# Informationsgenfinding: Partial match søgeteknikker

## Af Peter Havnø og Lizzi Schlander Hansen

### Indledning

Ved informationsgenfinding matches brugerens informationsbehov, repræsenteret ved hjælp af en query med repræsentationer af dokumenter og tekster i informationssystemerne - med det formål at forene brugeren med forfatternes dokumenter og tekster. En række funktioner indgår i informationsgenfindingssituationen, som beskrevet af Ingwersen (1991, s. 22):

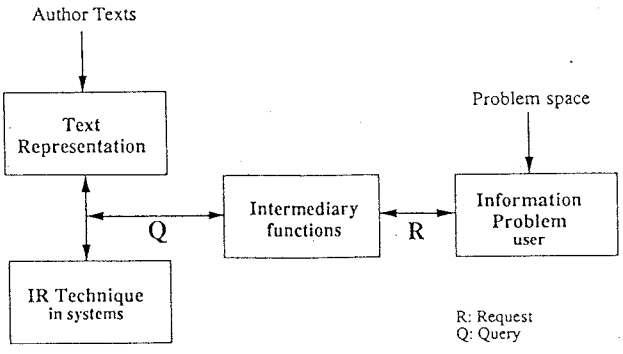


Fig. 1. The information retrieval system and IR interaction. (Fra Ingwersen, 1991, s.22)

De i figuren viste funktioner eller komponenter har alle afgørende betydning for hvorvidt og i hvilken grad informationsgenfindingen (i det følgende også benævnt information retrieval eller IR) vil være succesfuld: fra brugerens erkendelse og formulering af sit informationsbehov over intermediær-funktionens forståelse dels af brugerens informationsbehov, dels af IR systemets søgemuligheder, indre logik og praksis - til systemernes (f.eks. databaseværternes) IR-teknik(ker) og regler og praksis for tekst-repæsentation (samt forfatterne ikke at forglemme). Og i alle funktioner og komponenter er der forhindringer, der modvirker IR processen. Som Ingwersen fastslår:

> "An intriguing characteristic of IR is that its effectiveness - within the limits of present theories and IR models - is far from 100%." ... "The searcher, the IR system, and the IR researcher, 'does not know what he does not retrieve' - and will never know it." (Ingwersen, 1991, s.22).

Vi vil ikke i denne artikel gå ind i detaillerede diskussioner af brugernes, intermediærernes og forfatternes aktioner og interaktioner i IR processen. Vi vil nøjes med at pege på, at menneskelig kommunikation og

udveksling af information, der er en fundamental ingrediens i IR processen altid vil indebære muligheden for fejltolkninger.

Artiklen omhandler IR teknikker, der imidlertid har nøje sammenhæng med *repræsentation*, idet IR teknikken jo netop forener brugerens informationsbehov med forfatternes tekster i en match af *repræsentationer:* i søgeøjeblikket sammenlignes repræsentationer af tekster udtrykt i form af termer med repræsentationer af brugers informationsbehov, udtrykt i form af en query. Vi vil derfor i denne artikel inddrage en diskussion af repæsentationsformer og repræsentationens betydning for retrieval, inden vi går over til en gennemgang og diskussion af IR teknikker, hvor såvel de traditionelle exact match teknikker som de eksperimentelle partial match teknikker, der opererer ud fra term-vægtning, vil blive beskrevet og diskuteret ud fra relevansbegrebet, ligesom vi vil beskrive to forsøg med term-vægtning.

## 1. Informationsgenfinding: systemer og teknikker

Informationsgenfinding er en sammenligning mellem repræsentationer, nemlig repræsentationer af brugerens informationsbehov og repræsentationer af tekster. Idet disse repræsentationer substituerer henholdsvis brugerens behov og den fulde tekst, inklusive deres konceptuelle, kontekstuelle og fonetiske dimensioner, er de ideale krav til såvel repræsentationer som IR teknikker høje.

Noget nedslående beskriver Croft den nuværende teoretiske basis for IR således:

> "To put it simply, we do not know the best way of representing the content of text documents and the users´ information needs so that they can be compared and the relevant documents retrieved." (Croft, 1987, s.249).

Det er ellers på disse områder at den overvejende del af forskningen inden for informationsvidenskaben er sket. Resultaterne af denne forskning har formet og etableret paradigmer og praksis for elementer og funktioner i IR. Repræsentation - i form af klassifikation og/eller indeksering - er grundlæggende for alle eksisterende IR systemer. Klassifikationssystemerne baserer sig for de væsentligste og mest udbredtes vedkommende på samme grundlag, nemlig Dewey og UDC fra sidste del af 1800-tallet. Indeksering ved hjælp af ord, der repræsenterer emne og indhold i et dokument, og som er bærende for udnyttelsen af eksisterende elektroniske IR systemer, følger et begrænset antal etablerede praksiser, hvor især to er udbredte: på den ene side anvendelsen af kontrollerede emneord, påhæftet dokumentet af en menneskelig indeksør under anvendelse af en thesaurus, og på den anden side anvendelsen af ofte automatisk (elektronisk) genererede enkelt-termer i naturligt sprog, hentet fra dokumentet selv. Og repræsentation af brugerens informationsbehov i søgesituationen sker ved formaliseret og syntaktisk kontrolleret formulering af søgetermer under anvendelse af Boolsk logik. Ingwersen (1991, s.24-39) giver et nøjere overblik over disse IR praksiser, deres udvikling og karakteristika.

IR er baseret på repræsentation og på IR teknikker, og for begge er der altså etableret paradigmer og praksis. Og skønt enhver af disse praksiser har været genstand for undersøgelser og kritik, og det er påvist, at andre, nye systemer og metoder vil give bedre resultater, er de forblevet praksis. Som eksempel kan nævnes klassifikationssystemerne. Dewey og UDC med aflæggere er forblevet i brug, på trods af at de, skønt løbende reviderede, repræsenterer en model af verden, der grundlæggende er 100 år gammel. Disse systemer har en række påviste svagheder, hvoraf skal nævnes den implicitte konservatisme i systemerne og en strukturelt betinget vanskelighed ved på rimelig måde at optage nye fagdiscipliner og tværfaglige domæner (se blot indplaceringen af edb i DK5 som eksempel).

Når sådanne systemer og metoder forbliver paradigmer og praksis, skyldes det en række forhold, som påpeget af Belkin og Croft vedrørende exact match IR metoden (Belkin & Croft, 1987, s.113):

> "The traditional answers are that the investment in current systems is so great that changing them is economically unfeasible, that alternative techniques are untested in large-scale environments, and that the results of alternative techniques are not sufficiently better even in experimental environments to justify any changes".

IR indbefatter altså en fremherskende konservatisme m.h.t. paradigmer, systemer og teknikker. Man skifter ikke paradigmer og systemer overnight - det sker yderst sjældent.

Alligevel er der grund til at beskæftige sig med svaghederne ved de etablerede systemer og praksiser, især set ud fra en genfindingssynsvinkel, og med den kritik, der har været fremført.

## 2. Repræsentation

Som nævnt omfatter repræsentation i IR situationen repræsentation af brugerens informationsbehov, såvel som repræsentation af tekster/dokumenter i IR systemet. Til grund for repræsentation ligger begreberne *isness*, der relaterer sig til tekstens fysiske og andre formelle egenskaber (benævnt bibliografiske data ved dokumenter), og *aboutness*, der betegner beskrivelsen af, hvad en tekst handler om, og hvorved der kan lægges to synsvinkler: *author-aboutness* og *indexer-aboutness*.

### 2.1. Author-aboutness. Repræsentation ved NLR

Ved author-aboutness er grundlaget for repræsentationen forfatterens egne formuleringer i teksten, idet synspunktet er, at forfatteren ved sit ordvalg har givet teksten det informationsindhold, som den skal genkendes og genfindes på. Dette synspunkt danner baggrund for repræsentation ved enkelt-termer, hentet direkte fra teksten (Natural Language Representation, NLR), og som er udgangspunkter for Salton og beslægtede forskeres informationsbegreb:

> "Inherent in Salton's and alike researchers' information concept aboutness is associated with content bearing units in the text, generated by the author (Salton, 1968; Salton & McGill, 1983). Consequently, one may represent information by single terms derived directly from the document itself." (Ingwersen, 1991, s.19-20).

Og:

> "For the purposes of IR, this information content must and can be represented, in Salton's example and experiments, usually as a set of words each of which represents some aspect of the information content, all of them together representing the information content of the record as a whole. Thus, content and information seem closely linked, if not identical, information being an invariant attribute of the record which can be represented as a set of individual elements, each capable of being individually considered." (Belkin, 1978, s.62-63).

Enkelt-term NLR, hvor termer genereres fra dokumentets titel og abstrakt, er den mest udbredte indekseringsmetode i online IR systemer. Der sker ikke nogen filtrering eller tilføjelse af terminologi, som ikke er ophavets egen. Dermed er ophav og bruger bragt tættere på hinanden. Og dermed må man ved søgning jo så forlade sig på forfatterens terminologi på godt og ondt, hvor problemer med synoymer, homonymer spiller ind - samtidig med at problemerne med mulig forskellig terminologi brugere og ophav imellem fortsat eksisterer. Endvidere indebærer repræsentation ved NLR implicit inkonsistens i indeksering eftersom forfattere ikke kan forventes at *tilstræbe* konsistent terminologi!

Et andet væsentligt aspekt er, at termerne er revet ud af deres kontekst. Hver enkeltterm er uafhængig, tillægges absolut mening og semantisk korrespondence. Winograd og Flores (1986/1987) stiller afgørende spørgsmål ved den rationalistiske sprogopfattelse, der bl.a. ligger til grund for denne repræsentationsform:

> "Just what is the relationship between a sentence (or a word) and the objects, properties, and relations we observe in the world? Few philosophers adhere to the naive view that one can assume the presence of an objective reality in which objects and their properties are 'simply there'. They recognize deep ontological problems in deciding just what constitutes a distinct object or in what sense a relation or event 'exists'." (Winograd & Flores, 1986/1987, s.18).

### 2.2. Indexer-aboutness. Repræsentation ved kontrollerede emneord

Indexer-aboutness har udgangspunkt i teksten, men teksten fortolkes af indeksøren, hvorved indeksøren dels må placere teksten i et indholdsmæssigt fællesskab med andre tekster (ved hjælp af klassifikation), dels uddrage tekstens særlige emnemæssige karakteristika i forhold til andre tekster (ved hjælp af emneord, typisk kontrollerede emneord, hvortil er knyttet en thesaurus).

Brugen af kontrollerede emneord skulle indebære konsistens i repræsentationen, således at:

> "... new documents can be linked to old ones by consistent use of terms" (Ingwersen, 1991, s.27).

Dette forudsætter, at søgererne og indeksørerne (og til dels forfatterne) faktisk benytter samme terminologi, hvilket ikke altid er givet, blandt andet fordi et kontrolleret vokabularium implicit er konservativt, hvorved nye termer og ord ikke giver resultat, men også fordi forskelle i indeksørers og brugeres terminologi med baggrund i forskelle i kultur, uddannelse, fag og skoler/synspunkter spiller ind. En anden forudsætning er, at indekseringen faktisk er konsistent, hvilket undersøgelser viser langtfra altid er tilfældet: indeksørkonsistens varierer fra 10%-80% (Wormell, 1984, s.24).

Søgning på kontrollerede termer kræver brug af thesauri, og for så vidt at disse ikke er (og ikke kan være) detaljeret udtømmende, må brugerens informationsbehov, udtrykt i naturligt sprog, fortolkes og omsættes til en kontrolleret term - der ved at være synonym, bredere eller beslægtet må antages at repræsentere brugerens informationsbehov.

Hyppigt forekommende i IR-systemer er repræsentation ved såvel kontrollerede termer som ved enkelt-term NLR. Herved opnås en slags *author + indexer aboutness*, der giver bedre resultater end hver af de to former individuelt, uden dog at eliminere de anførte ulemper ved hver af repræsentationsformerne.

I ingen af de to repræsentationsformer (og dermed selvsagt heller ikke i en kombination af de to) indgår en vurdering af termernes relevans, idet alle termerne vægtes ens, uanset deres frekvens eller andre indikatorer for relevans (se afsnit 4 om vægtning af termer).

## 3. Teknikker til informationsgenfinding

Det er ved brug af IR-teknikker (eller match-funktioner om man vil), at brugerens informationsbehov, repræsenteret i queryen, matches med IR-systemets repræsentation af tekster. Belkin og Croft definerer IR systemer som:

> "... the means for identifying, retrieving, and/or ranking texts (or text surrogates or portions of texts), in a collection of texts, that might be relevant to a given query ... In particular, retrieval techniques adress the issue of comparing a representation of a query with representations of texts for the above purpose." (Belkin & Croft, 1987, s.109).

Ideelt skal IR-teknikken finde de og kun de dokumenter, der er relevante for en brugers informationsbehov. Query'en repræsenterer brugerens informationsbehov, og kan (afhængigt af teknik) formuleres i et logisk baseret sprog eller i naturligt sprog.

Skønt søgning med Boolsk logik er den altdominerende IR-teknik kommercielt, er der op gennem 60'erne og 70'erne foregået en omfattende forskning i andre teknikker.

Belkin og Croft giver en en oversigt over status for forskningen inden for IR-teknikker (1987). Artikelen opstiller en hierarkisk klassifikation over IR-teknik typer, hvis hierarki er vist her (i ændret opsætning):

- Exact match
- Partial match
  - Individual
    - Feature based
      - Formal
        - Vector-space
        - Probabilistic
        - Fuzzy sets
      - Ad-hoc
    - Structure based
      - Logic
      - Graphic
  - Network
    - Cluster
    - Browsing
    - Speading activation

### 3.1 Exact match teknikker

Exact match teknikken er den teknik, som vi kender fra de store databaser i form af Boolsk søgning. Her kræves et eksakt sammenfald mellem formuleringen af brugerens informationsbehov, udtrykt ved en query i form af enkelttermer eller fraser, og de termer eller fraser, der repræsenterer de tekster, som man søger efter. Boolsk søgning og de senere udvidelser med søgning på nærhedsoperatorer er først og fremmest udviklet via praksis og har ikke en teoretisk baggrund.

Belkin og Croft opregner en række ulemper ved disse teknikker:

> "The disadvantages of this type of technique are well known and well documented and a variety of aids

such as thesauri are required to achieve reasonable performance. In the simple case exact match searching: 1) misses many relevant texts whose representations match the query only partially; 2) does not rank retrieved texts; 3) cannot take into account the relative importance of concepts either whether the query or within the text; 4) requires complicated query logic formulation, and 5) depends on the two representations being compared having been drawn from the same vocabulary." (Belkin & Croft, 1987, s.113).

De angiver som begrundelse for at exact match stadig bruges i de kommercielle systemer først og fremmest økonomien, d.v.s. at der er investeret så store beløb i det nuværende software, at databaseværterne ikke vil skifte det ud, og desuden, at man ved hjælp af Boolsk logik kan specificere væsentlige aspekter af en forespørgsel. Ingwersen (1991) refererer desuden fra en undersøgelse af Kochen og Summitt, at værterne hævder, at resultaterne fra de forskelige forsøg i test baser ikke giver så væsentlig bedre resultater, at det kan begrunde indførelse af nye teknikker.

Exact match teknikken er knyttet nøje til de traditionelle indekseringsmetoder, og mulighederne for sortering efter termfrekvens, f.eks. i et abstract, udnyttes ikke. Relevans er derfor et enten/eller.

Der er dog gjort forsøg på at modificere exact-match søgning i form af udvidet Boolsk søgning i forbindelse med vector-space eller probability teknikker, med zoom teknikken hos ESA, der giver en liste over termerne i søgesæt, sorteret efter deres frekvens (Ingwersen, 1984) og med at bruge termpositions-angivelserne til at rangordne søgeresultater (Keen, 1991).

Endvidere har der fra 80'erne og frem foregået omfattende aktiviteter med at udvikle supportive og intelligente front-ends, der netop (i større eller mindre udstrækning) bringer løsninger på de rejste kritikpunkter - ved at omsætte brugerformulerede queries i naturligt sprog til f.eks. Boolsk logik, ved at stemme ord til deres betydningsbærende rod, ved at sætte trunkeringer og ved at rangordne dokumenter ved efterfølgende termfrekvensberegninger.

### 3.2. Partial match teknikker

Som en modsætning til exact match systemerne i de kommercielle systemer ses partial match teknikkerne, som i stor udstrækning er teoretisk baserede og afprøvet i mindre test-baser. En afgørende forskel er at man i partial match teknikkerne prøver ved brug af forskellige metoder at finde repræsentationer af de dokumenter, som har en lighed med en query (eller request), og at liste de fremfundne dokumenter i en prioriteret rækkefølge, som et udtryk for aftagende grad af lighed med query'en. I partial match formulerer man ikke Boolske relationer i query'en, man bruger blot en ustruktureret liste over søgetermer eller i ekspertsystemerne en request i naturligt sprog, der ved hjælp af stopordsliste og stemming (trunkering) omformes til søgetermer. Man forsøger derefter ved forskellige former for beregninger kvantitativt at udtrykke relevans som graden af lighed mellem query'en og repræsentationerne af dokumenterne i samlingen.

De partielle match teknikker omfatter en række grundlæggende forskellige metoder, der kan hovedopdeles i individuelle og i netværk baserede teknikker. Vi vil ikke her give en detaljeret beskrivelse af dem alle, men bringe en kort oversigt over de mest omtalte og sidst i afsnittet diskutere en af de eksperimentelt vel nok mest afprøvede, nemlig vector-space teknikken.

#### 3.2.1 Individuelle teknikker

De individuelle IR-teknikker underdeles i struktur- og featurebaserede, hvor man i de featurebaserede sammenligner enkelttræk eller termer i dokumenterne med queryen. De strukturelle metoder kan ses som en meget kompleks form for thesauri, baseret på domæneviden inden for det enkelte område, enten i form af logiske udsagn baseret på analyser af naturligt sprog eller grafiske ligheder mellem dokumenter og query.

De formelle featurebaserede teknikker opdeles i vector-space, sandsynlighedsbaserede og fuzzy sæt.

##### 3.2.1.1 Vector-space modellen

Den ældste er vector-space, der som Belkin og Croft fremhæver, har en intuitiv appel (1987, s.115), og som har været brugt i megen forskning. Modellen beskrives udførligt i afsnit 4.1. Vector-space er en af de teknikker, der benyttes i SMART-systemet, der er et test

system for forskellige metoder til automatisk indeksering og søgning. (Salton & McGill, 1983). Vector-space modellen kan bruges til at sammenkoble exact match og partiel match teknikker i form af den udvidede Boolske søgning, som ud fra en oprindelig Boolsk søgning ved hjælp af en beregning af ligheder producerer en prioriteret liste over søgeresultatet (se afsnit 4.1.1).

#### 3.2.1.2 Sandsynlighedsbaseret model

Den sandsynlighedsbaserede model minder meget om vector-space modellen. Formålet er under søgningen at finde de dokumenter, der har størst sandsynlighed for at være relevante i forhold til en query. Vanskeligheden består i at beregne sandsynligheden for relevans i en samling, som man ikke kender. Det kan ske ud fra en relevans-vurdering foretaget af en bruger på grundlag af et søgesæt, eller man kan bruge tf.idf vægte, som påvist af Croft og Harper (1979).

#### 3.2.1.3 Fuzzy sæt modellen

I fuzzy sæt modellen er det kun dokumenttermer, der vægtes. Query'en udføres som en standard Boolsk formulering. Rangordningen sker derefter efter følgende regel:

ELLER-relation: (A OR B): højeste vægt af A eller B i et dok.
OG-relation: (A AND B): laveste vægt af A eller B i et dok.
IKKE-relation: (NOT A): 1-vægten af A i et dok.

### 3.2.2 Netværksteknikker

I netværksteknikkerne, som Belkin og Croft opdeler i cluster, browsing og spreading activation, er genfindingen baseret på dokumenternes forbindelser til andre dokumenter, som har et lignende indhold. Disse relationer etableres så i form af klynger af dokumenter (svarende til en klassifikation), eller links, etableret ved hjælp af indeksering, thesauri, citationer, forfatterrelationer m.v., som en bruger får mulighed for at vælge i en browsing situation.

#### 3.2.2.1 Clustering

Et cluster (en klynge) er en gruppe dokumenter, hvis indhold har lighed. Cluster-teknikken søger principielt denne lighed mellem tekster og etablerer hierarkiske klynger af tekster med ensartet indhold af termer.

Klassisk , d.v.s. i henhold til Salton, formeres cluster-hierarkiet automatisk ved at opdele dokumentsamlingen i et antal store klynger ud fra ligheden m.h.t. index-termer. Klyngerne opdeles hver især i mindre klynger med højere grad af lighed o.s.v. Ved søgning sammenlignes query'en med det øverste hierarkiske niveau af klynger, og ud fra en sammenlignings-koefficient vælges det bedste, og sammenligning fortsætter nedefter klynge-niveau for klynge-niveau, ned til laveste niveau. Klynger på laveste niveau er herefter ranket i forhold til lighed med query'en, og i den bedst rankede klynge er dokumenterne ranket individuelt.

En variant af cluster-teknikken er 'nearest-neighbour' metoden, hvor et dokument er forbundet til dets nærmeste nabo-dokumenter i et netværk, der etableres ved hjælp af metoder til måling af term-lighed, og hvor klyngerne således genereres i søge-øjeblikket.

En barriere for videre udvikling af cluster-teknikken knytter sig til anvendelsen af *sproget* som udgangspunkt for vurdering af lighed:

> "However, the most severe problems encountered in relation to the clustering techniques, which can be seen as a form of automatic classification, are the issues of determining the *linguistic basis* for term associations and the provision of formal definitions between a pair of terms, and association among a class of terms" (Ingwersen, 1991, s.36).

Og her er vi atter ved et centralt punkt i IR-processen: brugen af sprog som repræsentation.

### 3.2.3 Relevans feedback

Mange af de ovennævnte teknikker bruges i de eksperimentelle systemer til at forbedre næste søgetrin. Når resultatet af den første søgning har været fremvist for brugeren, der har markeret, hvilke dokumenter, der er relevante, bruges disse dokumenter som udgangspunkt for en ny søgning, enten som ovenfor nævnt i sandsynlighedsbaseret model, eller i vector-space, hvor man ud fra de relevante dokumenter tilføjer nye termer til query'en, der så igen sammenlignes med dokumentsamlingen. Den samme teknik kan benyttes i netværksbaserede teknikker som clustering (Salton & McGill,

1983). Som Ingwersen påpeger, kræver dette imidlertid enten at brugeren specificerer den nye query, eller at der er fastlagt regler for hvilke del af en query der skal genbruges eller helt forkastes. (Ingwersen, 1991).

### 3.2.4 Vurdering af forsøgsmodeller

Partial match teknikkerne er som tidligere nævnt udviklet i eksperimentelle miljøer, hvor man har afprøvet systemerne på nogle standard testsamlinger. I forbindelse med disse testsamlinger har man også en samling query'er,som man kender svaret på, d.v.s. man ved hvor mange relevante dokumenter, der findes i samlingen. Man kan så vurdere sit system ud fra de recall/precision resultater, man får ved anvendelse af de forskellige teknikker.

Recall er forholdet mellem de relevante dokumenter, som er fundet ved søgningen og de relevante dokumenter i samlingen ialt. Recall forøges ved at anvende brede, høj-frekvente termer, der finder mange dokumenter.

Precision er forholdet mellem antallet af relevante dokumenter i søgningen og antallet af dokumenter i søgningen ialt. Precision forøges ved at benytte specifikke termer, som er gode diskriminatorer. Formuleringen af en query vil ofte være et kompromis mellem disse hensyn, og man forsøger derfor at finde systemer, der tilgodeser begge dele.

Belkin og Croft anvender den tommelfingerregel, at en forøgelse af recall og precision på 10 % er signifikant. Ud fra dette kommer de til de forsigtige konklusioner, at partiel match er bedre end exact match, og at sandsynlighedsbaseret teknik er den bedste feature-baserede teknik samt at klynge-teknikken kan opnå resultater, der svarer til de feature-baserede teknikker.

Der er imidlertid mange tvivlsspørgsmål.

Et fundamentalt spørgsmål: hvad er relevans egentlig?

Ifølge Ingwersen skal relevans forstås som en pragmatisk værdi, der er tæt knyttet til den individuelle brugers 'problem space' og 'state of knowledge'. Hvad der er relevant for den enkelte bruger, kan kun han afgøre. Unægtelig en subjektiv størrelse:

"Exactly identical queries from two users may often result in totally distinct relevance judgements." (Ingwersen, 1991, s.39).

Der er udsagn, der tyder på, at nogle teknikker måske er bedst egnede til bestemte typer databaser eller bestemte typer queri'es og andre teknikker til andre baser/queri'es (Salton & Buckley, 1988; Sparck Jones, 1973). Der tegner sig ikke noget klart billede, og i forsøgene på etablering af automatiske intermediarysystemer benytter man derfor ofte flere forskellige søgeteknikker, der hver især giver forskellige resultater, for at forsøge at forbedre det totale recall/precision. I I3R systemet benyttes således både probability og clustering, såvel som exact match (Ingwersen, 1991).

Når vi i det følgende vil se nærmere på vector-space modellen, er det således ikke fra overbevisende udsagn om, at dette er den bedste model, men snarere ud fra den tiltrækning, der ligger i en matematisk model, hvor resultatet umiddelbart kan beregnes, og dels at den i en form for udvidet Boolsk søgeteknik kan have appel til mange intermediærer, der har behov for en bedre præsentation af søgeresultater.

## 4. Vægtning af termer

I exact match søgning har alle termer og alle dokumenter lige stor vægt. Dokumenterne kan enten være relevante, hvis de indeholder samtlige de termer, som queryen består af, og de opfattes i så fald som lige relevante allesammen, eller de kan være irrelevante, uanset om de slet ikke indeholder nogen af querytermerne, eller blot mangler en enkelt.

I partial match forsøger man som tidligere nævnt at graduere søgeresultatet efter større eller mindre lighed, og i den forbindelse kommer vægtningen ind.

Automatisk vægtning af enkelt-termer ud fra den hyppighed hvormed de forekommer i en tekst baserer sig på Zipf's opdagelse af at frekvensen af en given term i en tekst mulipliceret med termens rank orden tilnærmer sig en konstant for teksten. Sparck Jones (1973) har ved forsøg med tre kendte testsamlinger påvist det forbedrede søgeresultat ved brugen af termvægtning.

De grundlæggende antagelser er:

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

Dette udtrykkes som reglen som en terms tf.idf vægt:

$$w = tf*idf$$

hvor tf=termens forekomst i dokumentet og idf (den inverterede dokument frekvens)=den inverterede funktion af termens forekomst i samlingens dokumenter.

For at tage hensyn til pkt. 2 ovenfor kan tf beregnes som dokumentets relative termforekomst. Sparck Jones hævder at:

> "document length weighting may be of some limited value. But it is certainly not worth much trouble" (Sparck Jones, 1973, s.631).

Salton og Buckley (1988) foreslår dog en faktor, som normaliserer længden af dokument vektorer, således at lange dokumenter med mange termer ikke har bedre chancer end korte. Denne normaliseringsfaktor har formlen:

$$w/\sqrt{\Sigma(wi)^2}$$

og indgår i vector-space formlen (se nedenfor).

Derimod er der bred enighed om at anvendelse af dokumentfrekvensen giver betydelig bedre resultater i recall/precision, iden det giver større vægt til de sjældne termer, og mindre til de hyppige termer, som indgår af hensyn til en rimelig recall.

Idf kan enten beregnes som 1/dokumentfrekvensen eller som det hyppigere ses:

$$\log(N/dfk)$$

hvor N=antallet af dokumenter i samlingen og dfk= dokumentfrekvensen.

Dokument-termernes vægte kan beregnes enten i forbindelse med indekseringen, hvorved vægten bliver en statisk størrelse eller ved selve søgningen, og man vil da i vægtningen hele tiden kunne tage hensyn til ændringer i databasens størrelse og sammensætning. I det følgende vil vi med dokument-vægtning udelukkende henføre til den sidste form, altså en dynamisk vægtning, som beregnes ved søgningen.

Man kan beregne vægte af både query-termer og dokument-termer eller kun af den ene del.

I stedet for en *beregning* af query-termernes vægte, kan man *tildele* query-termerne forskellige vægte, baseret på brugerens vurdering af de enkelte termers indbyrdes betydning.

Et dokuments vægt i relation til en query beregnes som summen af de termers vægte, som de har tilfælles. Dette udtryk kan benyttes til at opstille en rangorden for en samling dokumenter i relation til en query.

Der har i ovenstående kun været tale om vægtning af enkelttermer. Man kan naturligvis også vægte sammensatte termer, fraser o.lign. Fra kun at bruge vægtning af enkelttermer har man forsøgt sig med relaterede termer, baseret på statistisk samfaldende forekomster, vægtning af fraser, gruppering af termer a la thesaurus eller ud fra domæne-specifikke viden-baser. Alt dette har dog ikke medført overbevisende resultater, og Salton og Buckley konkluderer:

> "In reviewing the extensive literature accumulated during the past 25 years in the area of retrieval system evaluation, the overwhelming evidence is, that the judicious use of single-term identifiers is preferable to the incorporation of more complex entities extracted from the texts themselves or obtained from available vocabulary schedules" (Salton & Buckley, 1988, s.515).

### 4.1 Vector-space modellen

Vector-space modellen er en videreudbygning af den ovenfor beskrevne vægtning af termer. I formlen indgår ud over de enkelte termers td.idf vægte også en normalisering for forskellige længder af dokumenter.

I SMART systemet, hvor vector-space modellen som tidligere nævnt benyttes, opfattes hvert dokument som repræsenteret af en vektor af termer, hvor vægten af den enkelte term skal angive den vægt, hvori den indgår i dokumentet. Hvis en term er tilstede i et dokument, har den en positiv vægt, hvis den ikke er tilstede er vægten=0. På samme måde kan en query ses som en vektor bestående af termer. For en tydeligere forklaring af vector-space, se Nielsen, 1993.

Det er imidlertid ikke alle ord i dokumentet, der bruges. I forbindelse med en automatisk indeksering fjernes først alle de ord, som forekommer på en standard stopordsliste. De resterende reduceres til ordstammer (stems) ved hjælp af en "stemming" eller "conflation algoritme", der svarer til at man foretager en manuel trunkering (Lennon et.al, 1981). Salton indfører desuden en "term-discrimination model", da han hævder, at både de meget sjældne og de meget hyppige termer er dårlige "discriminators", og han erstatter de lav-frekvente ord med thesaurus-ord af en højere orden, og de højfrekvente med fraser (Salton, 1986). Tilsvarende behandles en request i naturligt sprog og omformes derved til en query.

I et exact match system skulle alle queryens termer være tilstede i dokumentet for at det blev fundet. Her kan man istedet angive en vilkårlig grad af lighed mellem dokumentets vektor og queryens vektor, som tilstrækkelig til at dokumentet repræsenteres som et resultat af søgningen. I beregningen af denne lighed benyttes cosinus relationen, som Salton forklarer således:

> "The cosine correlation measures the cosine of the angle between documents or between documents and queries, when these are viewed as vectors in the multidimensional term space of dimension t." (Salton & McGill, 1983, s.195).

Til beregning bruges formlen:

$$\text{COSINE}(\text{DOC}_i,\text{QUERY}_j) = \frac{\sum_{k=1}^{t} (\text{TERM}_{ik} \cdot \text{QTERM}_{jk})}{\sqrt{\sum_{k=1}^{t} (\text{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\text{QTERM}_{jk})^2}}$$

(Fra Salton & McGill, 1983, s.195)

I tælleren beregnes summen af vægten af de termer, som dokumentet og queryen har tilfælles. Vægten af termerne i dokumentet udtrykkes som deres tf.idf vægt. Nævneren fungerer som en normaliseringsfaktor, som udligner forskelle mellem længden af de enkelte dokumenter. Som resultat af denne beregning fås et tal mellem 0 og 1, og jo tættere på 1, jo større lighed er der mellem dokument og query.

Ved at beregne vector-space for alle dokumenter i en base, vil man således kunne opstille dem i en rangorden, hvor de højeste tal skulle give de dokumenter, der er de mest relevante i forhold til det, man søger. I SMART systemet lister man imidlertid ikke samtlige dokumenter i forhold til hver enkelt søgning, man angiver en grænseværdi, f.eks. 0,5 og kun dokumenter over denne grænseværdi vil blive fremvist. Man kunne også vælge som standard kun at fremvise de 10 eller 20 højest placerede dokumenter.

### 4.1.1 Udvidet Boolsk søgning

I SIRE systemet udføres først en almindelig Boolsk søgning. Dette resulterer i et søgesæt baseret på exact match mellem query og dokumenter. Ud fra dokumenterne i dette søgesæt beregnes så ligheden ved hjælp af cosinus relationer ud fra en "flad" query, hvor alle termer indgår som om de var "eller" relationer, d.v.s. en query bestående af (A or B) and (C or D), bliver til A or B or C or D. Derefter kan søgeresultatet præsenteres i faldende rangorden. (Salton & McGill, 1983)

Dette har senere ført til udviklingen af "extended Boolean retrieval system", som er en kombination af en konventionel Boolsk søgning, en fuzzy sæt søgning og en vektorberegningsmodel, hvorved man får fordelen af både den strukturerede Boolske logik, vægtning af termer og rangordning af søgeresultater (Salton, Fox & Wu, 1983; Salton, Fox & Voorhees, 1985, 1985).

## 5. Forsøg med termvægtning i praksis

Alle disse teknikker har først og fremmest været brugt i eksperimentelle systemer og er blevet vurderet ud fra en standard beregning af recall/precision uden direkte bruger-evaluering. Man kan stille sig spørgsmålet, om resultaterne er sammenlignelige med en "real life" situation: Svarer en rangordning af dokumenter baseret på statistiske forekomster af enkelttermer til en brugers rangordning ud fra vedkommendes "state of uncertainty", som førte til en query?

Er der forskel på author og indexer-aboutness, både med hensyn til repræsentation og konsistens i terminologi? Vil der være forskel på resultatet af beregningerne, hvis de bliver foretaget ud fra et forfatterformuleret abstract eller en titel i modsætning til et indeksørformuleret abstract og kontrollerede emneord?

For at afprøve teknikken har vi forsøgt at simulere en vector-space beregning på to spørgsmål indenfor henholdsvis det tekniske område med en søgning i INSPEC og indenfor patentområdet med en søgning i

WPI/WPIL. Undersøgelsens resultat skal tages med forbehold, og skal kun ses som en illustration af, hvordan en sådan undersøgelse kan gennemføres.

SIRE-forsøget (Salton & McGill, 1983) blev brugt som model, idet der først blev gennemført en almindelig søgning med Boolske operatorer. En reel vector-space model ville rangordne samtlige dokumenter, der indeholdt blot én af termerne i queryen, men her har vi som i SIRE valgt at benytte resultatet fra den Boolske søgning som udgangspunkt for rangordningen. Søgeresultaterne blev opstillet i den sædvanlige form med de nyeste dokumenter øverst på listen og fremlagt for de eksperter indenfor de pågældende fagområder, som oprindelig havde stillet forespørgslerne og eksperterne blev nu bedt om at rangordne referencerne efter relevans.

Vi beregnede dernæst vector-space værdier ud fra flere forskellige repræsentationer, idet vi dels for patenternes vedkommende testede forskellene mellem originaltitler og -abstracts og de indeksørformulerede tilsvarende repræsentationer. I INSPEC-posterne beregnede vi rangorden ud fra henholdsvis basic index (titel, abstracts, termer), titel + termer samt titel alene. Endvidere beregnede vi de simplere tf.idf vægte af alle poster både med og uden dokumentlængde.

Den rangorden af referencerne, som fremkom efter disse forskellige beregninger blev herefter sammenlignet med eksperternes rangordning af de samme poster.

### 5.1 Forsøgsresultater

De statistiske metoder ser ud til at kunne trække relevante dokumenter højt op på ranglisten. Der har dog i alle vore forsøg vist sig at være relevante dokumenter i bunden af listen også. Hvis man vil have det hele med må man altså hele listen igennem.

Der har ikke vist sig nogen væsentlig forskel på de tre beregningsmetoder, d.v.s. vector-space, tf.idf med dokumentlængde og tf.idf uden dokumentlængde. Der kan ikke på dette grundlag argumenteres for, at den ene metode er bedre end de andre. Hvis man anlægger en praktisk betragtning, kunne man foretrække tf.idf uden dokumentlængde, da den er den simpleste at beregne.

En sammenligning af forsøgene med termer fra de forskellige repræsentationer og forskellige dele af dokumenterne viser ikke de store afvigelser. En vægtning ud fra titel-forekomst alene er derfor også mulig, men man skal være opmærksom på, at vi på det tekniske område befinder os inden for fagområder, hvor der er tradition for gøre titlerne repræsentative for indholdet.

## 6. Konklusioner

Traditionel repræsentation ved enkelttermer (NLR) er repræsentation af mening uden kontekst og uden konsistens: Det samme objekt kan repræsenteres ved forskellige termer, og den samme term kan have forskellige betydninger. Ved IR er søgeren derfor i høj grad overladt til egen intuition, fantasi og overblik over variationer i (faglig) terminologi.

Ved repræsentation med kontrollerede emneord sker der en styring, således at det tilstræbes at det samme objekt repræsenteres ved én term eller frase, og at den samme term har en eller et antal kendte betydninger. Kontrollerede emneord indebærer sandsynlighed for begrænset konsistens som følge af forskelle i opfattelse af mening indeksører imellem og hos den samme indeksør over tid. Ved IR styres søgeren af en etableret terminologi, men den forudsættes behersket af søgeren.

Ved ingen af de traditionelle repræsentationsformer er det i IR situationen muligt at rangordne dokumenter efter relevanskriterier.

M.h.t. IR teknikker konkluderer Belkin og Croft (1987), at evalueringer har vist, at partial match teknikker er signifikant bedre end exact match teknikker (idet der med signifikans menes mindst 10% forbedring i precision og recall).

Generelt for alle de omtalte IR teknikker er, at relevans i forhold til brugerens informationsbehov baserer sig på en match af ord, en match mellem ord formuleret af brugeren (i naturligt sprog eller udtrykt ved en formaliseret logisk formulering) og ord, som repræsenterer teksternes indhold. Resultatet af denne match kan være, at alle fundne tekster er (lige) relevante (exact match) eller at brugeren præsenteres for en system-baseret rangordning af teksterne, hvor rangordningen udtrykker deres relevans (partial match) i forhold til query'en, og hvor relevansen er et estimat, en beregning på grundlag af termfrekvens.

Et fundamentalt spørgsmål er imidlertid om man uden videre kan sætte lighed mellem relevans og termfrekvens: Hvis man nu er interesseret i vector-space modellen, så vil ved partial match- søgninger (i titel, abstract eller den fulde tekst) denne artikel (for så vidt den er indekseret) blive stadig mere relevant, jo hyppigere termen vector-space er gentaget (i titel, abstract eller den fulde tekst) - hvorimod den artikel, der indledes således: 'Vector-space modellen (i det følgende omtalt som modellen) ...' sandsynligt kan risikere en lavere rangordning.

Vores forsøg med vægtningsmodeller baseret på de velkendte metoder tf.idf og vector-space viser da også, at modellerne nok kan bruges til at ordne en liste af databaseudskrifter efter relevans, og at en sådan liste vil præsentere nogle af de mest relevante dokumenter blandt de øverste på listen, men der vil også være mange relevante i bunden af listen. Man skal nok især være opmærksom på dette ved søgninger med lav precision.

Det ser ikke ud til at der er forskelle mellem tf.idf med dokumentlængde, tf.idf uden dokumentlængde og vector-space. Vi er enige med Sparck Jones (1973) i hendes vurdering: at dokumentlængden ikke ser ud til at have betydning, og at den simple tf.idf formel er lige så god som den mere komplicerede vector-space-formel, som benyttes af bl.a. Salton.

Der er altså trods alt ikke anledning til at afvise vægtningsmetoder som en mulighed for bedre præsentation af søgeudskrifter. En simpel løsning med vægtning ud fra tf.idf uden dokumentlængde kunne nemt programmeres, så man i lighed med SIRE systemet efter en Boolsk søgning kunne beregne vægten af de enkelte dokumenter og sortere dem i rangorden derefter. Det kunne endda ske udelukkende på grundlag af titler eller titler + termer, som hyppigt benyttes i forvejen. Brugeren ville så få præsenteret en liste, hvor der var større sandsynlighed for at finde relevante dokumenter i toppen af listen end ved en almindelig udskrift.

Der er naturligvis mange spørgsmål, som stadig mangler at blive besvaret, bl.a. spørgsmålet om hvilke typer databaser, modellerne er velegnede til, men de foreløbige resultater giver belæg for at arbejde videre med termvægtning til brug for rangordning af søgeudskrifter.

Iøvrigt vil den udvikling, der i dag sker m.h.t. udvikling af adaptive og intelligente front-ends, i stigende grad indeholde de muligheder, der ligger i partial match teknikkerne, men disse nye design vil fortsat i vid udstrækning operere op mod de store IR systemer på exact match betingelser. I de store IR-systemer (store, internationale databaser hos kommercielle værter) vil af bl.a. økonomiske årsager de kendte teknikker fortsat være fremherskende, dog med muligheden for udbygning med nye faciliteter som vi har set det med ESA-IRS' Zoom, Datastars Plurals og Dialogs Ranking.

## 7. Litteratur


Belkin, N.J. (1978). Information concepts for information science. Journal of Documentation, vol. 34, no. 1, s.55-85.

Belkin, N.J. & Croft, W.B. (1987). Retrieval techniques. Annual Review of Information Science and Technology, vol. 22, s.109-145.

Croft, W.B. (1987). Approaches to intelligent information retrieval. Information Processing and Management, vol. 23, no. 4, s.249-254.

Croft, W.B. & Harper, D.J.(1979). Using probabilistic models of document retrieval without relevance information. Journal of Documumentation, vol.35, s.285-295. Reprinted in: P.Willett (ed), Document Retrieval Systems. London: Taylor Graham,1988, s.161-171. (The Foundations of Information Science, vol.3).

Ingwersen, P. (1984). A cognitive view of three selected online search facilities. Online Review, vol.8, no.5, s.465-492.

Ingwersen, P. (1991). Intermediary Functions in Information Retrieval Interaction. København: Samfundslitteratur. (Ph.D. Serie 91/4). 169 s.

Keen, E.M. (1991). The use of term position devices in ranked output experiments. Journal of Documumentation, vol.47, no.1, s.1-22.

Lennon, M. et.al.(1981). An evaluation of some conflation algorithms for information retrieval. Journ. of Information Science, vol.3, s.177-183. Reprinted in: P. Willett (ed), Document Retrieval Systems. London:

Taylor Graham,1988, s.99-105. (The Foundations of Information Science, vol.3).

Nielsen, D. (1993). [Anmeldelse af] Informationsvidenskabelige grundbegreber: bibliotekss- og informationsvidenskab/ Birger Hjørland. DF-Revy, vol. 16, no. 3, s.84-86.

Salton, G. (1986). Another look at automatic text-retrieval systems. Communication of the ACM, vol.29, no.7, s.648-656.

Salton, G. & Buckley, C.(1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, vol.24, no.5, s.513-523.

Salton, G., Fox, E.A. & Voorhees, E.(1985). Advanced Feedback Methods in Information Retrieval. Journal of the American Society for Information Science, vol.36, no.3, s.200-210.

Salton, G., Fox, E. & Wu, H.(1983). Extended Boolean Information Retrieval. Communication of the ACM, vol.26, no.11, s.1022-1036.

Salton, G. & McGill, M.J. (1983). The SMART and SIRE Experimental Retrieval Systems. I: Salton, G. & McGill, M.J.: Introduction to modern information Retrieval. New York: McGraw-Hill, s.118-156. Reprinted in: P. Willett (ed), Document Retrieval Systems. London: Taylor Graham, 1988, s.192-229. (The Foundations of Information Science, vol.3).

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

Winograd, T. & Flores, F. (1986/1987). Understanding Computers and Cognition. A New Foundation for Design. Reading, Mass: Addison-Wesley. 207 s.

Wormell, I. (1984). Indexes and text analysis in automated IR-systems. Copenhagen: The Royal School of Librarianship. 38 s.