De evaluatie van kennissytemen.

P.J.F. Lucas

UU-CS-1998-32
September 1998

ISSN: 0924-3275

De Evaluatie van Kennissystemen*

Peter Lucas
Informatica-instituut, Universiteit Utrecht
Padualaan 14, 3584 CH Utrecht
E-mail: lucas@cs.uu.nl

Samenvatting

De evaluatie van kennissystemen behoort een integraal onderdeel te zijn van het ontwikkeltraject van kennissystemen. Vooral aan het einde van het ontwikkeltraject is evaluatie van belang, omdat inzicht in de kwaliteit van het onderliggende kennismodel noodzakelijk is, voordat het systeem kan worden ingevoerd in de organisatie. Momenteel krijgt de evaluatie van kennissystemen niet die aandacht waar het onderwerp, gegeven de practische belangrijkheid ervan, recht op heeft.

1 Inleiding

Een *kennissysteem* wordt gewoonlijk gedefinieerd als een computersysteem, waarin menselijke kennis is vastgelegd en kan worden toegepast voor het oplossen van problemen, op een wijze die min of meer lijkt op het menselijk redeneren. De kennis is vastgelegd in een *kennisbank*; probleemoplossen geschiedt met behulp van een verzameling *redeneermethoden*, meestal aangeduid als de *inference engine* [Lucas & Van der Gaag, 1991].

De kennisbank van een kennissysteem kan het beste opgevat worden als een *model* van de werkelijkheid. Deze werkelijkheid kan bijvoorbeeld de vakkennis van een specialist in een bepaald domein betreffen, maar kan ook afgeleid zijn uit de literatuur, of uit gegevens in een database door toepassing van methoden voor machinaal leren. Ook een specificatie van de functionaliteit van een fysisch, chemisch of biologisch process kan als basis dienen voor een kennisbank. Een kennissysteem wordt tegenwoordig vrijwel altijd volgens een bepaalde methodiek ontwikkeld, waarbij de nadruk in toenemende mate is komen te liggen op de ontwikkeling van kennismodellen [Wielinga et al., 1992]; de ontwikkeling van een kennissysteem is daarom tegenwoordig meer een modelleer- dan een programmeeractiviteit.

Aan het einde van het ontwikkeltraject is het gewoonlijk noodzakelijk te onderzoeken of het systeem voldoet aan vooraf gestelde eisen, dat wil zeggen, het systeem moet worden *geëvalueerd*. In dit artikel zal de huidige stand van zaken met betrekking tot de evaluatie van kennissystemen worden besproken.

2 Terminologie: verificatie en validering

In de literatuur over de evaluatie van kennissystemen wordt meestal aangesloten bij de terminologie die in de software engineering wordt gebruikt. Sommige onderzoekers hebben echter

*Verschenen in: *Informatie: maandblad voor de informatievoorziening*, juni 1997, jaargang 39, pp. 32–37.

ook terminologieën voorgesteld die hiervan verschillen (zie bijvoorbeeld [Hoppe & Meseguer, 1993], [Guida & Mauri, 1993] en [Mengshoel, 1993]); meestal wordt wel een onderscheid gemaakt tussen 'verificatie' en 'validering', net zoals in de software engineering [Boehm, 1979]. Hoewel de kenmerken van kennissystemen in een aantal opzichten verschillen van die van andere programmatuur, lijkt het niet verstandig al te zeer af te wijken van de terminologie in de software engineering. Om deze reden zal hierbij in dit artikel worden aangesloten.

Verificatie heeft betrekking op de structuur en logische correctheid van programmatuur, en op de vergelijking van het gerealiseerde systeem met de systeemspecificatie. In het geval van een kennissysteem zal de verificatie zich vooral richten op de kennisbank; de inference engine is immers meestal standaard gereedschap, zoals geboden door een expert system shell. *Validering* betreft het onderzoek of het systeem voldoet aan bepaalde eisen vanuit de toepassing. Bij een kennissysteem zal de validering vooral betrekking hebben op het probleemoplosgedrag van het systeem.

Verder wordt vaak een onderscheid gemaakt tussen *dynamische* en *statische* verificatie en validering [Sommerville, 1992]. In tegenstelling tot dynamische verificatie en validering, wordt een kennissysteem bij statische verificatie en validering niet uitgevoerd. In dit artikel zullen we vooral ingaan op de validering van kennissystemen, omdat dit onderwerp practisch gezien van groter belang is dan verificatie.

Evaluatie van een kennissysteem kan zowel gedurende de ontwikkeling van het systeem plaatsvinden, als na de afronding van de implementatie. Op deze laatste vorm van evaluatie zal in dit artikel de nadruk liggen, maar de andere vorm van evaluatie, die een rol speelt bij de verfijning van een kennissysteem is niet onbelangrijk [Lucas, 1994].

De bovengenoemde aandacht voor modellen deelt het vakgebied der kennissystemen met de natuurwetenschappen. Traditioneel bieden de vakgebieden van de statistiek en simulatie een grote variëteit aan methoden en wiskundige gereedschappen voor de evaluatie van modellen [Kreyszig, 1970; Shannon, 1975]. Het is dan ook niet verwonderlijk dat naast methoden en technieken uit de software engineering, een deel van de evaluatiemethoden, en dus ook de bijbehorende terminologie, afkomstig is uit de statistiek en simulatie.

3 De noodzaak tot evaluatie

De meeste programmatuur wordt voor de invoering in een organisatie aan een bepaalde vorm van evaluatie onderworpen. De complexiteit van zo'n evaluatie kan sterk wisselen [Boloix & Robillard, 1995]. Zo zullen sommige aspecten van een informatiesysteem dat informatie kan verschaffen over de beschikbaarheid van zitplaatsen in een vliegtuig moeilijk te evalueren zijn, bijvoorbeeld de afhandeling van gelijktijdig ingevoerde reserveringen, terwijl andere aspecten relatief eenvoudig te evalueren zijn. Zo zal er geen onenigheid bestaan over de voorwaarden voor het al dan niet bezet zijn van een zitplaats. Hoewel uitputtend testen van zo'n informatiesysteem onuitvoerbaar is, bestaat er in ieder geval overeenstemming over het al dan niet correct zijn van de informatie die het systeem in concrete gevallen verschaft.

Bij kennissystemen is de situatie echter aanzienlijk complexer. Beschouw, bijvoorbeeld, een medisch kennissysteem dat ondersteunt bij het stellen van een diagnose. Over de juistheid van het model dat als basis heeft gediend voor het kennissysteem, kunnen de meningen van medisch specialisten sterk uiteenlopen. Ook de diagnostische conclusies van het systeem zullen niet altijd volledig in overeenstemming zijn met de conclusies van een medisch specialist; medisch specialisten kunnen onderling sterk verschillen in hun diagnostische opvattingen.

Analoge situaties doen zich voor in andere probleemgebieden. Om inzicht te verkrijgen in de betrouwbaarheid van het advies van een kennissysteem is het derhalve van essentieel belang dat het systeem systematisch wordt geëvalueerd, alvorens het in gebruik wordt genomen. Bij zo'n evaluatie zullen diverse aspecten moeten worden onderzocht, niet alleen van het systeem, maar ook van de organisatie waarin het systeem zal worden ingezet. Een kennissysteem dat niet systematisch is geëvalueerd, is per definitie onbetrouwbaar.

Helaas krijgt het onderwerp 'evaluatie van kennissystemen' niet altijd de aandacht die het verdient. In methodieken voor de ontwikkeling kennissystemen, zoals CommonKADS, wordt veel aandacht besteed aan de ontwikkeling van kennismodellen [Schreiber et al., 1994]. Ten onrechte wordt echter vrijwel geen aandacht besteed aan methoden voor de evaluatie van kennissystemen. Veel onderzoekers lijken de mening toegedaan dat evaluatie overbodig wordt, zodra voor een goede methodiek van modelvorming is geopteerd. In de natuurwetenschappen, waar men goed op de hoogte is van de gevaren van blind geloof in eigen modellen, zou men zo'n opvatting ongetwijfeld verontwaardigd van de hand wijzen. Helaas blijkt dat ook bij veel bedrijven die gespecialiseerd zijn in de ontwikkeling van kennissystemen, evaluatie geen hoge prioriteit heeft.

4 Wie zijn bij een evaluatie betrokken?

Bij elke evaluatie zijn verschillende personen betrokken, die elk een aparte rol in het evaluatieproces vervullen:

- Deskundigen zijn eerst betrokken bij de modellering, en later bij een eerste grove evaluatie. Een groep deskundigen die niet betrokken is geweest bij de ontwikkeling van het systeem, kan een rol spelen in de uiteindelijke evaluatie.
- De toekomstige gebruikers van het systeem zullen een oordeel uitspreken over de practische bruikbaarheid van het systeem.
- De informaticus is geïnteresseerd in de nauwkeurigheid van het model, en hoe het systeem door de gebruiker optimaal benut kan worden.

Voor de informaticus is het belangrijk de verschillende rollen van deskundigen en gebruikers bij de evaluatie niet uit het oog te verliezen. Zo mogen bij het project betrokken deskundigen wel een uitspraak doen over de kwaliteit van de kennis in de kennisbank, maar hun oordeel over de toepasbaarheid van het systeem is van weinig belang. Bij dit laatste aspect is juist het oordeel van de gebruikers van veel groter belang.

5 Welke aspecten kunnen worden geëvalueerd?

Er is een groot aantal verschillen aspecten van een kennissysteem die onderwerp van een evaluatie kunnen zijn. Enkele aspecten, per onderwerp gegroepeerd, zijn opgenomen in Tabel 1 (zie ook [Wyatt & Spiegelhalter, 1990]). We zullen elk van deze aspecten hier kort aanstippen; de belangrijkste onderwerpen worden in de volgende paragraaf uitgebreider besproken.

Zowel deskundigen, literatuur als gegevens in een database spreken elkaar in bepaalde vakgebieden vaak tegen. Het is derhalve noodzakelijk inzicht te verkrijgen in de mate van deskundigheid van een 'deskundige', en de betrouwbaarheid van bepaalde literatuurbronnen

Modellen
betrouwbaarheid van kennisbronnen (deskundige, literatuur, database) juiste weergave van de werkelijkheid?
Kennisbank
volledigheid van de representatie van het model juiste kennisrepresentatieformalisme logische correctheid/consistentie
Redeneermethoden
coherentie van redeneringen mogelijkheid voor uitleg
Systeem
adequaatheid gebruikersinterface integratie met de omgeving gebruikersvriendelijkheid snelheid/geheugengebruik mate van ondersteuning bij probleemoplossen

Tabel 1: Aspecten voor de evaluatie van een kennissysteem.

of gegevens. Dit is geen eenvoudige opgave; in de volgende paragraaf zullen we hier verder aandacht aan besteden.

De mate waarin een model een juiste weergave is van de werkelijkheid kan pas na de implementatie van het kennissysteem systematisch worden onderzocht. Door echter voorafgaand aan de implementatie van een model als kennisbank nauwkeurig aandacht te besteden aan de structuur van het model, is het mogelijk enig inzicht te verkrijgen in de betrouwbaarheid ervan.

Bij de vertaling van een kennismodel naar de formalismen die geboden worden door een ontwikkelgereedschap, kan het voorkomen dat het meest geschikte formalisme voor het vastleggen van de kennis niet door het systeem geboden wordt. Bepaalde, mogelijk wezenlijke, kenmerken van het probleemdomein kunnen dan ontbreken in de uiteindelijke kennisbank. Hoewel formalismen in huidige gereedschappen meestal een grote uitdrukkingskracht bezitten, zodat dit probleem zich niet erg vaak voordoet, maakt dit wel duidelijk dat de huidige praktijk in het bedrijfsleven om één gereedschap tot standaard te verheffen, onbevredigende resultaten kan opleveren.

Onderzoek van de logische eigenschappen van een kennisbank wordt meestal niet ondersteund door gereedschappen. In de literatuur zijn wel enkele gereedschappen beschreven die gebruikt kunnen worden voor het opsporen van redundanties, ongewenste circulariteiten en potentiële bronnen van inconsistentie in een kennisbank [Perkins & Laffey, 1989; Polat & Guvenir, 1993]. Deze hulpmiddelen worden echter op dit moment niet op ruime schaal gebruikt. Uiteraard kunnen door nauwkeurige inspectie, alsmede door middel van dynamische verificatie van een kennisbank door verwerking van testproblemen, sommige semantische problemen opgespoord worden. Absolute zekerheid dat een kennisbank consistent is voor alle mogelijke invoergegevens, kan men zo uiteraard niet verkrijgen.

Hoewel vaak gebruik wordt gemaakt van standaard geboden redeneermethoden, kan het toch zinvol zijn de wijze van redeneren te evalueren. Immers, incorrecte redeneringen hoeven

niet altijd aanleiding te geven tot een incorrecte conclusie, maar het is wel waarschijnlijk dat zo'n redenering tot ongewenste conclusies aanleiding geeft. Als het redeneren ten bate van uitleg aan de gebruiker wordt gehanteerd, is het noodzakelijk dat het redeneergedrag door zowel deskundige als gebruiker op mate van coherentie wordt geëvalueerd.

Meer in het algemeen zal de gebruiker bepaalde wensen hebben ten aanzien van de gebruikersinterface, integratie met de omgeving, snelheid in gebruik en maximaal geheugengebruik. In al deze aspecten bestaat geen duidelijke verschil tussen kennissystemen en andere programmatuur. De mate waarin een systeem hulp biedt bij het oplossen van problemen vereist wel extra aandacht bij kennissystemen. In de volgende paragraaf zullen we hier uitgebreider op in gaan.

6 Studie-ontwerp

De centrale doelstelling van de validering van een kennissysteem is het verkrijgen van inzicht in de juistheid van het onderliggende kennismodel, gegeven de werkelijkheid die gemodelleerd is. Zoals hierboven al werd aangegeven, is het meestal niet mogelijk om dit in absolute zin vast te stellen [Gaschnig et al., 1983]. Veel kennissystemen worden toegepast als adviessysteem; de validering van zo'n kennissysteem zal meestal neerkomen op het vaststellen van de mate van overeenstemming van het advies dat het kennissysteem genereert met het advies van deskundigen. Wij zullen in dit verband spreken van de evaluatie van de *prestaties* van het systeem.

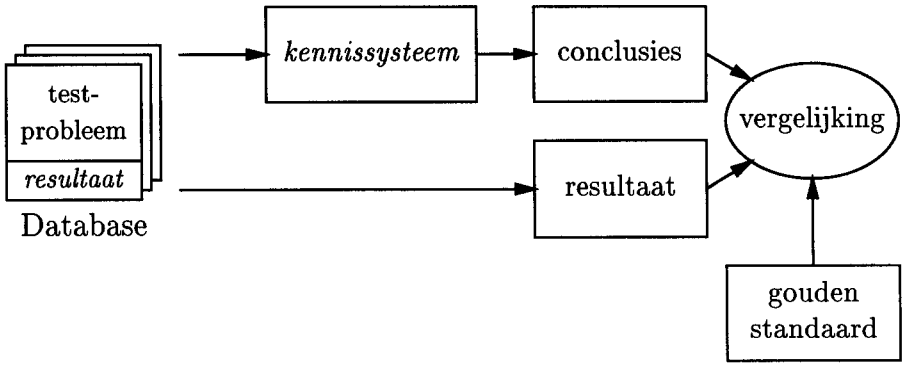
Uiteraard zal er bij een goed-uitgevoerde evaluatie altijd voor gezorgd worden dat de betrokkenen geen invloed kunnen uitoefenen op de prestaties van het systeem. Eigenlijk zou de evaluatie daarom niet door de ontwikkelgroep mogen geschieden; de deelnemers in de ontwikkelgroep zijn immers te goed op de hoogte van de inhoud van het systeem. Door echter zorgvuldig te werk te gaan, kan ongewenste beïnvloeding zoveel mogelijk voorkomen worden, ook als de ontwikkelgroep verantwoordelijk is voor de evaluatie; al dan niet bewuste vormen van fraude zijn dan niet helemaal uit te sluiten.

Het advies dat als vergelijkingsmateriaal wordt gebruikt, kan direct afkomstig zijn van één of meer deskundigen, beschreven staan in de literatuur, of vastgelegd zijn in een database. Dit advies wordt bij de validering gebruikt als referentiepunt; men spreekt in dit verband wel van een '*gouden standaard*'. Zoals hieronder zal worden besproken, kan het verkrijgen van inzicht in de hardheid van de 'gouden standaard' ook onderdeel zijn van studie. In deze paragraaf zullen enkele typische methoden voor validering worden besproken.

6.1 Laboratoriumvalidering

Bij een *laboratoriumvalidering* wordt het kennissysteem met opzet buiten de situatie, waarin het uiteindelijk gebruikt zal worden, geëvalueerd. Dit biedt de mogelijkheid de invloed van bepaalde parameters op de conclusies van het systeem, geïsoleerd van andere parameters, te onderzoeken. Een laboratoriumvalidering kan noodzakelijk zijn, indien het systeem schade kan veroorzaken in het geval van een incorrect uitgebracht advies. In de geneeskunde en luchtvaart, maar ook in andere toepassingsgebieden, kunnen dit soort overwegingen van belang zijn.

Bij de meting van de prestaties van een kennissysteem wordt gewoonlijk gebruik gemaakt van een database met probleemgevallen, waarbij voor elk probleemgeval een advies of resultaat in de database is opgenomen of wordt toegevoegd. Dit resultaat kan vergeleken worden met



Figuur 1: Evaluatie van de prestaties van een kennissysteem.

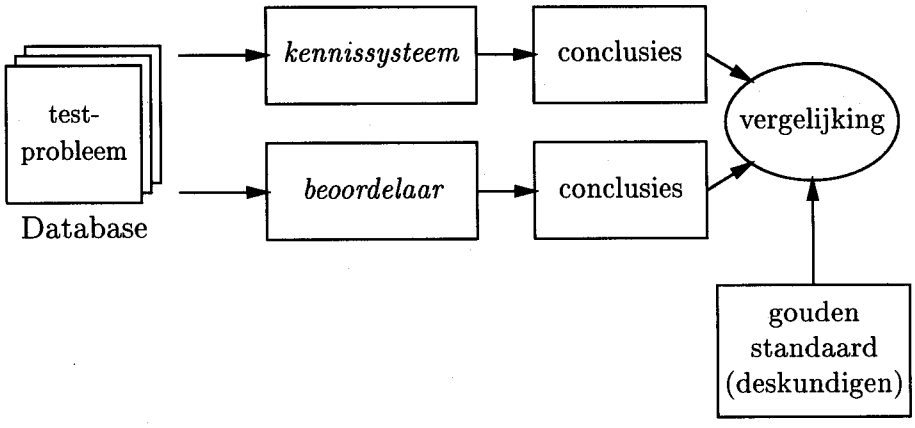
de conclusies van een kennissysteem, zoals schematisch weergegeven is in Figuur 1. De mate van overeenstemming wordt vastgesteld met behulp van een gouden standaard. Indien de evaluatie plaatsvindt met behulp van een al bestaande database, die niet speciaal voor dit doel is ontworpen en ingevuld, spreekt men van een *retrospectieve* evaluatie. Indien een database speciaal ten bate van de evaluatie is samengesteld, spreekt men van een *prospectieve* evaluatie. In sommige gevallen is het mogelijk testproblemen automatisch te genereren. In veel domeinen zal het meestal noodzakelijk zijn een deskundige te vragen een gegenereerd probleem van bijbehorend advies te voorzien.

Databases die bij retrospectieve evaluaties worden gebruikt, blijken verre van ideaal te zijn, omdat meestal gegevens die het systeem nodig heeft voor het trekken van zinvolle conclusies ontbreken, of omdat bepaalde gegevens onvoldoende precies zijn vastgelegd. Databases waarin problemen met vliegtuigen en acties om deze problemen te verhelpen zijn opgeslagen, zijn hiervan voorbeelden. De technische afdelingen van luchtvaartmaatschappijen maken intensief gebruik van dit soort databases; het zijn echte 'gebruiksdatabases' die slechts toepasbaar zijn voor het doel waarvoor ze ontworpen zijn. Deze databases bevatten veel ongestructureerde tekst, waarin de voor evaluatie benodigde informatie min of meer verborgen zit. Een vertaling met de hand naar een meer gestructureerde vorm is noodzakelijk om de database te kunnen gebruiken voor evaluatiedoeleinden.

Bij een prospectieve evaluatie kunnen de bovengenoemde problemen voorkomen worden. Een prospectieve evaluatie vereist echter veel tijd en extra inzet van personeel.

Van een betrouwbaar kennissysteem wordt verwacht dat het niet volstrekt andere adviezen geeft als de invoer van het systeem in zeer geringe mate verandert. Dit kan met behulp van *sensitiviteitsanalyse* worden onderzocht [Gaschnig et al., 1983], een techniek die vooral in de operations research en de mathematische besliskunde toegepast wordt [Shannon, 1975]. Vooral voor kennissystemen die een basis hebben in de kansrekening, zoals probabilistische-netwerk systemen, [Lucas & Van der Gaag, 1991], zal deze techniek bruikbaar zijn.

Wanneer een vergelijking gemaakt wordt tussen de voor probleemgevallen gegeven conclusies, en de conclusies van een kennissysteem, ontbreekt informatie over de prestaties die van het systeem verwacht mogen worden. Het is meestal onredelijk te verwachten dat een kennissysteem goed presteert in situaties waarin zelfs specialisten slecht presteren. Zelfs als van het kennissysteem betere prestaties verwacht worden dan van specialisten in het betreffende domein, is het noodzakelijk inzicht te verwerven in de prestaties van een specialist. Methoden voor kruisvalidering bieden hiervoor een oplossing. Bij *kruisvalidering* wordt het advies van het kennissysteem, aangevuld met de adviezen van deskundigen, blind beoordeeld



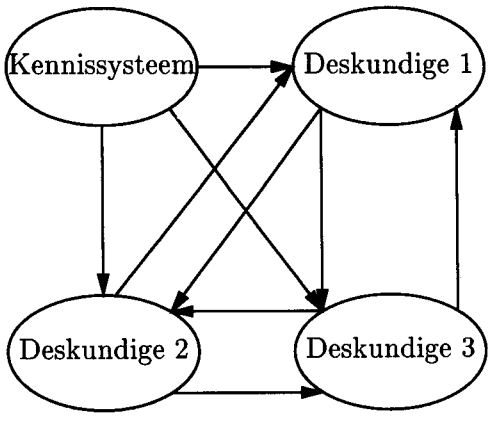
Figuur 2: Kruisvalidering van de prestaties van een kennissysteem.

door de deskundigen zelf (zie Figuur 2). Soms laat men de adviezen door een onafhankelijke groep van deskundigen beoordelen. In dat geval is het wenselijk dat de beoordelaars een verschillende mate van deskundigheid bezitten, zodat de invloed van de deskundigheid op de betrouwbaarheid van het advies bepaald kan worden. Dit is de zogenaamde 'inter-observer variabiliteit'. MYCIN, [Yu et al., 1979], en ONCOCIN, [Hickam et al., 1985], zijn twee systemen die een kruisvalidering ondergaan hebben. Kruisvalidering van een kennissysteem is door de grote hoeveelheid tijd en energie die geïnvesteerd moeten worden, in de praktijk slechts zelden uitgevoerd. De complexiteit van kruisvalidering blijkt uit Figuur 3, waarin zichtbaar wordt gemaakt welke adviesbronnen door een groep van drie deskundigen moeten worden beoordeeld, waarbij het aantal oordelen vermenigvuldigd moet worden met het aantal testgevallen.

In plaats van kruisvalidering kan ook gebruik gemaakt worden van de in de jaren 1960 bij de Rand Corporation ontwikkelde Delphi procedure [Shannon, 1975]. Bij de *Delphi procedure* krijgt elk lid van een panel van deskundigen hetzelfde probleem voorgelegd; hierover moet elk panellid zich een opinie vormen. Vervolgens wordt elke deskundige voorzien van de anonieme samenvattingen van de opinies van de medeleden van het panel. Elke deskundige wordt vervolgens verzocht een nieuwe opinie over het probleem te geven. In het ideale geval zullen de opinies van de deskundigen na enkele herhalingen convergeren. In de praktijk blijkt het nogal eens voor te komen dat er geen sprake is van convergentie naar een unieke, gemeenschappelijk oordeel. Indien de Delphi procedure wordt toegepast voor de validering van de prestaties van een kennissysteem, zullen de conclusies van het kennissysteem door het panel worden beoordeeld; het eindoordeel van het panel kan dan worden opgevat als het evaluatieresultaat.

6.2 Veldtest

Bij een *veldtest* wordt een kennissysteem geëvalueerd in de omgeving waarin het uiteindelijk zal worden ingezet. Meestal zal een veldtest voorafgegaan worden door een laboratoriumvalidering, zodat al voldoende inzicht verkregen is in de mogelijke prestaties van het systeem. Bij veel toepassingen is het ook dan echter niet verantwoord het systeem direct in te zetten in het productieproces. In dat geval kan men ervoor kiezen het systeem parallel te laten werken aan het reguliere productieproces, hetgeen, uiteraard, nogal wat extra inzet van personeel zal vereisen. De adviezen van het kennissysteem worden dan wel geregistreerd, en later



Figuur 3: Opzet van kruisvalidering.

gecontroleerd, maar niet opgevolgd.

Bij een goed opgezette veldtest is men ook geïnteresseerd in de vraag of het kennissysteem de gewenste ondersteuning biedt. Hiertoe is het noodzakelijk metingen te verrichten aan de prestaties van het personeel. De interpretatie van de uitkomsten van dit type onderzoek is moeilijk, aangezien de uitkomsten door diverse factoren kunnen worden beïnvloed. We zullen enkele van deze factoren kort bespreken (zie ook [Wyatt, 1992; Friedman & Wyatt, 1997]).

Het *Hawthorne effect* is een belangrijke factor om rekening mee te houden. Het is genoemd naar de Hawthorne fabriek in Chicago, waar onderzoekers in de jaren 1930 vaststelden dat de productiviteit verbeterde door het personeel bij feller licht te laten werken. Toen de verlichting bij toeval op een dag beneden het oorspronkelijke niveau was ingesteld, bleek de productiviteit ook te verbeteren. De uiteindelijke conclusie was dat de productiviteit alleen verbeterd was omdat het fabriekspersoneel wist dat hun prestaties werden gemeten [Roethligsburger & Dickson, 1939]. Een analoge factor kan de uitkomsten van een veldtest van een kennissysteem beïnvloeden, zodat de resultaten van de evaluatie met zorgvuldigheid moeten worden geïnterpreteerd.

Het *check-list effect* is een invloed die optreedt doordat gebruikers door het gebruik van het systeem expliciet gaan letten op de beschikbaarheid van bepaalde informatie. Dit effect alleen kan al een verbetering van de prestaties teweeg brengen.

Met het *carry-over effect* wordt het leereffect bedoeld dat na enige tijd optreedt bij gebruikers van een kennissysteem. Het systeem kan de gebruiker als het ware leren hoe een probleem moet worden opgelost, waardoor de prestaties van het personeel, ook als het systeem niet meer gebruikt wordt, verbeteren.

[De Dombal et al., 1991] geeft een beschrijving van een uitgebreid veldonderzoek, waarin het nut en de effecten op de gezondheidszorg van de toepassing van een kennissysteem voor de diagnose van buikklachten worden besproken. De effecten van de invoering van een kennissysteem waren nooit eerder op zo'n grote schaal onderzocht; tot op heden is niemand erin geslaagd deze prestatie te evenaren.

7 Meting van accuratesse

Veel methoden voor de meting van de prestaties van kennissystemen zijn afgeleid van vroeg werk op het gebied van probabilistische systemen [Habbema et al., 1978]. Het kennissysteem

wordt in dat geval als een blackbox beschouwd, waarbij gemeten wordt wat de prestaties van het systeem zijn in termen van het percentage correcte, incorrecte en niet geclassificeerde testproblemen; tevens worden de conclusies door het systeem getrokken worden, geclassificeerd als correct of incorrect [O'Keefe et al., 1987; Indurkhya & Weiss, 1989]. Bij de presentatie van de resultaten van een meting, moeten de resultaten voor individuele conclusies worden gepresenteerd, en niet slechts samenvattende resultaten, waarmee in de praktijk vaak ten onrechte volstaan wordt. Eventueel kan zo'n meting laagsgewijs plaatsvinden, waarbij telkens maar een deel van de beschikbare informatie aan het kennissysteem wordt aangeboden [Lucas, 1994]. Dit geeft een indruk van hoe goed het systeem in staat is om te gaan met onvolledige informatie.[1] Met behulp van statistische technieken, waarop in dit artikel niet verder wordt ingegaan, kan een indruk worden gekregen van de statistische spreiding van de resultaten van de meting [Kreyszig, 1970; O'Keefe et al., 1987].

Uiteraard is de toepassing van statistische technieken bij evaluatie alleen zinvol als het aantal testproblemen voldoende groot is. Het komt in de praktijk zeer vaak voor dat aan deze voorwaarde niet voldaan is.

8 Conclusies

De evaluatie van kennissystemen is op dit moment een ondergewaardeerd onderwerp. Er wordt vooral te weinig aandacht geschonken aan het practisch zo belangrijke onderwerp van validering. Een apart probleem is dat de technieken die bij de evaluatie van kennissystemen gebruikt worden, verspreid zijn over diverse disciplines: software engineering, statistiek, simulatie, psychologie en bepaalde toepassingsgebieden zoals medische informatica. Het is daarom niet eenvoudig literatuur over het onderwerp te vinden. Bovendien wordt er in de diverse vakgebieden verschillend gedacht over de evaluatie van kennissystemen. Toch zal het noodzakelijk zijn om voor de ontwikkeling van betrouwbare kennissystemen meer aandacht te besteden aan de evaluatie van kennissystemen dan nu zowel in academische kring als in het bedrijfsleven gebeurt.

Referenties

[Boehm, 1979] B.W. Boehm (1979). Software engineering: R & D trends and defense needs. In *Research Directions in Software Technology* (P. Wegner, ed.). MIT Press, Cambridge, Massachusetts.

[Boloix & Robillard, 1995] G. Boloix and P.N. Robillard (1995). A software system evaluation framework. *IEEE Computer*, **28**(12), 17–26.

[De Dombal et al., 1991] F.T. de Dombal, V. Dallos and W.A. McAdam (1991). Can computer-aided teaching packages improve clinical care in patients with acute abdominal pain? *British Medical Journal*, **302**, 1495–1497.

[Friedman & Wyatt, 1997] C.P. Friedman and J.C. Wyatt (1997). *Evaluation Methods in Medical Informatics*. New York: Springer-Verlag.

[Gaschnig et al., 1983] J. Gaschnig, P. Klahr, H. Pople, E. Shortliffe and A. Terry (1983). Evaluation of expert systems: issues and case studies. In *Building Expert Systems* (F. Hayes-Roth, D.A. Waterman and D.B. Lennart, eds.). Cambridge, MA: Addison-Wesley, pp. 241–280.

[1] De effecten van onvolledige informatie kunnen problematisch zijn bij een niet-monotoon redenerend systeem.

[Guida & Mauri, 1993] G. Guida and G. Mauri (1993). Evaluating performance and quality of knowledge-based systemens: foundations and methodology. *Transactions on Knowledge and Data Engineering*, **5**(2), 204–224.

[Habbema et al., 1978] J.D.F. Habbema, J. Hilden. and B. Bjerregaard (1978). The measurement of performance in probabilistic diagnosis I: the problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine*, **17**, 217–226.

[Hickam et al., 1985] D.H. Hickam, E.H. Shortliffe and M.B. Bishoff et al. (1985). The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Annals of Internal Medicine*, **103**, 928–936.

[Hoppe & Meseguer, 1993] T. Hoppe and P. Mesequer (1993). VVT terminology: a proposal. *IEEE Expert*, **8**(3), 48–55.

[Indurkhya & Weiss, 1989] N. Indurkhya and S.M. Weiss (1989). Models for measuring performance of medical expert systems. *Artificial Intelligence in Medicine*, **1**, 61–70.

[Kreyszig, 1970] E. Kreyszig (1970). *Introduction to Mathematical Statistics: Principles and Methods.* New York: John Wiley & Sons.

[Lucas, 1994] P.J.F. Lucas (1994). Refinement of the HEPAR expert system: tools and techniques. *Artificial Intelligence in Medicine*, **6**(2), 175–188.

[Lucas & Van der Gaag, 1991] P.J.F. Lucas and L.C. van der Gaag (1991). *Principles of Expert Systems*. Wokingman: Addison-Wesley.

[Mengshoel, 1993] O.J. Mengshoel (1993). Knowledge validation: principles and practice. *IEEE Expert*, **8**(3), 62–68.

[O'Keefe et al., 1987] R.M. O'Keefe,O. Balci and E.P. Smith (1987). Validating expert system performance. *IEEE Expert*, **4**, 81–89.

[Perkins & Laffey, 1989] W.A. Perkins and T.J. Laffey (1989). Knowledge base verification. In *Topics in Expert System Design* (G. Guida and C. Tasso, eds.), 353–376. Amsterdam: North-Holland.

[Polat & Guvenir, 1993] F. Polat and H. A. Guvenir (1993). UVT: a unification-based tool for knowledge-base verification. *IEEE Expert*, **8**(3), 69–75.

[Roethligsburger & Dickson, 1939] F.J. Roethligsburger and W.J. Dickson (1939). *Management and the Worker.* Cambridge, MA: Havard University Press.

[Schreiber et al., 1994] A.Th. Schreiber, B. Wielinga, R. de Hoog, H. Akkermans and W. van der Velde (1994). CommonKADS: a comprehensive methodology for KBS development. *IEEE Expert*, **9**(6), 28–37.

[Shannon, 1975] R.E. Shannon (1975). *Systems Simulation: the art and the science.* Englewood Cliffs, NJ: Prentice-Hall.

[Sommerville, 1992] I. Sommerville (1992). *Software Engineering.* Wokingham: Addison-Wesley.

[Yu et al., 1979] V.L. Yu, B.G. Buchanan, E.H. Shortliffe, et al. (1979). An evaluation of the performance of a computer-based consultant. *Computer Programs in Biomedicine*, **9**, 95–102.

[Wielinga et al., 1992] B.J. Wielinga, A.Th. Schreiber and J.A. Breuker (1992). KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition*, **4**, pp. 5–53.

[Wyatt & Spiegelhalter, 1990] J. Wyatt and D.J. Spiegelhalter (1990). Evaluating medical expert systems: what to test for and how? *Medical Informatics*, **15**(3), 205–217.

[Wyatt, 1992] J. Wyatt (1992). The evaluation of medical decision aids: why, what, where and how? *American Medical Informatics Association Spring Congress*, Portland, Oregon.