TARGET OG BOOLE

- anvendelse af TARGET-kommandoen og kognitiv quorumsøgning samt de bagvedliggende teorier

Af Karen Nedergaard

Abstract

Artiklen tager udgangspunkt i TARGET-kommandoen, der i december 1993 blev introduceret af databaseværten Dialog. Den praktiske anvendelse af TARGET-kommandoen bliver belyst og set i sammenhæng med udnyttelsen af viden om de kognitive strukturer, der findes i tilgængelige databaser. To søgemetoder kognitiv quorumsøgning og TARGET, er eksemplificeret ved søgning i INSPEC. Herudover bliver den formodede teoretiske baggrund for TARGET, nemlig vektorrumsmodellen, gennemgået.

1. Introduktion

Informationssøgning i databaser skete indtil december 1993 ved hjælp af information retrieval (IR) teknikken exact match. I december 1993 indførte Dialog TARGET-kommandoen, der er baseret på en form for partiel match IR teknik. "TARGET results are determined by a statistical analysis of the occurrences of your search terms to provide records that are relevant to your topic. These records are then ordered starting with the most relevant" (Dialog, 1993, s. 1).

Denne teknik bygger blandt andet på følgende fire elementer (Dialog, 1993, s. 2):

- The number of times each term is mentioned in the record
- Which terms appear in the record
- The proximity of your terms to each other
- How often your term(s) appears in the database

Dialog har ikke offentliggjort den algoritme, der ligger til grund for den statistiske beregning af dokumenters relevans i forhold til søgestrengen, men alt tyder på, at der benyttes en eller anden form for vektorrumsmodel, som tager hensyn til termers forekomst i dokumenter og i hele databasen. Der er ikke tale om nogen form for indholdsmæssig relevans, men en statistisk udregnet relevans, der blandt andet bygger på de ovennævnte elementer og tager udgangspunkt i ligheden mellem query og dokumenter.

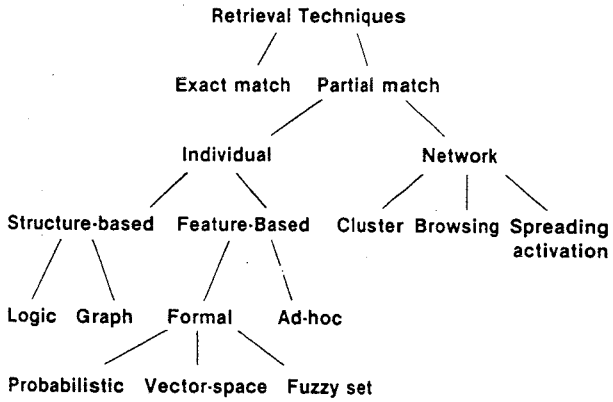
Indtil Dialog indførte TARGET-kommandoen blev partiel match teknikker kun afprøvet i mindre testdatabaser, startende med Saltons SMART system, med varierende resultat (Belkin og Croft, 1987). De store kommercielle systemer, som for eksempel Dialog og ESA-IRS, benytter den traditionelle exact match teknik, dog i de senere år raffineret med blandt andet ZOOM og RANK kommandoer der bearbejder søgeresultater, foretaget ved boolsk søgning, på en kvalitativ måde. Med TARGET er en partiel match teknik introduceret hos en stor databasevært, hvorfor det for første gang er muligt at afprøve bæredygtigheden af en partiel match teknik i et kommercielt system og lave sammenlignende undersøgelser af de to genfindingsteknikker.

Afsnit 2 vil behandle IR-teknikker overordnet og kort beskrive vektorrumsmodellen, for at give et billede af, hvad der ligger til grund for teknikken bag TARGET. Afsnit 3 handler om quorumsøgning, der tager hensyn til de forskellige kognitive strukturer, der er repræsenteret i IR-systemet. I afsnit 4 gives et eksempel på en søgning i INSPEC hos Dialog. Formålet er at finde ud af, om der er en signifikant forskel på søgeresultater ved anvendelse af de to IR-teknikker, og om en eventuel forskel kan udnyttes ved søgning. Er TARGET et hjælpemiddel til novicer, eller kan erfarne informationssøgere have udbytte af at anvende TARGET-kommandoen?

2. Informationsgenfindingsteknikker

Belkin og Croft (1987, s. 109) fremhæver IR-systemers formål som værende: "identifying, retrieving, and/or ranking texts (or surrogates or portions of texts), in a collection of texts, that might be relevant to a given query". Til dette formål kan man anvende mange forskellige teknikker, se nedenfor, med eller uden mulighed for at rangordne de fundne dokumenter.

Belkin og Croft (1987, s. 112) definerer genfindingsteknikker på denne måde: "a technique for comparing the query with the document representations". Belkin og Crofts klassifikation af genfindingsteknikker:



A classification of retrieval techniques. (Fra Belkin og Croft, 1987, s. 112)

Det følgende vil koncentrere sig om vektorrum, da det er denne model, der p.t. regnes for at ligge bag TARGET-kommandoen hos Dialog.

Den store mængde forsøg med partiel match teknikker har vist, skønt der er problemer ved at sammenligne søgesæt fra exact match med sæt fra partiel match, at partiel match teknikker giver bedre resultater. Belkin og Croft (1987, s. 124) betegner denne forskel som signifikant, og med signifikant, mener de, at der skal være en forbedring i precision og recall på mindst 10%.

2.1 Exact match teknikker

"Exact match techniques are those that require that the request model be contained, precisely as represented in the query formulation, within the text representation" (Belkin og Croft, 1987, s. 113). Boolsk logik er idag den oftest implementerede exact match teknik og den mest anvendte IR-teknik.

Søgning ved hjælp af exact match har dog flere begrænsninger:
- genfinder ikke de dokumenter, der kun delvist matcher query
- rangordner ikke de fundne dokumenter

- tager ikke højde for den relative vægt, som elementer i query eller dokumentet kan have
- boolsk logik kræver kompliceret query-formulering
- kræver at query og dokument har samme sprogbrug
- not-operatoren vil altid fjerne poster, der indholder relevante termer (dette er dog meningen i en quorumsøgning, hvor not-operatoren er nødvendig).

På trods af disse begrænsninger er det dog stadig exact match, der anvendes overalt. Økonomi er en vigtig grund til, at det stadig er exact match teknikker, de store databaseværter benytter, fordi det vil være bekosteligt at indføre en ny genfindingsteknik. En anden grund til at anvende exact match er den, at boolske queries er strukturerede, så de repræsenterer aspekter af brugerens informationsbehov (Ingwersen, 1992, s. 72). Indtil TARGET blev introduceret, har man forsøgt at kompensere for exact matchs begrænsninger ved hjælp af for eksempel muligheden for trunkering og kommandoerne ZOOM og RANK, der afhjælper problemerne med genfinding af kun delvist matchende dokumenter og rangordning af træk ved fundne dokumenter.

2.2 Partielle match teknikker

Partielle match teknikker adskiller sig fra exact match ved at være teoretisk funderede som følge af de mange forsøg i mindre testbaser og ved at rangordne søgeresultatet. Ved søgning med partielle match teknikker bruger man query i naturligt sprog, som ved hjælp af en stopordsliste og stemming (identificering af ordstammer) omformes til query og herefter sammenlignes med dokumenter i databasen. Ved brug af en matematisk model identificeres ligheder mellem query og dokumenter. De identificerede dokumenter vises i rækkefølge med dokumentet med den største lighed først (Belkin og Croft, 1987). TARGET er en partiel match teknik, der dog ikke kan behandle en query i naturligt sprog, da den ikke indeholder funktioner til at fjerne stopord og til at stemme. Derfor skal der ved brug af TARGET-kommandoen stadig foretages en form for query formulering, hvor man angiver informationsbærende termer, synonymer m.m.

Partielle match teknikker opdeles i individuelle og netværksbaserede teknikker. De individuelle teknikker baserer sig på relationer mellem dokumentkarakteristika, f.eks. ord, hvorimod de netværksbaserede baserer sig på relationer mellem dokumenter med lignende indhold, f.eks. klynger.

2.2.1 Vektorrumsmodellen - et eksempel på en partiel match teknik

Som eksempel på en individuel, feature-baseret teknik kan nævnes vektorrumsmodellen, der i korthed går ud på at dokumenter og queries repræsenteres som vektorer i et n-dimensionalt rum, (d_1, d_2, ..., d_n), hvor d_i angiver tilstedeværelsen eller fraværet af den i'te term og i overensstemmelse hermed har værdien 1 eller 0, og hvor n er antallet af indextermer i databasen.

Vektorer er linjestykker med en længde og en retning. Ligheden mellem to vektorer beregnes ved hjælp af et lighedsmål. Ved eksempelvis at udregne cosinus til vinklen mellem to vektorer opnås en indikation af ligheden mellem disse vektorer. Jo mindre vinkel, cosinus nærmer sig 1, jo større lighed. Der findes flere lighedsmål, cosinus til vinklen er blot et eksempel.

For at opnå en mere nuanceret repræsentation kan vektorernes koordinater, svarende til termerne, vægtes i query og/eller i et dokument. Der findes flere termvægtningsmodeller, bl.a. Saltons traditionelle metode, hvor frekvensen af termen i et dokument ses i forhold til det samlede antal termer i dokumentet. I det følgende vil Crofts termvægtningsmetode (Croft, 1993) blive anvendt. Croft ser frekvensen af en term i et dokument i forhold til frekvensen af den term, der forekommer hyppigst i dokumentet.

For at se termvægtene i forhold til andre termer og dokumenter udregnes termvægte ofte som tf.idf vægte, hvor tf, termfrekvens, er termens forhold til dokumentet og idf, inverteret dokumentfrekvens, termens forhold til hele samlingen.

Sparck Jones (1973, s. 621) fremsætter tre antagelser om termvægtnings betydning:

1. forekomsten af en term i et dokument er signifikant
2. forekomsten af en term i et kort dokument er mere signifikant end termens forekomst i et langt dokument
3. forekomsten af en sjælden term i et dokument er mere signifikant end forekomsten af en hyppig term.

For at tage hensyn til punkt 1 og 2 ses den rå termfrekvens i forhold til dokumentet, svarende til tf. Dette kan for eksempel ske ved at se på forekomsten af unikke termer, den hyppigst forekommende term eller alle termer i dokumentet. Punkt 3 tages i betragtning ved at se hele samlingens størrelse i forhold til forekomster af termen i samlingen, hvilket svarer til den inverterede dokument frekvens, idf.

I det følgende eksempel vil Crofts termvægtningsmetode og Cosinusligheden blive anvendt. I figur 1 ses formlen for Crofts termvægtningsmetode og i figur 2 formlen for Cosinusligheden.

Figur 1: Crofts termvægtningsmetode

I det følgende er:

tf_{ij} = frekvensen af term *i* i dokument *j*
$\max\{tf_{ij}\}$ = frekvens af hyppigst forekommende term i dokumentet
N = antallet af dokumenter
df_i = antallet af dokumenter der indeholder term *i*

$$tf \cdot idf = \frac{tf_{ij}}{\max\{tf_{ij}\}} \cdot \log\frac{N}{df_i}$$

Figur 2: Cosinusligheden

w_{qi} = den i'te queryterms tf.idf-vægt
w_{di} = den i'te dokumentterms tf.idf-vægt

$$sim(Q,D) = \frac{\sum w_{qi} \cdot w_{di}}{\sqrt{\sum (w_{qi})^2 \cdot \sum (w_{di})^2}}$$

Eksemplet består af en samling dokumenter, N, der indeholder 5 dokumenter og en query. I den viste matrix over samlingen er der suppleret med en enkelt oplysning, der normalt ikke indgår i den inverterede fil. Dette er $\max\{tf_{ij}\}$, frekvensen af den hyppigst forekommende term i dokumentet. Dette tal skal bruges for at kunne udregne termvægte ifølge Crofts termvægtningsmetode (figur 1). Værdien df_i, antallet af dokumenter, der indeholder en bestemt term, findes allerede idag, som det tal, der vises ved en expand kommando.

N=5	Q	D_1	D_2	D_3	D_4	D_5	df_i
t_1	1	2	1	2	0	1	4
t_2	2	3	2	1	1	0	4
t_3	0	1	0	1	2	4	4
t_4	1	1	1	0	0	1	3
$\max\{tf_{ij}\}$	2	3	2	2	2	4	

Termvægte for query og de fem dokumenter, udregnet ved hjælp af Crofts termvægtningsmetode (figur 1) bliver:

	Q	D_1	D_2	D_3	D_4	D_5
t_1	0,024	0,028	0,024	0,048	0,000	0,016
t_2	0,048	0,042	0,048	0,024	0,024	0,000
t_3	0,000	0,014	0,000	0,024	0,048	0,065
t_4	0,055	0,032	0,055	0,000	0,000	0,037

I figur 3 kan udregningen af termvægte for query og dokument 1 ses.

I figur 4 kan udregning af Cosinusligheden ses.

Figur 3: Udregning af termvægte for query og dokument 1

$$Q,t_1 = \frac{1}{4} \cdot \log\frac{5}{4} = 0{,}024 \qquad D_1,t_1 = \frac{2}{7}$$

$$Q,t_2 = \frac{2}{4} \cdot \log\frac{5}{4} = 0{,}048 \qquad D_1,t_2 = \frac{3}{7}$$

$$Q,t_3 = \frac{0}{4} \cdot \log\frac{5}{4} = 0{,}000 \qquad D_1,t_3 = \frac{1}{7}$$

$$Q,t_4 = \frac{1}{4} \cdot \log\frac{5}{3} = 0{,}055 \qquad D_1,t_4 = \frac{1}{7}$$

Disse termvægte anvendes når det, ved hjælp af Cosinuslighedsmålet, skal udregnes hvilke dokumenter, der ligner query mest.

Ligheden mellem query og dokumenter udregnet ved hjælp af Cosinusligheden er som følger:

Q,D_1	Q,D_2	Q,D_3	Q,D_4	Q,D_5
0,943	1,000	0,570	0,279	0,412

Figur 4: Udregning af cosinusligheden for query og dokumenterne

$$sim(Q,D_1) = \frac{(0{,}024 \cdot 0{,}028) + (0{,}048 \cdot 0{,}042) + (0{,}000 \cdot 0{,}014) + (0{,}055 \cdot 0{,}032)}{\sqrt{(0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2) \cdot (0{,}028^2+0{,}042^2+0{,}014^2+0{,}032^2)}} = 0{,}943$$

$$sim(Q,D_2) = \frac{(0{,}024 \cdot 0{,}024) + (0{,}048 \cdot 0{,}048) + (0{,}000 \cdot 0{,}000) + (0{,}055 \cdot 0{,}055)}{\sqrt{(0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2) \cdot (0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2)}} = 1{,}000$$

$$sim(Q,D_3) = \frac{(0{,}024 \cdot 0{,}048) + (0{,}048 \cdot 0{,}024) + (0{,}000 \cdot 0{,}024) + (0{,}055 \cdot 0{,}000)}{\sqrt{(0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2) \cdot (0{,}048^2+0{,}024^2+0{,}024^2+0{,}000^2)}} = 0{,}510$$

$$sim(Q,D_4) = \frac{(0{,}024 \cdot 0{,}000) + (0{,}048 \cdot 0{,}024) + (0{,}000 \cdot 0{,}048) + (0{,}055 \cdot 0{,}000)}{\sqrt{(0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2) \cdot (0{,}000^2+0{,}024^2+0{,}048^2+0{,}000^2)}} = 0{,}279$$

$$sim(Q,D_5) = \frac{(0{,}024 \cdot 0{,}016) + (0{,}048 \cdot 0{,}000) + (0{,}000 \cdot 0{,}065) + (0{,}055 \cdot 0{,}037)}{\sqrt{(0{,}024^2+0{,}048^2+0{,}000^2+0{,}055^2) \cdot (0{,}016^2+0{,}000^2+0{,}065^2+0{,}037^2)}} = 0{,}412$$

Disse lighedsmål viser, at denne dokumentsamling kan rangordnes som værende relevant for den givne query på følgende måde:

D_2, D_1, D_3, D_5, D_4

Dokument 2 er identisk med query og skal derfor rangordnes først og have en lighed på 1. Herefter følger dokumenterne 1 og 3, hvor 1 kommer først fordi hyppigheden af querytermer i dokumentet er større end i dokument 3. Dokument 4 kommer til sidst fordi det kun indeholder en queryterm i forhold til dokument 5, der indeholder to querytermer.

Vektorrumsmodellen er en meget robust partiel match teknik, der tager hensyn til både de søgte termer, men også de øvrige termer i dokumenterne, og både til termernes forkomst i dokumenter og i hele databasen. Dette medfører dog også, at det er en forholdsvis besværlig teknik at anvende i operationelle systemer, der skal både udregnes vægte for queryterme, men også for alle andre termer i dokumenterne. Derfor er det ikke sandsynligt, at TARGET anvender lige netop denne model for vektorrum, men en tillempet model, hvor der tages hensyn til querytermernes vægte og formodentligt også til antal ord i dokumenterne og i databasen. I et oprationelt system vil det tage for lang tid at anvende vektorrumsmodellen i sin fulde udstrækning, hvor der udregnes vægte for alle termer i alle dokumenter, der er relevante for query.

3. Onlinesøgemuligheder med exact match teknik og partiel match teknik

I det foregående er teknikken bag partiel match teknikken, vektorrum, omtalt for at give baggrund for dette afsnit, der handler om, hvordan den nye søgefacilitet hos Dialog, TARGET, kan udnyttes bedst muligt.

Dialog introducerede i december 1993 TARGET, som den kommando der skal redde os fra problemerne med at søge i fuldtekstdatabaser, med komplicerede kommandosprog, med boolske nærhedsoperatorer og sidst men ikke mindst det, at man traditionelt ikke kan få rangordnet poster. (Dialog, 1993). TARGET er for dem, der ikke kan udnytte den boolske søgelogik, og for dem, der skal søge i et fagområde, de ikke kender til. Sådan ser det umiddelbart ud, men sådan bør det ikke være. Kan TARGET være til nytte for IR-specialister? Kan man kombinere boolsk søgning med TARGET søgning på en fornuftig måde? Dette er spørgsmål, der så småt er ved at blive besvaret af IR-specialister. I det følgende vil en metode til at anvende TARGET og boolsk søgning blive foreslået. I denne forbindelse bliver det kognitive synspunkt inddraget i IR-processen, da det kan forbedre søgemulighederne ved traditionel boolsk søgning. Overlap, graden af enighed mellem to sæt af fundne dokumenter fra to forskellige søgninger udført for samme query (Pao, 1994), vil også blive inddraget, da flere undersøgelser viser at overlap ofte har stor relevans for query. Overlap kan dannes på flere måder, i det følgende vil forskellige kognitive strukturer og forskellige IR-teknikker blive eksemplificeret.

Kognitive strukturer har stor indflydelse på IR-processen, da der i IR-processen optræder mange aktører med individuelle kognitive strukturer. Med kognitiv menes, at hvert individ har sin egen måde at opfatte omverdenen på, afhængigt af sine tidligere erfaringer og viden (Ingwersen og Wormell, 1988). Ifølge Ingwersen (1994) er disse forskellige aktører i IR-processen: systemdesignere og producenter, udviklere af IR-teknik, konstruktører af indekseringsregler, indeksører, ophav til tekster og billeder, designere af intermediærmekanismer og brugere i en domænerelateret kontekst. Hver af disse aktører tilføjer sin egen omverdensopfattelse til IR-processen. Disse mange forskellige kognitive strukturer kan udnyttes i søgeøjemed.

Ingwersen foreslår, at man ved hjælp af disse aktørers forskellige kognitive strukturer kan afhjælpe visse problemer i IR-processen: "In order to reduce the uncertainties as well as the unpredictabilities in IR and ease the perception and inter-

pretation the cognitive viewpoint suggests to provide and make use of *different cognitive structures* during acts of communication, i.e. structures of *different cognitive and functional origin* - on both sides of the communication channel." (1994, s. 102)

I det følgende vil brugernes kognitive strukturer ikke blive behandlet, men der vil blive fokuseret på de kognitive stukturer, der er repræsenteret i databaser og i forskellige IR-teknikker.

I databaser findes der flere kilder til kognitive strukturer, for eksempel systemdesignere, ophav og indeksører. Det er på forhånd givet hvilke felters indhold, i en given database, der stammer fra ophavet eller fra indeksøren. I INSPEC, som vil blive brugt i et senere eksempel, er det indeksørens kognitive strukturer, der ligger bag ved descriptor og identifier felterne og forfatterens kognitive strukturer, der ligger bag ved titel og abstract felterne. I citationsindexer er det forfatterne, der er ophav til descriptorerne, men her er de kognitive strukturer fra de forfattere, der bliver citeret også repræsenteret. Med denne viden kan man udnytte disse forskellige kognitive strukturer i søgeøjemed som en form for polyrepræsentation. Jo flere forskellige synsvinkler der samtidigt findes på et dokument, jo større er chancen for, at der er match mellem brugerens kognitive struktur og de strukturer, der repræsenterer dokumentet.

Ifølge Ingwersen kan polyrepræsentationer forbedre muligheden for at genfinde dokumentrepræsentationer: "Intentional redundancy on both the system and user side is assumed to provide improved access to structures for both participants during IR interaction. The principle is proposed carried out by *polyrepresentative* means. By polyrepresentation is meant the representation in a variety of different forms of one information object, e.g. of an information requirement or a text entity." (Ingwersen, 1994, s. 105).

Ved at udnytte polyrepræsentationer kan der opnås et tilsigtet overlap, der kan give bedre søgeresultater, end hvis disse forskellige repræsentationer blev udnyttet hver for sig. Som eksempel nævner Ingwersen kontrollerede fraser og naturlige sprog: "... the combination - *the overlap* - of controlled index *phrases* and natural language representations from basic index yields the best retrieval results, better than the two seperately. The more variety in cognitive origin of method, the more different results." (Ingwersen, 1994, s. 105). Det vil sige, at udnytter man de kognitive strukturer, der findes, til at skabe et overlap, fås et bedre søgeresultat, end hvis disse strukturer blev udnyttet hver for sig.

Pao (1993) har studeret forskellige genfindingsundersøgelser og er nået frem til to konklusioner med hensyn til overlap. "First, whether the same topic was searched by two or more searchers, by the use of two or more information representations, by the use of two or more databases, or by the use of two search methods, the overlap between two or more parallel sets for the same query is small. Second, the overlap items were more likely to be judged as relevant than those non-overlap items." (Pao, 1993, s. 338).

Paos undersøgelse bekræfter, hvad Ingwersen skriver om, at udnyttes de kognitive strukturer, kan der opnås overlap, og at disse overlap giver bedre søgeresultater, end hvis de enkelte strukturer blev udnyttet individuelt. Dog skriver Pao også, at ved at bruge flere repræsentationer, databaser eller IR-teknikker, fås et overlap, men et lille overlap. Dette overlap bedømmes ofte som værende relevant. Det vil sige, at man ved at udnytte polyrepræsentationer med forskellige kognitive oprindelser kan få et lille, relevant overlap.

Som eksempel kan boolsk søgning og TARGET søgning anvendes. Boolsk søgning er den traditionelle, velafprøvede genfindingsteknik, som er logisk funderet, hvorimod TARGET er en ny genfindingsteknik, der baserer sig på statistiske beregninger. Dokumenter, der findes ved hjælp af flere forskellige genfindingsteknikker for samme query, vil sandsynligvis være mere relevante end

En quorumsøgning bestående af disse fire termer vil se ud som følger:

retrieval and text and full and information	=	s1
(retrieval and text and full) not s1	=	s2
(retrieval and text and information) not (s1 or s2)	=	s3
(retrieval and full and information) not (s1 or s2 or s3)	=	s4
(text and full and information) not (s1 or s2 or s3 or s4)	=	s5
(retrieval and text) not (s1 or s2 or s3 or s4 or s5)	=	s6
(retrieval and full) not (s1 or s2 or s3 or s4 or s5 or s6)	=	s7
(retrieval and information) not (s1 or s2 or s3 or s4 or s5 or s6 or s7)	=	s8
(text and full) not (s1 or s2 or s3 or s4 or s5 or s6 or s7 or s8)	=	s9
(text and information) not (s1 or s2 or s3 or s4 or s5 or s6 or s7 or s8 or s9)	=	s10
(full and information) not (s1 or s2 or s3 or s4 or s5 or s6 or s7 or s8 or s9 or s10)		s11

Herefter bør de enkelte termer søges alene, hvor der dog stadig trækkes allerede fundne hits fra.

Nu er alle kombinationmuligheder udnyttet og alle poster er unikke, fordi der hvergang er kombineret med *not*. S1 indeholder det mest relevante, fordi alle fire termer her er repræsenteret. S1 er mere relevant end s2, s2 mere relevant end s3, og så videre. Jo flere søgesæt der findes, jo mindre relevans har de i forhold til den oprindelige søgning. Hvis de fire termer blev kombineret med *or*, ville det give det samme som at lægge alle de fundne sæt sammen.

Hvis en søgning, der udnytter de forskellige kognitive strukturer i en database, udføres som en quorumsøgning, hvor der i alle kombinationer er forskellige kognitive strukturer repræsenteret, fås et forholdsvist lille søgesæt, når de forskellige søgesæt til sidst samles. Dette forholdsvis lille søgesæt med høj precision, kan frekvensanalyseres. Hvis der i denne analyse findes flere søgetermer, kan søgningen fortsættes. Fordelen ved denne metode er, at sammenhængen mellem query og dokumentrepræsentationer bibeholdes.

Ingwersen og Wormell (1988) fremhæver frekvensanalyse som et anvendeligt søgeredskab: "... the frequence analysis feeds back a conceptual cognitive structure representing actual relations between the entered search statement and the terms on the ranked list. The relationship exists due to the fact that the terms analysed originate from references also containing the search terms and consequently related to these." (Ingwersen og Wormell, 1988, s. 113). Fordelen ved at anvende frekvensanalyse er, at den kognitive sammenhæng mellem query og dokumentrepræsentationer bibeholdes, da det er de fundne dokumentrepræsentationers termer, der bliver analyseret. På baggrund af denne liste er det forholdsvis nemt for både novicer inden for IR og personer med lille emnemæssig domæneviden at udpege nye termer, der kan udvide den hidtidige søgning og dermed forøge recall ved hjælp af *or* kombinationer (Ingwersen og Wormell, 1988). Frekvensanalyse er det tætteste, vi kommer på relevansfeedback hos Dialog.

4. Target og kognitiv quorumsøgning i praksis

I figur 5 ses et eksempel på en søgning i INSPEC hos Dialog. I dette eksempel anvendes følgende query: *full text information retrieval*. Først foretages en søgning ved hjælp af TARGET-kommandoen. For at tilkendegive at det er fraser, der søges på, anvendes følgende søgestreng: *'full text' 'information retrieval'*, hvor '' indikerer, at det er en frase. Som alle TARGET søgninger giver denne søgning et resultat på 50 poster (figur 5, s1). Hvis følgende søgestreng var anvendt: *'information retrieval' 'full text'* var resultatet blevet de samme 50 poster, men i en anden rangordning. Det viser, at TARGET tager hensyn til den rækkefølge, man

dokumenter der kun findes ved brug af én genfindingsteknik.

Dette at udnytte flere repræsentationer eller teknikker, polyrepræsentationer, i IR-processen svarer til at udnytte indeksørinkonsistens, "... any two indexers indexing one and the same document individually, will select indexing terms which are most unlikely to be identical." (Ingwersen og Wormell, 1988, s. 107). Jo flere indeksører der indekserer et dokument, jo større chance er der for at genfinde det. Samtidig må det formodes, at de termer der bruges af flere indeksører, overlappet, er meget relevante, da der er flere individer med forskellige kognitive strukturer, der anvender disse termer til indeksering af det samme dokument.

Denne udnyttelse af viden om forskellige kognitive oprindelser kan udnyttes i IR-sammenhæng. Hvis en søgestreng er at finde i flere felter med forskellige kognitive oprindelser, er der stor sandsynlighed for, at dette dokument er relevant. Det er blandt andet dette, som TARGET-kommandoen udnytter. Ved at søge i basisregistret og inddrage termers tæthed til hinanden forsøger TARGET at få flere repræsentationer af dokumenterne inddraget. Hvis to søgetermer findes i titelfeltet og i descriptorfeltet, vil både ophavets og indeksørens kognitive strukturer blive udnyttet. En optimal udnyttelse sker, når begge termer står i samme felt, da TARGET også inddrager termers tæthed i rangordningen. Udover dette inddrager TARGET også termers forekomst i dokumentet og i hele samlingen.

Hvis en søgestreng søges i forskellige felter, der derefter kombineres, for eksempel som en quorumsøgning, kan intermediæren have kontrol med hvilke kognitive påvirkninger, der udnyttes i søgningen. På denne måde kan de forskellige kognitive strukturer i databasen udnyttes som polyrepræsentationer af det samme dokument til at forbedre udnyttelsen af dokumetrepræsentationerne og forbedre søgeresultatet.

I en quorumsøgning kombineres de mindste søgesæt med hinanden og en allerede udført søgning fjernes fra en ny søgning ved hjælp af *not*-operatoren. Denne søgemåde giver en form for rangordning. Dog ikke af enkelte poster, men af søgesæt. For eksempel en query der består af:

	items	postings
full	876	4639
text	508	1989
information	1050	9748
retrieval	497	2508

Først kombiners alle fire termer. Dette giver det snævreste, mest relevante, sæt, hvor alle fire termer er repræsenteret. Herefter kombineres de fire termer på forskellige måder, begyndende med de sjældnest forekommende termer først. I dette tilfælde forekommer *text* og *retrieval* næsten lige mange gange, ca. 500. Hvordan kan man adskille to termer, der forekommer i næsten lige mange dokumenter? Hvis postingsantallet anvendes, fås et tal for, hvor stor forekomst termen har i hele basen, og da må det være bedst at benytte den term, der forkommer hyppigst i databasen, da den er tungest i forhold til dokumentantallet. For eksempel står *text* ca. 4 gange i hvert dokument (2000/500) og *retrieval* ca. 5 gange i hvert dokument (2500/500), hvorfor det er bedre at benytte *retrieval* end *text*.

anvender søgetermerne i. En normal boolsk søgestreng i Dialogs basisregister kan se sådan ud: *ss full(w)text and information(w)retrieval*, hvor der søges på *full* og *text* ved siden af hinanden i nævnte rækkefølge og ligeledes *information* og *retrieval* ved siden af hinanden i den nævnte rækkefølge. Hvis søgningen skal foretages i en fuldtekst database, kan *full(w)text* og *information(w)retreival* kombineres med *(s)* for at finde begge søgetermer i det samme fuldtekstafsnit. En anden måde at kombinere de to søgetermer på er ved at benytte nærhedsopratoren *(n)*, der angiver, at termerne skal stå ved siden af hinanden i vilkårlig rækkefølge. Denne operator kan også anvendes med angivelse af afstand. Her kan man f.eks. anvende *(7n)*, svarende til en sætning, der angiver, at der må være 7 ord mellem søgetermerne.

Figur 5: Udsnit af søgehistorie fra Inspec hos Dialog

Set	Items	Description
S1	50	TARGET - 'FULL TEXT' 'INFORMATION RETRIEVAL'
S10	101	FULL(W)TEXT/1994:1995 AND INFORMATION(W)RETRIEVAL/1994:1995
S15	206	FULL(W)TEXT/1994:1995
S16	32	S15/TI
S17	190	S15/DE,ID
S22	1402	INFORMATION(W)RETRIEVAL/1994:1995
S23	139	S22/TI
S24	1392	S22/DE,ID
S27	9	S15/AB(7N)S22/AB
S28	2	S16 AND S17 AND S23 AND S24
S29	0	(S16 AND S17 AND S23) NOT S28
S30	0	(S16 AND S23 AND S24) NOT S28
S31	16	(S16 AND S17 AND S24) NOT S28
S32	9	(S17 AND S23 AND S24) NOT (S28 OR S31)
S33	0	(S16 AND S24) NOT (S28 OR S31 OR S32)
S34	0	(S17 AND S23) NOT (S28 OR S31 OR S32)
S35	2	S28 AND S27
S36	0	S31 AND S27
S37	1	S32 AND S27
S38	27	S28 OR S31 OR S32
S40	21	S1 AND S38
S41	21	TARGET - 'FULL TEXT' 'INFORMATION RETRIEVAL' *S40
S42	2	TARGET - 'FULL TEXT' 'INFORMATION RETRIEVAL' *S28
S43	16	TARGET - 'FULL TEXT' 'INFORMATION RETRIEVAL' *S31
S44	9	TARGET - 'FULL TEXT' 'INFORMATION RETRIEVAL' *S32

Ved en kognitiv quorumsøgning, hvor viden om databasens iboende, forskellige kognitive strukturer udnyttes, afgrænses de enkelte søgeargumenter til bestemte felter, med egen kognitiv oprindelse. For eksempel titel- og abstract-felterne, der repræsenterer ophavets kognitive strukturer eller descriptor- og identifier-felterne, der repræsenterer indeksørens kognitive strukturer. Herefter foretages en quorumsøgning med disse afgrænsede søgeargumenter (figur 5, s10 til s34).

En kognitiv quorumsøgning udføres enklest ved at undlade at anvende afgrænsningen til abstract-feltet. Ophavets kognitive strukturer er allerede repræsenteret ved titel-feltet. Abstract-feltet kan derimod anvendes til en yderligere underdeling af de kognitive søgesæt, der findes ved at anvende det på følgende måde: *s32 og full(w)text/ab(7n)information(w)retrieval/ab*. Hvis man ved en kombination af titel- og descriptor/identifierfelter har fået et forholdsvis stort søgesæt, der ønskes yderligere underdelt, kombineres det med søgetermerne afgrænset til abstractfeltet og med en nærhed på 7 ord.

I det viste eksempel er de- og id-felterne slået sammen, fordi begge felter repræsenterer indeksørens kognitive strukturer. Selvom identifier-feltets sproglige struktur ligner abstractets, foretager indeksøren en tilpasning af abstractord til identifiers, hvorfor det er indeksøren, der er 'ophav' til identifierfeltet. Derfor bliver disse to felter slået sammen i quorumsøgningen.

Ved at starte med en almindelig boolsk *and*-søgning i basisregistret fås et mål for hvor mange poster der *kan* findes ialt. I dette tilfælde er der ialt 101 poster, der indeholder søgetermerne. Hele quorumsøgningen vil ialt give 101 poster (s10), hvorimod den kognitive del kun giver 27 poster (s38).

TARGET søgningen og quorumsøgningen kan nu rangordnes på forskellige måder afhængigt af den allerede givne rangordning.

Den kognitive quorumsøgning falder i tre intervaller, de tre oprindelige søgesæt, s28, s31 og s32. Disse intervaller er hver især ordnet kronologisk, men kan hver især rangordnes ved hjælp af den rangordning, der er i TARGET søgningen (s1). Det første interval, s28, indeholder 2 poster, og disse 2 poster rangordnes nu i forhold til den rang, de har i TARGET søgningen. Herefter rangordnes s31 og s32 på tilsvarende måde. De poster, der er med i overlappet mellem de to søgninger, det vil sige findes i TARGET søgningen, rangordnes før de, der kun er i quroumsøgningen.

TARGET søgningen falder også i intervaller afhængigt af den relevansprocent, den enkelte post tildeles. Poster med ens relevansprocenter kan rangordnes ved hjælp af den kognitive quorumsøgning, hvor poster, der forekommer i den kognitive quorumsøgning, anses for værende mere relevante end poster med samme relevansprocent, der ikke findes i den kognitive quorumsøgning.

Denne metode til rangordning af søgesæt er forholdsvis tidskrævende. En anden metode til at underdele søgesæt er at kombinerer med abstract-felterne som nævnt ovenfor. Se s35, s36 og s37. I s37 forekommer 1 post, der indeholder søgetermerne i ab-feltet, hvorfor denne post bør rangordnes før de øvrige 8 poster fra s32. Endelig kan nedenstående metoder anvendes.

For det første kan de enkelte intervaller i quorumsøgningen rangordnes ved, at de hver især udføres som søgninger i TARGET: *'full text' 'information retrieval' *s31* (figur 5, s43), hvor s31 er det søgesæt, der indeholder et interval fra den kognitive quroumsøgning, og * indikerer, at s31 *skal* forekomme i det sæt, der rangordnes. For at få denne søgning rangordnet skal der tilføjes nogle søgetermer, og her er det mest naturligt at gentage de oprindelige søgetermer.

For det andet kan overlappet udnyttes. Som tidligere nævnt øges chancen for relevante dokumenter, hvis der anvendes forskellige kognitive strukturer og det deraf følgende overlap. I dette eksempel er ophavets og indeksørens kognitive strukturer bevidst udnyttet i quroumsøgningen. Ved at anvende to forskellige søgestrategier, der kan kombineres, kan der findes et overlap, som med stor sandsynlighed er relevant i forhold til query.

Overlappet mellem den kognitive quorumsøgning og TARGET søgningen er på 21 poster (figur 5, s40). Disse 21 poster bedømmes emnemæssigt relevante udfra en gennemgang af ti, de, id og ab-

felterne. Disse kan rangordnes ved at foretage en ny TARGET søgning: *'full text' 'information retrieval' *s40* (figur 5, s41). Denne metode resulterer i et rangordnet søgesæt med samme størrelse som s40, med formodet høj precision. Hvis disse 21 poster ikke er tilstrækkelige, kan frekvensanalyse af det fundne søgesæt anvendes til at finde nye søgetermer, der kan udvide søgningen og dermed øge recall. Hos Dialog kan man ved hjælp af kommandoen RANK få udført en frekvensanalyse på fraseindekserede felter, i dette tilfælde anvendes RANK-kommandoen til frekvensanalyse af desciptor- og identifier-felterne. Frekvensanalysen af descriptor- og identifierfelter i overlappet (figur 6) kan give et billede af hvilke fraser, der er anvendt i disse dokumenter, og hvilke fraser der dermed kan have relevans for query. På baggrund af en sådan liste vil det være forholdsvis nemt for en bruger at udpege en eller flere fraser, der kan bruges til udvidelse af søgningen, også selvom brugeren ikke har nogen stor domæneviden. Det er altid lettere at vælge ud fra en given liste end det er selv at finde relevante ord. De valgte fraser vil kunne gøre søgningen bredere uden at miste den emnemæssige relevans, der må formodes opnået ved dette overlap.

Figur 6: Frekvensanalyse af overlappet mellem den kognitve quorumsøgning og target-søgningen

1	19	FULL-TEXT DATABASES
2	15	INFORMATION RETRIEVAL
3	5	INFORMATION RETRIEVAL SYSTEM
4	3	BIBLIOGRAPHIC SYSTEMS
5	3	CD-ROMS
6	3	HYPERMEDIA
7	3	INDEXING
8	3	INFORMATION RETRIEVAL SYSTEMS
9	3	QUERY PROCESSING
10	2	ARITHMETIC COEFFICIENTS
11	2	DIALOG
12	2	FULL TEXT
13	2	FULL TEXT RETRIEVAL
14	2	FULL-TEXT DOCUMENTS
15	2	FULL-TEXT INFORMATION RETRIEVAL
16	2	HUMAN FACTORS
17	2	HYPERTEXT LINKS
18	2	INFORMATION NEEDS
19	2	INFORMATION RETRIEVAL SYSTEM EVALUATION

4.1 Target som ir-teknik

I det foregående er der vist en metode til at kombinere TARGET søgninger med enten traditionelle boolske søgninger eller med kognitive quorumsøgninger og på denne måde få forbedret søgeresultatet i forhold til det resultat, der ville være kommet ud af en boolsk søgning alene eller en TARGET søgning alene. Ved at anvende TARGET-kommadoen får man følgende muligheder:

Input search terms separated by spaces (e.g., DOG CAT FOOD). You can enhance your TARGET search with the following options:
- PHRASES are enclosed in single quotes
  (e.g., 'DOG FOOD')
- SYNONYMS are enclosed in parentheses
  (e.g., (DOG CANINE))
- SPELLING variations are indicated with a ?
  (e.g., DOG? to search DOG, DOGS)
- Terms that MUST be present are flagged with an asterisk
  (e.g., DOG *FOOD)

Q = QUIT H = HELP

Der er mulighed for at søge på enkelttermer, fraser, synonymer, trunkere eller at angive at termer skal være til stede i søgeresultatet. Den sidste facilitet svarer til en traditionel *and* kombination, der giver mulighed for at foretage en 'almindelig' *and*-søgning og få den rangordnet.

Der er lavet flere undersøgelser af boolsk søgning set i forhold til søgninger udført med TARGET-kommandoen.

Snow (1994) har sammenlignet boolsk søgning og TARGET søgning i Medline og Embase. Hun finder blandt andet følgende ulemper ved TARGET:

- ved onesearch er det ikke muligt at fjerne dubletter. Dette kan kun gøres med en normal rd-kommando, og så mistes rangordningen af de fundne dokumenter,
- der findes kun 50 poster, og er det en onesearch, er der stor sandsynlighed for, at der optræder dubletter blandt de 50 poster.

Dette kan afhjælpes ved at starte med at foretage en traditionel boolsk søgning, hvor dubletter fjernes, hvorefter der skiftes til TARGET, hvor der søges på det dubletfri søgesæt, som mærkes med * og suppleres med et par relevante søgeord. Dette burde give en søgning på 50 individuelle dokumenter rangordnet efter statistisk relevans. Snow (1994) fremhæver også frekvensanalyse af TARGET-søgesæt som en meget anvendelig facilitet.

Der er også mulighed for at rangordne præfix-søgninger som for eksempel au=ingwersen, p. Dette kan gøres ved at udføre søgningen som en traditionel boolsk søgning og derefter genbruge den i en TARGET søgning mærket med *. Denne søgning vil rangordne dokumenter med Ingwersen som eneste forfatter før dokumenter med fælles forfatterskab.

Tenopir og Cahn (1994, s. 46-47) fremhæver TARGET til følgende: indledende søgninger om et emne, søgning i fuldtekstdatabaser eller baser med lange abstracts, stor precision, at komme up-to-date med et emne, når boolsk logik giver for mange hits, når boolsk logik giver for få hits og når databasen ikke er kendt for søgeren.

Det må dog anbefales at bruge traditionel boolsk søgning ved verifikative søgninger.

Snow konkluderer at 'real searchers' bruger TARGET, eller i det mindste burde bruge TARGET. TARGET er blandt andet anvendelig, fordi den skaber et forholdsvis lille søgesæt med rangordning, der kan viderebearbejdes.

Snow fremhæver også nogle tips om den bagvedliggende teknik, som hun mener er vigtige at erindre i søgeprocessen:

- poster, hvor alle søgetermer er repræsenteret, rangordnes før poster, hvor kun nogle af søgetermerne er repræsenteret
- hvis termer står i en vis tæthed til hinanden, formodentlig 17 ord, tildeles de en stor vægt
- sjældne termer vægtes mere end hyppige termer.

Disse tips stemmer godt overens med Dialogs offentliggørelse af termfrekvens og tæthed som vigtige elementer i den statistiske beregning af relevans.

5. Konklusion

Med mulighed for at søge ved hjælp af to forskellige IR-tekniker er det interessant, om de to teknikker kan kombineres med et anvendeligt slutresultat. Eksemplet med anvendelse af de to forskellige IR-teknikker, kognitiv quorumsøgning og TARGET, viser, at man med fordel kan anvende TARGET.

Det er stadig uvist hvilken algoritme, der ligger bag TARGET-kommandoen, men der er flere og flere ideer om hvilke aspekter, der inddrages i relevansberegningen. For eksempel termers frekvens i dokumenter og i samlingen, querytermers tæthed i de relevante dokumenter og den rækkefølge querytermerne indtastes i. Selvom det ikke er helt klart, hvad der sker i TARGET kommandoen, kan den med fordel anvendes, hvis man ønsker et forholdsvist lille søgesæt, indeholdende nye dokumenter, der er rangordnet. Der er dog også ulemper forbundet med at anvende TARGET. For eksempel at der ved Onesearch ikke fjernes dubletter.

Hvis søgemetoden, med overlap mellem kognitiv quorumsøgning og TARGET søgning, anvendes, får man i første omgang et forholdsvis lille postantal med stor sandsynlighed for at de fundne poster er emnemæssigt relevante. Er dette søgesæt ikke stort nok, kan man, ved hjælp af frekvensanalyser af emneord i de fundne dokumentrepræsentationer, få søgetermer, der kan udvide det eksisterende søgesæt og dermed forøge recall uden at miste den kognitive sammenhæng mellem query og dokumentrepræsentationer.

Både den kognitive quorumsøgning og TARGET giver rangordnede resultater og med et relativt stort overlap. Derfor kan det konluderes, at både søgninger med anvendelse af quorum og TARGET kan give gode rangordnede resultater med emnemæssig relevans. Dette taler for, at man kan anvende TARGET eller quorumsøgning afhængig af, om man søger hos Dialog eller hos en anden vært, der ikke har en facilitet som TARGET.

Kan TARGET-kommadoen bruges til noget fornuftigt? Dette spørgsmål synes at være blevet besvaret bekræftende i det ovenstående. TARGET-kommandoen har gode anvendelsesmuligheder i de professionelle miljøer, der anvender Dialog, hvis man er opmærksom på de begrænsninger og muligheder, der ligger i TARGET-kommandoen, både anvendt alene og især i sammenhæng med boolske søgninger.

6. Litteratur

Belkin, N. J. & Croft, W. B. (1987). Retrieval techniques. ARIST vol. 22, s. 109-145.

Croft, B. (1993). Vector space model. Upubliceret materiale.

Dialog (1993). TARGET on Dialog: "How-to" guide. Palo Alto, U.S.A.: Dialog.

Ingwersen, P. (1992). Information Retrieval Interaction. London: Taylor Graham.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. Proceedings of the Seventheenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Edited by W. Bruce Croft and C. J. van Rijsbergen. London: Springer.

Ingwersen, P. og Wormell, I. (1988). Means to improved subject access and representation in modern information retrieval. Libri, Vol. 38, s. 94-119.

Pao, M. L. (1993). High precision by duplicate retrival. 14th National Online Meeting. Edited by Martha, E. Williams. Medford, New Jersey: Learned Information. s. 337-341.

Pao, M. L. (1994). Relevance odds of retrieval overlaps from seven search fields. Information Processing & Management, Vol. 30, s. 305-314.

Snow, B. (1994). TARGET for the biomedical searcher. Online. Vol 18, no. 6, s. 58-65.

Sparck Jones, K. (1973). Index term weighting. Information Storage and Retrieval, Vol. 9, s. 619-633.

Tenopir, C, Cahn, P. (1994). TARGET & FREESTYLE: DIALOG and Mead join the relevance ranks. Online. Vol 18, no 3, s. 31-47.