Fuzzy-søgning i DanBib - et relevant søgealternativ?

Af Jesper Bøgeskov

Abstract

I artiklen gennemgås principperne i søgeteknikken fuzzy logic som klassificeres i.f.t. eksisterende eksakte- og partielle match søgeteknikker. Der redegøres for et konkret laboratorie-eksperiment ved DBC, hvor fuzzy logic er forsøgt implementeret i DanBib. Herefter foretages søgninger i dels DanWeb, dels Fuzzy DanBib ud fra en udvalgt målgruppes selvvalgte søgetermer/-kombinationer. Dokumentrepræsentationerne præsenteres dernæst for målgruppen, der relevansbedømmer posterne ud fra en skala fra 1-3. Med baggrund i disse resultater, kombineret med mine egne evalueringer af Fuzzy DanBib, konkluderes det, at fuzzy-prototypen rummer en række indlysende søgefordele specielt overfor brugere med bevidst emneafgrænsede eller mudret emneafgrænsede informationsbehov. Fuzzy logic ses således som et relevant søgealternativ til almindelig boolesk logik.

Indledning

Søgning i bibliografiske søgesystemer har traditionelt set, i kraft af databasernes systemtekniske opbygning, medført en række typiske problemer for både slutbrugere og intermediærer. Dette skyldes ikke mindst søgesystemernes opbygning omkring eksakt match princippet samt udbredte brug af det tekstbaserede kommandosprog, CCL[1]. Specielt slutbrugere har ofte under onlinesøgninger haft svært ved at formulere "korrekte" søgestrenge, som systemet har kunnet acceptere – ofte har resultatet derfor været en lakonisk besked om "nul hits". Uden den tilstrækkelige viden om systemets opbygning vil slutbrugeren ofte opfatte denne besked som ensbetydende med, at systemet ikke rummer nogle dokumenter, der omhandler det pågældende emneområde (hvilket selvfølgelig ikke behøver at være tilfældet). I sådanne tilfælde vil brugeren forlade biblioteket med et uopfyldt informationsbehov, hvis da ellers ikke vedkommende vælger at konsultere bibliotekaren på stedet i håb om hjælp. En større viden på området omkring systemets opbygning, indekseringspraksis, brug af deskriptorer/keywords, etc. ville klart kunne hjælpe slutbrugeren i en sådan situation. Det er dog naturligvis ikke rimeligt at forlange, at den menige slutbruger skal beherske en så specifik systemmæssig viden. Hensigten må derfor klart være, at systemudviklerne på sigt letter brugernes adgang til søgesystemerne.

Et eksempel i rækken af sådanne udviklingsprojekter er DanBibs Fuzzy-prototype, "Fuzzy-søgning i DanBib, ver. 2.0", der er et samarbejde mellem DBC og forskere ved RUC's ISL-institut [2]. Prototypen har siden maj 1996 været undervejs ved DBC i forskellige reviderede udgaver, og det

er et projekt, der har ambitioner om på længere sigt at kombinere en forenklet, lettilgængelig brugergrænseflade med, ikke mindst, betydelig lettere og bedre søgemuligheder for systemets brugere. Planen er, at de erfaringer omkring fuzzy logic, som DBC med tiden høster ved prototypen, på sigt skal implementeres i den almindelige DanBib/DanWeb som et forbedrende søgealternativ. Hvad begrebet fuzzy logic er, hvordan søgeteknikken og det tilhørende fuzzy-semantiske netværk af emneord er forsøgt implementeret i DanBib, har jeg tænkt mig at redegøre nærmere for i denne artikel. På baggrund af teoretiske studier af fuzzy logic, min egen evaluering af prototypen samt en mini-test, der inddrager to forsøgspersoners relative relevansvurderinger af dokumentrepræsentationer fremfundet i hhv. DanWeb samt Fuzzy DanBib, vil der afslutningsvis blive konkluderet, hvorvidt fuzzy logic ses som et relevant søgealternativ til DanBib.

Indeksering og repræsentation

For at brugere af informationssystemer skal kunne finde frem til et konkret dokument, der omhandler et specifikt emne, kræves det, at andre på forhånd (som det første) har udtrykt eller beskrevet dokumentet v.h.a. en række emner eller begreber og omsat analysen til en repræsentation – en såkaldt "dokumentrepræsentation". Dokumentrepræsentationen forbindes normalt med en bibliografisk beskrivelse, hvor der udover den emnemæssige repræsentation i form af emneord og/eller abstract, er inkluderet formelle oplysninger om titel, forfatter, forlag, etc. De optræder fortrinsvis som indførsler i trykte bibliografier, kataloger eller, mere almindeligt i dag, elektronisk tilgængelige i databaser, som f.eks. DanBib. Dokumentrepræsentationerne opfattes i denne sammenhæng som stedfortrædende for selve dokumenterne.

Typer af indekseringssprog

Til at udtrykke dokumentets emner anvendes et "indekseringssprog", et sæt betegnelser, der kan være verbale udtryk, f.eks. emneord, eller symbolsk udtrykt i notationer, f.eks. tal eller bogstaver. Det valgte indekseringssprog kan være enten *ukontrolleret,* såkaldt naturligt sprog, eller *kontrolleret* ud fra en vedtagen, standardiseret liste. De kontrollerede indekseringssprog kan være ustrukturerede (f.eks. alfabetiske emneordslister) eller strukturerede (f.eks. klassifikationssystemer og thesauri med semantiske, hierarkiske relationer anført mellem emneordene). Karakteristisk for disse typer indekseringssprog er, at de begge rummer fordele såvel som ulemper. Hvor det kontrollerede er forudsigeligt, præcist og entydigt, men stift, så er det ukontrollerede mere udtryksfuldt og fleksibelt, men samtidig ofte flertydigt p.g.a. manglende morfologisk kontrol, kontrol af synonymer, homonymer, semantiske relationer m.v.

Indekseringsstrategi

Indeksering baseres implicit på en indekseringsstrategi, der inkluderer bestemte rammer og retningslinjer. Strategien udarbejdes på baggrund af en analyse af flg. elementer:

- *Brugerne:* deres informations- og søgebehov, søgeerfaring, deres rutiner, emnekendskab og brug af sproget.
- *Materialerne:* fysiske og emnemæssige karakteristika, dybde og omfang.
- *Emneområdet:* kompleksitet, anvendelse af terminologi.
- *Informationssystemet:* registrerings-, søge- og outputfaciliteter, hjælpe og browsingfaciliteter, grad af automation m.v.

Ovennævnte strategi fastsættes med det formål, at opnå en ensartet og korrekt indeksering, der er tilpasset databasens formål og rammer. Den indeholder typisk en vejledning omkring emneanalyse og *exhaustivitet,* indekseringssprog og *specificitet,* grad af prækoordination, brug af links (relationer mellem deskriptorer, som er stærkere end simpelt "sammenfald i indekseringen"), rolleindikatorer og vægtning. Indekseringsstrategien kan på baggrund af emneanalysen tage udgangspunkt i henholdsvis en refererende "dokumentorienteret" såvel som en tilpasset "brugerorienteret" indeksering.

Indekseringsniveau

Indekseringsniveauet i et system med bibliografiske poster udtrykkes normalt ud fra begreberne

"exhaustivitet" og "specificitet". Exhaustivitet er et udtryk for, hvor dækkende og udtømmende indekseringen beskriver dokumentets emne. Der skelnes i denne forbindelse mellem indeksering af et dokuments synsvinkel (herunder sværhedsgrad, kvalitet, ideologi/paradigme, forskningsmetode) og emnernes vigtighed: a) høj vigtigheds tærskel: et emne indekseres kun, hvis det er grundigt behandlet i dokumentet, eller b) lav vigtigheds tærskel: et emne indekseres, når det optræder i dokumentet. Høj grad af exhaustivitet øger muligheder for høj recall, men svækker typisk precision. Begrebet specificitet, eller detaljeringsgraden, udtrykker indekseringssystemets evne til at dække dokumenters emne specifikt, hvilket i høj grad har sammenhæng med, hvor detaljeret indekseringssystemet er underdelt.

DBC's indekseringsstrategi

DBC's officielle indekseringsstrategi[3] er, at tilstræbe indeksering på dokumentniveau ved tildeling af emneord til de bibliografiske poster i DanBib. Et dokument om f.eks. nattergale vil således blive indekseret med dette ord, og ikke det mere overordnede emneord spurvefugle eller fugle. Der tildeles normalt ikke emneord på analyseniveau, d.v.s. i.f.t. enkelte afsnit eller kapitler i dokumentet. Det kan dog gøres i specielle tilfælde som supplerende emneord, herunder analytiske emneord (f.eks. ved et sjældent optrædende emneord i DanBib og slang-, kæle-, samt kaldenavne – alle emneord, der skønnes at have en søgemæssig interesse for brugerne (Indeksering af faglitteratur, 1998, s. 17-18). DBC's indekseringsstrategi med at indeksere så specifikt som muligt, hvis dokumentet tilsiger dette, ses der i visse situationer bort fra - i de tilfælde hvor indeksøren skønner, at overspecificitet vil indvirke negativt på databasens effektivitet.

DanBibs indekseringsniveau

DBC's retningslinjer for god indekseringspraksis præger ved gennemgang af forskellige DanBib-poster i overvejende grad de poster DBC selv har indekseret, d.v.s typisk folkebiblioteks-relaterede poster fra FOLK-delbasen. Omvendt spores der i posterne fra FORSK-delbasen en udpræget uensartethed i de bibliografiske repræsentationers katalogiseringsniveau. Typisk identificerede problemer m.h.t. forskningsbibliotekernes katalogisering er

- forskellige minimumsregler for bibliografiske data
- forskellige navneformer
- katalogisering af monografiserie som periodicum eller som enkeltbind
  (Sinding, 1998).

Der er flere grunde til forskningsbibliotekernes forskelligartede katalogiseringer og emneindekseringer i.f.t DBC-praksis. Den vigtigste grund er, at bibliotekerne som oftest anvender enten deres egne klassifikationssystemer, DK5-systemet eller UDK – sidstnævnte primært indenfor universitetsbibliotekerne og generelt biblioteker med teknisk eller naturvidenskabeligt materiale. Det er således almindeligt blandt disse biblioteker, at "berige" posterne med supplerende emneord, danske som udenlandske, f.eks. tilførsel af lokale stednavne m.m., målrettet bibliotekets specielle domæner eller brugere. Af samme grund optræder flere dokumenter i DanBib med to indførsler – DBC-indførslen foruden den berigede post. Graden af exhaustivitet kan svinge meget fra ét forskningsbibliotek til et andet. Mens et bibliotek typisk vil tildele 2 emneord i snit til et dokument, tildeler et andet bibliotek måske op til 8-10 emneord. Ofte benyttes endvidere en række kontrollerede såvel som ukontrollerede emneord. En sådan uensartet og inkonsistent indeksering er naturligvis ikke ønskværdig, og opfattes af mange DanBib-brugere som et stigende og irriterende problem. Man kan sige, at variationen i brugen af emneord forstærker problemet omkring brugerens forventning til databasens udfaldsrum – det er således yderst vanskeligt at vide, hvorvidt man i en emnesøgning har fået alle de relevante dokumenter med.

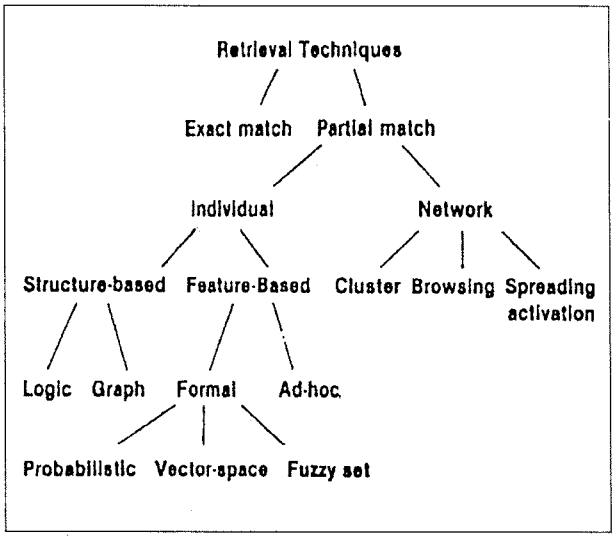
Fuzzy DanBibs beskrivelsesniveau

Fuzzy DanBibs dokumentrepræsentationer er meget sparsomme og samtidigt meget uensartede i beskrivelsesniveauerne. Posterne præsenteres obligatorisk med titel, emneord (DBC-kontrollerede, faglitterære emneord – i DanWeb søgbart med "ef=" eller "lef="), ID-nummer, databaseteknisk identifikator, samt i flere tilfælde forfatter, undertitel samt udgivelsesår. Baggrunden er bl.a., at det ikke har været DBC's primære intention at gøre prototypen, med sin nuværende bibliografiske præsentationsform, tilgængelig som søgealternativ. Det er principperne og de høstede erfarin-

ger ved fuzzy logic og emneords-fussyficeringerne, der her har været det mest interessante genstandsområde. Derfor det "fattige" beskrivelsesniveau, der beror på automatisk udtræk af bestemte katalogfelter fra DanBib-basen. En nærmere gennemgang af de forskellige typer søge- og matchteknikker vil nu blive foretaget.

Søgeteknikker

Søgeteknikker, eller genfindingsteknikkker (IR-teknikker) inddeles normalt i to overordnede kategorier, henholdsvis *eksakt* match- og *partial* matchteknikker, der adskiller sig væsentligt fra hinanden på en række punkter. Rent ydelses-/genfindingsmæssigt har forsøg gennem tiden påvist, at der er en tendens til, at søgning med partial matchteknikker giver bedre resultater end ved eksakt match - hermed menes forbedret recall og precision (Belkin & Croft, 1987, s. 124). Selvsamme forskere har i denne forbindelse konstrueret en anvendelig og overskuelig klassifikationsmodel for de forskellige søgeteknikker (fig. 1) – en model der siden hen er blevet et vigtigt referencepunkt for andre forskere og teoretikere indenfor informationsvidenskaben.



Figur 1: Klassifikation af søgeteknikker (Belkin & Croft, 1987, s. 112)

Når diskussionen falder på eksakt match vs. partial match, så er der grundlæggende en række forskellige systemtilgange/-opfattelser, der identificeres. Disse knytter sig til:

- Forskellige formodninger om relationen mellem brugerens request og query.
- Betydningen af eksakt sammenfald overfor delvist sammenfald mellem søgeformuleringen/forespørgslen og dokumentrepræsentationen.
- Diskussionen om humane vs. (halv-)automatiske intermediære funktioner – bl.a. set i forhold til partial match's grundidé med at ville undgå en human intermediær mekanisme p.g.a. inkonsistens i specielt referenceinterviewet (Ingwersen, 1992, s. 89).

I det følgende redegøres kort for de søgeteknikkers karakteristika, der primært har relevans for forståelsen og perspektiveringen af fuzzy logic.

Eksakt match

Ved eksakt match skal der være eksakt sammenfald mellem søgeformuleringen query og dokumentrepræsentationen. Matchteknikken benytter sig af de booleske operatorer (and, or, not – eller de danske oversættelser) til kombination eller fravalg af bestemte ord eller søgesæt. Man regner normalt "streng-søgning", "fuldtekst-søgning" og ikke mindst "boolesk søgning", som hørende under denne IR-teknik. Boolesk søgelogik er i dag den altoverskyggende implementerede søgeteknik i dagens kommercielle onlinedatabaser/kataloger – således også DanBib.

Boolesk logik
Princippet i boolesk logik er, at der tages udgangspunkt i en klassisk matematisk funderet sæt-teori, der involverer etableringen af systematiske relationer mellem elementer indenfor sættet, såvel som mellem elementer fra forskellige sæt. Karakteristisk for et sæt er, at dets enkelte elementer har fælles karakteristika - et tilfældigt element kan derfor enten "tilhøre" (udsagnet er sandt) eller "ikke tilhøre" (udsagnet er falsk) dette sæt. For at kunne udnytte denne viden i forbindelse med elektroniske bibliografiske søgesystemer, er man nødt til at omsætte udsagnet "sandt" eller "falsk" til binære tal, d.v.s. 0 (falsk) eller 1 (sandt) – computere vil ellers ikke kunne forstå disse udsagn om

medlemskab. Hvis vi overfører det til informationssøgning (IR), så gælder at for hver gang vi foretager en søgning, så dannes der et sæt af poster, med én eller flere karakteristika til fælles, f.eks. samme deskriptor. Den booleske søgelogik rummer en række klare fordele, foruden en del ulemper. Søgelogikken har gennem de sidste ca. 40 år været totalt dominerende ved de kommercielle databaseværter/systemproducenter - herunder Dialogs baser og DanBib. Årsagerne til dette er iflg. bl.a. Belkin & Croft (1987) flere, dels systemproducenternes uvilje mod nye investeringer og omlægninger af deres systemer, dels at den traditionelle målgruppe historisk set har været erfarne superbrugere (intermediærer eller forskere), tilfredse med systemerne.

Symptomatisk for boolesk logik og de booleske IR-systemer er, at de i mange år har afstedkommet en heftig debat af teknikkens fordele og ulemper – noget der gennem tiden har delt brugerne i to lejre.

Fortalerne fremhæver især flg. *fordele:*

- Mulighed for at kunne sammensætte sit søgeargument v.h.a. en kombination af to eller flere søgetermer eller søgesæt. Hermed opnår man at kunne sammensætte meget specifikke søgeargumenter, der afspejler bestemte aspekter af brugerens informationsbehov. Specielt hos brugere med en stor domæneviden har denne søgelogik sine fordele.
- Brug af højre-/venstretrunkeringer, maskeringer, liste/skan-funktioner, nærhedsoperatorer og intervalsøgningsmuligheder. Professionelle brugere prioriterer normalt disse funktionsområder meget højt, idet systemet hermed giver brugeren en række muligheder for virkelig at indsnævre sine søgninger til bestemte registre f.eks. DanBibs forfatter- eller titelregister. Anvendelsen kræver viden om de enkelte søgesystemers brug af syntaks og kommandosprog/søgekoder.

Modstanderne af den booleske logik understreger omvendt typisk flg. *ulemper:*

- Besværligt at formulere sine søgeforespørgsler "korrekt" – repræsentationerne, hhv. query og dokumentrepræsentationen, skal benytte det samme vokabular. Samtidig gør de fleste søgesystemer brug af tegnbaseret kommandosprog (CCL), der kræver specialiseret viden for at kunne udnytte søgesystemets muligheder fuldt ud. Konsekvensen er derfor ofte for slutbrugere, at intermediærer/bibliotekarer må bistå brugeren, der har det faktiske informationsbehov, i formulering af query. Hermed opstår potentielt en mulighed for fejltolkning af brugerens intention eller reelle informationsbehov.
- Søgeteknikken fremfinder ikke de dokumenter, som kun matcher query delvist. Booleske søgesystemer opdeler under søgninger databasen i to underdelte sæt: ét med de poster der 100% matcher query, og et andet der ikke gør. Man savner således typisk inddragelse af dokumenter i "gråzonen".
- For at kunne udnytte de booleske søgesystemer tilfredsstillende kræves brug af hjælpeværktøjer, såsom thesauri, emneordslister m.v.
- De enkelte søgetermer i query eller termer i dokumenterne vægtes ikke efter deres indbyrdes betydning og deraf relative vægt.
- Systemet rangordner ikke dokumenterne efter relevans – i visse søgesystemer rangordnes posterne kronologisk (f.eks. under Dialogs baser).
- Typisk opnås enten meget få (evt. slet ingen) eller alt for mange hits ved søgninger.

For at kunne forstå den booleske logik med dens indbyggede problemstillinger i et større perspektiv, er det nødvendigt at kigge på, hvordan dens logik og brug af kombinatorer egentlig matcher den menneskelige måde at tænke på – den menneskelige logik. Mange brugere, novicer som mere erfarne, har typisk problemer med at forstå den klare betydning med brugen af de booleske kombinatorer "og", "eller", "ikke". Selv fagligt veluddannede folk som f.eks. ingeniører og videnskabsfolk kommer fra tid til anden til at bytte rundt på kombinatorerne, når de benytter online databaser. Psykologer indenfor kognitionsforskningen har gennem tiden påvist, at mange mennesker/brugere misforstår de booleske operatorers virkning, fordi de ikke dagligt benytter disse logiske modeller (boolesk logik) i deres ræsonnementer, men snarere følger deres intuitive døm-

mekraft og tolker operatorerne i deres sproglige kontekst. Der er overordnet ingen tvivl om, at langt de fleste alm. brugere af booleske søgesystemer oplever og fremover vil opleve en række problemer i.f.m. deres søgninger. Der vil samtidig uden tvivl også være en række professionelle brugere, der ikke kan forestille sig andre måder at søge på i bibliografiske søgesystemer – set i lyset af deres tillærte viden om de booleske systemers opbygning og benyttede vokabular, mulighed for kombinatoriske søgninger, brug af trunkering/maskering, m.v. Med baggrund i gennemgangen af de eksakte matchteknikker, er det herefter naturligt at fortsætte med den anden kategori indenfor søgeteknikkerne, nemlig de partielle matchteknikker - gruppen af søgeteknikker, hvor bl.a. fuzzy sæt og fuzzy logic hører under.

Partial match

Partielle matchteknikker tager udgangspunkt i, at der ikke nødvendigvis skal (men kan) være eksakt sammenfald mellem query og dokumentrepræsentationen. Der kan ved formulering af request benyttes naturligt sprog, som v.h.a. stopordsliste, stemming og evt. automatiske søgethesauri omformes til en query, der sammenlignes med dokumentrepræsentationerne i basen. Der skal ikke benyttes nogen form for kombinatorer i søgeforespørgslen – søgningen er således uafhængig af ordenes rækkefølge. Teknikken adskiller sig fra eksakt match ved at være meget teoretisk funderet grundet de mange forsøg i testbaser, foruden søgeteknikkens statistisk genererede rangordning af søgeresultatet. Der opereres ofte med begrebet "best match", når snakken falder på partial matchteknikker (især i.f.m. søgemaskiner på Internet) – ved best match forstås, at der ved rangordningen af søgeresultatet tages hensyn til de dokumenter, der opfylder den booleske "og"-kombination – søgeresultatet præsenteres således efter faldende relevans. Det er en funktion der også er understøttet ved DanBibs fuzzy-prototype under rubrikken "Medtag relaterede poster" (mere herom senere ved gennemgang af prototypen). Før den nærmere beskrivelse af de enkelte partielle matchteknikker, vil det være på sin plads at fortælle lidt om termvægtning generelt, eftersom det i høj grad er noget af det, man forbinder med partial matchteknikkerne. Gradueringen af søgeresultatet efter større eller mindre lighed er således kun muligt i kraft af en statistisk genereret vægtning af termerne.

Termvægtning
Vægtning af termer kan dels knytte sig til query, dels selve dokumenterne/repræsentationerne. Dokument-termernes vægte beregnes enten i.f.m. indekseringen, hvorved vægten bliver en statisk størrelse, eller ved selve søgningen (dynamisk vægtning), og man vil da i vægtningen hele tiden tage hensyn til ændringer i databasens størrelse og sammensætning. Desuden kan der beregnes vægte af query-termerne, ligesom de også kan tildeles af brugeren, baseret på dennes vurdering af de enkelte termers indbyrdes betydning (relevans feedback). Et dokuments vægt i relation til en query, beregnes som summen af de termers vægte, de har tilfælles. Dette udtryk kan benyttes til at opstille en rangorden for en samling dokumenter i relation til en query (Havnø & Hansen, 1994, s. 48). Man kan udover vægtning af enkelttermer også vægte sammensatte termer, fraser, mv. Automatisk vægtning af enkelt-termer ud fra den hyppighed, hvormed de forekommer i en tekst, baserer sig på Zipf's[4] opdagelse af, at frekvensen af given term i en tekst multipliceret med termens rangorden tilnærmer sig en konstant for teksten. Sparck Jones har i 1973 ved forsøg med tre kendte testsamlinger påvist forbedrede søgeresultater ved brug af termvægtning (Havnø & Hansen, 1994, s. 47-48). De grundlæggende antagelser er:

1. Hyppigheden af en term i det enkelte dokument er signifikant.
2. Forekomsten af en term i et kort dokument er mere signifikant end forekomsten af samme term i et langt dokument.
3. Forekomsten af en sjælden term i et dokument er mere signifikant end forekomsten af en hyppig term.

Disse forhold udtrykkes normalt som en terms "tf.idf vægt": w=tf*idf, hvor "tf" er termfrekvensen i det enkelte dokument, mens "idf" (den inverterede dokumentfrekvens) er den inverterede funktion af termens forekomst i samlingens dokumenter, et mål for termens generelle hyppighed i basen. Vægtning af termer i query og dokumenter bør ideelt tage hensyn til termernes hyp-

pighed i dels dokumentet, dels i hele basen. Termfrekvensen "tf" indikerer termens/emnets signifikans for det pågældende dokument, mens "idf" angiver termens generelle hyppighed i hele basen, og dermed fungerer som indikator for termens evne som "diskriminator", d.v.s. værdi som søgeelement. "Idf" kan beregnes som 1/dokumentfrekvensen, eller som det hyppigere ses: log (N/dfk), hvor "N"=antal dokumenter i samlingen og "dfk"=dokumentfrekvensen. Man har for at foregribe den oplagte mulighed for høj rangordning ved lange dokumenter med mange termer, indført en såkaldt "normaliseringsfaktor", der bl.a. indgår i formlen for vektorrumsmodellen. Der er bred enighed om blandt forskerne, at anvendelsen af dokumentfrekvensen, "dfk", giver betydelig bedre resultater ved beregningerne af recall/precision, idet den giver større vægt til de sjældne termer, og mindre til de hyppige, som indgår af hensyn til en rimelig recall. DanBibs fuzzy-prototype bygger deres underliggende fuzzy-semantiske netværk på en statistisk udregnet vægtning af hver enkelt term i.f.t. andre relaterede emneord. De såkaldte "associationer" mellem emneordene, f.eks. mellem A og B, beror på en udregning af, hvor ofte det enkelte emneord A optræder sammen med B i de enkelte dokumenter. Prototypens såkaldte "emneords-associationer" udregnes mellem emneordene i det kontrollerede vokabular. Emneord A associerer således til emneord B med en grad, der bestemmes af, hvor mange gange A optræder i de samme dokumenter som B. Denne relation mellem A og B tildeles en værdi mellem 0 og 1 ved (i grove træk) at dividere antallet af dokumenter, hvor emneord A optræder sammen med B, med antallet af dokumenter hvor A forekommer (Andreasen, 1998, s. 16). Relationen mellem A og B kan i grove træk beregnes ud fra nedenstående formel. Der er dog i den aktuelle associationsudregning i prototypen anvendt en modificeret udgave af denne formel for at opnå en række "gode" egenskaber (Brugervejledning, 1999, s. 9). Denne modificering har primært til hensigt at nedtone bestemt hyppigt forekommende søgeords effekt som deskriptor, f.eks. termen "Danmark" (Nielsen, 1999) – således bliver associationen relativt svagere, jo hyppigere den associerede term er. Der er endvidere implementeret en modificering, der f.eks. associerer 10 sammenfald ud af 100 forekomster som stærkere, end en association baseret på 1 sammenfald ud af 10 forekomster.

I grove træk associerer emneord A til emneord B *(A : B)*, svarende til:

$$\frac{\textit{(Antal sammenfald imellem A og B)}}{\textit{(antal forekomster af A)}}$$

Af formlen ses det, at jo oftere emneordet A forekommer sammen med emneord B i indekseringen, jo stærkere vil associationen fra A til B være.

Ved tilbageblik på fig. 1's klassifikation, ses det, at de partielle match teknikker kan inddeles i to hovedtyper – de "individuelle-" og de "netværksbaserede matchteknikker". Karakteristika ved disse er iflg. Belkin & Croft (1987):

- *Individuelle matchteknikker:* Her baseres matchpå ligheder mellem forespørgsler og individuelle, enkelte dokumenters karakteristika.
- *Netværksbaserede matchteknikker:* Match baseres her på ligheder mellem dokumenter, hvor søgning foregår som navigering i et netværk af relaterede dokumenter.

Fuzzy-sæt modellen: Fuzzy-sæt tilhører gruppen af Individuduelle matchteknikker og modellen udgør en vigtig del af den samlede teori omkring fuzzy logic. I fuzzy-sæt modellen er det kun dokumenttermer, der vægtes (jf. f.eks. DanBibs fuzzy-prototype). Query'en udføres som en standard boolesk formulering, hvor de booleske operatorer repræsenteres eller "oversættes" ved såkaldte "aggregeringsfunktioner". I denne forbindelse svarer boolesk "og" til minimum opfyldelse af søgeformuleringen, mens "eller" svarer til maksimum opfyldelse af forespørgslen (Andreasen, 1998). Dokumenterne rangordnes herefter efter følgende regel, idet vægtene af termerne typisk vil være statistisk genereret (Havnø & Hansen, 1994):

- *ELLER*-relation (A eller B)→højeste vægt af emneord A eller B i et dokument.
- *OG*-relation (A og B)→laveste vægt af emneord A eller B i et dokument.

- *IKKE*-relation (ikke A)→1-vægten af emneord A i et dokument.

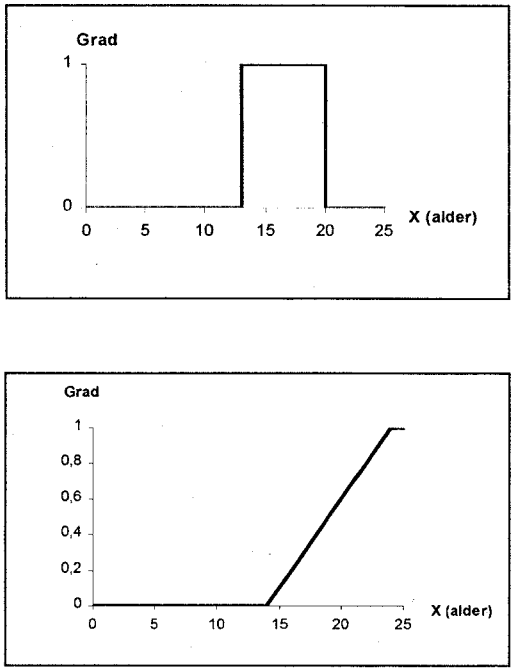
Med baggrund i skitseringen af de eksakte- såvel som de partielle matchteknikker, vil principperne i fuzzy logic efterfølgende blive gennemgået. Der redegøres i denne sammenhæng samtidig for bl.a. fuzzy-sæt, medlemskabsfunktioner og fuzzysemantiske netværk.

Fuzzy logic

Fuzzy logic som logisk begreb blev introduceret tilbage i 1965 af den amerikanske professor ved Berkeley universitetet i Californien, Lofti Zadeh, der arbejdede med, hvordan computere kunne lære at genkende håndskrift. I forbindelse med sit arbejde fik han ideen om, at gøre genkendelse af tegn til et spørgsmål om størst mulig overensstemmelse mellem disse, fremfor at sætte bogstaverne i klare (og brede) foruddefinerede kategorier (Nielsen, 1997). Filosofien var og er, at de enkelte tegn kan ligne et forudbestemt, og klart defineret bogstav på en prik eller til "en vis grad", og stadig tilhøre det enkelte bogstavs eller tegns gruppe. De enkelte bogstaver kan med andre ord tilhøre "gråzonen" og stadig være gruppemedlemmer. Selvom teorien om fuzzy logic er relativ ny indenfor informationsvidenskaben, og således ikke i særlig stor udstrækning gennem tiden har været afprøvet i testdatabaser eller operationelle søgesystemer, så er fuzzy logikkens teoretiske fundament allerede grundigt udforsket, bl.a. indenfor styring og regulering af industrielle systemer. På europæisk plan dukkede de første industrielle eksperimenter med fuzzy logic op ca. 1970 – dette var i England i.f.m. styringen af en dampgenerator, hvor konventionelle teknikker ikke slog til (von Altrock, 1995, s. 4). Efter 1980 kom Japan med på vognen, og de er i dag efterhånden blevet førende indenfor videreudviklingen af denne teknik. I Danmark startede F.L. Schmidt-koncernen i 1980'erne, som de første herhjemme, implementeringen af fuzzy logic i deres styringsanlæg (Schomacker, 1996).

Fuzzy logic, fuzzy-sæt og medlemskabsfunktioner

Man regner normalt fuzzy logic som en generalisering af boolesk logik, idet man udover de boolske sandhedsværdier sand og falsk (1 og 0) også indbefatter gradueringer mellem disse (Andreasen, 1998, s. 13). Hvis vi holder os til mængdebilledet eller sæt-teorien generelt, så er fuzzy logic udtryk for, at et element i mængden, sættet, har en medlemsgrad fra 0 til 1, svarende til graden af sandhed for, hvorvidt et medlem tilhører en mængde. Det er i denne forbindelse vigtigt at slå fast, at medlemsgraden "sandt til en vis grad" ikke må forveksles med "med en vis sandsynlighed" (Andreasen, 1998). Til fastlæggelse af sandhedsværdier benytter man normalt de såkaldte "medlemskabsfunktioner". Sandhedsværdien for eksempelvis udtryk som "teenager" kan fastlægges ved en funktion $m_{teenager}$, som illustreret i figur 2a, mens sandhedsværdien for udtrykket "voksen" f.eks. kan være fastlagt ved m_{voksen} i figur 2b.



Figur 2: Medlemsskabsfunktioner for »teenager« (2a) og »voksen« (2b) - (Andreasen, 1998, s. 14)

Eksemplet viser yderpunkterne, hvor man dels kan tilhøre, dels være udenfor gruppen af "teenagere", d.v.s. have grad 1 eller 0, eller man kan høre under gruppen af "voksne" i større eller mindre grad – afhængig af hvilken alder, man vælger at opfatte som svarende til værende helt voksen (i

dette eksempel fra 24 år). Fig. 2b viser som eksempel, at man som 18-årig må regnes for at tilhøre gruppen af voksne med medlemsgraden 0.4. Et andet eksempel til illustration af medlemskabsgrad kan være en gradinddeling af begrebet "høj" – vi opstiller et fuzzy-sæt "høj". I denne forbindelse opstår dilemmaet, hvornår regnes man for værende lav eller høj? For at kunne lave en sådan matematisk graduering, er man nødt til at fastlægge nogle subjektivt evaluerede grænseværdier, henholdsvis for lav og høj. Hvis vi vedtager, at en højde på 150 cm svarer til "lav", og en højde på 210 cm berettiger til prædikatet "høj", så kan en gradueret skala se ud på følgende måde (Tabel 1):

Højde (i cm):	Grad af medlemskab:
150	0.00
162	0.08
174	0.32
180	0.50
192	0.82
204	0.98
210	1.00

Tabel 1: Fuzzy-sættet "høj" (bearbejdet model efter Schneider et al., 1996, s. 25)

Til et sammensat udtryk, hvori der indgår flere simple udtryk, som f.eks. "A og B", hører også en medlemskabsfunktion $m_{A \text{ og } B}$. Denne defineres ud fra medlemskabsfunktionerne m_A og m_B for de indgående simple udtryk, idet man "aggregerer" disse med en funktion, der, som tidligere omtalt under *Fuzzy-sæt modellen,* repræsenterer den booleske operator "og". Andreasen (1998, s. 14) giver som eksempel, at man ved at benytte udtrykket minimum som aggregeringsfunktion, agg(min), for at udtrykke "voksen og teenager", således at der for alderen 18 (jf. fig. 2), som giver $m_{voksen}(18) = 0{,}4$ og $m_{teenager}(18) = 1$, fås $m_{voksen \text{ og } teenager}(18) = \text{minimum}(m_{voksen}(18), m_{teenager}(18)) = 0{,}4$. Havde vi i stedet valgt aggregeringsfunktionen agg(max), havde resultatet været maksi-mum($m_{voksen}(18)$, $m_{teenager}(18)$) = 1. Det er vigtigt i denne sammenhæng at pointere, at fuzzy logic rummer muligheder for valg af mange forskellige aggregeringsfunktioner, hvor aritmetisk (alm.) gennemsnit, som omtalt, er en ofte valgt og simpel aggregeringsfunktion (Andreasen, 1998).

Fuzzy forespørgsler

I forbindelse med traditionelle søgninger i databaser (eller dokumentbaser) er en forespørgsel et logisk udtryk, der ved ét eller flere kriterier, beskriver det, man interesserer sig for, og et svar er en mængde af de poster, der opfylder forespørgslen. I en fuzzy-forespørgsel foreligger i princippet en medlemskabsfunktion til hvert kriterium i forespørgslen. Hvert medlemskabsfunktion tager en post som argument, således at der til en forespørgsel F: "alder=voksen" og navn=Peter" (svarende til "giv mig en voksen, der hedder Peter"), hører to medlemskabsfunktioner $m_{alder=voksen}$ (X) og $m_{navn=Peter}$ (X). En given post hører til svaret på forespørgslen til en grad mellem 0 og 1, nemlig den grad, hvormed posten opfylder forespørgslens kriterier. Graden bestemmes af en medlemskabsfunktion for forespørgslen: m_F (X) = agg($m_{alder=voksen}$ (X), $m_{navn=Peter}$ (X)), hvor aggregeringsfunktionen "agg" f.eks. kan vælges som maksimum eller minimum (som tidligere omtalt). I forbindelse med ønsket om at opnå en fleksibilitet i søgemekanismerne, må man erkende, at hverken brug af minimum eller maksimum som aggregeringsfunktion er hensigtsmæssig. Funktioner der placerer sig imellem minimum og maksimum, de såkaldte "gennemsnitsfunktioner", er derfor ofte et bedre bud, hvis en rangordning af fremfundne poster efter grad af opfyldelse i.f.t. query ønskes (som f.eks. under DanBibs prototype, hvor alm. gennemsnit benyttes).

Til eksempel kan vi forestille os en boolesk medlemskabsfunktion for "navn=Peter", der er 1 (d.v.s. sand) for Peter og ellers 0, samt en medlemskabsfunktion for "alder=voksen" som i fig. 2b, hvormed vi opnår en fleksibel eller "blød" fortolkning af denne forespørgsel. Forespørgslen F bliver herefter F: "alder=voksen" og "navn=Peter", hvor Peter på 30 opfylder til graden 1, Lisa på 40 til graden 0.5, Peter på 18 til graden 0.75 og endelig Lisa på 18 til grad 0.25 (Andreasen, 1998, s. 15). Senere under gennemgangen af DanBibs fuzzy-prototype belyses, hvordan DBC har valgt at man skal kunne "opbløde" de enkelte søgekriterier.

Fuzzy logic til informationssøgning

Fuzzy sæt og fuzzy logic anvendt som søgeteknik tilhører, som nævnt, gruppen af de partielle matchteknikker, idet dokumenter (eller repræsentationer af disse), der kun matcher en query delvist, kan/vil blive fremfundet ved søgninger. Ét af formålene med at anvende fuzzy logic som søgeteknik er, at give brugeren hjælp i den situation, hvor vedkommende får 0 eller kun meget få poster. Specielt for slutbrugere vil fuzzysøgning være et alternativ, da de bl.a. slipper for eksplicit at formulere booleske søgeudtryk med brug af de booleske kombinatorer "og", "eller", "ikke" (Nielsen, 1997) – her vil der i stedet kunne blive tale om (hvis muligt) at indstille forskellige "skydeparametre" (som ved DanBibs fuzzy-prototype), der bl.a. kan repræsentere de tidligere omtalte aggregeringsfunktioner. Brugeren behøver endvidere ikke at spekulere på formulering af relaterede søgeord eller synonymer, da informationssøgesystemet i kraft af dets tilknyttede, underliggende "semantiske netværk" af relaterede søgeord, typisk emneord, rummer mulighed for mere eller mindre "automatisk spørgsmålsudvidelse". Ved søgningen foregår der således en udvidelse/inddragelse af fuzzy-sættets "medlemmer", relaterede emneord - hver især med større eller mindre medlemsgrad til query-termen. Query repræsenterer i denne forbindelse en medlemskabsfunktion i et fuzzy-sæt af emneord, og alt efter indstillingen af systemets søgeparametre, vil der blive søgt på query-termen, foruden udvidet med relaterede ord. I fald der ikke er nogle dokumentrepræsentationer, der matcher query-termen nøjagtigt, d.v.s. ingen dokumenter er indekseret med denne term, vil der stadig blive søgt på relaterede emneord, såfremt angivet under søgeindstillingerne. Det underliggende semantiske netværk, der afspejler relationerne mellem de enkelte fuzzy-sæt medlemmer (termer eller begreber), kan dannes på flere forskellige måder (Nielsen, 1997) (Nørgaard, 1998):

- *Ekspertviden* fra en thesaurus, en synonymordbog, eller en eksperts forståelse af sammenhænge.
- Man udbygger fuzzy-semantiske termnet ud fra en allerede eksisterende emnetildeling, f.eks. en klassifikationsmodel/-system.
- *Data-mining,* som er en analyse eller statistik over data, hvor f.eks. de indbyrdes termrelationer udregnes på baggrund af emneords sammenfald i dokumenterne eller repræsentationerne for disse.

DanBibs fuzzy-prototype benytter sig af sidstnævnte model, data-mining, til dannelsen af dets fuzzy-semantiske netværk. I denne forbindelse er det vigtigt at understrege, at fuzzy-søgeteknikken anvender et netværk af relaterede ord, der er baseret på, at de enkelte termrelationer udregnes i søgeøjeblikket – i modsætning til de à priori genererede termnetværk under de "netværksbaserede søgeteknikker", f.eks. klyngebaserede søgninger (Havnø & Hansen, 1994, s. 46). Hvis kort nogle af de vigtigste punkter skal sammenfattes, der knytter sig til anvendelsen af fuzzy logic i.f.m. informationssøgning, så drejer det sig primært om at finde ud af, hvordan de såkaldte "opblødningsfunktioner", der i øvrigt godt kan være flere funktioner, skal indrettes. Som eksempel på nogle opblødningsfunktioner kan nævnes følgende, der begge er fuldt implementeret i DanBibs fuzzy-prototype (Schomacker, 1996, s. 7):

- Fuzzy-semantiske netværk, der er et vægtet termnet, som angiver grad af overensstemmelse mellem netværkets enkelte termer.
- Simpel opblødning af forespørgsler, v.h.a. aggregering - d.v.s. elastisk overgang mellem "og" og "eller". I prototypen omtales funktionen som "samt".

Om det er lykkedes DBC at skabe et reelt, forbedrende og relevant søgealternativ til den traditionelle booleske søgning i DanBib, belyses senere under gennemgangen af fuzzy-prototypen foruden i afsnittene *Relativ relevansbedømmelser,* hvor en konkret mini-test giver fingerpeg herom.

DanBibs fuzzy-prototype

Søgeapplikationen *Fuzzy-søgning i DanBib – Prototype ver. 2.0* har DBC stillet til fri afbenyttelse for alle, der er interesseret i at prøve den. Prototypen findes på flg. URL:
http://www.isl.ruc.dk/cgi-bin/dbc_cgi.

Prototypens datagrundlag

En forudsætning for at fuzzy-søgeteknikken kan udvide sine søgninger med relaterede ord, er, at der er etableret et semantisk netværk, der kan tages udgangspunkt i. Det er dog vigtigt, at man fra systemproducent/-leverandørside forinden nøje gør sig klart, hvilke data der skal benyttes som netværkets datagrundlag - og ikke mindst hvorfor. DBC har til deres prototype valgt at benytte et udtræk af bibliografiske *faglitterære* poster hentet fra DanBib – i et omfang der svarer til ca. 45.000 artikelposter og 30.000 bogposter. Artikelposterne er alle indekseret efter DBC's emneordssystem – det drejer sig om artikelposter efter 1996. Bogposterne svarer i denne forbindelse til alle bøger optaget i seddelfortegnelsen siden 1987, hvor det daværende Bibliotekscentralen påbegyndte den verbale emneindeksering (Forsberg et al., 1999, s. 7). Indekseringen i DanBib og niveauet af dette samt DBC's generelle indekseringsstrategi er tidligere behandlet i afsnittet *Indeksering* og *repræsentation.* I løbet af de år, hvor DBC har arbejdet med fuzzy-projektet i forskellige prototype-versioner, har de gjort sig en række erfaringer, bl.a. m.h.t. hvilke data der er mest relevante til dannelsen af et termnet. Udviklingsholdet er i denne periode kommet frem til, at faglitteraturen må være den mest oplagte litteraturtype at bygge termnettet på, ud fra en common sense filosofi om, at virkeligheden, eller snarere "forholdet mellem virkelighedens komponenter, er fornuftig" (Forsberg et al., 1999, s. 7).

System-udviklerne hos DBC og ISL har i forbindelse med udviklingen af fuzzy-prototypen været bevidst om den måde eller det forløb, et faglitterært dokument gennemløber fra dets afsæt i forfatterens virkelighedsopfattelse til den senere bibliografiske repræsentation i en database. Medlemmer af projektgruppen diskuterer i denne forbindelse faglitteraturens afspejling af "virkelighedens komponenter" (Forsberg et al., 1999). Der opereres med en opdeling af forholdet mellem virkelighedens komponenter i en række "afspejlingsled", dels virkelighedens afspejling i litteraturen, dels dokumenternes afspejling eller repræsentation i de bibliografiske poster. Førstnævnte afspejlingsled, eller filter, knytter sig til ophavets opfattelse af virkeligheden, som sammen med forlagsredaktørerne sikrer os, at de beskrivelser af virkeligheden, som kommer til at foreligge som dokumenter, holder sig indenfor det normalt accepterede forklaringsrum. Med det litteraturgrundlag, der foreligger for termnettet, kommer så endnu et filtreringsled, idet ikke alle bøger optages i seddelfortegnelserne, ligesom heller ikke alle periodica indekseres i Dansk Artikelindeks. Det andet afspejlingsled er dokumenternes afspejling eller repræsentation i de bibliografiske poster – i denne forbindelse er det specielt deres emnerepræsentation, det drejer sig om. Det handler om at finde de ord, der bedst karakteriserer de komponenter eller de sammenhænge mellem komponenter, som dokumenterne "påstår" findes ude i virkelighedens verden. Med andre ord: Hvis virkelighedens komponenter, eller relationer, afspejles i dokumenterne, og dokumenterne afspejles i de bibliografiske poster, så afspejler de bibliografiske poster i sidste ende selv en del af virkeligheden (Forsberg et al., 1999, s. 8). Samlet kan man sige, at den fornuft vi går ud fra findes i virkeligheden, repræsenteres i dokumenterne og repræsentationen af disse afspejles i de bibliografiske poster. Indekseringen afspejler i denne forbindelse de meningsfulde sammenhænge fra virkelighedens verden – alt afhængig af indeksørens forståelse for forfatterens virkelighed og bagvedliggende hensigter med at forfatte sit værk. Hvorvidt brugeren efterfølgende netop opfatter indeksørens repræsentation (i form af emneord) som relevant, d.v.s. mener at dokumentet, det beskriver/repræsenterer, kan dække vedkommendes informationsbehov, afhænger herefter i høj grad af brugerens virkelighedsopfattelse og begrebsapparatur – d.v.s. videnstrukturer.

Jeg vil nu komme nærmere ind på det fuzzy-semantiske netværk og dannelsen af dets emneords-associationer.

Prototypens semantiske termnet

Hos DBC har man i.f.m. udviklingen af prototypen eksperimenteret med forskellige typer data fra de bibliografiske poster, henholdsvis data, der kan citeres direkte fra posten (titel, forfatter), foruden de data, der intellektuelt konstrueres af den bibliotekar/indeksør, der laver den bibliografiske post (f.eks. emneord). DBC er i denne forbindelse

nået frem til en erkendelse af, at de kontrollerede, faglitterære emneord [5] må være de bedste bibliografiske data til dannelsen af termnettet, der skal repræsentere dokumenternes afspejling af virkeligheden. Ved brug af denne type emneord undgås det, at DanBibs uensartede indeksering (jf. de mange "bidragsyderes"/forskningsbibliotekers inkonsistente indekseringspraksis) er stærkt medvirkende til at forplumre, hvad man som bruger af systemet med rimelighed kan forvente at finde i systemet. Man har endvidere fra DBC's side undladt at lade dokumenternes titler afspejles i termnettet. I en database som DanBib, hvor der er litteratur indenfor alle typer og indenfor alle emneområder, er der for stor spredning i kvaliteten af de forskellige citerende data-elementer. Hvem skulle f.eks. tro, at en titel som "Dræb de hvide elefanter" ikke er en opfordring til at gå på storvildtjagt, men en kritik af dansk u-landsbistand (Forsberg et al., 1999, s. 8). I denne forbindelse kunne det omkring fuzzy-prototypen dog overvejes, om ikke en hensyntagen til sammenfald mellem ord i emneords- samt titelfeltet, ville medføre en øget relevans for brugeren af de fremfundne dokumenter? Ved at benytte emneord fra et postkoordineret emneordssystem, opnås en række fordele, foruden et par problemer i søgemæssig henseende. Ulemperne eller de negative konsekvenser heraf, er i første omgang bl.a.:

1. alle emneord i en post har i princippet lige stor vægt, og der kan *ikke* ud af de kontrollerede emneord direkte læses noget om emnernes vægt i dokumentet.
2. emneordene angiver normalt ikke syntaktiske relationer, f.eks. viser brugen af deskriptoren "sociologi" *ikke,* om emneordet er brugt, fordi det er bogens emne, eller fordi "sociologi" er den anvendte metode eller synsvinkel i det indekserede dokument (Indeksering af faglitteratur, 1998, s. 16-17).

Fordele er der dog flere af, da netop denne type emneord er som skabt til anvendelse i edb-systemer. I modsætning til de prækoordinerede emneordssystemer, der kan bestå af dels "subject headings" (f.eks. Japan – samfundsforhold eller Japan – historie) og "komposita" (f.eks. *privatskole*bib*lioteker*), hvor kombinationen af søgetermer/ emnestrengen sker i "indekseringsfasen" (eller i konstruktionen af indekseringssproget) som aspekter eller facetter af det overordnede emne, så har det postkoordinerede emneordssystem den fordel, at sammenføjningen sker på "søgetidspunktet" ved kombination af emnestrengenes enkelt-elementer. Hermed opnås der mulighed for at kunne kombinere emneord og søge på meget specifikke aspekter ved et bestemt emne. Forsberg et al. (1999) nævner, at der i snit tildeles ca. 4 emneord pr. post i DBC's system (DanBib). Da netop et statistisk genereret termnet, som fuzzy-prototypens, baseres på antal af samforekomster mellem emneord, er det klart, at jo flere emneord, des bedre (og mere reelt) grundlag for udregningerne af termrelationerne.

Selvom det på en måde kan være nærliggende at betegne dette fuzzy-semantiske netværk som en thesaurus med en hierarkisk betydningsstruktur, så er det misvisende. Dette skyldes, at selvom A associerer til B med en vis grad, så behøver dette ikke nødvendigvis betyde at denne associationsgrad afspejler den betydningsmæssige relation mellem A og B, men blot at B potentielt er et godt alternativ i søgningen efter poster, der har med A at gøre. Det er dog vigtigt at pointere, at disse statistisk genererede relationer mellem emneordene sagtens kan indeholde semantiske relationer – der er dog ingen garanti for det. Man kan i fuzzy-systemet få vist en emneordsliste, der lister systemets anvendte kontrollerede emneord. Denne betegnes ofte som en "flad" thesaurus, da den netop ikke angiver de hierarkiske relationer mellem indekstermerne (Forsberg et al., 1999, s. 9).

Hvis fuzzy-termnettet skal sammenlignes med noget kendt, så er det mest nærliggende nok, at se det i.f.t. emneordsklynger. Klynger er dog som oftest lavet ved, at en bibliotekar har gennemgået ordforrådet i en liste af emneord, og derefter samlet et antal emneord i forskellige associative (sideordnede eller idè-associative) klynger. Mens disse klynger bygger over allerede eksisterende semantiske relationer, bliver thesauri derimod som regel lavet i forbindelse med dannelsen af selve ordforrådet, forstået på den måde, at man i forbindelse med at et nyt ord skal ind i ordforrådet, placerer det hierarkisk og e.v.t. associativt i.f.t. det allerede eksisterende ordforråd. I modsætning til disse to typer af "ordbøger", baserer fuzzy-prototypens

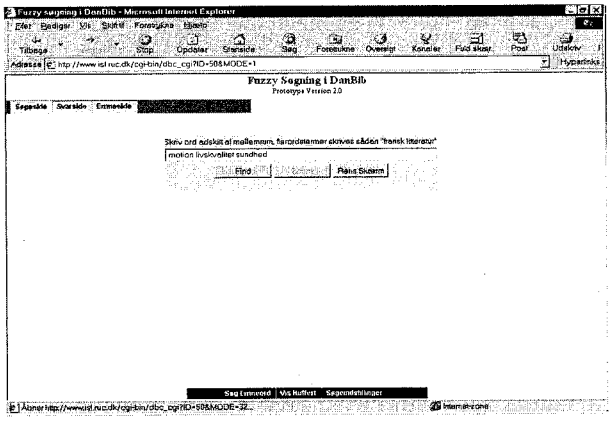
termnet sig derimod udelukkende på data-mining baseret på statistisk co-optræden i allerede eksisterende data - i form af bibliografiske poster. Man kan så spørge, hvorvidt dette semantiske netværk i sig selv har noget med fuzzy logic at gøre? Svaret på dette er, at det statistisk genererede semantiske netværk skaber grundlag for at danne medlemskabsfunktioner for emneordskriterier, og dermed grundlag for emneordssøgning med fuzzy logisk evaluering (Andreasen, 1998). Hvis et emneord som f.eks. "barndom" associerer til graden 0.6 til emneordet "børn", så vil en post P, der er indekseret med emneordet "børn", matche kriteriet "barndom" til graden: $m_{barndom}(P) = 0.6$.

Med baggrund i redegørelsen for nogle af de overvejelser, der ligger til grund for DBC's valg og etablering af det fuzzy-semantiske netværk, vil prototypen nu blive nærmere beskrevet.

Gennemgang af prototypen

Ved opstart præsenteres brugeren for et skærmbillede, der har links til applikationens 3 sider: *Søgesiden, Svarsiden og Emnesiden.* Disse er inddelt som fanebladssider med søgesiden forrest ved start af applikationen. Hertil kommer desuden 3 hjælpefunktioner, der er tilgængelige via knapperne forneden på skærmen "Søg Emneord", "Vis kuffert" og "Søgeindstillinger". Applikationens hovedsider samt de tilhørende hjælpefunktioner gennemgås nu punkt for punkt - i en rækkefølge, der er naturlig for søgeprocessen.

Søgesiden
Søgesiden fremtræder som vist i figur 3. Den indeholder blot et enkelt indtastningsfelt, hvori der kan indtastes forskellige søgeord. Alle de angivne søgeord tolkes som selvstændige søgekriterier – der kan dog benyttes flerordstermer eller fraser ved brug af citationstegn, som f.eks. i søgestrengen: DanBib "fuzzy logic". Der kan kun søges på kontrollerede emneord (Brugervejledning, 1999, s. 3) - fussyficeringen tager således udelukkende afsæt i disse. Hjælp til at finde de DBC-kontrollerede emneord fås med funktionen "Søg Emneord". Der er i prototypen ikke implementeret en stopordsliste. "Stemming" understøttes ligeledes ikke, idet der ikke er indbygget automatiske søgethesauri til hjælp i sådanne situationer.



Figur 3: Søgesiden

Søg Emneord

Denne hjælpefunktion giver brugeren adgang til en liste over de kontrollerede emneord, der anvendes i søgningen. Ved at klikke på knappen "Søg Emneord" åbnes et dialog-vindue. Under indtastningen af de enkelte tegn i dette felt, foretages der løbende en automatch-funktion på de kontrollerede emneord – det svarer til en "skan"- eller "liste"-funktion, der åbner sig for brugeren ved indtastningerne. Ved at vælge et ord fra listen og klikke på trykknappen "Tilføj" (eller dobbeltklikke på ordet), bliver dette føjet til søgetermerne. Efter valg af et ord på listen, og derefter klik på "Browse"-knappen, aktiveres emneords-browseren med den valgte søgeterm (mere herom senere under *Emnesiden*).

Søgeindstillingerne
Søgeindstillingerne kan justeres ved klik på knappen "Søgeindstillinger". Under normale omstændigheder burde det ikke være nødvendigt at justere disse indstillinger, da de er forudindstillet til at tilfredsstille de mest almindelige søgesituationer. Man kan som bruger bl.a. ændre på en simpel parameter, der tilpasser svaret, nemlig "Medtag antal poster pr. side" - der medtages fra 5 til 25 poster. Hertil kommer tre parametre, der kan anvendes til at tilrette den fleksible søgeteknik. Dette drejer sig om "Slider'en", "Medtag relaterede poster" samt "Medtag relaterede emneord":

Slider'en
Denne aggregerings-operator bestemmer hvor-

dan opfyldelsen af enkeltkriterier for en post sammensættes til en opfyldelse af posten som helhed. I normal boolesk logik bruger man typisk kun de to operatorer "og" og "eller" – i fuzzy logic kan man som nævnt graduere imellem disse. Da sandhedsværdierne er numeriske (mellem 0 og 1) repræsenteres en fuzzyaggregeringsfaktor normalt som en numerisk funktion. Et "og" kan repræsenteres som *minimum* og et "eller" som *maksimum.* I prototypen benyttes som standardvalg for aggregeringen et *aritmetisk gennemsnit*, svarende til almindelig gennemsnit – dette illustreres ved, at skydeparameteren er placeret i midterpositionen. Skydeparameterens yderpunkter repræsenterer hhv. almindeligt boolesk "eller" foruden "og". At flytte skyderen mod venstre (mod "eller"), vil systemet opfatte som et løsere krav til opfyldelsen af kriterierne i søgeforespørgslen, mod højre betyder omvendt strengere krav. Resultatet af disse ændringer af parametrene vil afspejles i graden til hvilken en post opfylder en forespørgsel, og til hvilke poster der medtages. Ved en indstilling med skyderen i midterpositionen, vil aggregeringen opfatte søgeforespørgslens ord som mere eller mindre ligeværdige. Hvis en søgeforespørgsel f.eks. har 4 emneord, bliver hver af disse således givet vægten 0.25 (1/4), med tre emneord 0.33 (1/3) o.s.v. Hvis skydeparameteren flyttes mod "eller", får det hyppigst forekommende ord i basen den største vægtning. Omvendt bliver resultatet, hvis parameteren flyttes mod "og", vægtes den mindst forekommende term højst. Dette svarer til den tidligere omtale af agg(max) og agg(min) under afsnittet *Fuzzy logic, fuzzy-sæt og medlemskabsfunktioner.*

Medtag relaterede poster
V.h.a. dette valgfelt kan man indstille om og i hvor høj grad poster, der ikke præcist matcher de eksakte søgekriterier, skal inddrages. Man angiver med andre ord, den mindste grad af opfyldelse for forespørgslen, der kan accepteres for poster. At medtage relaterede poster svarer hermed til, at poster, der kun opfylder forespørgslen til en grad mindre end 1, medtages. Som standard er denne indstilling sat til 0.5, hvilket svarer til, at der kan medtages poster, der kun indeholder ned til halvdelen af søgeordene. Har man f.eks. anvendt 4 søgeord, betyder det, at man giver tilladelse til at medtage poster, der indeholder 3 eller 2 af søgeordene, i søgeresultatet (Nørgaard, 1998, s. 33). Der kan medtages poster i intervallet 1 til 0.1, hvor graden bestemmer, i hvor høj grad posten relaterer til forespørgslen. Indstilles "Medtag relaterede poster" til f.eks. grad 0.6, medtages poster i svaret, blot de delvist opfylder kriterierne. Hvis den valgte aggregeringsfunktion er gennemsnit (standardvalget), er det gennemsnittet af sandhedsværdierne (værdier mellem 0 og 1) for kriterierne, der skal være mindst 0.6. Det er interessant at lægge mærke til, at hvis der fussyficeres over emneord, d.v.s. "Medtag relaterede emneord = nej" (svarende til værdien 1), så svarer "Medtag relaterede poster" til "best-match" – svarende til hvor mange af søgeordene, der obligatorisk skal medtages i søgeresultatet. Posterne vil efterfølgende blive rangordnet efter faldende antal emneord, der matcher søgeforespørgslen.

Medtag relaterede emneord
Denne funktion styrer fussyficeringen af emneord via det fuzzy-semantiske netværk. Med funktionen kan man således bede om at inddrage ord, der ikke direkte er angivet som søgeord, men som optræder som relateret eller associeret term til det pågældende søgeord. Emneords-kriterierne blødgøres med andre ord ved at acceptere et emneord, som er delvist knyttet til en post, blot posten indeholder et ord, der er associeret af det pågældende emneord. Præ-indstillingen er 0.2, hvilket betyder at hvert af de angivne emneord ekspanderes til en række ord, der er associerede mindst til graden 0.2. Jo mindre værdi, des bredere søgning – svarende til inddragelse af flere associerede søgeord.

Samlet om parametrene og kriterierne
Generelt kan man sige, at når der er tale om fuzzyforespørgsler med flere kriterier, så skal man som udgangspunkt opfatte kriterierne som sammensat med "og". Man kan så "fussyficere" forespørgslen, dels ved at acceptere forespørgslen som helhed kun er opfyldt "til en vis grad", dels ved at acceptere, at enkeltkriterier kun er opfyldt "til en vis grad". Det er hvad der udtrykkes med parametrene henholdsvis "Medtag relaterede poster" og "Medtag relaterede emneord". Opfyldelsen af et enkeltkriterium "til en vis grad" er kun muligt for søgeord, der er "Emneord" – som det fremgår af parameterens navn. Opfyldelsen af en fore-

spørgsel som helhed har intet at gøre med typerne af de enkelte kriterier. Den bestemmes således som en grad, der beregnes som en funktion af de enkelte kriteriers grad af opfyldelse. Parameteren "Medtag relaterede poster" kan benyttes til at indstille en minimum for den beregnede grad af opfyldelse for posten. "Slider'en" kan benyttes til at ændre på selve aggregeringsfunktionen.

Svarsiden
Resultatet af søgningen vises på denne svarside som en rangordnet liste, der øverst i hyperlinkform angiver de første 5-25 poster med titel samt grad af opfyldelse i.f.t. søgeforespørgslen som helhed. Længere nede på listen vises de mere detaljerede bibliografiske oplysninger om den enkelte post. Ved at klikke på den fremhævede post i skærmens top, springer man til disse oplysninger. Der er en høj grad af inkonsistens i fremvisningsniveauet på disse dokumentrepræsentationer. Obligatorisk rummer alle posterne i vis-formatet oplysninger om opfyldelsesgrad, titel, ID-nummer foruden de kontrollerede emneord. Derudover kan posterne i nogle tilfælde indeholde en undertitel, e.v.t. forfatter, udgivelsesår samt ISBN-nummer. Der gives ved fremvisning af søgeresultatet ikke nogen eksplicit mulighed for yderligere raffinering af søgningen, "relevans feedback". Den mulighed, der findes, er, at man som bruger kan browse sig igennem de hyperlinkede emneord for at se associationerne til og fra det valgte ord. Graden for opfyldelsen er et tal i intervallet [0;1], og udtrykker som tidligere nævnt i hvor høj grad posten indeholder termer i overensstemmelse med søgetermerne. Indstillingerne i valgfelterne (d.v.s. under Søgeindstillingerne) er afgørende for, i hvor høj grad søgningen aggregeres eller blødgøres, og dermed for hvor mange poster, der vises. Hvert emneord er i ovennævnte detaljerede beskrivelse igen et link, der kan føre brugeren videre til "Emnesiden", hvorved associationer i emneordsnettet til og fra det aktuelle ord kan undersøges.

Emnesiden
Denne emneside kan nås fra opstartsskærmen ved at klikke på fanebladet "Emneside". Første gang denne funktion benyttes er emneordslisten tom, og den viste dialogboks "Søg emneord" skal benyttes til indtastning af et ord, for at komme videre. Emnesiden kan også aktiveres ved klik på et emneord eller ved knapper hertil fra "Søg Emneord"- eller "Vis kuffert"-vinduerne. Ved at browse i dette emneordsnet fås et indblik i de enkelte emneords associationer i.f.t. hinanden. Det ord browseren er aktiveret med præsenteres centralt med ord, der associerer *til* det pågældende til venstre, og ord der associeres *af* det pågældende til højre. Når man klikker på et bestemt emneord gøres dette ord til et nyt centralt emneord i browseren.

Vis kuffert
Ud for emneord på svar- og emnesiden er vist et kuffertsymbol. Ved klik på et kuffertsymbol bliver dette mørkt, og det tilhørende ord gemmes i, hvad der svarer til en "kuffert" for emneord – til senere brug i søgning. Med knappen "Tilføj" kan ord i kufferten tilføjes på søgesiden.

Med baggrund i gennemgangen af fuzzy-prototypen vil forsøgspersonernes søgeforespørgsler og søgetermer nu blive afprøvet i søgeapplikationen, hvorefter en sammenlignende søgning i DanBib/DanWeb vil blive foretaget.

Søgninger i DanWeb vs. Fuzzy DanBib

For at kunne foretage sammenlignelige søgninger i de to systemer med efterfølgende evaluering af søgeresultaterne, må det sikres, at søgningerne foretages ud fra de samme forudsætninger. En tilpasning er nødvendig, hvis effekten af fussyficeringen overfor den booleske søgelogik skal vurderes på et så reelt grundlag som muligt. Flere metodiske overvejelser er gjort i denne sammenhæng, dels omkring selve søgningen, dels hvad angår den efterfølgende evaluering af de fremfundne dokumentrepræsentationer. Følgende forhold er relevante:

1. Præcist hvilken type(r) emneord består fuzzy-termnettet af, og hvordan benytter prototypen sig egentlig af disse ved søgninger og fussyficeringer?
2. Er der nogen bestemt afgrænsning på disse emneord, f.eks. kontrollerede eller ukontrollerede, årstalsbegrænsning, bestemte søgeregistre, m.v.?
3. Kan DanWeb indstilles til at søge præcist på de

samme emneord eller med samme søgeparametre som i fuzzy-prototypen?
4. Skal der benyttes søgekoder/kombinatorer i de to søgesystemer – og i givet fald, hvilke?
5. Systemernes forskellige repræsentationsniveauer med baggrund i dels de uensartede dokument-katalogiseringer (primært DanBibs FORSK-delbase) foruden eventuelle browserproblemer i.f.m. repræsentationernes, eller de enkelte katalogfelters, fremvisning.
6. Ved evalueringen af dokumentrepræsentationerne skal man nøje overveje det bibliografiske beskrivelsesniveau for posterne – der skal foretages en harmonisering mellem systemernes repræsentationer for at opnå et reelt sammenligningsgrundlag. Dette kan gøres på forskellig vis, eksempelvis ved at tilpasse den ene systemtypes repræsentation til den anden ud fra princippet om "laveste fællesnævner" – svarende i dette tilfælde til f.eks. at fjerne abstracts fra DanWeb-posterne. Eller det kan omvendt gøres ud fra "højeste fællesnævner" ved at erstatte de fremfundne fuzzy-prototype poster med DanWeb-poster for de samme dokumenter, og lade målgruppen relevansvurdere disse i den rækkefølge, de er fremfundet i fuzzy-prototypen. Sidstnævnte er ideelt set den bedste løsning, da det giver målgruppen de bedste forudsætninger for at foretage en relevansvurdering. Med baggrund i en sådan harmonisering af posterne, er det kun fuzzy-systemets søgefunktion, der evalueres.

Sidstnævnte harmoniseringsmetode er valgt her, da "laveste fællesnævner"-princippet ikke levner målgruppen nogen reel chance for relevansvurdering af dokumentrne – metoden rummer for megen tilfældighed. Grunden til valget af web-udgaven fremfor den tegnbaserede version er, dels at denne repræsenterer det nyeste tiltag indenfor DanBib-udviklingen, dels at jeg finder den væsentlig mere overskuelig og brugervenlig i sin opbygning i.f.t. FindMenu-versionen og endelig ikke mindst, den udnytter en række brugervenlige funktioner, der ligger indbygget i WWW-konceptet, herunder browsing og hyperlinks. Til søgningerne i web-udgaven er valgt den kommandobaserede søgemulighed af flere grunde. Den primære er, at der hermed gives større mulighed for at afgrænse sig til bestemte søgeregistre, så et bedre sammenligningsgrundlag kan opnås. Samtidig er der mulighed for at kombinere mere end to søgeudtryk, som er en begrænsning ved den "Enkle", menustyrede søgemulighed. Der vil omkring materialetyper blive tale om en efterfølgende udlugning af poster, der repræsenterer AV-materialer, da fuzzy-prototypens datagrundlag kun rummer faglitterære bøger fra 1987→samt tss.-/avisartikler fra 1996→. Indsnævring på materialetyper er således ikke muligt i øjeblikket i DanWeb. Eventuelle dubletter fjernes, da disse giver et skævt billede ved relevansvurderingen. Til søgningerne i DanWeb er følgende indstillinger valgt, da de på bedste vis er sammenlignelige med fuzzy-prototypens implementerede indstillinger:

DanWeb-søgeindstillinger:

År: større end eller lig med 1987 (1987-)
Sprog: Alle
Base: Alle (FOLK, FORSK, MARC)
Materialetype: Alle (bl.a. monografier, periodica, AV-materialer)
Artikler: Med (artikler indekseret i Dansk Artikelindeks)

DanWeb emneords-søgekode:

ef : alle faglitterære, (DBC)-kontrollerede emneord (enkeltord)

Fuzzy DanBib-søgeindstillinger:

Standard (agg-funktion i midterposition (alm. gennemsnit), relaterede poster/emneord til grad 0.5 / 0.2). Standard-indstillingen er som udgangspunkt valgt for at kunne teste systemets præindstillinger og deres effekt på søgeresultaterne. I tilfælde, hvor søgeformuleringen er resulteret i *for* store søgesæt, er der i en efterfølgende søgning justeret på parametrene (relaterede poster/emneord) for at nedbringe søgesættets størrelse.

Forsøgspersonerne med deres forskellige problemområder og tilhørende selvvalgte søgestrenge vil nu blive præsenteret. Der inddrages eksempler på søgeresultaterne fra de to systemer – de søgninger, der præsenteres inde i selve teksten markeret med ☺, indgår senere i relevansbedømmelserne. For en komplet skematisk liste over søgestrenge

samt søgeresultater, henvises til selve hovedopgaven (Bøgeskov, 1999). Søgestrenge begyndende med "ef=" markerer søgningerne fra DanWeb, mens der ikke benyttes søgekoder ved fuzzy-søgningerne. Tallene i parentes angiver antal hits før udlugningen af eventuelle AV-materialer eller dubletter. For en samlet oversigt over disse dokumentrepræsentationer i hhv. DanWeb-, Fuzzy DanBib-, og harmoniseret Fuzzy DanBib-format, henvises ligeledes til hovedopgaven (Bøgeskov, 1999).

Målgruppen repræsenteres ved to lærerstuderende, der læser på sidste år ved Aalborg Lærerseminarium. Første forsøgsperson er *SKP*, mens den anden er *JHO*. Søgningerne tager alle udgangspunkt i deltagernes egne formulerede søgetermer ud fra et selvkomponeret problemområde/opgaveformulering. Kravet til disse "opgaver" har været, at de skal afspejle en typisk opgaveformulering indenfor deres domæne fra undervisningen på Aalborgs Lærerseminarium. For at opnå et vist repræsentativt søgegrundlag, har det dog været nødvendigt at supplere deres søgekombinationer med yderligere et par kombinationer.

SKP's problemområde, søgeord/søgestrenge til en tænkt opgave

Problemområde: Livskvalitet = sundhed?
Det ønskes undersøgt, hvorvidt legemlig sundhed er direkte medvirkende til en øget livskvalitet.

SKP's udvalgte søgetermer:
sundhed, livskvalitet, sundhedskampagner, kampagner, børneopdragelse, handlekompetence, livsstil, levevis, kost, madvaner, motion, idræt.

De selvvalgte søgekombinationer (for flere søgekombinationer, se (Bøgeskov, 1999)):
- sundhed OG livskvalitet
- sundhedskampagner OG handlekompetence
- kampagner OG handlekompetence
- livsstil OG madvaner
- levevis OG madvaner
- kost OG motion
- kost OG motion OG sundhed

Fuzzy DanBib indstillinger: Standard (agg-funktion i midterposition (alm. gennemsnit), relaterede poster/emneord til grad 0.5 / 0.2).

SKP's søgestrenge	*Antal poster:*
ef=sundhed	*433 poster*
sundhed	393 poster
ef=livskvalitet	*239 poster*
livskvalitet	232 poster
ef=sundhed ELLER ef=livskvalitet	*655 poster*
ef=sundhed OG livskvalitet	*17 poster*
sundhed livskvalitet (el. i omvendt rækkefølge)	Systemfejl
ef=sundhed OG ef=livskvalitet OG ef=motion	☺ *1 post*
sundhed livskvalitet motion	☺ 31 poster
sundhed livskvalitet motion (rel. poster/emneord 0.3 / 0.8)	685 poster
sundhed livskvalitet motion (rel. poster/emneord 0.7 / 0.2)	1 post
ef=sundhed OG ef=motion OG ef=trivsel	☺ *3 poster (4)*
sundhed motion trivsel	☺ 39 poster
ef=sundhed OG ef=motion OG ef=kost	☺ *2 poster*
sundhed motion kost (rel. poster/emneord 0.8 / 0.2)	☺ 6 poster

Efterfølgende vil 2. forsøgspersons søgestrenge blive omtalt.

JHO's problemområde, søgeord/søgestrenge til en tænkt opgave

Problemområde:
Med udgangspunkt i fremmedsprogedes generelle situation i undervisningen forsøges en belysning af, hvordan danskundervisningen kan effektiviseres overfor fremmedsprogede. Der er fra JHO's side ikke blevet formuleret nogle enkelt søgetermer, hvorfor søgekombinationerne præsenteres, som de er valgt.

Selvvalgte søgekombinationer (for flere søgekombinationer, se (Bøgeskov, 1999)):

- fremmedsprogede OG tosprogede OG flygtninge
- modersmålsundervisning OG "dansk for fremmedsprogede"
- modersmålsundervisning OG tosprogede
- modersmålsundervisning OG flygtninge
- integration OG folkeskolen OG samfundet
- undervisningsdifferentiering OG dansk OG "skriftlig dansk"
- kultur OG kulturforskelle OG kulturbaggrund

Fuzzy DanBib indstillinger: Standard (agg-funktion i midterposition (alm. gennemsnit), relaterede poster/emneord til grad 0.5 / 0.2).

JHO's søgestrenge	*ANTAL POSTER:*
ef=tosprogede elever OG ef=danskundervisning	*5 poster*
ef=fremmedsprogede OG ef=flygtninge	☺ *5 poster*
fremmedsprogede flygtninge	1027 poster
fremmedsprogede flygtninge (rel. poster/emneord 0.7 / 0.2)	☺ 18 poster
ef=modersmålsundervisning OG ef=tosprogede elever OG ef=danskundervisning	☺ *4 poster (5)*
modersmålsundervisning "tosprogede elever"	☺ 9 poster (10)
ef=folkeskolen OG ef=undervisning OG ef=undervisningsdifferentiering	☺ *20 poster*
folkeskolen undervisning undervisningsdifferentiering (rel. poster/emneord 0.7/0.2)	☺ 394 poster

Opsummering på søgningerne

Søgningerne på forsøgspersonernes søgestrenge bekræfter, at fuzzy-prototypen i høj grad fremfinder en lang række ekstra poster, i.f.t. den almindelige booleske søgning i DanWeb. Dette er ikke specielt overraskende for et system, der tilhører de partielle matchteknikker. Både funktionen "relaterede poster" såvel som "relaterede emneord" kan benyttes til afgrænsning af søgesættenes størrelse såvel som til udvidelse med flere poster.

Alt efter hvordan de præcise søgeparametre er indstillet, samt hvor mange søgetermer, der er valgt, varierer søgesættenes størrelser meget. Der kræves en vis tilpasning af indstillingerne for ikke at drukne i søgesættenes størrelse. En effektiv nedbringelse af antal poster kan fås ved gradvist at nedsætte værdien af "relaterede poster" til et passende søgesæt opnås. Efter et par søgninger ses det hurtigt, om en eventuel begrænsning på udvidelsen i termnettet (d.v.s. nedsætte rel. emneord) er nødvendig som supplement.

Relevans

Relevans-begrebet har om noget begreb indenfor informationsvidenskaben været genstand for en utrolig stor opmærksomhed og forskning lige siden det dukkede op første gang i forbindelse med forsøgene med at evaluere ydelserne fra de første "Information Retrieval" (IR)-systemer i 1950'erne, herunder effektivitetsmålingerne af Uniterm-indekseringen [6] samt Cranfield I og II (Hjørland, 1995) (Kidmose & Møller, 1995). I løbet af denne periode har der været vidt forskellige tilgange til, hvordan dette svært definèrbare begreb skal evalueres i.f.t. brugerne og deres mere eller mindre klart erkendte/formulerede informationsbehov. De første evalueringsmetoder (startende slutningen af 1950'erne), og stadig meget benyttede, bygger på det systemdrevne synspunkt, hvor relevans opfattes som noget håndgribeligt og måleligt, der kan bedømmes objektivt – relevans udtrykkes v.h.a. recall- og precision udregninger.

De mere brugerorienterede relevansopfattelser, der i løbet af 1970'erne begynder at dukke op, tager udgangspunkt i, at relevans opfattes som værende dynamisk afhængig af brugerens *subjektive* bedømmelse af informationen i relation til sit informationsbehov. Det er således ved sidstnævnte metode, modsat førstnævnte, kun brugeren selv, der kan foretage en reel relevansbedømmelse. Det ligger ikke indenfor artiklens rammer at give en komplet detaljeret redegørelse for alle de forskellige relevansopfattelser, der gennem tiden har været praktiseret.

Forsøgspersonernes relative relevansbedømmelser

I det følgende vil de to forsøgspersoners relevansvurderinger af de fremfundne poster i hhv. DanWeb samt Fuzzy DanBib blive præsenteret. Der benyttes i denne forbindelse en skala fra √ til √√√ for at indikere graden af relevans. Relevansbedømmelserne foretages ved tildelinger svarende til:

√ : *"ikke relevant/uegnet"*
√√ : *"mindre relevant"* – nødvendigt at se det konkrete dokument.
√√√ : *"relevant"*

SKP's SØGEORD:	*DanWeb - boolesk OG-komb.:*	*Relevansvurdering DanWeb-poster:*	*Fuzzy DanBib*	*Relevansvurdering Fuzzy DanBib-poster*
sundhed *motion* *trivsel*	4 poster (- 1 dublet) → *3 poster*	√√: 2 poster √√√: 1 poster	39 poster → vis første *20 poster* (Indstillinger: standard)	√: 1 post √√: 9 poster √√√: 10 poster
sundhed *livskvalitet* *motion*	*1 post*	√√√: 1 post	31 poster → vis første 20 poster (Indstillinger: standard)	√: 5 poster √√: 6 poster √√√: 9 poster
sundhed *motion* *kost*	*2 poster*	√√√: 2 poster	(Indstillinger: rel. poster til grad 0.8, rel. emneord til grad 0.2) 8 poster (- 2 dubletter) → *6 poster*	√√: 3 poster √√√: 3 poster

JHO's SØGEORD:	*DanWeb - boolesk OG-komb.:*	*Relevansvurdering DanWeb-poster:*	*Fuzzy DanBib*	*Relevansvurdering Fuzzy DanBib-poster*
folkeskolen undervisning undervisnings-differentiering	*20 poster*	√: 11 poster √√: 6 poster √√√: 3 poster	(Indstillinger: rel. poster til grad 0.7, rel. emneord til grad 0.2) 394 poster → vis første *20 poster*	√: 10 poster √√: 7 poster √√√: 3 poster
flygtninge fremmedsprogede	*5 poster*	√: 3 poster √√: 1 post √√√: 1post	(Indstillinger: rel. poster til grad 0.7, rel. emneord til grad 0.2) *18 poster*	√: 6 poster √√: 5 poster √√√: 7 poster
modersmåls-undervisning "tosprogede elever" danskundervisning	*4 poster*	√: 0 poster √√: 1 post √√√: 3 poster	(Indstillinger: standard) 10 poster (- 1 dublet) → *9 poster*	√√: 2 poster √√: 3 poster √√√: 4 poster

Opsummering på relevans-vurderingerne

Det bemærkes ved disse relative relevans-vurderinger, at søgninger foretaget i Fuzzy DanBib og efterfølgende harmoniseret til DanWebs repræsentationsniveau, forholdsvist ofte bedømmes (høj)-relevante (√√√) af målgruppen. Ud af de 6 søgninger, der er blevet foretaget, rummer 4 Fuzzy DanBib-søgesæt vurderinger, hvor de fleste poster anslåes som relevante. Hvis vi inkluderer dokumenter vurderet til √√ eller √√√, er det helt op til 5 ud af 6 søgesæt, hvor mere end halvdelen af posterne lever op til dette. Disse søgninger skal ses overfor de booleske søgninger foretaget i DanWeb, hvor posterne ligeledes får en forholdsvis høj relevansvurdering. En del af grunden til denne lighed er, at flere af de højst rangordnede fuzzy-poster, der er blevet fremvist for målgruppen, går igen som DanWeb-poster (jf. fuzzy-prototypens "best-match"-understøttelse).

Søgningen på termerne "sundhed motion trivsel" (se evt. Bøgeskov, 1999, bilag 4-6) viser for fuzzy-posterne tydeligt en tendens af, hvad relaterede emneord kan bibringe. Flere af posterne er således indekseret med termer som "kost" eller "ernæring", hvilket forsøgspersonen i flere tilfælde har anført (overfor undertegnede) som en vigtig faktor ved relevansvurderingen. Når vi skal se nuanceret på resultaterne, så kan det iagttages, at JHO's søgeresultater er præget af forholdsvis lave relevansvurderinger, både hvad angår DanWeb-posterne samt fuzzy-posterne. Én af årsagerne til denne tendens er, at forsøgspersonen havde foretrukket, at søgeresultaterne, der inkluderer aspekter omkring undervisningsdifferentiering, i højere grad havde afspejlet aspekter omkring flygtninge og således ikke primært "svage danske elevers" læse- eller skrivevanskeligheder.

Jeg vil tillade mig at opsummere resultaterne fra min mini-test, der fra starten kun har været tænkt som en lille indikation af en tendens, og derfor naturligvis skal ses med de rette forbehold. Fuzzyficeringen i Fuzzy DanBib har en række oplagte fordele for brugeren i en søgesituation. Systemet rummer mulighed for inddragelse af poster, der ikke normalt vil blive indfanget ved en boolesk søgning, der *kun* understøtter "eksakt match" mellem query og dokumentrepræsentation. Dette giver en række fordele i forhold til opfyldelsen af brugerens mere eller mindre klart erkendte og/eller formulerede informationsbehov. Jeg mener, at systemet primært har sin force overfor

brugere med dels et *bevidst* emneafgrænset informationsbehov, dels et *mudret* emneafgrænset informationsbehov, da systemet rummer mulighed for høj recall ved fuzzy-søgningerne. Hvis der derimod er tale om et *verifikativt* informationsbehov, vil de booleske søgesystemer som DanBib eller DanWeb stadig være klart at foretrække p.g.a. deres forbedrede søgemulighed for høj precision. Specielt 1. forsøgsperson, SKP, besad tydeligvis en stor viden om sit emneområde og paradigmerne indenfor sit felt. Desuden var det tydeligt at se ved ham, men også ved JHO, at deres informationsbehov ændredes undervejs, var dynamisk, i kraft af at flere af de relaterede fuzzy-dokumenter, vakte deres interesse undervejs i relevansbedømmelsesprocessen. Ofte var det dog også andre oplysninger end lige netop emneordene, der afgjorde, at ét dokument blev vurderet højere. Typiske eksempler fra dokumentrepræsentationerne var oplysninger om form, udgivelsesår samt målgruppeniveau. Dette understreger blot, at beskrivelsesniveauet for dokumentrepræsentationerne spiller en vigtig rolle ved brugernes relevansbedømmelser.

Konklusion

Denne artikel indeholder en overordnet gennemgang af eksakte match- samt partielle match-søgeteknikker generelt, idet der tages udgangspunkt i en klassifikationsmodel udviklet af Belkin & Croft (1987). I modellen placerer fuzzy logic sig under de partielle matchteknikker, da søgeteknikken rummer mulighed for inddragelse af poster, der kun matcher query delvist. Søgeteknikken viser desuden, at den understøtter opfyldelse af booleske søgeudtryk, hvorfor fuzzy logic af samme grund ofte omtales som en generalisering af boolesk logik. Der redegøres i opgaven for et projekt ved DBC omkring en prototype, kaldet "Fuzzy-søgning i DanBib", der gennem et par år har været under udvikling som laboratorie-eksperiment i et samarbejde mellem DBC og RUC's ISL-institut. Undertegnedes erfaringer ved brug af denne prototype, kombineret med erfaringerne fra en mini-test, der foretages med to forsøgspersoner, giver samlet et billede af et informationssøgesystem, der giver mulighed for høj recall, og i denne sammenhæng primært vil understøtte brugere med et bevidst emneafgrænset eller mudret emneafgrænset informationsbehov. Fuzzy-prototypen benytter sig ved sine søgninger af et underliggende semantisk termnet, hvor associationerne mellem de kontrollerede, faglitterære emneord udregnes statistisk ud fra en formel, der tager hensyn til, hvor ofte ordene optræder sammen. Således vil dokumenter indekseret med emneord A, der samtidig indeholder emneord B, medføre en "associationsgrad" mellem A og B. Vigtigt i denne sammenhæng er det at fastslå, at det ikke automatisk er det samme som, at der er en semantisk relation mellem disse - dette *kan* dog være tilfældet. I denne forbindelse skal fuzzy-prototypens semantiske termnet ses primært som en afspejling af indeksørernes videnstrukturer, da det er de kontrollerede, faglitterære emneord, der benyttes som termnet og søgeudgangspunkt i prototypen. Samtidig ligger der overordnet en selektiv beslutning om, hvilke data der skal benyttes som grundlag for databasen. Der er derfor implicit en afspejling af også ledelsens, systemudviklernes og indekseringssystem-udviklernes videnstrukturer. Hvad angår indekseringen af DanBib, hvis emneordsdatagrundlag fuzzy-prototypen bygger på, så er den uensartet i alm. DanBib, men kontrolleret i prototypen. Indekseringsniveauet, d.v.s. den høje exhaustivitet, der præger mange af DanWeb posterne, får ikke nødvendigvis den ønskede effekt i prototypen. Dette skyldes, at prototypens relevansvurdering, rangordning af poster, typisk vil diskriminere fattige poster (poster med få emneord) fra DanWeb, der kan vise sig at være meget relevante, jf. høj specificitet. Systemet bygger således på en udvidelse af beslægtede poster, hvor poster med mange emneord, høj exhaustivitet, typisk vil blive tilgodeset, i og med der ved en søgeforespørgsel er en potentiel sandsynlighed for, at ét af disse emneord vil blive inddraget som relateret emneord. At systemet samtidig ikke i nødvendigt omfang er garant for en søgning med høj precision, skal primært ses ud fra de manglende søgemuligheder, der indsnævrer sig til søgning på de kontrollerede, faglitterære emneord. Således ingen mulighed for søgning i eksempelvis titel-, forfatter-, eller notefelterne. Min mini-test indikerer, at fuzzy-prototypen fremfinder forholdsvist høj-relevante dokumenter, dette bl.a. set ud fra inddragelsen af poster med relaterede emneord. Set i.f.t. de forskellige typer af relevans, som bl.a. Saracevic (1996) redegør for, er det tyde-

ligt, at fuzzy-prototypen i sin udformning understøtter flere forskellige måder, hvorpå brugeren kan relevansvurdere de fremfundne dokumenter. Der spores en understøttelse af dels kognitive, situationelle samt emnemæssige relevansaspekter. Dette skal ses ud fra, at systemet med sit semantiske emneordsnet og mulighed for browsing gennem emneordshierarkier let vil kunne inddrage relaterede dokumenter ved søgninger eller anspore brugeren til at søge på relaterede termer. Da brugerens informationsbehov samtidig ofte ikke er klart erkendt/formuleret, jf. etikette-effekten, og systemet tilgodeser en dynamisk udvikling af informationsbehovet hos brugeren, rummer systemet en række indlysende fordele som alternativ til den booleske søgning, som vi trods alt ikke helt vil eller kan undvære. Ved at afgrænse forsøgsgruppen til kun to personer, er der en lang række indlysende forbehold, der nødvendigvis må tages, når effekten af prototypens fussyficeringer skal konkluderes. Den metodiske fremgangsmåde med at harmonisere fuzzy-posterne til DanWeb-præsentationsniveau kunne derfor være interessant at benytte i en større empirisk undersøgelse med flere testpersoner inddraget. Jeg mener, at metodikken stadig holder og dermed udsiger noget om systemets evne til at inddrage yderligere relevante dokumenter v.h.a. sin funktion omkring "automatisk spørgsmålsudvidelse". Fuzzy DanBib har en række indlysende fordele implementeret, men der er samtidig en del mangler, der bør tages hensyn til ved en senere implementering af de høstede erfaringer i DanWeb. Jeg savnede bl.a. en raffineringsmetode, relevans feedback, med guidende forslag til at understøtte og hjælpe brugeren undervejs i søgningen – f.eks. i tilfælde med alt for store søgesæt, såvel som for små (e.v.t. 0 hits). Samtidig kunne der i.f.m. rangordningen af posterne efter faldende relevans e.v.t. tages hensyn til termsammenfald i emneords- samt titelfeltet - dette ville give en mere differentieret rangordning af posterne. Som det er nu, vil poster der f.eks. indeholder 2 søgetermer ud af 3 anførte (i søgeforespørgslen), alle blive rangordnet med samme grad 0.67 – dette er ikke en specielt nuanceret graduering.

Ellers vil jeg afrundingsvis sige, at fuzzy logic som implementeret søgeteknik eller søgealternativ i DanBib, lover godt for de kommende revisioner af DanWeb. Søgeteknikken indeholder en række indlysende søgefordele overfor brugerne.

Noter

1. CCL: Command Command Language, kommandosprog udviklet i EU-regi med det formål at standardisere og lette søgninger i databaser (Informationsordbogen, 2. udg., 1991)
2. ISL: Institut for Intelligente Systemer
3. iflg. Indeksering af faglitteratur, 1998.
4. Zipf, George K.: *Human Behaviour and the Principle of the Least Effort* (jf. Hjørland, 1995, s. 443)
5. Kontrollen af disse emneord foregår v.h.a. DBC's separate emneordsbase (Indeksering af faglitteratur, 1998, s. 15).
6. Der er spor af IR-forskning helt tilbage til 1953, hvor der i både USA og England blev udført tests med det formål at måle effektiviteten af Uniterm-indeksering i.f.t. konventionelle indekseringssystemer (jf. Ellis (1990), iflg. Kidmose & Møller, 1995, s. 45-46)

Referencer

1. *von Altrock, C. (1995): Fuzzy Logic and NeuroFuzzy Applications Explained.*- Upper Saddle River: Prentice-Hall PTR.- 350 sider : ill.
2. *Andreasen,* T. & *Schomacker,* T. *(1997): DanBib – a union catalogue applied for user friendly querying.* [online]. DBC.- Tilgængelig på Internet: <URL: http://www.dbc.dk/udvikl/ifla97v9.html> (Set: 28/1-1999)
3. *Andreasen,* T. (1998): *Fuzzy Logik i DanBib.*- I: Referencen, nr. 1, s. 13-19
4. *Belkin,* Nicholas J. & Croft, Bruce W. (1987): *Retrieval Techniques.*- I: Annual Review of Information Science and Technology (ARIST), vol. 22, s. 109-145
5. *Brugervejledning til Fleksibel søgning i DanBib – Prototype 2.0 (1999):* Brugervejledning og div. synopser/kopier af dias.- Ballerup: DBC (Udleveret DBC konference-materiale d. 29/1-1999)
6. *Bøgeskov,* J. (1999): *Fuzzy-søgning i DanBib – et relevant søgealternativ?* (DB-Aa Hovedopgave 1999)
7. *Carlson,* K. (1997): *Fuzzy logic – flexible*

searching: attempts of developing a more user friendly interface to DanBib. [online]. DBC.- Tilgængelig på Internet: <URL: http://www.dbc.dk/english/flexfuz.html> (Set: 19/1-1999)

8. *Forsberg,* A. [et al.] (1999): *En eksperimentel søgeflade til DanBib.*- I: Viden om, nr. 1.- s. 3-9 (Særnummer om Eksperimentelle søgemetoder)
9. *Frants,* Valery I. [et al.] *(1999): Boolean search: Current State and Perspectives.*- I: Journal of the American Society for Information Science, 50(1), s. 86-95
10. *Havnø,* P. & *Hansen,* Lizzi S. *(1994): Informationsgenfinding: Partial match søgeteknikker.*- I: Red. Anders Ørum, Biblioteksarbejde, nr. 41, s 41-54.- Aalborg: Danmarks Biblioteksskole
11. *Hjørland,* B. *(1993): Emnerepræsentation og informationssøgning – bidrag til en teori på kundskabsteoretisk grundlag.*- Borås: Valfrid.- 259 sider
12. *Hjørland,* B. *(1995): Informationsvidenskabelige grundbegreber (Bind I og II).*- 2. rev. udg..- København: Danmarks Biblioteksskole.- 457 sider
13. *Hjørland,* B. *(1996): Informationsbehov – en analyse af et vanskeligt begreb.*- I: Bogens Verden, August, s. 268-271
14. *Indeksering af faglitteratur (1998).*- Ballerup: DBC as.- 125 sider (DBC's indekseringsvejledninger)
15. *Ingwersen,* P. & *Willett,* P. *(1995): An Introduction to Algorithmic and Cognitive Approaches for Information Retrieval.*- I: Libri, vol. 45, s. 160-177
16. *Ingwersen,* P. & *Wormell,* I. *(1993): Informationsformidling i teori og praksis.*- 2. oplag.- Kbh: Munksgaard.- 104 sider
17. *Ingwersen,* P. *(1992): Information Retrieval Interaction.*- London: Taylor Graham Publishing.- 246 sider
18. *Kidmose,* B. & *Møller,* Christian Ø. *(1995): Relevant om relevans – teori og metodologi.*- I Biblioteksarbejde, nr. 45, s. 37-53
19. *Nielsen,* U. *(1997): Fuzzy logic i DanBib.*- I: Referencen, nr. 5, s. 5-10
20. *Nørgaard,* H. *(1998): Fuzzy logik.*- I: Red. Suzanne Hemmeth Christoffersen, Nye veje til viden – Om intelligente agenter og fuzzy logic, s. 29-34.- Ballerup: DBC (pamflet nr. 2)
21. *Projects in ISL:* Flexible Search in DanBib *(1999).* [online]. RUC.- Tilgængelig på Internet: <URL: http://www.isl.ruc.dk/projects.html> (Set: 19/1-1999)
22. *Saracevic,* T. *(1996): Relevance reconsidered '96.*- I: Red. Peter Ingwersen & Niels Ole Pors: Information Science – Integration in Perspective, s. 201-218 (Proceedings CoLIS 2 : Second International Conference on Conceptions of Library and Information Science, October 13-16, 1996)
23. *Schneider,* M. [et al.] *(1996): Fuzzy Expert System Tools.*- Chichester: John Wiley & Sons.- 198 sider
24. *Schomacker,* T. *(1996): Fuzzy logik – fleksibel søgning.*- I: Viden om, nr. 4.- s. 6-7
25. *Sinding,* E. *(1998): Søgevejledning og manual til DanBib – tegnbaseret.*- [Kbh]: Danmarks Biblioteksskole.- 49 sider