Zoning

Et alternativ til fuldtekstindeksering?

Af Anne Luther og Maria Bøtker Schmidt

Abstract

Artiklen belyser, hvordan den automatiske indekseringsproces kan forbedres. Den automatiske indekseringsproces gennemgås og det specificeres, hvor i den automatiske indekseringsproces tidligere forskning har undersøgt forbedringsmuligheder. Begrebet zoning *analyseres og der sættes fokus på, hvordan undersøgelsesmetoderne dokumentanalyse og interview anvendes til kvalitativt at bestemme hvilken* zone*, dvs. dokumentudsnit, der skal indgå i de videre analyser. Empirien er baseret på undersøgelser lavet i forbindelse med vores kandidatspeciale fra 2004, der bygger på en case fra medicinalvirksomheden Novo Nordisk. I WordStat gennemføres automatiske fuldtekst- og zoneindekseringer af dokumenttypen Standard Operating Procedures (SOPer). De automatiske indekseringer indgår sammen med vore manuelle indekseringer i en række komparative analyser, hvor vi analyserer indekseringerne i forhold til den tidligere skitserede automatiske indekseringsproces. Analysen viser, at kvaliteten af zoneindekseringerne er den bedste af de to automatiske indekseringer, særligt på grund af for mange fejl i fuldtekstindekseringerne, der kan føre til for megen støj i søgeresultatet. Det konkluderes, at zoneindekseringerne kan erstatte fuldtekstindekseringerne, idet titlerne er sigende og zonen derfor indeholder tilstrækkelige emnemæssige data i forhold til brugernes verifikative informationsbehov i forbindelse med genfinding af SOPerne. Sluttelig perspektiveres resultater og metoder. Det foreslås at videre forskning evaluerer forskellige metoder til valg af zone med henblik på at kunne anbefale en metodisk "best practice". Samtidig fremhæves, hvordan videre studier kan undersøge zonings anvendelsesmuligheder indenfor indeksering af skønlitteratur.*

Anne Luther er ph.d.-studerende ved Danmarks Biblioteksskole, København. alm@db.dk

Maria Bøtker Schmidt er cand.scient.bibl. ved Lindø Byg A/S. mabs@lindo-byg.dk

Indledning

Et velkendt karakteristika ved vores samtid er en eksplosive stigning i mængden af information (Burke, 2000; Jensen, 2001; Qvortrup, 2002). En udfordring for alle, der beskæftiger sig med Information Retrieval (IR). En gængs løsning til håndtering af den stigende informationsmængde er automatisk indeksering af fuldtekst-dokumenter. Automatisk indeksering virker (Anderson & Pérez-Carbello, 2001), men i takt med, at databaserne vokser i omfang, udstilles automatisk indekserings svagheder i større og større grad.

Et traditionelt problem i fuldtekstdatabaser er det store antal indekseringstermer, der ofte medfører et uhensigtsmæssigt højt recall, samtidig med manglende precision. Som en mulig løsning på det problem kan algoritmerne konstrueres, så indekseringen baseres på udvalgte dele af dokumenterne i stedet for den fulde tekst. Dermed nedbringes antallet af indekstermer og indekset reduceres. Kowalski og

Maybury (2000) arbejder i deres beskrivelse af automatisk indeksering med begrebet *zoning* som udtryk for den aktion, hvor det identificeres, hvilket udsnit af det enkelte dokument, der skal være genstand for indeksering.

At vælge et udsnit af et tekstdokument til indeksering er ikke nyt. Ved manuel indeksering er det praksis at inspicere særlige dele af dokumentet f.eks. titel, abstrakt, resumé og konklusion, da det ikke lader sig gøre, at indeksøren læser hele dokumentet fra start til slut (Lancaster, 1998). Ved automatisk indeksering viste en undersøgelse af 1.138 artikler i fire forskellige tidsskrifter, at abstracts i 96% af tilfældene er leksikalsk og intellektuel erstatning for den fulde tekst, de repræsenterer (Ries et al. [2001]). Den tråd tager vi op, og har fundet det interessant at se på, hvordan man udvælger den zone, der skal være objekt for den automatiske indeksering. I modsætning til Ries et al. [2001] skal zonen ikke udelukkende agere erstatning for det emnemæssig indhold i dokumentet, men zonen skal desuden indeholde relevante genfindingselementer i forhold til en specifik brugergruppe.

Artiklen er baseret på undersøgelser fra fra vores kandidatspeciale Forbedring af automatisk *indeksering* (Luther Madsen & Schmidt, 2004). Afdelingen Record Management Centre (RMC) ved medicinalvirksomheden Novo Nordisk udgjorde vores casegrundlag. RMC har til hovedopgave at arkivere virksomhedens dokumentation. Vi undersøgte, hvorvidt automatisk indeksering kan forbedres ved at indeksere en såkaldt *zone* (dvs. dokumentudsnit som beskrevet nedenfor) sammenlignet med fuldtekstindeksering. Ideen bygger på en hypotese om, at det er muligt kvalitativt at identificere en zone i et dokument, der kan være objekt for den automatiske indeksering i stedet for dokumentets fulde tekst. Det er et logisk ræsonnement, at en forudsætning for at kunne vælge en zone er, at dokumenttypen har en vis grad af lighed i den indre struktur, dvs. en ensartet opbygning som f.eks. tidsskriftsartikler, hvor elementerne titel, abstract, introduktion, analyse, konklusion, litteraturliste sædvanligvis indgår. Begrebet *zone* og aktionen *zoning* benyttes i overensstemmelse med idéerne i *Information storage and retrieval systems : Theory and Implementation* (Kowalski & Maybury, 2000). For at opnå en kvalitativ udvælgelse af en zone benyttede vi to kvalitative metoder; dokumentanalyse og interviews. Efterfølgende gennemførte vi komparative analyser af fuldtekst-, zone- og manuelle indekseringer, hvor vi anvendte både kvalitative og kvantitative aspekter i analyserne.

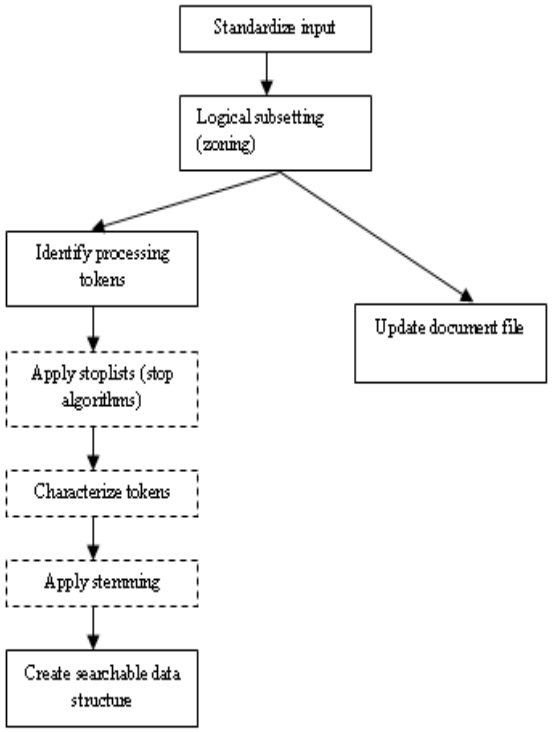
For at tydeliggøre, hvor i den automatiske indekseringsproces tidligere forskning har undersøgt forbedringsmuligheder, indledes artiklen med en gennemgang af den automatiske indekseringsproces. Beskrivelsen tager udgangspunkt i en figur (se Fig. 1) udviklet af Kowalski og Maybury (2000). Kowalski og Maybury tager imidlertid ikke stilling til, hvordan udvælgelsen af den aktuelle zone skal foregå. Dette er et centralt punkt for vores fokus og beskrives indgående i et selvstændigt afsnit. Efterfølgende formidles de metoder, vi anvendte til de komparative studier samt analysernes resultater. Afslutningsvis perspektiveres de anvendte metoder og undersøgelsernes resultater, dels med henblik på videre forskningsmuligheder og dels i forbindelse med praktisk anvendelse.

Den automatiske indekseringsproces

Kowalski og Maybury (2000, s. 58) definerer automatisk indeksering som:

> ”Automatic indexing is the capability for the system to automatically determine the index terms to be assigned to an item”

Figur 1’s venstre side; *identify processing tokens, apply stoplists, characterize tokens, apply stemming* og *create searchable data structure* udgør aktionerne i den automatiske indekseringsproces. Vi koncentrerer os om at beskrive indekseringsprocessen trin for trin i forhold til, hvordan den søgbare datastruktur opbygges og lagres, og forholder os løbende til, hvordan det sikrer en hensigtsmæssig genfinding.



Figur 1: Den automatiske indekseringsproces iflg. Kowalski og Maybury (2000) (note 1).

Figurens stiplede elementer indikerer, at det er valgfrit om aktiviteten indgår i indekseringsprocessen. Inden selve indekseringsprocessen påbegyndes, er det nødvendigt at konvertere informationssystemets objekter til det standardformat, der er læsbare af systemet (f.eks. ASCII eller Unicode) samt at identificere, hvilke dele af det enkelte dokument, der skal være objekt for indeksering (Kowalski & Maybury, 2000). Det vil sige, det besluttes, hvorvidt hele dokumentet er kandidat til indeksering eller blot dele deraf. Kowalski og Maybury kalder denne aktion l*ogical subsetting* eller *zoning*.

Når det er besluttet hvilken zone, der skal gennemgå den automatiske indekseringsproces er næste trin at identificere *tokens*, dvs. identificere de termer eller fraser, der skal indgå i indekset. Hvornår den enkelte token begynder og slutter bestemmes i algoritmen, hvor de nærmere regler for identifikation af tokens defineres. En typisk definition er, at en token identificeres, når en eller flere karakterer er adskilt af mellemrum (blanktegn) eller punktum (Harman, 1994; Anderson & Pérez-Carbello, 2001). Det overordnede mål for fasen er at fastlægge hvilke tokens, der er kandidater til indekset.

Foruden ovennævnte overordnede regelsæt for identifikationen af tokens, har andre aspekter indvirkning på, hvilke regler der opsættes for algoritmen. Behandlingen af tegnsætning, identifikation af tal, store eller små bogstaver, kombinationen tal og bogstaver, fraseindeksering samt ord bestående af et enkelt bogstav kan indgå i udarbejdelsen af det regelsæt, der styrer algoritmen. Bl.a. beskæftiger Harman (1994) og Anderson og Pérez-Carbello (2001) sig med disse aspekter og hvordan de kan forbedre identifikation af tokens til gavn for den automatiske indeksering.

Hvordan algoritmen til identifikation af tokens skal designes er forskellig afhængig af det domæne, systemet skal virke i samt de dokumenttyper, der skal indekseres. Er der f.eks. særlige tegn eller tal der er væsentlige for enten domænet eller en særlig dokumenttype? Det kan diskuteres om, det kan betale sig at udvikle den del af algoritmen, der identificerer tokens. Spørgsmålet er, hvor langvarig og ressourcetung en udviklingsproces det er, set i forhold til hvilke forbedringer der kan måles, sammenlignet med en standardalgoritme til identifikation af tokens.

Som det fremgår af Fig. 1 kan de potentielle tokens efterfølgende udsættes for en stopordsliste, der reducerer indeksets størrelse. En mulig metode til at finde frem til indholdet af stopordslisten er ved simpel optælling af termer i et dokumentkorpus, hvorved højfrekvente ord identificeres (Belew, 2000). Belew (2000) argumenterer for at inkludere højfrekvente termer i stopordslisten, da de almindeligvis ikke er sigende for indholdet i dokumentet. Ingwersen og Willett (1995) konstaterer, at stopord primært er funktionsord og mener, at en stopordsliste typisk består af de hundrede eller tohundrede mest højfrekvente termer, som forekommer så ofte, at det ikke er særlig sandsynligt, at de bliver brugt ved genfinding. Højfrekvente ord på dansk er termer som *og*, *at*, *har*, *en*, *i* osv., der har betydning for sætningens syntaks, men ingen reel værdi i søgeøjemed. Ordene beskriver ikke dokumentets emne og samtidig har de ingen diskriminerende værdi, idet de vil forekomme i alle dokumenter i databasen (Belew, 2000).

En forbedring af den automatiske indekseringsproces kan foregå ved at udarbejde en specifik stopordsliste, hvilket betyder at algoritmedesigneren kan udvide

en generel stopordsliste med de termer, der er højfrekvente for indholdet i en given database, eksempelvis hyppigt anvendte domænespecifikke termer. Man skal dog være påpasselig med at inkludere højfrekvente termer i stopordslisten, såfremt disse er gode indekseringstermer. Det vil sige, at selvom termen har en høj frekvens i det enkelte dokument, og dermed er et potentielt stopord, kan termen være særdeles anvendelig til at diskriminere dokumentet fra databasens øvrige dokumenter, såfremt disse ikke indeholder termen. Termen med høj frekvens kunne jo udtrykke dokumentets emneindhold i særlig grad, og dermed være en god indekseringsterm (Salton, 1997; Belew, 2000).

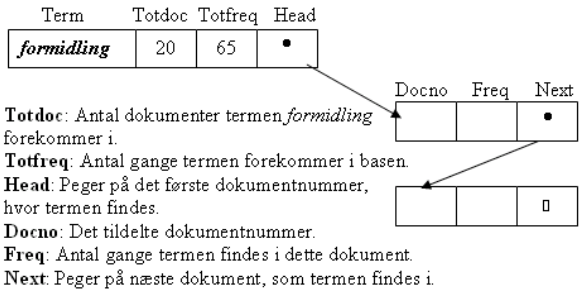
Det næste trin i indekseringsprocessen (Fig. 1) kaldes *karakteristik af tokens*. Elementet er essentielt ved IR-systemer, der anvender Natural Language Processing (NLP). Overordnet er idéen med NLP at identificere sproglige regler (og undtagelser) og udnytte denne viden med henblik på at automatisere processer. For at kunne karakterisere de enkelte tokens, udsættes de for lingvistiske analyser. Det overordnede problem ved lingvistiske analyser i IR-systemer, som i andre NLP-baserede systemer, er, at systemerne arbejder med dokumenter skrevet i naturligt sprog, der ikke umiddelbart kan sættes på formel (note 2). Det viser sig, at hvis der er regler, findes der også undtagelser (Smeaton, 1992). Algoritmedesignen kan vælge at arbejde med karakteristik af tokens på forskellige niveauer, for dermed at forbedre den automatiske indeksering. Spørgsmålet er om udbyttet står mål med de mange ressourcer, det kræver at analysere naturligt sprog med henblik på at designe en algoritme.

Sidste valgfrie trin i Fig. 1 kaldes *stemming*. At stemme betyder at finde termers rod ved at fjerne endelserne. Idéen om stemming bygger på hypotesen om, at termers ordstamme er det meningsbærende element (Porter [1980]; Hull, 1996; Paice, 1996; Kowalski & Maybury, 2000). Formålet med stemming er at få samlet semantisk relaterede termer, for dermed at give termen øget vægtning, der slår igennem i søgeresultatets ranking. En naturlig følge af stemming er øget recall (Belew, 2000). Foruden nævnte formål, reduceres det inverterede filsystem, fordi der bliver færre unikke termer efter endt stemming. En af problematikkerne ved stemming er, at ord uden semantisk lighed samles. Det karakteriseres som over-stemming (Porter, 2001). Antallet af fejl, som en stemmer (dvs. stemming-algoritme) medfører, afhænger i nogen grad af, hvor aggressivt den fungerer. Der findes forskellige stemmere med alternative regler for hvilke endelser, der skal stemmes bort.

Da stemming i visse tilfælde vil medføre fejl, skal algoritmedesigneren overveje, om disse fejl kan accepteres eller om indekseringsprocessen er bedre tjent med helt at fravælge stemming. Er det sidste tilfældet, er brugerne af systemet nødt til at påtage sig arbejdet med at samle semantisk relaterede termer. Enten ved trunkering eller ved at kombinere de forskellige mulige bøjningsformer af en term vha. boolesk søgning. Ifølge Ingwersen og Willett (1995) har evalueringer af stemming-algoritmer vist, at de producerer en acceptabel stemming for omkring 95% af termerne. Samtidig påpeger Ingwersen og Willett i overensstemmelse med Harman (1994), at det ikke er lykkedes at bevise en signifikant forbedring af søgeresultater fra IR-systemer, der anvender en stemming-algoritme sammenlignet med søgeresultater fra IR-systemer, der ikke anvender stemming.

Den automatiske indekseringsproces afsluttes med at kreere en søgbar datastruktur, dvs. et indeks. Hvordan indekset opbygges afhænger af systemets strategi for søgning. Kowalski og Maybury (2000) inddeler disse strategier i *statistical, natural language, concept* og *hyperlinks*. De statistiske strategier er de mest anvendte, og dækker over en række forskellige metoder (ibid.).

Figur 2 viser, hvordan en basisalgoritme fungerer på baggrund af traditionelle simple statistiske metoder. For hver term dannes en inverteret filstruktur, der indeholder oplysninger om antallet af dokumenter, som termen forekommer i, samt det antal gange termen forekommer i hele basen (Belew, 2000).



Figur 2: Basisstruktur i inverterede filsystemer (Belew, 2000).

Fordelen ved den inverterede filstruktur er, at den er alfabetisk, hvilket letter genfindingen i forhold til, hvis alle tekstfiler skulle gennemlæses. Et kendetegn ved statistiske algoritmer er, at de blandt andet baserer sig på frekvenser. Basis-algoritmen indeholder *absolutte frekvenser*, der er frekvenser, som viser, hvor ofte en term optræder i et dokument, eller hvor ofte den optræder i hele databasens dokumentkorpus (Lancaster, 1998). De absolutte frekvenser svarer til figurens *Freq* (hvor ofte en term optræder i et dokument) og *Totfreq* (hvor ofte en term optræder i hele databasen). De statistiske strategier er udtryk for de regler, som udviklerne kan implementere i indeksets algoritme. De absolutte frekvenser kan omsættes til et relativt frekvensmål, baseret på hvor ofte en term forekommer i hele databasen set i forhold til termfrekvensen i et bestemt dokument (Lancaster, 1998; Anderson og Pérez-Carballo, 2001).

Lige som ved de øvrige trin i den automatiske indekseringsproces er der foretaget undersøgelser med det formål at forbedre den søgbare datastruktur (se f.eks. Sparck Jones, 1973; Salton, 1997). Der er blandt andet forsket i at vægte søgeresultaterne, således at der tages hensyn til indholdet i hvert enkelt dokument (lokal vægtning), hele databasens indhold (global vægtning), en kombination af disse samt at korrigere for dokumentlængde (normalisering). Vægtning af indekseringstermer har betydning for i hvilken rækkefølge dokumenterne præsenteres (ranking).

Umiddelbart før den automatiske indekseringsproces påbegyndes skal det bestemmes, hvilke dokumentudsnit, der skal gennemgå indeksering; den såkaldte zoning. Kowalski og Maybury (2000) nøjes med at konstatere, hvad der sker i dette trin, hvorimod vi undersøger denne aktion nærmere i næste afsnit.

Valg af zone

Zoning er den logiske subdivision, der udføres på det standardiserede dokument (Kowalski og Maybury, 2000). Formålet med aktionen zoning er at identificere, de dele af dokumentet, der skal være genstand for indekseringen. Ifølge Kowalski og Maybury kan et systems algoritme være opsat til at udpege f.eks. titel, abstrakt, konklusion, referencer m.v. som indekseringsenheder for hvert dokument, men de beskæftiger sig ikke med, hvorfor og hvordan en aktuel zone identificeres.

Zoning vil naturligt afspejle dokumenternes indre struktur. Om det er hensigtsmæssigt, at algoritmen foretager en logisk subdivision, afhænger derfor af, om dokumenterne har en ensartet indre struktur dvs. en ensartet opbygning bestående af de samme identificerbare genkendelige elementer. F.eks. er virksomheders dokumentskabeloner og tidsskrifters form- og indholdskrav med til at sikre en ensartet opbygning (note 3).

Med udgangspunkt i den nævnte argumentation var det en forudsætning for vore undersøgelser at udvælge en dokumenttype med en ensartet indre struktur. Valget overlod vi til en ledende medarbejder ved Record Management Centre (RMC) ved Novo Nordisk, der som tidligere nævnt udgjorde vores case i undersøgelsen. Efter vores forskrifter fandt vedkommende, at dokumenttypen *Standard Operating Procedures (SOPer)* var en egnet type. SOPer beskriver rutinerne i en arbejdsgang. De er opbygget efter en skabelon, der sikrer en standardiseret form og indholdsopbygning for alle SOPer i hele virksomheden. RMC har 74 SOPer, der vedrører afdelingens arbejdsdomæne. Af rettighedsmæssige hensyn kunne vi få adgang til 31 SOPer ejet (forfattet og løbende revideret) af en medarbejder i RMC. Analysegrundlaget udgør således ca. 42% af afdelingens samlede antal SOPer.

Det er vores opfattelse, at valg af zone kan sammenlignes med det valg, en indeksør står overfor, når vedkommende skal bestemme, hvor i dokumentet de relevante genfindingselementer findes. Derfor er det typisk, at indeksører vælger særlige dele af dokumentet, som inspiceres nærmere f.eks. titel, abstrakt, resumé og konklusion (Lancaster, 1998). Denne identifikation kaldes *indexable matter* (Fidel, 1994; Anderson, 2002). Det er oplagt, at indeksøren overvejer, hvilke intellektuelle dele af et dokument, der skal repræsentere dokumentet ved indeksering (Fidel, 1994). Guidelines som hjælper indeksører til at identificere de dele af et dokument, som potentielt indeholder informationer til gavn ved indekseringen er ikke ualmindelige, f.eks. DS/INF 100:1994. Mai (2005) understreger, at anbefalingerne naturligvis er forskellige afhængig af, hvilken dokumenttype der indekseres. Han argumenterer for, at den dokumentorienterede tilgang til indeksering ikke er tilstrækkelig, men at tilgangen bør være domæneorienteret for på den måde at tage hensyn til brugerne og det domæne, som dokumentet indgår i (Mai, 2005). Det er ikke noget nyt forslag, at indekseringen bør inde-

holde et domæneanalytisk aspekt (se f.eks. Hjørland og Albrechtsen, 1995).

Det er oplagt at overføre disse ideer til aktionen valg af zone for på den måde at opnå en bedre forståelse for den kontekst SOPerne fungere i. Således har vi ikke gennemført en egentlig domæneanalyse, men tilstræbt et undersøgelsesdesign, hvor vi opnår en vis kontekstforståelse.

For at opnå projektets første målsætning: at bestemme den zone, der skal være genstand for den automatiske indeksering, anvender vi metoderne dokumentanalyse og interviews. Fokus for både interviews og dokumentanalyse er at indsamle empiri, der gør os i stand til at pege på den zone, som bedst udtrykker dokumenternes genfindingselementer i overensstemmelse med SOPernes arbejdsmæssige kontekst. Formålet med vores dokumentanalyse kan sammenfattes til tre hovedformål:

- At tegne et billede af SOPernes indre struktur
- At opnå forståelse for arbejdsgangene for at forstå, hvordan og hvornår SOPerne bruges
- At generere spørgsmål til interviewene

Det første mål er inspireret af Steinmark og Zangenberg (1998), der opererer med registrering og *journalisering*, hvor hensigten er at opnå en forståelse for dokumenttypens indre struktur, hvordan den er opbygget samt hvilke data de består af. Det andet og tredje mål kan genkendes, som elementer i den måde Bødker, Kensing og Simonsen (2000) anvender dokumentanalysen. Den skal give os en øget forståelse for arbejdsgangene i afdelingen eller relateret til afdelingens arbejdsområde. Desuden anvender vi dokumentanalysen som inspiration til forberedelse af interviewguide til de efterfølgende interviews (ibid.).

Formålet med interviewene er at få kendskab til, hvordan respondenterne anvender dokumenttypen SOPer. Ved udvælgelsen af respondenter er det derfor et ufravigeligt krav at, de anvender SOPer i deres arbejde. Desuden ønskede vi som et minimum at interviewe to medarbejdere, hvor én skulle have flere års erfaring og den anden skulle være forholdsvis nyansat i afdelingen. Kravet bygger på en hypotese om, at der kan være forskel på, hvordan/ hvornår en erfaren og en nyansat anvender SOPer. Udvælgelsen af respondenter er således et eksempel på en *informationsstyret respondentudvælgelse* (Skot-Hansen og Steffensen, 1995). En ledende medarbejder udpegede efter de nævnte forskrifter tre respondenter, der potentielt kan give et nuanceret billede af brugen af SOPer i afdelingen.

Dokumentanalysen og de semi-strukturerede interviews danner empirigrundlag for at kunne pege på den zone, der skal automatisk indekseres. For at have et fokus for dokumentanalyserne opstillede vi følgende kriterier for valg af zone.

- Zonen skal indeholde emnemæssige data
- Zonen skal indeholde formelle deskriptive data
- Zonen skal indgå i hovedparten af de analyserede SOPer

På undersøgelsestidspunktet var det enkelt at genfinde SOPer. Dokumenttypen er placeret i en separat database og via søgning på afdelingsnummer kan SOPer, der er gældende for den enkelte afdeling eller funktionsområde isoleres. En søgning på afdelingsnummer for RMC gav 74 SOPer – antallet af de tidligere nævnte gældende SOPer for afdelingens arbejdsområde. Det vil sige, at der på daværende tidspunkt ikke eksisterede et reelt genfindingsproblem. På trods af det, valgte vi at betragte udvælgelsen af den endelige zone i lyset af genfindingsaspektet, dog uden specifikt at undersøge søgeadfærden. På baggrund af interviewene anser vi informationsbehovet for at være af verifikativ karakter, dvs. kendetegnet ved, at brugerne ønsker at verificere eller lokalisere et kendt dokument, hvilket er en mindre kompleks søgeproces end hvis der var tale om enten et emnemæssigt informationsbehov eller et mudret informationsbehov (Ingwersen, 1992).

For at kunne bestemme en zone til brug ved de automatiske indekseringer identificerede vi ved dokumentanalyse potentielle zoner i forhold til de førnævnte kriterier og forudsætningen om et verifikativt informationsbehov. Valget faldt på *titelbladet* som objekt for den videre indeksering. Titelbladet består af deskriptive formelle data i form af *udarbejdet af, gældende for, ansvarlige* m.m. Derudover også af en titel, der ligeledes er deskriptiv, men som samtidig indeholder emnemæssige data. Der findes endnu en deskriptiv datatype i zonen, nemlig ISO-emne. ISO-emnet er en klassifikationskode, det henviser til et bestemt emneområde, defineret nærmere i en ISO-standard. Datatypen er både formel og emnebeskrivende. Datatypen kan være anvendelig i for-

bindelse med søgninger, hvor ønsket er at samle alt om et emne. En respondent gav udtryk for, at der er tilfælde, hvor det kan være svært at skelne mellem SOPen, der beskriver, hvordan afleveringen til afdelingen skal foregå og så den tilsvarende SOP for modtagelse. Førstnævnte skal benyttes af medarbejder udenfor afdelingen og sidstnævnte er til intern brug. I dette tilfælde kan ISO emnet være med til at afklare problematikken (note 4), og dermed blive en del af relevansvurderingen af søgesættet.

Titelbladet er den eneste zone, der indeholder deskriptive formelle datatyper. Datatypen er essentiel f.eks. til at identificere SOPer gældende for en specifik afdeling via afdelingsnummeret. Titelblad kan således ikke undværes i indekseringen, fordi de deskriptive formelle data optræder her. Interviewene peger på, at medarbejderne er fortrolige med indholdet i SOPerne bl.a. på grund af træning gennem SOP-spil. SOP-spillet spilles hver tredje måned, hvor alle fra afdelingen mødes "*... ved det store bord og så har vi en masse spørgsmål i en æske, hvor vi så trækker – læser spørgsmålet op og så svarer på det*".

Hvordan medarbejderne anvender de forskellige SOPer varierer afhængigt af, hvilken arbejdsopgave SOPen beskriver. Udvalgte SOPer har de i hånden under selve udførelsen af arbejdsgangen. Samtlige medarbejdere i afdelingen har indgående kendskab til SOPernes indhold, hvilket betyder at medarbejderne i tvivlstilfælde ofte spørger hinanden i stedet for at konsultere den aktuelle SOP. Eksemplerne understreger, at SOPerne er en velkendt dokumenttype og medarbejderne har et indgående kendskab til de enkelte SOPer.

Det er ikke nødvendigt at indeksere emnemæssigt ekshaustivt ved eksempelvis at vælge dokumentudsnittet *Arbejdsbeskrivelse* i SOPerne som indekseringszone. Medarbejdernes indgående kendskab til SOPerne, de emnemæssig sigende SOP titler og de formelle deskriptive datatyper vurderer vi er tilstrækkelig. Den manglende exhaustivitet kunne være problematisk for nye medarbejdere, men via interviewene ved vi, at nyansatte gennemgår intensiv SOP-træning. Oplæringen foregår over flere måneder og nye medarbejdere opnår hurtigt det fornødne kendskab til SOPerne.

Komparative analyser

I dokumentanalysen viste det sig, at 5 af de 31 SOPer, som vi fik adgang til som tidligere nævnt, var en engelsk udgave af den danske SOP. Når SOPerne også forelå på dansk valgte vi at udelade de engelske versioner. Et valg der betyder, at vi ikke vil komme ind på indekseringens sproglige problematikker. De komparative studier realiseres således gennem en række forskellige indekseringer på 26 dansksprogede SOPer. WordStat anvendes som redskab til at foretage de automatiske indekseringer, hvor de 26 SOPer indekseres to gange, hhv. en fuldtekstindeksering og en zoneindeksering. Når de enkelte SOPer lagres i separate tekstfiler, giver det os mulighed for at se termfrekvenserne i det enkelte dokument og dermed bedømme den lokale vægtning af hver enkelt term. Vi er således i stand til kvalitativt at sammenligne fuldtekstens frekvensliste med frekvenslisten for zonen.

Sideløbende gennemføres to automatiske indekseringer, der indekserer samtlige zoner i én tekstfil og samtlige fuldtekster i én tekstfil. Derved er det muligt at se de samlede termfrekvenser og synkront få oplyst, i hvor mange tekstfiler den enkelte term forekommer. Det gør det muligt at vurdere, den globale vægt (note 5), dvs. hvorvidt en term er diskriminerende for den enkelte SOP i forhold til samtlige de undersøgte SOPer.

WordStat har en række funktioner, hvoraf vi udnytter den del, som kan analysere termfrekvenser (SimStat). En fordel ved at anvende WordStat er, at det er velbeskrevet hvordan algoritmen arbejder. Der er tale om en simpel algoritme, der indekserer på enkeltordsniveau uden stemming. Når man undlader at ændre indstillingerne anvendes WordStats prædefinerede *exclusion list* automatisk. Det er en stopordsliste bestående af 130 engelske termer. Den influerer ikke på SOPernes automatiske indekseringer, da de er dansksprogede og ingen signifikante termer bliver fjernet (note 6).

Foruden de nævnte automatiske indekseringer, gennemførte vi en manuel indeksering af de 26 SOPer. Vi valgte en dokumentorienterede tilgang til indeksering, som den karakteriseres af Mai (1999). Dokumentorienteret indeksering foretages eksplicit, hvilket betyder, at man kun indekserer ved hjælp af de begreber, som er ekspliciteret i dokumentet. Der

foregår således udelukkende den form for fortolkning af indholdet, der ligger i at beslutte hvilke begreber, man anser som værende signifikante. Indeksørens udtrækning af nogle termer frem for andre er dermed udtryk for en vægtning af indholdet i dokumentet. En velkendt metode til at øge specificiteten er at kombinere termer vha. fraseindeksering (Lancaster, 1998). Vi valgte at lade fraser indgå i vores manuelle indekseringer, men anvendte ingen kontrolleret emneordsliste som hjælpemiddel til indekseringen. Dog kontrollerede vi navneord for entals- og flertalsendelser og verber for den tidsmæssige bøjning, hvor det var muligt

Valget af en dokumentorienteret tilgang er oplagt, idet vi ikke har fokus på problemstillinger forbundet med diskussionen manuel kontra automatisk indeksering. De manuelle indekseringer agerer i stedet målestok i forhold til zone- og fuldtekstindekseringerne i forbindelse med vægtning af signifikante termer. De termer vi tildeler ved den manuelle indeksering skal dermed gerne have en høj lokal vægtning i de automatiske indekseringer.

Resultater og konklusioner

Fokus for de komparative analyser er at undersøge zone- og fuldtekstindekseringerne i forhold til elementerne i den automatiske indekseringsproces, parallelt måles de i forhold til vores manuelle indekseringer.

Ved identifikation af tokens, altså det at bestemme de enheder (termer/ fraser/ tal m.v.), der skal indgå i indekset, forekommer flere fejl i fuldtekst- end i zoneindekseringerne, hvilket vi tilskriver længden af det indekserede. I den fulde tekst optræder simpelthen flere tokens, hvor eksempelvis tal og tegnsætning skaber problemer.

Begge indekseringer kan have gavn af en stopordsliste, fordi indekserne reduceres væsentligt. En generel stopordsliste vil gøre størst forskel ved fuldtekstindekseringerne. Zonen indeholder ikke ret mange funktionsord og gevinsten ved en generel stopordsliste vil derfor være minimal, når der zoneindekseres.

Ved karakteristik af tokens er fuldtekstindekseringerne de mest problemfyldte. Qua længden på det indekserede observeres flere eksempler på homonymer (ord, der udtales og skrives ens, men som betyder noget forskelligt) end i zoneindekseringerne. Der er derfor grund til at tro at antallet af fejl forbundet med homonymproblematikken er større ved fuldtekstindekseringerne end ved zoneindekseringerne.

Ved fuldtekstindekseringerne viser vores analyse, at stemming vil være en fordel, idet visse signifikante termer vil få en højere vægt, når endelserne fjernes. Vi har ikke undersøgt, hvor stor fejlprocent stemming vil indebære og vil derfor ikke endeligt anbefale, om der skal stemmes eller ej. Derimod observeres, at zonerne sandsynligvis ikke har gavn af stemming, fordi de fleste termer udelukkende optræder en gang i den enkelte zone og dermed har stemming ikke den samlende effekt, der ellers er formålet. Det kunne eventuelt være en fordel at stemme og dermed få samlet emnemæssigt semantisk relaterede SOPer, hvis man ser på det samlede indhold i databasen ved zoneindeksering. På grund af medarbejdernes verifikative informationsbehov, mener vi ikke, der er noget særligt behov for at samle SOPer omhandlende samme emne, og vi mener derfor, at stemming ikke er anvendeligt ved zoneindekseringerne. Denne konklusion har den positive effekt, at designprocessen forkortes samt at man undgår eventuelle stemming-fejl.

Den største kvalitative forskel mellem de to automatiske indekseringer er længden af de inverterede filstrukturer. Ikke overraskende har fuldtekstindekseringerne et langt større indeks end zoneindekseringerne. Fuldtekstindekseringerne indeholder samtlige signifikante termer, men indeholder samtidig rigtig mange ikke-signifikante termer, hvilket vil betyde megen støj i søgeresultatet. Når brugergruppens informationsbehov er verifikativt, vurderer vi desuden, at fuldtekstindekseringerne har en for høj grad af specificitet. Dette er således argumentet for, at selvom zoneindekseringerne mangler en tredjedel af de signifikante termer set i forhold til de automatiske indekseringer, er det ikke problematisk, idet de ekskluderede termer er meget specifikke og derfor ikke nødvendige kvalitativt bedømt.

I forbindelse med fortolkning af SOPernes indhold, har vi beregnet lokale og globale vægtninger for udvalgte signifikante termer. Det viste sig, at fuldtekstindekseringerne har de bedste lokale vægtninger pga. dokumentlængden. Zonen består af mange unikke termer, hvilket medfører, at der ikke er megen forskel på den lokale vægtning af signifikante og ikke-

signifikante termer. Denne problematik formindskes, når vi beregner de globale vægte for zonen.

Samlet set viste analyserne, at kvaliteten af zoneindekseringerne er den bedste af de to automatiske indekseringer, særligt på grund af for mange fejl i fuldtekstindekseringerne, der kan føre til for megen støj i søgeresultatet. Zoneindekseringerne kan erstatte fuldtekstindekseringerne for den undersøgte dokumenttype, idet titlerne er sigende og zonen derfor indeholder tilstrækkelige emnemæssige data i forhold til brugernes verifikative informationsbehov i forbindelse med genfinding af SOPerne.

Perspektivering

Afsnittet har to retninger. En retning hvor metoderne sættes i perspektiv ved at diskutere alternativer til det anvendte undersøgelsesdesign. Den anden retning vil pege på, hvor undersøgelsernes resultater kan bidrage videre i en forskningsmæssig og en praksiskontekst.

Det er oplagt at sætte indsamlingsmetoderne under kritisk lup for at evaluere metodernes anvendelighed og perspektiv i forhold til lignende undersøgelser. Til valg af zone benyttede vi dokumentanalyse og interview. Metoder der viste sig at være egnede. Dog mener vi, at der er behov for yderligere undersøgelser, som stiller spørgsmålstegn ved hvilken metode, der er bedst anvendelig til aktionen *valg af zone*. En anden metode kunne være at gennemføre en workshop med medarbejderne med det formål at indsamle viden om dokumenttypen og den kontekst dokumenterne fungerer i. Eksempelvis ved at få deltagerene til at beskrive, hvordan og hvornår de anvender SOPerne og diskutere eventuelle forskellige brugsmønstre, herunder hvordan de genfinder SOPerne i databasen. Det er interessant at se på fordele og ulemper ved forskellige indsamlingsmetoder samt at bemærke om, resultatet vil frembringe samme zonevalg ved de anvendte metoder. Hensigten med undersøgelserne ville være at kunne foreslå en metodisk "*best practice*" for valg af zone, som virksomheder f.eks. kan anvende, når de skal bestemme, hvilken zone, der skal automatisk indekseres i deres dokumenthåndteringssystem.

For at kunne bedømme om forskellige indekseringer i et system tilfører IR-systemet kvalitet, set i genfindingsøjemed, er det nødvendigt som i vores analyser at bedømme det i forhold til søgeadfærd og informationsbehov. Dette er et klassisk IR problemområde, som er vigtigt at inddrage i lignende studier. Det er oplagt at undersøge om valg af zone vil være forskellig afhængig af hvilke informationsbehov og hvilken søgeadfærd, der kan konstateres. Hvilken zone ville vi vælge hvis forudsætningen var et mudret informationsbehov?

De manuelle indekseringer valgte vi at gennemføre efter den dokumentorienterede tilgang, således tildeles den enkelte SOP udelukkende indekstermer, der i forvejen optræder i dokumentet uden fortolkning og kontrol. Et valg, som tjente vores fokus, hvor de manuelle indekseringer agerer målestok i forhold til de automatiske zone- og fuldtekstindekseringer. På den måde udelukker vi bevidst problemstillinger forbundet med diskussionen manuel kontra automatisk indeksering.

I videre studier er det oplagt at tage udgangspunkt i det faktum, at den bedste performance opnås ved en kombination af manuelle og automatiske indekseringsmetoder (Rowley, 1994). På baggrund af tidligere forskningsresultater er de fleste IR-systemer i dag baseret på en parallel indeksering af det enkelte dokument. I praksis betyder det, at systemerne indeholder to repræsentationer af den enkelte dokument. En repræsentation baseret på indeksørens metadataopmærkning, en anden på systemets automatiske indeksering af dokumentet. Det drejer sig således ikke om et valg mellem hhv. automatisk og manuel indeksering. De to indekseringsmetoder komplementerer hinanden. Et kardinalpunkt vil derfor være at bestemme kombinationer af indekseringsmetoder, der benytter den enkelte metodes styrker til forbedring af IR-systemer. En naturlig udvidelse af vore analyser vil derfor være at gennemføre de manuelle indekseringer eksempelvis efter den domæneorienterede tilgang (Mai, 2005), for på den måde at undersøge, hvilke indekseringstermer de manuelle indekseringer tilfører. Eksempelvis aspekter som dokumenttype, målgruppe, relaterede dokumenter osv. Hensigten med undersøgelserne ville være at komme med anvisninger på kombinationer af fuldtekst-, zone- og manuelle indekseringer indenfor det undersøgte domæne, der kan forbedre genfindingen.

Det er oplagt at afprøve ideen om zoning ved skønlitterære bøger, der traditionelt udgør en problematisk indekseringsenhed. Forlagene kunne udforme

et sæt fælles retningslinier for, hvilke elementer *bagsideteksten* skal bestå af f.eks. et handlingsresumé, hovedpersonernes navne, stedangivelse, hvor handlingen foregår osv. Ved at påtvinge forfatterne af bagsideteksten til at inkludere bestemte elementer er de ubevidst med til at tildele den skønlitterære bog en form for obligatoriske metadataelementer. Videre studier kunne afgøre dels om bagsideteksten er en potentiel zone til en forbedret emnemæssig indeksering af skønlitteratur, men også om katalogposten og zoneindekseringen i kombination er et bud på en forbedret indeksering af skønlitteratur.

Idéen med zoning er, automatisk at indeksere et udsnit af et dokument frem for at indeksere hele dokumentet fuldtekst. Målet er at forbedre system-performance ved at nedbringe støjen og samtidig beholde precision. En forudsætning for at kunne vælge en zone (et eller flere relevante dokumentudsnit) er, at dokumenttypen har en ensartet opbygning bestående af identificerbare genkendelige elementer. Vores undersøgelser tog udgangpunkt i et arbejdsmæssigt domæne, hvor det lykkedes at finde en dokumenttype, der var nogenlunde konsistent i sin opbygning. Skulle tankegangen om zoning videreudvikles kunne en strategi være, at domænet bevidst præger strukturen i de enkelte dokumenttyper ved at udstikke retningslinier for opbygningen. Det kendes allerede fra virksomheders dokumentskabeloner og tidsskrifters forfattervejledninger, men kunne gennemføres i langt større udstrækning som en bevidst strategi.

Noter

1. I lighed med Kowalski og Maybury (2000) behandles aktionen *update document file* ikke i artikelen, da det er udenfor vores fokus. Derimod koncentreres beskrivelsen til de enkelte aktioner i den automatiske indekseringsproces.

2. Trinnet behøver en algoritme, som kan identificere fraser. Samtidig kan der kun benyttes en mild eller ingen stopordsliste, idet stopordene typisk er funktionsord, der anvendes til identifikation af termers ordklasse (verber/substantiver/adjektiver osv.).

3. Kan der ikke umiddelbart identificeres genkendelige elementer må algoritmen enten behandle hele dokumentet som én zone (fuldtekstindeksering) eller inddele i zoner baseret på logisk subdivision i henhold til passagers længde. Eksempelvis ved at identificere de første 10 linier og de sidste 10 linier i et dokument (Lancaster, 1998).

4. SOPen om aflevering har ISO nummer 1610. SOPen om modtagelse har ISO nummer 0500.

5. Beregningerne foretages ud fra den forestilling, at dokumentkorpus alene består af de 26 analyserede SOPer.

6. De engelske stopord, som indgår i danske tekster er: at, i, for. Termer, der antageligt vil indgå i en dansk stopordsliste.

Litteraturliste

Anderson, JD (2002). Indexing, teaching of, *See*: Information retrieval design. *TheIndexer*, 23(1), 2-7.

Anderson, JD & Pérez-Carballo, J (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II : Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management*, 37, 255-277.

Belew, RK (2000). *Finding Out About : A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.

Burke, P (2000). *A Social History of Knowledge : From Gutenberg to Diderot*. Cambridge: Polity Press.

Bødker, K, Kensing, F & Simonsen, J (2000). *Professionelle IT-forundersøgelser : grundlaget for bæredygtig IT-anvendelser*. Frederiksberg : Samfundslitteratur.

DS/INF 100:1994 : Information og dokumentation : vejledning i indeksering : metoder til dokumentanalyse, emnebestemmelse og valg af emneord. Kbh : Dansk Standard.

Fidel, R (1994). User-centered indexing. *Journal of the American Society for Information Science*, 45(8), 572-576.

Harman, D (1994). Automatic indexing. I: R. Fidel (red.). *Challenges in Indexing Electronic Text and Images*. (s. 247-264). NJ : ASIS.

Hjørland, B & Albrechtsen, H (1995). Toward a new horizon in information science : domain-analysis. *Journal of American Society for Information Science*, 46 (6). s. 400-425.

Hull, DA (1996). Stemming algorithms : A case study for detailed evaluation. *Journal of the American Society for Information Science*. 47(1), 70-84.

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

Ingwersen, P & Willett, P (1995). An introduction to algorithmic and cognitive approaches for information retrieval. *Libri*, 45, 160-177.

Jensen, JB (2001). *Midt i en mellemtid : i overgangen fra det gamle til det nye samfund*. Viby J. : Jyllands-Postens Erhvervsbogklub.

Kowalski, GJ & Maybury, MT (2000). *Information Storage and Retrieval Systems : Theory and Implementation*. Boston, MA : Kluwer.

Lancaster, FW (1998). *Indexing and Abstracting in Theory and Practice*. London : The Library Association.

Luther Madsen, A & Schmidt, MB (2004). *Forbedring af automatisk indeksering : En kvalitativ vurdering af indeksering af en udvalgt zone versus fuldtekstindeksering med Novo Nordisk som praktisk eksempel*. Speciale. Aalborg : Danmarks Biblioteksskole.

Mai, J-E (1999). Deconstructing the indexing process. *Advances in Librarianship*, 23, 269-298.

Mai, J-E (2005). Analysis in indexing: document and domain centered approaches. *Information Processing and Management*, 41, 599-611.

Paice, CD (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, 47(8), 632-649.

Porter, MF [1980]. An algorithm for suffix stripping. Lokaliseret 19.12.2005 på WWW: http://www.tartarus.org/~martin/PorterStemmer/def.txt.

Porter, MF (2001). Snowball : A language for stemming algorithms. Lokaliseret 19.12.2005 på WWW: http://snowball.tartarus.org/texts/introduction.html.

Qvortrup, L (2002). *Det hyperkomplekse samfund : 14 fortællinger om informationssamfundet*. Kbh. : Gyldendal.

Ries, JE et al. [2001]. Comparing frequency of content-bearing words in abstracts and texts in articles from four medical journals: An exploratory study. Lokaliseret 02.01.2006 på WWW: http://jimries.com/MedInfo2001/MedInfo2001.pdf.

Rowley, J (1994). The controlled versus natural indexing languages debate revisited : a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), 108-119.

Salton, G (1997). A blueprint for automatic indexing. *SIGIR Forum*, 31(1), 23-36. Note: reprint fra *Sigir-Forum*. 1981. 16 (2).

Skot-Hansen, D & Steffensen, JB (1995). *Bibliotekssociologien : Introduktion til dens grundlag, principper og metoder*. Kbh. : Danmarks Biblioteksskole.

Smeaton, AF (1992). Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*. 35(3), 268-278.

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

Steinmark, C & Zangenberg, H (1998). *Elektronisk dokumenthåndtering : jura, metoder, eksempler*. Kbh. : Teknisk Forlag.