Polyrepræsentation som princip for indeksering og genfinding af videnskabelige fuldtekstdokumenter

Af Birger Larsen

Indledning

I denne artikel beskrives idégrundlaget bag forfatterens Ph.D.-projekt. Formålet med projektet er at udvikle og afprøve nye metoder til automatisk indeksering af fuldtekstdokumenter baseret på princippet om polyrepræsentation (Ingwersen, P., 1996). Metoderne anvender kombinationer af flere forskellige typer af repræsentationer på samme tid ved indeksering og søgning. Disse *poly*-repræsentationer dannes dels ud fra elementerne og strukturen i dokumenterne, dels ud fra andres fortolkninger af dokumenterne, f.eks. emneord tilføjet af menneskelige indeksører eller andre forfatteres citationer til dokumenterne. Der fokuseres i første omgang på videnskabelige dokumenter, da de indenfor visse fag har næsten samme overordnede struktur. Målet med projektet er at sikre bedre adgang til informationen i dokumenterne, og det ligger indenfor det forskningsfelt i Biblioteks- og Informationsvidenskaben der hedder *Information Retrieval* (IR), hvor hovedvægten ligger på design og evaluering af informationssystemer eller komponenter til informationssystemer. Artiklen er struktureret således: Første afsnit redegør for projektets teoretiske baggrund og relateret forskning. Dernæst fremlægges projektets forskningsspørgsmål efterfulgt af en beskrivelse af hvilke materialer og metoder der tænkes anvendt. Artiklen afsluttes med betragtninger over hvilken betydning et sådant projekt kan få.

Baggrund for projektet

Eksisterende modeller for indeksering og søgning af fuldtekstdokumenter

Det bliver i stigende grad almindeligt at dokumenter stilles til rådighed i elektronisk form i fuldtekst. Denne udvikling er bl.a. muliggjort af integrerede redskaber til produktion og distribution, stadigt billigere lagermedier samt mere effektive kommunikationskanaler. Blandt de mange millioner dokumenter der nu er tilgængelige i fuldtekst, findes også flere og flere videnskabelige dokumenter. På trods af at hele teksten i disse dermed kan gøres søgbar, er det dog ikke nødvendigvis blevet lettere for brugerne af informationssystemer at finde dokumenter der er relevante for deres informationsbehov. Det hænger blandt andet sammen med at mængden af dokumenter stiger. Da de første større samlinger af elektronisk fuldtekstdokumenter begyndte at dukke op i begyndelsen af 1980'erne blev de lagt op i den samme type informationssystemer som bliver benyttet til bibliografiske databaser. Disse er baseret på inverterede filsystemer, hvor dokumenterne repræsenteres af surrogater der er opbygget af en række elementer, f.eks. forfatter, titel, abstract, kontrollerede emneord, tidsskriftsnavn etc. som tilsammen repræsenterer (står i stedet for) selve dokumentet i sin helhed. Når fuldtekstdokumenter lægges ind i de inverterede filsystemer placeres hele teksten som regel i et enkelt

felt. Søgning i disse systemer foregår med booleske operatorer, som gør det muligt at kombinere flere søgetermer, eventuelt fra forskellige repræsentationer. Ud over de booleske operatorer findes ofte også nærhedsoperatorer, der retter sig mod søgning i større tekstafsnit. Kommandosproget er ofte ganske komplekst og giver hermed mange muligheder, dog kræver de inverterede filsystemer at søgetermerne skal forekomme i dokumenterne præcist som de er indtastet ('exact match' princip). I baser med fuldtekstdokumenter spiller nærhedsoperatorerne en afgørende rolle for søgningen, da de kan være med til at sikre at søgetermerne i en forespørgsel står indenfor en vis afstand af hinanden i teksten, og at der dermed formodentlig er en vis sammenhæng mellem dem. Antallet af nærhedsoperatorerne er da også typisk blevet udvidet i informationssystemer med fuldtekstdokumenter, så det f.eks. er muligt at søge indenfor den samme sætning eller det samme afsnit.

Man kunne have en forventning om at det ville være nemmere at finde relevante dokumenter, når det er muligt at søge i alle de originale ord i hele dokumentet i stedet for det mere begrænsede antal ord i dokumentsurrogaterne. På trods af de mere avancerede søgekommandoer som nærhedsoperatorer er det dog langtfra tilfældet. Forskningen i fuldtekstdatabaser viser, at søgning i hele teksten kan være en fordel, når det der søges efter kan beskrives med meget specifikke søgetermer, men at det som regel ikke er det ved brede emnesøgninger eller ved begreber, hvor der anvendes mange synonymer for begrebet i dokumenterne (Tenopir, C., 1985). Et andet problem ved søgning i hele teksten er, at søgetermerne nemt kan forekomme i dokumenter der kun perifert berører det søgte emne, eller at søgetermerne bruges i en anden betydning end den tilsigtede. Det har vist sig at ord i titlerne, abstracts, og tildelte emneord har en opsummerende funktion og reducerer den mangfoldighed og kompleksitet der findes i den fulde tekst. Det er derfor en fordel, hvis disse repræsentationer er indekseret i separate felter (og der med er søgbare) og ikke befinder sig i fuldtekstfeltet.

Parallelt med opblomstringen af de bibliografiske databaser er der indenfor forskningen i Information Retrieval (IR) siden 1960'erne forsket intensivt i at udvikle alternative metoder til indeksering af dokumenter. Her er tilgangen en helt anden, idet man, oftest ud fra forskellige matematiske modeller, har anvendt de statiske fordelinger af ordene i dokumenterne til automatisk udvælgelse og vægtning af indekseringstermer. Målet med denne forskning er at udvikle og evaluere algoritmer, som kan fremfinde og rangordne relevante dokumenter på baggrund af en antagelse om, at de ord der forekommer mange gange i et dokument udsiger noget om dokumentets emne. Forskningen er centreret om at udvikle de mest effektive algoritmer til automatisk indeksering af dokumenterne og design informationssystemer til genfinding af dem. Sammenlignet med de inverterede filsystemer har disse statistisk baserede systemer en række fordele. De giver bl.a. mulighed for at tillade at ikke alle søgetermer skal være tilstede i de fremfundne dokumenter og for at rangordne dokumenterne i forhold til sandsynligheden for at de er relevante i forhold til en forespørgsel ('best match' princip). Modsat de booleske systemer er det ikke afgørende i best match systemer, at brugerne behersker et kompliceret kommandosprog, da der kan søges i naturligt sprog. Resultaterne af forskningen ses i de store søgemaskiner på WWW, hvor stort set alle pionererne begyndte som akademiske projekter. Til trods for at IR-forskningen dermed har fundet anvendelse i håndteringen af de enorme tekstmængder på WWW er systemerne langt fra perfekte. Der forskes derfor med uformindsket kraft inden for denne tilgang, bl.a. indenfor rammerne af TREC (Text REtrieval Conferences) hvor forskellige systemer afprøves på meget store dokumentsamlinger (1). En lang række teorier og metoder er gennem årene blevet afprøvet, f.eks. brug af vektorer (Salton, G., 1988), probabilistisk indeksering (Robertson, S.E. and Sparck Jones, K., 1976), lingvistiske metoder (Smeaton, A.F., 1990). I forhold til de første systemer fra 1960'erne er der sket store forbedringer i systemernes effektivitet, og det er i dag muligt med en vis effektivitet at afsøge selv meget store samlinger af fuldtekstdokumenter. De senere år er der dog kun opnået marginale forbedringer; allerede i slutningen af 80'erne påpegede Croft, at systemer der baserer sig på ovennævnte tilgange formodentlig har nået grænsen for deres ydeevne, og at der er behov for helt nye måder at gå til værks på

(Croft, W. B. and Thompson, R. H., 1987). En del af forklaringen på denne stagnation kan måske findes i, at den overvejende del af de modeller der anvendes til automatisk indeksering ikke tager hensyn til sammenhængen mellem tekstens ord eller deres placering i dokumenterne, men som regel kun fokuserer på ordene hver for sig. Fra midt i 90'erne er der forsket i avancerede lingvistiske metoder (Natural Language Processing), hvor syntaksen mellem ordene og til en vis grad ordenes mening blev analyseret på sætnings- og afsnitsniveau og anvendt i IR-systemer (Smeaton, A. F., 1990). Desværre har dette vist sig at være mere kompliceret end ventet i praksis og denne forskning på et meget detaljeret mikroniveau har endnu ikke givet bedre resultater end de mere simple statistiske tilgange. Da de forskellige dele af dokumenterne har forskellige funktioner kan der dog være fornuft i at anvende strukturer i dokumenterne i forbindelse med indeksering og søgning. Formålet med dette projekt er bl.a. at undersøge mulighederne for at udnytte at visse dokumenttyper, f.eks. videnskabelige artikler, er opbygget i en række faste afsnit der går igen i mange af dokumenterne. Dette forhold bør kunne udnyttes aktivt ved indeksering og søgning, både i inverterede filsystemer og i best match systemer.

Teoretisk baggrund og relateret forskning

Projektet ligger inden for den tilgang til informationsvidenskab, der udgøres af det kognitive synspunkt på IR, som det fremstilles af f. eks. B. C. Brookes, Nicholas J. Belkin og Peter Ingwersen. Projektet tager udgangspunkt i at udføre empiriske undersøgelser af udvalgte dele af Ingwersens kognitive teori for IR (Ingwersen, P., 1996). Den traditionelle IR-forskning fokuserer først og fremmest på forbedring af informationssystemernes algoritmer, for at gøre informationssystemerne bedst mulige til at finde og rangordne de dokumenter, der er relevante i forhold til en forespørgsel. Set fra et kognitivt synspunkt er dette ikke tilstrækkeligt, da der ligger en kompliceret virkelighed med mange aktører bag både dokumenterne og forespørgslerne. Den kognitive teori for IR opstiller en bredere ramme, hvor forfatterne til dokumenterne, skaberne og brugerne af kontrollerede emneordslister, designere af informationssystemerne, samt brugerne af systemerne og den kontekst de indgår i inddrages for at forstå, hvad der foregår når en bruger anvender et informationssystem. Alle processer i informationssøgning anskues som resultater af *kognitive processer* mellem aktørerne. Hver aktør har sin egen model af verden, der påvirker forløbet af disse processer. For de menneskelige aktører kan modellerne ses som bestående af komplekse og dynamiske kognitive strukturer, der er opbygget gennem de oplevelser hver aktør har haft både individuelt og socialt. Da det netop er kognitive strukturer og processer, der foregår i hjernen på aktørerne er der ikke direkte adgang til dem. I alle forhold arbejdes der derfor i stedet med *repræsentationer* af disse: F.eks. anskues en nedskrevet tekst som en forfatters forsøg på at formidle en del af sin viden om et bestemt emne, men det er netop en repræsentation hvor de meget komplekse og dynamiske kognitive strukturer er reduceret og fikseret i en lineær tekst. Et andet eksempel er en brugers forespørgsel til en bibliotekar eller et informationssøgningssystem, der oftest er en kraftigt reduceret repræsentation af det problem der forsøges løst. Disse repræsentationer er af natur ikke perfekte, og der er stor *usikkerhed* og *uforudsigelighed* forbundet med dem. Det kan f.eks. bestå i at forfatterne af dokumenter og brugere af informationssøgningssystemer har forskellig sprogbrug, eller at forskellige indeksører aldrig vil kunne tildele præcist de samme kontrollerede emneord til de samme dokumenter, på trods af erfaring, træning og veludviklede hjælperedskaber (fænomenet er kendt som inter-indeksørinkonsistens (Mann, T., 1997)).

Usikkerhed og uforudsigelighed er en uundgåelig del af en kompleks virkelighed og er ikke i sig selv uhensigtsmæssig, men beriger tilværelsen med nuancer og facetter. For brugerne af de IR-systemer har det dog en række negative konsekvenser, f.eks. at det kan være svært at finde dokumenter der handler præcist om det emne der søges efter, eller at det kan være svært at finde alle dokumenter i en samling der omhandler et bestemt emne. Bibliotekerne har traditionelt forsøgt at løse disse problemer gennem forskellige former for standardiseringer og kontrol. De udførlige katalogiseringsregler, de store klassifikationssystemer og de kontrollerede emneordslister konstrueret efter bestemt regler og procedurer kan ses som en måde at håndtere en kompleks virkelighed på, og de

er forsøg på at sikre adgangen til information gennem en intellektuel indsats. Også IR-forskningen forsøger at reducere kompleksiteten og sikre adgangen til information, men her gøres det gennem udvikling af algoritmer som automatisk kan indeksere dokumenterne ved hjælp af computere. Set fra et kognitivt synspunkt er kompleksiteten helt naturlig og uundgåelig på grund af de mange aktører, der med hver deres baggrund påvirker processerne. Frem for at forsøge at reducere kompleksiteten mest muligt foreslår det kognitive synspunkt at usikkerheden og uforudsigeligheden i stedet udnyttes aktivt på en struktureret måde. Strukturering opnås ved at både dokumenterne og brugernes informationsbehov repræsenteres på en række forskellige måder (princippet om polyrepræsentation), og ved at de forskellige repræsentationer herefter udnyttes systematisk ved søgning:

> "The concept of polyrepresentation seeks to represent the current user's information need, problem and knowledge states and domain work task or interest in the form of contextual structures of causality. At the same time it implies that we should apply different cognitive and functional origin to the information ob-jects in the information space." (Ingwersen, P., 1996)

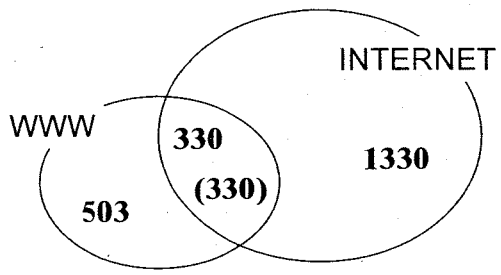
Ved at have en række adskilte repræsentationer af både brugerens behov og af dokumenterne, kan repræsentationerne kombineres systematisk i informationssystemets søgefunktion. Mellem disse sæt vil der være varierende grader af overlap, da hver repræsentation har rod i forskellige kognitive strukturer, og da der indenfor hver struktur er usikkerhed og uforudsigeligheden. Hypotesen i princippet om polyrepræsentation er, at disse overlap kan udnyttes til at opnå en bedre genfinding af relevante dokumenter:

> "...if different cognitive structures, in defiance of the inconsistency, do, in fact, retrieve overlapping information objects, this cognitive overlap presents more 'relevant/useful/...' information objects than each independent structure....[and]....the more different the cognitive structures producing an overlap are in time and by cognitive or functional type, the higher the probability of its 'relevance/usefulness/...'". (Ingwersen, P., 1996)

Siden teoriens fremsættelse er der endnu kun udført få empiriske studier af brugbarheden af den kognitive teori for IR og princippet om polyrepræsentation. Borlund har udviklet metoder til evaluering af interaktive systemer inspireret af teorien (Borlund, P., 2000a; Borlund, P., 2000b), og Ahlgren har udført et mindre studie af booleske søgestrategier afledt af princippet om polyrepræsentation (Ahlgren, P., 1999; Ahlgren, P., 1998). Som det er i dag bliver overlap mellem forskellige repræsentationer ikke udnyttet eksplicit i de eksisterende informationssystemer. Det kan anskueliggøres ved nedenstående søgning foretaget i LISA (online hos Dialog), hvor der søges efter nyere dokumenter omhandlende Internettet. I søgningen betragtes de to søgetermer INTERNET og WWW som nærsynonymer og kombineres derfor med Boolesk ELLER:

Set	Items	Description
S1	1660	INTERNET/2000
S2	833	WWW/2000
S3	2163	S1 OR S2

Efter at de i alt 2493 poster i sæt 1 og 2 kombineres med en Boolesk ELLER-operator returnerer informationssystemet kun 2163 poster i sæt 3. De 330 poster, som indeholder begge søgetermer, identificeres af informationssystemet to gange (en gang med WWW, og en gang med INTERNET), men når resultatet præsenteres frasorteres dubletterne automatisk, så hver post kun forekommer én gang. Situationen illustreres i Figur 1. For brugerne af de inverterede filsystemer er dette en særdeles hensigtsmæssig funktion, der betyder at man ikke skal se på de 330 poster to gange. Set fra et kognitivt synspunkt er de 330 poster i overlappet derimod særligt interessante, da der er to forskellige søgeord der peger på dem. I princippet om polyrepræsentation anbefales det, at der gøres brug af *intentional redundancy*, altså at udnytte at der kan refereres til det samme på flere forskellige måder, og derigennem reducere den usikkerhed og uforudsigelighed der er forbundet med søgetermerne hver for sig. I dette meget simple eksempel vil det derfor kunne anbefales at undersøge om posterne i overlappet er relevante før resten af posterne undersøges.



Figur 1: Overlap mellem repræsentationer (Kilde: LISA, online hos Dialog, 2001)

Som det ses af ovenstående citat er hypotesen i princippet om polyrepræsentation, at desto længere de repræsentationer der danner et overlap er fra hinanden i type og tid, desto større er sandsynligheden for, at dokumenterne i overlappet mellem dem er relevante. Det vil sige, at der sandsynligvis er mest at vinde ved at benytte forskellige repræsentationer, hvilket da også har været anvendt i praksis af bibliotekarer og informationsspecialister i lang tid. Når den Boolske OG-operator anvendes mellem forskellige repræsentationer finder den netop overlappet mellem dem. En række af empiriske undersøgelser anbefaler da også, at man anvender søgestrategier med en kombination af kontrollerede emneord og søgning i hele teksten/abstracts til søgninger, hvor der ønskes få, højrelevante dokumenter (Katzer, J. et al., 1982; Tenopir, C., 1985; Lancaster, F. W., 1991).

Der er kun udført få direkte studier af graden af overlap mellem repræsentationer og af overlappenes karakteristika. Et par af disse studier har fokuseret på citationssøgestrategier som alternativ til mere gængse emnesøgningsstrategier og analyseret overlappet mellem dokumenter fundet via citationssøgestrategier og via søgning i emneord, titler og abstracts. McCain har analyseret graden af overlap i et mindre eksperiment med 9 forespørgsler (McCain, K. W., 1989). Hun konstaterede, at overlappene mellem de to typer af repræsentationer var små: 10% i gennemsnit med variationer fra 1% til 26%. I et væsentligt større eksperiment med 89 forespørgsler undersøgte Pao graden af overlap, samt andelen af relevante dokumenter i overlappet og udenfor. Hun fandt at der var 28% af søgningerne, der ikke resulterede i noget overlap og at overlappet var på 5% i gennemsnit i de resterende søgninger. Analysen af selve overlappet afslørede, at andelen af relevante dokumenter var på hele 86%, hvilket er 6 til 8 gange mere end de dokumenter der kun blev fundet ved én af søgestrategierne. Dette resultat er meget markant, men er ikke i sig selv overraskende ud fra princippet om polyrepræsentation: De to typer af repræsentationer er meget forskellige, og et lille overlap med en høj andel af relevante dokumenter kan derfor forventes. Citationssøgestrategien i begge eksperimenter tog udgangspunkt i et eller flere kildedokumenter, der var udpeget af brugerne som relevante for forespørgslen (såkaldte 'seed documents'). Strategien var herefter at identificere de dokumenter der efterfølgende har citeret kildedokumenterne. En konklusion der kan drages af begge eksperimenter er, at en søgestrategi som identificerer dokumenter i overlappet mellem disse to typer af repræsentationer altså kan anvendes til at finde få, højrelevante dokumenter. Repræsentationer genereret udfra citationsdata og deres karakteristika i forhold til andre repræsentationer undersøges derfor eksplicit i dette projekt.

Da det er vigtigt for projektet, at der er mange forskellige repræsentationer af dokumenterne til rådighed i eksperimenterne, er det målet at anvende fuldtekstdokumenter frem for bibliografiske poster. I IR-forskningen anvendes normalt en række standard testsamlinger i eksperimenterne (test collections). Testsamlingerne bestod i begyndelsen af bibliografiske poster, men med TREC blev eksperimenterne skaleret op til store testsamlinger med fuldtekstdokumenter. Dokumenterne er typisk avisartikler da de har været relativt lette at få adgang til, men de er ikke specielt velegnede til dette projekt, da de ikke har nogen fast struktur som kan udnyttes i indekseringen. Swales har studeret den videnskabelige artikel som genre, både den historiske udvikling og nutidige artikler (Swales, J. M., 1990). Studierne viser at de videnskabelige artikler indenfor naturvidenskaberne typisk har formen: "Introduction–Method–Results–Discussion", og at der er markante forskelle på deres funktion, retoriske skrivestil og indhold i artiklerne. Hans analyse viser endvidere at de retoriske karakteristika er afhængige af "modenheden" af det fag artiklerne er skrevet indenfor: Hvis faget har vedtagne og etablerede konventioner vil der f.eks. være større forskel på

sektionerne. Videnskabelige dokumenter i fuldtekst fra et naturvidenskabeligt fag foretrækkes til eksperimenterne, da der er stor sandsynlighed for at der kan genereres en række forskellige repræsentationer udfra disse.

Forskningsspørgsmål

Det overordnede formål med projektet er at undersøge brugbarheden af princippet om polyrepræsentation i IR. Den centrale hypotese er, at man ved at arbejde med mange forskellige repræsentationer af de samme dokumenter kan konstruere IR-systemer, der er bedre til at genfinde relevante dokumenter. En række forskningsspørgsmål kan udledes udfra hypotesen:

- Kan der gennem automatisk indeksering skabes en række forskellige repræsentationer af videnskabelige fuldtekstdokumenter, der er anvendelige i eksperimenter med polyrepræsentation?
- Hvad er graden af overlap mellem de forskellige repræsentationer?
- Er der væsentligt flere relevante dokumenter i overlappene mellem polyrepræsentationer af videnskabelige fuldtekstdokumenter og positioneres relevante dokumenter højere i ranket output, end i repræsentationerne set enkeltvis?
- Hvilke typer af repræsentationer skaber overlap med størst andele af relevante dokumenter?
- Er andelen af relevante dokumenter større i overlap, hvor mange forskellige repræsentationer peger på dokumenterne?
- Kan flere forskellige repræsentationer og overlap imellem dem udnyttes systematisk i IR-systemer, og resulterer dette i bedre resultater end i traditionelle systemer?

Materiale og metoder

Da projektet ligger indenfor IR-forskningen og kan få konsekvenser herfor opstilles der i overensstemmelse med denne forskningstradition eksperimenter, hvor hypotesen og forskningsspørgsmålene undersøges empirisk. Til brug for eksperimenterne etableres et eksperimentelt informationssøgningssystem bestående af en database med et stort antal fuldtekstdokumenter, implementeret med modeller for automatisk indeksering, f. eks. baseret på vektorspace og/eller probabilistisk vægtning.

Dokumenter
Som allerede indikeret ovenfor foretrækkes videnskabelige dokumenter til eksperimenterne. Figur 2 viser hvilke elementer i artiklerne der sandsynligvis kan dannes repræsentationer ud fra. Listen er opdelt i *funktionsafsnit*, der er egentlige sektioner i artiklerne med hver deres funktioner som diskuteret ovenfor, og *strukturelementer*, der består af de elementer i artiklerne der træder ud fra den rå tekst og giver den struktur visuelt og indholdsmæssigt. Netop strukturelementerne er interessante, da de overskrifter og tabel- og figurtekster kan indeholde ord der er meget prægnante for artiklens emne. Overskrifterne har den dobbelte funktion, at de på det øverste niveau markerer begyndelsen på funktionsafsnittene (med ord der som regel ikke er prægnante for emnet), mens de på underliggende niveauer ofte opsummerer emnet i det pågældende afsnit.

Funktionsafsnit:
Titler
Abstract
Evt. emneord
Introduktionsafsnit
Metodeafsnit
Analyse/resultatafsnit
Konklusionsafsnit
Referencerne
Bilag
(Hele dokumentet i fuldtekst)

Strukturelementer:
Tabel- og figurtekster
Overskrifter (på forskellige niveauer)

Figur 2 : Elementer i artiklerne hvorfra der kan dannes repræsentationer

Målet er at generere repræsentationer fra elementerne i Figur 2 automatisk, dvs. med så lidt intellektuel arbejdsindsats som muligt. Der er i princippet to fremgangsmåder man kan anvende til dette, afhængigt af hvordan dokumenterne er formateret. Hvis dokumenterne foreligger i rå ASCII tekst, PDF eller PostScript, skal både funktionsafsnit og strukturelementer identificeres i indekse-

ringsprocessen, da ikke er bagvedliggende formateringskoder som markerer disse. Det kræver at der opstilles relativt komplicerede indekseringsprogrammer, som kan opmærke og udtrække de ønskede elementer fra rå tekst. Eksempler på dette kan ses i Researchindex (2), der automatisk indsamler videnskabelige artikler i disse formater fra Internet og analyserer dem med særlig fokus på automatisk identifikation af citationer og bibliografiske data, men ikke på dokumenternes struktur eller afsnit som sådan. Et andet eksempel er forskningen udført af Yves Chiaramella der netop fokuserer på at identificere dokumenternes struktur fra rå tekst og derefter konstruere hierarkier, der efterfølgende kan anvendes ved søgning (Chiaramella, Y., 2001). Fælles for begge projekter er, at de kræver relativt megen maskinkraft og ikke kan identificere elementerne med 100 procents nøjagtighed.

Foreligger dokumenterne derimod i strukturerede formater som SGML, eller XML (3) er strukturelementerne og visse af funktionsafsnittene opmærket med såkaldte 'tags', der er meget nemme at indeksere automatisk. Figur 3 viser et kort uddrag af en videnskabelig artikel formateret i SGML. Formateringen i form af tags ses som en kode omgivet af skarpe parenteser, f.eks. **<ABSTRACT>** og **</ABSTRACT>** der angiver begyndelse og slutning på abstractet. Som ved HTML-filer på WWW vises disse tags ikke når artiklen trykkes, men anvendes til at styre produktionen og udseendet af artiklen. I indekseringsprocessen i informationssøgningssystemer kan teksten for hvert tag indekseres og placeres i et bestemt felt. Da det kræver mange ressourcer at udvikle den bagvedliggende DTD (Document Type Definition), der definerer reglerne for hvert tag, anvender forlagene sædvanligvis den samme DTD i lang tid inden den ændres. Der er altså en stor konsistens i formatet i artikler i SGML/XML fra den samme forlægger, hvilket er en fordel når indekseringsrutinerne skal opstilles. Dog er det sådan at de forskellige forlæggere har hver deres DTD. På trods af at langt de fleste af de større akademiske forlag anvender SGML/XML i produktionen af videnskabelige tidsskrifter er det ikke så ligetil at få fat i større mængder af artikler i disse formater. Det skyldes at man som et biprodukt af trykkeprocessen kan producere PDF-filer ud fra SGML/XML-filerne. De har, set fra forlæggernes synspunkt bl.a. den fordel at de består af fikserede billeder, som ikke kan ændres når først de er lavet, hvilket er en garanti for at udskrifter af PDF-filerne bliver præcist som trykkeren har sat dem op (WhitePaper, 2000). Forlagene beholder derfor artiklerne formateret i SGML/XML internt i forlagene, hvilket betyder at det er nødvendigt at samarbejde med forlagene. Et eksempel på et sådant samarbejde er DeLIver projektet (4) der giver mulighed for søgning i artiklerne fra over 50 fysiktidsskrifter, hvor indekseringen af artiklerne og fremvisningen af dem er baseret på SGML.

```
...
<ABSTRACT>
<P>The traditional compensation model to explain the high resistivity properties of CdTe is based
on the presence of a deep acceptor level of the cadmium vacancy in the middle of the band gap. A
new compensation model based on a deep intrinsic donor level is presented. The compensation mo-
del is used together with an appropriate segregation model to calculate axial distributions of resistivi-
ty which are compared with spatially resolved resistivity measurements. The Te-antisite defect is
discussed as a possible origin cause of this intrinsic defect, which is also supported by theoretical
calculations. &copy; <EMPH TYPE="1">1998 American Institute of Physics.</EMPH><II>
&lsqb;S0021-8979(98)02224-5&rsqb;</II></P>
</ABSTRACT>
</FRONT>
<BODY>
<PART>
<CHAPTER>
<TITLE>I. INTRODUCTION</TITLE>
<P>CdTe has been of interest for many years regarding its potential in the use of room-temperature
x- and &ggr;-ray detectors.<CITEREF RID="r1" STYLE="superior">1</CITEREF> The re-
quirements of this application are a very low dark current and a high mobility. To obtain these mate-
rial parameters, the amount of impurities has to be compensated by the doping with chlorine. The
chlorine induces a compensation of the shallow levels in CdTe leading to high
...
</CHAPTER>
</PART>
...
```

*Figur 3: Uddrag af en artikel formateret i SGML. (Kilde: (Fiederle, M. et al., 1998) som den findes i DeLIver projektet (**http://dli.grainger.uiuc.edu/deliver/about.htm**))*

Det er generelt lettere at få adgang til artikler formateret i PDF, men i dette projekt satses der på at samarbejde med et forlag for at få adgang til en større mængde artikler i SGML/XML. Dels fordi der ikke er ressourcer til at programmere de mere komplicerede indekseringsalgoritmer, der fungerer på rå tekst, dels fordi det sandsynligvis vil give en større nøjagtighed at basere sig på SGML/-XML. Afhængigt af forlæggerens DTD vil visse af funktionselementerne ikke kunne indekseres direkte. Det kan derfor blive nødvendigt at bearbejde dokumenterne med avancerede søg-og-erstat-rutiner før de indekseres. For at sikre så stor realisme i eksperimenterne som muligt er det vigtigt at dokumenterne kommer fra det samme fagområde og at der er et stort antal af dem, således at det er muligt at stille realistiske spørgsmål, der kan besvares af dokumenterne i databasen.

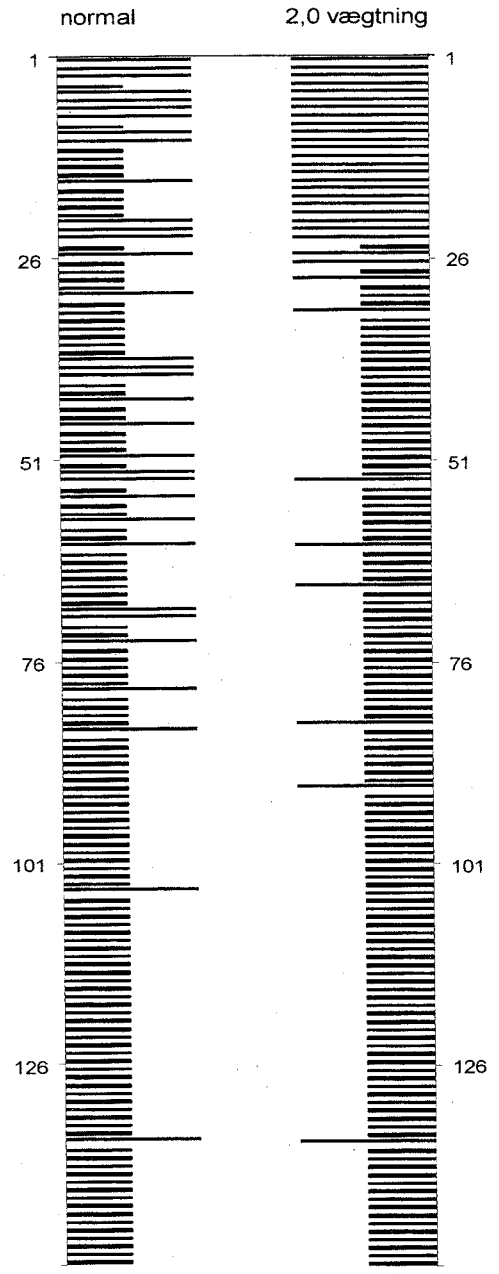
Eksperimentelt informationssøgningssystem

Dokumenterne skal lægges op i et eksperimentel informationssøgningssystem. Her anvendes InQuery, der er udviklet over en årrække af the Center for Intelligent Information Retrieval (CIIR) ved University of Massachusetts. InQuery er et probabilistisk informationssøgningssystem som ofte anvendes i IR-eksperimenter. Programmet har en række fordele i forhold til dette projekt:

- InQuery har en SGML-indekseringsalgoritme der kan modificeres til at kunne læse nye tags.
- InQuery understøtter indeksering og søgning i separate felter.

- InQuery har et udvidet sæt søgeoperatorer der understøtter både exact match og best match søgning.
- Umodificeret er InQuery så færdigudviklet et IR-system, at det er realistisk som sammenligningsgrundlag overfor en modificeret version, hvor polyrepræsentation er indarbejdet.

Aftalen med et forlag er endnu ikke i hus, og InQuery er i skrivende stund under afprøvning. Det er derfor ikke muligt at sige noget om eventuelle problemer med at indeksere de forskellige repræsentationer i fuldtekstdokumenter på nuværende tidspunkt. For at kunne teste InQuery er ca. 3500 bibliografiske poster i SGML lagt op i systemet. Posterne stammer fra Science Citation Index hos Dialog og dækker artikler produceret i Danmark i 1998 indenfor medicin. I denne testdatabase er titler, abstracts og emneord udtrukket fra posterne, og forskellige søgestrategier med overlap mellem repræsentationer har kunnet afprøves. Figur 4 viser virkningen af at vægte dokumenter i overlappet mellem to repræsentationer højere end resten af dokumenterne. Der er foretaget to søgninger i InQuery på 'cancer': én hvor overlappet mellem emneord og titler/abstracts vægtes 2 gange højere end resten af dokumenterne, og én med InQuerys normale vægtning. For at teste om InQuery kan håndtere denne type søgning også som best match anvendes den version af InQuery der rangordner outputtet. Resultatet af de to søgninger er 316 dokumenter rangordnet på to forskellige måder. De 150 første dokumenter fra hver søgning vises i Figur 4: de i alt 33 dokumenter i overlappet mellem de to repræsentationer er markeret med lange, vandrette streger. Resten af dokumenterne, hvor søgetermen cancer kun findes i én af repræsentationerne, vises som korte, vandrette strege. Alle 33 dokumenter i overlappet findes indenfor de første 135 dokumenter i rangordningen, men rækkefølgen og fordelingen af dem er forskellig i de to søgninger. Det ses at den dobbelte vægtning af overlappet betyder, at langt de fleste af dokumenterne heri rangordnes først, dog er der en mindre del af dem som spredes længere nede af listen. Det kan observeres at også InQuerys normale vægtning placerer dokumenter i overlappet højt, men at der er en større spredning mellem dem. Denne og andre indledende tests tyder på, at InQuery er i stand til at håndtere overlappene på den måde der er behov for i projektet.



Figur 4: Vægtning af overlap mellem repræsentationer. I højre side vises spredningen af dokumenter i overlappet med InQuerys normale vægtning. I venstre side er dokumenter i overkappet vægtet 2 gange højere end resten. (Kildedokumenter fra Science Citation Index, online hos Dialog, 2000)

Eksperimentel opstilling
Ved projektets begyndelse var tanken, at der skulle udarbejdes en grafisk brugergrænseflade til informationssystemet så eksperimenterne kunne udføres som interaktive søgeseancer med rigtige slutbrugere som testpersoner. Desværre har det vist sig at være for tidskrævende også at udvikle denne del af det eksperimentelle informationssøgningssystem i projektet. Det er i stedet tanken at anvende en mere klassisk model, hvor søgningerne udføres i laboratoriet, og dokumenterne efterfølgende præsenteres for potentielle brugere af systemet i udskrevet form.

En hel gren af IR-forskningen beskæftiger sig med at udvikle metoder til at evaluere informationssystemer. I dette projekt anvendes de metoder som Pia Borlund har udviklet til dette formål inspireret af det kognitive synspunkt (Borlund, P., 2000a; Borlund, P., 2000b). Centralt i disse metoder er anvendelsen af 'Simulated Work Tasks', der er korte rammefortællinger hvor testpersonerne bliver præsenteret for en arbejdsopgave, som de skal forsøge at løse. Fordelen ved de simulerede arbejdsopgaver er, at de giver mulighed for at testpersonerne kan danne sig deres eget indtryk af arbejdsopgaven og udvikle deres egne fortolkninger af den undervejs i eksperimenterne, samtidig med at der kan fastholdes en vis grad af eksperimentel kontrol. Metoderne kan ses som en hybrid mellem at anvende testpersonernes egne informationsbehov (hvor der er næsten ingen kontrol, men en stor grad af realisme), og at anvende foruddefinerede forespørgsler (hvor der er megen kontrol, men en lille grad af realisme). Metoderne er testet på samlinger af avisartikler og udviklet til at kunne anvendes til evaluering af interaktive informationssystemer, hvor systemerne interagerer med brugerne over flere omgange. De skal derfor tilpasses til situationen i dette projekt, hvor testpersonerne ikke kan interagere direkte med informationssystemet, og da det er en anden dokumenttype der anvendes. Særligt skal der lægges mange kræfter i at udarbejde de simulerede arbejdsopgaver, hvor viden om hvordan de simulerede arbejdsopgaver bedst konstrueres skal kombineres med faglig ekspertise fra fagfolk indenfor dokumenternes emneområde. Dette samarbejde er nødvendigt for at kunne opstille realistiske arbejdsopgaver, der samtidig er anvendelige til eksperimenterne.

Følgende scenario tænkes gennemført i eksperimenterne med hjælp fra forskere eller sidsteårsstuderende indenfor det fagområde som dokumenterne dækker:

- Et antal simulerede arbejdsopgaver udarbejdes i samarbejde med fagfolk.
- Et antal dokumenter identificeres med to versioner af InQuery (med og uden princippet om polyrepræsentation implementeret).
- Dokumenterne fra begge versioner samles sammen og præsenteres for testpersonerne, der vurderer om de er relevante i forhold til den opfattelse de har dannet sig af de simulerede arbejdsopgaver.

På baggrund af dette forventes det, at beregninger af de to versioners effektivitet kan udføres, samt at forskningsspørgsmålene kan besvares og afledte hypoteser enten afvises eller bekræftes.

Projektets betydning

Projektet er relevant på to områder: Det har betydning for IR-forskningen, da en ny tilgang til indeksering og søgning testes empirisk, og det kan få praktisk betydning for de der søger information i strukturerede fuldtekstdokumenter. Projektet udfører empiriske studier af princippet om polyrepræsentation, og kan bidrage med resultater, der kan indikere om denne tilgang er værd at udforske nærmere. I tilfælde af positive resultater kan det føre til yderligere eksperimenter og i sidste ende udvikling af egentlige informationssøgningssystemer, der er baseret på princippet. Projektet kan endvidere styrke og udvikle det kognitive synspunkt, da det udfører empirisk forskning indenfor dets teoretiske ramme, og kan siges at være en del af forskningsfronten, da interessen i at arbejde med strukturerede fuldtekstdokumenter er voksende i takt med at disse bliver alment tilgængelige (5). Projektet kan også være med til at etablere et egentligt miljø for IR-forskning på Danmarks Biblioteksskole og bidrage med erfaringer og ekspertise, så andre forskere på skolen kan udføre eksperimenter med fokus på andre aspekter. På lang sigt kan det betyde, at den danske biblioteksverden dermed får sin egen akademiske forskning i informationssøgningssystemer og brug af dem, uafhængigt af private interesser. Hermed kan man selv sætte dagsordenen og undersøge de problemer, der er vigtige for den danske biblioteksver-

den, frem for alene at basere sig på andres forskning og udvikling.

Projektet kan få en praktisk betydning på lang sigt, da det studerer en relativt simpel tilgang til automatisk generering af repræsentationer, der kan anvendes ved emnesøgning i strukturerede fuldtekstdokumenter. I første omgang fokuseres på naturvidenskabelige dokumenter, men ved nærmere studier af andre fags måder at skrive på kan princippet måske udvikles til også at fungere her. Ved positive resultater vil projektet kunne finde anvendelse i de hastigt voksende digitale biblioteker med videnskabelige artikler, der i dag stort set er uden effektive emnesøgningsredskaber. De store forlag er f.eks. i gang med at opbygge store portaler, hvor artikler fra alle deres videnskabelige tidsskrifter kan skaffes via portaler på Internet. Disse portalers søgeredskaber er dog stadig i stor udstrækning baseret på at understøtte verifikative søgninger, hvor de bibliografiske oplysninger på det ønskede dokument kendes i forvejen. Der er kun meget beskedne muligheder for at foretage egentlige emnesøgninger efter ikke-kendte dokumenter, bl.a. fordi forlagene ikke bruger ressourcer på at lade indeksører tildele emneord intellektuelt (6). Der er endvidere mulighed for at princippet måske også kan anvendes på ikke-videnskabelige dokumenter med en høj grad af struktur, f.eks. politirapporter, patientjournaler, juridiske dokumenter, patenter, tekniske manualer etc.

Noter

1. Flere oplysninger om TREC eksperimenterne kan findes på http://trec.nist.gov/.

2. Tankegangen bag er beskrevet i (Lawrence, S., Giles, C. L., and Bollacker, K., 1999). Systemet kan afprøves på Researchindex.com.

3. XML (eXtensible Markup Language) er delmængde af det mere komplicerede SGML (Standard Generalised Markup Language). SGML anvendes herefter for begge betegnelser.

4. Mere information kan findes på http://dli.grainger.uiuc.edu/deliver/about.htm. Af copyrightmæssige årsager er der desværre kun adgang til søgning i artiklerne fra campus på *University of Illinois at Urbana-Champaign.*

5. F.eks. er IR-gruppen ved Universitet i Dortmund under ledelse af Norbert Fuhr for nylig begyndt udvikle et søgesprog til søgning i XML-dokumenter (Fuhr, N. and Grossjohan, K., 2001).

6. Se f.eks. søgefunktionerne på *Oxford Journals* der giver adgang til ca. 175 videnskabelige tidsskrifter (http://rheumatology.oupjournals.org/searchall/) i fuldtekst eller *Wiley InterScience* der giver adgang til ca. 300 videnskabelige tidsskrifter i fuldtekst (http://www3.interscience.wiley.com/cgi-bin/advancedsearch).

Forfatteren ønsker at takke den danske afdeling af Dialog for generøs adgang til deres databaser samt Haakon Lund og Piet Seiden fra Danmarks Biblioteksskole for deres udstrakte hjælp i forbindelse med opsætningen af InQuery. Projektet der er beskrevet i denne artikel indgår som en del af TAPIR-projektet (Text Access Potentials for Interactive Information Retrieval) Projektet der er beskrevet i denne artikel indgår som en del af TAPIR-projektet (Text Access Potentials for Interactive Information Retrieval) der ledes af Professor Peter Ingwersen, og hører til under Institut for Informationsstudier ved Danmarks Biblioteksskole. TAPIR-projektet udforsker anvendelsen af mangeartede repræsentationsformer af videnskabelige fuldtekstdokumenter med det mål forbedre interaktive best match informationssystemer.

Referencer

Ahlgren, P. (1998): A note on search formulation redundancy. *Journal of Documentation*, 54(3), s. 352-354.

Ahlgren, P. (1999): On a cognitive search strategy. I: Aparac, T., Ingwersen, P. (Eds): *Digital libraries : interdisciplinary concepts, challenges and opportunities : proceedings of the Third International Conference on the Conceptions of the Library and Information Science, Dubrovnik, Croatia May 23-26, 1999*, s. 245-253.

Borlund, P. (2000a): Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1), s. 71-90.

Borlund, P. (2000b): *Evaluation of interactive information retrieval systems*. Åbo: Åbo Akademi University Press, 276 s.

Chiaramella, Y. (2001): Information retrieval and structured documents. I: Agosti, M., Crestani, F., and Pasi, G. (eds.): *Lectures on Information Retrieval: Third European Summer-School, ESSIR 2000, Varenna, Italy, September 11-15, 2000: Revised Lectures*. Berlin: Springer, s. 286-309. (Lecture Notes in Computer Science; 1980)

Croft, W. B. and Thompson, R. H. (1987): I^3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6), s. 389-404.

Fiederle, M. et al. (1998): Modified compensation model of CdTe. *Journal of Applied Physics*, 84(12), s. 6689-6692.

Fuhr, N. og Grossjohan, K. (2001): XIRQL: a query language for information retrieval in XML documents. I: Croft, W. B., Harper, D. J., and Zobel, J. (eds.): *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, September 9-13, 2001*. New York: The Association for Computing Machinery, s. 172-180.Ingwersen, P. (1996): Cognitive perspectives of information-retrieval interaction - elements of a cognitive IR theory. *Journal of Documentation*, 52(1), s. 3-50.

Katzer, J. et al. (1982): A study of the overlap among document representations. *Information Technology: Research and Development*, 1(4), s. 261-274.

Lancaster, F. W. (1991): *Indexing and abstracting in theory and practice*. London: Library Association, 328 s.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999): Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6), s. 67-71. (Online: http://www.neci.nj.nec.com/homepages/lawrence/papers/aci-computer98/aci-computer99.html)

Mann, T. (1997): `Cataloging must change!' and indexer consistency studies: misreading the evidence at our peril. *Cataloging and Classification Quarterly*, 23(3/4), s. 3-45.

McCain, K. W. (1989): Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science*, 40(2), s. 110-114.

Robertson, S. E. and Sparck Jones, K. (1976): Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), s. 129-146.

Salton, G. (1988): A simple blueprint for automatic Boolean query processing. *Information Processing and Management*, 24(3), s. 269-280.

Smeaton, A. F. (1990): Natural language processing and information retrieval. *Information Processing and Management*, 26(1), s. 21-186.

Swales, J. M. (1990): Research articles in English. (7), s. 110-176.

Tenopir, C. (1985): Full text database retrieval performance. *Online Review*, 9(2), s. 149-164.

White Paper (2000): XML and PDF: Why We Need Both - an Introduction to the Two Key Technologies for Electronic Publishing. 16 s. (Et White Paper fra Impressions, online: http://www.impressions.com/resources_pgs/SGML_pgs/XML_PDF.html).