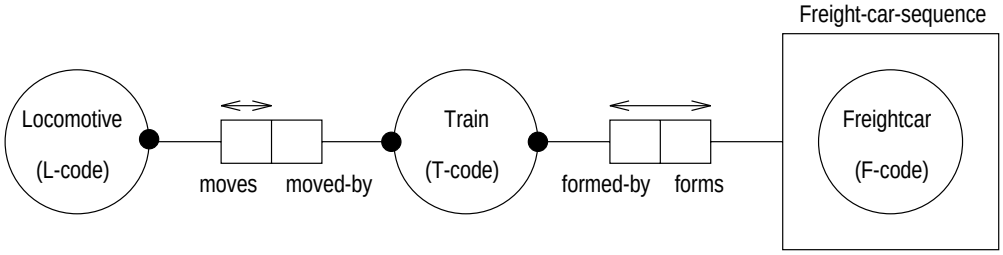
Inleiding

1.1 Wat is LISA-D

LISA-D staat voor **L**anguage for **I**nformation **S**tructure and **A**access **D**escription. Het is een taal die momenteel nog volop in ontwikkeling is. Met deze taal is het mogelijk om gegevens te modelleren. Men kan dan een informatiestructuur ontwikkelen. Via een CASE-tool is zelf te bepalen welk modelleringstechniek er voor de informatiestructuur gebruikt gaat worden. De basis voor het beschrijven en analyseren met LISA-D is de modelleringstechniek PSM. Een beschrijving van PSM en wat daar aan vooraf ging, is te vinden in achtereenvolgens [HW93] en [HPW92]. Als aanvulling daarop kan men [WHB92], [HW92] en [BHW91] lezen.

1.1.1 Analyse van informatiestructuren: padexpressies

Met LISA-D kan men de ontworpen informatiestructuur analyseren. De informatiestructuur levert daartoe het raamwerk. Voor de analyse worden padexpressies gebruikt. Wat deze zijn en hoe ze werken, wordt stapsgewijs aangegeven, te beginnen met een voorbeeld.



Figuur 1.1: Een administratie van de treinsamenstellingen

Voorbeeld 1.1.1

Bekijk de informatiestructuur in figuur 1.1 gepresenteerd als een PSM-schema. Als men over een treinsamenstelling iets weten wil, dan kan men vragen stellen zoals hieronder in de natuurlijke taal respectievelijk in de vorm van een LISA-D query:

Which locomotive moves a train formed by a sequence of freight-cars ?

LIST Locomotive moves Train formed-by Freight-car-sequence

□

Zoals men ziet vertoont de query veel overeenkomsten met de predicatoren en objecttypen in informatiestructuur in figuur 1.1. Een leuke bijkomstigheid is de mogelijkheid van zogenaamde

"query by navigation". Men wandelt gewoon langs objecttypen en feittypen, waarmee dan een query wordt gevormd.

Zulke queries worden hier door LISA-D vertaald in *padexpressies*. In zijn eenvoudigste vorm correspondeert een padexpressie met een pad in de informatiestructuur die begint met een objecttype en ook eindigt met een objecttype. Zij moet uiteraard wel resultaat opleveren. De padexpressies worden geëvalueerd met betrekking tot de huidige populatie van de informatiestructuur. De uitkomsten zijn hier in de vorm van *multisets* van *binaire relaties*. De uitkomsten van er tussen liggende instanties (die ook multisets zijn) worden bij de evaluatie gebruikt, maar worden na het bereiken van het eindresultaat weggegooid.
Er zijn verschillende soorten padexpressies. Beschrijvingen en definities van padexpressies en hoe deze uit queries vertaald worden zijn te vinden in [HPW93].

1.2 LISA-D en onzekerheid

Tot nu toe is LISA-D alleen geschikt om informatiesystemen mee te ontwikkelen. Er zijn echter nog andere systemen, namelijk information retrieval systemen en expertsystemen. De eigenschap van die systemen is dat ze kunnen redeneren met onzekerheid en de kennis die ze hebben. Echter is LISA-D niet geschikt om deze systemen te ontwikkelen, omdat LISA-D geen faciliteiten heeft om te kunnen redeneren met onzekerheid. Daar moet nog een oplossing voor worden gezocht. Eerst wordt kort aangegeven wat het principe van redeneren met onzekerheid in zowel expertsystemen als information retrieval systemen is. Daarna wordt het probleemgebied aangegeven waarvoor naar een oplossing wordt gezocht.

1.2.1 Expertsystemen en waarschijnlijkheid

Er zijn verschillende expertsystemen. Het merendeel ervan komt in de medische wereld voor. Zij hebben (bijna) allemaal twee dingen met elkaar gemeen:

- er wordt met onzekerheid geredeneerd, en
- de kennisdatabases hebben alle ongeveer dezelfde vorm. Dat wil zeggen, de kennis is er in opgeslagen als relaties tussen aanwijzingen en hypothesen in de vorm van regels **if** e **then** h **fi**.

Er wordt steeds met behulp van bovengenoemde regels geredeneerd. Maar meestal is de kennis niet absoluut zeker. Het ontbreekt soms aan gegevens en de kennis kan onvolledig zijn. Aan hypothesen kan een mate van onzekerheid hangen. Dat geldt net zo voor de onderzochte aanwijzingen. Opdat een systeem er mee werken kan worden deze onzekerheden uitgedrukt in zogenaamde "kansen", die dan aan alle hypothesen en aanwijzingen die samen de kennis vormen, worden toegekend. Het systeem kan dan zelf conclusies trekken met de kennis die het heeft. Wat als conclusie wordt beschouwd is bijvoorbeeld die met de hoogste kans. Diezelfde kans vormt het criterium voor het presenteren van die conclusie.

Met behulp van LISA-D kan men zoiets ook doen. Men kan dan vragen stellen, zoals hieronder in de vorm van een LISA-D query:

```
LIST Treatment of Disease as Diagnosis by-observing Evidence
```

Dit is mogelijk, ware het niet dat LISA-D geen faciliteiten heeft om met onzekerheid te kunnen redeneren. Voor het ontwikkelen van die faciliteiten moet men een onderzoek doen naar de modellen voor het redeneren met onzekerheid zoals deze in bestaande expertsystemen worden gebruikt.

1.2.2 Relevanties in Information Retrieval

De huidige trend in information retrieval systemen is systemen gebaseerd op hypertext. Een voorbeeld van zo'n systeem is ODILON, een dia documentatie systeem. Stel men wil dia's lenen.

name	expr	$\mu[\![\mathsf{expr}]\!](\mathsf{Pop}))$
object type	x	$\mathsf{Sqr} \cdot \mathsf{Multi} \cdot \mathsf{Pop}(x)$
predicator	p	$\{\!\{\langle v(p), v\rangle\uparrow^1 \mid v \in \mathsf{Pop} \cdot \mathsf{Fact}(p)\}\!\}$
reverse	$P^{\leftarrow}$	$\{\!\{\langle q, p\rangle\uparrow^n \mid \langle p, q\rangle \in^n \mu[\![P]\!](\mathsf{Pop})\}\!\}$
powerset	$\wp P$	$\mathsf{Sqr} \cdot \{\!\{i\uparrow^1 \mid i \subseteq \pi.1\,\mu[\![P]\!](\mathsf{Pop})\}\!\}$
concatenate	$P \circ Q$	$\bigcup_r \{\!\{\langle p, q\rangle\uparrow^{n\times m} \mid \langle p, r\rangle \in^n \mu[\![P]\!](\mathsf{Pop}) \wedge \langle r, q\rangle \in^m \mu[\![Q]\!](\mathsf{Pop})\}\!\}$
extend	$P \diamond Q$	$\bigcup_{r,s} \{\!\{\langle p, q\rangle\uparrow^{n\times m} \mid \langle p, r\rangle \in^n \mu[\![P]\!](\mathsf{Pop}) \wedge \langle q, s\rangle \in^m \mu[\![Q]\!](\mathsf{Pop})\}\!\}$

Tabel 1.1: Een overzicht van enkele padexpressies

Daarvoor moet men ze eerst even opzoeken. Men weet echter niet waar men ze moet vinden. Dan kan aan het systeem gevraagd worden om nummers van dia's over een specifiek onderwerp op te zoeken. Stel de dia's moeten gaan over Amerikaanse gebouwen. In het systeem zijn de volgende trefwoorden: Amerikaanse kunst, gebouwen in Amerikaanse stijl en Amerikaanse bouwwerken.
Als in het systeem strikte afleiding wordt toegepast, dan levert de vraag niets op. Laat het systeem nu zo zijn, dat deze op basis van *plausibele afleiding* werkt. Als er geen passend trefwoord is, of de vraagstelling is niet goed, dan kan het systeem zelf uitzoeken wat de gebruiker wil en/of bedoelt. Het systeem gaat dan zelf kijken in hoeverre de hem bekende gegevens *relevant* zijn met betrekking tot de vraagstelling. Als nou aan ieder trefwoord (als losse combinatie van woorden) een kans (*relevantie*) wordt toegekend, dan kan het systeem zelf bepalen wat relevant is.
Na enige tijd 'rekenen' kan het systeem alternatieven in de trant van suggesties aanbieden. Welke het systeem aanbiedt, hangt van de mate van relevantie af. In het hierboven gegeven voorbeeld kunnen de suggesties zijn: gebouwen in Amerikaanse stijl, Amerikaanse bouwwerken. De gebruiker bepaalt hierna zelf wat er dan gedaan moet worden.

Met behulp van LISA-D kan men zoiets ook doen. Dan stelt men vragen zoals hieronder in een LISA-D query:

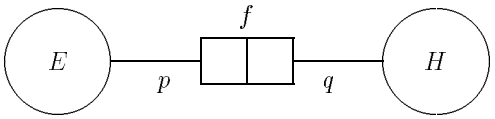
LIST Nummmer van Dia ABOUT Amerikaanse gebouwen

Hier moet echter voor LISA-D uitgezocht worden hoe daarin relevanties toegepast, gebruikt en/of berekend gaat worden. LISA-D heeft geen faciliteiten voor het toepassen van plausibele inferentie.

1.2.3 Het calculatiedomein onder padexpressies

In de inleiding is al gesproken over padexpressies. Een beschrijving daarvan is te vinden in [HPW93]. Voor een behandeling wordt er een aantal van die padexpressies hier in tabel ref-petable1 gegeven. De kreet $\mu[\![\mathsf{expr}]\!](\mathsf{Pop})$ geeft hier van een evaluatie van een padexpressie de semantiek weer. De semantiek bij de objecttype ziet er niet begrijpelijk uit, maar daar staat dat van de populatie van het objecttype x een multiset wordt gemaakt dat op zijn beurt weer wordt veranderd in een binaire relatie van multisets.
Als men de in de tabel gegeven padexpressies bekijkt, is te zien dat voor elke gegenereerde tupel wordt berekend hoe vaak deze in een multiset mag voorkomen. Voor dergelijke berekeningen is een aantal functies in gebruik:

- een functie die een constante toekent;
- een binaire functie voor de vermenigvuldiging;
- een telfunctie die een getal voor het aantal getelde elementen geeft.



Figuur 1.2: De informatiestructuur

Deze functies en eventueel andere zijn in het calculatiedomein onder padexpressies te vinden. Dit geeft het idee dat deze calculatiedomein misschien ook voor ander doeleinden is te gebruiken, zoals kansfuncties. Om dit te verduidelijken, wordt hieronder een voorbeeld gegeven.

Voorbeeld 1.2.1
Zij de informatiestructuur als gegeven in figuur 1.2. Zij

$$\begin{array}{lcl} \mathsf{Pop(E)} & = & \{e_1, e_2\} \\ \mathsf{Pop(H)} & = & \{h_1, h_2\} \\ \mathsf{Pop(f)} & = & \{\{p : e_1, q : h_1\}|_a, \{p : e_2, q : h_2\}|_b\} \end{array}$$

Dan is:

$$\mu\,[\![p \circ f \circ q^{\smile}]\!]\,(\mathsf{Pop}) = \left\{\!\!\left\{ \begin{array}{l} \langle e_1, h_1\rangle|_a \\ \langle e_2, h_2\rangle|_b \end{array} \right\}\!\!\right\}$$

Het gegeven voorbeeld (voorbeeld 1.2.1) geeft a de kans/relevantie weer voor de tupel $\langle e_1, h_1 \rangle$*. Voor b gaat het net zo. Er bestaat tevens de mogelijkheid dat aan de elementen in de populaties van de objecttypen ook kansen toegekend kunnen worden. Dit geeft aan dat wat de calculatiedomein betreft, er voor* LISA-D *meer mogelijkheden zijn.*
*Hier zij nog opgemerkt dat de gebruikte notaties voor het weergeven van die a en b niet officieel zijn. Het is geen standaard notatie in*LISA-D*, maar een dat voor het doel verzonnen is.* □

1.3 Uitbreiding van LISA-D

Het doel is om LISA-D uit te breiden met een faciliteit opdat er met onzekere kennis geredeneerd kan worden. Voor expertsystemen is dat een kansmodel en voor information retrieval een model voor de relevantie. In de voorgaande paragrafen is globaal aangegeven hoe bij information retrieval en in expertsystemen met onzekerheid wordt omgegaan.

Om LISA-D uit te breiden, moeten eerst de mogelijkheden worden onderzocht. Het onderzoek spitst zich toe op de volgende zaken:

- Welke (quasi-)probabilistische modellen zijn er in expertsystemen en hoe werken ze?
- Welke modellen zijn er voor de bepaling van relevanties in information retrieval? Uiteraard wordt ook de werking van die modellen onderzocht.
- Zijn in LISA-D de modellen toe passen zo dat LISA-D gebruikt kan worden voor expertsystemen en information retrieval systemen?

De modellen kunnen vooraf al geselecteerd worden voordat ze aan een uitgebreid onderzoek worden onderworpen. Voor toepassing van een model of theorie in LISA-D worden eisen aan dat model of die theorie gesteld:

1. Het is niet mogelijk tussenresultaten te bewaren. Bij de evaluatie worden als tussenresultaten multisets gevormd, maar na het bereiken van het eindresultaat worden ze meteen weg gegooid.

2. Het eindresultaat dat na evaluatie van een padexpressie wordt bereikt kan ergens (dynamisch) bewaard worden, maar hergebruik voor nieuwe evaluaties is niet mogelijk.

3. Het abstractieniveau zoals deze in LISA-D aanwezig is dient gehandhaafd te blijven. Het doel van dit onderzoek is om functies ten behoeve van het redeneren met onzekerheid of relevanties in het calculatiedomein onder te brengen, maar het is bijvoorbeeld niet de bedoeling om een informatiestructuur of iets dergelijks er bij in te betrekken.

4. De populaties bij de informatiestructuren zijn (enigzins) constant. Het is dus bijvoorbeeld niet mogelijk om resultaten van evaluaties van padexpressies in de populaties te verwerken. Dit onderstreept nog eens de eis dat hergebruik van resultaten niet mogelijk is.

Het onderzoek levert enigzins de volgende resultaten op:

- Een model afkomstig van expertsystemen dat in LISA-D toe te passen is;
- Een model dat wordt gebruikt bij information retrieval en die in LISA-D toe te passen is;
- Eisen en/of voorwaarden die bij toepassingen van de modellen mogelijk aan LISA-D gesteld kunnen worden.

Het is trouwens niet uitgesloten dat er ook wel een model kan zijn dat zowel voor information retrieval systemen als voor expertsystemen geschikt is. Ook is het mogelijk dat er **geen** geschikt model is.

Hoofdstuk 2

Expertsystemen en onzekerheid

Expertsystemen leiden uit de kennis die in de database is opgeslagen vaak informatie af. Echter is deze afleiding niet strikt, dat wil zeggen de afleiding gaat gepaard met een mate van onzekerheid. Dit is het geval bij expertsystemen. Om afleiding te kunnen doen waarbij onzekerheid een rol speelt, zijn afleidingstechnieken toegepast die per systeem verschillen. Deze technieken zullen uitgebreid worden besproken. Maar eerst wordt besproken hoe kennis is opgeslagen en hoe men daaraan een mate van onzekerheid aanhangt. Daarna wordt de waarschijnlijkheidstheorie die de oorsprong is van waaruit de afleidingstechnieken zijn ontwikkeld, behandeld. Tot slot worden de verschillende technieken behandeld dat voldoende informatie moet opleveren met betrekking tot de toepassing van die technieken in LISA-D.

2.1 Productieregels en afleiding

Bij expertsystemen wordt de kennis beschreven in de vorm van *productieregels* van de vorm **if** e **then** h **fi** waarin e een *aanwijzing* is dat elementair is of dat een combinatie is van atomaire condities die onderling met elkaar samenhangen door middel van de operatoren **and** en **or**. In de productieregel kan h bestaan uit samenhangende conclusies of slechts één conclusie. Als h slechts bestaat uit één conclusie, dan wordt h een *hypothese* genoemd. In het vervolg wordt daarvan uitgegaan.

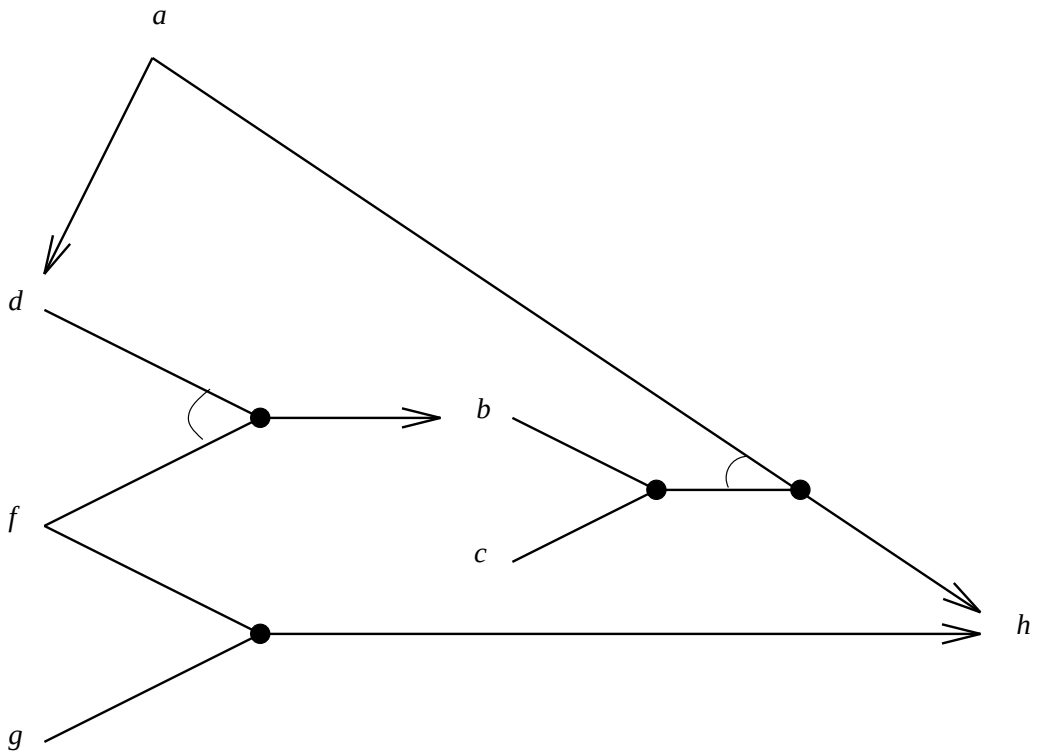
Om te zien hoe productieregels met elkaar samenhangen maakt men gebruik van een zogenaamde *afleidingsnetwerk* . Hieronder wordt een voorbeeld gegeven (uit [LG91]).

Voorbeeld 2.1.1
Stel we hebben de volgende regels:

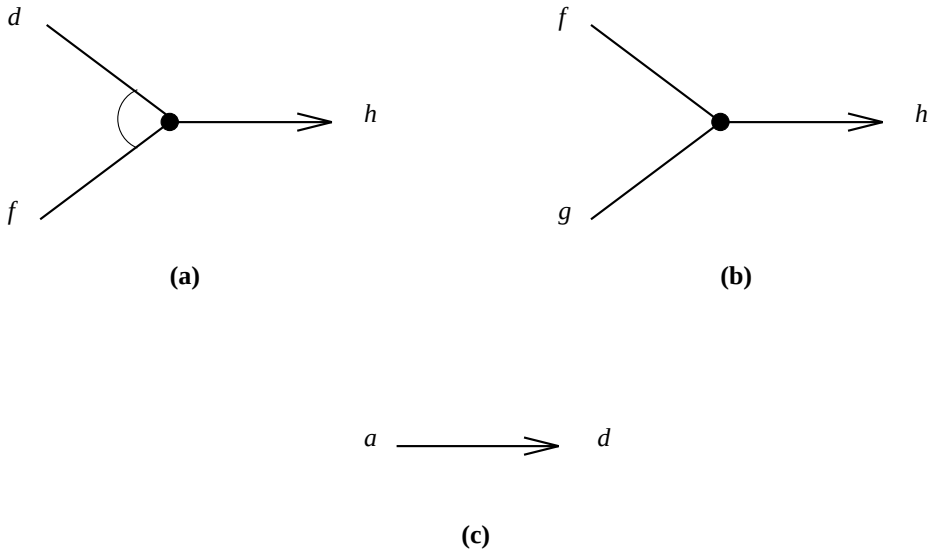
R_1: **if** a **and** (b **or** c) **then** h **fi**
R_2: **if** d **and** f **then** b **fi**
R_3: **if** f **or** g **then** h **fi**
R_4: **if** a **then** d **fi**

De samenhang tussen de regels kan men laten zien in een afleidingsnetwerk waarin gebruik gemaakt is van de "bouwblokken" zoals deze in figuur 2.2 zijn gegeven. Laten we stellen dat we als eindconclusie de hypothese h willen. Dan laat het afleidingsnetwerk in figuur 2.1 zien wat er aan regels in welke volgorde geëvalueerd moet worden, te beginnen bij a. Als er geen regel is waarin een aanwijzing een hypothese is, dan dient de gebruiker zelf iets voor de aanwijzing in te vullen. Zoals de figuur laat zien zijn er meer wegen die leiden tot de conclusie h. □

Tot nu toe is de situatie behandeld waarbij ervan uitgegaan is dat een productieregel **if** e **then** h **fi** als volgt wordt gelezen, namelijk dat als e zich zeker voordoet of beter gezegd absoluut waar is, de hypothese h als waar wordt bevestigd. In de praktijk wordt de hypothese zelden als volledig zeker



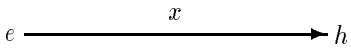
Figuur 2.1: Het afleidingsnetwerk waarin de productieregels zijn uitgebeeld, met als eindconclusie h



Figuur 2.2: De bouwblokken. Plaatje (a) representeert de **and**-operator, en regel R_2. Plaatje (b) representeert de **or**-operator en regel R_3. Plaatje (c) geeft de elementaire productieregel (R_4) weer.

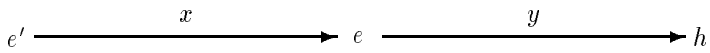
bevestigd. Men hanteert de regels met een mate van onzekerheid, dat meestal het gevolg is van onvolledige gegevens of onbetrouwbaarheid van een stukje kennis.

Om aan te geven dat aan een betreffende regel een mate van onzekerheid valt toe te schrijven, presenteert men een productieregel **if** e **then** h_x **fi** in een afleidingsnetwerk als volgt:

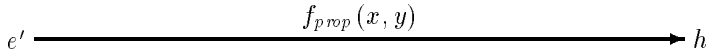


hier is x de graad waarmee h wordt bevestigd door de aanwijzing e (die absoluut waar is). Dit is nog maar het begin. Er zijn gevallen waarin onzekerheid bij de aanwijzing e een rol speelt en er is een geval met betrekking tot de hypothese h. Deze gevallen worden hieronder puntsgewijs behandeld.

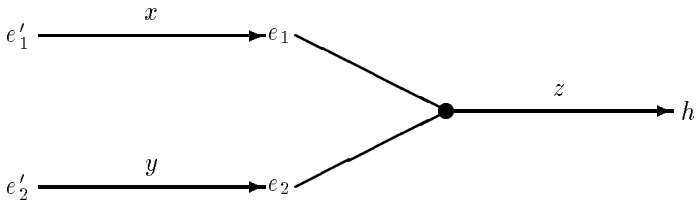
- De aanwijzing e kan een hypothese zijn van andere regels, of e kan worden bevestigd door de gebruiker die de status van e door zijn onderzoek bepaalt. Er moet dan een andere aanwijzing e' onderzocht worden dat e bevestigt. Dit kan ook gepaard gaan met een mate van onzekerheid. Deze situatie ziet er in een afleidingsnetwerk als volgt uit:

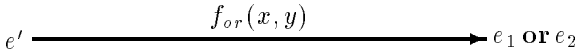
  Daarin wordt de hypothese h bevestigd in de graad y als e absoluut zeker is. Aangezien dit laatste niet zo is, moet de graad waarmee h wordt bevestigd afhankelijk gesteld worden van de graad waarmee e wordt bevestigd. De graad x moet dus naar h toe doorwerken. Dan is er een functie nodig dat x en y combineert en deze naar h propageert. Deze combinatiefunctie noemen we f_{prop}. Het afleidingsnetwerk kan dan als volgt worden (e is in e' opgenomen):

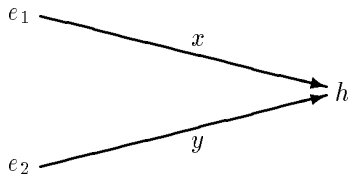
- De aanwijzing e kan een samengestelde aanwijzing zijn. Deze bestaat dan uit twee afzondelijke aanwijzingen verbonden door middel van de operator **and** of **or**. Zie als voorbeeld het netwerk hieronder voor de regel **if** e_1 **or** e_2 **then** h **fi**:

  Er is dan een functie nodig die de onzekerheden van de twee afzondelijke aanwijzingen combineert. Deze functie noemen we de combinatiefunctie voor samengestelde aanwijzingen. Er zijn er hier twee: f_{or} en f_{and}. Een deel van het bovenstaande plaatje kan dan vereenvoudigd worden en wordt dan als volgt:

- Een hypothese h kan in meerdere regels voorkomen. In elk van die regels kan h dan worden bevestigd, zij het dan door observatie van verschillende aanwijzingen. In het onderstaande plaatje is de situatie weergegeven:



Wat voor een onzekerheid moet in zo'n geval naar h gepropageerd worden? Een combinatiefunctie voor voor deze co-concluderende regels is dan nodig. Deze wordt f_{co} genoemd. Het bovenstaande plaatje kan dan als volgt worden vereenvoudigd:

$$e' = e_1 \,\mathbf{co}\, e_2 \xrightarrow{f_{co}(x,y)} h$$

Als naar het bovenstaande verhaal wordt gekeken, is er een aantal combinatiefuncties nodig, waarmee met onzekerheid kan worden gewerkt. In het kort zijn deze functies:

- De functie voor het propageren van de onzekerheid van de aanwijzingen: f_{prop};
- De functies voor samengestelde aanwijzingen: f_{or} and f_{and};
- De functie voor co-concluderende productieregels: f_{co}.

Hoe deze functies werken of wat ze moeten doen hangt af van het model dat voor het werken met onzekerheid gebruikt wordt. Dit model moet dan aangeven hoe ze werken moeten.

2.2 Waarschijnlijkheidstheorie

Waarschijnlijkheidstheorie is een van de eerste methoden waarmee men aan een uitspraak een mate van onzekerheid associeerde met betrekking tot de waarheid ervan. In de eerste expertsystemen die in de jaren '60 ontwikkeld werden, werd deze theorie toegepast. Het gebruik van de theorie wordt hier besproken, evenals wat deze inhoudt.

2.2.1 Kansrekening: definities en stellingen

In de kansrekening worden experimenten gedaan waarin zich mogelijke gebeurtenissen voordoen. We noemen hier een verzameling van alle mogelijke gebeurtenissen die bij een experiment kunnen voordoen, de *uitkomstenruimte* van een experiment. Als notatie voor deze uitkomstenruimte gebruikt men Ω. $e \subseteq \Omega$ is een verzameling van gebeurtenissen die zich voordoen. Om aan te geven dat het om een verzameling van gebeurtenissen die zich niet voordoen gaat, noteren we dit zo: $\bar{e}(= (\Omega \backslash e))$.

In de kansrekening is een aantal definities en/of stellingen van kracht (zie [LG91],[Maa88]), waarvan hier alvast een paar:

Definitie: De gebeurtenissen $e_1, \ldots, e_n \subseteq \Omega$, $n \geq 1$, worden *wederzijds uitsluitende* of *disjuncte gebeurtenissen* genoemd als $e_i \cap e_j = \emptyset$, $i \neq j$, $1 \leq i,j \leq n$.

Definitie: Zij Ω de uitkomstenruimte van een experiment. Als er een getal $P(e)$ aan elke deelverzameling $e \subseteq \Omega$ is toegekend zo dat

1. $P(e) \geq 0$
2. $P(\Omega) = 1$
3. $P(\bigcup_{i=1}^{n} e_i) = \sum_{i=1}^{n} P(e_i)$, als e_i, $i = 1, \ldots, n$, $n \geq 1$, disjuncte gebeurtenissen zijn

dan wordt de functie P een *kansverdeling* op de uitkomstenruimte Ω genoemd. Voor elke deelverzameling $e \subseteq \Omega$, noemt men $P(e)$ de *kans* dat een gebeurtenis optreedt.

Definitie: Zij Ω de uitkomstenruimte van een zeker experiment en zij P een kansverdeling op Ω. Voor elke h,$e \subseteq \Omega$ met $P(e) > 0$, is de *voorwaardelijke kans* van h gegeven e, notatie $P(h|e)$, als volgt gedefinieerd:

$$P(h|e) = \frac{P(h \cap e)}{P(e)}$$

2.2.2 Expertsystemen en kansrekening

De vorige paragraaf levert voldoende materiaal op om te gebruiken in een expertsysteem. Als men de produktieregel **if** e **then** h_x **fi** bekijkt, dan valt er voor x wel iets in te vullen. Als de gebeurtenis e zich voordoet, wat heeft deze voor invloed op de gebeurtenis h? Hoe beïnvloedt een aanwijzing e de kans $P(h)$ dat h waar is? Dit kan in het volgende plaatje duidelijk gemaakt worden:

$$e \xrightarrow{\quad P(h|e) \quad} h$$

Hier is nog een opmerking op zijn plaats. Een kansverdeling P werkt met verzamelingen, terwijl in de productieregels met logische conventies wordt gewerkt. Verzamelingen kunnen gelukkig wel naar logische formules vertaald worden:

Zij Ω een uitkomstenruimte. Voor elk gebeurtenis $e \subseteq \Omega$ definiëren we een predicaat e' zo dat $e'(x) =$ **true** dan en slechts dan als $x \in e$.
Wat doorsneden en verenigingen betreft:

- $e'(x \cap y) = e'(x) \wedge e'(y)$
- $e'(x \cup y) = e'(x) \vee e'(y)$

In de praktijk is $P(h|e)$ niet altijd vast te stellen. $P(e|h)$ daarentegen is wel vast te stellen. In bijvoorbeeld medische literatuur staat van een ziekte altijd de bijbehorende typische symptomen beschreven. Met behulp van de volgende stelling kan $P(h|e)$ nu wel vastgesteld worden:

Stelling (*Stelling van Bayes*)
Zij P een kansverdeling op een uitkomstenruimte Ω. Voor elke h,$e \subseteq \Omega$ zo dat $P(e) > 0$ en $P(h) > 0$, geldt:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

Er bestaat ook de mogelijkheid dat een gebeurtenis e niet van invloed is op de gebeurtenis h. Hier wordt h dan onafhankelijk van e genoemd. Dit is zo als $P(h|e) = P(h)$. Maar dit is niet altijd zo in het omgekeerde geval, dat is dat e onafhankelijk is van h. Er wordt een definitie gegeven van wat nou precies met onafhankelijkheid bedoeld wordt.

Definitie: De gebeurtenissen $e_1, \ldots, e_n \subseteq \Omega$ zijn *(onderling) onafhankelijk* als

$$P(e_{i_1} \cap \cdots \cap e_{i_k}) = P(e_{i_1}) \cdot \cdots \cdot P(e_{i_k})$$

voor elke deelverzameling $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, $1 \leq k \leq n, n \geq 1$. De gebeurtenissen $e_1, \ldots, e_n$ zijn *voorwaardelijk onafhankelijk* gegeven een gebeurtenis $h \subseteq \Omega$ als

$$P(e_{i_1} \cap \cdots \cap e_{i_k}|h) = P(e_{i_1}|h) \cdot \cdots \cdot P(e_{i_k}|h)$$

voor elke deelverzameling $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$.

Tot zover is er slechts gesproken over één e en één bijbehorende h. Hoe gaat men met waarschijnlijkheden te werk als er een verzameling $H = \{h_1, \ldots, h_n\}$ en een verzameling $E = \{e_1, \ldots, e_m\}$ te onderzoeken valt? Stel dat de aanwijzingen of **false** of **true** zijn, dan is de vraag hoe er een verzameling hypotheses $h \subseteq H$ wordt gevonden die het best aansluit bij de onderzochte verzameling van aanwijzingen $e \subseteq E$. In het meest simpele geval wordt voor elke deelverzameling $h \subseteq H$ de kans $P(h|e)$ uitgerekend en daarna de de deelverzameling h' met de hoogste kans geselecteerd. Echter kost dit exponentieel veel berekeningen, namelijk $2^n = \sum_{i=1}^{n} \binom{n}{i}$. Dit kan ook anders. Neem aan dat de hypothesen $h_i \in H, i = 1, \ldots, n$ disjunct zijn, en waarvoor geldt: $\bigcup_{i=1}^{n} h_i = \Omega$. Dan kan de volgende stelling gebruikt worden:

Stelling (*Stelling van Bayes*)
Zij P een kansverdeling op een uitkomstenruimte Ω. Zij $h_i \subseteq \Omega, i = 1, \ldots, n, n \geq 1$, wederzijds uitsluitende hypothesen met $P(h_i) > 0$, zo dat $\bigcup_{i=1}^{n} = \Omega$. Zij verder $e \subseteq \Omega$ zo dat $P(e) > 0$. Dan geldt de volgende eigenschap:

$$P(h_i|e) = \frac{P(e|h_i)P(h_i)}{\sum_{j=1}^{n} P(e|h_j)P(h_j)}$$

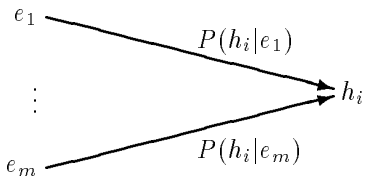
Nu kost het minder berekeningen, namelijk n^2.

Er zit echter een addertje onder het gras. Voor het gebruik van de formule in de bovenstaande stelling moet vooraf alle $P(e|h_j)$ bekend zijn voor elke verzameling van aanwijzingen $e \subseteq E$. Hierbij speelt nog het probleem dat $P(e|h_j)$ doorgaans niet altijd kan worden bepaald uit de voorwaardelijke kansen $P(e_i|h_j)$, $e_i \in e$. Aangezien e uit exponentieel veel combinaties van aanwijzingen bestaan kan, moeten er exponentieel veel kansen vooraf bekend zijn.
Dit kan ook anders. Neem aan dat de aanwijzingen $e_i \in E$ voorwaardelijk onafhankelijk zijn gegeven iedere hypothese $h_j \in H$. Dan kan men gebruik maken van de volgende stelling:

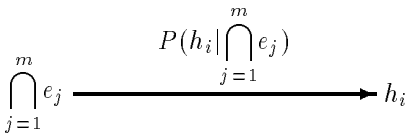
Stelling (*Stelling van Bayes*)
Zij P een kansverdeling op een uitkomstenruimte Ω. Zij $h_i \subseteq \Omega$, $i = 1, \ldots, n$, $n \geq 1$, wederzijds uitsluitende hypothesen met $P(h_i) > 0$, zo dat $\bigcup_{i=1}^{n} = \Omega$. Zij verder $e_{j_1}, \ldots, e_{j_k} \subseteq \Omega$, $1 \leq k \leq m$, $m \geq 1$ aanwijzingen zo dat zij voorwaardelijk onafhankelijk zijn gegeven ieder hypothese h_i. Dan geldt de volgende eigenschap:

$$P(h_i|e_{j_1} \cap \cdots \cap e_{j_k}) = \frac{P(e_{j_1}|h_i) \cdots P(e_{j_k}|h_i)P(h_i)}{\sum_{l=1}^{n} P(e_{j_1}|h_l) \cdots P(e_{j_k}|h_l)P(h_l)}$$

Deze stelling biedt een voordeel: het is te gebruiken als een combinatiefunctie voor co-concluderende regels. Bekijk namelijk het volgende afleidingsnetwerk:



Nu kan met de stelling van Bayes de gecombineerde invloed van de aanwijzingen $e_1, \ldots, e_m$ op de a priori kans van de hypothese h_i als volgt uitgerekend worden:



Voorzover is deze waarschijnlijkheidstheorie geschikt om te gebruiken in een expertsysteem. De theorie legt echter wel beperkingen op aan het probleemgebied waarin het systeem moet werken:

- De hypothesen $h_1, \ldots, h_n, n \geq 1$ zijn disjunct;
- Voor de hypothesen $h_1, \ldots, h_n$ geldt verder nog: $\bigcup_{i=1}^{n} h_i = \Omega$;
- De aanwijzingen $e_1, \ldots, e_m, m \geq 1$ zijn voorwaardelijk onafhankelijk gegeven iedere hypothese $h_i, 1 \leq i \leq n$.

De waarschijnlijkheidstheorie heeft zelf ook enige beperkingen:

- Ze kan geen beschrijving geven van wat de combinatiefunctie voor het propageren van een onzekere aanwijzing is;
- De combinatiefunctie voor samengestelde aanwijzingen is hier ook niet te beschrijven.

Verder speelt nog het probleem van de beschikbaarheid van de statistische gegevens waaruit de kansen moeten worden afgeleid. Dat probleem valt nog wel te omzeilen: men laat de experts kansen toekennen aan de door diezelfde experts opgestelde regels. Men maakt hier dan gebruik van *subjectieve waarschijnlijkheden.*

Als gevolg van de hierboven beschreven problemen is men later op zoek gegaan naar zogenaamde pseudo-waarschijnlijkheidsmodellen. Een aantal ervan wordt hierna besproken.

2.3 De subjectieve Bayesische methode

De subjectieve Bayesische methode is bedacht door de heren Duda, Hart en Nilsson ([DHN90]). De methode levert een oplossing voor het probleem hoe waarschijnlijkheden door het netwerk gepropageerd moet worden. Gesteld dat men van een aanwijzing de waarschijnlijkheid verandert. Dan zou dat leiden tot herberekeningen van waarschijnlijkheden voor bijbehorende hypothesen in regels waarin die veranderde aanwijzing voorkomt. Op hun beurt kunnen hypothesen weer aanwijzingen zijn in andere regels, enzovoorts. Dus wordt er gezocht naar een mechanisme voor het propageren van waarschijnlijkheden. Deze moet met de volgende zaken rekening houden:

- Aan ieder regel is een onzekerheid geassocieerd;
- De aanwijzing die geleverd wordt kan onzeker zijn;
- De door de expert geleverde waarschijnlijkheden kunnen inconsistent zijn.

De resultaten zijn een combinatiefunctie voor het propageren van onzekerheden in aanwijzingen f_{prop} en de combinatiefunctie voor co-concluderende produktieregels f_{co}. f_{prop} kon door de waarschijnlijkheidstheorie niet geleverd worden. f_{co} is hier een gemodificeerde versie van de combinatiefunctie zoals die in de waarschijnlijkheidstheorie is beschreven. Deze functies worden hier behandeld.

2.3.1 Een definitie van de te propageren getallen

In deze methode worden niet direct notaties van de waarschijnlijksheidstheorie gebruikt, slechts notaties die daaraan gerelateerd zijn. Het doel daarvan is om iets anders dan kansen te gebruiken dat ook beter te propageren is.
Een stel definities zoals ze in [LG91] beschreven staan en die hierna worden beschreven, zijn gebaseerd op de Stelling van Bayes. Met behulp van deze stelling krijgt men de volgende vergelijking:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

Er is ook een complementaire versie:

$$P(\bar{h}|e) = \frac{P(e|\bar{h})P(\bar{h})}{P(e)}$$

Door de eerste vergelijking met de tweede te delen wordt het resultaat:

$$\frac{P(h|e)}{P(\bar{h}|e)} = \frac{P(e|h)}{P(e|\bar{h})} \cdot \frac{P(h)}{P(\bar{h})}$$

Elke component in deze vergelijking heeft een betekenis. De componenten zijn in de definities weer te vinden.

Definitie: zij P een kansverdeling op een uitkomstenruimte Ω. Zij verder $h \subseteq \Omega$ zo dat $P(h) < 1$. De *a priori odds* van de gebeurtenis h, notatie $O(h)$, is als volgt gedefinieerd:

$$O(h) = \frac{P(h)}{P(\bar{h})} = \frac{P(h)}{1 - P(h)}$$

Definitie: zij P een kansverdeling op een uitkomstenruimte Ω. Zij verder $h, e \subseteq \Omega$ zo dat $P(e) > 0$ en $P(h|e) < 1$. De *a posteriori odds* van een hypothese h gegeven aanwijzing e, notatie $O(h|e)$, is als volgt gedefinieerd:

$$O(h|e) = \frac{P(h|e)}{P(\bar{h}|e)} = \frac{P(h|e)}{1 - P(h|e)}$$

Definitie: zij P een kansverdeling op een uitkomstenruimte Ω. Zij verder $h, e \subseteq \Omega$ zo dat $0 < P(h) < 1$ en $P(e|\bar{h}) > 0$. De *(positieve) waarschijnlijkheidsverhouding* λ gegeven h en e wordt gedefinieerd door:

$$\lambda = \frac{P(e|h)}{P(e|\bar{h})}$$

Hoe in de definities de odds en de waarschijnlijkheidsverhouding geïnterpreteerd moeten worden, laat zich niet makkelijk beschrijven. Daarom wordt voor het begrip een voorbeeld gegeven. Dit voorbeeld komt uit [Pea88].

Voorbeeld 2.3.1

Stel je wordt gewekt door het schelle geluid van het inbraakalarm. In welke mate geloof je dat er een inbraakpoging is geweest? Ter illustratie worden de volgende beoordelingen gemaakt:

1. *Er is 95% kans dat een inbraakpoging het alarm doet afgaan:* $P(\text{Alarm}|\text{Inbraak}) = 0.95$;
2. *Gebaseerd op het feit dat eerder wel eens vals alarm is gewest, is er een kleine (1%) kans dat het alarm door iets anders dan een inbraakpoging afgaat:* $P(\text{Alarm}|\neg\text{Inbraak}) = 0.01$;
3. *Er zijn misdaadcijfers gegeven die aangeven dat er een kans van 1 op 10000 is dat in een willekeurig huis in een willekeurige nacht wordt ingebroken:* $P(\text{Inbraak}) = 10^{-4} = 0.00001$.

Als deze aannamen worden samengevoegd waarbij de vergelijking $O(h|e) = \lambda \cdot O(h)$ *wordt gebruikt, dan is:*

$$\begin{aligned} O(\text{Inbraak}|\text{Alarm}) &= \lambda \cdot O(\text{Inbraak}) \\ &= \frac{P(\text{Alarm}|\text{Inbraak})}{P(\text{Alarm}|\not{\text{Inbraak}})} \cdot O(\text{Inbraak}) \\ &= \frac{0.95}{0.01} \cdot \frac{10^{-4}}{1 - 10^{-4}} \\ &= 0.0095. \end{aligned}$$

Met gebruikmaking van $P = O/(O+1)$ krijgt men dan:

$$\begin{aligned} P(\text{Inbraak}|\text{Alarm}) &= \frac{0.0095}{1+0.0095} \\ &= 0.00941. \end{aligned}$$

Dus de hypothese van inbraak dat wordt gesteund door de aanwijzing van het alarm doet de mate waarin je dit gelooft toenemen met een factor van bijna 100: van 1 op 10000 naar 94.1 op 10000.
Het feit dat de kans dat er is ingebroken kleiner dan 1% is, is niet verassend: eens in de drie maanden is er vals alarm. □

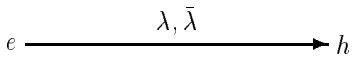
Essentieel in het voorbeeld is het feit dat op het moment dat het alarm afgaat, *niet direct* te bepalen is of er ingebroken is. Uit dit gegeven is $P(\text{Alarm}|\text{Inbraak})$ niet af te leiden. Beperkt men zich tot het (lokale) feit met betrekking tot de werking van het alarm in termen van waarschijnlijkheidsverhoudingen, en het (lokale) feit van waarschijnlijkheid van inbraak, dan is dit wel van invloed op de beoordeling van de mogelijkheid dat het alarm door een inbraakpoging is afgegaan.

Tot zover is er de situatie bekeken waarin een hypothese door een aanwijzing wordt gesteund. Er bestaat ook de situatie dat een hypothes wordt ontkend door de aanwijzing. Hieronder wordt daarvoor een begrip gedefinieerd.

Definitie: zij P een kansverdeling op een uitkomstenruimte Ω. Zij verder $h, e \subseteq \Omega$ zo dat $0 < P(h) < 1$ en $P(e|\bar{h}) < 1$. De *(negatieve) waarschijnlijkheidsverhouding* $\bar{\lambda}$ gegeven h en e wordt gedefinieerd door:

$$\bar{\lambda} = \frac{1 - \lambda \cdot P(e|\bar{h})}{1 - p(e|\bar{h})}$$

Nu kan de subjectieve Bayesische methode toegepast worden. Men moet hiertoe dan wel met elke produktieregel **if** e **then** h **fi** de *waarschijnlijkheidsverhoudingen* (*likelihood ratios*) associëren, zoals hieronder in het plaatje:



De expert geeft voor iedere regel de a priori kansen $P(h)$ en $P(e)$, en daarnaast nog de voorwaardelijke kansen $P(e|h)$ en $P(e|\bar{h})$. Uit deze laatste twee worden de waarschijnlijkheidsverhoudingen berekend. λ en $\bar{\lambda}$ kunnen ook worden gegeven, want daaruit kunnen weer uniek de voorwaardelijke kansen worden berekend. Er gelden namelijk de volgende eigenschappen:

$$\begin{aligned} O(h|e) &= \lambda \cdot O(h) \\ O(h|\bar{e}) &= \bar{\lambda} \cdot O(h) \\ P &= \frac{O}{O+1} \end{aligned}$$

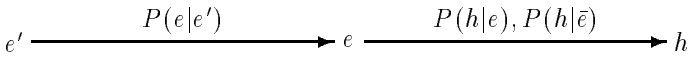
Dit kan bewezen worden met behulp van de regel van Bayes en de hierbovenstaande definities. In het geheel levert dit de (belangrijke) basis voor de functies die in de volgende paragraaf worden besproken.

2.3.2 De combinatiefuncties

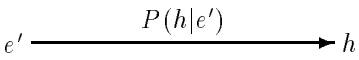
Met de in de vorige paragraaf gegeven definities is het nu mogelijk om de propagatiefunctie te construeren. Daarnaast is in de subjectieve Bayesische methode ook een combinatiefunctie voor co-concluderende productieregels beschreven. Deze en de overige functies worden hierna behandeld.

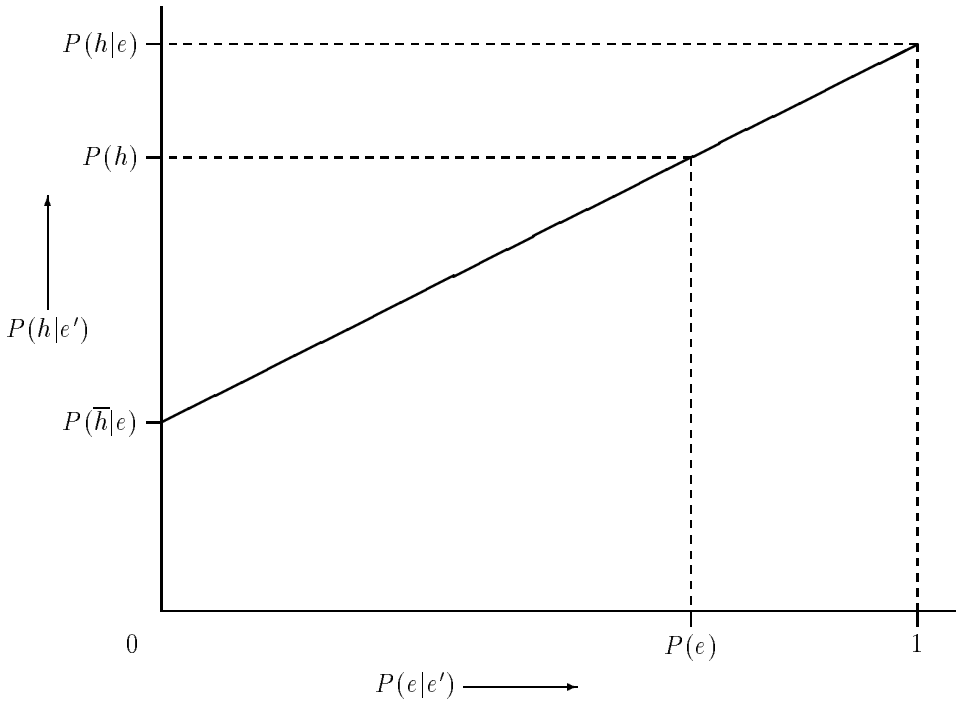
2.3.2.1 De combinatiefunctie voor het propageren van onzekerheden in aanwijzingen

Bij het beschrijven van een propagatiemechanisme wordt hier gebruik gemaakt van een produktieregel **if** e **then** h **fi** waarin e een hypothese is van een ander regel. Deze e wordt gebruikt als aanwijzing voor het bevestigen van de hypothese h. Veronderstel dat e wordt bevestigd door een te onderzoeken aanwijzing e' die bovendien onzeker is. Zie voor deze situatie het plaatje hieronder waarin de kans $P(e|e')$ al berekend is.



Nadat de regel is toegepast, wordt de situatie als volgt:

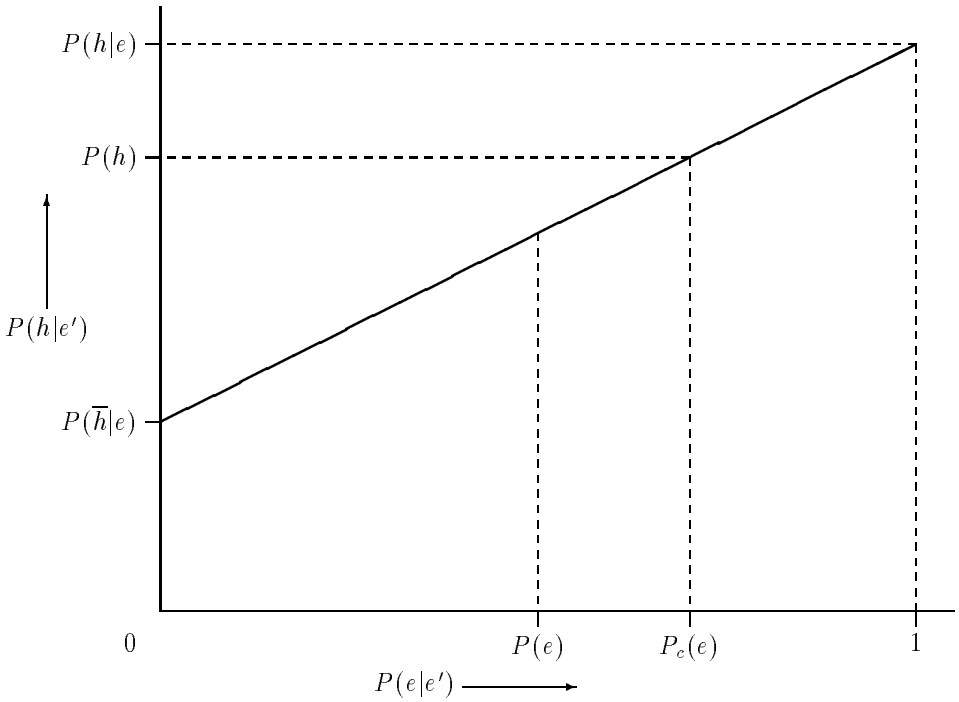




Figuur 2.3: Een grafische weergave van de interpolatiefunctie

Wat moet die $P(h|e')$ dan zijn? De expert kan deze niet geven, en de kansverdeling P is niet volledig gegeven. Immers we weten de relatie tussen e en e' niet. Dan moet er geschat worden. Er kan hier gebruik gemaakt worden van het volgende:

$$\begin{aligned}
P(h|e') &= P(h \cap e|e') + P(h \cap \bar{e}|e') \\
&= \frac{P(h \cap e \cap e')}{P(e')} \cdot \frac{P(e \cap e')}{P(e \cap e')} + \frac{P(h \cap \bar{e} \cap e')}{P(e')} \cdot \frac{P(\bar{e} \cap e')}{P(\bar{e} \cap e')} \\
&= \frac{P(h \cap e \cap e')}{P(e \cap e')} \cdot \frac{P(e \cap e')}{P(e')} + \frac{P(h \cap \bar{e} \cap e')}{P(\bar{e} \cap e')} \cdot \frac{P(\bar{e} \cap e')}{P(e')} \\
&= P(h|e \cap e')P(e|e') + P(h|\bar{e} \cap e')P(\bar{e}|e')
\end{aligned}$$



Figuur 2.4: De situatie met inconsistente kansen $P(h)$ en $P(e)$.

Als aangenomen wordt dat bekend is dat e absoluut waar of onwaar is, dan levert de aanwijzing e' dat relevant voor e is, geen nieuwe informatie voor de hypothese h. Dit levert dan het volgende op:

$$\begin{aligned} P(h|e') &= P(h|e)P(e|e') + P(h|\bar{e})P(\bar{e}|e') \\ &= P(h|e)P(e|e') + P(h|\bar{e})(1 - P(e|e')) \\ &= (P(h|e) - P(h|\bar{e}))P(e|e') + P(h|\bar{e}) \end{aligned}$$

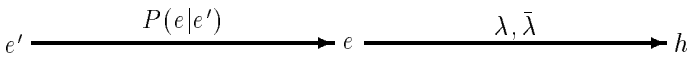
Het resultaat is een interpolatiefunctie voor $P(h|e')$. Figuur 2.3 geeft deze grafisch weer.

Voor elke produktieregel **if** e **then** h **fi** levert een expert de kansen $P(e)$, $P(h)$, $P(h|e)$, en$P(h|\bar{e})$. Deze kunnen onderling echter inconsistent zijn, omdat er geen onderliggende kansverdeling is. Een zo'n situatie waarin inconsistentie voorkomt, is gegeven in figuur 2.4. In die figuur is $P_c(e)$ de consistente waarde voor $P(e|e')$ met bijbehorende $P(h)$. Geeft men in de interpolatiefunctie voor $P(e|e')$ een waarde tussen $P(e)$ en $P(e')$, dan wordt $P(h|e') < P(h)$. Dit betekent echter dat bevestiging van e leidt tot ontkenning van h en dat is niet wat de regel **if** e **then** h **fi** uitdrukt. De oplossing is hier het opsplitsen van de interpolatiefunctie in twee verschillende functies, waarbij er voor gezorgd wordt dat met $P(e)$ de bijbehorende waarde $P(h)$ wordt verkregen. De opgesplitste functie wordt hieronder gegeven:

$$P(h|e') = \begin{cases} P(h|\bar{e}) + \dfrac{P(h) - P(h|\bar{e})}{P(e)} \cdot P(e|e') & \text{als } 0 \leq P(e|e') \leq P(e) \\ P(h) + \dfrac{P(h|e) - P(h)}{1 - P(e)} \cdot (P(e|e') - P(e)) & \text{als } P(e) < P(e|e') \leq 1 \end{cases}$$

het zal duidelijk zijn, dat deze functie als combinatiefunctie voor het propageren van onzekerheden in aanwijzingen kan dienen, oftewel f_{prop}.

In het begin is gezegd dat bij elke produktieregel λ en $\bar{\lambda}$ gegeven wordt zoals in het volgende plaatje:

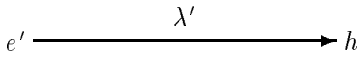


Met de volgende definitie hieronder:

Definitie: Zij P een kansverdeling op een uitkomstenruimte Ω, en zij O de bijbehorende odds als hiervoor gedefinieerd. Zij verder $h, e' \subseteq \Omega$. De *effectieve waarschijnlijkheidsverhouding* (*effective likelihood ratio*) λ', gegeven h en e', is als volgt gedefinieerd:

$$\lambda' = \frac{O(h|e')}{O(h)}$$

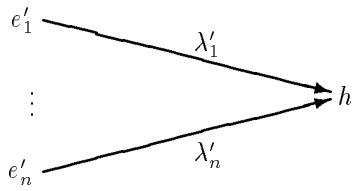
wordt het resultaat:



Het zal duidelijk zijn dat de effectieve waarschijnlijkheidsverhouding λ' tussen λ en $\bar{\lambda}$ in zit. λ' zit dichter bij λ als e in een bepaalde mate wordt bevestigd door e', en dichter bij $\bar{\lambda}$ als e in een bepaalde mate wordt ontkend door e'.

2.3.2.2 De combinatiefunctie voor co-concluderende produktieregels

In een afleidingsnetwerk komt het ook voor dat er meerdere produktieregels zijn die alle een en dezelfde hypothese hebben. Zie het plaatje hieronder.



Hierin is de onzekerheid in e_i gegeven een aanwijzing e_i' al naar h gepropageerd.

Het bepalen van $P(h|e')$ waarin e' de verzameling $\{e_1', \ldots, e_n'\}$ is, gaat analoog aan die in de waarschijnlijkheidstheorie. Hier wordt echter met odds en waarschijnlijkheidsverhoudingen gewerkt. Er kan gebruik worden gemaakt van de volgende regels:

$$\begin{array}{ll} O(h|e) = \lambda \cdot O(h) & \text{als } e \text{ absoluut waar is} \\ O(h|\bar{e}) = \bar{\lambda} \cdot O(h) & \text{als } \bar{e} \text{ absoluut waar is} \end{array}$$

Dit kan ook met meervoudige aanwijzingen. Onder de aanname dat de aanwijzingen e_i, $i = 1, \ldots, n$ *onderling onafhankelijk* zijn, gelden de volgende eigenschappen:

$$O(h|\bigcap_{i=1}^{n} e_i) = \left[\prod_{i=1}^{n} \lambda_i\right] \cdot O(h) \quad \text{als } e_i \text{ absoluut waar is voor alle } i = 1, \ldots, n$$

$$O(h|\bigcap_{i=1}^{n} \bar{e_i}) = \left[\prod_{i=1}^{n} \bar{\lambda}_i\right] \cdot O(h) \quad \text{als } \bar{e_i} \text{ absoluut waar is voor alle } i = 1, \ldots, n$$

Hier is λ_i en $\bar{\lambda}_i$ volgens de normale definities van λ en $\bar{\lambda}$, maar nu met e_i in plaats van e.

Nu kunnen ook hier de aanwijzingen e_i gegeven de aanwijzingen e'_i onzeker zijn. Als hier nu ook wordt verondersteld dat de e'_i *onderling onafhankelijk* zijn, dan kan de volgende formule worden gebruikt:

$$O(h|\bigcap_{i=1}^{n} e'_i) = \left[\prod_{i=1}^{n} \lambda'_i\right] \cdot O(h) \quad \text{waarin } \lambda'_i = \frac{O(h|e'_i)}{O(h)}$$

2.3.2.3 De combinatiefuncties voor samengestelde aanwijzingen

De subjectieve Bayesische methode geeft geen oplossing voor de functies f_{and} en f_{or}. De kansen voor de samenhangende aanwijzingen zijn niet vast te stellen omdat de onderlinge relaties tussen de aanwijzingen complex zijn. Hier dient er geschat te worden. In [LG91] worden de functies gegeven, met de opmerking erbij dat ze gebruikt zijn in een expertsysteem genaamd PROSPECTOR. De invullingen voor f_{and} respectievelijk f_{or} zijn:

$$\begin{aligned} P(e_1 \textbf{ and } e_2) &= \min\{P(e_1|e'), P(e_2|e')\} \\ P(e_1 \textbf{ or } e_2) &= \max\{P(e_1|e'), P(e_2|e')\} \end{aligned}$$

2.3.3 Tot slot

In de subjectieve Bayesische methode zijn voor enkele combinatiefuncties aannamen gemaakt:

- Voor de propagatiefunctie f_{prop} wordt aangenomen dat als e absoluut waar of onwaar is, dat de onderzochte aanwijzing e' geen nieuwe informatie geeft voor h.
- Voor het gebruik van de functie f_{co} geldt de aanname dat de onderzochte aanwijzingen onderling onafhankelijk zijn.

Er zijn nog enkele eigenschappen voor de combinatiefuncties niet genoemd. De functies zijn associatief en commutatief.

2.4 Het Certainty Factor model

Het model is ontworpen door Shortliffe en Buchanan ([BS84]) en is toegepast in het door hen onwikkelde medische experstsysteem MYCIN. In het model is er gekozen voor twee maten van onzekerheid, die geen kansen zijn:

- een maat van vertrouwen. Deze geeft aan in welke mate een onderzochte aanwijzing het vertrouwen in een zekere hypothese **versterkt**;
- een maat van wantrouwen. Deze geeft aan in welke mate een onderzochte aanwijzing het vertrouwen in een zekere hypothese **verzwakt**;

De reden hiervan is dat een expert aan een produktieregel een maat voor de onzekerheid toekent, die niet gebaseerd is op waarschijnlijkheidstheorie. Uit onderzoeken hiernaar is het volgende gebleken, dat logische implicaties van uitspraken die op waarschijnlijkheidstheorie zijn gebaseerd niet worden geaccepteerd. Als een expert aan een regel $P(h|e) = x$ toekent, dan geldt volgens de theorie $P(\bar{h}|e) = 1 - x$. Dat laatste slikt de expert niet. Het is niet de bedoeling een aanwijzing dat een hypothese ondersteunt te gebruiken als aanwijzing voor het ontkennen van diezelfde hypothese.

In de rest van deze paragraaf wordt het model uit de doeken gedaan. De genoemde maten worden hieronder gedefinieerd. Tevens worden de karakteristieke eigenschappen besproken.

2.4.1 Definities en karakteristieken

Definitie: Zij P een waarschijnlijkheidsfunctie gedefinieerd op een uitkomstenruimte Ω, en zij $h, e \subseteq \Omega$ zo dat $P(e) > 0$. De *maat voor het (toegenomen) vertrouwen* MB is een functie MB: $2^\Omega \times 2^\Omega \rightarrow [0,1]$, zo dat

$$\mathrm{MB}(h,e) = \begin{cases} 1 & \text{als } P(h) = 1 \\ \max\left\{0, \dfrac{P(h|e) - P(h)}{1 - P(h)}\right\} & \text{anders} \end{cases}$$

De *maat voor het (toegenomen) wantrouwen* MD is een functie MD: $2^\Omega \times 2^\Omega \rightarrow [0,1]$, zo dat

$$\mathrm{MD}(h,e) = \begin{cases} 1 & \text{als } P(h) = 0 \\ \max\left\{0, \dfrac{P(h) - P(h|e)}{P(h)}\right\} & \text{anders} \end{cases}$$

Hierop is een nieuwe maat gedefinieerd, namelijk de *zekerheidsfactor* (*certainty factor*). Deze is als volgt:

$$\mathrm{CF}(h,e) = \mathrm{MB}(h,e) - \mathrm{MD}(h,e)$$

De reden om dit zo te doen is om de mogelijkheid te hebben om een vergelijking te maken van de sterktes van de aanwijzingen van hypothesen waaruit "gekozen" moet worden. Hoe deze maten MD, MB, en CF zich tot elkaar verhouden, staat hieronder:

Karakteristieken met betrekking tot de maten van vertrouwen

1. Bereik van de maten:
   (a) $0 \leq \mathrm{MB}(h,e) \leq 1$
   (b) $0 \leq \mathrm{MD}(h,e) \leq 1$
   (c) $-1 \leq \mathrm{CF}(h,e) \leq 1$
2. Sterkte van de aanwijzingen en wederzijds uitsluitende hypothesen:
   Als aangetoond is dat h zeker is ($P(h|e) = 1$):
   (a) $\mathrm{MB}(h,e) = \dfrac{1 - P(h)}{1 - P(h)} = 1$
   (b) $\mathrm{MD}(h,e) = 0$
   (c) $\mathrm{CF}(h,e) = 1$

   Als aangetoond is dat de ontkenning van h zeker is ($P(\neg h|e) = 1$):
   (a) $\mathrm{MB}(h,e) = 0$
   (b) $\mathrm{MD}(h,e) = \dfrac{0 - P(h)}{0 - P(h)} = 1$
   (c) $\mathrm{CF}(h,e) = -1$

   Merk op dat dit het volgende oplevert: $\mathrm{MB}(\neg h, e) = 1$ dan en slechts dan als $\mathrm{MD}(h,e) = 1$ volgens de definities van MB en MD zoals hiervoor gegeven. Verder wordt met 1 het absolute vertrouwen (of wantrouwen) voor MB (of MD) bedoeld. Dus als $\mathrm{MB}(h_1, e) = 1$ en h_1 en h_2 zijn disjunct, dan is $\mathrm{MD}(h_2, e) = 1$.
3. Als een aanwijzing niets oplevert:
   (a) $\mathrm{MB}(h,e) = 0$ als h niet wordt bevestigd door e (e en h zijn dan onafhankelijk of e ontkent h)

(b) $\mathrm{MD}(h,e) = 0$ als h niet wordt ontkend door e (e en h zijn dan onafhankelijk of e bevestigt h)

(c) $\mathrm{CF}(h,e) = 0$ als e h noch ontkent noch bevestigt (e en h zijn dan onafhankelijk)

De zekerheidsfactor is een goede maat voor het toekennen van getallen door experts om de sterkte van de regels aan te geven:

- de expert geeft een positief getal ($\mathrm{CF} > 0$) om aan te geven dat de hypothese wordt bevestigd door de bekeken aanwijzing;
- de expert geeft een negatief getal ($\mathrm{CF} < 0$) als de aanwijzing leidt tot ontkenning van de hypothese;
- en als er geen aanwijzing is dat van invloed is op de beoogde hypothese, dan kiest men $\mathrm{CF} = 0$.

De expert kan niet zomaar getallen toekennen. Er zijn grenzen aan de som van de zekerheidsfactoren van disjuncte hypothesen. Een onderzoek van Buchanan en Shortliffe ([BS84]) levert de volgende grenzen:

- Als er k van de n disjuncte hypothesen h_i absoluut worden ontkend door een aanwijzing e, dan is de uitspraak:

$$\sum_{i=1}^{k} \mathrm{CF}(h_i, e) \geq -k \text{ (voor } h_i \text{ ontkend door } e)$$

- Als er k van de n disjuncte hypothesen h_i absoluut worden bevestigd door een aanwijzing e, dan is de uitspraak:

$$\sum^{k} \mathrm{CF}(h_i, e) \leq 1 \text{ (voor } h_i \text{ bevestigd door } e)$$

Dit laatste geeft een middel om nieuwe regels te onderzoeken die door experts zijn gegeven. Als de hypothesen disjunct zijn maar de som van de CF's is groter dan 1, dan is er wat fout. Dan moet of de expert de getallen veranderen, of de zaak moet zodanig worden genormaliseerd, dat de som van de CF's niet boven de 1 uitkomt. Dit komt het gedrag van het systeem echter niet ten goede. Dan rest nog de aanpassing van de regels zelf.

2.4.2 De combinatiefuncties

De hier te presenteren functies zijn hier niet formeel gedefinieerd, maar zijn benaderingen van wat de zekerheidsfactoren zouden moeten zijn. Een poging om vast te stellen wat een zekerheidsfactor moet zijn, wordt hierna besproken.

Stel voor elke s_k zijn $\mathrm{MB}(d_i, s_k)$ en $\mathrm{MD}(d_i, s_k)$ bekend, en e is een conjunctie van alle s_k. Het doel is om $\mathrm{CF}(d_i, e)$ te berekenen uit de MB's en de MD's, die voor iedere s_k bekend zijn.
Stel dat $e = s_1 \,\&\, s_2$ en e bevestigt d_i. Dan is:

$$\begin{aligned} \mathrm{CF}(d_i, e) = \mathrm{MB}(d_i, e) - 0 &= \frac{P(d_i|e) - P(d_i)}{1 - P(d_i)} \\ &= \frac{P(d_i|s_1 \,\&\, s_2) - P(d_i)}{1 - P(d_i)} \end{aligned}$$

Er is echter geen exacte representatie van $\mathrm{CF}(d_i, s_1 \,\&\, s_2)$ in alleen de termen $\mathrm{CF}(d_i, s_1)$ en $\mathrm{CF}(d_i, s_2)$. Om $P(d_i|s_1 \,\&\, s_2)$ te berekenen moeten ook de onderlinge relaties tussen s_1 en s_2, en overige relaties die er met d_i zijn, bekend zijn. Verder is er een extra probleem met de zekerheidsfactor. Volgens de definities moeten MD en MB van elkaar geïsoleerd zijn. De mogelijkheid bestaat

dat s_1 d_i bevestigt (MB > 0) maar dat s_2 d_i ontkent (MD > 0). Een gegeven CF$(d_i, s_1 \,\&\, s_2)$ moet deze situatie correct weergeven. Dit is niet te doen. Het enige wat men wel weet is dat CF$(d_i, s_2) \leq$ CF$(d_i, s_1 \,\&\, s_2) \leq$ CF(d_i, s_1). Verder is het van belang voor de commutativiteit van de berekeningen als MB en MD gescheiden worden behandeld.

De combinatiefuncties worden wel gegeven, met dien verstande dat ze slechts bij benadering zijn. Dit is een veel gebruikte techniek, bij gebrek aan beter en omdat de nodige kennis en informatie niet aanwezig zijn. Zij dienen zich wel te houden aan de hieronder gegeven criteria. In deze criteria is $e-$ een representatie van alle ontkennende aanwijzingen die aanwezig zijn, en $e+$ een representatie van alle bevestigende aanwijzingen die aanwezig zijn.

Criteria voor de definitie

1. Limieten:
   (a) MB$(h, e+)$ gaat naar 1 als bevestigende aanwijzingen worden gevonden, en is gelijk aan 1 als er een aanwijzing is dat h met zekerheid impliceert;
   (b) MB$(h, e-)$ gaat naar 1 als ontkennende aanwijzingen worden gevonden, en is gelijk aan 1 als er een aanwijzing is dat $\neg h$ met zekerheid impliceert;
   (c) CF$(h, e-) \leq$ CF$(h, e- \,\&\, e+) \leq$ CF$(h, e+)$.
2. Absolute bevestiging of ontkenning:
   (a) Als MB$(h, e+) = 1$, dan MD$(h, e-) = 0$, ongeacht de ontkennende aanwijzing in $e-$; ofwel CF$(h, e+) = 1$;
   (b) Als MD$(h, e-) = 1$, dan MB$(h, e+) = 0$, ongeacht de bevestigende aanwijzing in $e-$; ofwel CF$(h, e-)$ = -1;
   (c) Het geval waarin MB$(h, e+) =$ MD$(h, e-) = 1$ is tegenstrijdig en dientengevolge is CF niet gedefinieerd.
3. Commutativiteit: Als $s_1 \,\&\, s_2$ weergeeft dat eerst s_1 en dan s_2 is bekeken, dan:
   (a) MB$(h, s_1 \,\&\, s_2) =$ MB$(h, s_2 \,\&\, s_1)$;
   (b) MD$(h, s_1 \,\&\, s_2) =$ MD$(h, s_2 \,\&\, s_1)$;
   (c) CF$(h, s_1 \,\&\, s_2) =$ CF$(h, s_2 \,\&\, s_1)$.

   Dit is heel belangrijk. het geeft aan dat de volgorde waarin de aanwijzingen worden ontdekt niet uitmaakt met betrekking tot de mate vertrouwen of wantrouwen in een hypothese.
4. Ontbrekende informatie: Zij $s_?$ een potentiële aanwijzing, waarvan niet bekend is of deze waar of onwaar is, dan:
   (a) MB$(h, s_1 \,\&\, s_?) =$ MB(h, s_1);
   (b) MD$(h, s_1 \,\&\, s_?) =$ MD(h, s_1);
   (c) CF$(h, s_1 \,\&\, s_?) =$ CF(h, s_1).

   De regels van de vorm CF$(h\,, s_?) = x$ moeten kunnen worden overgeslagen als van $s_?$ niet vastgesteld kan worden, wat deze is. Als men iets niet weet dan moet men toch door kunnen werken. Men slaat gewoon de regels over waarvan de zekerheidsfactor niet vast te stellen is.

Er is eerder aangegeven dat MB en MD apart behandeld diende te worden. Het kan voorkomen dat zowel MB als MD niet 0 zijn. De zekerheidsfactor CF kan hier echter wat mee doen. Zij representeert dan het *netto vertrouwen*.
Omdat in het model met zekerheidsfactoren wordt gewerkt, is het prettig om formules voor de combinatiefuncties in termen van CF te beschrijven. Lucas en Van Der Gaag ([LG91]) hebben deze formules beschreven. Zij zijn eerder conventies voor het benaderen van het uiteindelijke resultaat zoals ook in andere theorieën gedaan wordt. Hier komen ze:

- De functie voor het propageren van onzekerheden f_{prop}:

$$\mathrm{CF}(h, e') \quad = \quad \mathrm{CF}(h, e) \cdot \max\{0, \mathrm{CF}(e, e')\}$$

- De functies voor samengestelde aanwijzingen:

  Voor de functie f_{and}:

$$\mathrm{CF}(e_1 \,\mathbf{and}\, e_2, e') \quad = \quad \min\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

  Voor de functie f_{or}:

$$\mathrm{CF}(e_1 \,\mathbf{or}\, e_2, e') \quad = \quad \max\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

- De functie voor co-concluderende aanwijzingen f_{co}:

$\mathrm{CF}(h, e_1 \,\mathbf{co}\, e_2) =$

$$\begin{cases} \mathrm{CF}(h,e_1) + \mathrm{CF}(h,e_2)\cdot(1-\mathrm{CF}(h,e_1)) & \text{als } \mathrm{CF}(h,e_i) > 0, i = 1,2 \\ \dfrac{\mathrm{CF}(h,e_1) + \mathrm{CF}(h,e_2)}{1 - \min\{|\mathrm{CF}(h,e_1)|, |\mathrm{CF}(h,e_2)|\}} & \text{als } -1 < \mathrm{CF}(h,e_1)\cdot\mathrm{CF}(h,e_2) \leq 0 \\ \mathrm{CF}(h,e_1) + \mathrm{CF}(h,e_2)\cdot(1-\mathrm{CF}(h,e_1)) & \text{als } \mathrm{CF}(h,e_i) < 0, i = 1,2 \end{cases}$$

2.4.3 De combinatiefunctie voor co-concluderende aanwijzingen

De functie f_{co} ziet er niet triviaal uit. De overige functies lijken veel op de functies in andere modellen, maar deze is anders. Buchanan en Shortliffe hebben de functie in eerste instantie wel zo gemaakt dat deze voldoet aan de criteria voor de definitie.

Er zijn echter wel problemen met deze functie. Om deze te kunnen behandelen, wordt eerst de functie f_{co} geven in termen van MB en MD.

$$\mathrm{MB}(h, e_1 \,\mathbf{co}\, e_2) = \begin{cases} 0 & \text{als } \mathrm{MD}(h, e_1 \,\mathbf{co}\, e_2) = 1 \\ \mathrm{MB}(h,e_1) + \mathrm{MB}(h,e_2)\cdot(1-\mathrm{MB}(h,e_1)) & \text{anders} \end{cases}$$

$$\mathrm{MD}(h, e_1 \,\mathbf{co}\, e_2) = \begin{cases} 0 & \text{als } \mathrm{MB}(h, e_1 \,\mathbf{co}\, e_2) = 1 \\ \mathrm{MD}(h,e_1) + \mathrm{MD}(h,e_2)\cdot(1-\mathrm{MD}(h,e_1)) & \text{anders} \end{cases}$$

Zoals ervoor gezegd zijn er problemen met die functie. Er valt het een en ander op te merken:

- De functie zorgt ervoor dat MD of MB altijd toeneemt ongeacht of er relaties zijn tussen een nieuwe aanwijzing en de al beschikbare aanwijzingen. In de praktijk hoeft dit niet zo te zijn. In de fysica komt het nog wel voor dat ieder van de twee observaties apart een hypothese steunen, terwijl de conjunctie van die observaties de hypothese ontkent. In het certainty factor model is er aangenomen dat zoiets zich niet voordoet, ofwel er wordt enige onafhankelijkheid verondersteld. In de formules hierboven komt dat ook enigzins tot uitdrukking:

$$\begin{aligned} \mathrm{MB}(h, e_1 \,\mathbf{co}\, e_2) &= \mathrm{MB}(h,e_1) + \mathrm{MB}(h,e_2)\cdot(1-\mathrm{MB}(h,e_1)) \\ &= \mathrm{MB}(h,e_1) + \mathrm{MB}(h,e_2) - \mathrm{MB}(h,e_2)\cdot\mathrm{MB}(h,e_1) \end{aligned}$$

  Als hier MB wordt beschouwd als functie voor het tellen van elementen, en **co** en $\cdot$ stellen respectievelijk $\cup$ en $\cap$ voor, dan lijkt het veel op dit: $|e_1 \cup e_2| = |e_1| + |e_2| - |e_1 \cap e_2|$, waarbij het telprincipe in de wiskunde de principe van inclusie en exclusie heet. Er is ook een overeenkomst met kansen in kanstheorie: $P(e_1 \cup e_2) = P(e_1) + P(e_2) - P(e_1 \cap e_2)$

- Buchanan en Shortliffe zijn met de functie nog wel voorzichtig geweest. De functie is eerst getest. Er werd een vergelijking gemaakt tussen twee zekerheidsfactoren, die hieronder staan:

  $CF^{\star}(h,e)$ = de berekende CF waarbij de definitie van CF in paragraaf 2.4.1 werd gebruikt (dus, er wordt gewerkt met "perfecte kennis", $P(h|e)$ en $P(h)$ zijn bekend);

  $CF(h,e)$ = de berekende CF met gebruikmaking van de functie f_{co} en de bekende MB's en MD's voor elke e_i waaruit e is samengesteld (dus $P(h|e)$ is niet bekend, maar wel $P(h|e_i)$ en $P(h)$ voor de berekening van $MB(h,e)$ en $MD(h,e)$).

  De resultaten zijn tevredenstellend. Er zijn verschillen tussen $CF^{\star}(h,e)$ en $CF(h,e)$, maar die zijn klein. De verschillen worden wel groter, als de functie f_{co} meerdere keren wordt toegepast.
  Wat de gevallen betreft waarin stukken aanwijzingen onderling sterk gerelateerd zijn met betrekking tot de beschouwde hypothese (dit wordt *voorwaardelijke afhankelijkheid* genoemd), zijn de uitkomsten navenant anders. Dat is ook wel te verwachten. Het geeft precies aan waarom de regel van Bayes zo slecht te gebruiken is.
  Hier zijn dus niet alle problemen zoals die zich bij de regel van Bayes in de zuivere vorm voordoen, de wereld uit. Het gaat echter om een nieuw kwantificatieschema. Er zijn ongeveer net zoals bij de subjectieve Bayesische methode aannamen gemaakt, dus kan er toch gebruik van gemaakt worden.

- Het Certainty Factor model stelt een voorwaarde aan regels. Net als bij de regel van Bayes moet men zich hier aan de voorwaarde van onafhankelijkheid van aanwijzingen houden. Afhankelijke stukken aanwijzingen die verspreid zijn over meerdere regels moeten dus in één enkele regel gestopt worden.
  Maar dat maakt het model onbruikbaar in toepassingen waar grote aantallen observaties moeten worden samengevoegd in één premisse van een regel, om onafhankelijke beslissingscriteria te garanderen.

- De functie f_{co} herkent geen verschijnselen die te maken hebben met wetten in de logica. Als bijvoorbeeld e_1 e_2 impliceert, dan is $CF(h, e_1 \,\mathbf{co}\, e_2) = CF(h,e_2)$ ongeacht wat $CF(h,e_1)$ is. De functie "weet" hier nergens wat van af. Dit maakt duidelijk dat de regels zeer zorgvuldig opgesteld moeten worden.

De bovenstaande opmerkingen maken duidelijk dat voor het toepassen van het Certainty Factor model de regels met zeer veel zorgvuldigheid moeten worden opgesteld. Dat is wat bewerkelijk maar verder geen onoverkomelijk probleem.
Maar hiermee zijn nog niet alle problemen opgelost. Dat is echter ook niet de bedoeling geweest. Het doel is niet de regel van Bayes te verbeteren, maar een model te bieden, waarmee kennis kan worden gerepresenteerd en toegepast, en dit in de speciale gevallen waarin:

- statistische gegevens ontbreken,
- inverse waarschijnlijkheden niet bekend zijn, en
- meestal voorwaardelijke onafhankelijkheid kan worden verondersteld.

2.5 De Dempster-Shafer theorie

In de jaren '60 legde Dempster de basis voor een nieuwe wiskundige theorie van onzekerheid. Shafer ([Sha76]) breidde deze in de jaren '70 uit tot wat nu de Dempster-Shafer theorie heet. Deze kan beschouwd worden als een generalisatie van de waarschijnlijkheidstheorie.
De reden van de ontwikkeling van de theorie is dat waarschijnlijkheidstheorie welbeschouwd niet in staat is onderscheid te maken tussen onzekerheid en onwetendheid, als gevolge van gebrek aan informatie. In de waarschijnlijkheidstheorie moet bij iedere individuele atomaire hypothese de kansen bekend zijn, voordat men überhaupt de relevante kansen kan berekenen. De Dempster-Shafer theorie kent maten van onzekerheid toe aan *verzamelingen* van (disjuncte) hypothesen en

kan uitspraken doen met betrekking tot de onzekerheid van *andere* verzamelingen van hypothesen. Op deze manier is de theorie in staat onderscheid te maken tussen onzekerheid en onwetendheid.

In eerste instantie werd de theorie niet speciaal ontwikkeld voor het redeneren met onzekerheid in expertsystemen. In de jaren '80 kwam men op het idee dat het daarvoor wel geschikt zou kunnen zijn. Onder andere Gordon en Shortliffe hebben de mogelijkheid van toepassing in MYCIN onderzocht ([GS90]).
Met betrekking tot de theorie valt nog wat te melden:

- Omdat de theorie eerst niet voor expertsystemen was bedoeld, zijn er geen combinatiefuncties beschreven. Zij ontbreken hier praktisch volledig;
- De theorie is qua rekenwerk zeer complex. De complexiteit is exponentieel.

Het blijkt dat voor toepassing in expertsystemen enige aanvullingen en modificaties noodzakelijk zijn. Maar zover is het nog niet. Eerst worden wat definities gegeven. De combinatieregel van Dempster wordt besproken en er worden ontbrekende combinatiefuncties beschreven.

2.5.1 De waarschijnlijkheidstoekenning: definities

Om met onzekerheid te kunnen werken wordt eerst een initiële verzameling van hypothesen gecreëerd. Daarna wordt voor elk stukje aanwijzing een mate van onzekerheid geassocieerd met deelverzamelingen van de oorspronkelijke verzameling hypothesen, totdat alle deelverzamelingen die door de combinatie van aanwijzingen worden aangesproken, een mate van onzekerheid hebben. De initiële verzameling van hypothesen wordt de *onderscheidingsverzameling* genoemd. Daarin worden de hypothesen als disjunct verondersteld. Hieronder volgt een aantal definities die betrekking hebben op de verzameling hypothesen (zie ook [LG91]).

Definitie: Zij Θ een onderscheidingsverzameling. Als aan elke deelverzameling $x \subseteq \Theta$ een getal $m(x)$ is toegekend zo dat:

1. $m(x) \geq 0$
2. $m(\emptyset) = 0$
3. $\sum_{x \subseteq \Theta} m(x) = 1$

dan wordt m een *waarschijnlijkheidstoekenning* over Θ genoemd. Voor elke deelverzameling $x \subseteq \Theta$ wordt het getal $m(x)$ de *waarschijnlijkheidsgetal* van x genoemd.

Definitie: Zij Θ een onderscheidingsverzameling en zij m een waarschijnlijkheidstoekenning over Θ. Een verzameling $x \subseteq \Theta$ wordt een *focaal element* in m genoemd als $m(x) > 0$. De *kern* van m, notatie $\kappa(m)$, is de verzameling van alle focale elementen in m.

Let hier op de overeenkomst tussen een waarschijnlijkheidstoekenning m en een kansverdeling P: $m : 2^\Theta \rightarrow [0, 1]$ en $P : \Theta \rightarrow [0, 1]$.

Hier volgen nog meer definities.

Definitie: Zij Θ een onderscheidingsverzameling en zij m een waarschijnlijkheidstoekenning over Θ. Dan is de *geloofwaardigheidsfunctie* die met m correspondeert de functie Bel: $2^\Theta \rightarrow [0, 1]$ die als volgt is gedefinieerd:

$$\text{Bel}(x) = \sum_{y \subseteq x} m(y)$$

voor elke $x \subseteq \Theta$.

Het verschil tussen Bel en m is dat $m(x)$ een vertrouwen in precies de verzameling x uitdrukt, terwijl Bel(x) het totale vertrouwen aangeeft in x en de deelverzamelingen daarvan.
De functie Bel heeft de volgende eigenschappen:

1. $\mathrm{Bel}(\Theta) = 1$ omdat $\sum_{y \subseteq x} m(y) = 1$
2. Voor alle $x \subseteq \Theta$ die uit precies één element bestaat, geldt: $\mathrm{Bel}(x) = m(x)$
3. Voor elke $x \subseteq \Theta$ geldt: $\mathrm{Bel}(x) + \mathrm{Bel}(\bar{x}) \leq 1$, omdat

$$\begin{aligned} \mathrm{Bel}(\Theta) &= \mathrm{Bel}(x \cup \bar{x}) \\ &= \mathrm{Bel}(x) + \mathrm{Bel}(\bar{x}) + \sum_{\substack{x \cap y \neq \emptyset \\ \bar{x} \cap y \neq \emptyset}} m(y) \\ &= 1 \end{aligned}$$

Verder geldt nog de ongelijkheid $\mathrm{Bel}(x) + \mathrm{Bel}(y) \leq \mathrm{Bel}(x \cup y)$ voor alle $x, y \in \Theta$.

Er zijn twee speciale gevallen van geloofwaardigheidsfuncties:

1. Stel men heeft een waarschijnlijkheidstoekenning die er zo uit ziet:

$$m(x) = \begin{cases} 1 & \text{als } x = \Theta \\ 0 & \text{anders} \end{cases}$$

Zo'n waarschijnlijkheidstoekenning beschrijft hier een gebrek aan aanwijzingen, oftewel onwetendheid wordt hier uitgedrukt. De geloofwaardigheidsfunctie wordt hier een *lege geloofwaardigheidsfunctie* genoemd.

2. Stel men heeft een waarschijnlijkheidstoekenning die er zo uit ziet:

$$m(x) = \begin{cases} 1 - c_l & \text{als } x = \Theta \\ c_l & \text{als } x = A \\ 0 & \text{anders} \end{cases}$$

Hierin is $a \subset \Theta$ en $0 < c_l < 1$ een constante. De geloofwaardigheidsfunctie die met deze waarschijnlijkheidstoekenning correspondeert heet hier een *enkelvoudige ondersteuningsfunctie* (*simple support function*).

De geloofwaardigheidsfunctie geeft voor alle verzamelingen x slechts een ondergrens met betrekking tot de huidge geloof in x aan. Maar er zijn ook verzamelingen y waarvoor $x \subseteq y$ en waaraan geloof is toegekend. Daarvoor is er een andere functie die hieronder wordt gedefinieerd.

Definitie: Zij Θ een onderscheidingsverzameling en zij m een waarschijnlijkheidstoekenning over Θ. Dan is de *aannemelijkheidsfunctie* die met m correspondeert de functie Pl: $2^{\Theta} \rightarrow [0, 1]$ die als volgt is gedefinieerd:

$$\mathrm{Pl}(x) = \sum_{x \cap y \neq \emptyset} m(y)$$

voor elke $x \subseteq \Theta$.

Semantisch gezien is Pl een indicatie voor het vertrouwen dat *niet* aan $\bar{x}$ is toegekend. Pl is dus de bovengrens.

Definitie: Zij Θ een onderscheidingsverzameling en zij m een waarschijnlijkheidstoekenning over Θ. Zij Bel de geloofwaardigheidsfunctie en Pl de aannemelijkheidsfunctie die beiden met m corresponderen. Voor elke $x \subseteq \Theta$ wordt het gesloten interval [Bel(x), Pl(x)] de *geloofwaardigheidsinterval* van x genoemd.

Wat het verschil Pl(x) – Bel(x) betreft, deze geeft het vertrouwen in de verzamelingen y waarvoor geldt $y \subseteq x$, weer. Dit geeft dan ook de onzekerheid met betrekking tot x weer.

2.5.2 De combinatieregel van Dempster

2.5.2.1 Een definitie

In de Dempster-Shafer theorie is een functie gegeven dat uit twee aanwijzingen met de bijbnehorende waarschijnlijkheidstoekenningen een nieuwe waarschijnlijkheidstoekenning berekent voor de gecombineerde invloed van deze aanwijzingen. Een definitie van deze functie wordt hieronder gegeven.

Definitie: (*combinatieregel van Dempster*)
Zij Θ een onderscheidingsverzameling en zij m_1 en m_2 waarschijnlijkheidstoekenningen over Θ. Dan is $m_1 \oplus m_2$ een functie $m_1 \oplus m_2 : 2^\Theta \rightarrow [0,1]$ die als volgt is gedefinieerd:

1. $m_1 \oplus m_2(\emptyset) = 0$,
2. $m_1 \oplus m_2(x) = \dfrac{\sum_{y \cap z = x} m_1(y) \cdot m_2(z)}{\sum_{y \cap z \neq \emptyset} m_1(y) \cdot m_2(z)}$ voor alle $x \neq \emptyset$.

$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is de functie $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 : 2^\Theta \rightarrow [0,1]$ die als volgt is gedefinieerd:

$$\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 = \sum_{y \subseteq x} m_1 \oplus m_2(y)$$

Hoe nou precies de combinatieregel werkt, wordt met een onderstaande voorbeeld getoond.

Voorbeeld 2.5.1
Zij Θ = {*hartaanval,pericarditis,longembolie, aneurysma-dissecans*}. *Beschouw de volgende waarschijnlijkheidstoekenningen over* Θ*:*

$$m_1(x) = \begin{cases} 0.6 & \text{als x} = \Theta \\ 0.4 & \text{als x} = \{hartaanval,pericarditis\} \\ 0 & anders \end{cases}$$

$$m_2(x) = \begin{cases} 0.5 & \text{als x} = \Theta \\ 0.5 & \text{als x} = \{longembolie\} \\ 0 & anders \end{cases}$$

Na toepassing van de combinatieregel van Dempster wordt er een nieuw waarschijnlijkheidstoekenning $m_1 \oplus m_2$ *verkregen, die het gecombineerde effect van* m_1 *en* m_2 *weergeeft. De resultaten staan in een doorsnedetabel in figuur 2.5 op bladzijde 28. Dit is echter geen goede tabel voor* $m_1 \oplus m_2$*. Daarin is* $m_1 \oplus m_2(\emptyset) > 0$*, en dat is niet volgens de definitie van de combinatieregel.* $m_1 \oplus m_2(\emptyset)$ *wordt 0 gemaakt en het getal 0.2 wordt verdeeld over de overige berekende waarschijnlijkheidsgetallen. De verdeelsleutel vormt de formule*

$$\sum_{y \cap z \neq \emptyset} m_1(y) \cdot m_2(z)$$

die als normalisatiefactor in de noemer staat. Het resultaat is een gecorrigeerde doorsnedetabel die in figuur 2.6 op bladzijde 28 is gegeven. □

2.5.2.2 Eigenschappen

Er is al eerder gezegd dat de Dempster-Shafer theorie wel eens geschikt kon zijn voor een toepassing in expertsystemen. De heren Gordon en Shortliffe ([GS90]) hebben de mogelijkheid van toepassing in het systeem MYCIN onderzocht. Een deel van dat onderzoek betrof een vergelijking van de combinatiefunctie in MYCIN met de combinatieregel van Dempster. Merk hier op dat niet eerst is vastgesteld dat de combinatieregel geschikt is als combinatiefunctie voor co-concluderende aanwijzingen. Dat komt nog, tot dan wordt dat verondersteld.
Er zijn drie categorieeën van gevallen vastgesteld waarin twee regels samen tot een conclusie komen.

m_2 / m_1	$\cdots$	$\{longembolie\}$ (0.5)	$\cdots$	Θ (0.5)
$\cdots$				
$\{hartaanval, pericarditis\}$ (0.4)		$\emptyset$ (0.2)		$\{hartaanval, pericarditis\}$ (0.2)
$\cdots$				
Θ (0.6)		$\{longembolie\}$ (0.3)		Θ (0.3)

Figuur 2.5: Een *foutieve* doorsnedetabel voor m_1 en m_2

m_2 / m_1	$\cdots$	$\{longembolie\}$ (0.5)	$\cdots$	Θ (0.5)
$\cdots$				
$\{hartaanval, pericarditis\}$ (0.4)		$\emptyset$ (0)		$\{hartaanval, pericarditis\}$ (0.25)
$\cdots$				
Θ (0.6)		$\{longembolie\}$ (0.375)		Θ (0.375)

Figuur 2.6: De *correcte* doorsnedetabel voor m_1 en m_2

Categorie 1 : Twee regels bevestigen of ontkennen beide een en dezelfde hypothese. In dit geval levert zowel de combinatieregel van Dempster als de CF-combinatiefunctie in MYCIN hetzelfde resultaat. Dat klopt ook wel: aan de hypothese als een deelverzameling van Θ dat bovendien uit precies één element bestaat, wordt een waarschijnlijkheidsgetal toegekend en het andere getal wordt toegekend aan Θ. De CF-combinatiefunctie is hier een speciaal geval van de combinatieregel van Dempster.

Categorie 2 : Een regel leidt tot ontkenning en de andere regel leidt tot bevestiging van een en dezelfde hypothese. In dit geval komt een aspect van de combinatieregel naar voren. Een aanwijzing die een deelverzameling A van Θ steunt, zorgt voor een verkleinend effect op het vertrouwen in elke deelverzameling die met A disjunct is.
Stel $A = \{streptococcus\}$, $s_1 = m_1(A)$, en $s_2 = m_2(A^c)$. Dan is:

$$\begin{aligned} m_1 \oplus m_2(A) &= \frac{s_1(1-s_2)}{1-s_1s_2} \\ m_1 \oplus m_2(A^c) &= \frac{s_2(1-s_1)}{1-s_1s_2} \\ m_1 \oplus m_2(\Theta) &= \frac{(1-s_1)(1-s_2)}{1-s_1s_2} \end{aligned}$$

$(1-s_1s_2)$ is hier de normalisatiefactor, omdat anders $m_1 \oplus m_2(\emptyset) > 0$ is.
Het getal s_1 wordt met factor $(1-s_2)/(1-s_1s_2)$ vermenigvuldigd en het getal s_2 met factor $(1-s_1)/(1-s_1s_2)$. Beide factoren zijn kleiner of gelijk aan 1. Hieruit valt af te leiden dat toepassing van de combinatieregel leidt tot een "verkleinde" steun voor *zowel* A als A^c. De CF-combinatiefunctie leidt hier maar tot een reductie van één van de CF's, de CF die absoluut gezien het grootst is.
Het verschil is het duidelijkst te zien als in het voorbeeld hierboven geldt dat $m_1(A) = s$ en $m_2(A^c) = s$. De CF-combinatiefunctie levert dan CF $= 0$, terwijl de combinatieregel van Dempster als resultaat heeft: $s(1-s)/(1-s^2) = s/(1+s)$ voor zowel A als A^c. Gordon en Shortliffe menen hier dat de combinatieregel meer realistisch de effecten van twee tegenstrijdige aanwijzingen weergeeft. Het is beter om aan te geven dat een hypothese tot de mogelijkheden behoort, dan dat een hypothese niet meedoet omdat de tegenstrijdige aanwijzingen daartoe geen aanleiding geven.

Categorie 3 : Twee regels hebben betrekking op twee verschillende hypothesen, die beiden in een en dezelfde onderscheidingsverzameling Θ zitten. Een voorbeeld: zij

$$m_1(x) = \begin{cases} 0.4 & \text{als } x = \{Staphylococcus\} \\ 0.6 & \text{als } x = \Theta \\ 0 & \text{anders} \end{cases}$$

$$m_3(x) = \begin{cases} 0.7 & \text{als } x = \{Streptococcus\}^c \\ 0.3 & \text{als } x = \Theta \\ 0 & \text{anders} \end{cases}$$

Wat $m_1 \oplus m_3$ dan wordt, is gegeven in de doorsnedetabel hieronder.

		m_3	
		$\{Streptococcus\}^c$ (0.7)	Θ (0.3)
m_1	$\{Staphylococcus\}$ (0.4)	$\{Staphylococcus\}$ (0.28)	$\{Staphylococcus\}$ (0.12)
	Θ (0.6)	$\{Streptococcus\}^c$ (0.42)	Θ (0.18)

Voor de combinatie waren de geloofwaardigheidsintervallen $[0.4, 1]$ respectievelijk $[0.7, 1]$ en na combinatie zijn ze $[0.4, 1]$ respectievelijk $[0.82, 1]$ (ga zelf na). Dus het bevestigen

van $\{Staphylococcus\}$ leidt ook tot de bevestiging van $\{Streptococcus\}^c$, een superset van $\{Staphylococcus\}$. Omgekeerd heeft de bevestiging van $\{Streptococcus\}^c$ geen invloed op $\{Staphylococcus\}$, een deelverzameling van $\{Streptococcus\}^c$.
Deze categorie is niet van toepassing voor de CF-combinatiefunctie, omdat deze maar voor slechts één hypothese tegelijkertijd een CF berekent.

2.5.3 De combinatiefuncties

De Dempster-Shafer theorie geeft geen definities voor combinatiefuncties. In eerste instantie was de theorie ook niet bedoeld voor toepassing in een expertsysteem. Dus ontbreken ze hier (bijna) volledig. In [LG91] is aangegeven dat de theorie door M. Ishizuka is toegepast in het systeem SPERIL. Daar zijn combinatiefuncties gebruikt, die hier worden behandeld (zie [LG91]).

Er is nu nog het probleem dat de Dempster-Shafer theorie voor een regel **if** e **then** h **fi** niet expliciet aangeeft wat voor informatie met de hypothese h moet worden geassocieerd. Het ligt voor de hand om er een waarschijnlijkheidstoekenning aan te hangen. Voor bijvoorbeeld de regel **if** e_1 **then** h **fi** wordt de waarschijnlijkheidstoekenning:

$$m_{e_1} = \begin{cases} 1 - c_1 & \text{als } x = \Theta \\ c_1 & \text{als } x = \{h\} \\ 0 & \text{anders} \end{cases}$$

De geloofwaardigheidsfunctie Bel_{e_1} die met m_{e_1} correspondeert is hier een enkelvoudige ondersteuningsfunctie (simple support function). Zoals hierboven beschreven wordt de situatie zo:

$$e_1 \xrightarrow{\quad m_{e_1} \quad} h$$

Men weet van de waarschijnlijkheidstheorie dat de volgende combinatiefuncties nodig zijn: f_{co}, f_{and}, f_{or} en f_{prop}. Deze functies worden hier puntsgewijs behandeld.

De functie f_{co}

Stel we hebben weer de regel **if** e_1 **then** h **fi** met de bijbehorende functies m_{e_1} en Bel_{e_1} als hierboven. Stel er is nu een tweede regel **if** e_2 **then** h **fi** die ook betrekking heeft op de hypothese h. Voor deze regel is de volgende waarschijnlijkheidstoekenning gegeven:

$$m_{e_2} = \begin{cases} 1 - c_2 & \text{als } x = \Theta \\ c_2 & \text{als } x = \{h\} \\ 0 & \text{anders} \end{cases}$$

Als we nu aannemen dat e_1 en e_2 met absolute zekerheid zijn vastgesteld, dan kan hier als combinatiefunctie de combinatieregel van Dempster gebruikt worden om de waarschijnlijkheidstoekenning $m_{e_1} \oplus m_{e_2}$ voor de h gebaseerd op e_1 en e_2 te berekenen.

De functie f_{prop}

Als in een regel **if** e **then** h **fi** de aanwijzing e niet met absolute zekerheid vastgesteld wordt, dan moet deze onzekerheid naar h doorwerken. Zie het plaatje hieronder waarin $\mathrm{Bel}_{e'}$ de mate van onzekerheid is waarmee e is bevestigd:

$$e' \xrightarrow{\quad \mathrm{Bel}_{e'}(e) \quad} e \xrightarrow{\quad m_e \quad} h$$

In deze situatie moet $\mathrm{bel}_{e'}(h)$ berekend worden. Dit kan geschieden nadat $m_{e'}(h)$ is berekend, en wel als volgt:

$$m_{e'}(h) = m_e(h) \cdot \mathrm{Bel}_{e'}(e)$$

Dit levert dus een combinatiefunctie voor het propageren van onzekerheid van aanwijzingen op.

De functies f_{or} en f_{and}

De combinatiefuncties voor samengestelde aanwijzingen zijn recht toe recht aan en zijn dus als volgt:

$$\begin{aligned} \mathrm{Bel}_{e'}(e_1 \textbf{ and } e_2) &= \min\{\mathrm{Bel}_{e'}(e_1),\, \mathrm{Bel}_{e'}(e_2)\} \\ \mathrm{Bel}_{e'}(e_1 \textbf{ or } e_2) &= \max\{\mathrm{Bel}_{e'}(e_1),\, \mathrm{Bel}_{e'}(e_2)\} \end{aligned}$$

2.5.4 Tot slot

De laatste drie combinatiefuncties zien er simpel uit maar ze zijn nog lang niet goed. In [LG91] wordt wel aangegeven dat er andere benaderingen zijn, maar die zijn ook niet (veel) beter.
De invullingen voor de laatste drie combinatiefuncties komen van M. Ishizuka, die ze in het systeem SPERIL heeft toegepast. Juist deze functies werden besproken, vanwege de grote overeenkomsten met functies in andere theorieën.

Een aspect is niet behandeld, maar al wel genoemd en dat is dat de combinatieregel qua berekeningen exponentiële tijd kan vergen, zeker als het gaat om grote onderscheidingsverzamelingen en veel focale elementen. Onder de voorwaarde dat de geloofwaardigheidsfuncties enkelvoudige ondersteuningsfuncties zijn, kan de rekentijd polynomiaal worden. In [GS90] wordt een schema gegeven waarmee lineaire rekentijd bewerkstelligd kan worden. Die schema is aangepast voor gebruik in het systeem MYCIN, maar er is verwezen naar [Bar81] voor de oorspronkelijke versie.

2.6 Netwerkmodellen

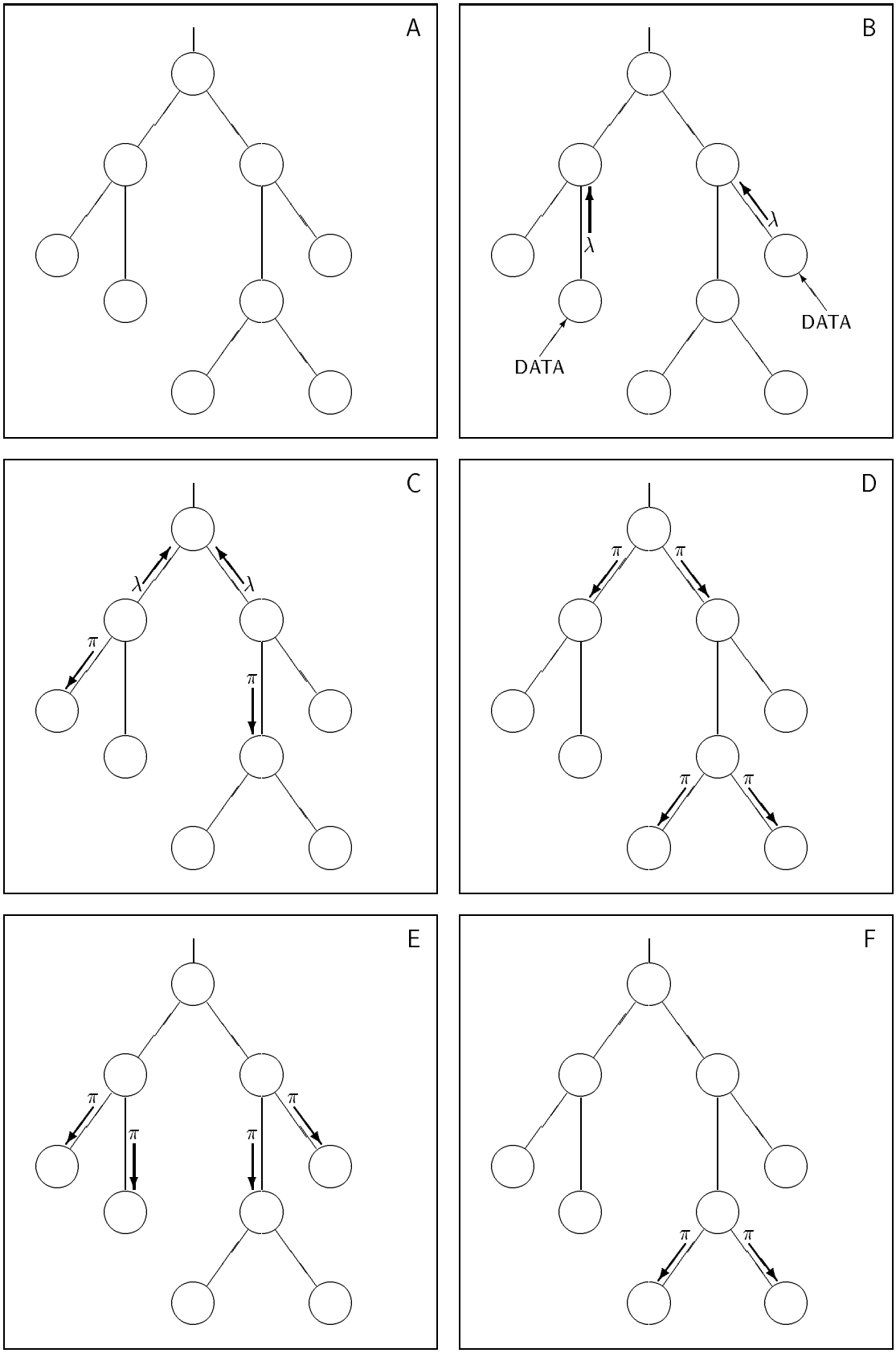
De nieuwste trend dat in de jaren '80 is opgekomen, zijn netwerkmodellen voor het gebruik in expertsystemen ten behoeve van het redeneren met onzekerheid. Netwerkmodellen zijn gebaseerd op graph-structuren, waarin de lijnen of pijlen de relatie tussen een tweetal punten aangeeft, en waarin de punten een aanwijzing of hypothese of een combinatie daarvan weergeven. Deze netwerkmodellen hebben een aantal karakteristieke eigenschappen ([LG91]):

1. Om een aanwijzing te propageren, dient het grafische deel van een *vertrouwensnetwerk* min of meer als een architectuur voor berekeningen.
2. Nadat een aanwijzing is verwerkt, verkrijgt men weer een vertrouwensnetwerk. Dus het vertrouwensnetwerk blijft invariant onder het propageren van aanwijzingen. Hierdoor is een recursieve verwerking van aanwijzingen mogelijk.
3. Elke keer als er een nieuwe aanwijzing is, wordt er een nieuwe "kansverdeling" berekend. Aan elke knoop in het netwerk worden getallen vernieuwd.

Er zijn twee modellen bekend ([LG91]). Ze worden hier slechts kort behandeld, daar ze voor LISA-D niet geschikt zijn. In LISA-D is recursie zoals dat hierboven beschreven is, niet mogelijk met multisets. En er wordt steeds een nieuwe "kansverdeling" berekend, waarbij de "kansen" bij de knopen of punten worden vernieuwd. Dit laatste verlaagt het niveau van abstractie (zie de eisen in paragraaf 1.3).

2.6.1 Het netwerkmodel van Kim en Pearl

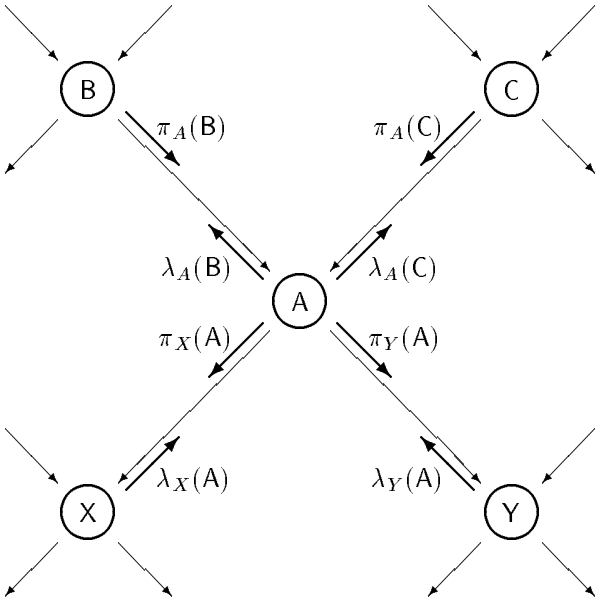
Het netwerkmodel van Kim en Pearl ([KP83]) is gebaseerd op een gerichte graph. De structuur van de graph is een boom met meervoudige vertakkingen, zogenaamd *polytree*. Als er knopen zijn die samen een gemeenschappelijke knoop hebben, dan is dat geen probleem als er voldaan is aan de eis dat er slechts één pad tussen elke twee punten is.
In de graph representeren de knopen de kennis in de vorm van hypothesen, aanwijzingen of iets dergelijks. Een pijl tussen twee knopen, bijvoorbeeld A $\longrightarrow$ B, geeft een relatie tussen die twee weer: A is van invloed op B. A wordt hier dan als "directe oorzaak" van B bedoeld.



Figuur 2.7: Een demonstratie van hoe een nieuwe aanwijzing door het netwerk wordt gepropageerd. Het netwerk dient als architectuur voor de "berichtenuitwisseling".

De graph wordt hier gebruikt als een structuur om aanwijzingen er doorheen te propageren. Elke knoop wordt als een processor beschouwd, die parameters genereert en deze 'verstuurt' naar zowel zijn 'ouders' (de λ-parameters) als naar zijn nakomelingen (de π-parameters). Verder worden er vaste voorwaardelijke kansen die aan de relatie van een knoop met zijn ouders toegekend, en dit voor elke knoop. Deze worden in een matrix bewaard en veranderen niet. Aan de hand van deze gegevens wordt er per knoop een vertrouwen berekend die bij die knoop hoort.

Hoe propagatie in zijn werk gaat is in stappen in figuur 2.7 uitgebeeld. Figuur 2.7a geeft de stabiele situatie weer. Er zijn geen nieuwe aanwijzingen die het vertrouwen in de bladeren beïnvloeden. Zogauw er enkele blaadjes worden "geactiveerd" door nieuwe gegevens (zie figuur 2.7b), dan produceren ze nieuwe λ-parameters die naar hun ouders worden gestuurd. In figuur 2.7c reageren de ouders erop door nieuwe λ-parameters te produceren en die naar hun eigen ouders te sturen. Ook worden er nieuwe π-parameters gemaakt voor die nakomelingen waarvan ze **niet** de nieuwe λ-parameters hebben ontvangen. Dus als een nakomeling een nieuw λ-parameter heeft gestuurd, dan krijgt deze **niet** een nieuwe π-parameter terug.
Dan krijgt de wortel nieuwe λ-parameters waarop deze nieuwe π-parameters terugstuurt (zie figuur 2.7d). Het proces gaat hierna net zo lang door totdat alle parameters verwerkt zijn (figuur 2.7e en f).
Figuur 2.8 toont in detail in welke richtingen de parameters lopen. Het verschil met figuur 2.7 is



Figuur 2.8: Een knoop uit een netwerk, waarin alle mogelijke parameters zijn weergegeven die kunnen worden uitgewisseld. Hier heeft knoop A meerdere ouders.

nog dat er hier nu met meervoudige ouders wordt gewerkt.

Waar nog niet over gesproken is, maar dat nu goed uitgelegd kan worden, is wat de knopen nu precies representeren. De bladeren stellen aanwijzingen uit de buitenwereld. Van die bladeren zijn de ouders representanten voor hypothesen en hun ouders stellen een combinatie van die hypothesen voor, enzovoort.
Om enigzins voor te stellen hoe het werkt, wordt hier een voorbeeld gegeven (uit [Pea90]).

> *Meneer Holmes werd op zijn werk gebeld door zijn buurvrouw, die hem waarschuwde dat zij een inbraakalarm vanuit de richting van zijn huis hoorde. Als hij zich gereedmaakt om snel huiswaarts te gaan, herinnert Holmes zich dat onlangs het alarm was afgegaan door een aardbeving. Naar huis rijdend hoort hij het radiobericht waarin een aardbeving op 200 mijl afstand wordt gemeld.*

Holmes beschouwt twee episoden die mogelijk potentiële oorzaken zijn voor het afgaan van het alarm, een poging tot inbraak en een aardbeving. Ook al kan veilig worden aangenomen dat inbraken onafhankelijk zijn van aardbevingen, het radiobericht verkleint nog de mogelijkheid van inbraak. Het bovenstaand voorbeeld geeft een idee van wat er achter het model zit, namelijk het menselijk redeneren. Details over het model zijn te vinden in [Pea86] of [Pea90]. Een samenvatting staat in [KP83].

2.6.2 Het netwerkmodel van Lauritzen en Spiegelhalter

Het netwerkmodel van Kim en Pearl en andere modellen hebben de eigenschap dat in de onderliggende graph de propagatie van een aanwijzing tegen de richting van de pijlen in geschiedt. Dit bracht Lauritzen en Spiegelhalter ertoe om een nieuw netwerkmodel te ontwikkelen ([LS90] of [LS88]). Het netwerkmodel heeft enkele kenmerken:

- Het grafische deel ontstaat uit een transformatie van een grafische deel van een een of ander vertrouwensnetwerk.
- Het schema voor het propageren van een aanwijzing is op deze nieuwe representatie gebaseerd en is strikt probabilistisch.

Het transformatieproces en het propagatieproces wordt hier besproken. Hoe het vernieuwen van de waarschijnlijkheden gebeurt wordt buiten beschouwing gelaten. De representatie van de kansverdeling wordt in het kort wel behandeld. Het netwerkmodel is deels gebaseerd op de grafentheorie, daarom worden er eerst een paar definities gegeven.

Definitie: Zij $G = (V(G), E(G))$ een ongerichte graph met $E(G)$ een eindige verzameling van lijnen (V_i, V_j), $V_i, V_j \in V(G)$.
Een *cykel* is een pad van tenminste lengte 1 van V_0 naar V_0, $V_0 \in V(G)$.
Een cykel is *elementair* als al zijn punten verschillend zijn.
Een *koord* van een elementaire cykel $V_0, V_1, \ldots, V_k = V_0$ is een lijn (V_i, V_j), $i = (j \pm 1) \bmod (k+1)$.

Definitie: Een ongerichte graph is *decomponeerbaar* als alle elementaire cykels van lengte $k \geq 4$ een koord bevatten.

Er moet van een oorspronkelijk vertrouwensnetwerk het grafische deel worden getransformeerd in een decomponeerbare graph. Dat het een decomponeerbare graph moet zijn, is vanwege de constructie van een kansverdeling en het propagatieproces.
Het grafische deel van een vertrouwensnetwerk is meestal een acyclische gerichte graph. De transformatie van zo'n graph G gaat als volgt:

1. Voeg aan de graph G lijnen toe, zodanig dat van elke punt geen van zijn directe voorgangers "alleen" is. Dat wil zeggen: "trouw" van elk punt diens ouders;
2. Laat de richtingen van de pijlen vervallen, ofwel maak de graph ongericht;
3. Voeg tot slot aan elke elementaire cykel van tenminste lengte 4 een koord toe.

Dit levert een decomponeerbare graph op (die niet uniek is).

Voor een te bepalen bijbehorende representatie van een kansverdeling wordt er hier gebruik gemaakt van klieken. De definitie wordt hieronder gegeven:

Definitie: Zij $G = (V(G), E(G))$ een ongerichte graph. Een *kliek* van G is een subgraph $H = (V(H), E(H))$ van G zo dat voor elke twee verschillende punten $V_i, V_j \in V(H)$ geldt dat $(V_i, V_j) \in E(H)$.
Een kliek H van G wordt *maximaal* genoemd als er in G geen kliek H' bestaat dat van H verschilt zo dat H een subgraph van H' is.

Om een nieuwe representatie van de kansverdeling te verkrijgen moeten de punten en de klieken van een decomponeerbare graph geordend worden. De punten worden genummerd volgens het principe van de zogenaamde *maximum cardinality search*. Dit gaat als volgt:

1. Kies een van de punten en geef hem het nummer 1;
2. Nummer de resterende punten in oplopende volgorde; het eerstvolgende nummer wordt steeds toegekend aan het punt met de meeste al genummerde buren.

Hierna worden de klieken in de graph achtereenvolgens genummerd in de volgorde van het hoogst genummerde punt. Door een eigenschap dat de klieken door deze ordening met elkaar hebben die bekend staat als het *running intersection property* kan een kansverdeling op de hele graph bepaald worden uit de zogenaamde *marginale kansverdelingen* op de klieken en hun doorsneden. Marginale kansverdelingen zijn locale kansverdelingen: ze zijn slechts geassocieerd met de klieken en hun punten daarin.

De decomponeerbare graph en de bijbehorende kansverdeling vormen samen een *decomponeerbare vertrouwensnetwerk*. Voordat er verder wordt gegaan, merken we nog op dat dit netwerk maar één keer gemaakt hoeft te worden. Er wordt met kopieën daarvan gewerkt.
Nu er zo'n netwerk is, kan het propagatieproces beschouwd worden. Als er een aanwijzing is dat een van de punten als variabele waar maakt, dan wordt er het volgende gedaan:

1. Nummer opnieuw de punten volgens het principe van de maximum cardinality search, te beginnen bij de door de aanwijzing "geactiveerde" punt;
2. Maak opnieuw een ordening van de klieken zoals dat hierboven is beschreven. Deze ordening geeft de volgorde aan waarin de aanwijzing door de klieken dient te worden gepropageerd;
3. De marginale kansverdelingen worden vernieuwd in dezelfde volgorde als de aanwijzing door de klieken wordt gepropageerd;
4. Na vernieuwing wordt het "geactiveerde" punt uit de graph verwijderd.

Bij nog een nieuwe aanwijzing wordt het bovenstaande procedé nog eens uitgevoerd op de resterende graph. De graph wordt zo steeds verkleind.

Dit is slechts een summiere beschrijving van het netwerkmodel van Lauritzen en Spiegelhalter. Er is niet beschreven wat die marginale kansverdelingen zijn, hoe ze vernieuwd worden en hoe daarmee een kansverdeling op de hele graph wordt bepaald. De geïnteresseerde lezer wordt voor een gedetailleerde uitleg verwezen naar [LS88] of [LS90]. Een minder formele en minder gedetailleerde versie die echter goed leesbaar is, staat in [LG91].

Hoofdstuk 3

Relevanties in information retrieval

3.1 Inleiding

De kern van een information retrieval systeem is het vergaren van bruikbare documenten in overeenstemming met de informatiebehoeften van de gebruiker. Er komt heel wat bij kijken om zo'n systeem te ontwikkelen. Dit wordt duidelijk als het information retrieval systeem wordt vergeleken met een conventionele database systeem met betrekking tot de verschillen tussen deze systemen. De belangrijkste verschillen zijn:

- De kennisrepresentatie. In een database systeem zijn de objecten waarmee gewerkt wordt tupels, een verzameling van attribuutwaarden. In een information retrieval systeem zijn deze objecten *documenten*. In dit systeem wordt niet rechtstreeks met documenten gewerkt, maar met representanten daarvan. Er moet voor information retrieval dan een geschikte methode worden ontworpen om de documenten te karakteriseren. Een geschikte kennisrepresentatie wordt gezocht. Een "perfecte" representatie is nog niet gevonden. Tot nu toe voldoet het principe van trefwoorden of sleutelwoorden het best.
- De formulering van de vraag (*query*). In een database systeem kunnen vragen zo precies worden geformuleerd dat deze de informatiebehoeften van de gebruiker volledig dekt. Het resultaat zal zijn bij elkaar vergaarde tupels die exact voldoen aan de selectiecriteria. Het is absoluut zeker dat de tupels relevant zijn voor de vraag.
  Daarentegen is in information retieval een vraag lang niet zo precies te formuleren dat deze de informatiebehoeften van de gebruiker weergeeft. Het probleem is hier hoe de informatiebehoefte te representeren. Dit is net zo moeilijk als een representatie te vinden voor documenten. Er zijn wel vraagtalen ontwikkeld die een verzoek vertaalt, maar die zijn niet krachtig genoeg dat de vertaling ook de informatiebehoefte weergeeft.
  En wat de relevantie betreft: er is geen maat om vast te stellen in hoeverre een document relevant is voor een verzoek. Naar een geschikt criterium voor de bepaling van de relevantie moet worden gezocht.

Er is veel onderzoek gedaan naar methoden om relevanties te bepalen. Talloze afkeidingsmodellen zijn ontworpen. afkeidingsmodellen omdat bepaling van de relevanties samenhangt met de gebruikte afkeidingstechniek. In dit hoofdstuk worden enkele modellen besproken:

- Het afleidingsnetwerk voor document retrieval van Turtle en Croft ([TC90]);
- Het model van Wong en Yao ([WY91]).
- Het index expressie vertrouwensnetwerk model van Bruza en Van der Gaag ([BG92]).

3.2 Het probabilistische afleidingsmodel

Het probabilistische afleidingsmodel is ontwikkeld door de heren Wong en Yao ([WY91]). Dit model heeft een aantal kenmerken:

- Het basis voor het model is een Bayesische afleidingsnetwerk (zie [KP83], [Pea86] of [Pea88]). Het model kan hier als een speciaal soort netwerk worden beschouwd. Dit netwerk wordt gekarakteriseerd door de drie lagen die het bevat. Deze lagen corresponderen met queries, documenten en basisconcepten. Dit netwerk is gekozen omdat hierin nog mogelijkheden voor probabilistische afleiding kunnen worden onderzocht. Verder is zo'n netwerk uit te breiden tot meer complexere versies;
- Met betrekking tot waarschijnlijkheden wordt hier een *epistemologische gezichtspunt* gehanteerd. De waarschijnlijkheden worden hier als maten van geloof opgevat. De relevanties worden hierop gebaseerd en wel zo dat een mate van relevantie van een aanwijzing (document) voor een propositie (query) wordt geïnterpreteerd als de maat van geloof dat de aanwijzing de gegeven propositie in een of andere conceptruimte steunt. Hier kan de conceptruimte bijvoorbeeld een verzameling van index termen zijn.
  De tegenhanger van dit gezichtspunt is het *frequentiegezichtspunt* (*aleatorische gezichtspunt*). Bij dit gezichtspunt worden bijvoorbeeld documenten geteld op basis van een bepaalde index term of een ander representant. De keuze voor de epistemologische gezichtspunt wordt nader toegelicht.

Het is de bedoeling geweest dat men door het model een beter begrip krijgt van de concepten die in information retrieval worden gebruikt. Het model is slechts een compleet en coherent raamwerk waarin nog een discussie mogelijk is over vraagstukken als documentrepresentatie en formulering van de vraagstelling.

Er is aangegeven dat voor de waarschijnlijkheden een epistemologische gezichtspunt zal worden gehanteerd. Dit zal hierna worden beargumenteerd. Daarna zal het model uitgebreid worden behandeld.

3.2.1 Het aleatorische versus epistemologische gezichtspunt

Het conventionele probabilistische model is met betrekking tot waarschijnlijkheden gebaseerd op het frequentiegezichtspunt (aleatorische gezichtspunt). In dit model wordt gepoogd de relevantie van een document te meten door documenten met een en dezelfde beschrijving te tellen. Een document wordt beschreven door en binaire vector, $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ waarin $x_i = 0$ of 1 aangeeft of de i-de index term afwezig of aanwezig is. Als men $x_i = 1$ als een label van een gebeurtenis interpreteert en $x_i = 0$ als een ontkenning van de gebeurtenis, dan kan een document worden gezien als een instantie van een conjunctie van gebeurtenissen die corresponderen met de aanwezigheid of afwezigheid van een index term in de document. Wat de bepaalde informatiebehoeften van een gebruiker betreft kan men twee gebeurtenissen onderscheiden, namelijk *relevant* en *niet-relevant*. Men kan inzien dat een document daar ook een instantie van zijn kan. Hierop kan een waarschijnlijkheid van een document met gegeven beschrijving x worden uitgedrukt als $P(relevant|\,\mathrm{x})$. Hier is P een kansverdeling dat is gedefinieerd op de gebeurtenissenruimte die van belang is.

Het zal duidelijk zijn dat de relatie tussen een document en een query afgeleid kan worden aan de hand van de aanwezigheid of afwezigheid van zekere index termen in de document tekst tezamen met de informatie hoe de termen zijn verspreid. Blijft er nu nog over hoe de waarschijnlijkheid bepaald wordt. Het meest voor de hand liggend is de klassieke waarschijnlijkheidstheorie. Echter is bekend dat aan deze theorie haken en ogen zijn:

- Voor een correcte berekening van de waarschijnlijkheden moeten vooraf veel statistische gegevens bekend zijn. Dat is er meestal niet;
- Verder speelt nog het probleem van afhankelijkheden tussen de gebeurtenissen onderling. Om dit te omzeilen worden onafhankelijkheidsaannamen gemaakt wat echter tot gevolg heeft dat het probleemgebied waarin de theorie kan worden toegepast zeer beperkt is.

Elders in de scriptie (in het hoofdstuk over expertsystemen) is de theorie al uitgebreid beschreven. Aanvullende informatie kan daar worden gevonden.

Er worden dus voor de toepassing van de waarschijnlijkheidstheorie strikte onafhankelijkheidsaannamen gemaakt. Dit maakt het model eigenlijk slechts geschikt voor het geval dat documenten binair worden gerepresenteerd. Dit geeft aan dat het model niet bruikbaar is. In dit 'afgekeurde' model is het aleatorische gezichtspunt gehanteerd. Deze biedt kennelijk geen mogelijkheden met betrekking tot verdere ontwikkelingen. Maar hiermee is nog niet gezegd dat het epistemologische gezichtspunt geschikt is. Dat moet nog duidelijk worden. Om aan te geven dat het epistemologische gezichtspunt aanknopingspunten biedt, wordt hier information retrieval vergeleken met patroonherkenning. Patroonherkenning omdat het conventionele probabilistische model deze als uitgangspunt heeft. De vergelijking levert de volgende verschillen tussen patroonherkenning en information retrieval op:

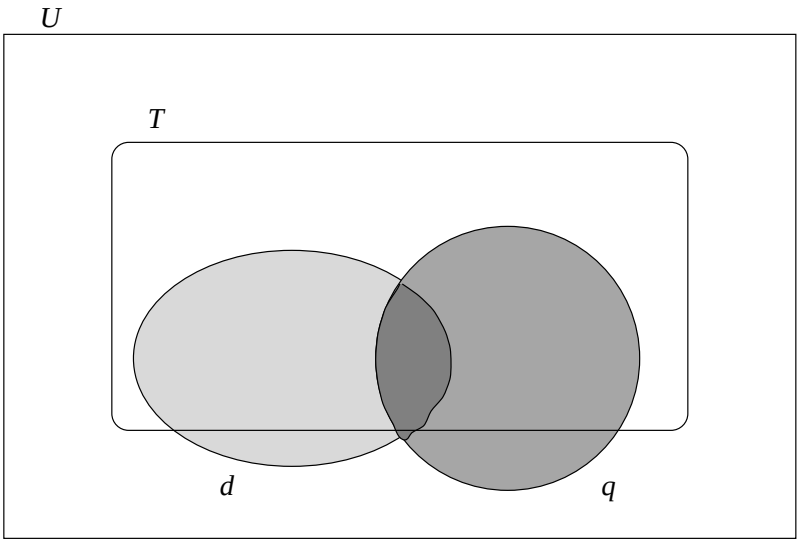
- Bij patroonherkenning speelt de semantiek geen rol terwijl deze in information retrieval een sleutelrol vervult;
- In patroonherkenning zijn patronen in sommige herkenningsproblemen doorgaans onafhankelijk. Deze onafhankelijkheid is een voordeel bij de bepaling van samengestelde waarschijnlijkheden. Deze kan dan redelijkerwijs worden bepaald uit de al bekende waarschijnlijkheden. In information retrieval is van onafhankelijkheid geen sprake. Index termen zijn doorgaans afhankelijk.

Bij patroonherkenning kan men dan wel van een aleatorische gezichtspunt gebruik maken, maar voor information retrieval is het nauwelijks geschikt. Daarentegen wordt bij het epistemologische gezichtspunt rekening gehouden met de semantieken van documenten. Het is een geschikt uitgangspunt waarop het gebruik van waarschijnlijkheidstheorie in information retrieval kan worden onderzocht.

De waarschijnlijkheden die worden gezocht zijn gedefinieerd op basis van de semantische relatie tussen documenten en queries die in een conceptruimte als proposities worden beschouwd. Deze waarschijnlijkheden worden als maten van geloof opgevat. Bijvoorbeeld de kans $P(t|d) = P(d \cap t)/P(d)$ kan worden bepaald door de semantieken van document d en index term t en geeft de mate van geloof weer dat een iemand die indexeert toekent aan de relaties tussen t en d. Op dezelfde manier geeft de kans $P(q|t) = P(q \cap t)/P(t)$ de mate van geloof van de gebruiker in de relatie tussen q en t weer. Dit idee vormt het uitgangspunt voor het model van Wong en Yao. Overigens moet ervoor gewaakt worden dat de beschreven duale concepten van waarschijnlijkheid elkaar niet tegenspreken. Anders gezegd, degene die indexeert moet met betrekking tot de te gebruiken index termen enigzins rekening houden met de vraagstellingen van de gebruiker(s). Tot slot zij nog opgemerkt dat het vanuit het epistemologische gezichtspunt niet uitgesloten is dat de statistische gegevens kunnen worden gebruikt voor het bepalen van de waarschijnlijkheden. Dit kan zinvol zijn, zeker in de gevallen dat kennis wordt gerepresenteerd in termen van gegevens van voorkomen van die kennis.

3.2.2 Het afleidingsmodel in details

In het model worden queries en documenten op een bepaalde manier geïnterpreteerd. Er wordt verondersteld dat er een *ideale* conceptruimte U bestaat waarin de elementen worden beschouwd als *elementaire concepten*. Elke propositie kan in deze conceptruimte opgevat worden als een deelverzameling van concepten (zie figuur 3.1). Het heeft zijn nut om deze representatie te gebruiken. De bekende logische notaties als conjunctie, disjunctie, ontkenning en implicatie worden op deze manier 'vertaald' in termen van doorsnede, vereniging, complement nemen en inclusie. Deze laatste notaties zijn meer toepasselijk in information retrieval. Blijft nu nog over de vraag wat die *conceptruimte* nu inhoudt. Deze kan opgevat worden als een kennisruimte waarin documenten, index termen en queries als deelverzamelingen in voorkomen.
In dit model worden enkele notaties gebruikt die hier worden ingevoerd. Er wordt hierbij verondersteld dat er een kansverdeling P op de conceptruimte U is gedefinieerd. Hoe P is ingevuld, is niet bekend. Bij deze definitie wordt van het epistemologische gezichtspunt uitgegaan.



Figuur 3.1: (i) een ideale conceptruimte U; (ii) een kennisdeelruimte T; (iii) en een document en query representatie

Definitie: Zij H, E proposities in het conceptruimte U. Als E als aanwijzing wordt beschouwd en E impliceert H, dan wordt deze relatie tussen E en H als volgt bepaald:

$$\Psi(E \rightarrow H) \quad =_{\text{def}} \quad P(H \mid E) = \frac{P(H \cap E)}{P(E)} \tag{3.1}$$

Als voor elke deelruimte $K \subseteq U$ het geloof in E dat H steunt slechts is gebaseerd op K, dan wordt dit geloof als volgt gedefinieerd:

$$\Psi(E \rightarrow H \mid K) \quad =_{\text{def}} \quad P(H \mid E \cap K) = \frac{P(H \cap E \cap K)}{P(E \cap K)} \tag{3.2}$$

In de definitie is de deelruimte K het bekende gedeelte binnen binnen de ideale conceptruimte U afhankelijk van bepaalde applicaties waarin dit bekende gedeelte wordt toegepast (denk K als verzameling trefwoorden of iets anders).

Op het gebied van information retrieval wordt een document d (of een query q van een gebruiker) conceptueel gerepresenteerd als een deelverzameling van U zoals dat in figuur 3.1 is weergegeven. Op deze manier kunnen documenten d en queries q als proposities in U als uitkomstenruimte worden beschouwd. Als nu document d als aanwijzing wordt beschouwd, dan wordt de mate waarin d relevant voor q is weergegeven in een *geloofwaardigheidsfunctie* zoals deze in in de definitie (zie vergelijking 3.1) is gedefinieerd:

$$\Psi(d \rightarrow q) = \Psi(d \rightarrow q \mid U) \quad =_{\text{def}} \quad P(q \mid d \cap U) = \frac{P(q \cap d)}{P(d)} \tag{3.3}$$

Daarin is P weer een kansverdeling die op U is gedefinieerd. In deze vergelijking is uit de bovenstaande gegeven definitie $\Psi(d \rightarrow q) = \Psi(d \rightarrow q \mid U)$ af te leiden.
In het kort geeft de functie $\Psi(d \rightarrow q)$ de onzekerheid van de implicatie $d \rightarrow q$ weer. De functie kan in het model in elk geval beschouwd worden als een maat voor de relevantie van document d voor de query q in de ideale conceptruimte U.
Ook al worden documenten en queries beschouwd als deelverzamelingen in de conceptruimte U, over de *ideale* concepten in U is maar heel weinig bekend. Toch is voor de documenten en de queries een representatie nodig om een model voor information retrieval te ontwikkelen. In de praktijk is slechts bekend hoe documenten worden gerepresenteerd in een kennisdeelruimte $T \subseteq U$, dat als een verzameling *basisconcepten* is gedefinieerd. T kan zijn gedefinieerd door een verzameling trefwoorden of zinnen. Als zo'n $T \subseteq U$ is gegeven, dan kan de maat voor de relevantie als volgt

worden uitgedrukt:

$$\begin{aligned}\Psi(d \rightarrow q) &= \Psi(d \rightarrow q \,|\, U) \\ &= P(q \,|\, d \cap U) \\ &= \frac{1}{P(d)} \cdot \left[P(d \cap q \cap T) + P(d \cap q \cap \bar{T})\right] \\ &= P(q \,|\, d \cap T)P(T \,|\, d) + P(q \,|\, d \cap \bar{T})P(\bar{T} \,|\, d) \\ &= \Psi(d \rightarrow q \,|\, T)\Psi(d \rightarrow T) + \Psi(d \rightarrow q \,|\, \bar{T})\Psi(d \rightarrow \bar{T}) \end{aligned} \tag{3.4}$$

In deze vergelijking is $\bar{T} = U \backslash T$. Deze vergelijking geeft duidelijk aan dat men bij de constructie van een kennisrepresentatie een deelruimte T probeert te vinden zó dat $P(T \,|\, d) = 1$ (of $P(\bar{T} \,|\, d) = 0$) voor elk document d. In dit geval is dan

$$\Psi(d \rightarrow q) = \Psi(d \rightarrow q \,|\, T) \tag{3.5}$$

In de praktijk is een constructie van zo'n representatie niet reëel. Van de werkelijkheid is slechts gedeeltelijk kennis beschikbaar. T is dus doorgaans niet volledig, maar partieel. Als gevolg hiervan kan de relevantie van d voor q niet nauwkeurig bepaald worden. Deze relevantie zal wel nauwkeuriger bepaald kunnen worden als de kennis die door T wordt gerepresenteerd toeneemt.
Omdat over U weinig bekend is maar de kennisdeelruimte T daarentegen wel, dan moet de relevantie van d voor q bepaald worden met behulp van kennis dat over T beschikbaar is. $\Psi(d \rightarrow q)$ moet dan geschat worden met behulp van $\Psi(d \rightarrow q \,|\, T)$. Stel $T = t_1 \cup t_2 \cup \ldots \cup t_n$ als verzameling met de t_i's als bekend zijnde basisconcepten. In de wetenschap dat over U weinig bekend is, kan worden aangenomen dat

$$\Psi(d \rightarrow q) \approx \Psi(d \rightarrow q \,|\, T) \tag{3.6}$$

In het algemeen hoeven de basisconcepten niet disjunct te zijn. Dit kan gevolgen hebben voor het model dat hier wordt beschreven. Daarom moeten twee gevallen nader bekeken worden:

1. Alle basisconcepten zijn onderling disjunct, ofwel $t_i \cap t_j = \emptyset$ voor $i \neq j$; of
2. de basisconcepten zijn niet disjunct.

Deze gevallen worden hierna verder uitgewerkt.

3.2.2.1 Disjuncte concepten

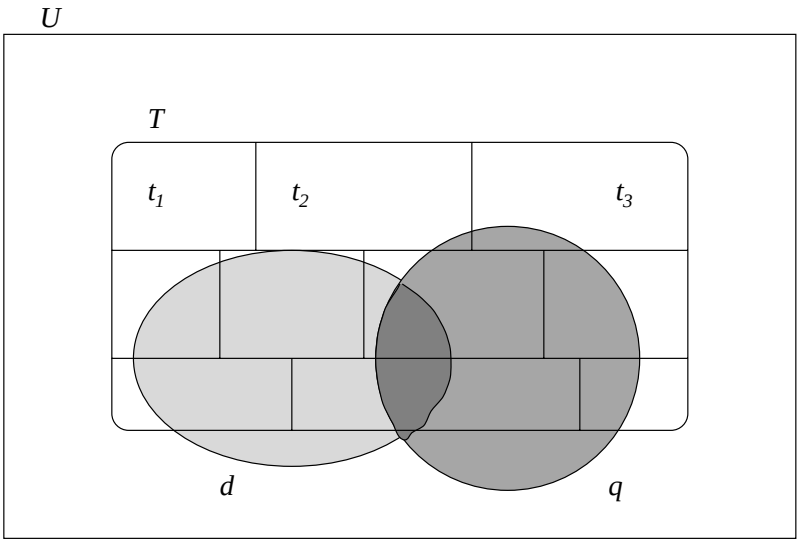
Hier wordt het geval bekeken dat de basisconcepten onderling disjunct worden verondersteld, ofwel $t_i \cap t_j = \emptyset$ voor $i \neq j$. In figuur 3.2 is dit geïllustreerd. De vergelijking 3.6 wordt dan:

$$\begin{aligned}\Psi(d \rightarrow q) &\approx \Psi(d \rightarrow q \,|\, T) \\ &= P(q \,|\, d \cap T) \\ &= \sum_t \frac{P(d \cap q \cap T)}{P(d \cap T)} \end{aligned} \tag{3.7}$$

Hierin wordt over de verzameling basisconcepten de som genomen. Hiermee is het dan mogelijk om de mate van relevantie voor een individueel document te schatten, afhankelijk van de representatie van documenten en queries in de kennisdeelruimte T.
Er zijn twee gevallen waarop documenten en queries in T gerepresenteerd kunnen worden:

1. Stel elk document d wordt gerepresenteerd door de vereniging van enkele t's (dus voor elke t geldt: òf $t \cap d = t$ òf $t \cap d = \emptyset$), ofwel:

$$d = \bigcup_{t \cap d \neq \emptyset} t \tag{3.8}$$



Figuur 3.2: Een kennisdeelruimte T opgesplitst in disjuncte concepten

Ook de query q kan door een vereniging van enkele t's gerepresenteerd worden:

$$q = \bigcup_{t \cap q \neq \emptyset} t \tag{3.9}$$

In de praktijk komt het echter vrijwel nauwelijks voor dat alle elementaire in een bepaalde t in een document of query voorkomen, laat staan dat dat ook bij de overige documenten en queries zo is. Dus $t \not\subseteq d$ en $t \not\subseteq q$ in tegenstelling tot de gemaakte veronderstelling bij vergelijking 3.8 en 3.9.

2. Hier wordt in tegenstelling tot het voorgaande nu een meer realistische representatie van documenten en queries beschouwd. Er wordt hier gebruik gemaakt van een zogenaamde *boom-afhankelijkheidsbenadering* (*tree-dependence approximation*). Wat dat is, wordt in paragraaf 3.2.2.1.1 uitgelegd. Uitgaande hiervan kan de kans $P(d \cap q \cap t)$ als volgt worden uitgedrukt:

$$\begin{aligned} P(d \cap q \cap t) &= P(t)P(d \mid t)P(q \mid d \cap t) \\ &\approx P(t)P(d \mid t)P(q \mid t) \\ &= P(d \cap t)\frac{P(q \cap t)}{P(t)} \end{aligned} \tag{3.10}$$

Het bovenstaande benadering komt overeen met de aanname dat wat waarschijnlijkheden betreft, d en q onafhankelijk zijn bij een gegeven t, wat zich als volgt laat uitdrukken in de volgende vergelijking:

$$\begin{aligned} P(d \cap q \cap t) &= P(t)P(d \cap q \mid t) \\ &\approx P(t)P(d \mid t)P(q \mid t) \\ &= P(d \cap t)P(q \mid t) \end{aligned} \tag{3.11}$$

Deze aanname is consistent, omdat in de praktijk het leggen van indexen op documenten gewoonlijk onafhankelijk is van het leggen van indexen op queries in tekstuele vorm.
Als nu vergelijking 3.10 in vergelijking 3.7 wordt gesubstitueerd dan wordt de relatie voor $d \rightarrow q$ met betrekking tot de relevantie gegeven door

$$\Psi(d \rightarrow q) \approx \sum_t \left[\frac{P(d \cap t)}{P(d \cap T)} \cdot \frac{P(q \cap t)}{P(t)} \right] \tag{3.12}$$

Nu is de vraag of de gebruikte boom-afhankelijkheidsbenadering in vergelijking 3.10 wel correct is. Dat dat in orde is, kan men als volgt inzien. Door de gemaakte aannamen dat

$t_i \cap t_j = \emptyset$ voor $i \neq j$ kan men de maat voor de relevantie zoals deze door vergelijking 3.7 is gegeven, als volgt schrijven:

$$\begin{aligned} \Psi(d \rightarrow q) &\approx \sum_t \frac{P(d \cap q \cap t)}{P(d \cap T)} \\ &= \sum_t \left[\frac{P(d \cap t)}{P(d \cap T)} \cdot P(q \mid d \cap t) \right] \\ &= \sum_t [\Psi(d \rightarrow t \mid T) \Psi(t \rightarrow q \mid d)] \end{aligned} \tag{3.13}$$

In deze bovenstaande vergelijking geeft de term $\Psi(t \rightarrow q \mid d)$ semantisch gezien de representatie van een query weer. Daar staat echter ook dat de specificatie van de query nauw verwant moet zijn met de karakteristieken van elk document, wat er op neer komt dat de gebruiker a priori kennis van die documenten moet hebben. Dat is in feite onmogelijk. Men kan daarom gerust aannemen dat de specificatie van een query niet afhangt van de documenten, ofwel $\Psi(t \rightarrow q \mid d)$ is onafhankelijk van d, zo dat:

$$\Psi(t \rightarrow q \mid d) \approx \Psi(t \rightarrow q) \approx \Psi(t \rightarrow q \mid T) \tag{3.14}$$

Dan kan de vergelijking 3.13 als volgt worden weergegeven:

$$\Psi(d \rightarrow q) \approx \sum_t [\Psi(d \rightarrow t \mid T) \Psi(t \rightarrow q \mid T)] \tag{3.15}$$

En zie: dit is gelijk aan vergelijking 3.12 dat met behulp van een boom-afhankelijkheidsbenadering is verkregen. Die gebruikte benadering is dus niet verkeerd.

In de bovenstaande analyse tot dusver kan in de vergelijking 3.15 de eerste term $\Psi(d \rightarrow t \mid T)$ opgevat worden als een representatie voor document d en de tweede term $\Psi(t \rightarrow q \mid T)$ als een *expliciete* representatie van de query q in de kennisdeelruimte T. De formule geeft dus een natuurlijke interpretatie van de betekenis van het indexeren van documenten en het formuleren van queries in information retrieval. Ook komt hieruit naar voren dat het indexeerproces en de queryformulering onafhankelijke processen zijn.
Nu moeten aan die termen nog waarden toegekend worden. In het meest ideale geval worden de waarden voor de termen aangeleverd door menesen die indexeren en gebruikers die queries opstellen met behulp van de semantiek van documenten, queries en index termen. Het is namelijk zo dat met die waarden deze personen hun geloof in de door hen gelegde relaties weergeven.
Er dient daarbij zeker wel rekening gehouden te worden met eisen en axioma's waaran voldaan dient te worden. De opgegeven waarden die kansen zijn, dienen consistent te zijn en enige betekenis te hebben. Men kan bijvoorbeeld bij het indexeren aangeven dat als een document d niets anders dan concept t steunt, dat deze persoon dan zegt dat $\Psi(d \rightarrow t \mid T) = 1$. Maar als de document d het concept t_1 meer steunt dan het concept t_2, dan moet het zo zijn dat $\Psi(d \rightarrow t_1 \mid T) > \Psi(d \rightarrow t_2 \mid T)$. Dit geldt ook voor de gebruiker die queries opstelt. In de praktijk is dit vrij moeilijk te realiseren. Er zijn systemen die hierbij hulp kunnen bieden. Wong en Yao zien het liefst een geautomatiseerd proces. In dit model wordt zo'n systeem voor het automatische indexeerproces hierna beschreven.

Wil men met behulp van vergelijking 3.12 $\Psi(d \rightarrow q)$ berekenen kunnen, dan moet eerst bekend zijn wat de kwantiteiten $P(d \cap t)/P(d \cap T)$ en $P(q \cap t)/P(t)$ zijn.
Stel de kennisdeelruimte $T = t_1 \cup t_2 \cup \ldots \cup t_n$ is verkregen door middel van het indexeren met trefwoorden. De basisconcepten t_i corresponderen dan met trefwoorden (index termen). Hier moet nog een **opmerking** geplaatst worden: in de beschrijving van het gehele systeem wordt het genoemde indexeerproces aangehouden.
Zij $f(d,t)$ een notatie voor de frequentie van voorkomen van trefwoorden t in document d. Onder de aanname dat de kans $P(d \cap t)$ bij benadering evenredig is met de frequentie van voorkomen $f(d,t)$, wordt het volgende verkregen:

$$P(d \cap t) \approx \eta f(d,t), \tag{3.16}$$

en

$$P(d \cap T) \approx \eta \sum_t f(d,t). \tag{3.17}$$

In deze vergelijkingen is η een normalisatiefactor. Men verkrijgt dan dus:

$$\frac{P(d \cap t)}{P(d \cap T)} \approx \frac{f(d,t)}{\sum_t f(d,t)} = \hat{f}(d,t). \tag{3.18}$$

Hier wordt de ratio $P(d \cap t)/P(d \cap T)$ benaderd door de *genormaliseerde* frequentie van voorkomen $\hat{f}(d,t)$, die door middel van het automatische indexeerproces verkregen kan worden. Met de resultaten tot nu toe kan de maat voor de relevantie als volgt worden geschreven:

$$\begin{aligned} \Psi(d \rightarrow q) &\approx \sum_t [\Psi(d \rightarrow t \,|\, T)\Psi(t \rightarrow q \,|\, T)] \\ &= \sum_t \left[\frac{P(d \cap t)}{P(d \cap T)} \cdot \frac{P(q \cap t)}{P(t)} \right] \\ &= \sum_t \left[\hat{f}(d,t) \cdot \frac{P(q \cap t)}{P(t)} \right]. \end{aligned} \tag{3.19}$$

Nu moet er nagegaan worden wat in de vergelijking die $P(t)$ moet zijn. Het indexeerproces met trefwoorden levert niet de informatie over de kans $P(t)$ van een een individueel trefwoord. Wat men wèl weet is hoeveel keren een trefwoord is gebruikt. Als men dit statistisch beschouwt, dan kan redelijkerwijs worden verondersteld dat $P(t)$ bij benadering evenredig is met het totaal aantal keren dat t in alle documenten voorkomt. Preciezer geformuleerd,

$$P(t) \approx \eta \sum_{d'} f(d',t), \tag{3.20}$$

Hierin wordt over alle documenten de som genomen. Als nu de vergelijkingen 3.19 en 3.20 worden samengevoegd, dan levert dat als resultaat op:

$$\Psi(d \rightarrow q) \approx \frac{1}{\eta} \sum_t \left[\hat{f}(d,t) \cdot \frac{P(q \cap t)}{\sum_{d'} f(d',t)} \right]. \tag{3.21}$$

Volgens Wong en Yao is in deze vergelijking de kwantiteit $1/\sum_{d'} f(d',t)$ algemeen bekend als de *geïnverteerde frequentie van voorkomen van documenten* (*inverse document frequency (idf)*) die in modellen die op vectoren zijn gebaseerd, voorkomt.

De kans $P(q \cap t)$ kan hier op de zelfde manier als met de kans $P(d \cap t)$ worden geschat met behulp van het indexeerproces met trefwoorden. Zij $f(q,t)$ een notatie voor de frequentie van voorkomen van trefwoorden t in de tekst van query q. Men kan dan aannemen dat

$$P(q \cap t) \approx \eta f(q,t). \tag{3.22}$$

Hier is η weer de normalisatiefactor. Uit de vergelijkingen 3.21 en 3.22 volgt dan voor de maat voor de relevantie:

$$\Psi(d \rightarrow q) \approx \sum_t \left[\hat{f}(d,t) \cdot \frac{f(q,t)}{\sum_{d'} f(d',t)} \right]. \tag{3.23}$$

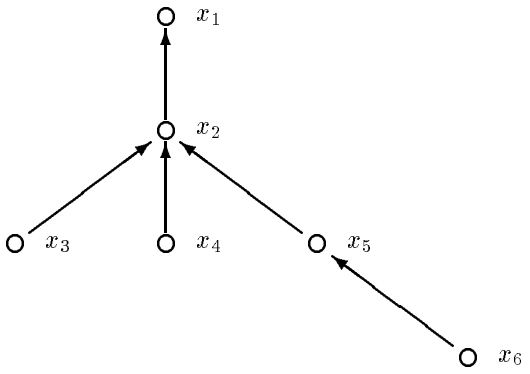
3.2.2.1.1 Intermezzo: boom-afhankelijkheid

In het artikel van de heren Chow en Liu ([CL68]) wordt gesproken over een manier om een discrete kansverdeling te benaderen. Daarbij kwam de benadering met behulp van afhankelijkheidsbomen aan de orde. Aangezien dit onderwerp in het probabilistische afleidingsmodel ook ter sprake is gekomen wordt hier een korte bespreking gegeven.

Zij $P(\mathbf{x})$ een samengestelde kansverdeling met n discrete variabelen $x_1, x_2, \ldots, x_n$. $\mathbf{x}$ is een notatie voor de vector $(x_1, x_2, \ldots, x_n)$. De bedoeling is dat de samengestelde kansverdeling wordt bepaald door het produkt van kansverdelingen van componenten van $\mathbf{x}$. Er kunnen echter legio uitkomsten worden ingevuld. Daarom wordt er een beperking gelegd op de mogelijkheden en wordt er slechts gekeken naar componenten die zelf tweede-orde kansverdelingen zijn. Dat wil zeggen, $P(\mathbf{x})$ wordt slechts beschouwd als zijnde een produkt van de kansverdelingen $P(x_i \mid x_j)$. In dit geval gaat men iets verder: er zijn $n(n-1)/2$ van die kansverdelingen, waarvan er maar $n-1$ kunnen worden gebruikt. Wat $P(\mathbf{x})$ dus kan worden is:

$$P(\mathbf{x}) \;=\; \prod_{i=1}^{n} P(x_{m_i} \mid x_{m_{j(i)}}) \quad , \quad 0 \leq j(i) < n,$$

waarin $m_1, m_2, \ldots, m_n$ een onbekende permutatie van de integers $1, 2, \ldots, n$, en $P(x_i \mid x_0) = P(x_i)$ per definitie. Elke variabele in de formule hierboven kan worden gekoppeld aan hoogstens één variabele.
De kansverdeling zoals deze in de gegeven vergelijking is gerepresenteerd, wordt een kansverdeling van de **eerste-orde boom-afhankelijkheid** genoemd. De paren die uit de verzameling $\mathbf{x}= \{x_i \mid i = 1, 2, \ldots, n\}$ worden gevormd en de bijbehorende afbeelding $j(i)$ met $0 \leq j(i) < n$ wordt de *afhankelijkheidsboom* genoemd. Dat men dit zo noemt wordt duidelijk als men de afbeelding grafisch weergeeft. Figuur 3.3 geeft daar een voorbeeld van (met $n = 6$).

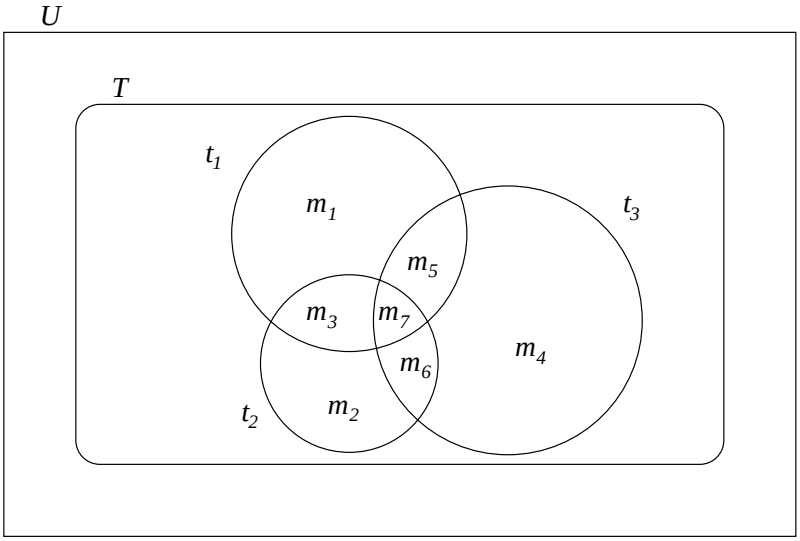


$$P(x) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2)P(x_6|x_5)$$

Figuur 3.3: Voorbeeld van een afhankelijkheidsboom

3.2.2.2 Niet-disjuncte concepten

Hier wordt het geval bekeken dat de basisconcepten niet disjunct zijn. Dit is ook in overeenstemming met de praktijk waarin bij automatische generatie van de indexen de basisconcepten (trefwoorden) meestal semantisch aan elkaar gerelateerd zijn. Toch is daar wat aan te doen. Zij $T = t_1 \cup t_2 \cup \ldots \cup t_n$ met t_i's als basisconcepten. Dan kan daaruit een verzameling van 2^n disjuncte *atomaire concepten* construeren:



Figuur 3.4: De atomaire concepten die uit niet-disjuncte basisconcepten zijn gegenereerd.

$$\begin{aligned}
m_0 &= \bar{t}_1 \cap \bar{t}_2 \cap \bar{t}_3 \cap \cdots \cap \bar{t}_{n-1} \cap \bar{t}_n \\
m_1 &= t_1 \cap \bar{t}_2 \cap \bar{t}_3 \cap \cdots \cap \bar{t}_{n-1} \cap \bar{t}_n \\
m_2 &= \bar{t}_1 \cap t_2 \cap \bar{t}_3 \cap \cdots \cap \bar{t}_{n-1} \cap \bar{t}_n \\
m_3 &= t_1 \cap t_2 \cap \bar{t}_3 \cap \cdots \cap \bar{t}_{n-1} \cap \bar{t}_n \\
&\cdots \\
m_{2^n-2} &= \bar{t}_1 \cap t_2 \cap t_3 \cap \cdots \cap t_{n-1} \cap t_n \\
m_{2^n-1} &= t_1 \cap t_2 \cap t_3 \cap \cdots \cap t_{n-1} \cap t_n
\end{aligned} \tag{3.24}$$

Het principe wordt enigzins duidelijk als men naar figuur 3.4 kijkt (voor $n = 3$).
Elk van deze atomaire concepten is een deelverzameling van de elementaire concepten in U. Nu de m's disjuncte concepten zijn, kan de maat voor de relevantie net zo als in de vergelijkingen 3.12 en 3.15 als volgt worden gedefinieerd:

$$\begin{aligned}
\Psi(d \rightarrow q) &\approx \sum_m [\Psi(d \rightarrow m|T)\Psi(m \rightarrow q|T)] \\
&= \sum_m \left[\frac{P(d \cap m)}{P(d \cap T)} \cdot \frac{P(q \cap m)}{P(m)} \right].
\end{aligned} \tag{3.25}$$

Hierin wordt over de verzameling van niet-lege atomaire concepten de som genomen.
Nu kan men $\Psi(d \rightarrow q)$ gelijk berekenen ware het niet dat niet bekend is wat die kwantiteiten $P(d \cap m)/P(d \cap T)$ en $P(q \cap m)/P(m)$ zijn. Daarnaast is er nog het probleem dat er geen gegevens zijn over hoe deze kwantiteiten afhangen van de atomaire concepten. Bij automatische indexgeneratie wordt slechts gebruik gemaakt van de basisconcepten (trefwoorden), dus de $t_i, i = 1, \ldots, n$. Dit biedt dus geen uitkomst. Desondanks kan er een methode worden ontwikkeld om die $\Psi(d \rightarrow q)$ te schatten waarbij er dan afgegaan wordt op de kennis over de basisconcepten (trefwoorden) in de kennisdeelruimte T. Op basis hiervan wordt in het hiernavolgende afgeleid wat de genoemde kwantiteiten zijn.

3.2.2.2.1 Een schatting voor $P(d \cap m)/P(d \cap T)$

Om een schatting te kunnen maken van $P(d \cap m)/P(d \cap T)$ is het handig om van de volgende benadering gebruik te maken:

$$\frac{P(d \cap m)}{P(d \cap T)} \approx \sum_t \frac{P(d \cap m \cap t)}{P(d \cap T)} \tag{3.26}$$

Als de t's werkelijk disjuncte concepten zijn, dan is de rechter term in de vergelijking exact. Als er nu van een boom-afhankelijkheidsbenadering gebruik wordt gemaakt, dan kan men voor $P(d \cap m \cap t)$ invullen:

$$P(d \cap m \cap t) \;=\; P(t)P(d \,|\, t)P(m \,|\, d \cap t) \quad \approx \quad P(t)P(d \,|\, t)P(m \,|\, t). \tag{3.27}$$

Wordt dit in vergelijking 3.26 ingevuld, dan wordt het volgende verkregen:

$$\begin{aligned} \frac{P(d \cap m)}{P(d \cap T)} \quad &\approx \quad \sum_t \left[\frac{P(d \cap t)}{P(d \cap T)} \cdot \frac{P(m \cap t)}{P(t)} \right] \\ &= \quad \sum_t \left[\Psi(d \rightarrow t \,|\, T) \Psi(t \rightarrow m \,|\, T) \right] \\ &\approx \quad \sum_t \left[\hat{f}(d,t) \cdot \frac{P(m \cap t)}{P(t)} \right] \end{aligned} \tag{3.28}$$

Nu moet er voor $P(m \cap t)$ nog iets ingevuld worden. Zij D_m een notatie voor de *maximale* deelverzameling van documenten waarin de doorsnede van de documenten met de basisconcepten die in m zitten niet leeg is, en voor het geval dat basisconcepten niet in m zitten is de genoemde doorsnede leeg. Wiskundig beschreven is D_m dus als volgt:

$$D_m \quad = \quad \{d \,|\, d \cap t_i \neq \emptyset \text{ als } t_i \cap m \neq \emptyset \;\wedge\; d \cap t_i = \emptyset \text{ als } t_i \cap m = \emptyset\} \tag{3.29}$$

Door gebruik te maken van de relatie tussen de documenten d en de atomaire concepten m zoals dit in vergelijking 3.29 is beschreven, en dan de aanname te gebruiken dat $P(d \cap t)$ bij benadering evenredig is met $f(d,t)$ zoals dat bij vergelijking 3.16 is gebruikt, kan men hier redelijkerwijs aannemen dat:

$$P(m \cap t) \quad \approx \quad \sum_{d' \in D_m} P(d' \cap t) \;\approx\; \eta \sum_{d' \in D_m} f(d',t) \tag{3.30}$$

Dan kunnen we nu met behulp van de vergelijkingen 3.20 en 3.30 het volgende afleiden:

$$\frac{P(m \cap t)}{P(t)} \quad \approx \quad \frac{\displaystyle\sum_{d' \in D_m} f(d',t)}{\displaystyle\sum_d f(d,t)} = h(t)g(t,m), \tag{3.31}$$

waarin

$$h(t) = \frac{1}{\displaystyle\sum_d f(d,t)} \quad , \quad g(t,m) = \sum_{d' \in D_m} f(d',t). \tag{3.32}$$

Merk op dat hier weer de geïnverteerde frequentie van voorkomen van documenten en de frequentie van voorkomen van een term binnen documenten komt kijken. Er zijn trouwens bepaalde verzamelingen documenten waarbij er moet worden gewerkt met een verzameling M van niet-lege atomaire concepten, ofwel

$$M \quad = \quad \{m \,|\, D_m \neq \emptyset\} \tag{3.33}$$

Nu kan er een afleiding gemaakt worden voor $P(d \cap m)/P(d \cap T)$. Uit de vergelijkingen 3.28 en 3.31 valt dan af te leiden:

$$\begin{aligned} \frac{P(m \cap t)}{P(d \cap T)} \quad &\approx \quad \sum_t \left[\hat{f}(d,t) h(t) g(t,m) \right] \\ &= \quad \sum_t \left[\hat{\mathbf{F}}_{dt} \mathbf{H}_{tt} \mathbf{G}_{tm} \right] \;=\; \hat{\mathbf{W}}_{dm} \end{aligned} \tag{3.34}$$

waarin

$$\hat{\mathbf{F}} = \begin{bmatrix} \hat{f}(d_1,t_1) & \hat{f}(d_1,t_2) & \dots & \hat{f}(d_1,t_n) \\ \hat{f}(d_1,t_1) & \hat{f}(d_1,t_2) & \dots & \hat{f}(d_1,t_n) \\ \vdots & \vdots & & \vdots \\ \hat{f}(d_1,t_1) & \hat{f}(d_1,t_2) & \dots & \hat{f}(d_1,t_n) \end{bmatrix}, \tag{3.35}$$

$$\mathbf{H} = \begin{bmatrix} h(t_1) & 0 & \dots & 0 \\ 0 & h(t_2) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & h(t_n) \end{bmatrix}, \tag{3.36}$$

$$\mathbf{G} = \begin{bmatrix} g(t_1,m_1) & g(t_1,m_2) & \dots & g(t_1,m_u) \\ g(t_2,t_1) & g(t_2,m_2) & \dots & g(t_2,m_u) \\ \vdots & \vdots & & \vdots \\ g(t_n,t_1) & g(t_n,m_2) & \dots & g(t_n,m_u) \end{bmatrix}, \tag{3.37}$$

en

$$\hat{\mathbf{W}} = \hat{\mathbf{F}}\mathbf{H}\mathbf{G}. \tag{3.38}$$

De matrix $\hat{\mathbf{F}}$ geeft de genormaliseerde frequentie van voorkomen van elke term voor elk document. Matrix $\mathbf{H}$ geeft de geïnverteerde frequentie van voorkomen van documenten (voor elk term). Matrix $\mathbf{G}$ lijkt veel op de matrix $\hat{\mathbf{F}}$ met één verschil dat daar naar de atomaire concepten is gekeken.
Als nu de vergelijking 3.34 in vergelijking 3.25 wordt gesubstitueerd, dan wordt het volgende verkregen:

$$\Psi(d \rightarrow q) \approx \sum_{m \in M} \left[\hat{\mathbf{W}}_{dm} \cdot \frac{P(q \cap m)}{P(m)} \right]. \tag{3.39}$$

3.2.2.2.2 Een schatting voor $P(q \cap m)/P(m)$

Nu moet er nog een afleiding voor $P(q \cap m)/P(m)$ gevonden worden waarbij weer van de kennis over de basisconcepten in de kennisdeelruimte T gebruik wordt gemaakt. Op dezelfde manier als in de vergelijkingen 3.26 en 3.27 een benadering voor $P(d \cap m)/P(d \cap T)$ is gemaakt, kan voor $P(q \cap m)/P(m)$ worden afgeleid:

$$\begin{aligned} \frac{(q \cap m)}{P(m)} &\approx \sum_t \frac{P(q \cap m \cap t)}{P(m)} \\ &= \sum_t \frac{P(t)P(q \mid t)P(m \mid q \cap t)}{P(m)} \\ &\approx \sum_t \frac{P(t)P(q \mid t)P(m \mid t)}{P(m)} \\ &= \sum_t \left[\frac{P(m \cap t)}{P(m)} \cdot \frac{P(q \cap t)}{P(t)} \right] \\ &= \sum_t \left[\Psi(m \rightarrow t \mid T)\Psi(t \rightarrow q \mid T) \right]. \end{aligned} \tag{3.40}$$

In de paragraaf over disjuncte concepten is er een schatting voor $P(q \cap t)$ gegeven, waarin $P(q \cap t)$ evenredig is met de frequentie van voorkomen $f(q,t)$ van de trefwoorden t in de tekst van de query zoals weergegeven in vergelijking 3.22. Dus als nu de vergelijkingen 3.22, 3.31 en 3.40 met elkaar worden gecombineerd, wordt het volgende resultaat verkregen:

$$P(q \cap m) \approx \eta \sum_t \left[f(q,t)h(t)g(t,m) \right] \tag{3.41}$$

Als er nu een $(1 \times n)$-matrix $\mathbf{V}$ wordt gedefinieerd dat de frequenties van voorkomen van trefwoorden in de tekst van de query representeert, ofwel

$$\mathbf{V} \quad = \quad [f(q,t_1), f(q,t_2), \ldots, f(q,t_n)]\,, \tag{3.42}$$

dan kan $P(q \cap m)$ als volgt worden herschreven:

$$P(q \cap m) \quad \approx \quad \eta \sum_t [\mathbf{V}_t \mathbf{H}_{tt} \mathbf{G}_{tm}] \; = \; \eta(\mathbf{VHG})_m\,. \tag{3.43}$$

Verder kan $P(m)$ als volgt worden benaderd (hier wordt vergelijking 3.28 gebruikt):

$$P(m) \approx \sum_d P(d \cap m) \quad \approx \quad \sum_d \left[\sum_t P(d \cap t) \cdot \frac{P(m \cap t)}{P(t)} \right]. \tag{3.44}$$

Nu bekend is dat $P(d \cap t) \approx \eta f(d,t)$ en $P(m \cap t)P(m) \approx h(t)g(t,m)$, kan de voorgaande vergelijking worden herschreven tot:

$$\begin{aligned} P(m) \quad &\approx \quad \eta \sum_d \left[\sum_t f(d,t)h(t)g(t,m) \right] \\ &= \quad \eta \sum_d \left[\sum_t \mathbf{F}_{dt} \mathbf{H}_{tt} \mathbf{G}_{tm} \right] \\ &= \quad \eta \sum_d \mathbf{W}_{dm}\,, \end{aligned} \tag{3.45}$$

waarin $\mathbf{F}_{dt} = f(d,t)$ een element is van de *niet-genormaliseerde* matrix van frequenties van voorkomen van trefwoorden $\mathbf{F}$, en

$$\mathbf{W} \quad = \quad \mathbf{FHG}. \tag{3.46}$$

3.2.2.3 Een formule voor de bepaling van de relevantie

Nu alle gegevens bekend zijn, kan nu een definitieve uitdrukking voor de maat voor de relevantie $\Psi(d \rightarrow q)$ afgeleid worden. Uit de vergelijkingen 3.39, 3.43 en 3.45 valt af te leiden:

$$\begin{aligned} \Psi(d \rightarrow q) \quad &\approx \quad \sum_{m \in M} \left[\hat{\mathbf{W}}_{dm} \cdot \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}} \right] \\ &= \quad \sum_{m \in M} \hat{\mathbf{W}}_{dm} \mathbf{q}_m \; = \; (\hat{\mathbf{W}}\mathbf{q})_d\,, \end{aligned} \tag{3.47}$$

waarin q_m een $(u \times 1)$-matrix is die als volgt is gedefinieerd:

$$\mathbf{q}_m \quad = \quad \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}}. \tag{3.48}$$

Over deze formule valt nog iets te zeggen. Als in die formule de correcties voor de geïnverteerde document frequentie (inverse document frequency, idf), dat is $1/\sum_{d'} \mathbf{W}_{d'm}$, worden weggelaten uit de vergelijking 3.48, dan wordt $\Psi(d \rightarrow q)$ vereenvoudigd tot (kijk ook naar vergelijking 3.47):

$$\Psi(d \rightarrow q) \quad \approx \quad \sum_{m \in M} \hat{\mathbf{W}}_{dm} \hat{\mathbf{q}}_m \; = \; (\hat{\mathbf{W}}\hat{\mathbf{q}})_d\,, \tag{3.49}$$

waarin $\hat{\mathbf{q}}$ een $(u \times 1)$-matrix is met $\hat{\mathbf{q}}_m = (\mathbf{VHG})_m$.
Volgens Wong en Yao is de formule in vergelijking 3.49 op enkele verschillen in normalisaties na

nagenoeg *identiek* aan de formule die voor het bepalen van de maat in het Generalized Vector Space Model (GVSM) wordt gebruikt, waarin bovendien de correcties voor de geïnverteerde document frequentie niet wordt gebruikt.

Nu lijkt het alsof de formule voor $\Psi(d \rightarrow q)$ zomaar uit de lucht komt vallen. Maar de correctheid ervan kan wel even gecontroleerd worden, door hem gewoon uit te proberen op disjuncte basisconcepten. Stel dus dat de basisconcepten disjunct zijn. Dan is elke t_i een atomair concept, ofwel $M = \{t_1, t_2, \ldots, t_n\}$. In dat geval wordt de matrix $\mathbf{HG}$ dat door de vergelijkingen 3.36 en 3.37 is gedefinieerd, een eenheidsmatrix $\mathbf{I}$:

$$(\mathbf{HG})_{tt'} \;=\; h(t)g(t,t') \quad = \quad \frac{P(t \cap t')}{P(t)} = \begin{cases} 1 & \text{als } t = t' \\ 0 & \text{als } t \neq t' \end{cases}, \tag{3.50}$$

en dus wordt:

$$\hat{\mathbf{W}} = \hat{\mathbf{F}}\mathbf{HG} = \hat{\mathbf{F}}\mathbf{I} = \hat{\mathbf{F}} \quad , \quad \hat{\mathbf{W}}_{dt} = \hat{\mathbf{F}}_{dt}, \tag{3.51}$$

$$\mathbf{W} = \mathbf{FHG} = \mathbf{FI} = \mathbf{F} \quad , \quad \mathbf{W}_{dt} = \mathbf{F}_{dt} \tag{3.52}$$

en

$$\mathbf{q}_t = \frac{(\mathbf{VHG})_t}{\sum_{d'} \mathbf{W}_{d't}} \quad = \quad \frac{(\mathbf{V})_t}{\sum_{d'} \mathbf{F}_{d't}} = \mathbf{V}_t \mathbf{H}_{tt}. \tag{3.53}$$

Dan wordt de vergelijking 3.47 onmiddellijk vereenvoudigd tot:

$$\begin{aligned} \Psi(d \rightarrow q) \quad &\approx \quad \sum_{t \in M} \hat{\mathbf{W}}_{dt} \mathbf{q}_t \;=\; \sum_{t} \hat{\mathbf{F}}_{dt} \mathbf{V}_t \mathbf{H}_{tt} \\ &= \quad \sum_{t} \left[\hat{f}(d,t) \cdot \frac{f(q,t)}{\sum_{d'} f(d',t)} \right]. \end{aligned} \tag{3.54}$$

En zie: dit is gelijk aan de maat voor de relevantie zoals deze in vergelijking 3.23 is gegeven, waarbij is verondersteld dat de basisconcepten disjunct zijn.

Het beschreven systeem heeft nog een eigenschap dat het vermelden waard is. Laat nu $\Psi(d \rightarrow q)$ worden uitgedrukt in termen van de Ψ's. Als nu de vergelijkingen 3.25, 3.28 en 3.40 worden gecombineerd, dan wordt het resultaat:

$$\begin{aligned} \Psi(d \rightarrow q) \quad &\approx \quad \sum_{m \in M} \left[\Psi(d \rightarrow m \,|\, T) \Psi(m \rightarrow q \,|\, T) \right] \\ &= \quad \sum_{m \in M} \sum_{t} \left[\Psi(d \rightarrow t \,|\, T) \Psi(t \rightarrow m \,|\, T) \Psi(m \rightarrow t \,|\, T) \Psi(t \rightarrow q \,|\, T) \right]. \end{aligned}$$

Deze formule ziet er interessant uit, want daar staat dat het mogelijk en toegestaan is om meer dan twee aanwijzingen te combineren om de relevantie van de relatie tussen d en q te krijgen.

3.2.3 Een voorbeeldtoepassing

De waslijst aan formules draagt eigenlijk niet bij aan de duidelijkheid van het model en de inzicht daarin. Een voorbeeld kan wonderen doen wat het begrip kweken betreft. Daarom is hieronder een toepassing gegeven waarin enkele (afgeleide) formules worden gebruikt. Het voorbeeld komt overigens uit [WY91].

Beschouw een verzameling documenten $D = \{d_1, d_2, d_3, d_4, d_5\}$. Daarbij is een matrix van frequenties van voorkomen van trefwoorden $\mathbf{F}$ gegeven waarin de frequenties niet genormaliseerd zijn:

$\mathbf{F} =$

	t_1	t_2	t_3
d_1	2	0	1
d_2	1	0	0
d_3	0	1	0
d_4	2	1	0
d_5	1	0	2

De genormaliseerde matrix van frequenties van voorkomen van trefwoorden $\hat{\mathbf{F}}$ zoals deze volgens de vergelijkingen 3.18 en 3.35 is gedefinieerd, ziet er als volgt uit:

$\hat{\mathbf{F}} =$

	t_1	t_2	t_3
d_1	2/3	0	1/3
d_2	1	0	0
d_3	0	1	0
d_4	2/3	1/3	0
d_5	1/3	0	2/3

Dit krijgt men uit de matrix $\mathbf{F}$ als men in die matrix de elementen deelt door de som van de elementen in de rijen.
Nu kan voor de verzameling documenten een verzameling $M = \{m_1, m_2, m_3, m_5\}$ van niet-lege atomaire concepten worden geproduceerd. Daarin is

$$m_1 = t_1 \cap \bar{t}_2 \cap \bar{t}_3 \quad , \quad m_2 = \bar{t}_1 \cap t_2 \cap \bar{t}_3$$
$$m_3 = t_1 \cap t_2 \cap \bar{t}_3 \quad , \quad m_5 = t_1 \cap \bar{t}_2 \cap t_3$$

Dus als de vergelijking 3.29 wordt toegepast, krijgen we $D_{m_1} = \{d_2\}$, $D_{m_2} = \{d_3\}$, $D_{m_3} = \{d_4\}$ en $D_{m_5} = \{d_1, d_5\}$. Deze resultaten zijn te verkrijgen, door een m_i over de rijen in matrix $\hat{\mathbf{F}}$ te leggen, en als de doorsnede met een rij niet leeg is, dan pakt men de bij de rij behorende document en stopt die in D_{m_i}.
Met behulp van vergelijking 3.32 worden de volgende matrices verkregen:

$\mathbf{H} =$

	t_1	t_2	t_3
t_1	1/6	0	0
t_2	0	1/2	0
t_3	0	0	1/3

en

$\mathbf{G} =$

	m_1	m_2	m_3	m_5
t_1	1	0	2	3
t_2	0	1	1	0
t_3	0	0	0	3

Met de bovenstaande matrices wordt dan

$\mathbf{W} = \mathbf{FHG} =$

	m_1	m_2	m_3	m_5
d_1	1/3	0	2/3	2
d_2	1/6	0	1/3	1/2
d_3	0	1/2	1/2	0
d_4	1/3	1/2	7/6	1
d_5	1/6	0	1/3	5/2

$\hat{\mathbf{W}} = \hat{\mathbf{F}}\mathbf{HG} =$

	m_1	m_2	m_3	m_5
d_1	1/9	0	2/9	2/3
d_2	1/6	0	1/3	1/2
d_3	0	1/2	1/2	0
d_4	1/9	1/6	7/18	1/3
d_5	1/18	0	1/9	5/6

Nu aan de 'documentenzijde' alle gegevens bekend zijn, kan er nu naar de query gekeken worden. Tel de voorkomens van trefwoorden in de tekst van de query en stop ze in een frequentie matrix $\mathbf{V}$ van voorkomen van trefwoorden:

$$\mathbf{V} = \quad [f(q,t_1), f(q,t_2), f(q,t_3)] \quad = (2,0,1)$$

De trefwoordconcepten kunnen zowel disjunct als niet disjunct zijn. Dit moet per geval bekeken worden:

Disjuncte concepten. Per document moet de relevantie voor de query bepaald worden. Dit kan met behulp van de vergelijking 3.23 (of 3.54):

$$\begin{aligned}\Psi(d_1 \rightarrow q) \quad &\approx \quad \sum_t \frac{\hat{\mathbf{F}}_{d_1 t}\mathbf{V}_t}{\sum_d \hat{\mathbf{F}}_{dt}} = \sum_t \hat{\mathbf{F}}_{d_1 t}\mathbf{V}_t\mathbf{H}_{tt} \\ &= \quad (\tfrac{2}{3}\cdot 2\cdot \tfrac{1}{6}) + (0\cdot 0\cdot \tfrac{1}{2}) + (\tfrac{1}{3}\cdot 1\cdot \tfrac{1}{3}) \;=\; \tfrac{3}{9}.\end{aligned}$$

Op de zelfde manier gebeurt dit ook voor de overige documenten:

$$\begin{aligned}\Psi(d_2 \rightarrow q) \;\approx\; \tfrac{3}{9} \quad , \quad \Psi(d_3 \rightarrow q) \;\approx\; 0 \\ \Psi(d_4 \rightarrow q) \;\approx\; \tfrac{2}{9} \quad , \quad \Psi(d_5 \rightarrow q) \;\approx\; \tfrac{3}{9}.\end{aligned}$$

In information retrieval worden documenten (meestal) naar mate van relevantie gesorteerd gepresenteerd. Met de waarden van Ψ hierboven is hier de ordening

$$\left\{ \begin{array}{l} d_1 \\ d_2 \\ d_5 \end{array} \right\}, \quad \{d_4\}, \quad \{d_3\}$$

Niet-disjuncte concepten. Hier wordt er dan met atomaire concepten gewerkt. Voor de bepaling van $\Psi(d \rightarrow q)$ wordt vergelijking 3.47 gebruikt. De trefwoorden zijn hier namelijk niet disjunct. Er moeten eerst nog enkele gegevens vergaard worden.
Met behulp van de matrices $\mathbf{V}$, $\mathbf{H}$ en $\mathbf{G}$ wordt bepaald:

$$\mathbf{VHG} \quad = \quad (1/3, 0, 2/3, 2).$$

En door in matrix $\mathbf{W}$ kolomsgewijs de elementen op te tellen, krijgt men:

$$\begin{aligned}\sum_d \mathbf{W}_{dm_1} \;=\; 1 \quad , \quad \sum_d \mathbf{W}_{dm_2} \;=\; 1 \\ \sum_d \mathbf{W}_{dm_3} \;=\; 3 \quad , \quad \sum_d \mathbf{W}_{dm_5} \;=\; 6.\end{aligned}$$

Met deze resultaten en met behulp van vergelijking 3.48 wordt

$$\mathbf{q} \quad = \quad (1/3, 0, 2/9, 1/3)^T,$$

waarin T aangeeft dat het een getransponeerde matrix is, dat wil zeggen de gegeven rijvector is een kolomvector.
Nu alle gegevens bekend zijn kan dit in vergelijking 3.47 ingevuld worden:

$$\begin{aligned}\Psi(d_1 \rightarrow q) \quad &\approx \quad \sum_{m \in M} \hat{\mathbf{W}}_{dm}\mathbf{q}_m \\ &= \quad (\tfrac{1}{9}\cdot\tfrac{1}{3}) + (0\cdot 0) + (\tfrac{2}{9}\cdot\tfrac{2}{9}) + (\tfrac{2}{3}\cdot\tfrac{1}{3}) \;=\; \tfrac{25}{81}.\end{aligned}$$

Voor de overige documenten gaat dit net zo:

$$\begin{aligned}\Psi(d_2 \rightarrow q) \approx \tfrac{24}{81} \quad , \quad \Psi(d_3 \rightarrow q) \approx \tfrac{9}{81} \\ \Psi(d_4 \rightarrow q) \approx \tfrac{19}{81} \quad , \quad \Psi(d_5 \rightarrow q) \approx \tfrac{26}{81}.\end{aligned}$$

Uitgaande van deze waarden van Ψ wordt er de volgende lijst van documenten gepresenteerd:

$$\{d_5\},\ \{d_1\},\ \{d_2\},\ \{d_4\},\ \{d_3\}.$$

3.2.4 Slotopmerking

In het voorgaande is een model beschreven met een daarbij beschreven systeem voor het (automatisch) genereren van waarschijnlijkheden bij automatische indexering. Bij het beschreven systeem is echter de query niet behandeld. Er is aangenomen dat met een query wordt verwezen naar trefwoorden. De query wordt daar namelijk gegeven middels de frequenties van voorkomen van trefwoorden t, $\mathbf{V}_t = f(q,t)$, in de tekst van de query. dan wordt er impliciet aangenomen dat de gebruiker bij de formulering van de vraagstelling gebruik maakt van die trefwoorden.
Wong en Yao geven aan dat dit in de praktijk niet geldt. Het komt zelden voor dat met de trefwoorden precies de informatiebehoeften van de gebruiker worden gedekt. Wong en Yao hebben om dit probleem op te lossen een methode ontwikkeld waarop de formulering van de vraagstelling wordt gebaseerd. Dit wordt hier niet behandeld.

3.3 Het afleidingsnetwerkmodel voor document retrieval

Turtle en Croft ([TC90]) hebben een retrieval model ontwikkeld dat gebaseerd is op netwerken. Netwerkrepresentaties worden al wel vaker gebruikt in information retrieval. Hier is de beschrijving enigzins formeel in de zin dat er een basisstructuur wordt gegeven die ook op andere modellen van toepassing is.
Het model moet de volgende mogelijkheden ondersteunen en in die situaties gebruikt kunnen worden:

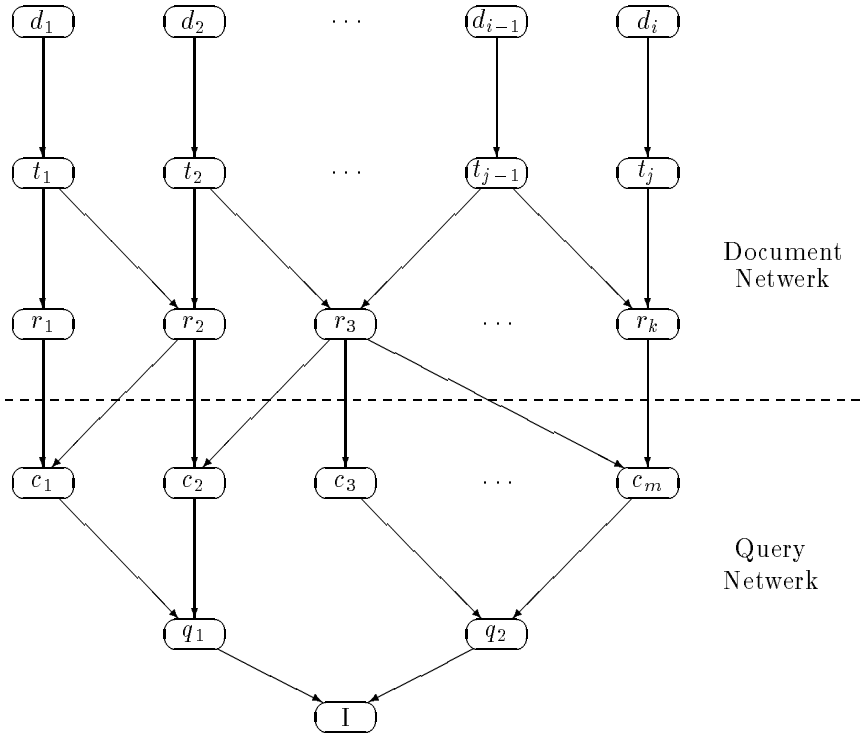
- Voor documenten worden meervoudige representatieschema's gebruikt. Voor een gegeven query komt het voor dat per representatieschema steeds verschillende documenten worden verkregen. Deze verschillen horen er niet te zijn;
- Resultaten van queries en verschillende soorten daarvan moeten te combineren zijn. Gegeven een omschrijving van een informatiebehoefte formuleren de gebruikers verschillende queries die die informatiebehoefte representeren, met als gevolg dat verschillende documenten worden verkregen bij elke query. De representatie van de query speelt hierbij semantisch gezien een grote rol;
- Er is een goede koppeling van de termen of concepten die in queries worden gebruikt met die concepten die aan documenten worden toegekend. Als de semantiek van concepten bij queries anders is als bij documenten, dan is de kans zeer groot dat men slechte resultaten krijgt.

Het formele retrieval model dat dit allemaal ondersteunt, integreert voorgaande modellen (voorlopers) in een theorethisch raamwerk. Men denke aan Boolese retrieval modellen, conventionele probabilistische modellen en dergelijke.

In het formele model is het de bedoeling dat meervoudige representaties van documenten en queries worden behandeld als aanwijzingen die dan worden gecombineerd om de kans te bepalen dat een document voorziet in de informatiebehoefte van de gebruiker.
De onderliggende structuur in het model is een Bayesische afleidingsnetwerk (zie [KP83], [Pea86] of [LS88]). Het idee om het netwerk te gebruiken bestaat al lang en verscheidene modellen zijn er op gebaseerd. Door dit netwerk te gebruiken kunnen bestaande modellen hier gegeneraliseerd worden. Verder biedt het netwerk de mogelijkheid dat verschillende kennisbronnen kunnen worden geïntegreerd in het raamwerk.

In het hiernavolgende wordt het model besproken. De basis wordt uitgewerkt en er wordt aangegeven hoe het model moet worden gebruikt. Onder meer het gebruik van het model voor Boolese of probabilistische retrieval wordt beschreven.

Turtle en Croft hebben ook hun model vergeleken met andere (probabilistische) modellen. Tevens hebben zij een uitbreiding van het model besproken. Maar dit wordt hier uit oogpunt van overzichtelijkheid niet behandeld. Het gaat meer om het idee, het principe.



Figuur 3.5: Het inferentienetwerk als basis in document retrieval

3.3.1 De basis van het model

De basis van het model is een document retrieval afleidingsnetwerk, die in figuur 3.5 grafisch is weergegeven. Het netwerk bestaat uit twee componenten:

- Een document netwerk.
  Deze representeert een verzameling documenten waarbij er verscheidene representatieschema's worden gebruikt. Het netwerk hoeft maar één keer gemaakt te worden en de structuur wordt niet door de query gewijzigd;

- Een query netwerk.
  Deze bestaat uit één knoop dat de informatiebehoefte representeert en een of meer queryrepresentaties die de informatiebehoefte weergeven. Het query netwerk wordt voor elke informatiebehoefte opgebouwd. Tijdens de behandeling van een query kan dat netwerk worden gewijzigd. Er kunnen nieuwe queries worden toegevoegd of bestaande queries worden meer gedetailleerd om de informatiebehoefte beter weer te geven.

Het document netwerk en het query netwerk worden aan elkaar gekoppeld via de verbindingen tussen de representatieconcepten en de queryconcepten. In het afleidingsnetwerk moet men alle knopen behalve het blad als aanwijzingen beschouwen. Zij zijn dan òf waar (true) òf niet waar (false). De knopen nemen met betrekking toto hun status deze waarden aan. De genoemde netwerken worden hieronder beschreven.

3.3.1.1 Het document netwerk en het query netwerk

In deze paragraaf worden van de netwerken de knopen en de relaties daartussen beschreven, te beginnen met het document netwerk. Het document netwerk bestaat uit een aantal verschillende knopen:

Documentknopen. Elke documentknoop (d_i) representeert een document in de verzameling. Deze correspondeert met de gebeurtenis dat een bepaald document wordt bekeken. De manier van representatie hangt af van de documenten en hoe ze worden gebruikt. In de veronderstelling dat een document als object goed gedefinieerd is, kan voor de representatie een abstracte type worden gebruikt (journaal, boek, ...).

Tekstrepresentatieknopen. In tegenstelling tot de documentknopen die documenten abstract representeren, zijn tekst(representatie-)knopen (de t_j's) gekoppeld aan een specifieke tekstrepresentatie van documenten. Hier gaat het meer om de tekstinhoud van de documenten. Het is mogelijk om met meer tekstknopen een document te representeren (plaatjes, audio), of tekstknopen zijn van toepassing op meerdere documenten, maar voor de duidelijkheid wordt verondersteld dat voor elk document een tekst is en omgekeerd.
Dat een tekst behoort bij een document wordt aangegeven middels een pijl van de documentknoop naar de tekstknoop.

Representatieknopen. De representatieknopen (r_k's) worden verdeeld naar deelverzamelingen die elk corresponderen met een representatietechniek dat op de tekst van de documenten is toegepast. Een tekstrepresentatie "information retrieval" kan zijn een zinsnede uit de tekst van een document, of het is een aan een document aangebrachte index term. Dit zijn twee verschillende concepten die dan ook met twee knopen worden aangegeven. De ene knoop correspondeert met de gebeurtenis dat "information retrieval" uit de tekst van de deelverzameling documenten is gehaald, en de andere knoop met de gebeurtenis dat "information retrieval" als index term aan een deelverzameling van documenten is toegekend. Hier wordt trouwens de concepten als disjunct beschouwd ook al lijkt het er soms niet op.
Een pijl van een tekstknoop naar een representatieknoop geeft aan dat een concept is toegepast op de tekst van een document. Er wordt verondersteld dat de verbinding strikt is, dat wil zeggen het concept is òf toegepast òf niet toegepast.
Men kan afvragen of het aantal representatieschema's geen problemen oplevert. Er zijn talloze representatieschema's. Dit zal wel meevallen. In de praktijk past men op een echte verzameling documenten maar een beperkt aantal representatietechnieken toe. Het domein van een representatieschema zal ook niet te groot worden, doorgaans is de omvang niet groter dan het aantal woorden in de hele documentverzameling.

Het query netwerk zoals in figuur 3.5 is weergegeven, heeft queryconceptknopen als wortels, als enige blad de knoop dat de informatiebehoefte weergeeft, en daartussen de queryknopen. Een beschrijving van die knopen is als volgt:

De informatiebehoefte. De knoop I dat de informatiebehoefte representeert kan niet goed omschreven worden. De interpretatie van de informatiebehoefte hangt van de gebruiker af. De informatiebehoefte wordt wel uitgedrukt via queries, zij het niet expliciet. De knoop correspondeert met de gebeurtenis dat aan de queries wordt voldaan. De pijl van de queryknopen naar de knoop I geeft de afhankelijkheid van de informatiebehoefte van de queries weer.

Queryknopen. De queryknopen (q_p's) geven de queries weer, die de informatiebehoefte (enigzins) beschrijven. De formulering van een query kan op basis van een beschrijving van een natuurlijke taal zijn. Men denke aan trefwoorden, zinsnedes, of Boolese representaties, uittreksels van documenten of iets anders. Alle queryknopen tezamen geven de informatiebehoefte weer. Hoe meer queries, des te beter. De queryknoop is afhankelijk van het concept waarin de query is beschreven.

Queryconceptknopen. De queryconceptknopen (c_m's) representeren de elementaire concepten waarmee de informatiebehoefte kan worden uitgedrukt. De queries maken gebruik van deze concepten. Elke queryconceptknoop correspondeert met de gebeurtenis dat een queryconcept op de representatieconcept(en) is af te beelden. De queryconceptknopen definiëren een afbeelding tussen de concepten die de verzameling documenten representeren en de concepten die in de queries worden gebruikt. Een pijl van een representatieknoop naar een queryconceptknoop geeft de afbeelding weer.

In het eenvoudigste geval formuleert men queries in termen van representanten van documenten. Dan zijn de queryconcepten gelijk aan de representatieconcepten. In de overige gevallen vertoont een queryconcept een overeenkomst met een representatieconcept dat uit andere representatieconcepten is afgeleid.

Het zal duidelijk dat in de bovenstaande tekst nu ook impliciet is beschreven hoe het query netwerk aan het document netwerk is gekoppeld. Die koppeling is niet van invloed op het document netwerk zelf. De structuur blijft onveranderd, de specificaties van de voorwaardelijke kansen die aan de knopen hangen eveneens.

Er is niet aangegeven hoe hier de onzekerheid ingebracht wordt. In het afleidingsnetwerk is op de wortels na bij elke knoop voorwaardelijke kansen gespecificeerd. Dus voor bijvoorbeeld een representatieknoop r_k die afhankelijk is van de tekstknopen wordt voor elke tekstknoop de kans $P(r_k|t_j)$ gespecificeerd. Hetzelfde gebeurt met de queryknopen, tekstknopen en de knoop dat de informatiebehoefte representeert.
De documentknopen vormen de wortels. Voor elk van deze knopen wordt een a priori kans $P(d_i)$ gespecificeerd. Details over de invullingen voor die kansen evenals hoe ze gerepresenteerd worden, worden later gegeven. Tot zover kan er van uitgegaan worden dat ze kunnen worden gebruikt.

3.3.1.2 Gebruik van het afleidingsnetwerk

We kunnen nu bekijken hoe het netwerk wordt gebruikt bij het bepalen van de kansen (relevanties). Het basisprincipe is dat als aan de documenten de a priori kansen en aan de interne knopen de voorwaardelijke kansen zijn toegekend, er voor elke knoop in het netwerk een a posteriori kans berekend kan worden. Als er waarden van de variabelen in het netwerk bekend worden, dan worden er voor de resterende knopen de kansen opnieuw berekend naar aanleiding van die nieuwe "aanwijzing".

Het netwerk als geheel representeert de afhankelijkheid van de informatiebehoefte van de gebruiker van de verzameling documenten. Die afhankelijkheid wordt bepaald door de representaties van de queries en de documenten. Eerst wordt er op basis van van de queries een query netwerk opgebouwd die dan aan het document netwerk wordt gekoppeld, waarna voor elke knoop in het query netwerk het bijbehorende geloof (de kans) wordt berekend. Initieel heeft de knoop dat de informatiebehoefte representeert een kans dat aan de informatiebehoefte is voldaan zonder dat er ook maar één specifiek document is bekeken.
Nu wordt er van elk document d_i in de verzameling *apart* de kans berekend dat deze aan de informatiebehoefte voldoet. Dus als document d_i wordt bekeken, dan worden alle overige documenten $d_j, j \neq i$ als niet bekeken beschouwd. Als aanwijzing dat document d_i wordt bekeken voert men $d_i = true$ in. Heeft men van elke document de kans (relevantie) bepaald, dan worden de documenten gerangschikt naar hoogte van de relevantie, waarna de rangschikking (of een deel ervan) wordt gepresenteerd.

Het hoeft niet persé zo te zijn dat elk document apart wordt bekeken, men kan ook deelverzamelingen van documenten bekijken. Alleen zijn er voor een gegeven deelverzameling van n documenten zo'n 2^n deelverzamelingen mogelijk. Qua complexiteit van het benodigde rekenwerk is deze alternatief niet echt bruikbaar. Men kan dan het beste bij aparte documenten blijven.
Het kan zijn dat de opgegeven rangschikking niet nauwkeurig genoeg is. Dan kan men nieuwe informatie aan het query netwerk toevoegen, of men verfijnt de structuur van het netwerk om de betekenis van de queries beter te karakteriseren. Naast de al gegeven queries nog nieuwe queries toevoegen kan ook, maar dat leidt tot een nieuw op te bouwen query netwerk met nieuw te bepalen kansen. Bij elk van de hiervoor genoemde mogelijkheden begint het proces met de documenten weer van voren af aan.

3.3.2 Onzekerheid in het model

Er is een afleidingsnetwerk als basis voor het model beschreven en hoe deze gebruikt moet worden. Bij het gebruik van het netwerk zijn waarschijnlijkheden ter sprake gekomen. De invulling

en berekening van waarschijnlijkheden zijn daar echter niet uitgewerkt. Ook is niet vermeld hoe waarschijnlijkheden gerepresenteerd worden. Dit gebeurt hier alsnog. Eerst wordt er een representatietechniek besproken. Een toepassing ervan in Boolese en probabilistische retrieval systemen voor eventuele gebruik bij het model wordt behandeld. Daarna wordt aangegeven wat de invulling voor de waarschijnlijkheid per knoop is.

3.3.2.1 Canonische link matrices

In het gehele afleidingsnetwerk moet voor de knopen uitgezonderd die die grafisch gezien geen wortels zijn, voor elke knoop een kans geschat worden afhankelijk van wat de verzameling van waarden voor diens ouderlijke knopen zijn. Dus als een knoop a een verzameling ouders $\pi_a = \{p_1, \ldots, p_n\}$ heeft, dan moet de kans $P(a \mid p_1, p_2, \ldots, p_n)$ bepaald worden. Men kan die kansen coderen in een link matrix. Er zijn zo'n beetje 2^n combinaties van ouders waarover over het geheel de kans dat $a = \textbf{true}$ of $a = \textbf{false}$ moet worden bepaald. De link matrix krijgt dan een omvang van 2×2^n voor n ouders. Dit is duur, zeker als per knoop het aantal ouders groot is. De link matrix is dus slechts praktisch in het geval dat voor een knoop het aantal ouders klein is. Verondersteld dat dat laatste zo is, dan moet er een oplossing worden gezocht voor het schatten van de afhankelijkheid van een knoop van zijn ouders, en hoe die schattingen in een bruikbaar vorm te krijgen is. Er is er een gevonden. Men maakt gebruik van een vorm van *canonische link matrices*. Hoe dat werkt, wordt hier gedemonstreerd met Boolese operatoren. Daarbij wordt gelijk getoond hoe dan een kans uit zo'n matrix is af te leiden.

Voorbeeld 3.3.1

Stel een knoop Q *heeft drie ouders,* A, B *en* C *waarvoor geldt:*

$$P(\mathrm{A} = \textbf{true}) = a, \quad P(\mathrm{B} = \textbf{true}) = b, \quad P(\mathrm{C} = \textbf{true}) = c.$$

Bekijk nu een **or***-combinatie, dat wil zeggen,* $\mathrm{Q} = \mathrm{A} \vee \mathrm{B} \vee \mathrm{C}$*. Dit suggereert een link matrix van de vorm*

$$L_{\textbf{or}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Als men update procedures in gesloten vorm gebruikt, krijgt men

$$\begin{aligned} P(\mathrm{Q} = \textbf{true}) &= (1-a)(1-b)c + (1-a)b(1-c) + (1-a)bc + a(1-b)(1-c) \\ &\quad + a(1-b)c + ab(1-c) + abc \\ &= 1 - (1-a)(1-b)(1-c) \end{aligned}$$

wat overeenkomt met de regel voor disjunctieve combinaties van gebeurtenissen waarvan niet bekend is dat ze onderling onafhankelijk zijn. Soortgelijke matrices kunnen worden gemaakt voor **and** *(*$P(\mathrm{Q} = \textbf{true}) = abc$*), en* **not** *(*$P(\mathrm{Q} = \textbf{true}) = 1 - a$*) of iets anders. Dit gaat als volgt. Om de matrix voor de* **or***-combinatie af te leiden, maakt men een waarheidstabel die er zo uitziet:*

canonische produkt												canonische som							
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	A	B	C	Q	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
0	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*1*	*0*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*1*	*0*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*1*	*0*
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*1*

De waarheidstabel is bekend van de Boolese algebra evenals de kreten canonische som en canonische produkt. Kijkt men in de tabel onder de canonische som naar de diagonaal, dan

staat daar de invulling van de onderste rij van de link matrix. Ditzelfde kan men ook doen met de tabel voor het canonische produkt. Dan staat daar in geïnverteerde vorm de invulling voor de bovenste rij van de link matrix. De handige (snelle) manier is om de kolom van Q *te gebruiken.* □

3.3.2.2 Probabilistische retrieval

Nu er een representatietechniek in de vorm van link matrices beschreven is, kan nu de toepassing ervan in een probabilistisch systeem beschreven worden.
Voor probabilistische retrieval wordt aan elk ouder een gewicht toegekend en zo heeft ook het kind er een. In de canonische link matrix die hier nu als een *gewogen-som matrix* wordt beschouwd, hangt het geloof in het kind Q af van de specifieke ouders die waar (true) zijn. Ouders met een grotere gewicht dan de overige ouders hebben meer invloed op het geloof (in Q). Een voorbeeld van een gewogen-som matrix wordt hier in het volgende voorbeeld gegeven.

Voorbeeld 3.3.2

Net als in het vorige voorbeeld (voorbeeld 3.3.1) wordt er naar het antwoord op de uitspraak $\mathrm{Q} = \mathrm{A} \vee \mathrm{B} \vee \mathrm{C}$ *gezocht.*
Stel de ouders A, B *en* C *hebben respectievelijk de gewichten* $w_a, w_b, w_c, \geq 0$, *en het kind* Q *heeft gewicht* $0 \leq w_q \leq 1$. *Verder is* $t = w_a + w_b + w_c$. *Dan wordt er een gewogen-som link matrix verkregen die er zo uitziet:*

$$\begin{pmatrix} 1 & \frac{(w_a+w_b)w_q}{t} & \frac{(w_a+w_c)w_q}{t} & \frac{w_a w_q}{t} & \frac{(w_b+w_c)w_q}{t} & \frac{w_b w_q}{t} & \frac{w_c w_q}{t} & 1-w_q \\ 0 & \frac{w_c w_q}{t} & \frac{w_b w_q}{t} & \frac{(w_b+w_c)w_q}{t} & \frac{w_a w_q}{t} & \frac{(w_a+w_c)w_q}{t} & \frac{(w_a+w_b)w_q}{t} & w_q \end{pmatrix}.$$

Evaluatie van deze link matrix levert dan als resultaat:

$$P(\mathrm{Q} = \mathbf{true}) \quad = \quad \frac{(w_a \mathrm{a} + w_b \mathrm{b} + w_c \mathrm{c}) w_q}{t}.$$

Daarin geven a, b *en* c *met de waarde* 0 *of* 1 *aan of de desbetreffende ouder waar (true) is. Vergelijk deze matrix met de* **or***-matrix in het vorige voorbeeld en kijk aldaar ook naar de waarheidstabel.* □

In principe kan de link matrix in het voorbeeld gebruikt worden voor toepassingen met verscheidene modellen voor gewichten zoals *term frequentie binnen documenten (tf)*, *geïnverteerde document frequentie (idf)*, of beide (*tf.idf*). Deze begrippen zijn in information retrieval vrijwel bekend. Voor de betekenissen van die begrippen wordt er een voorbeeld gegeven, waarin (*tf.idf*) wordt toegepast.

Voorbeeld 3.3.3

Zij w_a, w_b, w_c *genormaliseerde tf-waarden voor* A, B *en* C, *en* idf_q *de genormaliseerde idf-gewicht voor* Q. *Zij voorts*

$$w_q \quad = \quad idf_q \cdot (w_a + w_b + w_c). \tag{3.55}$$

Als de link matrix in het vorige voorbeeld wordt gebruikt, en A *wordt bekeken (*$\mathrm{A} = \mathrm{true}$*), dan wordt het geloof in* Q *gegeven door:*

$$\begin{aligned} \mathrm{bel}(\mathrm{Q}) &= \frac{w_a w_q}{w_a + w_b + w_c} \\ &= \frac{tf_a \cdot idf_q \cdot (w_a + w_b + w_c)}{w_a + w_b + w_c} \\ &= tf_a \cdot idf_q \end{aligned}$$

wat een vorm van een tf.idf gewicht is. Voor een goed begrip van de termen tf en idf wordt er een voorbeeld invulling bij gegeven. Stel er zijn weer de drie ouders A, B *en* C, *en er zijn nu drie knopen* Q, R *en* S *die elk één term in hun domein hebben. Dan zijn er de volgende matrices waarin de relaties tussen de ouders en de knopen via de termen zijn gegeven:*

$$\mathbf{T} = \begin{array}{c|ccc} & \text{Q} & \text{R} & \text{S} \\ \hline \text{A} & 2 & 0 & 1 \\ \text{B} & 1 & 0 & 0 \\ \text{C} & 0 & 1 & 1 \end{array}, \quad \hat{\mathbf{T}} = \begin{array}{c|ccc} & \text{Q} & \text{R} & \text{S} \\ \hline \text{A} & 2/3 & 0 & 1/3 \\ \text{B} & 1 & 0 & 0 \\ \text{C} & 0 & 1/2 & 1/2 \end{array}, \quad \mathbf{I} = \begin{array}{c|ccc} & \text{Q} & \text{R} & \text{S} \\ \hline \text{Q} & 1/3 & 0 & 0 \\ \text{R} & 0 & 1 & 0 \\ \text{S} & 0 & 0 & 1/2 \end{array}.$$

De matrix **T** *geeft term frequenties voor elk paar van ouder/kind. In matrix* $\hat{\mathbf{T}}$ *staan de genormaliseerde term frequenties. De matrix* **I** *geeft de geïnverteerde document frequenties. Die vindt men door de kolommen van matrix* **T** *op te tellen en te inverteren.*
Uit deze gegevens zal blijken dat voor de bekeken A bel(Q) $= \frac{2}{9}$. □

Zoals men ziet is in het voorbeeld in principe de afhankelijkheid van een representatieknoop van zijn ouderlijke tekstknopen geïllustreerd. Als een document wordt bekeken, dan nemen al de representatieknopen waaraan het document is gekoppeld het *tf.idf* gewicht aan dat met het document/term paar is geassocieerd. Het gewicht hoort specifiek bij de term dat aan het document is toegekend.

Turtle en Croft hebben de vergelijking 3.55 gegeven zonder dat er een formele basis ervoor is. Ze hebben wel de betekenis of de bedoeling van de componenten in de vergelijking omschreven.
Het component idf_q geeft het maximale geloof dat bij de knoop Q haalbaar is. Bedenk hier dat een term aan meerdere documenten kan worden toegekend. Het component hangt af van de verdeling van de term in de verzameling documenten.
Het component $(w_a + w_b + w_c)$ normaliseert de gewichten van de ouders. In de gewogen-som matrix in voorbeeld 3.3.2 staat deze term in de noemers.

3.3.2.3 Boolese retrieval

Hier wordt beschreven hoe in het model Boolese retrieval kan worden toegepast met behulp van canonische matrices. Een variant van Boolese retrieval komt hier ook aan bod.

Voor de beschrijving van de toepassing wordt verondersteld dat de queryconcepten gelijk aan de representatieconcepten zijn, zodat de queryconcepten uit het netwerk kunnen worden weggelaten. Tevens wordt aangenomen dat als een document d_i wordt bekeken, de overige documenten d_j niet wordt bekeken (d_j = false, $j \neq i$).
De toepassing op het basismodel gaat als volgt:

1. Gebruik bij elke representatieknoop een canonische **or**-matrix. Als een document wordt bekeken, dan wordt bij elke representatieconcept waar het document aangehangen is, bel(r_i)=1. Bij de overige concepten is bel(r_j)=0;

2. Bouw voor de query een expressieboom op. De wortel van deze graaf is de query, en alle pijlen zijn naar de wortel gericht. De bladeren zullen de representatieconcepten zijn en de tussenliggende knopen corresponderen met operatoren voor de expressie. Gebruik dan voor elk zo'n operatorknoop een canonische link matrix. Hang dan deze boom aan het document netwerk;

3. Maak nu gebruik van de evaluatieprocedure die in paragraaf 3.3.1.2 is beschreven. Bekijk elk document een voor een en bewaar voor het bekeken document het geloof in de queryknoop. Elke document waarvoor bel(Q) = 1 is relevant voor de query, en de overige documenten waarvoor bel(Q) $<$ 1 niet.

Als naast de bovengenoemde aannamen nog wordt verondersteld dat de ouderlijke knopen in de matrices voor Boolese operatoren slechts de waarden 0 of 1 aannemen (*binaire indexering*), dan kan bel(Q) ook alleen maar de waarden 0 of 1 aannemen. In dit geval simuleert het afleidingsnetwerk exact een conventioneel Boolese systeem.
Het hoeft trouwens niet persé zo te zijn dat voor documenten die niet worden bekeken, de waarden 0 moet zijn. Dan beschouwt men alleen de documenten waarvoor bel(Q) = 1 relevant voor de query en de overige documenten hebben dan een klein bel(Q) $\neq$ 0.

Een variant op het beschreven systeem is Boolese retrieval. Als nu de termen de waarden als gewichten in het bereik $[0, 1]$ aannemen, kunnen de gewichten worden beschouwd als de kans dat

de term aan een document is toegekend.
Net als bij probabilistische retrieval in de vorige paragraaf, kan de interpretatie van de Boolese operatoren op deze manier ook met gewogen indexen geschieden. Bij de representatieconcepten worden de **or**-link matrices vervangen door de gewogen-som matrices, met daarin de geschikte *tf*- en *idf*-gewichten (*tf* staat voor term frequentie, *idf* voor geïnverteerde document frequentie). Als in dit geval een document wordt bekeken, dan nemen al de representatieknopen waaraan het document is gekoppeld, als waarde het gewicht *tf.idf* aan dat met het betreffende document/term paar is geassocieerd, en bij de overige representatieknopen wordt bel = 0. Deze gewichten worden dan gecombineerd met behulp van een uitdrukking in gesloten vorm zoals in voorbeeld 3.3.1 in paragraaf 3.3.2.1, om tenslotte voor de query bel(Q) te bepalen.
Samengevat komt het er op neer dat de *tf.idf* gewichten worden geïnterpreteerd als kansen en worden gecombineerd waarbij er dan de normale regels voor ontkenning, disjunctie of conjunctie van verzamelingen in de uitkomstenruimte wordt gebruikt.

3.3.2.4 Schatting van de waarschijnlijkheden

Nu de canonische linkmatrices in paragraaf 3.3.2.1 zijn beschreven en hoe ze bij Boolese of probabilistische retrieval te gebruiken zijn, kan er nu een invulling worden gegeven voor de waarschijnlijkheden voor de knopen in het netwerk. Hieronder worden de knopen puntsgewijs behandeld.

Documentknopen. Deze zijn de wortels in de graaf in figuur 3.5. Met deze knopen worden a priori kansen geassocieerd. De kans wordt 1/(omvang verzameling documenten).

Tekstknopen. In de beschrijving van het document netwerk is aangegeven dat de tekstknoop bij precies één documentknoop hoort en omgekeerd. Dus is de tekstknoop volledig afhankelijk van de documentknoop. Omdat er maar één ouder is kan er een link matrix worden gebruikt; t_j = true als d_i = true, dus

$$L_{\text{text}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Merk op dat een onderscheid tussen de documentknoop en de tekstknoop niet vereist is voor het basismodel. Voor de duidelijkheid kan men de tekstknopen wel weglaten. De tekstknopen zijn wel vereist, als een tekst op meerdere documenten van toepassing is. Als het toegestaan is dat documentknopen tekstknopen delen, dan is een **or**-matrix geschikt. t_j wordt true zo gauw er een ouder wordt bekeken.

Representatieknopen. In paragraaf 3.3.2.1 en verder is het gebruik van link matrices voor representatieconcepten beschreven. Voor Boolese retrieval met binaire indexering en ongewogen termen kan een **or**-matrix worden gebruikt. In het geval er met gewichten als term frequentie (*tf*) en geïnverteerde document frequentie (*idf*) of beide (*tf.idf*) wordt gewerkt, wordt er een gewogen-som link matrix gebruikt.

Queryconceptknopen. Hier is eigenlijk het probleem dat de verbindingen tussen de representatieconcepten en de queryconcepten niet duidelijk zijn. Hier speelt de semantiek een heel belangrijke rol. Hier is dan de vraag hoe dan met betrekking tot de waarschijnlijkheden de afhankelijkheid van de concepten in de query van de representatieconcepten geschat moet worden. Als wordt aangenomen dat de concepten identiek zijn en de gebruiker kent de representatieconcepten op basis waarvan dan ook de queries worden geformuleerd, dan kan men dezelfde link matrices gebruiken als die voor de tekstknopen.
In de praktijk geeft een slechte schatting van de afhankelijkheid van de queryconcepten van de representatieconcepten slechte resultaten. Er wordt onderzoek gedaan naar middelen waarmee de schattingen kunnen worden verbeterd.

Queryknopen. Over de afhankelijkheid van de queryknopen van de queryconceptknopen kan men kort zijn. Voor Boolese queries wordt de procedure die in de paragraaf over Boolese retrieval is beschreven, gebruikt. Voor probabilistische queries maakt men gebruik van gewogen-som matrices. In beide gevallen kunnen de waarden in de matrices worden aangepast op basis van de informatie over de relatieve belangrijkheid van de queries.

Informatiebehoefte. De informatiebehoefte kan algemeen worden uitgedrukt als een klein aantal queries van verschillende soorten (Bools, m-op-n afbeelding, op basis van kansen, natuurlijke taal, ...). De queries kunnen worden gecombineerd met gebruikmaking van een gewogen-som matrix met gewichten, die kunnen worden aangepast, zo dat ze het oordeel van de gebruiker met betrekking tot de belangrijkheid of de compleetheid van de individuele queries weergeven.

3.4 Het index expressie vertrouwensnetwerk model

In information retrieval is het in het algemeen moeilijk om te bepalen of documenten voor een gegeven query relevant zijn. Als de query wordt geformuleerd in termen die aan documenten wordt toegekend, dan is het niet zo'n probleem, maar zo gauw de gebruiker die deze query heeft opgesteld een andere interpretatie heeft van de gebruikte termen of er wordt een eigen jargon gebruikt waarin de informatiebehoefte wordt uitgedrukt, dan is het een groot probleem om de juiste documenten te krijgen. Men kan om dit probleem op te lossen twee dingen doen:

1. Men laat een expert de relaties tussen termen voor documenten en termen in queries leggen. Daarvoor moet de expert de gebruikers "kennen" en weten wat de karakterisatie van de documenten is. Op basis daarvan moet de expert dan met betrekking tot de relevantie beoordelen of er relaties gelegd worden en hoe sterk die relaties zijn. De sterkte wordt meestal uitgedrukt in termen van waarschijnlijkheden, gewichten en dergelijke. Retrieval systemen die op deze manier zijn gebouwd heten retrieval systemen op *empirische basis*.
2. De semantieken van de documenten en de queries in het retrieval proces betrekken. Dit is een vrij nieuwe richting in information retrieval waarnaar nu volop onderzoek gedaan wordt.

Bruza en Van der Gaag hebben onderzoek gedaan op het gebied van semantieken. In dat onderzoek is als uitgangspunt het redeneren met onzekerheid door middel van logische afleiding genomen. Het principe van logische afleiding is hetzelfde als afleiding dat in de wiskundige logica bekend is ([Vel84]). Het idee van redeneren met logische afleidingen is afkomstig van Rijsbergen ([Rij86]). Het als voorlopig resultaat van onderzoek hier te presenteren mechanisme voor het ontsluiten van informatie is op het genoemde uitgangspunt gebaseerd. Het mechanisme wordt hier een *verfijningsmachine* (*refinement machine*) genoemd. De voorlopige resultaten staan ook in [BG92]. Het zijn voorlopige resultaten omdat de verfijningsmachine niet volledig uitontwikkeld is.
De verfijningsmachine werkt concreet volgens het principe van een index expressie vertrouwensnetwerk model. Deze machine heeft een aantal kenmerken:

- zij werkt met en op index expressies;
- zij werkt met afleidingsregels voor zowel strikte (absolute) afleiding als plausibele afleiding. Hiermee wordt gepoogd uit een karakterisatie van docuemnten de query af te leiden;
- Bij invoer van een document als aanwijzing bepaalt zij met behulp van een vertrouwensnetwerk de relevantie van het document voor een gegeven query.

De verfijningsmachine en dus het onderliggende model zal hierna worden besproken, zij het niet zo formeel. Daar waar het nodig is zal er formeel worden gewerkt. Voor een preciezere beschrijving wordt verwezen naar [BG92].

3.4.1 De beschrijvingstaal: index expressies

Wil een verfijningsmachine uit een karakterisatie van documenten de query af kunnen leiden, moet er een taal zijn waarmee de documenten worden beschreven. Hetzelfde geldt ook voor queries.
Men begon eerst met de karakterisatie van documenten met behulp van trefwoorden of index termen als algemene kreet. Met deze *term descriptoren* gaat dat wel aardig goed, alleen in grote databases met veel documenten als informatie objecten wordt het zeer moeilijk om nog documenten te onderscheiden. Daarop zijn de mogelijkheden voor karakterisatie uitgebreid door trefwoorden

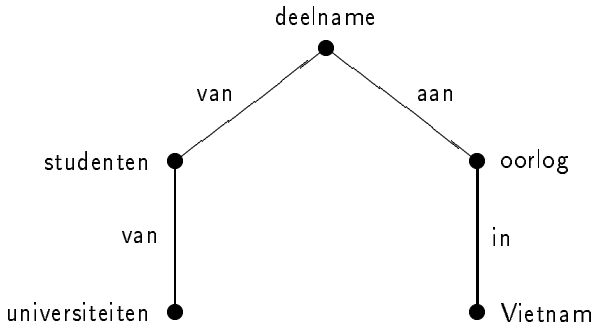
samen te stellen. De frase computer programmeren is dus bijvoorbeeld uit de trefwoorden computer en programmeren samengesteld. Deze *samengestelde term descriptoren* (*term phrase descriptor*) hebben zo meer onderscheidingskracht. De frase computer programmeren is specifieker dan diens losse componenten.
De beschreven karakterisaties zijn automatisch te genereren. Alleen de generatie van samengestelde descriptoren levert een voorraad descriptoren waarvan een klein deel van betekenis is en het resterende deel in principe in de prullenbak kan, omdat de betekenis ver te zoeken is. Deze descriptoren zijn naar alle waarschijnlijkheid gebrekkig en hebben dus een aanvulling nodig.

Omdat de samengestelde term descriptoren nu ook niet geheel geschikt zijn, zijn deze uitgebreid tot wat nu *index expressie descriptoren* zijn. Deze index expressies bestaan ook hier uit samengestelde termen, maar met dit verschil dat tussen de termen *connectoren* toegepast worden. Deze connectoren zijn voegwoorden zoals of, bij, met, en, als, naar, enzovoorts. Samengestelde termen krijgen zo opeens meer betekenis. De expressie expert in bommen 'zegt' meer dan de frase expert bommen.
Index expressies hebben hier een bepaalde structuur. De beschrijvingstaal voor deze expressies laat zich dan ook formeel definiëren zoals hieronder gedaan is.

Definitie: Zij T een verzameling termen en C een verzameling connectoren. De taal $\mathcal{L}$(T,C) van index expressies over T en C wordt door de volgende syntax gedefinieerd (in uitgebreide BNF):

$$\begin{aligned} &\mathsf{Expr} \rightarrow \epsilon \mid \mathsf{Sexpr} \\ &\mathsf{Sexpr} \rightarrow \mathsf{Term}\,\{\mathsf{Connector}\ \mathsf{Sexpr}\}^{\star} \\ &\mathsf{Term} \rightarrow \mathsf{t}\,,\ \mathsf{t} \in \mathrm{T} \\ &\mathsf{Connector} \rightarrow \mathsf{c}\,,\ \mathsf{c} \in \mathrm{C} \end{aligned}$$

Hierin stelt ϵ een lege index expressie voor.



Figuur 3.6: Een voorbeeld van een expressie en diens boomachtige structuur

In de definitie correspondeert de term t met een zelfstandig naamwoord, een bijvoeglijke naamwoord of een gezegde. De connector c geeft het soort relatie tussen twee termen weer. De soorten connectoren zijn hier beperkt tot deze die de samenhang in een bepaalde context zetten en de zogenaamde *null-connector* (notatie: $\circ$) die daar voorkomt waar een connector in feite niet hoeft. Deze null-connector is erbij geplaatst om de structuur van de index expressies te bewaren volgens de definitie van de taal $\mathcal{L}$(T,C).
De structuur van de index expressies is boomachtig. Deze kan voor een index expressie worden verkregen door element voor element de index expressie af te lopen. In het geval men bij een connector is, kan men op basis van belangrijkheid van deze connector besluiten de boom dieper of breder te maken. Het idee erachter is de visie dat sommige connectoren een sterkere relatie tussen twee termen bewerkstelligt dan andere. Die connectoren die een sterkere binding maken leiden tot het dieper maken van de structuur. Een voorbeeld van de structuur van de index expressie deelname van studenten van universiteiten aan oorlog in Vietnam staat in figuur 3.6
Dat de structuur van een index expressie van belang is, wordt in de volgende paragrafen duidelijk.

3.4.1.1 Machtsverzameling van index expressies

Voor information retrieval wordt hier gebruik gemaakt van *Machtsverzamelingen van index expressies*. Zo'n machtsverzameling bestaat uit een verzameling van alle index subexpressies van een gegeven index expressie. Wat nou in een gegeven expressie precies subexpressies zijn, wordt gedefinieerd door een *is-subexpressie-van* relatie die met $\trianglelefteq$ wordt aangegeven. De machtsverzameling $\mathcal{P}$ voor een gegeven index expressie i wordt dan gedefinieerd als $\mathcal{P} = \{j | j \trianglelefteq i\}$. Het volgende voorbeeld laat zien hoe $\trianglelefteq$ werkt en geeft gelijk aan wat dan de machtsverzameling concreet is.

Voorbeeld 3.4.1

Beschouw de index expressie deelname van studenten van universiteiten aan oorlog in Vietnam. *Als men de index expressie grafisch beschouwt dan wordt de subexpressie gerepresenteerd door een subgraph van de index expressie in figuur 3.6. De index expressies* oorlog in Vietnam *en* deelname van studenten aan oorlog *zijn subexpressies van* deelname van studenten van universiteiten aan oorlog in Vietnam. *Daaren tegen is* studenten in Vietnam *geen subexpressie. Er is geen bijbehorende subgraph van de graph in figuur 3.6. Als naar de machtsverzameling wordt gekeken, dan vormen de elementen daarin een structuur dat een* tralie *is met* $\trianglelefteq$ *als onderliggende ordening. Een grafische representatie van de machtsverzameling van de bovengenoemde index expressie is gegeven in figuur 3.7.* □

De omvang van een machtsverzameling kan nadelig uitpakken. Als een index expressie i_n uit n termen bestaat, dan wordt het aantal subexpressies $s(i_n) = 2^{n-1} + n$ in het ergste geval. Dit geval betreft de structuur van een index expressie als zijnde een ondiepe boom met veel vertakkingen. In het beste geval is $s(i_n) = \frac{1}{2}n(n+1) + 1$ voor een structuur van een expressie als zijnde een boom met slechts één pad waarop alle subexpressies liggen. In de praktijk komt het gelukkig weinig voor dat een expressie een ondiepe brede boom als structuur heeft. Een polynomiale omvang van de machtsverzameling is normaal, dus de bovengrens valt nog mee.

Tot nu toe is van één informatieobject (document) de indexexpressie als karakterisatie daarvan beschouwd. In een information retrieval systeem is een verzameling van informatieobjecten te vinden en dus wordt er ook een verzameling van bijbehorende karakterisaties in de vorm van index expressies gegenereerd. In deze *kern* $\mathcal{I}$ van index expressies heeft elke expressie een bijbehorende machtsverzameling. Wat dan een "machtsverzameling" van $\mathcal{I}$ wordt, is een vereniging van machtsverzamelingen van index expressies in de kern, ofwel:

$$\bigcup_{i \in \mathcal{I}} \mathcal{P}(i)$$

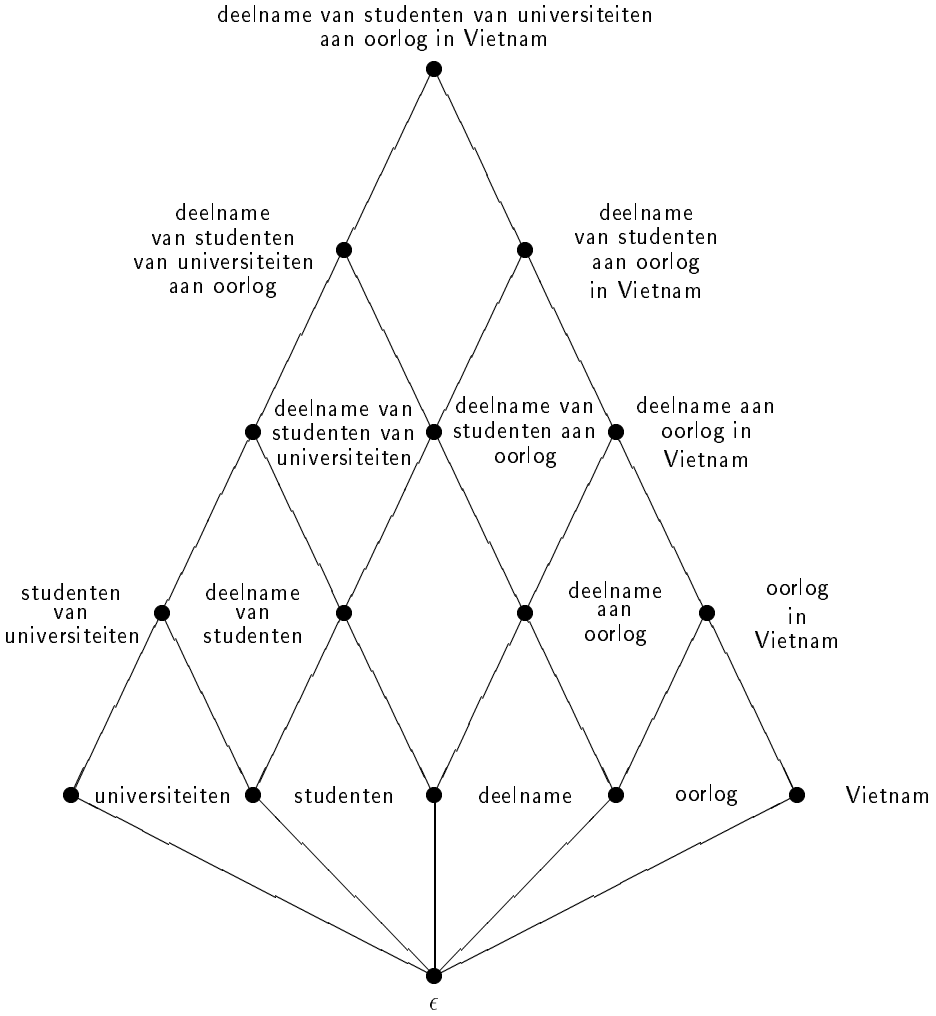
Dit is niet zo triviaal. De machtsverzameling van de index expressies lucht vervuiling in Holland en effecten van vervuiling in rivier hebben de index expressies vervuiling en ϵ gemeen. De vereniging van de machtsverzamelingen van de genoemde index expressies levert een structuur op die in figuur 3.8 is weergegeven. Daarin zijn de termen afgekort. Deze structuur wordt een *lithoïde* genoemd, wegens overeenkomsten met een kristallijnenstructuur.

3.4.2 Regels voor afleiding

Nu er een taal is gedefinieerd, waarmee documenten kunnen worden gekarakteriseerd en queries kunnen worden geformuleerd, kan de verfijningsmachine proberen uit de karakterisatie van documenten de gegeven query af te leiden. Daartoe staat er een aantal afleidingsregels tot diens beschikking. Voor de verfijningsmachine zijn er de strikte afleidingsregels *Modus Continens*, *Modus Generans*, *Modus Substituens* en de plausibele afleidingsregels *plausibele verfijning* en *plausibele substitutie*. Deze regels worden in de volgende paragrafen besproken.

3.4.2.1 Strikte afleidingsregels

Hier worden de afleidingsregels *Modus Continens*, *Modus Generans*, en *Modus Substituens* behandeld. De verfijningsmachine kan concreet gezien tot nu toe alleen met de afleidingsregel *Modus Continens* werken. Voor de overige regels wordt nog onderzoek gedaan. Hieronder volgt een puntsgewijze beschrijving van de afleidingsregels.



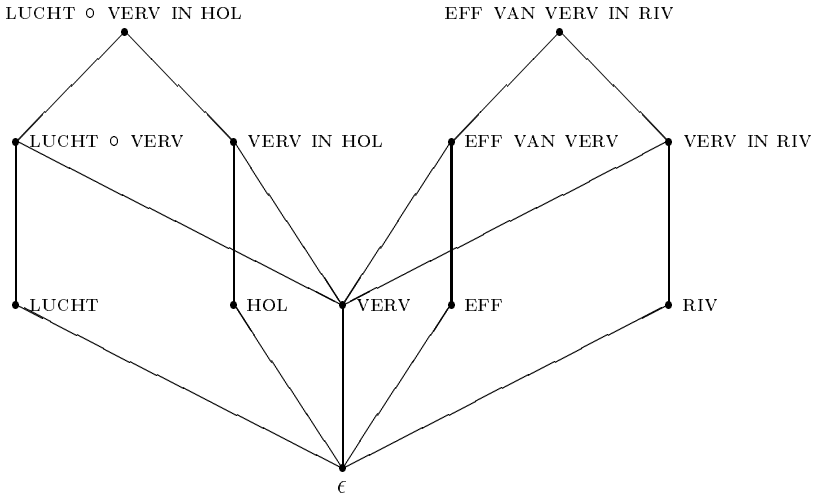
Figuur 3.7: Een voorbeeld van een machtsverzameling van index expressies, geordend volgens de relatie is-subexpressie-van

- *Modus Continens*
  Het principe achter *Modus Continens* is dat van een gegeven index expressie die op een object betrekking heeft, ook de subexpressie betrekking heeft op het object. Als de expressie vervuiling in rivier op een object slaat, dan slaat ook vervuiling en rivier op het object. Het algemene principe is weergegeven in de definitie hieronder.

  Definitie: Zij i en j index expressies in de taal $\mathcal{L}$(T,C) en zij $\trianglelefteq$ de is-subexpressie-van relatie over $\mathcal{L}$(T,C). Als j een subexpressie van i is, dan kan j uit i afgeleid worden, ofwel:

$$j \trianglelefteq i \quad \Rightarrow \quad i \vdash_{\mathrm{MC}} j$$

- *Modus Generans*
  Het basisprincipe van deze afleidingsregel is de *generalisatie*. Om echter iets te generaliseren, moet men eerst weten wat er algemener gemaakt kan worden. Met behulp van een ISA-relatie kan worden aangegeven welke termen er gegeneraliseerd worden. Als bijvoorbeeld zalm ISA vis is gegeven, dan kan men afleiden dat als een object over zalm gaat, deze dan ook over vis gaat. De volgende definitie geeft de werking van de afleidingsregel weer.



Figuur 3.8: De lithoïde als vereniging van machtsverzamelingen

Definitie: Zij $\mathcal{L}$(T,C) de taal voor de index expressies en zij $i, j \in \mathcal{L}$(T,C). Zij ISA $\subseteq$ T $\times$ T. Als i ISA j, dan kan j uit i worden afgeleid, ofwel:

$$i \text{ ISA } j \quad \Rightarrow \quad i \vdash_{\text{MG}} j$$

Over de ISA-relatie moet nog opgemerkt worden dat deze met betrekking tot homoniemen problemen kan geven. Als bijvoorbeeld gegeven is dat kraan ISA tappunt en kraan ISA hijsmachine, dan kan het gebruik van *Modus Generans* wel eens vervelende resultaten opleveren. Voorzichtigheid met het gebruik van deze regel is dus geboden.

- *Modus Substituens*
  In deze afleidingsregel is het pricipe het vervangen van subexpressies uit een expressie door andere subexpressies. Dit gaat bijvoorbeeld als volgt. In het voorbeeld bij de regel *Modus Continens* kon vervuiling worden afgeleid uit vervuiling in rivier. Dan kan *Modus Substituens* worden toegepast om voor een object dat over effecten van VERVUILING IN RIVIER op milieu gaat, af te leiden dat deze ook over effecten van VERVUILING op milieu gaat. Dit principe is in de onderstaande definitie weergegeven.

  Definitie: Zij k en i index expressies in de taal $\mathcal{L}$(T,C) zo dat i een index subexpressie van k is. Zij verder k_i^j de index expressie met daarin i gesubstitueerd door j. Dan geldt het volgende:

$$i \vdash j \quad \Rightarrow \quad k \vdash_{\text{MS}} k_i^j$$

Met deze regel wordt de mogelijkheid van contextvrije substitutie geboden. Het probleem is dan echter dat door substitutie allerlei mogelijke index expressies verkregen kunnen worden, ook deze die nergens op slaan. Uit bijvoorbeeld de expressie effecten van vervuiling van rivier in Australië kan men effecten van rivier afleiden. De context is echter zoek.
Om dergelijke "wilde" substituties te voorkomen, moeten beperkende voorwaarden aan de regel gesteld worden. Er zijn tot nu toe twee mogelijkheden voor beperking:

1. substitutie door generalisatie:

$$i \text{ ISA } j \;\Rightarrow\; k \vdash_{\text{MS}} k_i^j$$

   Dus in de expressie k mag i slechts door j vervangen worden als j een generalisatie van i is.

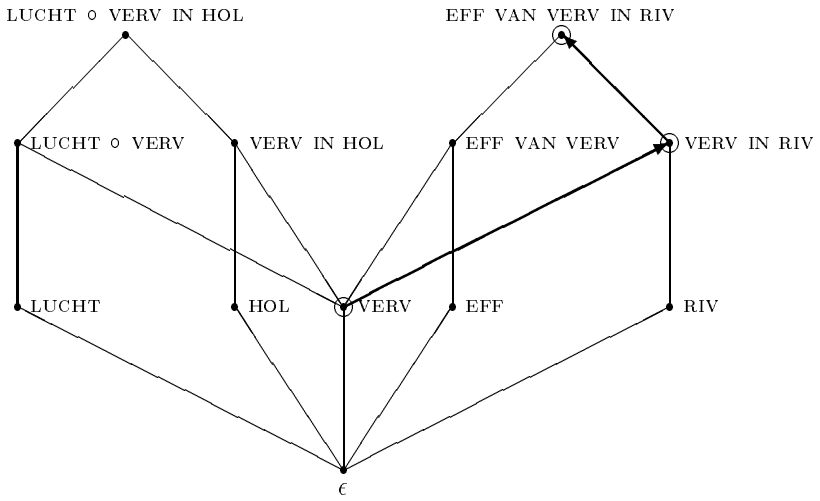
2. Afleiding uit een opeenvolging van termen.
   Substituties mogen dan alleen plaatsvinden met subexpressies waarin slechts null-connectoren voorkomen. Van bijvoorbeeld de expressie kleine ∘ groene ∘ marsmannen is de context nog aanwezig in de (sub-)expressie groene ∘ marsmannen. Als men zich beperkt tot steeds de eerste term-connector paar weg te halen (zoals in de genoemde expressies), dan gaat de afleiding in praktijk vaak goed. Zoals nu hierboven beschreven is hoe de subexpressies verkregen moet worden, kan men nu substitutie toepassen dat een geldige afleiding oplevert:

$$\text{invasie van KLEINE} \circ \text{GROENE} \circ \text{MARSMANNEN} \vdash \text{invasie van GROENE} \circ \text{MARSMANNEN}$$

   Bij deze manier van substitutie wordt er dus meer op de context gelet.

3.4.2.2 Plausibele afleidingsregels

Het kan gebeuren dat met strikte afleiding niet datgene bereikt wordt, als men wel wil. Men kan dan met de beschikbare index expressies niet uit de voeten. Aanvullingen zijn dus gewenst.
Een middel is om index expressies te *verfijnen*. Verfijning wordt hier bereikt door aan een expressie een term-connector paar toe te voegen. Het principe wordt duidelijk als men naar figuur 3.9 kijkt. Daarin kan de expressie vervuiling worden verfijnd in vervuiling in rivier die op zijn beurt weer in



Figuur 3.9: Het principe van verfijnen van index expressies

effecten van vervuiling in rivier kan worden verfijnd.
Dergelijke verfijningen kan men direct verkrijgen door gebruik te maken van een geïnverteerde $\trianglelefteq$-relatie over de taal $\mathcal{L}$(T,C) van index expressies. De verfijning kan echter ook door de geïnverteerde ISA-relatie gedefinieerd worden. De expressie vis kan bijvoorbeeld worden verfijnd tot zalm. De precieze definitie van de verfijning is hieronder gegeven.

Definitie: zij i en j index expressies in de taal $\mathcal{L}$(T,C). We zeggen dat i *verfijnd* kan worden tot j (notatie $i \rightarrow\!\!\!\!\rightarrow j$) dan en slechts dan als aan een van de volgende voorwaarden wordt voldaan:

1. $i \trianglelefteq j$ en voor alle index expressies k zo dat $i \trianglelefteq k \trianglelefteq j$ geldt: $k = i \vee k = j$
2. i ISA j en voor alle index expressies k zo dat i ISA k ISA j geldt: $k = i \vee k = j$

Nu kan verfijning worden gebruikt voor plausibele afleiding. Door verfijning te gebruiken wordt door de afleidingen index expressies verkregen die van origine geen karakterisaties van informatie-objecten zijn, en die niet door strikte afleiding te verkrijgen zijn. Er zijn twee regels voor plausibele afleiding. Deze zijn:

- *Plausibele afleiding door Verfijning*
  Het principe van verfijning wordt hier als volgt gebruikt. Als men aanneemt dat in de verzameling informatieobjecten die door vervuiling wordt gekarakteriseerd er objecten zijn die gaan over vervuiling van rivier, dan zou men op basis daarvan de index expressie vervuiling van rivier uit vervuiling kunnen afleiden. Over de zekerheid van deze afleiding kan nu nog niets gezegd worden, maar het kan wel als plausibel worden beschouwd. De definitie van deze afleidingsregel is als volgt:

  Definitie: zij i en j index expressies in de taal $\mathcal{L}$(T,C). Als i verfijnd kan worden tot j, dan is j uit i af te leiden en wel plausibel, ofwel:

  $$i \rightarrowtail j \quad \Rightarrow \quad i \mid\sim_{\mathrm{PR}} j$$

- *Plausibele substitutie*
  Deze afleidingsregel heeft sterke overeenkomsten met de afleidingsregel *Modus Substituens*. Hier kan worden volstaan met een definitie.

  Definitie: zij k en i index expressies in de taal $\mathcal{L}$(T,C) zo dat i een index subesxpressie van k is. Zij verder k_i^j de index expressie met i gesubstitueerd door j. Dan:

  $$i \rightarrowtail j \quad \Rightarrow \quad k \mid\sim_{\mathrm{PR}} k_i^j$$

Voor het inzicht in de werking van de plausibele afleidingsregels wordt er een voorbeeld gegeven.

Voorbeeld 3.4.2
Met behulp van de plausibele afleidingsregels wordt getoond hoe de index expressie metalen *uit de index expressie* vervuiling van rivieren *kan worden afgeleid. Een afleiding met behulp van* plausibele afleiding door verfijning *gaat als volgt:*

$$\begin{aligned} &\text{vervuiling van rivieren} && \vdash_{\mathrm{MC}} \\ &\text{vervuiling} && \mid\sim_{\mathrm{PR}} \\ &\text{vervuiling door metalen} && \vdash_{\mathrm{MC}} \\ &\text{metalen} \end{aligned}$$

Een afleiding waarin plausibele substitutie *wordt gebruikt, gaat als volgt:*

$$\begin{aligned} &\text{effecten van VERVUILING VAN RIVIEREN} && \vdash_{\mathrm{MC}} \\ &\text{effecten van VERVUILING} && \mid\sim_{\mathrm{PS}} \\ &\text{effecten van VERVUILING DOOR METALEN} && \vdash_{\mathrm{MC}} \\ &\text{effecten van METALEN} \end{aligned}$$

□

3.4.2.2.1 Problemen met plausibele afleiding

Men kan zich nu afvragen hoe goed de regels zijn. Daartoe worden ze gewoon uitgeprobeerd. Beschouw de objecten O_1, O_2 en O_3. De karakterisaties (notatie χ) van elke object staat hieronder.

$$\begin{aligned} \chi(O_1) &= \{\text{rivier} \circ \text{vervuiling in Australië}\} \\ \chi(O_2) &= \{\text{effecten van vervuiling in rivier}\} \\ \chi(O_3) &= \{\text{lucht} \circ \text{vervuiling in Holland}\} \end{aligned}$$

Stel nu dat de query rivier ∘ vervuiling is. Op basis van intuïtie kan men zeggen dat O_1 en O_2 relevant zouden zijn terwijl O_3 dat niet is.
De query wordt nu in de verfijningsmachine gevoerd. Deze mechanisme kan werken met de afleidingsregels *Modus Continens* (MC), *Modus Generans* (MG), *Modus Substituens* (MS), *Plausibele afleiding door Verfijning (Refinement)* (PR) en *Plausibele Substitutie* (PS). De verfijningsmachine wordt gestart en er wordt geprobeerd de query uit de gegeven karakterisaties af te leiden:

- uit de karakterisatie van O_1:

$$\text{rivier} \circ \text{vervuiling in Australië} \quad \vdash_{\text{MC}} \quad \text{rivier} \circ \text{vervuiling}$$

- uit de karakterisatie van O_2:

$$\begin{aligned} \text{effecten van vervuiling in rivier} \quad &\vdash_{\text{MC}} \\ \text{vervuiling} \quad &\mid\!\sim_{\text{PR}} \\ \text{rivier} \circ \text{vervuiling} \end{aligned}$$

- uit de karakterisatie van O_3:

$$\begin{aligned} \text{lucht} \circ \text{vervuiling in Holland} \quad &\vdash_{\text{MC}} \\ \text{vervuiling} \quad &\mid\!\sim_{\text{PR}} \\ \text{rivier} \circ \text{vervuiling} \end{aligned}$$

Hieruit blijkt dat O_1 met zekerheid relevant is, er is daar geen plausibele afleiding aan te pas gekomen. Als men aan O_2 en O_3 de waarschijnlijkheid van relevantie wil toekennen, dan zal de verfijningsmachine aan zowel O_3 als O_2 gelijke kans toekennen. Er is in de afleidingen van beide objecten precies dezelfde plausibele afleidingsstap gedaan. Dekans voor O_2 zal vrij hoog zijn gezien de overeenkomsten met de query. Dus de verfijningsmachine geeft als antwoord op de query O_1, O_2 en O_3 terug.
Het zal duidelijk zijn dat dit niet goed kan zijn. Het probleem zit hem in het feit dat door toepassing van de afleidingsregels belangrijke delen van de context van de initiële karakterisatie van een object verloren gaat. Beschouw bijvoorbeeld de afleiding hieronder:

$$\text{effecten van vervuiling in rivier} \quad \vdash_{\text{MC}} \quad \text{vervuiling}$$

Daarin wordt de context dat vervuiling noemt gewoon weggegooid en kan dan verder niet meer voor de afleiding gebruikt worden.

3.4.3 Het index expressie vertrouwensnetwerk

In de vorige paragraaf is aangegeven dat door gebruik van plausibele substitutie en plausibele verfijning (plausibele afleiding door verfijning) de resultaten voor een gegeven query niet correct kunnen zijn. De verfijningsmachine zou dus niet met die afleidingsregels moeten werken.
Plausibele afleiding moet dus op een andere manier gebeuren. De oplossing ligt in het redeneren met een vertrouwensnetwerk. Met behulp van zo'n netwerk kan men probabilistische uitspraken doen. Bij het netwerk zijn nog twee algoritmen van belang:

- een algoritme waarmee met behulp van het netwerk de waarschijnlijkheden die van belang zijn worden berekend, en
- een algoritme voor het propageren van aanwijzingen (*evidence propagation*). Dat wil zeggen, door in het netwerk een aanwijzing in te voeren, worden de (locale) kansverdelingen herberekend.

Het netwerk dient als architectuur voor de berekeningen die door de algoritmen worden uitgevoerd. In het als graph beschouwde netwerk zijn de punten de variabelen waarbij de bijbehorende informatie (kansen of iets anders) wordt opgeslagen, en drukken de lijnen (of pijlen) een relatie tussen

paren punten uit. Dit is slechts een algemeen principe. De interpretatie van de punten en lijnen (pijlen) hangt af van het soort algoritme dat voor het redeneren met het netwerk wordt gebruikt. Bekende algoritmen zijn die van Pearl ([KP83], [Pea86] of [Pea90], [Pea88]) en van Lauritzen en Spiegelhalter ([LS88] of [LS90]). Een korte samenvatting van deze genoemde algoritmen is in het hoofdstuk over expertsystemen en onzekerheid te vinden.

3.4.3.1 Constructie van het netwerk

Voordat tot de feitelijke constructie overgegaan wordt, wordt er eerst een formele definitie van het vertrouwensnetwerk gegeven. Tevens wordt er aangegeven hoe dan voor het netwerk in zijn geheel als graph de samengestelde kansverdeling wordt berekend.
De definitie van het netwerk is als volgt.

Definitie: Een vertrouwensnetwerk is een tupel $B = (G, \Gamma)$ waarin

1. $G = (V(G), A(G))$ een *gerichte acyclische graph* is met de punten $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, en
2. $\Gamma = \{\gamma_{V_i} | V_i \in V(G)\}$ een verzameling van niet-negatieve reële functies $\gamma_{V_i} : \{v_i, \neg v_i\} \times \{c_{\pi(V_i)}\} \rightarrow [0, 1]$ is, zo dat voor elke configuratie $c_{\pi(V_i)}$ van de verzameling ouders $\pi(V_i)$ van V_i in G geldt dat $\gamma_{V_i}(\neg v_i | c_{\pi(V_i)} = 1 - \gamma_{V_i}(v_i | c_{\pi(V_i)}$, $i = 1, \ldots, n$. γ_{V_i} wordt hier een *toekenningsfunctie (voor voorwaardelijke kansen)* genoemd.

De verzameling Γ kan erg groot worden. Als namelijk in een graph met n een punt i $n-1$ ouders heeft, dan zijn er 2^{n-1} combinaties van ouders en zijn er dus even zoveel functies γ_{V_i} nodig. Dit zal echter wel meevallen. Men zal zien dat in het grafische deel van het netwerk elk punt hoogstens twee ouders heeft, op de wortels na.

Nu zijn er toekenningsfuncties voor voorwwardelijke kansen gedefinieerd, maar hoe dan een (algehele) *samengestelde kansverdeling* wordt gedefinieerd, is nog de vraag. Er is een zo'n kansverdeling, maar dan moet er eerst aan een belangrijke voorwaarde voldaan zijn. Er wordt verondersteld dat het grafische deel van het vertrouwensnetwerk alle *onafhankelijkheidsrelaties* tussen de statistische variabelen die men in de graph onderscheidt representeert. Dit wil zeggen, een tweetal punten waartussen geen lijn loopt worden als onafhankelijk van elkaar beschouwd. En als tussen een tweetal punten een pijl loopt, dan is het ene punt van waaruit de pijl vertrekt onafhankelijk van het andere punt (en de rest die daarmee verbonden is).
In dit geval kan men voor een vertrouwensnetwerk $B = (G, \Gamma)$ met in de graph G de verzameling punten $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, de volgende Keywordsamengestelde kansverdeling definiëren:

$$Pr(C_{V(G)}) \quad = \quad \prod_{V_i \in V(G)} \gamma_{V_i}(V_i | C_{\pi(V_i)})$$

$C_{V(G)}$ stelt de conjunctie van de V_i voor en $C_{\pi(V_i)}$ de conjunctie van ouders van V_i. V_i wordt afwisselend gebruikt als punt en als *probabilistische variabele*, dat wil zeggen het punt V_i wordt als variabele beschouwd die de waarden $\{v_i, \neg v_i\}$ die *true* respectievelijk *false* voorstellen, aanneemt.

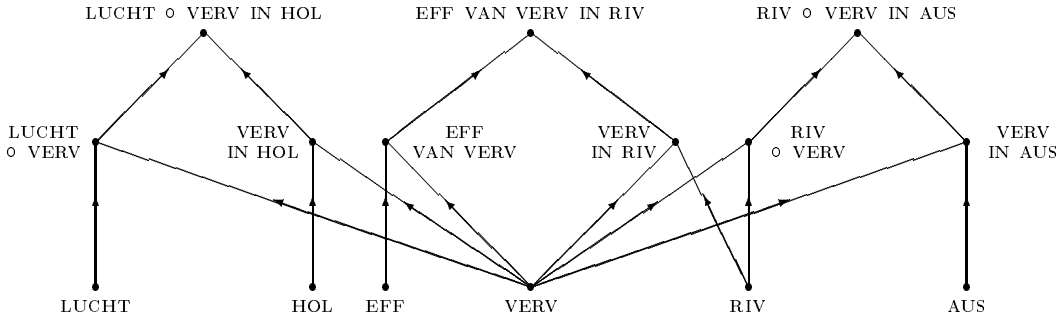
3.4.3.1.1 Het grafische deel

In paragraaf 3.4.1.1 is een machtsverzameling van index expressies ingevoerd. De vereniging van machtsverzamelingen van index expressies voor een gegeven kern als verzameling van index expressies werd een lithoïde. Deze lithoïde kan gebruikt worden voor het zogenaamde query by navigation. De punten (knopen) in de lithoïde kunnen als focus dienen met voor elke focus een te bepalen bijbehorende collectie van relevante documenten.
Als nu elke knoop in de lithoïde als een probabilistische variabele wordt beschouwd, dan geeft dit het idee dat de lithoïde gebruikt kan worden als grafisch deel van het netwerk. Hier is bijvoorbeeld een probabilistische variabele een punt met index expressie vervuiling in rivier die als een variabele wordt beschouwd, met de waarden {*vervuiling in rivier*, $\neg$*vervuiling in rivier*} die *true* respectievelijk *false* voorstellen.
De transformatie van de lithoïde in een grafisch deel van het netwerk gaat als volgt:

1. In de machtsverzameling van index experessies zijn de index expressies partieel geordend door de is-subexpressie-van relatie $\trianglelefteq$. Dus de punten als probabilistische variabelen zijn ook partieel geordend. Deze ordening is zo geschikt, dat de lithoïde kan worden genomen als zijnde een ongerichte topologie van het grafische deel van het netwerk.

2. De lijnen in de graph die is verkregen, geven een partiële ordening op de probabilistische variabelen weer. Maak deze lijnen gericht zo dat de richting wordt bepaald door van de geïnverteerde $\trianglelefteq$-relatie gebruik te maken. Men wil bijvoorbeeld weten wat de kans van een gegeven index expressie vervuiling in rivier is, gegeven de los van elkaar gegeven termen vervuiling en rivier.

3. De lege index expressie ϵ kan verwijderd worden. Deze is niet van nut voor retrieval.



Figuur 3.10: Een voorbeeld van een gerichte graph

Zie voor de duidelijkheid figuur 3.10 voor de informatieobjecten die in paragraaf 3.4.2.2.1 gegeven zijn.

3.4.3.1.2 Toekenning van waarden aan variabelen

Nu er een grafisch deel van het netwerk is, moet er nog aangegeven worden wat voor waarden de toekenningsfuncties bij elke probabilistische variabele kunnen aannemen. De graph heeft wortels en de overige punten. Voor elk van deze soorten punten wordt hieronder de invulling beschreven.

- De wortels
  Deze variabelen hebben geen ouders, dus krijgen zij a priori kansen toegekend. Deze variabelen corresponderen met unaire index expressies, of termen. Het ligt enigzins voor de hand om de frequenties van voorkomen van termen binnen documenten (term frequenties) te gebruiken. Dus voor elke variabele T voor een term t wordt dan de waarde voor de toekenningsfunctie $\gamma_T(t)$ die bij T hoort berekend en wel als volgt:

  $$\gamma_T(t) \quad \approx \quad \eta f(t),$$

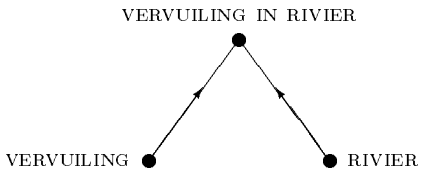
  waarin η een een of andere normalisatiefactor is, en $f(t)$ de termfrequentie is. De complementaire functie $\gamma_T(\neg t)$ wordt bepaald door de vergelijking $\gamma_T(\neg t) = 1 - \gamma_T(t)$.
  Voor bijvoorbeeld de termen in in de lithoïde in figuur 3.10 krijgt men de volgende waarden voor γ_T:

t	$f(t)$	$\gamma_T(t)$
vervuiling	3	0.33
rivier	2	0.22
effecten	1	0.11
Australië	1	0.11
lucht	1	0.11
Holland	1	0.11

Dit idee is vrij algemeen in information retrieval. In het model van Wong en Yao ([WY91]) wordt het ook gebruikt (zie ook elders in dit hoofdstuk).

- De overige punten.
  Voor deze punten zijn twee gevallen te onderscheiden, te weten punten met wortels als ouders en punten met ouders die geen wortels zijn. Deze gevallen worden hieronder behandeld.

  – Punten met wortels als ouders.
    Voor het gemak wordt er een stukje uit de lithoïde in figuur 3.10 gehaald, dat hieronder gegeven wordt:

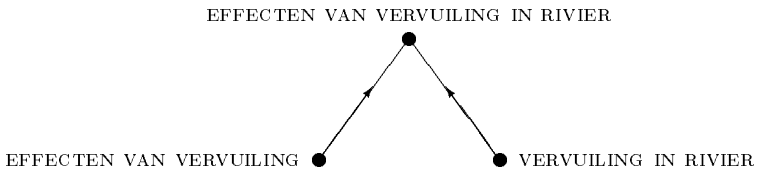


    Voor het gemak wordt `VERVUILING IN RIVIER` afgekort tot `V IN R`. Voor de functie γ`V IN R` zijn acht waarden te specificeren; de vier functiewaarden

$$\begin{aligned}
\gamma\texttt{V IN R}(\text{vervuiling in rivier}|\text{vervuiling} \wedge \text{rivier}) &= w \\
\gamma\texttt{V IN R}(\text{vervuiling in rivier}|\neg\text{vervuiling} \wedge \text{rivier}) &= x \\
\gamma\texttt{V IN R}(\text{vervuiling in rivier}|\text{vervuiling} \wedge \neg\text{rivier}) &= y \\
\gamma\texttt{V IN R}(\text{vervuiling in rivier}|\neg\text{vervuiling} \wedge \neg\text{rivier}) &= z
\end{aligned}$$

    en de complementaire waarden. De eerste toekenningsfunctie heeft de volgende betekenis. Gegeven dat men weet dat een object O over vervuiling gaat en ook dat O over rivier gaat. De kans dat O dan over vervuiling in rivier gaat is dan gelijk aan w.
    De waarde die wordt ingevuld hangt af van het feit hoe vaak een connector in een verzameling index expressies voorkomt. Een waarde $w \approx 0.06$ geeft bijvoorbeeld aan dat 6% van de index expressies de connector in bevat.
    Wat de overige waarden betreft: het is bewezen (in [BG92]) dat als een van de bovenstaande functie een waarde groter dan 0 is, dat dan de overige bovenstaande functies noodzakelijkerwijs de waarden 0 aanneemt.

  – Punten met ouders die geen wortels zijn.
    Hier gaat het om variabelen die index expressies met meer dan twee termen representeren. Voor het gemak wordt ook hier een stukje uit de lithoïde in figuur 3.10 hierheen gehaald:



    Wat wordt dan de waarden voor de bijbehorende toekenningsfuncties? Uit de constructie van de lithoïde zal blijken dat een index expressie van n termen steeds wordt gevormd door een combinatie van twee onderliggende index subexpressies van $n-1$ termen die elkaar in één term overlappen ($n \geq 3$). Dus de index expressie effecten van vervuiling in rivier wordt gecombineerd door de index expressies effecten van vervuiling en vervuiling in rivier op basis van de term vervuiling die in beide expressies voorkomt.
    Het is uit onderzoek gebleken dat de boven beschreven combinatie vrijwel eenduidig gebeurt. Dus kan men voor twee index expressies i en j die tot k wordt gecombineerd

voor de functies γ_K de volgende waarden invullen:

$$\begin{aligned}\gamma_K(k|i \wedge j) &= 1\\ \gamma_K(k|\neg i \wedge j) &= 0\\ \gamma_K(k|i \wedge \neg j) &= 0\\ \gamma_K(k|\neg i \wedge \neg j) &= 0\end{aligned}$$

De waarden voor de laatste drie functies volgt weer noodzakelijkerwijs.

3.4.3.2 Gebruik van het netwerk

Nu zo'n netwerk geheel geconstrueerd is, kan er nu aangegeven worden hoe het door de verfijningsmachine gebruikt wordt. Eerst wordt aangegeven hoe er wordt bepaald of een informatieobject relevant is voor een gegeven query. Dit is in de onderstaande definitie beschreven.

Definitie: Zij $\mathcal{L}$(T,C) de taal van index expressies. Zij O een object, $\chi(O) \subseteq \mathcal{L}$(T,C) diens karakterisatie en zij $q \in \mathcal{L}$(T,C) een verzoek. Zij verder Pr een samengestelde kansverdeling gedefinieerd op $\mathcal{L}$(T,C). Dan is de *waarschijnlijkheid van relevantie* van O voor q, notatie $P_{\mathrm{Rel}}(O,q)$, als volgt gedefinieerd:

$$P_{\mathrm{Rel}}(O,q) \quad = \quad \max\{Pr(q|i) \mid i \in \chi(O)\}$$

Een alternatieve definitie van de waarschijnlijkheid van relevantie zou zijn door de gehele karakterisatie van een document als aanwijzing te nemen, ofwel:

$$P_{\mathrm{Rel}}(O,q) \quad = \quad \{Pr(q|\chi(O)\}$$

Op welk van deze twee de verfijningsmachine de beste resultaten geeft, wordt nog onderzocht.

Het mechanisme voor strikte afleiding van de verfijningsmachine kan goed worden gecombineerd met deze probabilistische benadering. Als een object O en een verzoek q opnieuw beschouwd wordt en er is een index expressie $i \in \chi(O)$ waaruit q door strikte afleiding wordt verkregen, dan is de waarschijnlijkheid van q gegeven i maximaal, ofwel:

$$i \vdash q \quad \Rightarrow \quad Pr(q|i) = 1$$

In dit geval is dus de waarschijnlijkheid van relevantie van O voor q gelijk aan 1. Verder zal uit het geval dat slechts $Pr(q|i) = 0$ voor elke $i \in \chi(O)$ de verfijningsmachine de conclusie trekken dat O niet relevant voor q is.

Het was met dit netwerk de bedoeling dat het een oplossing was voor plausibele afleiding. Deze afleiding moet hier dan nog even gedefinieerd worden.
Met behulp van een voorbeeld wordt getoond hoe plausibel redeneren met het index expressie vertrouwensnetwerk gaat. Kijk nog eens naar figuur 3.10. Uit de twee afzonderlijke termen vervuiling en rivier zou de binaire index expressie vervuiling in rivier afgeleid worden met kans Pr(vervuiling in rivier | vervuiling $\wedge$ rivier). In termen van logica is dit gelijk aan de plausibele afleidingsstap

$$\text{vervuiling, rivier} \quad \mid\!\sim \quad \text{vervuiling in rivier}$$

Deze afleidingsstap kan worden gezien als een stap waarin de connector in is "gegokt"; hoe sterk die gok is, hangt af van de bijbehorende voorwaardelijke kans.
Deze voorbeeldafleiding kan worden gegeneraliseerd tot een mechanisme voor plausibele afleiding voor de verfijningsmachine. Deze afleiding is als volgt gedefinieerd:

Definitie: Zij $i_1, \ldots, i_n$, $n \geq 1$, en k index expressies in $\mathcal{L}$(T,C). Dan geldt het volgende:

$$Pr(k \mid i_1 \wedge \cdots \wedge i_n) > 0 \quad \Rightarrow \quad i_1, \ldots, i_n \mid\!\sim_{\mathrm{PI}} k$$

Deze afleidingsregel wordt *plausibele afleiding door deductie* genoemd.

Hoe komt men nu aan een waarde voor Pr?. In het geconstrueerde netwerk zijn slechts *lokale* voorwaardelijke kansen gegeven. Voor elke $\gamma(k|i \wedge j)$ is er van uitgegaan dat de index expressies i en j absoluut waar of onwaar zijn. De index expressies kunnen op hun beurt weer van andere index expressies afhangen.
Als nu een query k in de vorm van een index expressie is gegeven, dan moet voor elk document de Pr bepaald worden. Deze verkrijgt men door de karakterisatie van het desbetreffende document (object) door het netwerk te propageren. Daartoe is een algoritme vereist. Bruza en Van der Gaag hebben het algoritme van Lauritzen en Spiegelhalter ([LS90]) gebruikt. De Pr worden bekend, waarna de grootste Pr als maat voor relevantie van een document voor de query (dus $P_{\mathrm{Rel}}(O, k)$) wordt genomen.

3.4.4 Slotopmerkingen

In de voorgaande paragrafen is de verfijningsmachine en het onderliggende model beschreven. Zo veel als er geschreven is over de verfijningsmachine, concreet kan hij nog niet alles. Tot nu toe kan de verfijningsmachine alleen met de regels *Modus Continens* en *plausibele afleiding door deductie* concreet werken. Dat lijkt niet veel, maar toch doet hij het aardig. Het voorbeeld dat in paragraaf 3.4.2.2.1 is nu goed opgelost (dankzij het vertrouwensnetwerk).

Hoofdstuk 4

Integraties in padexpressies

LISA-D is een taal dat in te zetten is bij informatiesystemen. Er kunnen gegevensmodellen als informatiestructuren in weergegeven worden. Een informatiestructuur kan (deels) gecontroleerd worden op fouten. Tevens kan deze gepopuleerd worden met de bedoeling om queries er op los te laten. Dit is te vergelijken met SQL (Structured Query Language), alleen het basisprincipe verschilt.
LISA-D is echter (nog) niet in te zetten bij information retrieval systemen en expertsystemen. De taal bevat namelijk geen faciliteiten om te kunnen redeneren met onzekere kennis. Expertsystemen en information retrieval systemen maken gebruik van modellen om onzekerheid van een kennis te representeren en om op een bepaalde manier met die onzekerheid om te gaan. LISA-D moet dan faciliteiten krijgen die de mogelijkheid bieden een model toe te passen en ermee te werken. In het onderzoek dat hiernaar gedaan wordt wordt dan de vraag gesteld:

> Kan dat? Wat moet er gedaan worden om die mogelijkheid te bieden en zijn er soms voorwaarden of eisen? En als het niet kan, waarom dan niet?

Om hier antwoord op te geven, wordt er een basismodel ontwikkeld. Met dit model moet dan nagegaan worden of met LISA-D modellen uit information retrieval systemen en expertsystemen kunnen worden toegepast en of daar dan ook mee te werken valt.
Er zijn inmiddels verscheidene modellen die bij expertsystemen en information retrieval systemen gebruikt worden, onderzocht. Er zijn drie zaken die expertsystemen met elkaar gemeen hebben:

1. de kennisrepresentatie. Kennis wordt gerepresenteerd in de vorm van productieregels. De relaties tussen de productieregels laat zich visualiseren door middel van een afleidingsnetwerk;
2. Het propagatiemechanisme. Zonder uitzondering worden aanwijzingen door het netwerk naar de hypothesen toe gepropageerd. De combinatiefunctie voor het propageren van onzekerheden in aanwijzingen speelt hierbij een belangrijke rol;
3. De combinatiefuncties, een voor het propageren van onzekerheden in aanwijzingen, twee voor samengestelde aanwijzingen middels **or** en **and**, en een voor co-concluderende regels.

Het onderzoek naar de bruikbaarheid van LISA-D voor expertsystemen zal zich op de drie genoemde punten richten. Aangezien tussen de modellen veel overeenkomsten zijn, wordt voor het onderzoek één van de modellen geselecteerd die als testvehikel kan dienen.
Daarentegen zijn de modellen van information retrieval systemen onderling zeer verschillend. Bij elk model is de manier waarop voor een document de relevantie wordt bepaald verschillend. De ene model gebruikt canonische link matrices of gewogen-som matrices ([TC90]), bij het andere model wordt met gewone breuken gewerkt, en weer een ander model gebruikt netwerkmodellen ([KP83, LS88]). Overeenkomsten zijn er niet, alleen dat bij elk model steeds de relevantie van een document gegeven een query bepaald wordt. Het gaat hier te ver om elk model gedetailleerd uit te werken. Er zal een model worden geselecteerd voor nader onderzoek.

4.1 Keuze van modellen

Er zal hier een model dat bij expertsystemen wordt gebruikt, worden geselecteerd zodat het als een zogenaamde testvehikel kan dienen. Het is de bedoeling dat er duidelijkheid wordt verschaft over de mogelijkheden en onmogelijkheden van LISA-D. Hetzelfde wordt ook gedaan met een model dat bij information retrieval systemen wordt gebruikt.

De keuze van de modellen geschiedt op basis van het volgende:

- Over het model is veel bekend, dat wil zeggen aspecten aan het model zijn zoveel mogelijk uitgewerkt. Het model kan dan vrijwel direct worden gebruikt.
- Het model is "eenvoudig" van opzet. De gedachte erachter is dat als een "simpel" model niet met LISA-D te gebruiken is, dat het met "moeilijke" modellen ook wel eens niet goed kan gaan.
- Het model is representatief voor varianten ervan. Dit is op basis van de gedachte dat elk model zodanig "uniek" is dat het een volledig eigen implementatie in LISA-D behoeft. De verschillen tussen de modellen komen al naar voren in maten voor onzekerheid.
- In het inleidende hoofdstuk is op bladzijde 4 is een aantal eisen gesteld die gerespecteerd moeten worden als een model wordt toegepast in LISA-D.

4.1.1 Een model voor expertsystemen

Er zijn vier modellen waar een keus uit gemaakt kan worden:

- De waarschijnlijkheidstheorie;
- De subjectieve Bayesische methode van Duda, Hart en Nilsson ([DHN90]);
- Het zekerheidsfactor model van Buchanan en Shortliffe ([BS84]);
- De Dempster-Shafer theorie.

De netwerkmodellen van Kim en Pearl, Lauritzen en Spiegelhalter ([LS88, KP83]) worden buiten beschouwing gelaten, omdat bij deze modellen het noodzakelijk is dat de gebruikte gegevens worden vernieuwd. Een padexpressie (of meerdere) dat het model moet simuleren, zal in LISA-D de gegevens in de onzekerheidspopulatie moeten vernieuwen.
Er zal met het zekerheidsfactormodel worden gewerkt. De subjectieve Bayesische methode is een variant hiervan. Het verschil zit hem in de representatie van onzekerheid. Bij het ene model wordt met zekerheidsfactoren gewerkt en bij het andere met odds. De functies verschillen inhoudelijk wel, maar het principe blijft hetzelfde.
De Dempster-Shafer theorie is ook een variant op het zekerheidsfactormodel. Daar wordt gewerkt met geloofwaardigheidsfuncties. Een probleem daar is echter dat de combinatieregel als combinatiefunctie rekentechnisch complex is. Gordon en Shortliffe geven in [GS90] wel aan dat het onder bepaalde voorwaarden is te gebruiken.
De waarschijnlijkheidstheorie is niet geschikt omdat het functies ontbeert die voor het gebruik in een expertsysteem min of meer vereist zijn.

4.1.2 Een model voor information retrieval systemen

Turtle en Croft ([TC90]) hebben een afleidingsnetwerk voor document retrieval ontwikkeld en daarbij aangegeven hoe de relevantie moet worden bepaald. Het model is echter vrij globaal beschreven. Relevantie worden gerepresenteerd in de vorm van canonische link matrices die bovendien op verscheidene plaatsen in het netwerk van formaat verschillen. De canonische link matrices representeren in feite een Boolese formule. In LISA-D zou men uit een multiset de probabilistische variabelen met de bijbehorende waarden (true, false) kunn halen, maar het is niet duidelijk en op

het eerste gezicht niet echt mogelijk, hoe men precies kan "zien" waar de waarden van de variabelen in de formule ingezet moeten worden. Dit model is daarom niet zo geschikt.

Bruza en Van der Gaag ([BG92]) hebben een index expressie vertrouwensnetwerkmodel ontwikkeld. Dit model kan niet goed gebruikt worden, omdat er bij het bepalen van de relevantie van een document een netwerkmodel wordt gebruikt. Voor het berekenen van de relevanties wordt bijvoorbeeld het model van Lauritzen en Spiegelhalter ([LS88]) gebruikt. In het hoofdstuk over modellen in expertsystemen is aangegeven dat dit niet bruikbaar is in LISA-D. Dit model valt jammer genoeg ook af. Het model is echter wel formeel, en het idee van een taal met daarop de afleidingsmechanismen ziet er bruikbaar uit.

Het laatste model dat nu dan over blijft, is met probabilistische afleidingsmodel van Wong en Yao ([WY91]). Dit model is gedetailleerd uitgewerkt en kan in principe zo gebruikt worden. Een paar kleine variaties in het model maakt dat het model een representatie wordt van het *gegeneraliseerde vectorruimte model* (*Generalized Vector Space Model* (GVSM)).
Uit een voorbeeldtoepassing van dit model blijkt dat met matrices wordt gewerkt.Dit laat zich in een informatiestructuur met een populatie goed representeren, zoals later zal blijken. Het zal duidelijk zijn dat voor dit model gekozen zal worden.
De genoemde modellen zijn trouwens niet de enige modellen in information retrieval. Er zijn nog andere modellen maar die zijn hier niet bestudeerd. We noemen er een paar: een probabilistisch model voor integratie van IR en databases van Fuhr ([Nor93]), niet-klassieke logica van Rijsbergen ([Rij86]).

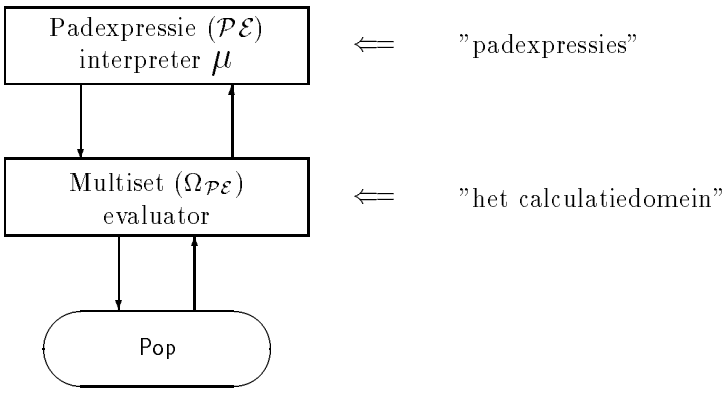
4.2 Het basismodel: representatie van onzekerheid

In LISA-D kan kennis als een populatie van een informatiestructuur als databasemodel weergegeven worden. Maar om een maat dat de onzekerheid van een stukje kennis weergeeft toe te kennen aan het desbetreffende stukje kennis, daartoe biedt LISA-D geen faciliteiten. Deze moeten eerst ontwikkeld worden. Er wordt hier eerst een basismodel gegeven, dat zoveel mogelijk algemeen te gebruiken is voor allerlei methoden en modellen waarin op bepaalde manieren met onzekerheid wordt gewerkt. Verder zal er worden gelet op de eigenschappen en kenmerken van het werken met het basismodel in LISA-D. Er zullen hier enkele voorbeelden gegeven worden die de aspecten verduidelijken.
Tot zover is er over algemeenheden gesproken, er zullen ook specifieke aspecten zijn. Men kan hier slechts iets over zeggen door een model uit expertsystemen en een model uit information retrieval systemen uit te proberen.

4.2.1 Het calculatiedomein onder padexpressies

Het is de bedoeling dat als LISA-D faciliteiten krijgt, dat deze faciliteiten dan in het calculatiedomein onder padexpressies worden geïmplementeerd. Wat houdt eigenlijk dit calculatiedomein in?
De formulering van een padexpressie $\mathcal{PE}(\mathcal{I})$ is nauw gerelateerd met de informatiestructuur $\mathcal{I}$. De padexpressie correspondeert met een pad in de informatiestructuur. Het pad begint en eindigt in een objecttype. Bij evaluatie van zo'n padexpressie worden de tussentijdse instanties van objecten gebruikt, maar komen in het eindresultaat niet meer voor. Op deze manier zijn de resultaten steeds in de vorm van paren (binaire relaties). Om het verlies aan informatie door weglaten van tussentijdse instanties te compenseren, worden de genoemde paren in een multiset opgenomen. Bekijk hiertoe voorbeeld 1.2.1 op bladzijde 4 waarin de multiset voor de leesbaarheid als tabel is gegeven.

De multiset-benadering van padexpressies valt onder het calculatiedomein. Als we dit als een module van het systeem LISA-D zien, dan wordt daar een multiset van binaire (inhomogene) relaties als resultaat van een gegeven padexpressie bepaald. Operaties op multisets vallen ook in dit domein. Voor het verkrijgen van de multisets wordt gebruik gemaakt van de populatie (Pop) van de desbetreffende informatiestructuur ($\mathcal{I}$). In figuur 4.1 wordt de relatie van het calculatiedomein met andere onderdelen weergegeven. Het domein van multisets van binaire (inhomogene) relaties wordt gegeven door $\Omega_{\mathcal{PE}}$ en is als volgt gedefinieerd: $\Omega_{\mathcal{PE}} = \{X \mid X \text{ is een multiset over } \Omega \times \Omega\}$.



Figuur 4.1: Een hiërarchie waarin de (input-/output-)relaties tussen de componenten expliciet zijn weergegeven.

Hier is Ω het universum van instanties die in de populatie van een informatiestructuur kunnen voorkomen. De functie μ geeft van een evaluatie van een padexpressie de semantiek weer. Deze functie is formeel gedefinieerd als $\mu : \mathcal{PE} \times \mathsf{POP} \rightarrow \Omega_{\mathcal{PE}}$. POP is een verzameling van alle populaties en is gedefinieerd als $\mathsf{POP} = \mathcal{O} \rightarrow \wp(\Omega)$. Daarentegen is Pop een afbeelding $\mathsf{Pop} : \mathcal{O} \rightarrow \wp(\Omega)$ die van een gegeven objecttype de populatie (als (machts-)verzameling) levert.

Mogelijkheden voor uitbreiding

Het is de bedoeling dat de formele definities van padexpressies en de bijbehorende operaties daarop onveranderd blijven. Dit is om het abstractieniveau te handhaven. Padexpressies zijn slechts een schakel in het geheel waaruit LISA-D bestaat. Met padexpressies zijn andere zaken gerelateerd en een "ingreep" in padexpressie kan dan van invloed zijn op die zaken. Dan zijn multisets en de populatie van de informatiestructuur de enige plaatsen waar uitbreiding van LISA-D met faciliteiten voor het redeneren met onzekere kennis mogelijk is.

Onzekere kennis verkrijgt men door met een gegeven een maat voor onzekerheid te associëren. En om met onzekere kennis te werken komt er ook nog het een en ander bij kijken. Om LISA-D uit te breiden moet er op de volgende vragen antwoord gegeven worden:

- Hoe associeert men met een gegeven een maat voor de onzekerheid?
- Als meerdere stukken onzekere kennis tezamen nieuwe gegevens oplevert, hoe is dan de maat voor de onzekerheid die bij het nieuwe gegeven hoort te bepalen?

4.2.2 Toekenning van een maat voor de onzekerheid

Multisets worden gecreëerd met behulp van de populatie van de informatiestructuur. Als nu met elementen in de populatie maten voor onzekerheid geassocieerd worden, dan kunnen bij creatie van multisets deze maten meegenomen worden.
De populatie is een (machts-)verzameling. In de verzameling zijn de elementen (instanties) **uniek**. Dan kan dus met iedere instantie slechts eenduidig een maat voor onzekerheid geassocieerd worden. In de informatiestructuur zijn alleen objecttypen te populeren, wat ook blijkt uit de definities van POP en Pop. Dus is het niet mogelijk een maat voor onzekerheid te associëren met iets dat niet "bestaat".
Er bestaat de mogelijkheid dat in eerste instantie aan een element uit de populatie nog geen maat voor onzekerheid toegekend zal worden. In het geval het wel de bedoeling was dat in de populatie van een objecttype aan elk element een waarde wordt toegekend, moet er een soort *neutraal element* zijn, die dan toegekend kan worden. Het wordt hier een neutraal element genoemd, omdat dit element dan kan worden gebruikt in situaties waarin de invloed niet erg gewenst is. Wat deze

neutrale element is hangt af van het soort berekeningsoperaties dat toegepast wordt. Zo'n element valt dus niet direkt te geven en wordt dan voor het gemak $\xi_{\mathcal{U}}$ genoemd.
Er bestaat de mogelijkheid dat het element per objecttype verschillend moet zijn om zo ter plekke een berekening te beïnvloeden. Dit element is voor optelling anders dan voor vermenigvuldiging. Uit voorzorg wordt er dus rekening mee gehouden dat de berekeningsoperaties verschillend zijn.

Definitie 4.2.1

Zij X *een objecttype in een informatiestructuur dat gegeven is. Het lokale standaardelement is een functie* $\xi : \mathcal{O} \rightarrow \mathcal{U}$ *die als volgt is gedefinieerd:*

$$\xi(\mathrm{X}) \quad = \quad \begin{cases} v, & \text{toe te kennen door de gebruiker} \\ \xi_{\mathcal{U}}, & \text{anders} \end{cases}$$

waarin v *een getal is dat door de gebruiker zelf kan worden ingevuld.* □

In de bovenstaande definitie is $\mathcal{U}$ het domein waarin de maten voor onzekerheid worden gerepresenteerd. Dit domein kan Booleans, breuken, reële getallen of iets anders zijn. Dat hangt af van het model dat wordt gebruikt.
Nu kan volledig beschreven worden wat als maat voor onzekerheid aan een element van een populatie toegekend kan worden.

Definitie 4.2.2

Zij een informatiestructuur en diens populatie gegeven en zij X *een objecttype in de structuur. Zij tevens* $x \in \mathsf{Pop}(X)$. *Een* onzekerheidstoekenning *(Engels:* assignment of uncertainty*)* $\mathcal{A}u$ *is een functie* $\mathcal{A}u : (\Omega \cup \wp(\Omega)) \rightarrow \mathcal{U}$ *die als volgt is gedefinieerd:*

$$\mathcal{A}u(x) \quad = \quad \begin{cases} m, & \text{in te vullen door de gebruiker} \\ \xi(\mathrm{X}), & \text{anders} \end{cases}$$

□

Het zal duidelijk zijn dat de functie $\mathcal{A}u$ niet bijectief is. Een element m kan aan meerdere instanties toegekend worden. De relatie tussen een element van een populatie en de bijbehorende onzekerheidstoekenning moet dus expliciet gegeven zijn. Een mogelijkheid om dit te bewerkstelligen is de definities van Pop en POP aanpassen. Echter kan LISA-D faciliteiten hebben dat van deze afbeeldingen gebruik maakt. Bijvoorbeeld het controleren of de populatie voldoet aan de constraints is een van die faciliteiten. Een aanpassing van de definities zou dus kunnen leiden tot aanpassingen in de faciliteiten. Het is dus het beste om een nieuwe soort populatie te definiëren. Deze kan worden gebaseerd op de populaties Pop en POP.

Definitie 4.2.3

Zij X *een objecttype in een gegeven informatiestructuur. De* onzekerheidspopulatie UPop *waarin met de elementen onzekerheidstoekenningen worden gerelateerd is een afbeelding* $\mathsf{UPop} : \mathcal{O} \rightarrow (\Omega \cup \wp(\Omega)) \times \mathcal{U}$ *die als volgt is gedefinieerd:*

$$\mathsf{UPop}(X) \quad = \quad \{(x, \mathcal{A}u(x)) \mid x \in \mathsf{Pop}(\mathrm{X})\}$$

waarvoor de volgende eigenschappen gelden:

(1) $\{(x, \mathcal{A}u(x)), (y, \mathcal{A}u(y))\} \subseteq \mathsf{UPop}(\mathrm{X}) \wedge (x = y) \Rightarrow \mathcal{A}u(x) = \mathcal{A}u(y)$
(2) $\mathrm{Y} \in \mathcal{O} \wedge \mathrm{Y}\ \mathsf{Gen}\ \mathrm{X} \Rightarrow \mathsf{UPop}(\mathrm{X}) \subseteq \mathsf{UPop}(\mathrm{Y})$
(3) $\mathrm{Y} \in \mathcal{O} \wedge \mathrm{Y}\ \mathsf{Spec}\ \mathrm{X} \Rightarrow \mathsf{UPop}(\mathrm{Y}) = \{(y, \mathcal{A}u(y)) \mid y \in \mathsf{Pop}(\mathrm{Y}) \wedge (y, \mathcal{A}u(y)) \in \mathsf{UPop}(\mathrm{X})\}$

□

De verzameling van alle mogelijke onzekerheidspopulaties kan nu ook gedefinieerd worden: $\mathsf{UPOP} = \mathcal{O} \rightarrow (\Omega \cup \wp(\Omega)) \times \mathcal{U}$. Men kan zich afvragen of het gewenst is dat bij **elk** objecttype met de elementen van de populatie onzekerheidstoekenningen moeten worden gedaan. Gelukkig lossen de definities van $\mathcal{A}u$ en ξ dat probleem op. Door gewoon niets toe te kennen zal er als alternatief het lokale standaardelement $\xi_{\mathcal{U}}$ worden ingevuld.

Voorbeeld 4.2.1

Zij een objecttype B *gegeven en zij de populatie van* B *als volgt:*

$$\mathsf{Pop}(\mathrm{B}) \quad = \quad \{b_1, b_2, b_3, b_4\}$$

Als de gebruiker aan b_1 *kans* 0.7 *en aan* b_3 *kans* 0.95 *heeft toegekend, dan is*

$$\mathsf{UPop}(\mathrm{B}) \quad = \quad \{(b_1, 0.7), (b_2, \xi(\mathrm{B})), (b_3, 0.95), (b_4, \xi(\mathrm{B}))\}$$

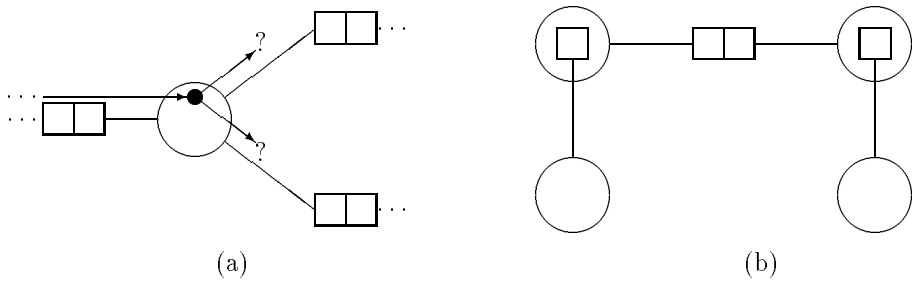
□

4.2.2.1 Probleemgevallen

Er zijn bepaalde gevallen waarbij het gebruik van een onzekerheidstoekenning problemen kan geven. De problemen kunnen zich voordoen met predicatoren en gegeneraliseerde objecttypen.

Predicatoren

Wat hier zal worden besproken is eigenlijk geen probleem, maar meer een ongemak. Het zal duidelijk zijn dat volgens de definitie van PSM de predicatoren geen objecttypen zijn. Toch zou men willen dat daar ook een onzekerheidstoekenning bij gespecificeerd kan worden. In een padexpressie geeft men namelijk aan welke pad gevolgd wordt. Als men bij een objecttype is, en de populatie speelt in meerdere feittypen een rol, dan kan slechts via de predicatoren aangegeven worden welke informatie dan doorgegeven moet worden. Zie voor de duidelijkheid figuur 4.2(a). Bij het be-

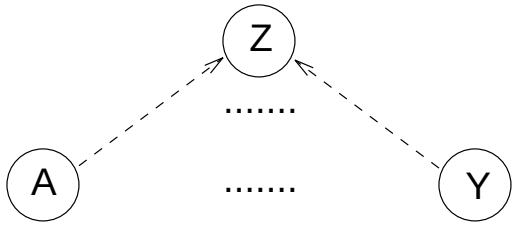


Figuur 4.2: Plaatje (a) geeft het probleem van versturen van gegevens weer. Plaatje (b) representeert een voorbeeld informatiestructuur door toepassen van de genoemde oplossing (zie tekst).

treffende objecttype kan wel een onzekerheidstoekenning worden gegeven, maar daarin moet dan een functie zijn opgenomen die dan aan de padexpressie moet aflezen welke kant men opgaat, om daarmee te bepalen welke gegevens meegegeven worden. Dat is vrij moeilijk, zo niet onmogelijk. Verder is er het probleem dat ook nog voor elke richting gegevens moet worden opgeslagen.
De oplossing is vrij simpel. Wat men moet doen is de predicator objectificeren. Het is dan in principe een *unaire feittype* geworden, en daar kan dan wel weer een onzekerheidstoekenning bij gegeven worden. De informatiestructuur kan hierdoor wel eens niet mooi worden zoals uit figuur 4.2(b) wel blijkt.

Gegeneraliseerde objecttypen

Deze objecttypen kunnen problemen veroorzaken met betrekking tot de onzekerheidstoekenning. Het probleem is hier het overerven van de gegevens. Zie figuur 4.3. Als met een element uit de populatie van A een onzekerheidstoekenning wordt geassocieerd, dan vindt men deze onzekerheidstoekenning ook weer bij hetzelfde element in de populatie van Z. Omgekeerd gaat het verhaal **niet** op.
Volgens de definitie van generalisatie worden de populaties van A, Y en eventueel andere objecttypen in de populatie van Z opgenomen. Dit geldt ook voor UPop. Bij de definitie van de onzekerheidspopulatie is rekening gehouden met generalisatie. Dus de onzekerheidstoekenningen



Figuur 4.3: Het objecttype Z als generalisatie van A, Y en eventueel meer objecten.

worden overgeërfd.
Als men niet wil dat de onzekerheidspopulatie van enkele objecttypen wordt overgeërfd, dan is daar een oplossing voor: zorgen dat van die objecttypen geen onzekerheidspopulaties bestaan waarin men aan een element een maat voor de onzekerheid heeft toegekend. Verder moet men zonodig het lokale standaardelement aanpassen om er voor te zorgen dat met betrekking tot het gegeneraliseerde objecttype daar de onzekerheidstoekenningen dan nog juist zijn.

4.2.3 Werken met onzekerheid

Nu er de mogelijkheid is gecreëerd is om met elementen van de populaties onzekerheidstoekenningen te relateren, moet er nu de mogelijkheid worden gecreëerd om met die onzekerheidstoekenningen te gaan rekenen. Hierbij is het de bedoeling dat een opgegeven padexpressie als een soort reken-voorschrift zal dienen.

4.2.3.1 De basis

Er moet echter eerst een middel zijn om tussenresultaten en het eindresultaat te bewaren. Bij evaluatie van een padexpressie zijn de tussenresultaten in standaardvorm, namelijk multisets van binaire relaties waarmee dan steeds mee verder gerekend kan worden. Hetzelfde moet ook voor tussentijds berekende maten voor onzekerheid gelden. Om een representatievorm voor de onzekerheid te maken moet er met enkele zaken rekening gehouden worden:

- Bij elke tupel dat in een multiset als resultaat van een geëvalueerde padexpressie voorkomt moet een bijbehorende berekende maat voor de onzekerheid gerelateerd worden;
- De berekende maat voor de onzekerheid kan sterk samenhangen met de frequentie van voorkomen van een tupel in een multiset. Dit is met kanstheorie het geval. In voorbeeld 4.2.1 is aan b_1 kans 0.7 toegekend, wat betekent dat $P(b_1 \text{ komt 1 keer voor}) = 0.7$ en $P(b_1 \text{ komt 0 keer voor}) = 0.3$. Beide waarden kunnen bij de bepaling van een nieuwe maat voor de onzekerheid van belang zijn.

Een idee is om een *rijtje* (*sequence*) te maken waarbij de frequentie van voorkomen van een tupel als index op het rijtje gebruikt kan worden. Een constructie van rijtjes gaat op dezelfde manier als met multisets. In [HPW93] is een multiset geïntroduceerd als een functie $M : X \rightarrow \mathbb{N}$. Als $C(a,n)$ een expressie is zó dat de uitspraak $\forall_{a \in X} \exists!_{n \in \mathbb{N}} [C(a,n)]$ geldt, dan is de multiset als functie zodanig gedefinieerd dat een multiset is te schrijven als $M \equiv \{\!\{a\uparrow^n \mid C(a,n)\}\!\}$.
Een rijtje wordt geïntroduceerd als een functie $S : \mathbb{N} \rightarrow Y$ waarin op elke positie $n \in \mathbb{N}$ een waarde $u \in Y$ voorkomt. Als $C(n,s)$ een expressie is zó dat de uitspraak $\forall_{n \in \mathbb{N}} \exists!_{s \in Y} [C(n,s)]$ geldt, dan is het rijtje als functie zodanig gedefinieerd dat het rijtje is te schrijven als $S \equiv \{\!\{s_n \mid C(n,s)\}\!\}$. Een meer geschikte manier om een rijtje op te schrijven is $S =< s_0, s_1, s_2, \ldots, s_n >$, $n \in \mathbb{N}$.

Het is de bedoeling dat met elk tupel een rijtje wordt gerelateerd. Welke waarde in het rijtje bij de tupel behoort, wordt bepaald door de frequentie van voorkomen van de tupel. Om het principe waarvoor rijtjes worden gebruikt duidelijk uit te drukken, noemen we een rijtje in het vervolg een *frequentieverdeling* (Engels: *frequency distribution*). De frequentieverdeling kan heel groot worden, in principe zelfs oneindig zijn. Immers, er is geen maximum voor de frequentie van voorkomen van een tupel.

Een evaluatie van een padexpressie leverde steeds een multiset van binaire relaties op. $\Omega_{\mathcal{PE}}$ is het domein van de mogelijke multisets. Nu er ook frequentieverdelingen in multisets worden opgenomen, moeten de operaties die op die multisets betrekking hebben worden aangepast. Hieronder volgen de aangepaste operaties die in [HPW93] vermeld staan.

- Coërcie van multisets naar sets en omgekeerd:

$$\begin{aligned} \mathsf{Set}(M) &\equiv \{\, y \mid y \in M \,\} \\ \mathsf{Multi}(S) &\equiv \{\!|\, y\uparrow^1 \mid y \in S \,|\!\} \end{aligned}$$

- Coërcie van multisets naar multisets van binaire tupels en omgekeerd:

$$\begin{aligned} \mathsf{Sqr}(M) &\equiv \{\!|\, (\langle x, x\rangle, d)\uparrow^n \mid (x,d) \in^n M \,|\!\} \\ \pi_1(M) &\equiv \{\!|\, (x,d)\uparrow^n \mid (\langle x, y\rangle, d) \in^n M \,|\!\} \\ \pi_2(M) &\equiv \{\!|\, (y,d)\uparrow^n \mid (\langle x, y\rangle, d) \in^n M \,|\!\} \end{aligned}$$

Aangezien nu de multisets steeds bestaan uit paren van elementen (tupels, singles) en de bijbehorende frequentieverdelingen, is het handig ook een operatie te hebben die alleen elementen oplevert. Zo'n operatie noemen we Elem:

$$\mathsf{Elem}(M) \equiv \{\!|\, x\uparrow^n \mid (x,d) \in^n M \,|\!\}$$

Een verzameling van frequentieverdelingen is niet zinvol. De frequentieverdelingen kunnen misschien wel ergens van nut zijn, maar daar kan momenteel niets over gezegd worden.

4.2.3.2 Padexpressies

Het zal duidelijk zijn dat een evaluatie van een padexpressie multisets van paren van elementen met frequentieverdelingen oplevert. Daartoe moet $\Omega_{\mathcal{PE}}$ worden aangepast. Als nu de verzameling van mogelijke frequentieverdelingen wordt gedefinieerd als $\mathsf{SEQU} = \mathbb{N} \rightarrow \mathcal{U}$, dan wordt de aangepaste versie van $\Omega_{\mathcal{PE}}$ als volgt:

$$\Psi_{\mathcal{PE}} = \{\, X \mid X \text{ is een multiset over } (\Omega \times \Omega) \times \mathsf{SEQU} \,\}$$

Ook de functie μ die de semantiek van padexpressies weergeeft moet aangepast worden. Om aan te geven dat er met andere zaken gewerkt wordt, hernoemen we die functie als ν:

$$\nu : \mathcal{PE} \times \mathsf{UPOP} \rightarrow \Psi_{\mathcal{PE}}$$

De functie ν verschilt slechts van μ in frequentieverdelingen. Als in de uitkomst van ν de verdelingen weggelaten worden, moet er dezelfde gegevens staan als in de uitkomst van μ, ofwel:

$$\mu[\![P]\!](\mathsf{Pop}) = \mathsf{Elem}(\nu[\![P]\!](\mathsf{UPop})),$$

waarin P een gegeven padexpressie is.

Het is nu de vraag hoe bij elk soort padexpressie de frequentieverdelingen er inhoudelijk uitzien. Daar kan geen eenduidig antwoord op gegeven worden omdat de specificatie van de frequentieverdelingen afhangt van de zogenaamde "kansmodellen" dat bij expertsystemen en information retrieval systemen wordt gebruikt. De specificatie is per model verschillend. Om te laten zien wat met de frequentieverdelingen precies de bedoeling is wordt er een voorbeeld gegeven. Er wordt eerst een overzicht gegeven van de meest voorkomende padexpressies en de semantieken daarvan. In het overzicht komen de functies (of operaties) 1_{SEQU} en seq voor die een eenheidsverdeling of neutrale verdeling levert respectievelijk een frequentieverdeling creëert. De rest spreekt voor zich. De Engelse benaming van padexpressies wordt gehandhaafd.

name	expr	$\mathcal{V}[\![\text{expr}]\!](\mathsf{UPop})$
empty path	$\varnothing_{\mathcal{PE}}$	$\varnothing$
neutral path	$1_{\mathcal{PE}}$	$1_{\Omega \times \Omega \times \mathsf{SEQU}}$
constant	c	$\mathsf{Sqr}(\{[(c, 1_{\mathsf{SEQU}})]\})$
multiset	X	$\mathsf{Sqr}(X)$
objecttype	x	$\{\!\![(\langle y, y\rangle, \mathsf{seq}(d))\uparrow^1 \mid (y,d) \in \mathsf{UPop}(x)]\!\!\}$
predicator	p	$\{\!\![(\langle v(p), v\rangle, 1_{\mathsf{SEQU}})\uparrow^1 \mid v \in \mathsf{Pop} \circ \mathsf{Fact}(p)]\!\!\}$
reverse	$P^{\leftarrow}$	$\{\!\![(\langle q, p\rangle, s)\uparrow^n \mid (\langle p, q\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})]\!\!\}$
concatenate	$P \circ Q$	$\bigcup_{r, s \odot t} \left\{\!\!\left[(\langle p, q\rangle, s \odot t)\uparrow^{n \times m} \;\middle|\; \begin{array}{l}(\langle p, r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge \\ (\langle r, q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})\end{array}\right]\!\!\right\}$
front	$\not\leftarrow P$	$\mathsf{Sqr}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

Voorbeeld 4.2.2

In het vorige voorbeeld (voorbeeld 4.2.1) is er een onzekerheidspopulatie gegeven waarin aan elk element een kans is toegekend. Die kans moet nog worden vertaald in een frequentieverdeling. Om enige relatie met kanstheorie te houden, wordt een verdeling $S =< s_0, s_1, s_2, \ldots >$ geconstrueerd, waarin voor een $(b_i, \mathcal{A}u(b_i)) \in \mathsf{UPop}(\mathrm{B})$

$$\begin{aligned} s_0 &= P(b_i\text{komt niet voor}) = 1 - \mathcal{A}u(b_i), \\ s_1 &= P(b_i\text{komt 1 keer voor}) = \mathcal{A}u(b_i), \\ s_k &= P(b_i\text{komt k keer voor}) = 0. \end{aligned}$$

In het vorige voorbeeld is $\mathsf{UPop}(\mathrm{B}) = \{(b_1, 0.7), (b_2, \xi(\mathrm{B})), (b_3, 0.95), (b_4, \xi(\mathrm{B}))\}$*. Dan wordt:*

$$\mathcal{V}[\![\mathrm{B}]\!](\mathsf{UPop}) \quad = \quad \begin{array}{|c|c|l|} \hline b_1 & b_1 & < 0.3, 0.7, 0, \ldots > \\ b_2 & b_2 & < 1 - \xi(\mathrm{B}), \xi(\mathrm{B}), 0, \ldots > \\ b_3 & b_3 & < 0.05, 0.95, 0, \ldots > \\ b_4 & b_4 & < 1 - \xi(\mathrm{B}), \xi(\mathrm{B}), 0, \ldots > \\ \hline \end{array}$$

□

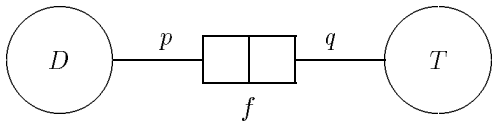
Er zijn naast de padexpressies in het overzicht ook andere padexpressies. Bij enkele van deze padexpressies is een met een gegeven tupel te associëren frequentieverdeling niet eenvoudigweg te geven (bijvoorbeeld bij concatenatie $\circ$, vereniging $\cup$). Er zijn ook padexpressies bij waarbij het op het eerste gezicht geen zin heeft om een met een gegeven tupel te associëren verdeling te berekenen (Sum, Cnt). Deze en andere padexpressies zijn gegeven in [HPW93]. Het is al eerder gezegd dat de omschrijving van de semantiek van de padexpressies afhangt van het model waarin deze expressies (kunnen) worden gebruikt.

Opmerking 4.2.1

Om te voorkomen dat er padexpressies overblijven die niet aan de orde komen, wordt hier aangenomen dat deze padexpressies multisets opleveren waarin met elk tupel de neutrale verdeling is geasocieerd. □

4.2.3.3 Extra operaties op frequentieverdelingen

Net zoals er verschillende soorten padexpressies zijn om allerlei bruikbare multisets te krijgen met de inhoud in gewenste vorm, is het de bedoeling dat zoiets ook met frequentieverdelingen gebeurt. De invulling van de verdeling moet kunnen worden gemodificeerd. Hier gaat het om operaties waarvan de werking niet recht toe recht aan via padexpressies is te beschrijven. Hieronder volgt een voorbeeld.



Figuur 4.4: De informatiestructuur

Voorbeeld 4.2.3

In de informatiestructuur in figuur 4.4 is de relatie tussen Documenten en Termen (trefwoorden) gegeven. Een onzekerheidspopulatie geeft aan welke documenten met welke termen zijn gerelateerd en het bijbehorende getal geeft aan hoe vaak een term in het document van toepassing is. Zij zo'n populatie als volgt:

$$\begin{aligned}
\mathsf{UPop}(D) &= \{(d_1,1),(d_2,1)\} \\
\mathsf{UPop}(T) &= \{(t_1,1),(t_2,1)\} \\
\mathsf{UPop}(f) &= \{(\{\mathrm{p}{:}d_1,\mathrm{q}{:}t_1\},2),(\{\mathrm{p}{:}d_1,\mathrm{q}{:}t_2\},1),(\{\mathrm{p}{:}d_2,\mathrm{q}{:}t_1\},1),(\{\mathrm{p}{:}d_2,\mathrm{q}{:}t_2\},1)\}
\end{aligned}$$

Voor een goede overzicht laten we met behulp van een padexpressie de relatie zien:

$$\mathcal{V}[\![D \circ p \circ f \circ q^{\leftarrow} \circ T]\!](\mathsf{UPop}) \;=\; \begin{array}{|c|c|l|} \hline d_1 & t_1 & <0,2,0,\ldots> \\ \hline d_1 & t_2 & <0,1,0,\ldots> \\ \hline d_2 & t_1 & <0,1,0,\ldots> \\ \hline d_2 & t_2 & <0,1,0,\ldots> \\ \hline \end{array}$$

Maar stel nu dat het gewenst is dat de frequentie van voorkomen van een term in een document wordt genormaliseerd en dan voor alle documenten en termen, hoe bereikt men dit dan? Hoe krijgt men dus voor elkaar dat voor $\langle d_1,t_1\rangle \in \mathsf{Elem}(\mathcal{V}[\![D \circ p \circ f \circ q^{\leftarrow} \circ T]\!](\mathsf{UPop}))$ het bijbehorende getal in de frequentieverdeling $\frac{2}{3}$ is in plaats van 2? ($3 = \sum_{t \in \mathsf{Pop}(T)} x\,, (\{\mathrm{p}{:}d_1,\mathrm{q}{:}t\},x) \in \mathsf{UPop}(f)$) □

Er zijn manieren om voor elkaar te krijgen dat de frequentieverdelingen worden gemodificeerd. Het is echter niet kort en krachtig te omschrijven wat die manieren zijn. In de volgende paragrafen wordt uitgebreid op de mogelijkheden ingegaan.

4.2.3.3.1 Het "verdubbelen" van objecttypen

De titel van deze paragraaf geeft eigenlijk al aan hoe het invoegen van een nieuwe multiset geschiedt. Er wordt gebruik gemaakt van de eigenschap van objecttypen, namelijk dat $\mu[\![x]\!](\mathsf{Pop}) = \mu[\![x \circ x]\!](\mathsf{Pop})$ voor een objecttype x. Dit verhaal gaat echter niet op in het geval dat frequentieverdelingen in het spel komen. Stel $\mathsf{UPop}(x) = \{(x_1, 0.5)\}$ en verdelingen worden geconcateneerd door indexsgewijs de elementen met elkaar te vermenigvuldigen, dan is

$$\mathcal{V}[\![x]\!](\mathsf{UPop}) \;=\; \begin{array}{|c|c|l|} \hline x_1 & x_1 & <0,0.5,0,\ldots> \\ \hline \end{array}\;,$$

en

$$\mathcal{V}[\![x \circ x]\!](\mathsf{UPop}) \;=\; \begin{array}{|c|c|l|} \hline x_1 & x_1 & <0,0.25,0,\ldots> \\ \hline \end{array}\;.$$

Zij zijn dus niet gelijk.
Men ziet aan het voorbeeldje wel hoeveel invloed zo'n verdubbeling heeft op de inhoud van de uiteindelijke frequentieverdeling. Het zogenaamde "verdubbelen" van slechts objecttypen heeft een reden. Als P een padexpressie is waarin zo'n objecttype is verdubbeld en Q een padexpressie waarin dat niet voorkomt, dan geldt:

$$\mu[\![Q]\!](\mathsf{Pop}) \;=\; \mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop})\;.$$

Dit wil zoveel zeggen dat er niet aan het basisprincipe van padexpressie wordt gerommeld wat de creatie van tupels van de vorm $\langle a,b\rangle$ betreft. LISA-D moet ook gewoon zonder uitbreiding voor

kansmodellen werken. Alleen met de frequentieverdelingen kan iets gedaan worden.
Het idee is tot zover het gebruiken van de populatie van een objecttype om tupels te creëren en daar zelf frequentieverdelingen mee te associëren. Wat voor verdelingen men wil hebben moet men kunnen omschrijven. Aangezien objecttypen uniek zijn, kan aan een objecttype een functie aangehangen worden, waarin dan wordt gespecificeerd hoe de verdelingen er uit zouden moeten komen te zien, en het resultaat verpakken in een multiset.

Definitie 4.2.4

Zij x een objecttype en tevens een padexpressie. Dan is $\mathsf{cfd}(x)$ *weer een padexpressie waarvan de semantiek als volgt is:*

$$\mathcal{V}[\![\mathsf{cfd}(x)]\!](\mathsf{UPop}) = \begin{cases} (\{[\![(\langle y,y\rangle, d)\uparrow^1 \mid y \in \mathsf{Pop(x)} \wedge d = \mathrm{fd}(x)]\!\} \circ \mathcal{V}[\![x]\!](\mathsf{UPop})) & , \text{als } x \in \mathcal{O} \\ \mathcal{V}[\![x]\!](\mathsf{UPop})) & , \text{anders} \end{cases}$$

waarin fd *een functie* $\mathrm{fd} : \mathcal{O} \rightarrow \mathsf{SEQU}$ *is, die voor elk instantie van het gegeven objecttype x de bijbehorende frequentieverdeling levert. Deze padexpressie noemt men* verandering (change) *van frequentiedistributie.* □

De haakjes staan in de omschrijving niet voor niets er omheen. Het maakt semantisch gezien veel uit als men kijkt naar bijvoorbeeld de padexpressies $(P \circ Q) \cup R$ en $P \circ (Q \cup R)$.

Voorbeeld 4.2.4

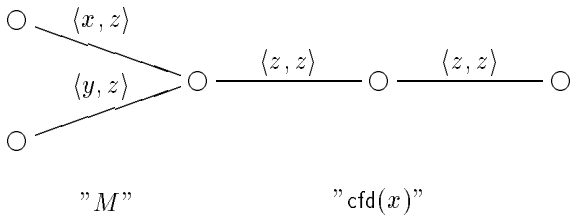
In dit vervolg op het vorige voorbeeld (voorbeeld 4.2.3) lossen we hier nu het probleem op. Zij $\langle y,y\rangle$ de tupel waaraan een verdeling toegekend moet worden ($y \in \mathsf{Pop}(D)$).

$$\mathrm{fd}(D) = \begin{cases} \mathsf{seq}\left(\frac{1}{\sum_{v(p)=y} \mathcal{A}u(v)}\right) & , \text{als } \mathsf{Fact}(p) = f \wedge \mathsf{Base}(p) = D \wedge (v, \mathcal{A}u(v)) \in \mathsf{UPop}(f) \\ 1_{\mathsf{SEQU}} & , \text{anders} \end{cases}$$

Als nu in de padexpressie van het vorige voorbeeld D vervangen wordt door $\mathsf{cfd}(D)$*, dan levert dat het volgende resultaat op:*

$$\mathcal{V}[\![\mathsf{cfd}(D) \circ p \circ f \circ q^{\leftarrow} \circ T]\!](\mathsf{UPop}) = \begin{array}{|c|c|l|} \hline d_1 & t_1 & <0, \frac{2}{3}, 0, \ldots> \\ \hline d_1 & t_2 & <0, \frac{1}{3}, 0, \ldots> \\ \hline d_2 & t_1 & <0, \frac{1}{2}, 0, \ldots> \\ \hline d_2 & t_2 & <0, \frac{1}{2}, 0, \ldots> \\ \hline \end{array}$$

Hier wordt voor het gemak verondersteld dat bij concatenatie de verdelingen op een simpele wijze worden gecombineerd. □



Figuur 4.5: Schematische weergave van "aaneen te rijgen" tupels als resultaat van padexpressie $M \circ \mathsf{cfd}(x)$ (zonder verdelingen).

4.2.3.3.2 Wijzigen van verdelingen in multiset

In de vorige paragraaf is beschreven hoe men de invulling van de definitieve frequentieverdelingen kan sturen door op strategische plaatsen enkele nieuwe verdelingen mee te laten rekenen. In figuur 4.5 is het principe grafisch weergegeven. Toch kent dat principe een beperking: hoe is in de multiset M zoals in figuur 4.5 daar de frequentieverdeling bij de tupels afzonderlijk te wijzigen?

Voorbeeld 4.2.5

Stel $M = \{[(\langle x,z\rangle, <0,1,0,\ldots>)\uparrow^1, (\langle y,z\rangle, <0,0.6,0,\ldots>)\uparrow^1]\}$. *Daarin is* z *een benaming van de aanwijzing* a **or** b. *In het zekerheidsfactormodel moet voor die aanwijzing de grootste zekerheidsfactor bepaald worden. Hoe krijgt men voor elkaar dat de multiset* M *wordt veranderd in* $M' = \{[(\langle x,z\rangle, <0,1,0,\ldots>)\uparrow^1, (\langle y,z\rangle, <0,1,0,\ldots>)\uparrow^1]\}$ *?* □

Dit probleem kan opgelost worden. Het idee is als volgt. Zoals figuur 4.5 aangeeft is er een objecttype x die met multiset M is te combineren. De volgende stappen moeten worden ondernomen:

1. Gebruik de functie cfd die bij x hoort en laat deze het volgende doen:
   (a) als $z \in \mathsf{Pop}(x)$, dan zoek je het bijpassende element of vergaar je de elementen uit M. Dat laatste is bijvoorbeeld in wiskundige termen:
   $$S = \{[(\langle a,y\rangle, d)\uparrow^n | (\langle a,y\rangle, d) \in^n M \wedge y = z]\} ;$$
   (b) laat een operatie los op die verzameling zó dat het 1 frequentieverdeling oplevert die dan kan worden geassocieerd met instantie z.

   Dit doe je dan voor elk instantie in de populatie van x;
2. Zorg er dan voor dat uit M en x een multiset M' ontstaat zó dat M' in principe M is, maar dan met frequentieverdelingen geproduceerd door fd(x).

Het overvoeren van M in multiset M' is vrij gemakkelijk. We definiëren daartoe een variant op de concatenatie.

Definitie 4.2.5

Zij P *en* Q *padexpressies, Dan zijn* $P \vdash Q$ *en* $P \dashv Q$ *weer padexpressies, waarvan de semantieken als volgt zijn gedefinieerd:*

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
leads to	$(P \vdash Q)$	$\bigcup_{r,t} \left\{ \left[(\langle p,q\rangle, t)\uparrow^{n\times m} \middle\| \begin{array}{c} (\langle p,r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \ \wedge \\ (\langle r,q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop}) \end{array} \right] \right\}$
follows from	$(P \dashv Q)$	$\bigcup_{r,s} \left\{ \left[(\langle p,q\rangle, s)\uparrow^{n\times m} \middle\| \begin{array}{c} (\langle p,r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \ \wedge \\ (\langle r,q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop}) \end{array} \right] \right\}$

Wat er met de frequentieverdelingen gebeurt als een tupel $\langle p,q\rangle$ *op meerdere manieren gevormd kan worden hangt af van het model waarin deze padexpressies worden gebruikt.*
□

Dus een padexpressie $(M \vdash \mathsf{cfd}(x))$ levert een multiset M' met tupels uit M maar met verdelingen geproduceerd door fd(x).
Hier zij nog opgemerkt dat haakjes om de expressie staan. Dat is noodzakelijk omdat met betrekking tot frequentieverdeling het zo is dat $(P \vdash Q) \circ R \not\equiv P \vdash (Q \circ R)$. Het hier beschreven principe van veranderen van frequentieverdelingen in de multiset M gebeurt hier momenteel slechts via het principe van concatenatie. Andere combinaties van padexpressies zoals vereniging, doorsnede lijken in eerste instantie niet zinvol. Als blijkt dat ze toch nodig zijn, dan kunnen ze later alsnog gedefinieerd worden.

De functie fd geeft zelf niet aan welke multiset er gebruikt moet worden om daarmee de nieuwe verdelingen te bepalen. Dit moet men zelf in de body van die functie aangeven. Men kan op twee manieren aangeven wat voor multiset men gebruik wil maken:

- door een padexpressie op te geven, en dan het resultaat van evaluatie ervan gebruiken; of
- men maakt van een *standaard multiset* als database gebruik. De inhoud daarvan wordt gegeven door aan te geven van welk deel van de padexpressie waarin op dit moment de expressie $\mathsf{cfd}(x)$ wordt geëvalueerd, het resultaat in de multiset moet. Bijvoorbeeld: van de padexpressie $P \circ Q \circ \mathsf{cfd}(x) \circ R$ moet het resultaat van evaluatie van expressie Q in de standaard multiset voorkomen.

De eerste mogelijkheid is vrij triviaal. Echter worden er dan padexpressies in het calculatiedomein betrokken. De tweede mogelijkheid vereist een definitie van een standaard multiset, en een definitie van een middel om in een padexpressie aan te geven van welk gedeelte het resultaat in de standaard multiset hoort. Er wordt nu één definitie gegeven waarin beide aspecten tegelijk worden behandeld.

Definitie 4.2.6

Zij P een padexpressie. Dan is $\mathsf{sv}\,P$ *weer een padexpressie waarvan de semantiek als volgt is gedefinieerd:*

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!]\,(\mathsf{UPop})$
save	$\mathsf{sv}\,P$	$\{[\![x\uparrow^n \mid x \in^n \mathcal{V}[\![P]\!]\,(\mathsf{UPop}) \wedge x \in^m \mathsf{STORAGE}]\!]\}$

waarin STORAGE *de naam is voor de* standaard multiset *en waarin voor m geldt: $m \in \mathbb{N} \setminus \{0\}$.* □

In deze definitie ligt m niet geheel vast. Er is ook niet aan gegeven **hoe** de multiset STORAGE wordt gevuld, maar wel **dat** moet gelden:

$$\forall_{x \in \nu[\![P]\!]\,(\mathsf{UPop})}\,[x \in \mathsf{STORAGE}]\ .$$

Informeel gezegd is $\mathsf{STORAGE} = f(\mathcal{V}[\![P]\!]\,(\mathsf{UPop}))$, waarin f een functie is dat de manier van vullen van STORAGE nader specificeert. Een eenduidige specificatie valt niet te geven omdat (nog) niet duidelijk is waar de standaard multiset voor wordt gebruikt.
Nu kan in de body van de functie fd naast de padexpressie ook van STORAGE gebruik gemaakt worden. Nu kan ook gedemonstreerd worden hoe het gebruik van de padexpressies met sv en cfd gedacht wordt.

Voorbeeld 4.2.6

In dit voorbeeld wordt het probleem dat in voorbeeld 4.2.5 is aangekaart, opgelost. Zij hier weer $M = \{[(\langle x, z\rangle, <0,1,0,\ldots>)\uparrow^1, (\langle y, z\rangle, <0,0.6,0,\ldots>)\uparrow^1]\}$. Zij A een objecttype met $\mathsf{Pop}(A) = \{z\}$, zodat bereikt wordt dat

$$M \circ \mathcal{V}[\![A]\!]\,(\mathsf{UPop}) \;=\; \left\{\!\!\left[\begin{array}{l}(\langle x, z\rangle, <0,1,0,\ldots>)\uparrow^1\\(\langle y, z\rangle, <0,0.6,0,\ldots>)\uparrow^1\end{array}\right]\!\!\right\}$$

Er verandert in principe dus niks. Het is echter de bedoeling dat de verdelingen voor alle tupels $\langle a, b\rangle$ in M gelijk zijn. Beschouw hier nu ook M als zijnde een padexpressie. Als we nu de volgende padexpressie X opschrijven:

$$X \;=\; (\mathsf{sv}\,M \vdash \mathsf{cfd}(A))\ ,$$

en we specificeren voor objecttype A de functie fd *als volgt:*

$$\mathrm{fd}(A) \;=\; \begin{cases} 1_{\mathsf{SEQU}} & , \text{als } y \notin \mathsf{STORAGE} \\ \mathsf{seq}(\max\{s[m] \mid (\langle a, y\rangle, s) \in^m \mathsf{STORAGE}\}) & , \text{anders} \end{cases}\ ,$$

en we evalueren expressie X, waarbij hier wordt aangenomen dat eerst $\mathsf{sv}\,M$ *wordt geëvalueerd (omdat anders de functie* fd *niet goed werkt), dan krijgt men het volgende resultaat (ga na):*

$$\mathcal{V}[\![X]\!]\,(\mathsf{UPop}) \;=\; \left\{\!\!\left[\begin{array}{l}(\langle x, z\rangle, <0,1,0,\ldots>)\uparrow^1\\(\langle y, z\rangle, <0,1,0,\ldots>)\uparrow^1\end{array}\right]\!\!\right\}$$

□

In het bovenstaande voorbeeld wordt de samenhang tussen de nieuwe padexpressies enigzins duidelijk. Er staat echter ook nog een vraag open. In het voorbeeld kwam al naar voren wat er eerst geëvalueerd moest worden, sv of cfd. Welk van deze twee heeft hoogste prioriteit? Het maakt heel wat uit of men met een padexpressie $\mathsf{sv}\, P \circ \mathsf{cfd}(x)$ van doen heeft of met $\mathsf{sv}(P \circ \mathsf{cfd}(x))$. In het tweede geval moet eerst $P \circ \mathsf{cfd}(x)$ geëvalueerd worden voordat de uitkomst in de standaard multiset verwerkt kan worden. Dit is verder vrij triviaal. In het eerste geval is het de vraag of eerst $\mathsf{sv}\, P$ gedaan moet worden omdat voor $\mathrm{cfd}(x)$ die gegevens mogelijk nodig zijn, of dat men gewoon eerst $\mathsf{cfd}(x)$ kan uitrekenen. Wat de oplossing zal zijn hangt af van de vraag of er slechts met verdelingen wordt gewerkt binnen het bereik van de expressie P, of dat er "van buitenaf" verdelingen in het proces wordt betrokken. Het antwoord daarop is vrij simpel: $\mathsf{sv}\, P$ geeft aan dat het resultaat van expressie P in de standaard multiset moet, en niet P met iets dat er aan hangt, zoals $P \circ x$. Dan zou dit ook moeten gelden voor frequentieverdelingen.
Terug naar de vraag welke het eerst moet worden uitgevoerd, sv of cfd. Het antwoord wordt hieronder in een uitspraak gegeven.

Propositie 4.2.1

1. In een padexpressie waarin sv en cfd naast elkaar voorkomen, heeft sv bij evaluatie **hogere** prioriteit dan cfd. cfd is volgens het volgende punt niet van invloed op het resultaat van evaluatie van een sv. Hogere prioriteit wil dus zeggen: deze padexpressie wordt het eerst uitgevoerd.
2. Als X op $\mathsf{sv}\, P$ op Y een padexpressie is, dan wordt $\mathsf{sv}\, P$ geëvalueerd en wel alsof X en Y niet bestaan (ofwel doe alsof men slechts met expressie $\mathsf{sv}\, P$ te maken heeft).
3. Wil men bij evaluatie van een padexpressie $\mathsf{sv}\, P$ verdelingen "van buitenaf" gebruiken, dan dient men daarvoor de expressie cfd te gebruiken. Voorbeeld: $\mathsf{sv}(P \circ \mathsf{cfd}(x))$.

Een andere relatie tussen sv en cfd is die van *onderlinge afhankelijkheid*. Een goede werking van de functie fd staat of valt met de correcte inhoud van STORAGE. Daar kan echter geen enkele uitspraak over worden gedaan. Het enige dat men weet is dat als de functie van een objecttype x $\mathrm{fd}(x)$ van STORAGE gebruik maakt, dat $\mathrm{fd}(x)$ en STORAGE één ding gemeenschappelijk hebben:

$$\exists_{(\langle a,b\rangle,d)\in \mathsf{STORAGE}} \left[b \in \mathsf{Pop}(x) \vee a \in \mathsf{Pop}(x)\right] .$$

Deze relatie is er omdat men anders haast geen verdelingen in een multiset kan veranderen. Men moet in de functies fd er zelf op letten dat men met de juiste elementen in de standaard multiset werkt. Het is zeer wel mogelijk dat bijvoorbeeld $\{[(\langle a,b\rangle, d), (\langle b,l\rangle, f)]\} \subseteq \mathsf{STORAGE}$, terwijl men naar die tupels zoekt waarin b voorkomt. Moet men de "rechterkant" ($\pi_2(\mathsf{STORAGE})$) of de "linkerkant" ($\pi_1(\mathsf{STORAGE})$) hebben? De volgorde van cfd en sv in een padexpressie bepaalt de keus.

4.3 LISA-D en IR: het probabilistische afleidingsmodel

Er zal hier worden geprobeerd het model voor information retrieval systemen te integreren in LISA-D en dan wel op het niveau van padexpressies. Aan een aantal aspecten van het model wordt hier aandacht besteed waaronder het disjunct en niet-disjunct zijn van concepten, de queryformulering, sorteren naar mate van relevantie en het automatisch indexeren van documenten. Bij de integratie in padexpressies zal gebruik worden gemaakt van de voorbeeldtoepassing in paragraaf 3.2.3 op bladzijde 50.

4.3.1 Query: formulering en bruikbaarheid

Er wordt hier gekeken naar hoe in de praktijk naar documenten wordt gezocht en hoe een verzoek om bepaalde documenten middels een LISA-D query wordt gedaan. Hier is het van belang hoe zo'n query op het niveau van padexpressies er uit ziet.
Er wordt om documenten gevraagd waarbij erna het onderwerp waarop deze documenten betrekking moeten hebben, wordt gegeven. Het retrieval proces bestaat uit het voor **elk** document

nagaan of deze relevant is, dat wil zeggen of deze voldoet aan het verzoek. Als het verzoek niet nauwkeurig of specifiek genoeg is, dan is de mate waarin het document voldoet aan het verzoek bepalend. Dus de relevantie wordt bepaald. Dit principe geldt niet alleen in het probabilistische afleidingsmodel, maar ook in het model van Turtle en Croft en het model van Bruza en Van der Gaag.

Een verzoek is in LISA-D gemakkelijk geformuleerd:

LIST Documenten OVER Onderwerp

Als men van de documenten iets specifieks wil zoals auteur, titel of iets dergelijks, dan kan dat ook in de query meegegeven worden.
Op het niveau van padexpressies zou de query bijvoorbeeld er zo uit kunnen zien:

Documenten ∘ ⋯ ∘ Onderwerp

Dat tussen de rondjes stippellijntjes staan, is vanwege het feit dat niet bekend is wat het woord OVER voor staat. Het is namelijk een verpakking van iets waar men niets van hoeft te weten.
Men begint dus eerst met een collectie documenten waarop in tussentijdse stappen die in de padexpressie worden aangegeven handelingen worden uitgevoerd, om dan uiteindelijk bij het onderwerp uit te komen en men intussen een selectie van documenten heeft overgehouden die betrekking op het onderwerp kan hebben. Dit proces gaat vrijwel analoog aan het hierboven beschreven retrieval proces. Met de query zijn dus geen problemen te verwachten. Dat wil zeggen men hoeft met het verzoek geen kunsten uit te halen om retrieval te bewerkstelligen.

4.3.1.1 Het modelleren van de query in de informatiestructuur

In LISA-D moeten gegevens in de vorm van een populatie van de informatiestructuur aanwezig zijn, wil een query als padexpressie resultaat opleveren. Dit betekent concreet dat het **onderwerp** waarop de documenten betrekking moeten hebben, in de informatiestructuur en diens populatie moet zijn opgenomen. Dit is een vervelende zaak, omdat er heel wat onderwerpen mogelijk zijn. Als elke document wordt gekarakteriseerd door q index termen en er zijn n documenten, dan zijn er $\sum_{k=1}^{q} \binom{n}{k}$ onderwerpen mogelijk ($n \geq q$). De hoeveelheid kan in totaal exponentieel zijn.
Een direkte oplossing ligt niet voor de hand. Een idee is dat men de invoer zou kunnen scannen op een opgegeven reeks termen en die termen een voor een in een multiset plaatsen, maar de concrete uitwerking ervan is niet duidelijk. Vooralsnog maken we het ons makkelijk door de queries expliciet in de informatiestructuur te modelleren.

4.3.2 De informatiestructuur

Figuur 4.6 is een informatiestructuur waarin de volgende gegevens zijn gemodelleerd:

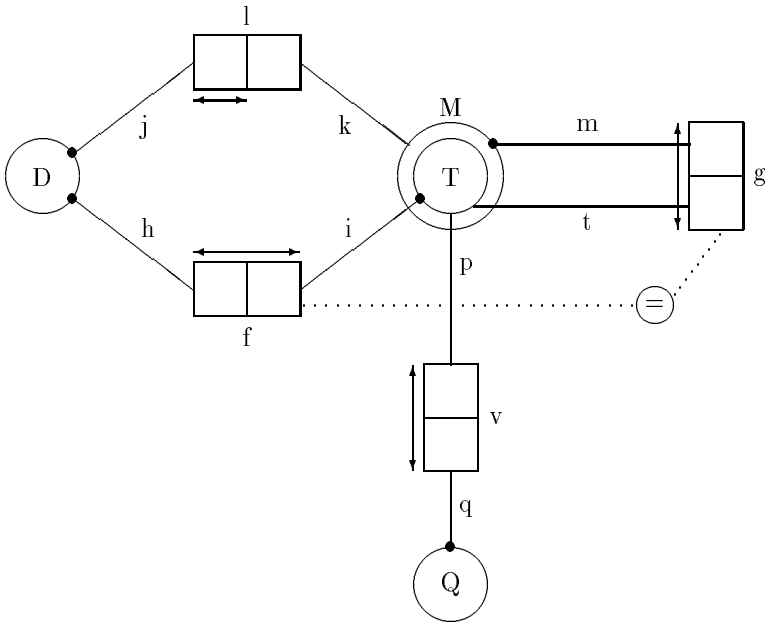
- Objecttype D stelt de verzameling documenten voor;
- Objecttype T stelt de verzameling index termen voor;
- Voor het geval dat de index termen niet disjunct zijn, zijn er nog de atomaire concepten die in de powertype M zijn vertegenwoordigd. Dat klopt eigenlijk niet helemaal: in het artikel ([WY91]) is $m_i = t_1 \cap \ldots \cap t_n$ en in de informatiestructuur is $m_i = \{t_1, \ldots t_n\}$;
- De objecttype Q bevat de onderwerpen die in queries kunnen worden opgenomen;
- Het feittype f geeft de karakterisaties van de documenten middels de concepten in de vorm van index termen;
- Het feittype v drukt de relaties tussen onderwerpen en de termconcepten uit;

- De feittypen l en g zijn een verhaal apart. De totale rol constraint bij objecttype D en de uniciteitsconstraint zorgt ervoor dat elk document in de collectie precies één atomair concept heeft. Als dat bekend is, dan kan precies aangegeven worden welke termen met de atomaire concepten gerelateerd zijn en het moet precies die termen zijn waarmee de met die atomaire concepten gerelateerde documenten zijn gekarakteriseerd. Om aan te geven wat er eigenlijk bedoeld wordt, staat hieronder een update request dat ook met LISA-D mogelijk is:

  ADD D 'd_1' ASSOCIATED-WITH T 't_1' ASSOCIATED-WITH M ASSOCIATED-WITH THAT D

  De matrix **G** in de voorbeeldtoepassing is op deze manier ingevuld. De vergelijking 3.29 op bladzijde 47 speelt hierbij een cruciale rol.
  Opmerking: feittype g is hier **niet** de impliciete feittype $\in_M$!

De informatiestructuur is ook geldig voor disjuncte concepten. Alleen kan in dat geval de powertype M en de feittypen l en g weggelaten worden.



Figuur 4.6: Een informatiestructuur voor de nodige gegevens voor toepassing van het probabilistische afleidingsmodel.

4.3.3 De onzekerheidspopulatie

In paragraaf 3.2.3 op bladzijde 50 is een voorbeeld toepassing van het model gegeven. Een studie naar deze toepassing leert, dat er met natuurlijke getallen en breuken gewerkt wordt. De relevanties van documenten voor een query worden ook in termen van breuken gegeven. Dus kan er veilig gesteld worden dat het domein voor de onzekerheid als volgt is:

$$\mathcal{U} = \mathbb{Q} .$$

Het bepalen van de relevantie geschiedt door matrices te vermenigvuldigen. Deze operatie is een combinatie van vermenigvuldigen en optellen. Zoals de eenheidsmatrix bestaat uit nullen en op de diagonaal enen, is voor de vermenigvuldiging het eenheidselement (neutraal element) 1. Het ligt dan voor de hand om voor het neutrale element te kiezen:

$$\xi_{\mathcal{U}} = 1$$

Over de onzekerheidspopulaties kan nu de volgende belangrijke opmerking gemaakt worden.

Propositie 4.3.1 Volgens de definitie van de onzekerheidspopulatie is aan de instanties van alle objecttypen initieel de waarde $\xi_{\mathcal{U}}$ toegekend. Voor alle objecttypen met uitzondering van de feittypen f en v moet dit worden gehandhaafd. Dit geldt ook voor uitbreidingen van de informatiestructuur (figuur 4.6).

De populatie van feittype f geeft de relaties tussen documenten en termen (trefwoorden) weer. Bij de relaties moet kunnen worden aangegeven hoe vaak een term met een document is gerelateerd. De populatie van feittype v geeft aan welke termen in de query voorkomen.
In beide gevallen moet de gebruiker (of wie dan ook) de mogelijkheid krijgen de relaties correct weer te geven. De waarden die meegegeven worden, zijn waarden uit $\mathbb{N}\setminus\{0\}$. De waarde 0 kan weggelaten worden omdat een gepopuleerde relatie document–term en/of query–term vanzelfsprekend minstens één ($\xi_{\mathcal{U}}$) keer voorkomt.

Voorbeeld 4.3.1

In de voorbeeldtoepassing op bladzijde 50 staan matrices met hun invullingen gegeven. Deze gegevens zijn hier gemakkelijk in de populatie van de informatiestructuur (figuur 4.6) op te nemen. Waar wat verwerkt is, wordt vermeld.
Elk document, term of atomair concept of query is uniek en komt derhalve $\xi_{\mathcal{U}} = 1$ keer voor.

$$\begin{aligned}
\mathsf{UPop}(\mathrm{D}) &= \{(d_1,1),(d_2,1),(d_3,1),(d_4,1),(d_5,1)\}\\
\mathsf{UPop}(\mathrm{T}) &= \{(t_1,1),(t_2,1),(t_3,1)\}\\
\mathsf{UPop}(\mathrm{Q}) &= \{(q,1)\}\\
\mathsf{UPop}(\mathrm{M}) &= \{(\{t_1\},1),(\{t_2\},1),(\{t_1,t_2\},1),(\{t_1,t_3\},1)\}
\end{aligned}$$

De invulling van de matrix $\mathbf{F}$ *wordt in de populatie van feittype* f *verwerkt:*

$$\mathsf{UPop}(\mathrm{f}) = \left\{ \begin{array}{lll}
(\{\mathrm{h}:d_1,\mathrm{i}:t_1\},2), & (\{\mathrm{h}:d_1,\mathrm{i}:t_3\},1), & (\{\mathrm{h}:d_2,\mathrm{i}:t_1\},1),\\
(\{\mathrm{h}:d_3,\mathrm{i}:t_2\},1), & (\{\mathrm{h}:d_4,\mathrm{i}:t_1\},2), & (\{\mathrm{h}:d_4,\mathrm{i}:t_2\},1),\\
(\{\mathrm{h}:d_5,\mathrm{i}:t_1\},1), & (\{\mathrm{h}:d_5,\mathrm{i}:t_3\},2) &
\end{array} \right\}$$

De invulling van de matrix $\mathbf{V}$ *wordt in de populatie van feittype* v *verwerkt:*

$$\mathsf{UPop}(\mathrm{v}) = \{(\{\mathrm{p}:t_1,\mathrm{q}:q\},2),(\{\mathrm{p}:t_3,\mathrm{q}:q\},1)\}$$

Wat de populatie van feittype l *is, hangt af van wat de verzamelingen D_m voor elk van de atomaire concepten $m \in \mathsf{Pop}(\mathrm{M})$ zijn. Er is vastgesteld dat*

$$D_{m_1} = \{d_2\},\quad D_{m_2} = \{d_3\},\quad D_{m_3} = \{d_4\},\quad D_{m_5} = \{d_1,d_5\}$$

voor $m_1 = \{t_1\}$, $m_2 = \{t_2\}$, $m_3 = \{t_1,t_3\}$, $m_5 = \{t_1,t_3\}$. Dit wordt als volgt als populatie van feittype l *vertaald:*

$$\mathsf{UPop}(\mathrm{l}) = \left\{ \begin{array}{ll}
(\{\mathrm{j}:d_1,\mathrm{k}:\{t_1,t_3\}\},1), & (\{\mathrm{j}:d_2,\mathrm{k}:\{t_1\}\},1),\\
(\{\mathrm{j}:d_3,\mathrm{k}:\{t_2\}\},1), & (\{\mathrm{j}:d_4,\mathrm{k}:\{t_1,t_2\}\},1),\\
(\{\mathrm{j}:d_5,\mathrm{k}:\{t_1,t_3\}\},1) &
\end{array} \right\}$$

De matrices $\mathbf{H}$, $\mathbf{G}$, $\hat{\mathbf{F}}$, $\mathbf{W}$ *en* $\hat{\mathbf{W}}$ *zijn niet in de populatie verwerkt. Maar daar zijn ze ook niet voor bedoeld. Het zijn gegevens die voor de bepaling van de relevantie voor elk document nodig zijn.* □

4.3.4 Beschrijving van frequentieverdelingen en operaties daarop

Deze paragraaf gaat over de invulling van de frequentieverdeling. Verder zijn er nog padexpressies waarvan de werking met betrekking tot de verdelingen nog gespecificeerd moet worden. Voor elk document moet bij een gegeven query de relevantie bepaald worden. Als er precies een query is, dan is er maar één zo'n waarde te bepalen. Zijn er meerdere (verschillende) queries, dan is voor elke query een waarde dat bij het document behoort.
Ongeacht de frequentie van voorkomen van een element in een multiset dat de lokatie van de relevantie in de frequentieverdeling bepaalt, ongeacht waar dat getal staat moet steeds dezelfde relevantie uit de verdeling te halen zijn. Dan kan op elke lokatie dezelfde waarde worden ingevuld. Hieronder volgt een overzicht van hoe de verdelingen er uit horen te zien.

- De neutrale verdeling: $1_{\mathsf{SEQU}} = < \xi_{\mathcal{U}}, \ldots >$
- Objecttypen: als $(y, d) \in \mathsf{UPop}(x)$, dan is $\mathsf{seq}(d) = < d, d, d, \ldots >$
- Concatenatie van padexpressies: als $(\langle p, r \rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})$ en $(\langle r, q \rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})$, dan is
$$s \odot t = < s[0] * t[0], s[1] * t[1], \ldots, s[n \times m] * t[n \times m], \ldots >$$
Als de tupel $\langle p, q \rangle$ op meerdere manieren kan worden gevormd, dan wordt de verdeling berekend volgens het principe dat bij vereniging wordt beschreven.
- Vereniging van padexpressies: als $(\langle p, q \rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})$ en $(\langle p, q \rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})$, dan is
$$s \cup t = < s[0] + t[0], s[1] + t[1], \ldots, s[k] + t[k], \ldots >$$
voor $(\langle p, q \rangle, x) \in^{n+m} \mathcal{V}[\![P \cup Q]\!](\mathsf{UPop})$, $k = n + m$.

Tot zover zijn dit de belangrijkste omschrijvingen. Daarmee is een verzameling padexpressies voor het zekerheidsfactormodel semantisch volledig gedefinieerd. Hier wordt even een overzicht van de bedoelde padexpressies gegeven.

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
empty path	$\varnothing_{\mathcal{PE}}$	$\varnothing$
neutral path	$1_{\mathcal{PE}}$	$1_{\Omega \times \Omega \times \mathsf{SEQU}}$
constant	c	$\mathsf{Sqr}(\{[(c, 1_{\mathsf{SEQU}})]\})$
multiset	X	$\mathsf{Sqr}(X)$
objecttype	x	$\{\!\{(\langle y, y \rangle, \mathsf{seq}(d))\uparrow^1 \mid (y, d) \in \mathsf{UPop}(x)\}\!\}$
predicator	p	$\{\!\{(\langle v(p), v \rangle, 1_{\mathsf{SEQU}})\uparrow^1 \mid v \in \mathsf{Pop} \circ \mathsf{Fact}(p)\}\!\}$
reverse	$P^{\leftarrow}$	$\{\!\{(\langle q, p \rangle, s)\uparrow^n \mid (\langle p, q \rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})\}\!\}$
concatenate	$P \circ Q$	$\bigcup_{r, s \odot t} \left\{\!\!\left[(\langle p, q \rangle, s \odot t)\uparrow^{n \times m} \;\middle|\; \begin{array}{l} (\langle p, r \rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge \\ (\langle r, q \rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop}) \end{array} \right]\!\!\right\}$
union	$P \cup Q$	$\left\{\!\!\left[(\langle p, q \rangle, s \cup t)\uparrow^{n+m} \;\middle|\; \begin{array}{l} (\langle p, q \rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge \\ (\langle p, q \rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop}) \end{array} \right]\!\!\right\}$
distinct	$\mathsf{ds}\ P$	$\mathsf{Multi}(\mathsf{set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$
front	$\not{f}\ P$	$\mathsf{Sqr}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

4.3.5 Het bepalen van de relevantie

In paragraaf 3.2.3 op bladzijde 50 is een voorbeeldtoepassing van het model gegeven. In de volgende paragrafen zal ernaar verwezen worden. Bij toepassing van de padexpressies moeten voor elk document dezelfde relevanties te krijgen zijn. In de volgende paragrafen wordt ingegaan op de middelen om term frequenties, geïnverteerde document frequenties en dergelijke te krijgen.
In het probabilistische afleidingsmodel is sprake van disjuncte en niet-disjuncte concepten (termen). Er zal in de volgende paragrafen het geval van niet-disjuncte concepten worden uitgewerkt. Gaat dat namelijk goed, dan gaat het met disjuncte concepten zeker goed.

4.3.5.1 Term frequentie binnen documenten

Om de matrix $\hat{\mathbf{F}}$ te krijgen, moet de matrix $\mathbf{F}$ met de frequenties van voorkomen van termen binnen een document genormaliseerd worden. Deze frequentie wordt verkregen met behulp van de vergelijking 3.17

$$P(d \cap T) \approx \eta \sum_t f(d, t)$$

en $\hat{\mathbf{F}}$ wordt dan verkregen met behulp van vergelijking 3.18 verkregen:

$$\frac{P(d \cap t)}{P(d \cap T)} \approx \frac{f(d,t)}{\sum_{t} f(d,t)} = \hat{f}(d,t).$$

Om de frequenties te krijgen, moet er met de objecttype D een functie fd geassocieerd worden. Zij $d \in \mathsf{Pop}(\mathrm{D})$. Dan is de specificatie van fd(D) als volgt:

$$\mathrm{fd(D)} = \begin{cases} \mathsf{seq}\left(\dfrac{1}{\sum_{(\langle d,y\rangle,t)\in^{m}\mathsf{STORAGE}} t[m]}\right) & \text{, als } (d,t) \in \pi_1(\mathsf{STORAGE}) \\ 1_{\mathsf{SEQU}} & \text{, anders} \end{cases} \tag{4.1}$$

Het gebruik van de functie fd en de invulling van de standaard multiset wordt weergegeven in de volgende padexpressie waarmee dan de matrix $\hat{\mathbf{F}}$ in de vorm van een multiset kan worden verkregen:

$$\mathsf{cfd(D)} \circ \mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow})$$

4.3.5.2 Geïnverteerde document frequentie

Matrix **H** bevat de geïnverteerde frequentie van voorkomen van documenten (inverse document frequency). Deze frequenties verkrijgt men met de vergelijking 3.20:

$$P(t) \approx \eta \sum_{d'} f(d',t),$$

en dan deze te inverteren. Semantisch staat er dat per term het aantal documenten die met die term zijn gekarakteriseerd moet worden geteld. Om de frequenties te krijgen moet er met de objecttype T een functie fd geassocieerd worden. Zij $t \in \mathsf{Pop}(\mathrm{T})$. Dan is de specificatie van fd(T) als volgt:

$$\mathrm{fd(T)} = \begin{cases} \mathsf{seq}\left(\dfrac{1}{\sum_{(\langle y,t\rangle,v)\in^{m}\mathsf{STORAGE}} v[m]}\right) & \text{, als } (t,v) \in \pi_2(\mathsf{STORAGE}) \\ 1_{\mathsf{SEQU}} & \text{, anders} \end{cases} \tag{4.2}$$

Het gebruik van de functie fd en de invulling van de standaard multiset wordt weergegeven in de volgende padexpressie:

$$\mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd(T)}$$

4.3.5.3 Relaties tussen termen, documenten en atomaire concepten

Als termen niet disjunct zijn, dan worden uit de termen atomaire concepten gegenereerd die onderling wel disjunct zijn. Aan de hand van de termen die met de documenten gerelateerd zijn wordt bepaald welk atomair concept met welk document wordt gerelateerd. Een verzameling D_{m_i} bevat de documenten die met term m_i gerelateerd zijn.
De matrix **G** geeft aan hoe vaak in een atomaire concept m_i de termen t_j gebruikt worden bij de documenten die in de verzameling D_{m_i} voorkomen. In de informatiestructuur (figuur 4.6) zijn verzamelingen D_{m_i} middels feittype l als een relatie tussen documenten en atomaire concepten m_i gemodelleerd. De populatie van het feittype geeft de daadwerkelijke relaties weer.
Het is nu de vraag hoe men aan de gegevens komt zoals dat in matrix **G** staat. Dat zou in principe in feittype g gepopuleerd moeten zijn, maar daar staat alleen welke termen in welke concepten voorkomen. Er is een oplossing. Er is de relatie tussen documenten en termen en er is de zojuist beschreven relatie tussen documenten en concepten. Als men nu de padexpressie G evalueert waarin

$$G = \mathrm{M} \circ \mathrm{k} \circ \mathrm{l} \circ \mathrm{j}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow} \circ \mathrm{T}\,, \tag{4.3}$$

dan levert dit als resultaat op:

$$\nu[\![G]\!](\mathsf{UPop}) \;=$$

$$\left\{\left[\begin{array}{lll} (\langle m_1,t_1\rangle,<1,1,\ldots>)\uparrow^1, & & \\ & (\langle m_2,t_2\rangle,<1,1,\ldots>)\uparrow^1, & \\ (\langle m_3,t_1\rangle,<2,2,\ldots>)\uparrow^1, & (\langle m_3,t_2\rangle,<1,1,\ldots>)\uparrow^1, & \\ (\langle m_5,t_1\rangle,<3,3,\ldots>)\uparrow^2, & & (\langle m_5,t_3\rangle,<3,3,\ldots>)\uparrow^2 \end{array}\right]\right\}$$

Vergelijk deze multiset met de matrix $\mathbf{g}$. Een ander padexpressie dat met betrekking tot de verdelingen hetzelfde resultaat oplevert, is

$$\mathrm{T}\circ\mathrm{i}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow}\circ\mathrm{D}\circ\mathrm{j}\circ\mathrm{l}\circ\mathrm{k}^{\leftarrow}\circ\mathrm{M} \;=\; G^{\leftarrow} \tag{4.4}$$

Nu is het nog maar een "kleine" stap om multisets a la matrices $\mathbf{W}$ en $\hat{\mathbf{W}}$ te krijgen, zoals dat in het volgende voorbeeld getoond wordt.

Voorbeeld 4.3.2

In de voorgaande paragrafen zijn de mogelijkheden beschreven om de geïnverteerde document frequentie, term frequentie en een matrix als multiset te creëren. Met deze middelen wordt hier geprobeerd middels padexpressies multisets à la matrices $\mathbf{W}$ *en* $\hat{\mathbf{W}}$ *te krijgen.*
$\mathbf{W}$ *wordt gecreëerd uit de matrices* $\mathbf{F}$, $\mathbf{H}$ *en* $\mathbf{G}$*:* $\mathbf{W} = \mathbf{FHG}$*. De multiset dat hetzelfde resultaat oplevert, wordt verkregen door de volgende padexpressie te evalueren:*

$$X \;=\; \mathrm{D}\circ\mathsf{sv}(\mathrm{h}\circ\mathrm{f}\circ\mathrm{i}^{\leftarrow})\circ\mathsf{cfd}(\mathrm{T})\circ\mathrm{i}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow}\circ\mathrm{D}\circ\mathrm{j}\circ\mathrm{l}\circ\mathrm{k}^{\leftarrow}\circ\mathrm{M} \tag{4.5}$$

Ga na dat met de gegeven populatie in voorbeeld 4.3.1 de evaluatie van X *het volgende resultaat oplevert:*

$$\nu[\![X]\!](\mathsf{UPop}) \;=\; \left\{\left[\begin{array}{ll} (\langle d_1,m_1\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, & (\langle d_4,m_1\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, \\ (\langle d_1,m_3\rangle,<\tfrac{2}{3},\tfrac{2}{3},\ldots>)\uparrow^1, & (\langle d_4,m_2\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1, \\ (\langle d_1,m_5\rangle,<2,2,\ldots>)\uparrow^4, & (\langle d_4,m_3\rangle,<\tfrac{7}{6},\tfrac{7}{6},\ldots>)\uparrow^2, \\ (\langle d_2,m_1\rangle,<\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1, & (\langle d_4,m_5\rangle,<1,1,\ldots>)\uparrow^2, \\ (\langle d_2,m_3\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, & (\langle d_5,m_1\rangle,<\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1, \\ (\langle d_2,m_5\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^2, & (\langle d_5,m_3\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, \\ (\langle d_3,m_2\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1, & (\langle d_5,m_5\rangle,<\tfrac{5}{2},\tfrac{5}{2},\ldots>)\uparrow^4, \\ (\langle d_3,m_3\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1 & \end{array}\right]\right\}$$

Matrix $\hat{\mathbf{W}}$ *wordt gecreëerd uit de matrices* $\hat{\mathbf{F}}$, $\mathbf{H}$ *en* $\mathbf{G}$*:* $\hat{\mathbf{W}} = \hat{\mathbf{F}}\mathbf{HG}$*. De multiset dat hetzelfde resultaat oplevert, wordt verkregen door de volgende padexpressie te evalueren:*

$$Y \;=\; \mathsf{cfd}(\mathrm{D})\circ\mathsf{sv}(\mathrm{h}\circ\mathrm{f}\circ\mathrm{i}^{\leftarrow})\circ\mathsf{cfd}(\mathrm{T})\circ\mathrm{i}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow}\circ\mathrm{D}\circ\mathrm{j}\circ\mathrm{l}\circ\mathrm{k}^{\leftarrow}\circ\mathrm{M} \tag{4.6}$$

Met de gegeven populatie in voorbeeld 4.3.1 levert evaluatie van Y *het volgende resultaat op:*

$$\nu[\![Y]\!](\mathsf{UPop}) \;=\; \left\{\left[\begin{array}{ll} (\langle d_1,m_1\rangle,<\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1, & (\langle d_4,m_1\rangle,<\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1, \\ (\langle d_1,m_3\rangle,<\tfrac{2}{9},\tfrac{2}{9},\ldots>)\uparrow^1, & (\langle d_4,m_2\rangle,<\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1, \\ (\langle d_1,m_5\rangle,<\tfrac{2}{3},\tfrac{2}{3},\ldots>)\uparrow^4, & (\langle d_4,m_3\rangle,<\tfrac{7}{18},\tfrac{7}{18},\ldots>)\uparrow^2, \\ (\langle d_2,m_1\rangle,<\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1, & (\langle d_4,m_5\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^2, \\ (\langle d_2,m_3\rangle,<\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, & (\langle d_5,m_1\rangle,<\tfrac{1}{18},\tfrac{1}{18},\ldots>)\uparrow^1, \\ (\langle d_2,m_5\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^2, & (\langle d_5,m_3\rangle,<\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1, \\ (\langle d_3,m_2\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1, & (\langle d_5,m_5\rangle,<\tfrac{5}{6},\tfrac{5}{6},\ldots>)\uparrow^4, \\ (\langle d_3,m_3\rangle,<\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1 & \end{array}\right]\right\}$$

□

4.3.5.4 Het query gedeelte

In de voorbeeldtoepassing onder het kopje "niet-disjuncte concepten" moeten de gegevens voor de query bepaald worden. Dit moet dan volgens vergelijking 3.48:

$$\mathbf{q}_m = \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}}.$$

De matrices $\mathbf{V}$ is verwerkt in de populatie van feittype v. Van de matrices $\mathbf{G}$ en $\mathbf{H}$ is beschreven hoe aan die gegevens te komen is. Maar hoe komt men aan de noemer in de bovenstaande vergelijking? In vergelijking 3.45 staat hoe men daaraan komt:

$$\begin{aligned} P(m) &\approx \eta \sum_d \left[\sum_t f(d,t)h(t)g(t,m) \right] \\ &= \eta \sum_d \left[\sum_t \mathbf{F}_{dt}\mathbf{H}_{tt}\mathbf{G}_{tm} \right] \\ &= \eta \sum_d \mathbf{W}_{dm}. \end{aligned}$$

In voorbeeld 4.3.2 is een manier van "berekening" van matrix $\mathbf{W}$ gedemonstreerd. Dus met de padexpressie $X = \mathrm{D} \circ \mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow} \circ \mathrm{M}$ zijn de nodige gegevens voorhanden. Die gegevens moeten ingezet worden, maar waar?
De bedoeling is dat per document voor een gegeven query q de relevantie bepaald wordt. Dus er uitkomst van evaluatie van de bijbehorende padexpressie zal een multiset moeten zijn met tupels van de vorm $\langle d_i, q \rangle$ er in. De relevantie moet bepaald worden volgens vergelijking 3.47:

$$\begin{aligned} \Psi(d \rightarrow q) &\approx \sum_{m \in M} \left[\hat{\mathbf{W}}_{dm} \cdot \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}} \right] \\ &= \sum_{m \in M} \hat{\mathbf{W}}_{dm}\mathbf{q}_m = (\hat{\mathbf{W}}\mathbf{q})_d, \end{aligned}$$

waarin q_m een $(u \times 1)$-matrix is die als volgt is gedefinieerd:

$$\mathbf{q}_m = \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}}.$$

De multiset dat $\hat{\mathbf{W}}$ voorstelt, bevat tupels $\langle d_i, m_j \rangle$. Om $\hat{\mathbf{W}} \cdot q_m$ uit te rekenen moet er een padexpressie voor q_m komen die bij evaluatie een multiset oplevert met tupels $\langle m_j, q \rangle$ er in. In de vergelijking voor q_m staat $\mathbf{VHG}$. Vertaald naar een padexpressie begint deze bij q en eindigt deze bij m_j. Dat moet precies andersom, dus $\mathbf{G}^T\mathbf{H}\mathbf{V}^T$. De T staat voor transponeren ofwel het verwisselen van rijen met kolommen en omgekeerd.
Een padexpressie dat bij evaluatie een multiset á la $\mathbf{G}^T\mathbf{H}\mathbf{V}^T$ oplevert, is:

$$Z = \mathrm{M} \circ \mathrm{k} \circ \mathrm{l} \circ \mathrm{j}^{\leftarrow} \circ \mathrm{D} \circ \mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{j}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{p} \circ \mathrm{v} \circ \mathrm{q}^{\leftarrow} \circ \mathrm{Q} \qquad (4.7)$$

Het resultaat van evaluatie van expressie Z is als volgt (ga na):

$$\nu[\![Z]\!](\mathsf{UPop}) = \left\{\!\!\left[\begin{array}{l} (\langle m_1, q \rangle, < \frac{1}{3}, \frac{1}{3}, \ldots >)\uparrow^1 \\ (\langle m_3, q \rangle, < \frac{2}{3}, \frac{2}{3}, \ldots >)\uparrow^1 \\ (\langle m_5, q \rangle, < 2, 2, \ldots >)\uparrow^4 \end{array} \right]\!\!\right\}$$

Dit is dan volgens de uitkomst van matrix $\mathbf{VHG}$ in het voorbeeldtoepassing op bladzijde 50.

Voorbeeld 4.3.3

Tot zover ontbreekt nog de mogelijkheid om de noemer in de vergelijking voor q_m te berekenen. Met de tot dusver bereikte resultaten kan toch iets gedaan worden. Zij Y de padexpressie in voorbeeld 4.3.2 en zij Z de padexpressie als in de paragraaf hiervoor genoemd. Dan is $Y \circ Z$ weer een padexpressie die na evaluatie ervan het volgende resultaat oplevert:

$$\mathcal{V}[\![Y \circ Z]\!](\mathsf{UPop}) \;=\; \mathcal{V}[\![Y]\!](\mathsf{UPop}) \circ \mathcal{V}[\![Z]\!](\mathsf{UPop}) \;=$$

$$\left\{\begin{array}{ll}
(\langle d_1, m_1\rangle, <\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1, & (\langle d_4, m_1\rangle, <\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1,\\
(\langle d_1, m_3\rangle, <\tfrac{2}{9},\tfrac{2}{9},\ldots>)\uparrow^1, & (\langle d_4, m_2\rangle, <\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1,\\
(\langle d_1, m_5\rangle, <\tfrac{2}{3},\tfrac{2}{3},\ldots>)\uparrow^4, & (\langle d_4, m_3\rangle, <\tfrac{7}{18},\tfrac{7}{18},\ldots>)\uparrow^2,\\
(\langle d_2, m_1\rangle, <\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1, & (\langle d_4, m_5\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^2,\\
(\langle d_2, m_3\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1, & (\langle d_5, m_1\rangle, <\tfrac{1}{18},\tfrac{1}{18},\ldots>)\uparrow^1,\\
(\langle d_2, m_5\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^2, & (\langle d_5, m_3\rangle, <\tfrac{1}{9},\tfrac{1}{9},\ldots>)\uparrow^1,\\
(\langle d_3, m_2\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1, & (\langle d_5, m_5\rangle, <\tfrac{5}{6},\tfrac{5}{6},\ldots>)\uparrow^4,\\
(\langle d_3, m_3\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1 &
\end{array}\right\} \circ$$

$$\left\{\begin{array}{l}
(\langle m_1, q\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1\\
(\langle m_3, q\rangle, <\tfrac{2}{3},\tfrac{2}{3},\ldots>)\uparrow^1\\
(\langle m_5, q\rangle, <2,2,\ldots>)\uparrow^4
\end{array}\right\} \;=\; \left\{\begin{array}{l}
(\langle d_1, q\rangle, <\tfrac{41}{27},\tfrac{41}{27},\ldots>)\uparrow^{18}\\
(\langle d_2, q\rangle, <\tfrac{23}{18},\tfrac{23}{18},\ldots>)\uparrow^{10}\\
(\langle d_3, q\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1\\
(\langle d_4, q\rangle, <\tfrac{26}{27},\tfrac{26}{27},\ldots>)\uparrow^{11}\\
(\langle d_5, q\rangle, <\tfrac{95}{54},\tfrac{95}{54},\ldots>)\uparrow^{18}
\end{array}\right\}$$

Wat hier is uitgerekend, is overeenkomstig de formule voor $\Psi(d \to q)$ in vergelijking 3.49:

$$\Psi(d \to q) \;\approx\; \sum_{m \in M} \hat{\mathbf{W}}_{dm}\hat{\mathbf{q}}_m \;=\; (\hat{\mathbf{W}}\hat{\mathbf{q}})_d,$$

waarin $\hat{\mathbf{q}}$ een $(u\times 1)$-matrix is met $\hat{\mathbf{q}}_m = (\mathbf{VHG})_m$. Dit is de vergelijking zonder $1/\sum_{d'} \mathbf{W}_{d'm}$ en volgens Wong en Yao ([WY91]) is die op enkele verschillen in normalisaties na nagenoeg identiek aan de formule voor het gegeneraliseerde vectorruimte model *(*Generalized Vector Space Model, GVSM*).* □

Er is één plek waar de noemer in de vergelijking voor q_m in een padexpressie verwerkt kan worden. In voorbeeld 4.3.3 wordt al een beeld gegeven van hoe de padexpressie eruit ziet en wat er van verwacht wordt. De plaats om $\sum_{d'} \mathbf{W}_{d'm}$ te laten berekenen is objecttype M. Immers, de padexpressie Y in het voorbeeld "beschrijft" de matrix $\hat{\mathbf{W}}$ en als de expressie cfd(D) nu D was, dan had men de padexpressie X om matrix $\mathbf{W}$ te krijgen.
Er moet nu een specificatie voor de functie fd gegeven worden. Deze functie wordt geassocieerd met objecttype M. Zij $m \in$ Pop(M). Dan is de specificatie van fd(M) als volgt:

$$\mathrm{fd(M)} \;=\; \begin{cases} \mathsf{seq}\left(\dfrac{1}{\displaystyle\sum_{(\langle x,m\rangle,t)\in^p \mathsf{STORAGE}} t[p]}\right) & \text{, als } (m,t) \in \pi_2(\mathsf{STORAGE}) \\ 1_{\mathsf{SEQU}} & \text{, anders} \end{cases} \qquad (4.8)$$

De padexpressie waarin de functie wordt gebruikt en waarvoor de nodige gegevens in de standaard multiset wordt geplaatst, is als volgt:

$$W \;=\; \mathsf{cfd(D)} \circ \mathsf{sv}(\mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd(T)} \circ \mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow}) \circ \mathsf{cfd(M)} \qquad (4.9)$$

4.3.5.5 Relevanties voor documenten: een demonstratie

In de voorgaande paragrafen zijn manieren beschreven om de nogige gegevens te vergaren. Nu kan geprobeerd worden het uiteindelijke doel te bereiken, namelijk het bepalen van de relevantie

voor elk document gegeven een query. In vergelijking 3.47 op bladzijde 49 is de formule voor het bepalen van die relevanties beschreven. Deze vergelijking volgt hieronder:

$$\begin{aligned}\Psi(d \rightarrow q) &\approx \sum_{m \in M} \left[\hat{\mathbf{W}}_{dm} \cdot \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}} \right] \\ &= \sum_{m \in M} \hat{\mathbf{W}}_{dm} \mathbf{q}_m = (\hat{\mathbf{W}}\mathbf{q})_d,\end{aligned}$$

waarin q_m een $(u \times 1)$-matrix is die als volgt is gedefinieerd:

$$\mathbf{q}_m = \frac{(\mathbf{VHG})_m}{\sum_{d'} \mathbf{W}_{d'm}}.$$

In voorbeeld 4.3.2 is beschreven hoe men aan matrix $\hat{\mathbf{W}}$ komt. Voor q_m is een beschrijving in de vorige paragraaf gegeven.
Hier wordt nu een padexpressie geconstrueerd die $\hat{\mathbf{W}}\mathbf{q}$ nabootst. Eerst wordt naar de delen voor $\hat{\mathbf{W}}$ en q_m gekeken.

Een expressie voor $\hat{\mathbf{W}}$

Hier wordt eerst een padexpressie gegeven dat $\hat{\mathbf{W}}$ en een gedeelte van de formule voor q_m nabootst. Zij de padexpressie W als volgt:

$$W = \mathsf{cfd}(\mathrm{D}) \circ \mathsf{sv}(\mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{M}) \quad (4.10)$$

De delen van deze padexpressie hebben de volgende functies:

- $\mathsf{cfd}(\mathrm{D})$ levert de termfrequenties,zodat het stukje expressie $\mathsf{cfd}(\mathrm{D}) \circ \mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow})$ een matrix $\hat{\mathbf{F}}$ oplevert;
- $\mathsf{sv}(\mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow})$ stelt hier de matrix $\mathbf{W}$ voor (zie voorbeeld 4.3.2). Het resultaat van evaluatie van deze expressie wordt bovendien in STORAGE opgeslagen ten behoeve van $\mathsf{cfd}(\mathrm{M})$;
- $\mathsf{cfd}(\mathrm{M})$ levert een gedeelte van de gevens voor q_m, namelijk $\frac{1}{\sum_{d'} \hat{\mathbf{W}}_{d'm}}$;
- $\mathsf{cfd}(\mathrm{T})$ levert de matrix $\mathbf{H}$ met de geïnverteerde document frequenties. Het stukje expressie $\mathsf{sv}(\mathrm{h} \circ \mathrm{f} \circ \mathrm{i}^{\leftarrow})$ zorgt voor de nodige gegevens;
- De deelexpressie $\mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow} \circ \mathsf{cfd}(\mathrm{M})$ bootst de matrix $\mathbf{G}$ na.

Een expressie voor q_m

Zij de padexpressie Z dat de uitkomst $\mathbf{G}^T\mathbf{H}\mathbf{V}^T$ moet nabootsen als volgt:

$$Z = \mathrm{M} \circ \mathrm{k} \circ \mathrm{l} \circ \mathrm{j}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{j}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{p} \circ \mathrm{v} \circ \mathrm{q}^{\leftarrow} \circ \mathrm{Q} \quad (4.11)$$

De delen van deze padexpressie hebben de volgende functies:

- $\mathsf{cfd}(\mathrm{T})$ levert de matrix $\mathbf{H}$ met de geïnverteerde document frequenties. Het stukje expressie $\mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow})$ zorgt voor de nodige gegevens;
- De deelexpressie $\mathrm{M} \circ \mathrm{k} \circ \mathrm{l} \circ \mathrm{j}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow})$ bootst de matrix $\mathbf{G}^T$ na;
- De deelexpressie $\mathrm{p} \circ \mathrm{v} \circ \mathrm{q}^{\leftarrow} \circ \mathrm{Q}$ produceert de matrix $\mathbf{V}^T$

De noemer van de formule voor q_m wordt in de expressie W verzorgd door $\mathsf{cfd}(\mathrm{M})$.

Nogmaals: term frequenties binnen documenten

Er is met cfd(D) echter een probleem. Voor de functie fd die met objecttype D is geassocieerd zijn gegevens nodig die echter niet beschikbaar zijn, om de eenvoudige reden dat er geen operatie sv is gebruikt die de nodige juiste gegevens in de standaard multiset laat opslaan.
Een blik op de padexpressies W en Z leert het volgende: er zijn momenteel 3 sv-operaties in de padexpressies waarvan er twee onderling wezenlijk verschillen. In de standaard multiset kunnen elementen van de vorm $(\langle d_i, m_j \rangle, \ldots)$ en $(\langle f_k, t_n \rangle, \ldots)$ voorkomen. f_k is hier de instantie van de populatie van feittype f (zie informatiestructuur op bladzijde 90). Dat zijn niet de correcte gegevens voor de functie fd(D).
Er is een oplossing mogelijk maar deze hangt af van hoe de gegevens in de standaard multiset worden verwerkt. Er zijn hier twee mogelijkheden. Als $\mathsf{sv}(P)$ een padexpressie is, dan is :

1. $\mathsf{STORAGE} = \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

2. $\mathsf{STORAGE} = (\mathsf{STORAGE} - \mathcal{V}[\![P]\!](\mathsf{UPop})) \cup \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

Omdat er meerdere sv-operaties op hetzelfde niveau in de padexpressie zitten (bijvoorbeeld $\mathsf{sv}(f) \circ \mathsf{cfd}(g) \circ \mathsf{sv}(h)$), en volgens propositie 4.2.1 altijd eerst uitgevoerd moeten worden voordat een cfd-operatie wordt uitgevoerd, is bij de eerste mogelijkheid het risico van verlies van of te weinig informatie te groot. Dus valt hier de keus op de tweede mogelijkheid:

$$\mathsf{STORAGE} = (\mathsf{STORAGE} - \mathcal{V}[\![P]\!](\mathsf{UPop})) \cup \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![P]\!](\mathsf{UPop}))) \tag{4.12}$$

Nu de tweede mogelijkheid wordt toegepast, kunnen elementen van de vorm $(\langle d_i, m_j \rangle, \ldots)$ en $(\langle f_k, t_n \rangle, \ldots)$ tegelijkertijd in de standaard multiset voorkomen. De gegevens die voor de met objecttype D geassocieerde functie fd nodig zijn, moeten in STORAGE te onderscheiden zijn. Dat kan alleen als elementen van de vorm $(\langle \ldots, d_i \rangle, \ldots)$, $d_i \in \mathsf{Pop(D)}$ in STORAGE worden verwerkt. Dit heeft tot gevolg dat de specificatie van fd(D) moet worden herzien. Deze was gemaakt om met elementen van de vorm $(\langle d_i, \ldots \rangle, \ldots)$ te werken. Hieronder volgt een herziene specificatie van de functie fd die met objecttype D is geassocieerd.

$$\mathrm{fd(D)} = \begin{cases} \mathsf{seq}\left(\dfrac{1}{\sum\limits_{(\langle y,d\rangle,t)\in^m \mathsf{STORAGE}} t[m]}\right) & \text{, als } (d,t) \in \pi_2(\mathsf{STORAGE}) \\ 1_{\mathsf{SEQU}} & \text{, anders} \end{cases} \tag{4.13}$$

In deze specificatie is de tupel $\langle y, d \rangle$ nu correct, en de projectie ook (2 in plaats van 1).

Er zijn twee mogelijkheden om de gegevens ten behoeve van de functie fd(D) in de standaard multiset te verwerken:

1. In de padexpressie Z de deelexpressie $\mathrm{f} \circ \mathrm{h}^{\leftarrow}$ vervangen door $\mathsf{sv}(\mathrm{f} \circ \mathrm{h}^{\leftarrow})$;

2. Een expressie $\mathsf{sv}(\mathrm{D} \circ \mathrm{h} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow})$ met W te concateneren, waarbij W achteraan staat.

Beide mogelijkheden voldoen. Voor de overzichtelijkheid en de structuur wordt voor de tweede mogelijkheid gekozen. Als in de padexpressies W en Z en de aanvulling van W alle sv-operaties zou zijn geëvalueerd en men werkt met de populatie die in voorbeeld 4.3.1, dan levert dit de

volgende standaard multiset op:

$$\mathsf{STORAGE} = \left\{\!\!\left[\begin{array}{ll}
(\langle d_1,d_1\rangle, <3,3,\ldots>)\uparrow^1, & (\langle d_1,m_1\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1,\\
(\langle d_2,d_2\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_1,m_3\rangle, <\tfrac{2}{3},\tfrac{2}{3},\ldots>)\uparrow^1,\\
(\langle d_3,d_3\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_1,m_5\rangle, <2,2,\ldots>)\uparrow^1,\\
(\langle d_4,d_4\rangle, <3,3,\ldots>)\uparrow^1, & (\langle d_2,m_1\rangle, <\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1,\\
(\langle d_5,d_5\rangle, <3,3,\ldots>)\uparrow^1, & (\langle d_2,m_3\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1,\\
 & (\langle d_2,m_5\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_1,\mathrm{i}:t_1\},t_1\rangle, <2,2,\ldots>)\uparrow^1, & (\langle d_3,m_2\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_2,\mathrm{i}:t_1\},t_1\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_3,m_3\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_4,\mathrm{i}:t_1\},t_1\rangle, <2,2,\ldots>)\uparrow^1, & (\langle d_4,m_1\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_5,\mathrm{i}:t_1\},t_1\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_4,m_2\rangle, <\tfrac{1}{2},\tfrac{1}{2},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_3,\mathrm{i}:t_2\},t_2\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_4,m_3\rangle, <\tfrac{7}{6},\tfrac{7}{6},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_4,\mathrm{i}:t_2\},t_2\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_4,m_5\rangle, <1,1,\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_1,\mathrm{i}:t_3\},t_3\rangle, <1,1,\ldots>)\uparrow^1, & (\langle d_5,m_1\rangle, <\tfrac{1}{6},\tfrac{1}{6},\ldots>)\uparrow^1,\\
(\langle\{\mathrm{h}:d_5,\mathrm{i}:t_3\},t_3\rangle, <2,2,\ldots>)\uparrow^1, & (\langle d_5,m_3\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1,\\
 & (\langle d_5,m_5\rangle, <\tfrac{5}{2},\tfrac{5}{2},\ldots>)\uparrow^1
\end{array} \right]\!\!\right\}$$

Op basis van deze standaard multiset wordt nu getoond wat voor resultaten de padexpressies $\mathsf{cfd}(\mathrm{M})$, $\mathsf{cfd}(\mathrm{T})$ en $\mathsf{cfd}(\mathrm{D})$ opleveren.

$$\mathcal{V}[\![\mathsf{cfd}(\mathrm{D})]\!](\mathsf{UPop}) = \left\{\!\!\left[\begin{array}{c}
(\langle d_1,d_1\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1\\
(\langle d_2,d_2\rangle, <1,1,\ldots>)\uparrow^1\\
(\langle d_3,d_3\rangle, <1,1,\ldots>)\uparrow^1\\
(\langle d_4,d_4\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1\\
(\langle d_5,d_5\rangle, <\tfrac{1}{3},\tfrac{1}{3},\ldots>)\uparrow^1
\end{array} \right]\!\!\right\}$$

$$\mathcal{V}[\![\mathsf{cfd}(\mathrm{T})]\!](\mathsf{UPop}) = \left\{\!\!\left[\begin{array}{c}
(\langle t_1,t_1\rangle, <\frac{1}{2+1+2+1},\tfrac{1}{6},\ldots>)\uparrow^1\\
(\langle t_2,t_2\rangle, <\frac{1}{1+1},\tfrac{1}{2},\ldots>)\uparrow^1\\
(\langle t_3,t_3\rangle, <\frac{1}{1+2},\tfrac{1}{3},\ldots>)\uparrow^1
\end{array} \right]\!\!\right\}$$

$$\mathcal{V}[\![\mathsf{cfd}(\mathrm{M})]\!](\mathsf{UPop}) = \left\{\!\!\left[\begin{array}{c}
(\langle m_1,m_1\rangle, <\frac{1}{\frac{1}{3}+\frac{1}{6}+\frac{1}{3}+\frac{1}{6}},1,\ldots>)\uparrow^1\\
(\langle m_2,m_2\rangle, <\frac{1}{\frac{1}{2}+\frac{1}{2}},1,\ldots>)\uparrow^1\\
(\langle m_3,m_3\rangle, <\frac{1}{\frac{2}{3}+\frac{1}{3}+\frac{1}{2}+\frac{7}{6}+\frac{1}{3}},\tfrac{1}{3},\ldots>)\uparrow^1\\
(\langle m_5,m_5\rangle, <\frac{1}{2+\frac{1}{2}+1+\frac{5}{2}},\tfrac{1}{6},\ldots>)\uparrow^1
\end{array} \right]\!\!\right\}$$

Nu de elementaire gegevens bekend zijn, kunnen nu de expressies $\mathsf{sv}(\mathrm{D}\circ\mathrm{h}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow})\circ W$ en Z volledig geëvalueerd worden. Voor elk relevant deelexpressie ervan wordt een multiset getoond. Al die multisets worden tot één multiset als uitkomst van $\mathcal{V}[\![\mathsf{sv}(\mathrm{D}\circ\mathrm{h}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow})\circ W\circ Z]\!](\mathsf{UPop})$ samengevoegd. $\mathcal{V}[\![Z]\!](\mathsf{UPop})$ is in de paragraaf over het query gedeelte uitgewerkt. De padexpressie $\mathsf{sv}(\mathrm{h}\circ\mathsf{sv}(\mathrm{f}\circ\mathrm{i}^{\leftarrow})\circ\mathsf{cfd}(\mathrm{T})\circ\mathrm{i}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow}\circ\mathrm{D}\circ\mathrm{j}\circ\mathrm{l}\circ\mathrm{k}^{\leftarrow})$ is op enkele objecttypen na vrijwel identiek aan expressie X in voorbeeld 4.3.2. De uitwerking van die expressie komt ook daar vandaan.

$$\mathcal{V}[\![\mathsf{sv}(\mathrm{D}\circ\mathrm{h}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow})\circ W\circ Z]\!](\mathsf{UPop}) = \tag{4.14}$$

$$\mathcal{V}[\![\mathsf{sv}(\mathrm{D}\circ\mathrm{h}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow})\circ\mathsf{cfd}(\mathrm{D})]\!](\mathsf{UPop})\circ$$

$$\mathcal{V}[\![\mathsf{sv}(\mathrm{h}\circ\mathsf{sv}(\mathrm{f}\circ\mathrm{i}^{\leftarrow})\circ\mathsf{cfd}(\mathrm{T})\circ\mathrm{i}\circ\mathrm{f}\circ\mathrm{h}^{\leftarrow}\circ\mathrm{D}\circ\mathrm{j}\circ\mathrm{l}\circ\mathrm{k}^{\leftarrow})]\!](\mathsf{UPop})\circ$$

$$\mathcal{V}[\![\mathsf{cfd}(\mathrm{M})]\!](\mathsf{UPop})\circ\mathcal{V}[\![Z]\!](\mathsf{UPop}) =$$

$$\left\{\!\!\left[\begin{array}{l}(\langle d_1,d_1\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1\\(\langle d_2,d_2\rangle\,,\mathsf{seq}(1))\uparrow^1\\(\langle d_3,d_3\rangle\,,\mathsf{seq}(1))\uparrow^1\\(\langle d_4,d_4\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1\\(\langle d_5,d_5\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1\end{array}\right]\!\!\right\} \circ \left\{\!\!\left[\begin{array}{ll}(\langle d_1,m_1\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1, & (\langle d_4,m_1\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1,\\(\langle d_1,m_3\rangle\,,\mathsf{seq}(\frac{2}{3}))\uparrow^1, & (\langle d_4,m_2\rangle\,,\mathsf{seq}(\frac{1}{2}))\uparrow^1,\\(\langle d_1,m_5\rangle\,,\mathsf{seq}(2))\uparrow^4, & (\langle d_4,m_3\rangle\,,\mathsf{seq}(\frac{7}{6}))\uparrow^2,\\(\langle d_2,m_1\rangle\,,\mathsf{seq}(\frac{1}{6}))\uparrow^1, & (\langle d_4,m_5\rangle\,,\mathsf{seq}(1))\uparrow^2,\\(\langle d_2,m_3\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1, & (\langle d_5,m_1\rangle\,,\mathsf{seq}(\frac{1}{6}))\uparrow^1,\\(\langle d_2,m_5\rangle\,,\mathsf{seq}(\frac{1}{2}))\uparrow^2, & (\langle d_5,m_3\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1,\\(\langle d_3,m_2\rangle\,,\mathsf{seq}(\frac{1}{2}))\uparrow^1, & (\langle d_5,m_5\rangle\,,\mathsf{seq}(\frac{5}{2}))\uparrow^4,\\(\langle d_3,m_3\rangle\,,\mathsf{seq}(\frac{1}{2}))\uparrow^1 & \end{array}\right]\!\!\right\} \circ$$

$$\left\{\!\!\left[\begin{array}{l}(\langle m_1,m_1\rangle\,,\mathsf{seq}(1))\uparrow^1\\(\langle m_2,m_2\rangle\,,\mathsf{seq}(1))\uparrow^1\\(\langle m_3,m_3\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1\\(\langle m_5,m_5\rangle\,,\mathsf{seq}(\frac{1}{6}))\uparrow^1\end{array}\right]\!\!\right\} \circ \left\{\!\!\left[\begin{array}{l}(\langle m_1,q\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^1\\(\langle m_3,q\rangle\,,\mathsf{seq}(\frac{2}{3}))\uparrow^1\\(\langle m_5,q\rangle\,,\mathsf{seq}(2))\uparrow^4\end{array}\right]\!\!\right\} = \left\{\!\!\left[\begin{array}{c}(\langle d_1,q\rangle\,,\mathsf{seq}(\frac{25}{81}))\uparrow^{18}\\(\langle d_2,q\rangle\,,\mathsf{seq}(\frac{16}{54}))\uparrow^{10}\\(\langle d_3,q\rangle\,,\mathsf{seq}(\frac{1}{9}))\uparrow^1\\(\langle d_4,q\rangle\,,\mathsf{seq}(\frac{19}{81}))\uparrow^{11}\\(\langle d_5,q\rangle\,,\mathsf{seq}(\frac{1}{3}))\uparrow^{18}\end{array}\right]\!\!\right\}$$

Deze waarden in de verdelingen komen overeen met de uitkomsten van respectievelijk $\Psi(d_1 \rightarrow q)$, $\Psi(d_2 \rightarrow q)$, $\Psi(d_3 \rightarrow q)$, $\Psi(d_4 \rightarrow q)$, $\Psi(d_5 \rightarrow q)$ in de voorbeeldtoepassing in paragraaf 3.2.3.

4.3.6 Aspecten van het afleidingsmodel

In de voorgaande paragrafen is het gebruik van het afleidingsmodel in LISA-D gepresenteerd. Het is nu zinvol om wat details te bekijken.

4.3.6.1 "Automatisch" indexeren

Wong en Yao zijn bij het bepalen van de formules er vanuit gegaan dat er een indexeerproces met index termen als trefwoorden "geautomatiseerd" plaats vond. In de vergelijkingen worden van factoren als geïnverteerde document frequentie en term frequentie binnen documenten gebruik gemaakt. Het is de bedoeling dat deze factoren niet met de hand bijgehouden hoeft te worden als er termen worden toegevoegd of verwijderd.
Met de objecttypen in de informatiestructuur van het probabilistische afleidingsmodel zijn functies fd geassocieerd die van de inhoud van de multisets afhangen. Op hun beurt hangen deze multisets weer af van de onzekerheidspopulatie. Wat er in de onzekerheidspopulatie komt te staan, hangt af van de gebruiker die waarden aan instanties toekent. Bij het aanbrengen van indexen moeten die waarden door de gebruiker worden toegekend.
Een automatisch indexeerproces, dat wil zeggen dat een computer zelf indexen legt op documenten, is hier mogelijk. Hier zou dan wel de onzekerheidspopulatie in betrokken moeten worden. Het is hier niet van belang hoe dat gaat.

4.3.6.2 Sorteren naar maat van relevantie

Nu het mogelijk is om voor elk document de maat voor de relevantie te bepalen, kan er nu gekeken worden naar operaties op de resultaten. Een zo'n operatie is het sorteren naar maat van relevantie. In LISA-D is daarvoor een operatie die dit bewerkstelligt, genaamd *ordening*. In het volgende voorbeeld wordt de werking van die operatie getoond.

Voorbeeld 4.3.4
Op het niveau van padexpressie werkt de ordeningsoperatie ψ als volgt: Als P een padexpressie is zo dat

$$\mu[\![P]\!](\mathsf{Pop}) = \begin{array}{|c|c|}\hline d_1 & \frac{3}{9}\\ d_2 & \frac{3}{9}\\ d_4 & \frac{2}{9}\\ d_5 & \frac{3}{9}\\ \hline\end{array},$$

dan is

$$\mu[\![\psi\,(P,\geq)]\!]\,(\mathsf{Pop}) \quad = \quad \begin{array}{|c|c|}\hline \langle d_1,d_2,d_5,d_4\rangle & \langle d_1,d_2,d_5,d_4\rangle \\ \langle d_1,d_5,d_2,d_4\rangle & \langle d_1,d_5,d_2,d_4\rangle \\ \langle d_2,d_1,d_5,d_4\rangle & \langle d_2,d_1,d_5,d_4\rangle \\ \langle d_2,d_5,d_1,d_4\rangle & \langle d_2,d_5,d_1,d_4\rangle \\ \langle d_5,d_1,d_2,d_4\rangle & \langle d_5,d_1,d_2,d_4\rangle \\ \langle d_5,d_2,d_1,d_4\rangle & \langle d_5,d_2,d_1,d_4\rangle \\ \hline \end{array}$$

□

Men kan ongewild meerdere ordeningen krijgen zoals uit het voorbeeld blijkt. Dit verschijnsel kan vermeden worden door eerst de documenten te *groeperen* waarna dan pas de ordeningsoperatie gebruikt wordt.

Voorbeeld 4.3.5

Op het niveau van padexpressie is er een groepeeroperatie φ die als volgt werkt: als P en Q padexpressies zijn zo dat

$$\mu[\![P]\!]\,(\mathsf{Pop}) \;=\; \begin{array}{|c|c|}\hline d_1 & \frac{3}{9} \\ d_2 & \frac{3}{9} \\ d_4 & \frac{2}{9} \\ d_5 & \frac{3}{9} \\ \hline \end{array}\;, \quad \mu[\![Q]\!]\,(\mathsf{Pop}) \;=\; \begin{array}{|c|c|}\hline d_1 & q \\ d_2 & q \\ d_4 & q \\ d_5 & q \\ \hline \end{array}\;,$$

dan is

$$\mu[\![\varphi\,(Q,P)]\!]\,(\mathsf{Pop}) \quad = \quad \begin{array}{|c|c|}\hline \{d_1,d_2,d_5\} & \frac{3}{9} \\ \{d_4\} & \frac{2}{9} \\ \hline \end{array}\;.$$

Als dan dit resultaat wordt geordend, dan krijgt men het volgende:

$$\psi\left(\begin{array}{|c|c|}\hline \{d_1,d_2,d_5\} & \frac{3}{9} \\ \{d_4\} & \frac{2}{9} \\ \hline \end{array}\;,\;>\right) \quad = \quad \begin{array}{|c|c|}\hline \langle\{d_1,d_2,d_5\},\{d_4\}\rangle & \langle\{d_1,d_2,d_5\},\{d_4\}\rangle \\ \hline \end{array}\;.$$

Nu is er een eenduidige uitkomst. □

De te volgen strategie is eerst de resultaten naar relevantie te groeperen en dan te sorteren. In de voorbeelden is er gewerkt met de interpreter μ en zijn de gegevens voor gebruik met die interpreter als zodanig gegeven. In werkelijkheid moeten de volgende stappen ondernomen worden:

- De waarden uit de verdelingen halen en deze met een gedeelte van de tupel samenvoegen tot een nieuwe tupel;
- Nieuwe padexpressies beschrijven voor elementen met frequentieverdelingen erin op basis van de groepeer- en sorteerfuncties.

In het afleidingsmodel zijn slechts de relevanties van belang, niet de gehele verdelingen op zich. Deze waarden zijn nodig om te kunnen groeperen, en bij sorteren om arithmetische vergelijkingen ten behoeve van het sorteercriterium te kunnen maken.

Definitie 4.3.1

Zij Q een padexpressie. Dan zijn $\mathsf{Prob}\,Q$, $\mathsf{Pr}\,Q$ *en* $\mathsf{Val}\,Q$ *weer padexpressies waarvan de semantieken als volgt zijn gedefinieerd:*

name	expr	$\mathcal{V}[\![\text{expr}]\!](\mathsf{UPop})$
probability (of)	$\mathsf{Prob}\,Q$	$\{\!\![(\langle t[m], p\rangle, t)\uparrow^m \mid (\langle p, q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})]\!\!\}$
probability related	$\mathsf{Pr}\,Q$	$\mathcal{V}[\![(\mathsf{Prob}\,Q)^{\leftarrow}]\!](\mathsf{UPop})$
measures	$\mathsf{Val}\,Q$	$\mathcal{V}[\![\mathcal{f}\,(\mathsf{Prob}\,Q)]\!](\mathsf{UPop})$

□

Nu kunnen er nieuwe definities van de groepeer- en sorteeroperaties gegeven worden. Echter is het met frequentieverdelingen niet zo eenvoudig zoals uit het volgende voorbeeld blijkt.

Voorbeeld 4.3.6

Er wordt aangenomen dat er een voorlopige definitie van groeperen en sorteren is. Zij P en Q padexpressies die bij evaluatie de volgend uitkomsten leveren:

$$\mathcal{V}[\![P]\!](\mathsf{UPop}) = \begin{array}{|c|c|c|}\hline d_1 & q & v \\ d_4 & q & w \\ \hline\end{array} \quad , \quad \mathcal{V}[\![Q]\!](\mathsf{UPop}) = \begin{array}{|c|c|c|}\hline d_1 & d_1 & y \\ d_4 & d_4 & z \\ \hline\end{array}$$

Dan levert een groepeeroperatie het volgende op:

$$\mathcal{V}[\![\varphi(Q,P)]\!](\mathsf{UPop}) = \begin{array}{|c|c|c|}\hline \{d_1, d_4\} & q & x \\ \hline\end{array}$$

Wat moet hier die verdeling x zijn: v, w, y of z ? Of toch iets anders? Als q waarden uit de frequentieverdelingen w en v waren dan zou men voor x een van die verdelingen kunnen invullen. Dit gaat slechts goed als de verdelingen gelijk zijn, dat wil zeggen dat op elke positie in de verdelingen w en v dezelfde waarden staan. □

Het is kennelijk raadzaam om een neutraal verdeling te kiezen als men geen problemen met de groepeeroperatie wil. Hetzelfde geldt ook voor de sorteeroperatie.
Hieronder worden de algemene definities gegeven voor de genoemde operaties voor het werken met elementen die verdelingen bevatten.

Definitie 4.3.2

Zij P, G en S padexpressies. Dan zijn $\varphi(P,G)$, $\psi(P,S)$, $\Upsilon(P)$ en $\Xi(P)$ weer padexpressies waarvan de semantieken als volgt zijn gedefinieerd:

- Groeperen (group)*:*

$$\mathcal{V}[\![\varphi(P,G)]\!](\mathsf{UPop}) =$$

$$\mathsf{Multi}\left(\{(\langle K_g, g\rangle, 1_{\mathsf{SEQU}}) \mid g \in \mathsf{Elem}(\pi_2(\mathcal{V}[\![G]\!](\mathsf{UPop}))) \wedge K_g \neq \varnothing \wedge\}\right)$$

waarin:

$$K_g = \{x \in \mathsf{Elem}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))\mid \langle x, g\rangle \in \mathsf{Elem}(\mathcal{V}[\![G]\!](\mathsf{UPop}))\}$$

- Ontgroepen (ungroup)*:*

$$\mathcal{V}[\![\Upsilon(P)]\!](\mathsf{UPop}) = \mathsf{Sqr}(\biguplus(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))) \ ,$$

waarin

$$\biguplus(M) = \bigcup_x \{\![(x,v)\uparrow^m \mid (A,v) \in^m M \wedge x \in A]\!\}$$

- Ordening (ordering)*:*

$$\mathcal{V}[\![\psi(P,S)]\!](\mathsf{UPop}) = \mathsf{Sqr}(\{\!\![(s, 1_{\mathsf{SEQU}})\uparrow^1 \mid s \text{ is compatibel met } S \text{ over } P \text{ in } \mathsf{Pop}]\!\!\})$$

waarin voor s dat compatibel is met sorteercriterium S over P in de populatie Pop *moet gelden:*

1. *in s komen alle elementen in* $\mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop}))$ *voor en dan ook nog net zo vaak:* $\mathsf{Lin}(s) = \mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop}))$.
2. *de volgorde van de elementen in s is niet in strijd met de ordeningsregels in S:*

$$0 \leq i < j < |s| \quad \Rightarrow \quad \exists_{y_1,y_2} \left[\begin{array}{l} \langle s[i], y_1 \rangle \in \mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop})) \wedge \\ \langle s[j], y_2 \rangle \in \mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop})) \wedge \\ \langle y_2, y_1 \rangle \notin \mathsf{Elem}(\mathcal{V}[\![S]\!](\mathsf{UPop})) \end{array} \right]$$

De operatie Lin *maakt hier van rijtjes weer multisets:*
$\mathsf{Lin}(\langle x_1, \ldots, x_n \rangle) = \bigcup_{1 \leq i \leq n} \{[x[i]]\}$.

- Ontleding (unordering)*:*
zij $x \in \mathsf{Elem}(\mathcal{V}[\![P]\!](\mathsf{UPop}))$. *Dan is*

$$C_x = \left\{ [\![(x[i], v)\uparrow^n \;\middle|\; (x, v) \in^n \pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})) \wedge (1 \leq i \leq |x|)]\!] \right\})$$

waarvoor de volgende eigenschap geldt:

$$\mathsf{Elem}(C_x) = \mathsf{Lin}(x)\uparrow^n$$

Dan is de semantiek van $\Xi(P)$ *als volgt:*

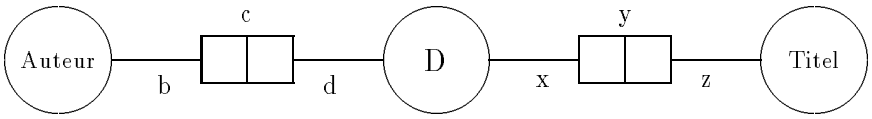
$$\mathcal{V}[\![\Xi(P)]\!](\mathsf{UPop}) = \mathsf{Sqr}\left(\bigcup_{x \in \mathsf{Elem}(\nu[\![P]\!](\mathsf{UPop}))} C_x \right)$$

□

Het overgrote deel van de definities is ontleend aan [HPW93].
Volledigheidshalve is naast sorteren en groeperen ook de omgekeerde versies gegeven. Het moet nu geen probleem meer zijn om documenten naar relevantie te groeperen en te sorteren.

4.3.6.3 Uitbreiden van mogelijkheden van retrieval

Tot nu toe zijn slechts documenten als zijnde archiefnummers op te vragen. Van een document zijn ook gegevens zoals soort, auteur(s), titel, jaar van uitgave en dergelijke bekend. Deze zaken zouden ook opgevraagd moeten kunnen worden. De informatiestructuur van het probabilistische afleidingsmodel moet dan wel even uitgebreid worden door aan documenten extra gegevens te koppelen zoals dat bijvoorbeeld in figuur 4.7 alvast gedaan is. Het vragen naar dergelijke specifieke informatie over de documenten kan problemen opleveren. Met behulp van het volgende voorbeeld wordt dit toegelicht.



/hfill

Figuur 4.7: Een voorbeeld informatiestructuur waarin nieuwe objecttypen aan D zijn gekoppeld.

Voorbeeld 4.3.7

In dit voorbeeld wordt gewerkt met de populatie die in paragraaf 4.3.1 is gegeven, en de bijbehorende informatiestructuur in figuur 4.6. We nemen aan dat de informatiestructuur is uitgebreid met een objecttype Auteur *en een feittype* c *dat de relatie tussen de objecttypen* D *en* Auteur *weergeeft (zie als voorbeeld figuur 4.7). Natuurlijk hoort ook de populatie uitgebreid te worden:*

$$\begin{array}{lcl} \mathsf{Pop}(\text{Auteur}) & = & \{a_1, a_2, a_3\} \\ \mathsf{Pop}(c) & = & \left\{ \begin{array}{lll} \{\text{b} : a_1, \text{d} : d_1\}, & \{\text{b} : a_1, \text{d} : d_2\}, & \{\text{b} : a_2, \text{d} : d_3\}, \\ \{\text{b} : a_1, \text{d} : d_4\}, & \{\text{b} : a_3, \text{d} : d_5\} & \end{array} \right\} \end{array}$$

Dit zijn weliswaar geen onzekerheidspopulaties, maar daarin is met elke instantie het neutrale element geassocieerd.
Zij Q een padexpressie zo dat

$$\mathcal{V}[\![Q]\!]\,(\mathsf{UPop}) \;=\; \left\{\!\!\left[\begin{array}{l} (\langle d_1, q\rangle\,, <\tfrac{5}{9}, \tfrac{5}{9}, \ldots >)\uparrow^2 \\ (\langle d_2, q\rangle\,, <\tfrac{3}{9}, \tfrac{3}{9}, \ldots >)\uparrow^1 \\ (\langle d_4, q\rangle\,, <\tfrac{2}{9}, \tfrac{2}{9}, \ldots >)\uparrow^1 \\ (\langle d_5, q\rangle\,, <\tfrac{3}{9}, \tfrac{3}{9}, \ldots >)\uparrow^2 \end{array}\right]\!\!\right\}.$$

Dan levert evaluatie van de padexpressie Auteur $\circ$ b $\circ$ c $\circ$ d$^{\leftarrow}$ $\circ$ Q *het volgende resultaat op:*

$$\mathcal{V}[\![\text{Auteur} \circ \text{b} \circ \text{c} \circ \text{d}^{\leftarrow} \circ Q]\!]\,(\mathsf{UPop}) \;=\; \left\{\!\!\left[\begin{array}{l} (\langle a_1, q\rangle\,, <\tfrac{10}{9}, \tfrac{10}{9}, \ldots >)\uparrow^4 \\ (\langle a_3, q\rangle\,, <\tfrac{3}{9}, \tfrac{3}{9}, \ldots >)\uparrow^2 \end{array}\right]\!\!\right\}.$$

In het eindresultaat zijn enkele elementen uit Q opgegaan in één element en is er een bijbehorend nieuwe verdeling gegeven □

In het eindresultaat in het voorbeeld zijn enkele elementen uit Q opgegaan in één element en is er een bijbehorend nieuwe verdeling gegeven. Men kan zich afvragen of dit de bedoeling is. Wong en Yao ([WY91]) reppen niet over dergelijke zaken, alleen maar over documenten. Wat dus een document inhoudt is hier de vraag. De inhoud van de verdeling dat met een document is geassocieerd, geeft de relevantie van het document gegeven een query aan. Deze verdeling hoort per document gehandhaafd te blijven, ongeacht of men vraagt naar namen van Auteurs, titels of iets dergelijks.
Er is een oplossing in de vorm van een padexpressie dat *confluentie* (*confluence*) heet. Deze expressie is beschreven in [HPW93] en er wordt hiervan een variant gegeven dat met verdelingen ook werkt.

Definitie 4.3.3

Zij $P_1, P_2, \ldots, P_n$ en Q padexpressies. Dan is $[P_1, \ldots, P_n \,|\, Q]$ weer een padexpressie waarvan de semantiek als volgt is gedefinieerd:

$$\mathcal{V}[\![[P_1, \ldots, P_n \,|\, Q]]\!]\,(\mathsf{UPop}) \;=$$

$$\bigcup_{(x,t)\in\pi_1(\mathcal{V}[\![Q]\!]\,(\mathsf{UPop}))} \left\{\!\!\left[(\langle\langle x_1, \ldots, x_n\rangle\,, x\rangle\,, t)\uparrow^{k_1\times\cdots\times k_n} \;\middle|\; \forall_{1\le i\le n}\left[\langle x_i, x\rangle \in^{k_i} \mathsf{Elem}(\mathcal{V}[\![p_i]\!]\,(\mathsf{UPop}))\right]\right]\!\!\right\}$$

Deze padexpressie wordt confluentie *(*confluence*) genoemd.* □

Voorbeeld 4.3.8

In het vorige voorbeeld (4.3.7) zijn padexpressies gebruikt die ook hier worden gebruikt. Zij Q als in dat voorbeeld en zij $P =$ Auteur $\circ$ b $\circ$ c $\circ$ d$^{\leftarrow}$ $\circ$ D. Dan is $[P \,|\, Q]$ weer een padexpressie dat bij evaluatie het volgende resultaat oplevert:

$$\mathcal{V}[\![[P \,|\, Q]]\!]\,(\mathsf{UPop}) \;=\; \left\{\!\!\left[\begin{array}{l} (\langle\langle a_1\rangle\,, d_1\rangle\,, <\tfrac{5}{9}, \tfrac{5}{9}, \ldots >)\uparrow^1 \\ (\langle\langle a_1\rangle\,, d_2\rangle\,, <\tfrac{3}{9}, \tfrac{3}{9}, \ldots >)\uparrow^1 \\ (\langle\langle a_1\rangle\,, d_4\rangle\,, <\tfrac{2}{9}, \tfrac{2}{9}, \ldots >)\uparrow^1 \\ (\langle\langle a_3\rangle\,, d_5\rangle\,, <\tfrac{3}{9}, \tfrac{3}{9}, \ldots >)\uparrow^1 \end{array}\right]\!\!\right\}.$$

De titels kunnen er ook in opgenomen worden. Men definieert daartoe een expressie R dat titels met documenten relateert en evalueert daarna de expressie $[P, R \,|\, Q]$. □

4.4 LISA-D en expertsystemen

In expertsystemen is kennis in de vorm van productieregels. De relatie tussen de regels laat zich grafische weergeven in een afleidingsnetwerk. De structuur is echter niet vast maar dynamisch.

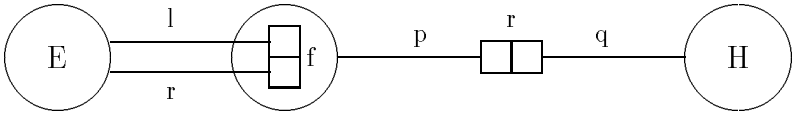
Deze kennis laat zich daarom wat moeilijk weergeven in een gegevensmodel. Een model is wel nodig wil men in LISA-D met onzekerheid werken, want de manier van associatie van een maat van onzekerheid aan een stukje kennis is van belang. Een afleidingsnetwerk geeft grafisch de samenhang tussen de productieregels weer. Het netwerk laat dan tevens gelijk zien in welke volgorde met welke functies de uiteindelijke maat voor de onzekerheid berekend moet worden. Dus de productieregels vormen tezamen grafisch gezien een architectuur voor het berekeningsproces. In een poging de kennis te modelleren zullen de productieregels bekeken worden. De combinatiefuncties hangen er nauw mee samen en zullen daarom ook bekeken worden. Deze analyse zou dan tot een te presenteren gegevensmodel moeten leiden. De (on-)mogelijkheden met dit model wordt daarna bediscussieerd, om zo op deze manier het basismodel waardoor men met LISA-D in staat is expertsystemen te ontwikkelen, te toetsen.
Er wordt in de voorbeelden die worden gebruikt om het een en ander duidelijk te maken van het *zekerheidsfactormodel* (*Certainty Factor model*) van Buchanan en Shortliffe ([BS84], [SB90]) gebruik gemaakt. Het model is eenvoudig en daarvan is de meeste informatie verkrijgbaar.

4.4.1 Productieregels

De kennisdatabase van een expertsysteem bevat productieregels van de vorm **if** e **then** h **fi**. Bij deze regels zijn enige aspecten te onderscheiden:

1. Er bestaan mogelijk relaties tussen de regels. Een hypothese h als conclusie in een regel kan deel uitmaken van de aanwijzing e in een andere regel;
2. Een aanwijzing in een regel kan samengesteld zijn. In principe kan een aanwijzing uit n deelaanwijzingen bestaan, met $1 \leq n \leq m$, $m \in \mathbb{N}$. Het aantal is eindig. Deze deelaanwijzingen worden met elkaar verbonden middels de operatoren **and** en **or**;
3. Meerdere regels kunnen één en dezelfde hypothese als conclusie hebben (co-concluderende regels).

Het eerste punt levert niet zoveel moeilijkheden op. Men stopt de aanwijzingen en hypothesen bij elkaar in een verzameling waaruit dan de elementen voor het maken van al dan niet samengestelde aanwijzingen kan worden betrokken.
De overige twee punten moeten kritisch beschouwd worden. Hoe kan bereikt worden dat als er een samengestelde aanwijzing is of co-concluderende regels zijn, dat precies de juiste combinatiefunctie voor het berekenen van de bijbeborende mate van de onzekerheid wordt gebruikt? In de toekomstige informatiestructuur waarin de regels zullen worden gemodelleerd moet de onderscheid duidelijk gemaakt zijn. Zie het hiernavolgende voorbeeld dat het probleem illustreert.



Figuur 4.8: De informatiestructuur waarin een productieregel expliciet is gemodelleerd

Voorbeeld 4.4.1

We beschouwen een simpel expertsysteem die in zijn kennisdatabase slechts enkele productieregels heeft. Zij deze regels als volgt:

R_1 : **if** a **and** b **then** h **fi**
R_2 : **if** f **or** g **then** l **fi**

*In figuur 4.8 is deze regel enigzins expliciet weergegeven, met een linker (*l*) en een rechter (*r*) argument van een propositie (*p*) dat een conclusie impliceert (*q*). Normaal werkt een expertsysteem met een grote hoeveelheid productieregels, echter om deze alle apart te modelleren*

gaat te ver. Er is tot nu toe nog geen geschikte representatie voor de regels. De populatie van deze informatiestructuur is in termen van LISA-D *als volgt:*

$$\begin{array}{lcl} \mathsf{Pop(E)} & = & \{a, b, f, g\} \\ \mathsf{Pop(H)} & = & \{h, l\} \\ \mathsf{Pop(f)} & = & \{\{\mathrm{l} : a, \mathrm{r} : b\}, \{\mathrm{l} : f, \mathrm{r} : g\}\} \\ \mathsf{Pop(r)} & = & \{\{\mathrm{p} : \{\mathrm{l} : a, \mathrm{r} : b\}, \mathrm{q} : h\}, \{\mathrm{p} : \{\mathrm{l} : f, \mathrm{r} : g\}, \mathrm{q} : l\}\} \end{array}$$

Stel de padexpressie P is $(\mathrm{l} \cup \mathrm{r}) \circ \mathrm{A} \circ \mathrm{p} \circ \mathrm{R} \circ \mathrm{q}^{\leftarrow}$*. Dan levert de evaluatie van deze padexpressie als resultaat:*

$$\mu[\![P]\!](\mathsf{Pop}) \;=\; \left\{\!\!\left[\begin{array}{c} \langle a, h\rangle \\ \langle b, h\rangle \\ \langle f, l\rangle \\ \langle g, l\rangle \end{array} \right]\!\!\right\}$$

Aan de regels is duidelijk te zien welke combinatiefuncties er bij elke regel gebruikt moet worden, maar dat is aan de multiset van tupels niet te zien. De uitkomst kan men ook als zodanig interpreteren dat men denkt met co-concluderende regels te maken te hebben. □

Aan het voorbeeld te zien is het kennelijk noodzakelijk om de co-concluderende regels expliciet te modelleren, dat wil zeggen dat er expliciet moet worden aangegeven dat de regels een en dezelfde hypothese als conclusie hebben. Verder moeten de regels ook zo opgebouwd worden dat daaruit precies blijkt of men met een operator **and** of **or** te maken heeft.

4.4.1.1 De manier van opbouwen van productieregels

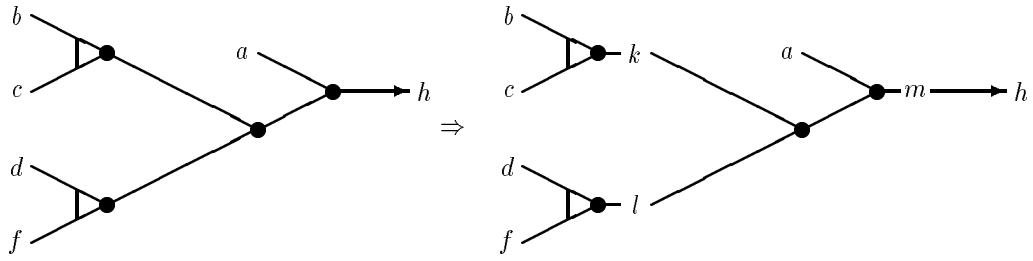
Het onderscheiden van operatoren **or** en **and** in samengestelde aanwijzingen en gemeenschappelijke hypothesen in regels is van belang. Hier wordt getoond hoe dit te bewerkstelligen is. Stel men heeft de volgende regels:

R_1 : **if** a **or** (b **and** c) **or** (b **and** c) **then** h **fi**
R_2 : **if** g **then** h **fi**

In regel R_1 wordt eerst onderscheid gemaakt tussen **and** en **or**. De aanwijzing wordt uiteengerafeld, en de deelaanwijzingen die men overhoudt krijgen een unieke naam. Aldus,

R_1' :
k = b **and** c
l = d **and** f
m = a **or** k **or** l
if m **then** h **fi**

Nu zijn de **or** en **and** duidelijk van elkaar gescheiden zodat nu op een 'simpele' manier na te gaan is welke combinatiefunctie voor welke component gebruikt moet worden. In de volgende afleidingsnetwerkjes is grafisch te zien hoe de regel eigenlijk omgevormd is.



Hoe zo'n regel ontleed dient te worden hangt af van de prioriteit van de operatoren of de haakjes die in de samengestelde aanwijzing voorkomt. Deze aspecten moeten uiteraard gerespecteerd worden. De samengestelde aanwijzingen a **or** (b **and** c) en (a **or** b) **and** c zijn logisch gezien niet hetzelfde.

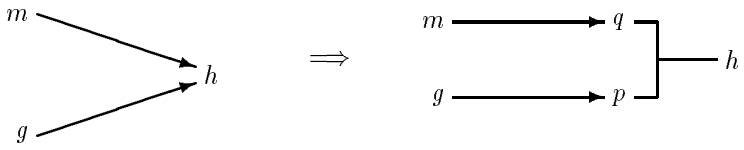
Nu moet nog het probleem van co-concluderende regels opgelost worden. De regels

$$\begin{array}{ll} R_1' : & \textbf{if } m \textbf{ then } h \textbf{ fi} \\ R_2' : & \textbf{if } g \textbf{ then } h \textbf{ fi} \end{array}$$

hebben een gemeenschappelijke hypothese als conclusie. Voordat daar iets aan gedaan wordt moet er rekening worden gehouden met de onzekerheid van de aanwijzingen die naar de hypothese h wordt gepropageerd. Er is een combinatiefunctie voor het propageren van aanwijzingen die de waarde die de onzekerheid van aanwijzingen weergeeft doorberekent naar de hypothesen. Deze informatie mag niet verloren gaan. Het is verstandig in elk regel de hypothese te hernoemen met een unieke naam en een nieuwe samenstelling van een aanwijzing in het leven te roepen dat weergeeft welke regels co-concluderend zijn. Aldus,

$$\begin{array}{ll} R_1'' : & \textbf{if } m \textbf{ then } q \textbf{ fi} \\ R_2' : & \textbf{if } g \textbf{ then } p \textbf{ fi} \\ R_3 : & h = p \textbf{ co } q \end{array}$$

De volgende afleidingsnetwerkjes geven deze omvorming grafisch weer.



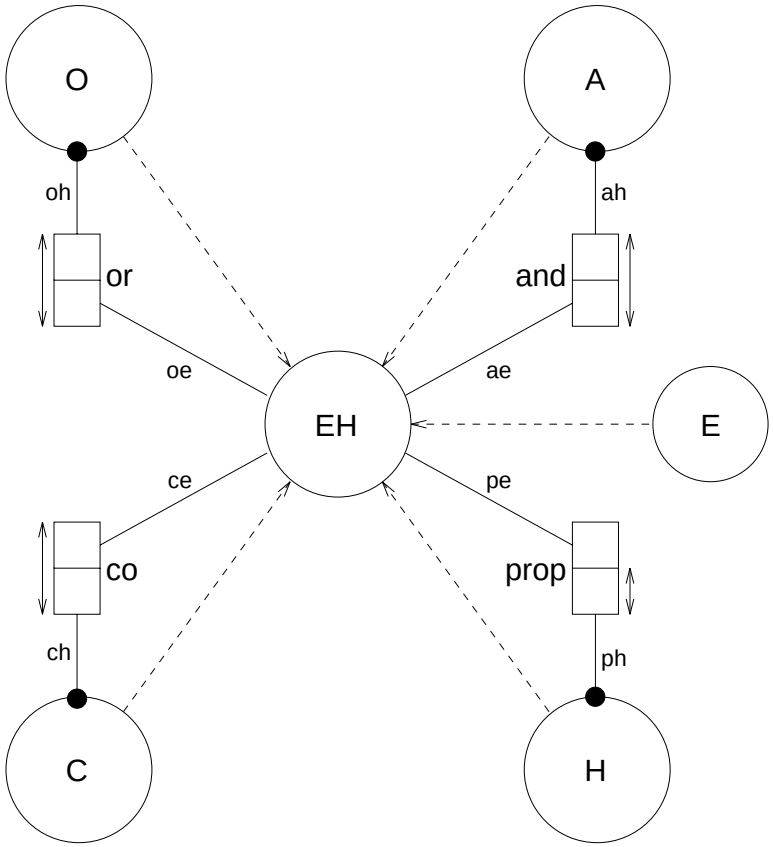
Hiet ziet er wat vreemd uit, maar dat heeft een reden. De extra hypothesen q en p zijn gemaakt om ervoor te zorgen dat de informatie bij propagatie ook verwerkt wordt. Aan de regels R_1' en R_2 kan een maat voor onzekerheid zijn toegekend. Dat moet ook ergens weer staan.
Hier zij trouwens opgemerkt dat de in het rechtse netwerk gebruikte grafische symbool voor de co-conclusie een eigen bedenksel is.

Een nadeel aan deze werkwijze is dat de directe relatie tussen een hypothese en de aanwijzing dat middels een regel gelegd is, verloren gaat. Door een toevoeging van een tussenstap zijn de hypothesen en aanwijzingen enigszins van elkaar gescheiden. Zo kan met behulp van een hypothese niet direct de bijbehorende aanwijzingen worden gevonden en omgekeerd. Dit is van wezenlijk belang omdat bij sommige experstsystemen zoals MYCIN ([BS84]) een geproduceerd resultaat verklaard kan worden.

4.4.1.2 De informatiestructuur

Nu in de vorige paragraaf de transformatie van de regels is beschreven, zijn we nu in staat de productieregels in een PSM-schema te modelleren. Er zal hier voor een goede inzicht een voorbeeld gegeven worden waarin de informatiestructuur wordt gepopuleerd, en hoe men uit de "aanwijzingen" de "hypothesen" kan krijgen.

Voorbeeld 4.4.2
De regels die hier worden verwerkt in de populatie, komen uit voorbeeld 2.1.1 op bladzijde



Figuur 4.9: De informatiestructuur waarin de verschillende componenten waaruit het afleidingsnetwerk kan bestaan is verwerkt.

7. De populatie van die regels in de informatiestructuur in figuur 4.9 is als volgt:

$$
\begin{array}{lcl}
\mathsf{Pop}(E) & = & \{a, c, f, g\} \\
\mathsf{Pop}(H) & = & \{h\} \\
\mathsf{Pop}(O) & = & \{i, m\} \\
\mathsf{Pop}(A) & = & \{k, p\} \\
\mathsf{Pop}(C) & = & \{b, d, j, l\} \\
\mathsf{Pop}(EH) & = & \{a, b, c, d, f, g, h, i, j, k, l, m, p\} \\
\mathsf{Pop}(\text{and}) & = & \{\{\text{ae}{:}\,d, \text{ah}{:}\,k\}, \{\text{ae}{:}\,f, \text{ah}{:}\,k\}, \{\text{ae}{:}\,a, \text{ah}{:}\,p\}, \{\text{ae}{:}\,i, \text{ah}{:}\,p\}\} \\
\mathsf{Pop}(\text{or}) & = & \{\{\text{oe}{:}\,g, \text{oh}{:}\,m\}, \{\text{oe}{:}\,f, \text{oh}{:}\,m\}, \{\text{oe}{:}\,b, \text{oh}{:}\,i\}, \{\text{oe}{:}\,c, \text{oh}{:}\,i\}\} \\
\mathsf{Pop}(\text{co}) & = & \{\{\text{ce}{:}\,j, \text{ch}{:}\,h\}, \{\text{ce}{:}\,l, \text{ch}{:}\,h\}\} \\
\mathsf{Pop}(\text{prop}) & = & \{\{\text{pe}{:}\,k, \text{ph}{:}\,b\}, \{\text{pe}{:}\,a, \text{ph}{:}\,d\}, \{\text{pe}{:}\,m, \text{ph}{:}\,l\}, \{\text{pe}{:}\,p, \text{ph}{:}\,j\}\}
\end{array}
$$

Men kan nu middels padexpressies allerlei informatie opvragen over de productieregels en diens structuur. Bijvoorbeeld:

- *Wat zijn de aanwijzingen waaruit met behulp van de operator* **and** *samengestelde aanwijzingen zijn gemaakt:* $P = ae \circ ah^{\leftarrow}$
- *Wat zijn de aanwijzingen waaruit met behulp van de* **or***-operator samengestelde aanwijzingen zijn gemaakt:* $R = oe \circ oh^{\leftarrow}$
- *Welke aanwijzingen leiden tot welke hypothesen:* $S = pe \circ ph^{\leftarrow}$
- *Welke regels zijn co-concluderend:* $T = ce \circ ch^{\leftarrow}$

Naast deze vragen kan men nog moeilijker vragen stellen, zoals de vraag tot welke hypothesen men komt als men slechts de elementaire aanwijzingen heeft, ofwel in de vorm van een padexpressie:

$$X = E \circ (P \cup R \cup S \cup T)^{+} \circ H$$

Een evaluatie van deze padexpressie levert als resultaat:

$$\mu[\![X]\!](\mathsf{Pop}) \;=\; \begin{array}{|c|c|}\hline \mathrm{a} & \mathrm{h} \\ \mathrm{c} & \mathrm{h} \\ \mathrm{f} & \mathrm{h} \\ \mathrm{g} & \mathrm{h} \\ \hline \end{array}$$

□

In het voorbeeld is om de gegevens er uit te halen gebruik gemaakt van een padexpressie dat *afsluiting* (Engels: *closure*) heet. Hieronder wordt de formele omschrijving van deze padexpressie gegeven zoals dat in [HPW93] staat:

name	expr	$\mu[\![\mathsf{expr}]\!](\mathsf{UPop})$
closure	P^+	$\mu[\![\mathsf{ds}(\bigcup_{n\in\mathbb{N}} \mathsf{closure}(n,P))]\!](\mathsf{Pop})$

waarin

$$\begin{aligned} \mathsf{closure}(0,P) &= P\ , \\ \mathsf{closure}(n+1,P) &= \mathsf{closure}(n,P)\circ P\ . \end{aligned}$$

Deze padexpressie lijkt wel bruikbaar te zijn, waarover later meer.

4.4.2 Toepassing van het zekerheidsfactormodel

Het is nu de bedoeling het zekerheidsfactormodel toe te passen. In de voorgaande paragrafen is beschreven hoe de productieregels worden gepopuleerd. Er zal geprobeerd worden met behulp van die regels zekerheidsfactoren voor de hypothesen te bepalen. Voor het zover is, moet nog het volgende gedaan worden:

- invullingen van onzekerheidspopulaties geven;
- invullingen van frequentieverdelingen beschrijven en met betrekking tot die verdelingen operaties die in padexpressies voorkomen, specificeren;
- de combinatiefuncties inbrengen.

4.4.2.1 De onzekerheidspopulatie

In de onzekerheidspopulaties die worden gespecificeerd, moet aan elementen die **geen** a priori aanwijzingen zijn, een waarde toegekend worden waaruit dat duidelijk blijkt. Dan is er via de frequentieverdeling ook een middel om na te gaan of voor dat element een zekerheidsfactor berekend is of niet.
Verder moet er aan de elementen die a priori aanwijzingen voorstellen, een waarde kunnen worden toegekend dat uitdrukt dat de waarheid of onwaarheid van die elementen niet vast te stellen is. Buchanan en Shortliffe hielden in hun model rekening met de mogelijkheid van ontbrekende informatie.
In het zekerheidsfactormodel ([BS84]) neemt een zekerheidsfactor de volgende waarden aan:
$-1 \leq \mathrm{CF} \leq 1$. De bovengenoemde waarden mogen niet in het bereik van de zekerheidsfactor liggen.

Voor het neutrale element wordt de volgende waarde gekozen:

$$\xi_u \;=\; -5\ .$$

Over de onzekerheidspopulaties kan nu de volgende belangrijke opmerking gemaakt worden.

Propositie 4.4.1 Volgens de definitie van de onzekerheidspopulatie is aan de instanties van alle objecttypen initieel de waarde $\xi_{\mathcal{U}}$ toegekend. Voor alle objecttypen met uitzondering van objecttypen E en prop moet dit worden gehandhaafd. Dit geldt ook voor uitbreidingen van de informatiestructuur (figuur 4.9).

De onzekerheidspopulatie van de objecttype E bevat instanties die a priori aanwijzingen voorstellen. Aan deze instanties moet initieel een andere waarde toegekend worden dan $\xi_{\mathcal{U}}$. Initieel kan er over de waarheid of onwaarheid van de aanwijzingen nog niets gezegd worden. De definitie van het lokale standaardelement biedt daartoe de mogelijkheid. We kiezen hiervoor een waarde:

$$\xi(\mathrm{E}) = -3 \ .$$

De gebruiker kan hier later alsnog de zekerheidsfactoren invullen.
De gebruiker kan ook aan de hypothese als conclusie van een aanwijzing een zekerheidsfactor toekennen. Deze waarde zegt iets over de betrouwbaarheid van de hypothese in het geval de aanwijzing absoluut waar is. De populatie van het feittype prop bevat elementen die de regels als unieke relaties tussen hypothesen en aanwijzingen voorstellen. De gebruiker kan daar aan die elementen de zekerheidsfactoren toekennen. In het geval dat dat niet gebeurt, worden de regels als absoluut betrouwbaar beschouwd.

4.4.2.2 Beschrijving van frequentieverdelingen en operaties daarop

Deze paragraaf gaat over de invulling van de frequentieverdeling. Verder zijn er nog padexpressies waarvan de werking met betrekking tot de verdelingen nog gespecificeerd moet worden.
Voor elk hypothese moet er een zekerheidsfactor bepaald kunnen worden.

Propositie 4.4.2 Het is de bedoeling dat voor een hypothese precies één zekerheidsfactor wordt berekend. Als zo'n zekerheidsfactor bekend is, moet dat het enige en juiste getal zijn. Als bijvoorbeeld $\{[(\langle a,h\rangle,d)\uparrow^1,(\langle b,h\rangle,f)\uparrow^1]\} \in M$ met M een multiset en h is een hypothese, dan geldt: $d = f$

Ongeacht de frequentie van voorkomen van een element in een multiset dat de lokatie van de zekerheidsfactor in de frequentieverdeling bepaalt, ongeacht waar dat getal staat moet steeds dezelfde zekerheidsfactor uit de verdeling te halen zijn. Dan kan op elke lokatie dezelfde waarde worden ingevuld.
Hieronder volgt een overzicht van hoe de verdelingen er uit horen te zien.

- De neutrale verdeling: $1_{\mathsf{SEQU}} = <\xi_{\mathcal{U}},\ldots>$
- Objecttypen: als $(y,d) \in \mathsf{UPop}(x)$, dan is $\mathsf{seq}(d) = <d,d,d,\ldots>$
- Concatenatie van padexpressies: als $(\langle p,r\rangle,s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})$ en $(\langle r,q\rangle,t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})$, dan is

$$s \odot t = \begin{cases} t, & \text{als } s = 1_{\mathsf{SEQU}} \\ s, & \text{als } t = 1_{\mathsf{SEQU}} \\ <s[0]*t[0], s[1]*t[1], \ldots, s[n\times m]*t[n\times m], \ldots>, & \text{anders} \end{cases}$$

  Als de tupel $\langle p,q\rangle$ op meerdere manieren kan worden gevormd, dan verandert er niets aan de verdeling (propositie 4.4.2). Zie de beschrijving van de vereniging.

- Vereniging van padexpressies: als $(\langle p,q\rangle,s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})$ en $(\langle p,q\rangle,t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})$, dan is

$$s \cup t = \begin{cases} t \ , \text{als } s = 1_{\mathsf{SEQU}} \\ s \ , \text{als } t = 1_{\mathsf{SEQU}} \\ t \ , \text{anders (aanname: } s = t) \end{cases}$$

  voor $(\langle p,q\rangle,x) \in^{n+m} \mathcal{V}[\![P \cup Q]\!](\mathsf{UPop})$.

Tot zover zijn dit de belangrijkste omschrijvingen. Daarmee is een verzameling padexpressies voor het zekerheidsfactormodel semantisch volledig gedefinieerd. Hier wordt even een overzicht van de bedoelde padexpressies gegeven.

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
empty path	$\varnothing_{\mathcal{PE}}$	$\varnothing$
neutral path	$1_{\mathcal{PE}}$	$1_{\Omega \times \Omega \times \mathsf{SEQU}}$
constant	c	$\mathsf{Sqr}(\{[(c, 1_{\mathsf{SEQU}})]\})$
multiset	X	$\mathsf{Sqr}(X)$
objecttype	x	$\{\!\![(\langle y, y\rangle, \mathsf{seq}(d))\uparrow^1 \mid (y,d) \in \mathsf{UPop}(x)]\!\!\}$
predicator	p	$\{\!\![(\langle v(p), v\rangle, 1_{\mathsf{SEQU}})\uparrow^1 \mid v \in \mathsf{Pop} \circ \mathsf{Fact}(p)]\!\!\}$
reverse	$P^{\leftarrow}$	$\{\!\![(\langle q, p\rangle, s)\uparrow^n \mid (\langle p, q\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})]\!\!\}$
concatenate	$P \circ Q$	$\bigcup_r \left\{\!\!\left[(\langle p, q\rangle, s \odot t)\uparrow^{n \times m} \;\middle|\; \begin{array}{c}(\langle p, r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge \\ (\langle r, q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})\end{array}\right]\!\!\right\}$
union	$P \cup Q$	$\left\{\!\!\left[(\langle p, q\rangle, s \cup t)\uparrow^{n+m} \;\middle|\; \begin{array}{c}(\langle p, q\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge \\ (\langle p, q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})\end{array}\right]\!\!\right\}$
distinct	$\mathsf{ds}\ P$	$\mathsf{Multi}(\mathsf{set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$
front	$\not\vdash P$	$\mathsf{Sqr}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

4.4.2.3 De combinatiefuncties

In deze paragraaf wordt beschreven hoe de combinatiefuncties van het zekerheidsfactormodel operationeel worden gemaakt. Er zijn vier combinatiefuncties, te weten:

- een functie voor het propageren van aanwijzingen;
- twee functies voor samengestelde aanwijzingen (**or** en **and**);
- een functie voor co-concluderende regels.

Deze functies worden middels de functie fd geïmplementeerd. De specificatie van de functies hangt af van de padexpressies en hoe ze worden geëvalueerd. Om de zekerheidsfactoren te bepalen, begint men bij de a priori aanwijzingen en propageert die naar de eindconclusie toe. Vertaald naar padexpressies begint men bij de aanwijzingen en eindigt men bij de hypothesen. Dit is niet verplicht. Men kan ook bij de hypothesen beginnen, maar het rekenproces is dan achterwaarts. We maken hier een keus.

Propositie 4.4.3 Een padexpressie wordt hier **van links naar rechts** geëvalueerd, te beginnen bij de aanwijzingen. Als $P \circ Q$ een padexpressie is, dan wordt eerst P geëvalueerd en dan ook van links naar rechts.

Deze aanname zal gelden voor het hele model. Om de problematiek met de functie fd te verduidelijken wordt er hieronder twee padexpressies gegeven.

- Zij $X = \mathsf{sv}(\mathrm{EH} \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}$. De informatiestroom is hier van links naar rechts. De functie fd die van STORAGE gebruik maakt, kijkt in de standaard multiset naar de rechterkant van de tupels $\langle a, b\rangle$.
- Zij $X = \mathrm{H} \circ \mathrm{ph} \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathsf{sv}(\mathrm{pe}^{\leftarrow} \circ \mathrm{EH})$. Hier is de informatiestroom van rechts naar links. De functie fd die van STORAGE gebruik maakt, kijkt in de standaard multiset naar de linkerkant van de tupels $\langle a, b\rangle$.

Als in een padexpressie een operatie sv is, dan wordt er gegevens in de standaard multiset geplaatst. Er moet alleen nog worden aangegeven hoe dat gebeurt. Er zijn twee mogelijkheden. Als $\mathsf{sv}(P)$ een padexpressie is, dan:

1. $\mathsf{STORAGE} = \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$.
   Dit is een eenvoudige oplossing. Het is alleen de vraag of men niet te weinig gegevens heeft;
2. $\mathsf{STORAGE} = (\mathsf{STORAGE} - \mathcal{V}[\![P]\!](\mathsf{UPop})) \cup \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![P]\!](\mathsf{UPop})))$.
   Deze ziet er wat ingewikkelder uit. Wat hier gewoon gebeurt, is oude gegevens verwijderen en nieuwe gegevens ervoor in de plaats zetten. Hier kunnen er dit keer wat meer gegevens in zitten. Het nadeel is dat de standaard multiset na een volledig geëvalueerde padexpressie weer moet worden geleegd ($\mathsf{STORAGE} = \varnothing_{\mathcal{PS}}$).

Welke de beste is, moet nog blijken.

4.4.2.3.1 Propagatie

De functie fd waarin de propagatiefunctie wordt gespecificeerd, wordt gekoppeld aan het feittype prop (zie figuur 4.9). In de onzekerheidspopulatie van het feittype kunnen door de gebruiker aan de instanties zekerheidsfactoren worden toegekend. Bij de specificatie van de functie fd moet daar rekening mee worden gehouden.

Stel men heeft de productieregel **if** a **then** b **fi**, met als bijbehorende zekerheidsfactor $\mathrm{CF}(b,a) = y$. De aanwijzing a kan ook onzeker zijn. Wat de zekerheidsfactor voor a is, moet worden nagegaan. In de onderstaande tabel worden de mogelijke waarden aangegeven. Voor hypothese b wordt gelijk aangegeven wat dan de netto zekerheidsfactor is.

	Mogelijke waarden in verdeling			
	nog niet bekend	kan niet bepalen	zekerheidsfactor	
Aanwijzing a	$\xi_{\mathcal{U}}$	$\xi(\mathrm{E})$	$-1 \le \mathrm{CF} < 0$	$0 \le \mathrm{CF} \le 1$
Hypothese b	$\frac{\xi_{\mathcal{U}}}{y} \cdot y$	$\frac{\xi(\mathrm{E})}{y} \cdot y$	0	$\mathrm{CF} \cdot y$

De eerste twee waarden voor hypothese b zien er wat komisch uit, maar het heeft een reden. $\xi_{\mathcal{U}}$ is een neutraal element. Elk andere waarde wordt bij concatenatie met het neutrale element ongehinderd doorgegeven. Wat de waarde $\xi(\mathrm{E})$ betreft: deze geeft aan dat de waarheid niet vastgesteld kan worden. Dan moet $\mathrm{CF}(b,a)$ geneutraliseerd worden.
Nu kan de functie fd gespecificeerd worden. Zij $\langle x,x\rangle$ de tupel waaraan een verdeling moet worden toegekend. Zij tevens $(x,y) \in \mathsf{UPop}(\mathrm{prop})$.

$$\mathrm{fd}(\mathrm{prop}) = \begin{cases} \mathsf{seq}\left(\frac{1}{y}\right) & , \text{als } v = <\xi(\mathrm{E}),\xi(\mathrm{E}),\ldots> \wedge (x,v) \in \pi_2(\mathsf{STORAGE}) \\ \mathsf{seq}(\max(0,v[n])) & , \text{als } \begin{pmatrix} (x,v) \in^n \pi_2(\mathsf{STORAGE}) \wedge \\ v \neq <\xi(\mathrm{E}),\xi(\mathrm{E}),\ldots> \wedge \\ v \neq 1_{\mathsf{SEQU}} \end{pmatrix} \\ \mathsf{seq}\left(\frac{\xi_{\mathcal{U}}}{y}\right) & , \text{anders} \end{cases} \tag{4.15}$$

De padexpressie dat hierbij hoort om de juiste gegevens in de standaard multiset te krijgen en om er achter komen wat de waarde voor de hypothese is, is als volgt:

$$\mathsf{sv}(\mathrm{EH} \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}$$

4.4.2.3.2 Samengestelde aanwijzingen: de or–operatie

De functie fd waarin de combinatiefunctie voor samengestelde aanwijzingen wordt gespecificeerd, wordt gekoppeld aan het objecttype O (zie figuur 4.9). De onzekerheidspopulatie van dat objecttype bevat slechts elementen waarin met de instanties het neutrale element $\xi_{\mathcal{U}}$ is geassocieerd. De

gebruiker kan en mag daar geen andere waarden toekennen, in verband met de overerving van de gegevens. Bij de specificatie van de functie fd hoeft daar geen rekening mee gehouden te worden.

Stel men heeft de productieregel **if** a **or** b **then** h **fi**. Dan moet eerst de zekerheidsfactor voor a **or** b bepaald worden. Het onderstaande tabel geeft de mogelijkheden.

b	a		
	$\xi_{\mathcal{U}}$	ξ(E)	CF(a,e)
$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$
ξ(E)	$\xi_{\mathcal{U}}$	ξ(E)	CF(a,e)
CF(b,e)	$\xi_{\mathcal{U}}$	CF(b,e)	max(CF(b,e),CF(a,e))

In het geval dat voor een van de aanwijzingen a of b de zekerheidsfactor de waarde ξ(E) heeft, dan kan men gewoon de maximum nemen als aan de de volgende **voorwaarde** voldaan is: $\xi(\mathrm{E}) < -1$. -1 is de ondergrens van de "echte" zekerheidsfactoren. Voor alle zekerheid kan men het beste eerst die ξ(E) er uit filteren, voordat men het maximum bepaalt. In de onderstaande specificatie wordt dat laatste toegepast.
Zij k de naam van de samengestelde aanwijzing a **or** b. Dan is $k \in \mathsf{Pop}(\mathrm{O})$. De specificatie van de functie fd(O) is als volgt:

$$\mathrm{fd(O)} \quad = \quad \begin{cases} \mathsf{seq}\left(\max\left\{ v[n] \;\middle|\; \begin{array}{c}(\langle x,k\rangle, v) \in^n \mathsf{STORAGE} \wedge \\ v \neq <\xi(\mathrm{E}),\xi(\mathrm{E}),\ldots>\end{array} \right\}\right), \\ \qquad\qquad \text{als } \left(\begin{array}{l}(k,v) \in \pi_2(\mathsf{STORAGE}) \wedge \\ \forall_{(\langle x,k\rangle,t)\in\mathsf{STORAGE}}[t \neq 1_{\mathsf{SEQU}}]\end{array}\right) \\ 1_{\mathsf{SEQU}}, \\ \qquad\qquad \text{anders} \end{cases} \tag{4.16}$$

De padexpressie dat hierbij hoort om de juiste gegevens in de standaard multiset te krijgen en om er achter komen wat de waarde voor zo'n samengestelde aanwijzing is, is als volgt:

$$(\mathsf{sv}(\mathrm{EH} \circ \mathrm{oe} \circ \mathrm{or} \circ \mathrm{oh}^{\leftarrow}) \vdash \mathsf{cfd}(\mathrm{O}))$$

4.4.2.3.3 Samengestelde aanwijzingen: de and–operatie

De functie fd waarin de combinatiefunctie voor samengestelde aanwijzingen wordt gespecificeerd, wordt gekoppeld aan het objecttype A (zie figuur 4.9). De onzekerheidspopulatie van dat objecttype bevat slechts elementen waarin met de instanties het neutrale element $\xi_{\mathcal{U}}$ is geassocieerd. De gebruiker kan en mag daar geen andere waarden toekennen, in verband met de overerving van de gegevens. Bij de specificatie van de functie fd hoeft daar geen rekening mee gehouden te worden.

Stel men heeft de productieregel **if** a **and** b **then** h **fi**. Dan moet eerst de zekerheidsfactor voor a **and** b bepaald worden. Het onderstaande tabel geeft de mogelijkheden.

b	a		
	$\xi_{\mathcal{U}}$	ξ(E)	CF(a,e)
$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$	ξ(E)	$\xi_{\mathcal{U}}$
ξ(E)	ξ(E)	ξ(E)	ξ(E)
CF(b,e)	$\xi_{\mathcal{U}}$	ξ(E)	min(CF(b,e),CF(a,e))

In het geval dat voor een van de aan wijzingen a of b de zekerheidsfactor de waarde ξ(E) heeft, dan kan men gewoon het minimum nemen als aan de de volgende **voorwaarde** voldaan is: $\xi(\mathrm{E}) < -1$. -1 is de ondergrens van de "echte" zekerheidsfactoren. Voor alle zekerheid kan men het beste eerst op de waarde ξ(E) testen, voordat men het minimum neemt. In de onderstaande specificatie wordt dat laatste toegepast.
Zij k de naam van de samengestelde aanwijzing a **and** b. Dan is $k \in \mathsf{Pop}(\mathrm{A})$. De specificatie van

de functie fd(A) is als volgt:

$$\text{fd(A)} \quad = \quad \begin{cases} < \xi(\text{E}), \xi(\text{E}), \ldots > , \\ \qquad\qquad \text{als} \left(\begin{array}{l} (k,v) \in \pi_2\,(\mathsf{STORAGE}) \wedge \\ \exists_{(\langle y,k\rangle,t)\in \mathsf{STORAGE}} [t = < \xi(\text{E}), \xi(\text{E}), \ldots >] \end{array} \right) \\ \mathsf{seq}\left(\min\left\{ v[n] \;\middle|\; (\langle x,k\rangle, v) \in^n \mathsf{STORAGE} \right\}\right) , \\ \qquad\qquad \text{als} \left(\begin{array}{l} (k,v) \in \pi_2\,(\mathsf{STORAGE}) \wedge \\ \neg\exists_{(\langle y,k\rangle,t)\in \mathsf{STORAGE}} [t = < \xi(\text{E}), \xi(\text{E}), \ldots > \vee t = 1_{\mathsf{SEQU}}] \end{array} \right) \\ 1_{\mathsf{SEQU}} , \\ \qquad\qquad \text{anders} \end{cases} \tag{4.17}$$

De padexpressie dat hierbij hoort om de juiste gegevens in de standaard multiset te krijgen en om er achter komen wat de waarde voor zo'n samengestelde aanwijzing is, is als volgt:

$$(\mathsf{sv}(\text{EH} \circ \text{ae} \circ \text{and} \circ \text{ah}^{\smile}) \vdash \mathsf{cfd}(\text{A}))$$

4.4.2.3.4 Co-concluderende regels

De functie fd waarin de combinatiefunctie voor samengestelde aanwijzingen wordt gespecificeerd, wordt gekoppeld aan het objecttype C (zie figuur 4.9). De onzekerheidspopulatie van dat objecttype bevat slechts elementen waarin met de instanties het neutrale element $\xi_{\mathcal{U}}$ is geassocieerd. De gebruiker kan en mag daar geen andere waarden toekennen, in verband met de overerving van de gegevens. Bij de specificatie van de functie fd hoeft daar geen rekening mee gehouden te worden.

Stel men heeft de productieregels **if** a **then** h **fi** en **if** b **then** h **fi**. Deze regels zijn co-concluderend en dus moet uit de afzonderlijke zekerheidsfactoren een nieuwe zekerheidsfactor voor hypothese h worden bepaald. We doen hier even voor het gemak alsof h de naam is voor $a\,\mathbf{co}\,b$. In de onderstaande tabel staan de mogelijke waarden voor h.

b \ a	$\xi_{\mathcal{U}}$	$\xi(\text{E})$	$\text{CF}(a,e)$
$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$	$\xi_{\mathcal{U}}$
$\xi(\text{E})$	$\xi_{\mathcal{U}}$	$\xi(\text{E})$	$\text{CF}(a,e)$
$\text{CF}(b,e)$	$\xi_{\mathcal{U}}$	$\text{CF}(b,e)$	$\text{CF}(h, b\,\mathbf{co}\,a)$

De invulling voor $\text{CF}(h, b\,\mathbf{co}\,a)$ is echter te ingewikkeld om hier volledig in de tabel te beschrijven. Voor de omschrijving wordt verwezen naar het hoofdstuk over expert systemen waarin een paragraaf over het zekerheidsfactormodel is opgenomen.
Als van een van de aanwijzingen a of b de waarheid niet vast te stellen is (bijvoorbeeld $\text{CF}(a,e) = \xi(\text{E})$), dan telt deze gewoon niet mee in het eindresultaat ($\text{CF}(h, b\,\mathbf{co}\,a) = \text{CF}(b,e)$).
Zij h de naam van a **or** b. Dan is $h \in \mathsf{Pop}(\text{C})$. De specificatie van de functie fd(C) is als volgt:

$$\text{fd(C)} \quad = \quad \begin{cases} < \xi(\text{E}), \xi(\text{E}), \ldots > , \\ \qquad\qquad \text{als} \left(\begin{array}{l} (h,v) \in \pi_2\,(\mathsf{STORAGE}) \wedge \\ \forall_{(\langle x,h\rangle,t)\in \mathsf{STORAGE}} [t = < \xi(\text{E}), \xi(\text{E}), \ldots >] \end{array} \right) \\ \mathsf{seq}\left(\text{co}\left(0, \left\{ v[n] \;\middle|\; \begin{array}{l} (\langle x,h\rangle, v) \in^n \mathsf{STORAGE} \wedge \\ v \neq < \xi(\text{E}), \xi(\text{E}), \ldots > \end{array} \right\}\right)\right) , \\ \qquad\qquad \text{als} \left(\begin{array}{l} (h,v) \in \pi_2\,(\mathsf{STORAGE}) \wedge \\ \neg\forall_{(\langle y,h\rangle,t)\in \mathsf{STORAGE}} [t = < \xi(\text{E}), \xi(\text{E}), \ldots >] \wedge \\ \neg\exists_{(\langle y,h\rangle,t)\in \mathsf{STORAGE}} [t = 1_{\mathsf{SEQU}}] \end{array} \right) \\ 1_{\mathsf{SEQU}} , \\ \qquad\qquad \text{anders} \end{cases} \tag{4.18}$$

De functie co wordt hieronder beschreven op een manier dat bij transformationeel programmeren ([Par90]) gehanteerd wordt.

mode $\mathcal{U}$ = **real**;
mode uset = **set**($\mathcal{U}$,=);

type co = **uset**, co:
sort uset
funct ($\mathcal{U}$ i, **uset** R) $\mathcal{U}$ co;

laws $\mathcal{U}$ i, y, **uset** R:
$co(i, \varnothing)$ $\equiv$ i ,
$co(i, R \cup \{y\})$ $\equiv$ **if** $i > 0 \wedge y > 0$
then $co(i + y * (i - 1), R)$
elsif $i < 0 \wedge y < 0$
then $co(i + y * (i - 1), R)$
else $co\left(\frac{i + y}{1 - \min(|i|, |y|)}, R\right)$
fi

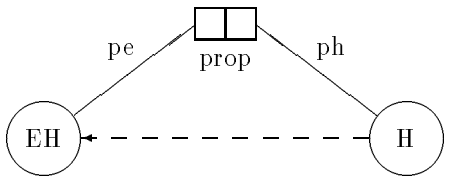
endoftype

De padexpressie dat hierbij hoort om de juiste gegevens in de standaard multiset te krijgen en om er achter komen wat de waarde voor de hypothese is, is als volgt:

$$(\mathsf{sv}(\mathrm{EH} \circ \mathrm{ce} \circ \mathrm{co} \circ \mathrm{ch}^{\leftarrow}) \vdash \mathsf{cfd}(\mathrm{C}))$$

4.4.2.4 Berekening van zekerheidsfactoren voor hypothesen

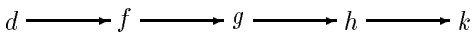
Zoals in de paragraaf over de informatiestructuur enigzins gedemonstreerd is, is om de a priori aanwijzingen naar hypothesen te "propageren" een padexpressie gebruikt dat afsluiting heet. Er moet hiervan nog een versie komen, met frequentieverdelingen erin. Echter, met frequentieverdelingen is het niet zo triviaal als het op het eerste gezicht lijkt. In de volgende paragraaf wordt de problematiek in het geval van propagatie bekeken, dat representatief is voor de overige soorten regels.



Figuur 4.10: Een informatiestructuur dat propagatie van aanwijzingen weergeeft.

4.4.2.4.1 Propagatie

Veronderstel dat er slechts productieregels zijn waar alleen propagatie in voorkomt (dus geen samengestelde aanwijzingen en geen co-concluderende regels). Neem zo'n reeks regels zoals in het afleidingsnetwerk hieronder.



Deze regels worden gepopuleerd in de informatiestructuur in figuur 4.10. De onzekerheidspopulaties zien er als volgt uit:

$$\begin{array}{lcl} \mathsf{UPop}(\mathrm{E}) & = & \{(d, 0.5)\} \\ \mathsf{UPop}(\mathrm{H}) & = & \{(f, \xi_{\mathcal{U}}), (g, \xi_{\mathcal{U}}), (h, \xi_{\mathcal{U}}), (k, \xi_{\mathcal{U}})\} \\ \mathsf{UPop}(\mathrm{EH}) & = & \{(f, \xi_{\mathcal{U}}), (g, \xi_{\mathcal{U}}), (h, \xi_{\mathcal{U}}), (k, \xi_{\mathcal{U}}), (d, 0.5)\} \\ \mathsf{UPop}(\mathrm{prop}) & = & \left\{ \begin{array}{l} (\{\mathrm{pe}{:}\, d, \mathrm{ph}{:}\, f\}, 1.0), (\{\mathrm{pe}{:}\, f, \mathrm{ph}{:}\, g\}, 0.8), \\ (\{\mathrm{pe}{:}\, g, \mathrm{ph}{:}\, h\}, 0.9), (\{\mathrm{pe}{:}\, h, \mathrm{ph}{:}\, k\}, -1.0) \end{array} \right\} \end{array}$$

Als men nu voor padexpressie P de expressie $\mathsf{sv}(\mathrm{EH} \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}$ kiest, dan wordt

$$\mathcal{V}[\![P]\!](\mathsf{UPop}) \quad = \quad \begin{array}{|c|c|c|} \hline d & f & < 0.5, 0.5, \ldots > \\ f & g & 1_{\mathsf{SEQU}} \\ g & h & 1_{\mathsf{SEQU}} \\ h & k & 1_{\mathsf{SEQU}} \\ \hline \end{array}$$

Als er nu een tweede padexpressie Q is en die is identiek aan P, dan levert een evaluatie van de padexpressie $P \circ Q$ het volgende resultaat op (volgens de specificatie van de functie fd(prop)):

$$\mathcal{V}[\![P \circ Q]\!](\mathsf{UPop}) \quad = \quad \begin{array}{|c|c|c|} \hline d & g & < 0.5, 0.5, \ldots > \\ f & h & 1_{\mathsf{SEQU}} \\ g & k & 1_{\mathsf{SEQU}} \\ \hline \end{array}$$

Dit is niet goed. Bij tupel $\langle d, q \rangle$ hoort de verdeling $< 0.4, 0.4, 0.4, \ldots >$ te staan. Het probleem is dat bij evaluatie van padexpressie Q de verdeling $< 0.5, 0.5, 0.5, \ldots >$ voor instantie f niet als invoer beschikbaar is, zoals uit $\mathcal{V}[\![P]\!](\mathsf{UPop})$ blijkt (zie tupel $\langle f, g \rangle$). Dit zal hetzelfde resultaat zijn voor padexpressie Q.
Stel dat de genoemde verdeling als invoer voor f bij evaluatie van padexpressie Q wél beschikbaar is. Dan:

$$\mathcal{V}[\![P \circ Q]\!](\mathsf{UPop}) \quad = \quad \begin{array}{|c|c|c|} \hline d & g & < 0.5, 0.5, \ldots > \ \odot \ < 0.4, 0.4, \ldots > \\ f & h & 1_{\mathsf{SEQU}} \\ g & k & 1_{\mathsf{SEQU}} \\ \hline \end{array}$$

en $< 0.5, 0.5, \ldots > \ \odot \ < 0.4, 0.4, \ldots > = < 0.2, 0.2, \ldots >$. Als nu met de hand met de propagatiefunctie de zekerheidsfactor voor hypothese g wordt nagerekend:

$$\begin{array}{l} \mathrm{CF}(f, e) \ = \ \max(0, \mathrm{CF}(d, e)) \ = \ 0.5\,, \\ \mathrm{CF}(g, e) \ = \ \max(0, \mathrm{CF}(f, e)) \ * \ 0.8\,, \end{array}$$

dan is $\mathrm{CF}(g, e) \ = \ 0.4$ en dat staat niet in de frequentieverdeling, bij tupel $\langle d, g \rangle$.

De afsluiting is met betrekking tot verdelingen in zijn huidige vorm niet zo geschikt. Samengevat zijn er twee problemen die overwonnen moeten worden:

- overdracht van verdelingen van de ene padexpressie naar de ander padexpressie voor gebruik aldaar;
- foute berekening van de verdelingen bij concatenatie.

Er kan echter wel een variant op de afsluiting worden gemaakt, die dan wel beter werkt.
Om het probleem van foute berekening op te lossen, wordt er gebruik gemaakt van een variant op de concatenatie genaamd "leads to", die is gedefinieerd in definitie 4.2.5. Als nu in plaats van $P \circ Q$ de expressie $P \vdash Q$ wordt genomen, dan levert de evaluatie ervan het volgende resultaat op (onder de voorwaarde dat voor expressie Q de nodige informatie beschikbaar is):

$$\mathcal{V}[\![P \circ Q]\!](\mathsf{UPop}) \quad = \quad \begin{array}{|c|c|c|} \hline d & g & < 0.4, 0.4, \ldots > \\ f & h & 1_{\mathsf{SEQU}} \\ g & k & 1_{\mathsf{SEQU}} \\ \hline \end{array}$$

Nu staat voor tupel $\langle d, g\rangle$ wel de juiste waarde als zekerheidsfactor in de verdeling.
Het ligt dus voor de hand dat de variant op de afsluiting gebruik maakt van "leads to". Er is trouwens ook een versie van "leads to" die dan de andere kant op werkt, genaamd " follows from". Voor het gemak en de volledigheid wordt op basis van die operatie ook een variant op de afsluiting gegeven.

Definitie 4.4.1

Zij P een padexpressie. Dan zijn $P^{\vdash}$ en $P^{\dashv}$ weer padexpressies waarvan de semantieken als volgt zijn gedefinieerd:

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!]\,(\mathsf{UPop})$
forward propagation	$P^{\vdash}$	$\mathcal{V}[\![\mathsf{ds}(\bigcup_{n\in\mathbb{N}}\ \mathsf{propforward}(P,n))]\!]\,(\mathsf{UPop})$
backward propagation	$P^{\dashv}$	$\mathcal{V}[\![\mathsf{ds}(\bigcup_{n\in\mathbb{N}}\ \mathsf{propbackward}(P,n))]\!]\,(\mathsf{UPop})$

waarin

$$\begin{aligned} \mathsf{propforward}(P,0) &= P\ , \\ \mathsf{propforward}(P,n+1) &= (\mathsf{propforward}(P,n) \vdash P)\ , \end{aligned}$$

$$\begin{aligned} \mathsf{propbackward}(P,0) &= P\ , \\ \mathsf{propbackward}(P,n+1) &= (P \dashv \mathsf{propbackward}(P,n)). \end{aligned}$$

□

Voor alle duidelijkheid wordt opgemerkt dat voor het zekerheidsfactormodel de voorwaartse propagatie gebruikt wordt. Dit omdat er een uitspraak is gedaan over de evaluatie van de padexpressie (propositie 4.4.3). Evaluatie van de padexpressie geschiedt van links naar rechts.

Het laatste probleem dat nog moet worden opgelost, is de overdracht (propagatie) van de informatie in frequentieverdelingen van de ene padexpressie naar de andere padexpressie. Als nu in een multiset bij de tupels de frequentieverdelingen bekend zijn, is het de vraag waar ze precies weer gebruikt worden. Dat is daar waar de hypothesen als aanwijzingen zijn opgenomen, ofwel objecttype EH. Met de instanties van de populatie van EH moeten dan die verdelingen worden geassocieerd. Hiervoor wordt dan gebruik gemaakt van de padexpressie operatie cfd. Dus door in de padepxressie P expressie EH door cfd(EH) te vervangen zou men er al zijn. Echter moet nog de functie fd die met objecttype EH wordt geassocieerd worden gespecificeerd.
Zij $x \in \mathsf{Pop}(\mathrm{EH})$, dan is

$$\mathrm{fd}(\mathrm{EH}) = \begin{cases} v & , \text{als } (x,v) \in \pi_2\,(\mathsf{STORAGE}) \\ 1_{\mathsf{SEQU}} & , \text{anders} \end{cases}$$

Nu moet nog in de padexpressie worden aangegeven wat in de standaard multiset verwerkt moet worden. Zij de padexpressie $P = \mathsf{sv}(\mathrm{EH} \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}$ en zij de padexpressie $Q = \mathsf{sv}(\mathsf{cfd}(\mathrm{EH}) \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}$. Men heeft pas de correcte verdelingen door te geven als de padexpressie P **geheel** geëvalueerd is. Dus moet om die verdelingen te kunnen gebruiken de hele expressie gesaved worden. Als nu $\mathsf{sv}\,P$ wordt geëvalueerd, dan

$$\mathcal{V}[\![P]\!]\,(\mathsf{UPop}) = \begin{array}{|c|c|c|} \hline d & f & <0.5, 0.5, \ldots> \\ \hline f & g & 1_{\mathsf{SEQU}} \\ \hline g & h & 1_{\mathsf{SEQU}} \\ \hline h & k & 1_{\mathsf{SEQU}} \\ \hline \end{array} =: \mathsf{STORAGE}$$

Dan is

$$\mathcal{V}[\![Q]\!]\,(\mathsf{UPop}) = \begin{array}{|c|c|c|} \hline d & g & <0.4, 0.4, \ldots> \\ \hline f & h & 1_{\mathsf{SEQU}} \\ \hline g & k & 1_{\mathsf{SEQU}} \\ \hline \end{array} = \mathcal{V}[\![(\mathsf{sv}\ P \vdash Q)]\!]\,(\mathsf{UPop})$$

En dat is precies datgene dat gewenst is.
Als men de afsluiting wil gebruiken dan neemt men de expressie $T = \mathsf{sv}\ P$. Als men de onzekerheidspopulatie aan het begin van deze paragraaf heeft, dan levert evaluatie van $T^{\vdash}$ als resultaat:

$$\mathcal{V}[\![T^{\vdash}]\!](\mathsf{UPop}) \;=\; \begin{array}{|c|c|c|}\hline d & f & <0.5,0.5,\ldots> \\ f & g & 1_{\mathsf{SEQU}} \\ g & h & 1_{\mathsf{SEQU}} \\ h & k & 1_{\mathsf{SEQU}} \\ d & g & <0.4,0.4,\ldots> \\ f & h & 1_{\mathsf{SEQU}} \\ g & k & 1_{\mathsf{SEQU}} \\ d & h & <0.36,0.36,\ldots> \\ f & k & 1_{\mathsf{SEQU}} \\ d & k & <-0.36,-0.36,\ldots> \\ \hline \end{array}$$

4.4.2.4.2 Het algemene geval

Er is niet alleen propagatie, maar ook co-conclusie en samengestelde aanwijzingen. In de paragraaf over combinatiefuncties zijn padexpressies gegeven. Deze worden hier bij elkaar geveegd, waarbij in elk van de expressies de expressie EH wordt vervangen door cfd(EH). Bij evaluatie van de verzamelde expressie worden alle verdelingen bekend die dan doorgegeven moeten worden, dus moet over het geheel van die padexpressies een operatie sv gezet worden. Dus:

$$X \;=\; \mathsf{sv}\left(\begin{array}{l}(\mathsf{sv}(\mathsf{cfd}(\mathrm{EH})\circ \mathrm{pe})\circ \mathsf{cfd}(\mathrm{prop})\circ \mathrm{ph}^{\leftarrow}\circ \mathrm{H})\cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH})\circ \mathrm{oe}\circ \mathrm{or}\circ \mathrm{oh}^{\leftarrow})\vdash \mathsf{cfd}(\mathrm{O}))\cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH})\circ \mathrm{ae}\circ \mathrm{and}\circ \mathrm{ah}^{\leftarrow})\vdash \mathsf{cfd}(\mathrm{A}))\cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH})\circ \mathrm{ce}\circ \mathrm{co}\circ \mathrm{ch}^{\leftarrow})\vdash \mathsf{cfd}(\mathrm{C}))\end{array}\right) \tag{4.19}$$

Er is nog geen definitieve uitspraak gedaan over hoe de gegevens in de standaard multiset worden opgeslagen. In paragraaf 4.4.2.3 was er keus uit twee mogelijkheden:

$$\begin{array}{ll}\textbf{(1)} & \mathsf{STORAGE} \;=\; \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![Z]\!](\mathsf{UPop}))) \\ \textbf{(2)} & \mathsf{STORAGE} \;=\; (\mathsf{STORAGE} - \mathcal{V}[\![Z]\!](\mathsf{UPop}))\cup \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![Z]\!](\mathsf{UPop})))\end{array}$$

Hier is Z een willekeurige padexpressie. Bekijk nu de padexpressie X. Er zijn vier operaties waar uit een en dezelfde standaard multiset de gegevens gebruikt worden (cfd(EH)). Kiest men voor het eerste mogelijkheid van opslag, dan zijn de gegevens na één keer opvragen door een "lokale" save-operatie sv gewist. De tweede keus is de beste oplossing, dus

$$\mathsf{STORAGE} \;=\; (\mathsf{STORAGE} - \mathcal{V}[\![Z]\!](\mathsf{UPop}))\cup \mathsf{Multi}(\mathsf{Set}(\mathcal{V}[\![Z]\!](\mathsf{UPop}))) \tag{4.20}$$

4.4.2.5 Een proef op de som

Aangezien de toepassing van het zekerheidsfactormodel in LISA-D vrij ingewikkeld is en de kans op fouten daardoor aanwezig is, wordt de constructie hier getoetst.

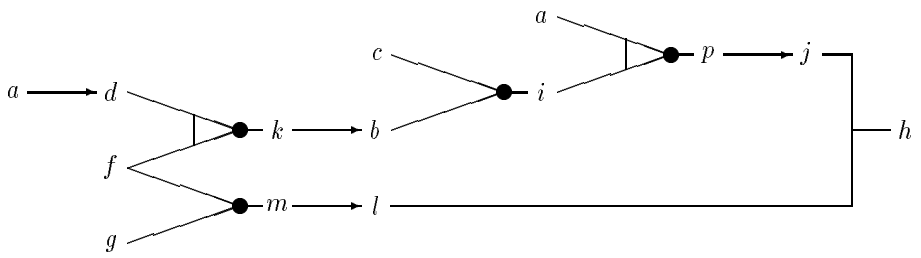
De onzekerheidspopulatie

Voor de creatie van de onzekerheidspopulatie wordt gebruik gemaakt van de populatie in voorbeeld 4.4.2. De met de instanties te associëren zekerheidsfactoren zijn ontleend aan een voorbeeldUitwerking van het zekerheidsfactormodel in [LG91]. Hier worden van de relevante objecttypen de relevante onzekerheidspopulaties gegeven.

$$\begin{array}{rcl}\mathsf{UPop}(\mathrm{E}) & = & \{(a,1.0),(c,0.5),(f,0.7),(g,-0.4)\} \\ \mathsf{UPop}(\mathrm{EH}) & = & \left\{\begin{array}{l}(a,1.0),(b,\xi_{\mathcal{U}}),(c,0.5),(d,\xi_{\mathcal{U}}),(f,0.7),(g,-0.4), \\ (h,\xi_{\mathcal{U}}),(i,\xi_{\mathcal{U}}),(j,\xi_{\mathcal{U}}),(k,\xi_{\mathcal{U}}),(l,\xi_{\mathcal{U}}),(m,\xi_{\mathcal{U}}),(p,\xi_{\mathcal{U}})\end{array}\right\}\end{array}$$

$$
\begin{aligned}
\mathsf{UPop}(\text{prop}) &= \left\{\begin{matrix}(\{\text{pe}:a,\text{ph}:d\},0.75),(\{\text{pe}:k,\text{ph}:b\},0.6),\\(\{\text{pe}:m,\text{ph}:l\},0.4),(\{\text{pe}:p,\text{ph}:j\},0.8)\end{matrix}\right\}\\
\mathsf{UPop}(\text{or}) &= \left\{\begin{matrix}(\{\text{oe}:f,\text{oh}:m\},\xi_{\mathcal{U}}),(\{\text{oe}:g,\text{oh}:m\},\xi_{\mathcal{U}}),\\(\{\text{oe}:b,\text{oh}:i\},\xi_{\mathcal{U}}),(\{\text{oe}:c,\text{oh}:i\},\xi_{\mathcal{U}})\end{matrix}\right\}\\
\mathsf{UPop}(\text{and}) &= \left\{\begin{matrix}(\{\text{ce}:a,\text{ch}:p\},\xi_{\mathcal{U}}),(\{\text{ce}:i,\text{ch}:p\},\xi_{\mathcal{U}}),\\(\{\text{ce}:d,\text{ch}:k\},\xi_{\mathcal{U}}),(\{\text{ce}:f,\text{ch}:k\},\xi_{\mathcal{U}})\end{matrix}\right\}\\
\mathsf{UPop}(\text{co}) &= \{(\{\text{ce}:j,\text{ch}:h\},\xi_{\mathcal{U}}),(\{\text{ce}:l,\text{ch}:h\},\xi_{\mathcal{U}})\}
\end{aligned}
$$

De onzekerheidspopulatie geeft grafisch gezien het netwerk in figuur 4.4.2.5 weer. Deze netwerk is een gemodificeerde versie van het netwerk in figuur 2.1 op bladzijde 8.



Figuur 4.11: Een gemodificeerde versie van het afleidingsnetwerk in het hoofdstuk over expert systemen

Padexpressies: evaluaties

In paragraaf 4.4.2.4.2 staat een padexpressie X in vergelijking 4.19 die hier wordt gebruikt. De bedoeling is dat de expressie $X^{\vdash}$ wordt geëvalueerd. Dat wordt hier in stapjes gedaan, waarbij de definitie van propforward wordt gebruikt (zie definitie 4.4.1). Het is dan zo beter te zien of er geen fouten zijn gemaakt. Aan het netwerk in figuur 4.4.2.5 te zien is het langste pad 7 stappen lang. Dus moeten de functies $\mathsf{propforward}(X,n)$ voor $n=0,\ldots,6$ worden geëvalueerd. Zoals in propopsitie 4.4.3 is aangegeven geschiedt evaluatie van links naar rechts.

$$\mathcal{V}[\![\mathsf{propforward}(X,0)]\!](\mathsf{UPop}) = \mathcal{V}[\![X]\!](\mathsf{UPop}) =$$

$$\left\{\begin{array}{lll}
(\langle a,d\rangle, <0.75,0.75,\ldots>)\uparrow^1, & (\langle d,k\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle f,k\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle f,m\rangle, <0.7,0.7,\ldots>)\uparrow^1, & (\langle g,m\rangle, <0.7,0.7,\ldots>)\uparrow^1, & (\langle k,b\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle m,l\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle c,i\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle b,i\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle l,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle i,p\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle a,p\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle p,j\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle j,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1 &
\end{array}\right\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,1)]\!](\mathsf{UPop}) = \mathcal{V}[\![(X\vdash X)]\!](\mathsf{UPop}) =$$

$$\left\{\begin{array}{lll}
(\langle a,k\rangle, <0.75,0.75,\ldots>)\uparrow^1, & (\langle c,p\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle b,p\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle d,b\rangle, <0.42,0.42,\ldots>)\uparrow^1, & (\langle f,b\rangle, <0.42,0.42,\ldots>)\uparrow^1, & (\langle k,i\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle f,l\rangle, <0.28,0.28,\ldots>)\uparrow^1, & (\langle g,l\rangle, <0.28,0.28,\ldots>)\uparrow^1, & (\langle m,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle i,j\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle a,j\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle p,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1
\end{array}\right\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,2)]\!](\mathsf{UPop}) = \mathcal{V}[\![((X\vdash X)\vdash X)]\!](\mathsf{UPop}) =$$

$$\left\{\begin{array}{lll}
(\langle a,b\rangle, <0.42,0.42,\ldots>)\uparrow^1, & (\langle d,i\rangle, <0.5,0.5,\ldots>)\uparrow^1, & (\langle f,i\rangle, <0.5,0.5,\ldots>)\uparrow^1,\\
(\langle \mathbf{f},\mathbf{h}\rangle, \mathbf{1}_{\mathsf{SEQU}})\uparrow^1, & (\langle g,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle k,p\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle c,j\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle b,j\rangle, 1_{\mathsf{SEQU}})\uparrow^1, & (\langle i,h\rangle, 1_{\mathsf{SEQU}})\uparrow^1,\\
(\langle \mathbf{a},\mathbf{h}\rangle, \mathbf{1}_{\mathsf{SEQU}})\uparrow^1 & &
\end{array}\right\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,3)]\!]\,(\mathsf{UPop}) \;=\; \mathcal{V}[\![(((X \vdash X) \vdash X) \vdash X)]\!]\,(\mathsf{UPop}) \;=$$

$$\left\{\left[\begin{array}{lll}(\langle a,i\rangle\,,<0.5,0.5,\ldots>)\uparrow^1, & (\langle d,p\rangle\,,<0.5,0.5,\ldots>)\uparrow^1, & (\langle f,p\rangle\,,<0.5,0.5,\ldots>)\uparrow^1,\\ (\langle k,j\rangle\,,1_{\mathsf{SEQU}})\uparrow^1, & (\langle c,h\rangle\,,1_{\mathsf{SEQU}})\uparrow^1, & (\langle b,h\rangle\,,1_{\mathsf{SEQU}})\uparrow^1\end{array}\right]\right\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,4)]\!]\,(\mathsf{UPop}) \;=\; \mathcal{V}[\![((((X \vdash X) \vdash X) \vdash X) \vdash X)]\!]\,(\mathsf{UPop}) \;=$$

$$\left\{\left[\begin{array}{ll}(\langle a,p\rangle\,,<0.5,0.5,\ldots>)\uparrow^1, & (\langle d,j\rangle\,,<0.4,0.4,\ldots>)\uparrow^1,\\ (\langle f,j\rangle\,,<0.4,0.4,\ldots>)\uparrow^1, & (\langle k,h\rangle\,,1_{\mathsf{SEQU}})\uparrow^1\end{array}\right]\right\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,5)]\!]\,(\mathsf{UPop}) \;=\; \mathcal{V}[\![(((((X \vdash X) \vdash X) \vdash X) \vdash X) \vdash X)]\!]\,(\mathsf{UPop}) \;=$$

$$\{[(\langle a,j\rangle\,,<0.4,0.4,\ldots>)\uparrow^1,\ (\langle d,h\rangle\,,1_{\mathsf{SEQU}})\uparrow^1,\ (\langle \mathbf{f},\mathbf{h}\rangle\,,\mathbf{1}_{\mathsf{SEQU}})\uparrow^1]\}$$

$$\mathcal{V}[\![\mathsf{propforward}(X,6)]\!]\,(\mathsf{UPop}) \;=\; \mathcal{V}[\![((((((X \vdash X) \vdash X) \vdash X) \vdash X) \vdash X) \vdash X)]\!]\,(\mathsf{UPop}) \;=$$

$$\{[(\langle \mathbf{a},\mathbf{h}\rangle\,,<\mathbf{0.568},\mathbf{0.568},\ldots>)\uparrow^1]\}$$

In enkele multisets zijn elementen vet gedrukt. Dat zijn elementen die meer dan één keer zouden voorkomen als al die multisets worden verenigd. Volgens de definitie van $X^{\vdash}$ kan elke element maar één keer voorkomen. Dan is er een probleem. De vetgedrukte elementen zijn verschillend in frequentieverdelingen. Voor bijvoorbeeld de tupel $\langle a,h\rangle$ zijn er de verdelingen 1_{SEQU} en $<0.568,0.568,\ldots>$. Volgens de definitie van de vereniging ($\cup$) zal uiteindelijk één van deze twee verdelingen met die tupel worden geassocieerd. Dat wordt zeer vervelend als de verkeerde verdeling wordt meegegeven.
De expressie $X^{\vdash}$ zal bij evaluatie wat de verdelingen betreft niet gegarandeerd de juiste uitkomst geven. Er is een manier om daar wat aan te doen. De standaard multiset waarvan de inhoud in de volgende paragraaf wordt gegeven, bevat wel de juiste informatie. Na evaluatie van $X^{\vdash}$ staan daar voor elke hypothese x de laatste berekeningen van de zekerheidsfactoren in de verdelingen dat met de tupels $\langle w,x\rangle$ zijn gerelateerd. Van deze standaard multiset wordt nog één keer gebruik gemaakt. Als als padexpressie wordt genomen:

$$Z \;=\; (X^{\vdash} \vdash \mathsf{cfd}(\mathrm{EH}))\ , \tag{4.21}$$

dan wordt bij evaluatie ervoor gezorgd dat de juiste verdelingen met de tupels worden geassocieerd. Dus:

$$\{[(\langle f,h\rangle\,,<0.568,0.568,\ldots>),\ (\langle a,h\rangle\,,<0.568,0.568,\ldots>)]\} \subseteq \mathcal{V}[\![Z]\!]\,(\mathsf{UPop})$$

En passant is hiermee nog een ander probleempje opgelost. Hadden we namelijk als padexpressie $\not\vdash X^{\vdash}$, dan zouden er tupels meerdere keren voorkomen, maar dan wéér met verschillende verdelingen. Hier zij nog opgemerkt dat die padexpressie tupels met slechts hypothesen erin oplevert. De aanwijzingen zijn er uitgefilterd.

De standaard multiset

De padexpressie X bevat een aantal sv-expressies waardoor de standaard multiset wordt gevuld. Tabel 4.4.2.5 geeft een overzicht van de inhoud van STORAGE. Per stap wordt aangegeven wat de zekerheidsfactor in de verdeling per tupel is. De invulling is op basis van de evaluatie van de functie $\mathsf{propforward}(X,n)$. De lege plekken geven aan dat men met een neutrale verdeling te maken heeft.

4.4.2.6 Uitbreiding mogelijkheden voor het zekerheidsfactormodel

Met de padexpressie Z in vergelijking 4.21 is men nu in staat voor elke hypothese h een zekerheidsfactor te bepalen. Deze padexpressie levert een multiset op met daarin elementen van de vorm $(\langle e,h\rangle\,,<cf,\ldots>)$. In de volgende paragrafen wordt bekeken of met het resultaat wat mee te doen valt.

$x \in$ Elem(STORAGE)	waarden in bijbehorende frequentieverdelingen						
	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
$\langle a,d\rangle$	0.75	0.75	0.75	0.75	0.75	0.75	0.75
$\langle d,k\rangle$		0.7	0.7	0.7	0.7	0.7	0.7
$\langle f,k\rangle$		0.7	0.7	0.7	0.7	0.7	0.7
$\langle f,m\rangle$	0.7	0.7	0.7	0.7	0.7	0.7	0.7
$\langle g,m\rangle$	0.7	0.7	0.7	0.7	0.7	0.7	0.7
$\langle k,b\rangle$			0.42	0.42	0.42	0.42	0.42
$\langle m,l\rangle$		0.28	0.28	0.28	0.28	0.28	0.28
$\langle c,i\rangle$				0.5	0.5	0.5	0.5
$\langle b,i\rangle$				0.5	0.5	0.5	0.5
$\langle l,h\rangle$							0.568
$\langle a,p\rangle$					0.5	0.5	0.5
$\langle i,p\rangle$					0.5	0.5	0.5
$\langle p,j\rangle$						0.4	0.4
$\langle j,h\rangle$							0.568
$\langle a,\{\text{pe}:a,\text{ph}:d\}\rangle$	0.75	0.75	0.75	0.75	0.75	0.75	0.75
$\langle k,\{\text{pe}:k,\text{ph}:b\}\rangle$			0.7	0.7	0.7	0.7	0.7
$\langle m,\{\text{pe}:m,\text{ph}:l\}\rangle$		0.7	0.7	0.7	0.7	0.7	0.7
$\langle p,\{\text{pe}:p,\text{ph}:j\}\rangle$						0.5	0.5

Tabel 4.1:

4.4.2.6.1 Onderscheid bevestigde en verworpen hypothesen

Evaluatie van de padexpressie Z (in vergelijking 4.21) levert een multiset op met alle al dan niet directe relaties tussen aanwijzingen en hypothesen er in op, met de bijbehorende zekerheidsfactoren. De graadmeter om te zien welke hypothese bevestigd of verworpen is, is dezekerheidsfactor in de frequentieverdeling. Een positieve waarde bevestigt een hypothese, een negatieve verwerpt het. Om met een padexpressie op basis van het zekerheidsfactor als criterium een selectie van de elementen in de multiset toe te passen is niet gemakkelijk. Men zou kunnen sorteren met als criterium dat CF > 0 of CF < 0, maar dan krijgt men meerdere rijtjes met hetzelfde probleem als eerst bij het probabilistische afleidingsmodel (zie paragraaf 4.3.6.3). Dit keer biedt vooraf groeperen geen soelaas. Dan is de enige oplossing nieuwe padexpressies te definiëren.

Definitie 4.4.2

Zij P een padexpressie. Dan zijn confirm P *en* disconfirm P *weer padexpressies waarvan de semantiek als volgt is gedefinieerd:*

name	expr	$\mathcal{V}[\![\text{expr}]\!]\,(\text{UPop})$
confirmation	confirm P	$\left\{\left[(\langle p,q\rangle,s)\uparrow^m \;\middle|\; (\langle p,q\rangle,s)\in\mathcal{V}[\![P]\!]\,(\text{UPop}) \wedge 0 < s[m] \leq 1\right]\right\}$
disconfirmation	disconfirm P	$\left\{\left[(\langle p,q\rangle,s)\uparrow^m \;\middle|\; (\langle p,q\rangle,s)\in\mathcal{V}[\![P]\!]\,(\text{UPop}) \wedge -1 \leq s[m] < 0\right]\right\}$

□

In de definitie van de padexpressies is aangegeven bij welke zekerheidsfactoren een hypothese verworpen of bevestigd is. Wat daar staat is niet definitief. Er wordt hierbij nog opgemerkt dat de waarde 0 in geen van de gavallen wordt gebruikt. Dat is ook wel zo veilig, want wat is dan de status van een hypothese?
Met deze nieuwe padexpressies en de expressie Z kan men de volgende padexpressies maken met de omschreven resultaten:

1. Bevestigde hypothesen: PE = confirm Z

2. Alle hypothsen behalve de bevestigde: $PE \;=\; Z - \mathsf{confirm}\, Z$
3. verworpen hypothesen: $PE \;=\; \mathsf{disconfirm}\, Z$
4. Alle hypothsen behalve deze die verworpen zijn: $PE \;=\; Z - \mathsf{disconfirm}\, Z$

In het tweede en laatste geval verkrijgt men ook elementen waaarin in de bijbehorende verdelingen de zekerheidsfactor 0 staat. Met padexpressie Z wordt trouwens nog steeds de expressie in vergelijking 4.21 bedoeld.
Hier wordt de padexpressie operator $-$ gebruikt, waarvoor met betrekking tot frequentieverdelingen nog geen omschrijving is gegeven. Dat gebeurt alsnog:

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!]\,(\mathsf{UPop})$
minus	$P - Q$	$\left\{\left[\,(\langle p,q\rangle\,, s-t)\!\uparrow^{n-m} \;\middle|\; \begin{array}{l}(\langle p,q\rangle\,, s) \in^{n} \mathcal{V}[\![P]\!]\,(\mathsf{UPop}) \wedge \\ (\langle p,q\rangle\,, t) \in^{m} \mathcal{V}[\![Q]\!]\,(\mathsf{UPop})\end{array}\,\right]\right\}$

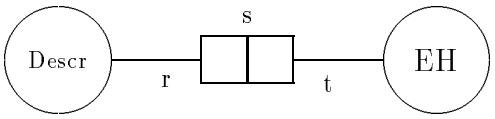
waarin

$$s-t \;=\; \begin{cases} s & , \text{als } n-m>0 \\ 1_{\mathsf{SEQU}} & , \text{anders} \end{cases}$$

Deze operatie kan bij de padexpressie veilig gebruikt worden, omdat de elementen in de multiset uniek zijn dankzij de transitieve aflsuiting die in expressie Z voorkomt.

4.4.2.6.2 Verklaren/Traceren van resultaten

Niet elke element in de multiset is zinvol. Sommige hypothesen en/of aanwijzingen zijn extra gecreëerde instanties om productieregels te kunnen populeren. De hypothesen en aanwijzingen waarvan een omschrijving is gegeven zijn van belang. De koppeling van een omschrijving met een aanwijzing/hypothese in de informatiestructuur zou volgens de aanvulling in figuur 4.4.2.6.2 kunnen zijn. Tezamen met de nieuwe padexpressies in de vorige paragraaf kan de omschrijving van



Figuur 4.12: Een uitbreiding van de informatiestructuur waarin de productieregels zijn gepopuleerd.

een bepaalde (soort) hypothese worden verkregen.
Alle omschrijvingen van hypothsen verkrijgt men met de padexpressie Descr $\circ$ r $\circ$ s $\circ$ t$^{\leftarrow}$ $\circ$ EH $\circ$ $Z^{\leftarrow}$.
De padexpressie Z moet gereverseerd worden, omdat deze elementen oplevert waarin de tupels met aanwijzingen begint. De objecttype EH kan veilig in de expressie worden opgenomen: alleen a priori aanwijzingen hebben een zekerheidsfactor en die aanwijzingen komen niet in het eindresultaat voor omdat ze geen hypothesen van een een of ander regel zijn.

Soms zoekt men niet naar hypothesen, maar naar aanwijzingen die tot de al dan niet verworpen of bevestigde hypothesen. Men wil wel eens de uitkomst traceren, of via de omschrijvingen verklaard hebben wat de aanleiding tot een of ander hypothese is geweest. De zinvolle aanwijzingen zijn die aanwijzingen die een omschrijving hebben. Een 'simpele' padexpressie om dit resultaat te bereiken, is Descr $\circ$ r $\circ$ s $\circ$ t$^{\leftarrow}$ $\circ$ EH $\circ$ Z. Echter is in de elementen van het eindresultaat niet de juiste verdelingen met de tupels geassocieerd, en daarnaast beïnvloeden de frequentieverdelingen bij a priori aanwijzingen de verdelingen in de multiset als resultaat van Z. De a priori aanwijzingen zitten in de populatie van objecttype EH. Dus er wordt gezocht naar elementen van de vorm $(\langle e,h\rangle\,, s)$ met aanwijzing e er in, maar met verdelingen t die in elementen van de vorm $(\langle \ldots, e\rangle\,, t)$ zitten. Een padexpressie die ervoor zorgt dat elementen van de vorm $(\langle e, \ldots .\rangle\,, t)$ in de uitkomst zit, is

$$\not{f}\, Z \;\cap\; \not{f}\, (Z^{\leftarrow})$$

Nu is er een probleem! Als $(\langle p,p\rangle, s) \in \mathcal{V}[\![\mathcal{f}\ Z]\!](\mathsf{UPop})$ en $(\langle p,p\rangle, t) \in \mathcal{V}[\![\mathcal{f}\ (Z^{\leftarrow})]\!](\mathsf{UPop})$, wat is dan $s \cap t$? Het is mogelijk dat $s \neq 1_{\mathsf{SEQU}}$ en $t \neq 1_{\mathsf{SEQU}}$, maar zeker is dat $s \neq t$! De bedoeling is eigenlijk dat

$$s \cap t \ = \ \begin{cases} s \text{ , als } t = 1_{\mathsf{SEQU}} \\ t \text{ , als } s = 1_{\mathsf{SEQU}} \\ s \text{ , als } s = t \text{ (aanname)} \end{cases}$$

Een andere oplossing of padexpressie is er niet. De kern van het probleem zit in het feit dat uit populatie niet meer de oorspronkelijke regels terug te vinden zijn. Als nu de originele regels ergens waren gepopuleerd, dan zou er een oplossing zijn. Tot dan blijft het probleem bestaan wat nu $s \cap t$ is als $s \neq t$ en het zijn geen neutrale verdelingen.
We merken terloops op dat afgezien van de frequentieverdelingen in de gegeven padexpressie niet helemaal goed is: de a priori aanwijzingen komen niet in de uitkomst voor.

Hoofdstuk 5

Integratie: een evaluatie

Nu is er een basismodel geconstrueerd waarop een model uit expertsystemen en een model uit information retrieval systemen is opgebouwd en uitgeprobeerd. Nu kan er op basis van de resultaten in het vorige hoofdstuk een antwoord worden gegeven op de vraag die al eerder is gesteld:

> Is het met LISA-D mogelijk een quasi-probabilistisch model toe te passen en er mee te werken? Wat moet er gedaan worden om die mogelijkheid te bieden en zijn er voorwaarden en/of eisen? En als het niet kan, waarom niet?

Voordat op die vragen antwoorden gegeven worden, laten we eerst de ervaringen met een gekozen model uit expertsystemen de revue passeren. Hetzelfde gebeurt met een model uit information retrieval systemen. Tenslotte wordt op een rijtje gezet wat het zogenaamde basismodel waarmee LISA-D moet worden uitgebreid, inhoudt.

5.1 Expertsystemen

Bij integratie in padexpressies zijn de belangrijkste kenmerken van enkele modellen bij expertsystemen behandeld:

- kennisrepresentatie,
- propagatie van (onzekerheden in) aanwijzingen, en
- de combinatiefuncties.

Deze kenmerken zijn van toepassing op het zekerheidsfactormodel ([BS84]), de subjectieve Bayesische methode ([DHN90]), en de Dempster-Shafer theorie ([LG91]). Deze modellen zijn in het hoofdstuk over expertsystemen behandeld. Er zijn weliswaar onderling verschillen wat betreft de maten voor onzekerheid en de omschrijving van de combinatiefuncties, maar de principes zijn nagenoeg hetzelfde.
Bij integratie in padexpressies is slechts naar het zekerheidsfactormodel gekeken. De reden daarvan is, dat de modellen onderling verschillen in maten voor onzekerheid en dat werkt in de omschrijvingen van padexpressies met betrekking tot frequentieverdelingen door.

5.1.1 Het propagatiemechanisme

Het is bij het zekerheidsfactormodel gelukt om een padexpressie te construeren zo dat voor elke hypothese gegeven de aanwijzingen de juiste zekerheidsfactoren worden bepaald. Deze padexpressie wordt hieronder gegeven (zie vergelijkingen 4.21 en 4.19):

$$Z \;=\; (X^{\vdash} \vdash \mathsf{cfd}(\mathrm{EH}))\ ,$$

waarin

$$X = \mathsf{sv}\left(\begin{array}{l}(\mathsf{sv}(\mathsf{cfd}(\mathrm{EH}) \circ \mathrm{pe}) \circ \mathsf{cfd}(\mathrm{prop}) \circ \mathrm{ph}^{\leftarrow} \circ \mathrm{H}) \cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH}) \circ \mathrm{oe} \circ \mathrm{or} \circ \mathrm{oh}^{\leftarrow}) \vdash \mathsf{cfd}(\mathrm{O})) \cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH}) \circ \mathrm{ae} \circ \mathrm{and} \circ \mathrm{ah}^{\leftarrow}) \vdash \mathsf{cfd}(\mathrm{A})) \cup \\ (\mathsf{sv}(\mathsf{cfd}(\mathrm{EH}) \circ \mathrm{ce} \circ \mathrm{co} \circ \mathrm{ch}^{\leftarrow}) \vdash \mathsf{cfd}(\mathrm{C}))\end{array}\right).$$

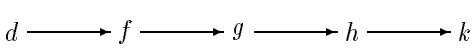
In paragraaf 4.4.2.5 is deze padexpressie uitgeprobeerd en met succes. De correcte evaluatie van de padexpressie is echter zeer afhankelijk van de manier waarop padexpressies worden geëvalueerd. In propositie 4.4.3 op bladzijde 111 is aangenomen dat de padexpressie letterlijk van links naar rechts wordt geëvalueerd. Dit is **zeer belangrijk** omdat

1. de wijze waarop met de padexpressie propagatie van aanwijzingen is bewerkstelligd op die aanname gebaseerd is. De padexpressie Z levert tupels van de vorm $\langle e, h\rangle$ op. Propagatie moet dan van links naar rechts. Om die propagatie te bewerkstelligen zijn in de padexpressie Z de operaties $\vdash$ gebruikt op basis van de wijze van evaluatie. Was de evaluatie van padexpressies van rechts naar links en Z zou tupels van de vorm $\langle h, e\rangle$ opleveren, dan moet de operatie $\dashv$ worden gebruikt;
2. de manier waarop voor de combinatiefuncties de gegevens correct uit STORAGE vergaard worden, van deze aanname afhangt. In de specificaties van de functies fd wordt de informatie vergaard op basis van wat in de tupels $\langle p, q\rangle$ staat. Er wordt voor het verkrijgen van de waarden in de gezochte verdelingen ofwel gezocht naar tupels met die ene q aan de rechterkant (π_2), of naar tupels met die ene p aan de linkerkant (π_1).

Een wijziging van de wijze van evaluatie van padexpressies wordt dus gestraft met de volgende problemen die dan opdoemen:

1. De functies fd die specificaties van de combinatiefuncties bevatten, halen de verkeerde gegevens uit de standaard multiset;
2. De operatie $\vdash$ in de padexpressie Z die er juist voor moest zorgen dat de juiste verdelingen met de betreffende tupels worden geassocieerd, heeft nu het omgekeerde effect: de verkeerde verdelingen worden meegegeven.

Propagatie gaat dus volledig de mist in.
Er is nog een reden om aan de wijze van evaluatie van padexpressies vast te houden. Als in het netwerk in figuur 5.1 d een a priori aanwijzing is en k de te bereiken hypothese, dan gaat het fout als men niet eerst voor tupel $\langle d, f\rangle$ een bijbehorende verdeling bepaalt. Als men eerst voor tupel $\langle f, g\rangle$ een verdeling zou bepalen, dan krijgt men de foute verdeling omdat de informatie over de hypothese f die hier als aanwijzing wordt gebruikt nog niet bekend is. Hier moet men dus wat padexpressies betreft echt netjes van links naar rechts werken.



Figuur 5.1: Een netwerk waarin d een a priori aanwijzing is en k een hypothese

5.1.2 Kennisrepresentatie

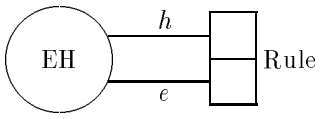
Bij expertsystemen is kennis gerepresenteerd in de vorm van productieregels. De relaties tussen de regels kunnen worden gevisualiseerd in een afleidingsnetwerk.
De regels kunnen in hun oorspronkelijke vorm niet zomaar in de populatie van een een of ander informatiestructuur worden opgeslagen om de volgende redenen:

- In een multiset is de onderscheid tussen de regels verdwenen. Aan de tupels is niet te zien of men met samengestelde aanwijzingen of co-concluderende regels te maken heeft;

- Omdat de onderscheid tussen de regels verdwenen is, is het niet mogelijk om te bepalen welke combinatie er nu gebruikt moet worden.

De productieregels moeten dus worden omgevormd. De beschreven manier waarop de regels moeten worden omgevormd is succesvol. De omgevormde produvtieregels kunnen wél duidelijk worden onderscheiden. De verschillende soorten regels kunnen in de informatiestructuur zoals in figuur 4.9 gemodelleerd en gepopuleerd worden. Daarnaast is het ook mogelijk geworden op de juiste plaats aan te geven welk van de combinatiefuncties gebruikt moet worden. Het feit dat het propagatiemechanisme (goed) werkt, getuige de proef met het zekerheidsfactormodel in paragraaf 4.4.2.5 spreekt boekdelen.

Er is echter een probleempje dat nog opgelost moet worden. Uit de gepopuleerde omgevormde productieregels zijn de oorspronkelijke regels niet meer terug te vinden. Dit is van belang als men het proces dat tot een hypothese leidde, wil reconstrueren voor verklaringen of een contrôle van de berekening van waarden. In paragraaf 4.4.2.6.2 is aangegeven waar de oorspronkelijke regels voor nodig zijn. De oorpsronkelijke regels moeten ergens zijn gepopuleerd. De informatiestructuur kan bijvoorbeeld uitgebreid worden op de manier zoals dat in figuur figuur 5.1.2 gedaan is. Daar krijgt men weer het probleem dat men niet kan "zien" of men met samengestelde aanwijzingen van doen heeft of met co-concluderende regels.



Figuur 5.2: Uitbreiding van de informatiestructuur om de oorspronkelijke regels te populeren.

5.1.3 Specificatie van combinatiefuncties

Met behulp van de expressie cfd zijn de combinatiefuncties in te zetten om de nodige verdelingen te produceren. De functies zijn ondergebracht in de specificaties van de functies fd. Daarvoor is het nodig dat voor elk combinatiefunctie een (uniek) objecttype is waaraan zo'n functie fd (fd : $\mathcal{O} \rightarrow$ SEQU) waarin de combinatiefunctie is gespecificeerd, kan worden gehangen.
Een succesvolle specificatie hangt van de volgende voorwaarden af:

- Er wordt rekening gehouden met de onzekerheidspopulatie van het desbetreffende objecttype;
- De wijze waarop de gegevens die voor de functies nodig zijn uit de standaard multiset worden gehaald aan de hand van de tupels. Het is maar net of men op tupels met inhoud $\langle d, q \rangle$ of $\langle q, d \rangle$ zoekt. Een succesvol gebruik van de functie is afhankelijk van het feit of de juiste projectie (π) wordt gebruikt.

Op deze voorwaarden is niet formeel te controleren. Fouten maken is hier zeer mogelijk.

5.2 Information retrieval

Als testvehikel is als model in information retrieval systemen het probabilistische afleidingsmodel gekozen. Dit model is gekozen om de details die erover bekend zijn. De varianten van dit model zijn het vectorruimte model (Vector Space Model, VSM) en het gegeneraliseerde vectorruimte model (Generalized Vector Space Model, GVSM).
In eerste instantie was niet helemaal duidelijk waar naar gekeken moet worden, maar nu kan daar wel iets concreets over gezegd worden:

- kennisrepresentatie,
- bepaling van de relevanties, en
- enkele extra mogelijkheden.

5.2.1 Kennisrepresentatie

In tegenstelling tot het zekerheidsfactormodel is bij het probabilistische afleidingsmodel de kennis duidelijk afgebakend. Er zijn slechts documenten, de representatie ervan in termconcepten (index termen), en de queries die ook met de termconcepten zijn gerelateerd. Deze kennis heeft een vaste structuur en is niet zo dynamisch als de kennis in de vorm van productieregels is. Het is dus een fluitje van een cent om de kennis in zijn geheel met zijn duidelijk te onderscheiden compenenten te modelleren in een informatiestructuur en te populeren.
Wat die structuur van de kennis betreft is dit niet alleen maar het geval in het probabilistische afleidingsmodel, maar ook in andere modellen zoals het afleidingsnetwerk voor document retrieval van Turtle en Croft ([TC90]) en het index expressie vertrouwensnetwerkmodel van Bruza en Van der Gaag ([BG92]). Met het laatstgenoemde model is het wat lastiger, maar niet onmogelijk.

5.2.1.1 De representatie van de query

Op basis van de query worden de relevante documenten bepaald. Echter moet eerst de query gekoppeld worden met de concpten waarmee de documenten worden gerepresenteerd. Bij het gebruik van het probabilistische afleidingsmodel in LISA-D komt het er op neer dat de query expliciet in de informatiestructuur moet worden gepopuleerd. Op deze manier kunnen dan de relaties tussen de query en de concepten worden vastgelegd. In paragraaf 4.3.1.1 op bladzijde 89 is aangegeven dat de populatie van de queries exponentieel groot kan worden.
Een simpele oplossing is er niet. In modellen zoals het afleidingsnetwerk voor document retrieval en het index expressie vertrouwensnetwerk model wordt een expressieboom voor een gegeven query gegenereerd alvorens om een koppeling met de concepten waarmee de documenten worden gerepresenteerd, mogelijk te maken. Bij het probabilistische afleidingsmodel is het niet zo erg, maar het probleem is hier eigenlijk hoe de query is te vertalen in termconcepten.

5.2.2 Het bepalen van de relevanties

Om in het probabilistische afleidingsmodel de relevanties van documenten te bepalen moeten eerst de geïnverteerde document frequenties, term frequenties binnen documenten, en gegevens over de relaties tussen documenten en atomaire concepten en relaties tussen termconcepten en atomaire concepten bepaald worden. Tot dusver is het gelukt om deze gegevens via de padexpressie cfd te verkrijgen. Al de benodigde gegevens in de vorm van frequentieverdelingen konden in de functies fd die met de juiste objecttypen waren te associëren, gespecificeerd worden. Met de sv-operaties konden de nodige gegevens voor de bepaling van de frequentieverdelingen vergaard worden. Het resultaat is een padexpressie R dat een multiset oplevert waarin in elke element $(\langle d_i, q\rangle, v)$ de relevantie van het document d_i in de frequentieverdeling v zit. De padexpressie R is als volgt (zie vergelijking 4.14 op bladzijde 99):

$$R \quad = \quad \nu[\![\mathsf{sv}(\mathrm{D} \circ \mathrm{h} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow}) \circ W \circ Z]\!]\,(\mathsf{UPop})$$

waarin volgens de vergelijkingen 4.10 en 4.11

$$W \quad = \quad \mathsf{cfd}(\mathrm{D}) \circ \mathsf{sv}(\mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{i}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{i} \circ \mathrm{f} \circ \mathrm{h}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{j} \circ \mathrm{l} \circ \mathrm{k}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{M})$$

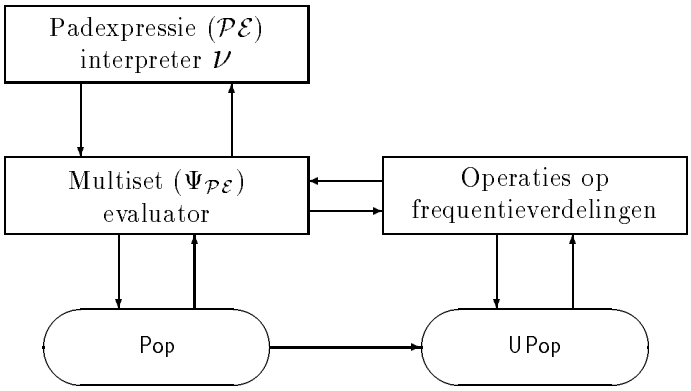
en

$$Z \quad = \quad \mathrm{M} \circ \mathrm{k} \circ \mathrm{l} \circ \mathrm{j}^{\leftarrow} \circ \mathrm{D} \circ \mathrm{h} \circ \mathsf{sv}(\mathrm{f} \circ \mathrm{j}^{\leftarrow}) \circ \mathsf{cfd}(\mathrm{T}) \circ \mathrm{p} \circ \mathrm{v} \circ \mathrm{q}^{\leftarrow} \circ \mathrm{Q}\ .$$

Merk op dat in deze padexpressies geen operaties zoals $\vdash$ en $\dashv$ voorkomen. Er is hier ook geen eis aan de wijze van evaluatie van expressies gesteld.

5.3 Het basismodel

Het basismodel bevat zaken als onzekerheidspopulatie, frequentieverdelingen en algemeen bruikbare padexpressies. Deze zaken worden hier besproken. Het onderzoek heeft zich gericht op het calculatiedomein onder padexpressies. De genoemde componenten kunnen er in ondergebracht worden, met als resultaat een structuur in figuur 5.3 dat een uitbreiding is van figuur 4.1 op bladzijde 78.



Figuur 5.3: Een hiërarchie waarin de (input-/output-)relaties tussen de componenten expliciet zijn weergegeven.

5.3.1 De onzekerheidspopulatie

In LISA-D zijn de gegevens de instanties van de populatie van een een of andere informatiestructuur. Om met de instanties maten voor onzekerheid te associëren is de onzekerheidspopulatie UPop gedefinieerd (zie definitie 4.2.3).
Uit gebruik van de onzekerheidspopulatie bij het zekerheidsfactormodel en de probabilistische afleidingsmodel is gebleken, dat precies wordt aangegeven bij welk objecttype aan instanties van de populatie door de gebruiker maten voor onzekerheid wordt toegekend, en bij welke objecttypen dat niet toegestaan is. Er is nog geen manier om dat formeel te regelen. Een idee is om een verzameling $\mathcal{M}$ van objecttypen te definiëren zo dat als een objecttype x in $\mathcal{M}$ zit, het toegestaan is om $\mathsf{UPop}(x)$ te wijzigen.
Om deze verzameling als contrôlemechanisme toe te passen, moet de definitie van de onzekerheidspopulatie aangepast worden.

Definitie 5.3.1

Zij X *en* Y *objecttypen in een gegeven informatiestructuur. De* onzekerheidspopulatie UPop *waarin met de elementen onzekerheidstoekenningen worden gerelateerd is een afbeelding* $\mathsf{UPop} : \mathcal{O} \rightarrow (\Omega \cup \wp(\Omega)) \times \mathcal{U}$ *die als volgt is gedefinieerd:*

$$\mathsf{UPop}(X) \quad = \quad \{(x, \mathcal{A}u(x)) \mid x \in \mathsf{Pop(X)}\}$$

waarvoor de volgende eigenschappen gelden:

(1) $\{(x, \mathcal{A}u(x)), (y, \mathcal{A}u(y))\} \subseteq \mathsf{UPop(X)} \wedge (x = y) \Rightarrow \mathcal{A}u(x) = \mathcal{A}u(y)$
(2) $\mathrm{Y}\ \mathsf{Gen}\ \mathrm{X} \Rightarrow \mathsf{UPop(X)} \subseteq \mathsf{UPop(Y)}$
(3) $\mathrm{Y}\ \mathsf{Spec}\ \mathrm{X} \Rightarrow \mathsf{UPop(Y)} = \{(y, \mathcal{A}u(y)) \mid y \in \mathsf{Pop(Y)} \wedge (y, \mathcal{A}u(y)) \in \mathsf{UPop(X)}\}$
(4) $\mathrm{X} \notin \mathcal{M} \Rightarrow \forall_{(x,v)\ \in\ \mathsf{UPop(X)}}\ [v = \xi_{\mathcal{U}}]$
(5) $\mathrm{X} \notin \mathcal{M} \wedge \mathrm{Y}\ \mathsf{Gen}\ \mathrm{X} \Rightarrow \forall_{y\ \in\ \mathsf{Pop(X)}}\ [(y, \xi_{\mathcal{U}}) \in \mathsf{UPop(Y)}]$
(6) $\mathrm{X} \notin \mathcal{M} \wedge \mathrm{Y}\ \mathsf{Spec}\ \mathrm{X} \Rightarrow \forall_{y\ \in\ \mathsf{Pop(Y)}}\ [(y, \xi_{\mathcal{U}}) \in \mathsf{UPop(Y)}]$

Hierin is $\mathcal{M}$ de verzameling objecttypen waarvan de onzekerheidspopulaties veranderd mogen worden. □

5.3.2 Werken met frequentieverdelingen

In deze paragraaf wordt de toepassing van de frequentieverdelingen onder de loep genomen. Zaken als padexpressies en extra mogelijkheden om allerlei operaties op verdelingen toe te passen worden hier besproken.

$\vdots$	$\vdots$
$\|s\|$	: de lengte van de verdeling: hoeveel elementen zitten er in totaal in
$s[i]$	: de waarde in de verdeling op plaats i
$s \odot t$	: concatenatie van verdelingen
$s \otimes t$	: cartesisch produkt van verdelingen (extend)
$s \cup t$	: vereniging
$s \cap t$	: doorsnede
$s - t$	: verschil
$\mathsf{seq}(v)$	: creatie van verdeling uit v
$s = t$	: vergelijken van verdelingen

Tabel 5.1: Een overzicht van de basisoperaties op frequentieverdelingen.

5.3.2.1 Basisoperaties

In de omschrijvingen van de semantieken van padexpressies worden frequentieverdelingen betrokken. Verder worden in de functies fd frequentieverdelingen opgebouwd uit andere verdelingen. Het is hier dan handig om een overzicht te geven van de access-operaties op verdelingen en operaties ten behoeve van het combineren van verdelingen tot een verdeling.
Van deze operaties kunnen geen definities worden gegeven omdat ze per quasi-probabilistisch model verschillen qua werking. Vergelijk bijvoorbeeld de omschrijving van $\cup$ in het zekerheidsfactormodel met die in het probabilistische afleidingsmodel. In het overzichtje op bladzijde 130 zijn s en t verdelingen, v een waarde uit $\mathcal{U}$ en i een natuurlijk getal.

5.3.2.2 Problemen met padexpressies

Er zijn padexpressies waarvan in de omschrijvingen niet eenduidig aangegeven kan worden wat voor een verdeling met een tupel geassocieerd moet worden. Een zo'n geval doet zich voor bij expertsystemen. In paragraaf 4.4.2.5 is vastgesteld dat het mogelijk is dat meerdere elementen met een en dezelfde tupel kunnen voorkomen, maar dat de verdelingen verschillen. De definitie van de vereniging bood daar geen uitkomst. Het was zelfs mogelijk dat er verkeerde verdelingen worden meegegeven. Dat probleem is daar met een truc in vergelijking 4.21 opgelost door achter padexpressie X een expressie cfd(EH) achter te plakken. Het kan niet de bedoeling zijn, dat met trucs gewerkt wordt. De padexpressie dat vereniging voorstelt moet ook zo gebruikt kunnen worden. Er zijn padexpressies waarin vereniging optreedt: concatenatie, cartesisch produkt (extend), afsluiting, ontgroepen, ontleden (unorder), confluentie en de hier gedefinieerde varianten op die expressies.
Het probabilistische afleidingsmodel kent ook problemen met de vereniging. Het probleem doet zich daar echter niet voor tijdens de berekening van de relevanties, maar bij presentatie van de gewenste resultaten (zie paragraaf 4.3.6.3). Daar is een oplossing (of truc?) voor gevonden in de vorm van een padexpressie dat confluentie heet.

5.3.2.3 Operaties op verdelingen: extra mogelijkheden

Padexpressies dienen als rekenvoorschrift voor de te berekenen verdelingen. Echter kan men met padexpressies alleen niet alle mogelijke verdelingen krijgen die men wil. Er zijn daartoe enkele nieuwe padexpressie-operaties in het leven geroepen, namelijk sv en cfd. Als aanvulling hierop zijn nog de operaties $\vdash$ en $\dashv$. Deze operaties zijn noodzakelijk als men in een multiset letterlijk de verdelingen wil vervangen.
Via de expressie cfd wordt een functie fd die met een door cfd aangegeven objecttype is geassocieerd, aangeroepen. In zo'n functie kan men specificeren wat men voor verdelingen wil met dien verstande dat men slechts als gegevensbron de onzekerheidspopulatie van het desbetreffende objecttype heeft, en de standaard multiset waarin bepaalde gegevens zitten, aangegeven via de sv-operatie. Met de functie fd kan men combinatiefuncties van expertsystemen specificeren, en bij information retrieval systemen zijn het geïnverteerde document frequenties, term frequenties binnen documenten en

dergelijke die met fd gespecificeerd kunnen worden. Aan deze opzet en het succes ervan zijn enkele voorwaarden verbonden:

- De functie fd is te associëren met een (uniek) objecttype;
- Voor de specificatie van fd is een **taal** vereist.

Tot dusver zijn alle specificaties in het probabilistische afleidingsmodel en het zekerheidsfactormodel wiskundig opgesteld. Wat die taal is, is hier dus niet zo van belang, maar wel van belang is dat die taal in LISA-D "ingebed" moet worden.

5.3.3 Standaard padexpressies

Als men van het zekerheidsfactormodel en van het probabilistische afleidingsmodel de gebruikte soorten padexpressies naast elkaar legt, dan zijn er padexpressies die met betrekking tot frequentieverdelingen semantisch verschillen, maar ook padexpressies waarin de omschrijvingen voor de te produceren frequentieverdelingen gelijk zijn. Van die laatste categorie wordt hier een overzicht gegeven. Er wordt ook een overzicht gegeven van soorten padexpressies die weliswaar niet gebruikt zijn in de modellen, maar waarvan het voor de hand ligt wat de omschrijvingen van de verdelingen zullen zijn.

- Overzicht van standaard padexpressies:

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
empty path	$\varnothing_{\mathcal{PE}}$	$\varnothing$
neutral path	$1_{\mathcal{PE}}$	$1_{\Omega \times \Omega \times \mathsf{SEQU}}$
constant	c	$\mathsf{Sqr}(\{[(c, 1_{\mathsf{SEQU}})]\})$
multiset	X	$\mathsf{Sqr}(X)$
objecttype	x	$\{\!\![(\langle y, y\rangle, \mathsf{seq}(d))\uparrow^1 \mid (y,d) \in \mathsf{UPop}(x)]\!\!\}$
predicator	p	$\{\!\![(\langle v(p), v\rangle, 1_{\mathsf{SEQU}})\uparrow^1 \mid v \in \mathsf{Pop} \circ \mathsf{Fact}(p)]\!\!\}$
reverse	$P^{\leftarrow}$	$\{\!\![(\langle q, p\rangle, s)\uparrow^n \mid (\langle p, q\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop})]\!\!\}$
front	$\not\mathbf{f}\, P$	$\mathsf{Sqr}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))$

- Overzicht van padexpressie met "triviale" omschrijvingen van de verdelingen:

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
count	Cnt P	$\mathsf{Sqr}\{\!\![(\lvert\mathcal{V}[\![P]\!](\mathsf{UPop})\rvert, 1_{\mathsf{SEQU}})\uparrow^1]\!\!\}$
sum	Sum P	$\mathsf{Sqr}\left\{\!\!\left[(v, 1_{\mathsf{SEQU}})\uparrow^1 \;\middle|\; v = \sum_{x \in^m \mathsf{Elem}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))} x \times m\right]\!\!\right\}$
max	Max P	$\mathsf{Sqr}\{\!\![(v, 1_{\mathsf{SEQU}})\uparrow^1 \mid v = \max(\mathsf{Set}(\mathsf{Elem}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))))]\!\!\}$
min	Min P	$\mathsf{Sqr}\{\!\![(v, 1_{\mathsf{SEQU}})\uparrow^1 \mid v = \min(\mathsf{Set}(\mathsf{Elem}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))))]\!\!\}$
powerset	$\wp P$	$\mathsf{Sqr}\{\!\![(v, 1_{\mathsf{SEQU}})\uparrow^1 \mid v \subseteq \mathsf{Elem}(\pi_1(\mathcal{V}[\![P]\!](\mathsf{UPop})))]\!\!\}$

Er is ook een aantal padexpressies dat in een van de overzichten thuis zou horen, maar met die padexpressies moet men voorzichtig zijn. Als testvehikels zijn het probabilistische afleidingsmodel en het zekerheidsfactormodel gebruikt, maar hoe zouden de omschrijvingen van die padexpressies zijn, als men nu eens met kanstheorie of een ander model werkte? De padexpressies die om die reden niet in de bovenstaande overzichten voorkomen, zijn concatenatie ($\circ$), Cartesisch produkt ($\diamond$), verschil ($-$), doorsnede ($\cap$), vereniging ($\cup$), groepering (φ), ordenen (ψ), ontgroepen (Υ), ontleden van ordening (Ξ), confluentie ($[\ldots]$) en afsluiting (P^+).

5.3.4 Nieuwe padexpressies

In het voorgaande hoofdstuk is een aantal nieuwe padexpressies gedefinieerd, elk met een bepaalde functie. Een opsomming van die nieuwe padexpressies wordt hieronder gegeven, met waar nodig enkele opmerkingen erbij.

- Padexpressies ten behoeve van creatie van extra mogelijkheden om allerlei operaties op frequentieverdelingen toe te passen:
  - de expressie cfd:

    $$\mathcal{V}[\![\mathsf{cfd}(x)]\!](\mathsf{UPop}) \;=\; \begin{cases} (\{[\![(\langle y,y\rangle, d)\uparrow^1 \mid y \in \mathsf{Pop(x)} \;\wedge d = \mathrm{fd}(x)]\!\} \circ \mathcal{V}[\![x]\!](\mathsf{UPop})) & , \text{als } x \in \mathcal{O} \\ \mathcal{V}[\![x]\!](\mathsf{UPop})) & , \text{anders} \end{cases}$$

    waarin fd een functie $\mathrm{fd} : \mathcal{O} \rightarrow \mathsf{SEQU}$ is, die voor elk instantie van het gegeven objecttype x de bijbehorende frequentieverdeling levert.
  - de expressie sv:

    $$\mathcal{V}[\![\mathsf{sv}\,P]\!](\mathsf{UPop}) \;=\; \{[\![x\uparrow^n \mid x \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \wedge x \in^m \mathsf{STORAGE}]\!\}$$

    waarin STORAGE de naam is voor de standaard multiset en waarin $m \in \mathbb{N} \setminus \{0\}$. m ligt hier niet vast, het hangt er van af hoe STORAGE wordt gevuld.
- Padexpressies die associaties van frequentieverdelingen met tupels "sturen":

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
leads to	$(P \vdash Q)$	$\bigcup_{r,t} \left\{ [\![(\langle p,q\rangle, t)\uparrow^{n\times m} \mid \begin{array}{c}(\langle p,r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \;\wedge \\ (\langle r,q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})\end{array}]\!] \right\}$
follows from	$(P \dashv Q)$	$\bigcup_{r,s} \left\{ [\![(\langle p,q\rangle, s)\uparrow^{n\times m} \mid \begin{array}{c}(\langle p,r\rangle, s) \in^n \mathcal{V}[\![P]\!](\mathsf{UPop}) \;\wedge \\ (\langle r,q\rangle, t) \in^m \mathcal{V}[\![Q]\!](\mathsf{UPop})\end{array}]\!] \right\}$

  Deze expressies zijn bedoeld om er voor te zorgen dat in een multiset alle elementen de juiste verdelingen krijgen, al dan niet in combinatie met de expressie cfd.
  Echter kan men deze padexpressies te pas en te onpas gebruiken, met alle gevolgen van dien. Bij de definitie van deze padexpressies op bladzijde 86 is dat enigzins toegelicht. Die toelichting was gelijk de aanleiding om haakjes te gebruiken. Het is bijvoorbeeld mogelijk de expressie $(\mathsf{sv}\,P \dashv \mathsf{cfd}(x))$ te construeren, maar de expressie cfd is hier in feite nutteloos.
  Het liefst worden deze expressies niet gebruikt, maar dan gaat het propagatiemechanisme de mist in. In de expressie Z wordt veelvuldig in multisets de in de elementen voorkomende verdelingen vervangen.
- Varianten van de afsluiting ten behoeve van het propagatiemechanisme (bij expertsystemen):

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!](\mathsf{UPop})$
forward propagation	$P^{\vdash}$	$\mathcal{V}[\![\mathsf{ds}(\bigcup_{n\in\mathbb{N}} \mathsf{propforward}(P,n))]\!](\mathsf{UPop})$
backward propagation	$P^{\dashv}$	$\mathcal{V}[\![\mathsf{ds}(\bigcup_{n\in\mathbb{N}} \mathsf{propbackward}(P,n))]\!](\mathsf{UPop})$

  waarin

$$\begin{aligned} \mathsf{propforward}(P,0) &= P \;, \\ \mathsf{propforward}(P,n+1) &= (\mathsf{propforward}(P,n) \vdash P) \;, \end{aligned}$$

$$\begin{aligned} \mathsf{propbackward}(P,0) &= P \;, \\ \mathsf{propbackward}(P,n+1) &= (P \dashv \mathsf{propbackward}(P,n)) \;. \end{aligned}$$

Hier wordt voor alle duidelijkheid nog vermeld dat forward propagation alleen van nut is als de expressies van links naar rechts worden geëvalueerd, zoals dat in propositie 4.4.3 op bladzijde 111 is vastgelegd. Backward propagation is slechts van nut als evaluatie van padexpressies van rechts naar links geschiedt. Hierbij moet nog vermeld worden dat de wijze van evaluatie voor het propagatiemechanisme van wezenlijk belang is. Als men deze propagatiemechanisme voor andere doeleinden gebruikt, hoeft het principe van evaluatie niet persé vereist te zijn.

- Expressies die te gebruiken zijn om waarden uit frequentieverdelingen te halen ten behoeve van padexpressie operaties zoals sorteren (ordenen), groeperen, maar ook om ze te tonen (zonodig met confluentie operatie):

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!]\,(\mathsf{UPop})$
probability (of)	$\mathsf{Prob}\,Q$	$\{\!\![(\langle t[m],p\rangle,t)\uparrow^m \mid (\langle p,q\rangle,t)\in^m \mathcal{V}[\![Q]\!]\,(\mathsf{UPop})\}\!\!\}$
probability related	$\mathsf{Pr}\,Q$	$\mathcal{V}[\![(\mathsf{Prob}\,Q)^{\leftarrow}]\!]\,(\mathsf{UPop})$
measures	$\mathsf{Val}\,Q$	$\mathcal{V}[\![\mathcal{f}\,(\mathsf{Prob}\,Q)]\!]\,(\mathsf{UPop})$

- Expressies die in principe bedoeld zijn voor gebruik bij een expertsysteemmodel, in dit geval specifiek het zekerheidsfactormodel:

name	expr	$\mathcal{V}[\![\mathsf{expr}]\!]\,(\mathsf{UPop})$
confirmation	$\mathsf{confirm}\ P$	$\left\{\!\!\left[(\langle p,q\rangle,s)\uparrow^m \;\middle|\; \begin{array}{c}(\langle p,q\rangle,s)\in\mathcal{V}[\![P]\!]\,(\mathsf{UPop})\ \wedge\\ 0<s[m]\leq 1\end{array}\right]\!\!\right\}$
disconfirmation	$\mathsf{disconfirm}\ P$	$\left\{\!\!\left[(\langle p,q\rangle,s)\uparrow^m \;\middle|\; \begin{array}{c}(\langle p,q\rangle,s)\in\mathcal{V}[\![P]\!]\,(\mathsf{UPop})\ \wedge\\ -1\leq s[m]<0\end{array}\right]\!\!\right\}$

Het kan handig zijn dergelijke operaties bij information retrieval te gebruiken om bijvoorbeeld van alle relevante documenten die documenten te selecteren die minimaal relevant zijn (of iets dergelijks). De definitie van deze padexpressies verschilt dan natuurlijk. Bij information retrieval werkt men niet met zekerheidsfactoren.

5.4 Conclusies

Om met succes een model toe te passen in LISA-D is gebleken dat men in hoofdzaak met een aantal zaken rekening te houden heeft:

1. Een specificatietaal voor de functie fd. Het is noodzakelijk dat er een taal in LISA-D ingebed wordt, omdat bij de bepaling van de relevanties of zekerheidsfactoren deze functie onontbeerlijk is. Wat die taal is, is (nog) niet zo belangrijk omdat alle specificaties wiskundig zijn opgesteld;

2. De kennisrepresentatie. In de toepassing van het probabilistische afleidingsmodel in LISA-D is het tot nu nodig om ook de query in de informatiestructuur te modelleren en te populeren. Bij het zekerheidsfactormodel kunnen de productieregels mits omgevormd worden gemodelleerd en gepopuleerd in een informatiestructuur, maar er is behoefte aan originele regels voor doeleinden zoals het reconstrueren van het propagatieproces dat tot een bepaalde hypothese leidde. De gebruiksmogelijkheden van het zekerheidsfactormodel wordt door het gemis aan regels in originele vorm bemoeilijkt;

3. De wijze van evaluatie van padexpressies. Een voorschrift voor de wijze van evaluatie is dat als in een padexpressie naast een cfd-operatie ook een sv-operatie voorkomt, dat de sv-operaties het eerst worden uitgevoerd.(zie propositie 4.2.1 op bladzijde 88) Een tweede voorschrift dat alleen van toepassing is op het zekerheidsfactormodel is dat een padexpressie ven links naar

rechts diende te worden geëvalueerd. Dit is te behoeve van het geconstrueerde propagatiemechanisme die slechts onder deze aanname goed werkt. Theoretisch is het mogelijk, maar er zijn praktische bezwaren. Het is een beperking op de implementatie van LISA-D, en als men dit toch doorvoert, dient elk model er mee te werken;

4. De problematiek van definities van padexpressies met betrekking tot verdelingen. Bij zowel het zekerheidsfactormodel als het probabilistische afleidingsmodel levert de vereniging als padexpressie problemen op. Bij het tweede genoemde model is daar niet veel aan te doen, omdat de vereniging in het proces van bepaling van relevanties een belangrijke rol speelt. Bij het eerstgenoemde model hindert die vereniging de bepaling van de zekerheidsfactoren, ondanks het feit dat die vereniging daar wel nodig is. Het is niettemin wenselijk dat het probleem met de vereniging wordt opgelost, omdat verschillende padexpressies impliciet van de vereniging gebruik maken, zoals concatenatie, confluentie en afsluiting.
   Hier zij nog opgemerkt dat de vereniging niet de enige probleemexpressie is, maar het is wel de belangrijkste.

5.4.1 Toepassen van modellen: doen of niet

Theoretisch is het mogelijk om modellen uit expertsystemen toe te passen. Het belangrijkste onderdeel is het propagatiemechanisme. Het succes ervan hangt af van de wijze van evaluatie van padexpressies. Praktisch gezien is het een groot bezwaar. Volgens [HPW93] wordt er al gewerkt aan een implementatie van LISA-D en daar wordt nu een extra beperking aan opgelegd. Voert men de wijze van evaluatie toch door, dan zit men er aan vast. Elk model die men in LISA-D wil toepassen moet er dan maar mee werken.
Theoretisch kan gesteld worden dat het mogelijk is het zekerheidsfactormodel zoals deze nu is uitgewerkt, in LISA-D toe te passen is, maar het is geen goed idee.

Het is daarentegen zeer wel mogelijk om het probabilistische afleidingsmodel in LISA-D toe te passen, omdat dit model niet van de wijze van evaluatie afhangt (op de voorrangsregeling met betrekking tot sv en cfd na). Een probleem dat nog wel speelt (afgezien van de noodzaak van een taal in LISA-D), is hoe de query in het retrievalproces wordt betrokken. Momenteel is men genoodzaakt de query in de informatiestructur op te nemen en te populeren.

5.4.2 Suggesties voor verder onderzoek

Geen van de modellen die hier als testvehikel zijn gebruikt, is echt tot in de puntjes uitgewerkt. Zo ontbreken er enkele definities van padexpressies, maar dat heeft een reden. De omstandigheden waarin de niet-gedefinieerde padexpressies kunnen worden gebruikt, zijn niet bekend. Een advies is hier om een systeem waarin bijvoorbeeld het probabilistische afleidingsmodel wordt gebruikt, volledig in LISA-D uit te werken. Tegelijk kan dan gekeken worden naar het niveau van LISA-D queries. Of zo'n model met queries ook te gebruiken is, hangt af van het feit of men de nieuwe padexpressies in een query kan aangeven en dus gebruiken.

Het geconstrueerde basismodel voor uitbreiding van LISA-D is nu niet meteen hét perfecte model. Zo zou het gebruik van de expressie cfd in combinatie met een sv-operatie anders geregeld kunnen worden. Het is ook de vraag of het basismodel wel volledig genoeg is. Daarnaar moet dan verder onderzoek naar gedaan worden. Andere modellen dan het probabilistische afleidingsmodel en varianten ervan moeten op geschiktheid voor toepassing in LISA-D worden onderzocht. Modellen die een propagatiemechanisme vereisen, vallen af wegens het probleem met de wijze van evaluatie van padexpressies, tenzij men een propagatiemechanisme voor een model weet te verzinnen dat zonder die eis kan werken.

Er is een model dat niet eenmaal is beschouwd om in aanmerking te komen voor toepassing in LISA-D, maar die wel de grondslag is voor de meeste quasi-probabilistische modellen, en dat de kanstheorie. Dit is wel wezenlijk anders dan een model uit information retrieval of expertsystemen, maar daarom kan het nog wel mogelijk zijn dat het toepasbaar is in LISA-D, waarom ook niet?

Bibliografie

[Bar81] J.A. Barnett. Computational methods for a mathematical theory of evidence. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence IJCAI-81*, pages 868–875, Vancouver, BC, Canada, 1981.

[BG92] P.D. Bruza and L.C. van der Gaag. Index Expression Belief Networks for Information Disclosure. Technical Report 92-23, Department of Information Systems, University of Nijmegen, Nijmegen, The Netherlands, 1992.

[BHW91] P. van Bommel, A.H.M. ter Hofstede, and Th.P. van der Weide. Semantics and verification of object-role models. *Information Systems*, 16(5):471–495, October 1991.

[BS84] B.G. Buchanan and E.H. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, Massachusetts, 1984.

[CL68] C.K. Chow and C.N. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, IT-14(3):462 – 467, May 1968.

[DHN90] Richard O. Duda, Peter E. Hart, and Nils J. Nilsson. Subjective Bayesian methods for rule-based inference systems. In Glenn Shafer and Judea Pearl, editors, *Readings in uncertain reasoning*, chapter 5, pages 274–281. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.

[GS90] Jean Gordon and Edward H. Shortliffe. The Dempster-Shafer Theory of Evidence. In Glenn Shafer and Judea Pearl, editors, *Readings in uncertain reasoning*, chapter 6, pages 529–539. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.

[HPW92] A.H.M. ter Hofstede, H.A. Proper, and Th.P. van der Weide. Data Modelling in Complex Application Domains. In P. Loucopoulos, editor, *Proceedings of the Fourth International Conference CAiSE'92 on Advanced Information Systems Engineering*, volume 593 of *Lecture Notes in Computer Science*, pages 364–377, Manchester, United Kingdom, May 1992. Springer-Verlag.

[HPW93] A.H.M. ter Hofstede, H.A. Proper, and Th.P. van der Weide. Formal definition of a conceptual language for the description and manipulation of information models. *Information Systems*, 18(7), 1993.

[HW92] A.H.M. ter Hofstede and Th.P. van der Weide. Formalisation of techniques: chopping down the methodology jungle. *Information and Software Technology*, 34(1):57–65, January 1992.

[HW93] A.H.M. ter Hofstede and Th.P. van der Weide. Expressiveness in conceptual data modelling. *Data & Knowledge Engineering*, 10(1):65–100, February 1993.

[KP83] J.H. Kim and J. Pearl. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence IJCAI-83*, pages 190–193, Karlsruhe, West-Germany, 1983.

[LG91] P.J.F. Lucas and L.C. van der Gaag. *Principles of Expert Systems.* Addison-Wesley, Reading, Massachusetts, 1991.

[LS88] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. *The Journal of the Royal Statistical Society*, 50:157–224, 1988.

[LS90] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their applications to expert systems. In Glenn Shafer and Judea Pearl, editors, *Readings in uncertain reasoning*, chapter 6, pages 415–448. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.

[Maa88] H. Maassen. Stochastiek. A reader for students, Mathemathic Institute, Faculty of Mathemathics & Informatics, University of Nijmegen, Nijmegen, The Netherlands, 1988. In Dutch.

[Nor93] Norbert Fuhr. A Probabilistic Relational Model for the Integration of IR and Databases. *Association for the Computing Machinery - SIGIR'93*, June 1993.

[Par90] H. Partsch. *Specification and Transformation of Programs - a Formal Approach to Software Development.* Springer-Verlag, Berlin, Germany, 1990.

[Pea86] J. Pearl. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*, 29(3):241–88, 1986.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Palo Alto, California, 1988.

[Pea90] J. Pearl. Fusion propagation, and structuring in belief networks. In Glenn Shafer and Judea Pearl, editors, *Readings in uncertain reasoning*, chapter 6, pages 366–414. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.

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

[SB90] Edward H. Shortliffe and Bruce G. Buchanan. A Model of Inexact Reasoning in Medicine. In Glenn Shafer and Judea Pearl, editors, *Readings in uncertain reasoning*, chapter 5, pages 259–273. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.

[Sha76] G. Shafer. *A Mathemathical Theory of Evidence.* Princeton University Press, 1976.

[TC90] Howard Turtle and W. Bruce Croft. Inference Networks for Document Retrieval. In Jean-Luc Vidick, editor, *Proceedings of the International Conference on Research and Development in Information Retrieval*, volume 13, pages 1–24, 11 West 42nd Street, New York, NY 10036, 1990. The Association of the Computing Machinery.

[Vel84] Wim Veldman. Introductie tot de logica. A reader for students, Mathemathic Institute, Faculty of Mathemathics & Informatics, University of Nijmegen, Nijmegen, The Netherlands, 1984. In Dutch.

[WHB92] Th.P. van der Weide, A.H.M. ter Hofstede, and P. van Bommel. Uniquest: Determining the Semantics of Complex Uniqueness Constraints. *The Computer Journal*, 35(2):148–156, April 1992.

[WY91] S.K.M. Wong and Y.Y. Yao. A probabilistic inference model for information retrieval. *Information Systems*, 16(3):301–321, 1991.

Auteur Index

Index

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z