

MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP
VID BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN
2006:124
ISSN 1404-0891

En thesaurus som ledsagare -
En jämförande studie av tre sökstrategiers inverkan på återvinningsresultatet i en
bibliografisk databas.

LENA HAGBERG
JOHANNA MÜNTZING



HÖGSKOLAN I BORÅS

© **Författarna**

Mångfaldigande och spridande av innehållet i denna uppsats
– helt eller delvis – är förbjudet utan medgivande.

Svensk titel: En tesaurus som ledsagare - En jämförande studie av tre sökstrategiers inverkan på återvinningsresultatet i en bibliografisk databas.

Engelsk titel: The thesaurus as a companion - A comparative study of three search strategies and their influence on information retrieval results in a bibliographic database.

Författare: Lena Hagberg och Johanna Müntzing

Kollegium: 2 och 4

Färdigställt: 2006

Handledare: Nasrine Olson

Abstract:

This Master's thesis is a comparative study of information retrieval results between three distinct search strategies in simulated automatic query expansion in a bibliographic database. Our purpose is to investigate which of the search strategies score the most effective precision and to what extent the same relevant documents are retrieved (overlapped). A thesaurus attached to the database is used to select appropriate descriptors for the baseline query formulations which subsequently are expanded with hierarchical relations. The search strategies are s1: A baseline query with two or three descriptors, s2: The baseline descriptors combined with at least one Narrower Term, s3: The baseline descriptors combined with Narrower Term and at least one Broader Term. A Document Cutoff Value of 15 is used and only the 15 highest ranked documents are judged by relevancy. The measurements used are precision for effectiveness and Jaccard's index for overlap. In terms of precision, results reveal that s1 scores the highest value (average 84,8 %) with s2 and s3 in decreasing order (average 81,94 % and 61,41 % respectively). The overlap varies greatly depending on topic and the average is between s1 and s2 78,81 %, between s2 and s3 58,48 % and between s3 and s1 40,41 %. In short, average precision decreases as well as average overlap. The use of thesaurus in the applied strategy of automatic query expansion is not recommended in this specific database, if the aim is to increase precision. However, in single searches with the structure like s1 the thesaurus can be of assistance in the selection of specific search terms.

Nyckelord: Information Retrieval, Query Expansion, tesaurus, bibliografisk databas, kontrollerad vokabulär, informationsåtervinning.

Innehållsförteckning

1 INLEDNING	1
2 DISPOSITION	1
3 PROBLEMFÖRMULERING	2
3.1 SYFTE.....	2
3.1.1 FRÅGESTÄLLNINGAR.....	3
3.1.2 AVGRÄNSNINGAR.....	3
4 TIDIGARE FORSKNING	4
5 DATABASER OCH DATABASVÄRDAR	9
5.1 SOCIOLOGICAL ABSTRACTS.....	10
5.1.1 DATABASEN I SITT SAMMANHANG: CHALMERS BIBLIOTEK.....	11
6 IR	13
6.1 IR MODELLER.....	13
6.1.1 BOOLESKA MODELLEN.....	14
6.1.2 VEKTORMODELLEN.....	15
6.2 RELEVANSBEDÖMNING.....	16
6.3 PRECISION OCH RECALL.....	17
7 QUERY EXPANSION	18
7.1 AUTOMATISK QUERY EXPANSION.....	19
7.2 INTERAKTIV QUERY EXPANSION.....	19
7.3 MANUELL QUERY EXPANSION.....	20
8 KONTROLLERAD VOKABULÄR	22
9 METOD	24
9.1 FORMULERING AV TOPIC.....	24
9.2 SÖKSTRATEGI.....	25
9.3 STRATEGI FÖR RELEVANSBEDÖMNING.....	25
9.4 BERÄKNING AV PRECISION OCH ÖVERLAPPNING.....	26
9.4.1 GPRD.....	26
9.4.2 JACCARDS INDEX.....	27
9.5 SAMMANSTÄLLNING AV RESULTAT.....	28
9.6 METODOLOGISKA PROBLEM.....	28

10 RESULTAT OCH ANALYS	29
10.1 ANALYS AV SÖKNINGEN I DATABASEN.....	29
10.2 ANALYS AV GPRD.....	29
10.3 ANALYS AV ÖVERLAPPNINGEN.....	31
11 SLUTSATSER OCH SLUTDISKUSSION	34
11.1 FAKTORER SOM KAN HA PÅVERKAT RESULTATET.....	35
11.2 FÖRSLAG TILL VIDARE STUDIER.....	35
12 SAMMANFATTNING	36

1 Inledning

I dagens informationssamhälle pågår en diskussion om vilka krav som kan ställas på medborgare då det gäller informationskompetens. Detta innefattar förmågan att söka, värdera och ta till sig information som ofta återfinns via hemsidor, sökmotorer och databaser tillgängliggjorda på Internet. Som blivande bibliotekarier och informationsförmedlare ser vi databaser som några av de viktigaste redskapen för att hitta information. För att kunna utnyttja dessa fullt ut är det i många fall en fördel att ha en viss förståelse för hur återvinningen av informationen går till. Vid val av ämne till magisteruppsatsen tog vi därför tillvara på tillfället att fördjupa våra kunskaper om informationsåtervinning som inom biblioteks- och informationsvetenskap ingår i forskningsområdet Information Retrieval (IR). I föreliggande uppsats har vi undersökt vilken inverkan tre olika sökstrategier kan ha på återvinningsresultatet vid sökning i en bibliografisk databas. Undersökningen utfördes efter ett verkligt behov, då Chalmers tekniska högskolas bibliotek sökte studenter som var intresserade av att undersöka ”deras” databas Sociological Abstracts. Denna databas förknippas inte med de traditionella ämnena som studeras vid Chalmers, men bibliotekarierna som arbetar där har statistik som visar att den används. Detta fann vi intressant och för att knyta an till Chalmers valde vi att utföra sökningar med ämnesmässigt teknisk karaktär, kombinerade med de ämnen som databasen främst innehåller, liksom sociologi och beteendevetenskap.

2 Disposition av uppsatsen

I kapitel tre beskrivs problemformuleringen och uppsatsens ämne placeras i sitt sammanhang, samt att syfte, frågeställningar och avgränsningar redovisas. Efter detta ges i kapitel fyra en översikt av den forskning vi finner relaterad till valt ämne. I kapitel fem beskrivs olika sorters databaser, inklusive Sociological Abstracts. Därefter följer kapitel sex till åtta som innefattar uppsatsens teoretiska tyngdpunkt. Kapitel sex ger en genomgång av två vanliga IR-modeller samt behandlar de två vanliga måtten precision och recall, som kan användas för beräkning av återvinningseffektiviteten. I detta kapitel tas även problematiken kring relevansbedömning upp. Kapitel sju fokuserar på de tre vanligaste formerna av query expansion, samt hur en sökstrategi bör byggas upp. Eftersom kontrollerad vokabulär är ett viktigt redskap i studien ges en övergripande bild av framförallt en thesaurus i kapitel åtta. Därefter följer i kapitel nio en kronologisk genomgång av vårt metodologiska tillvägagångssätt. I detta kapitel exemplifieras hur måtten Jaccards index och GPRD används i studien. Jaccards index används i denna studie för att mäta likheten mellan sökningars återvinningsresultat och GPRD är ett mått för att mäta den genomsnittliga precisionen vid observerade relevanta dokument.¹ Kapitel tio presenterar studiens resultat varvat med analys av insamlad data. I kapitel elva redogörs för vilka slutsatser vi kommit fram till samt förslag till framtida studier och slutligen följer en sammanfattning i kapitel tolv.

¹GPRD beskrivs mer ingående i kapitel 9.4.1 och Jaccards index i kapitel 9.4.2.

3 Problemformulering

Uppsatsens ämne problematiseras i detta kapitel och syftet med uppsatsen klargörs. Frågeställningar som formulerats för att uppnå syftet redovisas samt att de avgränsningar som anses nödvändiga att ta upp avslutar kapitlet.

Målet med informationssökning är att få bästa möjliga återvinningsresultat och möjligheterna för detta ökar om det finns en väl genomtänkt sökstrategi bakom. Problematiken ligger i att informationssystem återvinner information på olika sätt samt att olika sökstrategier ger olika återvinningsresultat beroende på använd metod. Kunskap om hur sökstrategier fungerar i olika system, samt vilka alternativa sökstrategier som kan användas vid sökning, är därför till hjälp för användaren då denna söker i systemet. Denna problematik har flertalet forskare inom IR försökt att utröna, bland annat genom experimentella studier där sökstrategier testas och jämförs med varandra i olika typer av system. Denna uppsats ansluter sig till nämnda forskning genom att tre sökstrategiers prestanda vid sökning i den bibliografiska databasen Sociological Abstracts jämförs med varandra. En vanlig strategi för valet av termer till sökningarna kan vara användningen av en tesaur och så är även fallet i denna uppsats eftersom rankningen av de återgivna träffarna i den aktuella databasen sker efter deskriptorer².

I denna studie har vi valt en sökstrategi som simulerar automatisk query expansion, det vill säga automatisk expansion³ av sökfrågor. Sökstrategierna som vi kallar s1, s2 och s3 har konstruerats med hierarkiska relationer. S1 består av två till tre deskriptorer, som därefter expanderas i s2 med Narrower Terms (som här efter förkortas NT), samt s3 där deskriptorerna från s1 expanderas med både NT och Broader Terms (som här efter förkortas med BT). Resultat från den tidigare forskningen⁴ indikerar att användning av NT ger hög precision vid expansion av sökfrågor om man jämför med sökningarna då expansionstermerna består av andra relationer, likt exempelvis relaterade termer (som här efter förkortas RT). Expansion med BT kan ge högre recall i förhållande till andra relationer vilket indikerar att BT ger en omfattande återvinning. Greenberg⁵ uppmärksammar att det finns ett behov att ytterligare undersöka hur ekvivalenta, hierarkiska och associativa relationer påverkar återvinningseffektiviteten för att kunna förbättra automatisk query expansion. Detta behov i kombination med ovan nämnda resultat gör att vi finner det relevant och intressant att utföra en undersökning av sökstrategier innehållande hierarkiska relationer i form av endast BT och NT.

3.1 Syfte

Syftet med denna uppsats är att undersöka hur olika sökstrategier presterar när det gäller återvinningseffektivitet i Sociological Abstracts. För att göra detta jämförs tre olika sökstrategier (s1, s2 och s3) och dess prestationer relativt varandra i fråga om återvinning av relevanta dokument.

²En tesaur innehåller deskriptorer, det vill säga kontrollerad vokabulär som är standardiserade termer vilka ofta rekommenderas till användning vid indexerings och sökning. En tesaur visar relationerna/släktskapen mellan deskriptorerna som används vid sökning och/eller indexerings för att identifiera dokument i en databas. (Se kap. 8.) Okontrollerad vokabulär är motsatsen till kontrollerad vokabulär och kan exemplifieras med naturligt språk.

³För mer ingående beskrivning av automatisk expansion se kapitel 7.1, samt metoden för vår simulering av detta i kapitel 9.2.

⁴Om tidigare forskning se kap. 4.

⁵Greenberg, Jane 2001. Automatic query expansion via lexical-semantic relationships, *Journal of the American Society for Information Science and technology*, vol. 52, no. 5, s. 402-415.

Vi vill även undersöka i vilken utsträckning samma relevanta dokument återvinns med de tre strategierna och se vilka eventuella mönster som kan utläsas.

3.1.1 Frågeställningar

1. Vilken av sökstrategierna s1, s2 och s3 har högst genomsnittlig precision vid observerade relevanta dokument (GPRD)?
2. I vilken utsträckning återvinns samma relevanta dokument, med hjälp av de olika sökstrategierna?

3.1.2 Avgränsningar

Vid studier av databaser finns det inom IR-forskningen olika angripssätt. Två av de mest förekommande är system- respektive användarorienterade utvärderingar av databaser.⁶ När det gäller den användarorienterade utvärderingen undersöks exempelvis hur väl informationssystemet möter användarens krav och preferenser (exempelvis gällande gränssnitt med mera). I den systemorienterade utvärderingen läggs fokus på själva systemet, exempelvis dess prestanda.

I denna undersökning mäts endast återvinningseffektiviteten av de tre sökstrategierna och därmed exkluderas hur användaren uppfattar databasen utifrån gränssnitt med mera. Vi utgår endast från systemet och dess premisser för sökning.

⁶Van Rijsbergen, C. J. *Information retrieval*, <http://www.dcs.gla.ac.uk/Keith/Preface.html> [2006-04-02]. Se kap. 1 Introduktion.

4 Tidigare forskning

Undersökningar av återvinningsresultatet, och sökstrategiers inverkan på detta, utförs i databaser i antingen en testmiljö eller en operationell miljö.⁷ Relevansbedömning och sökstrategier är viktiga faktorer i dessa studier och sökstrategierna formuleras efter topicet⁸ samt består ofta av en initial sökning, vilken därefter expanderas på lite varierande sätt. Vi har inte funnit någon studie med samma sökstrategi som används i denna uppsats, men i detta kapitel redogörs det för några studier som är av intresse. Gemensamt för dessa är att de alla har en tesaurus eller en annan form av kontrollerad vokabulär som termkälla vid sökningarna.

Användning av testtesaurusar i två fulltextdatabaser

Kristensen och Järvelin redogör i artikeln *The Effectiveness of a Searching Thesaurus in Free-Text Searching in a Full-Text Database*⁹ för en undersökning om en tesaurus inverkan på sökresultatet i en finsk fulltextdatabas.

En så kallad ”searching thesaurus” konstruerades med NT, BT och relaterade termer (som härefter förkortas RT) samt synonymer (som härefter förkortas SYN). Denna tesaurus, som enligt artikelförfattarna inte är en tesaurus i vanlig benämning, konstruerades för att förbättra återvinningsresultatet genom att förse användarna med alternativa termer, exempelvis SYN till sökningarna. Syftet var att möjliggöra en systematisk breddning eller avgränsning av termer och få del av de fördelar kontrollerad vokabulär medför eftersom sökningarna utfördes i en fulltextdatabas med okontrollerad vokabulär. Författarna utvärderade tesaurusens inverkan på precisions- och relativ recallnivåerna¹⁰ hos framtagna dokument och jämförde återvinningsresultatet mellan en initial sökning, en sökning då SYN lagts till samt en sökning med både SYN och RT.

Evalueringen utfördes i en operationell databas tillhörande BASIS information retrieval and management system med cirka 34 000 tidningsartiklar inom området ekonomi. Sökfrågorna formulerades av fem journalister och med dessa frågor som grund formulerade författarna de initiala sökningarna. De använde sig av de ord som journalisterna uttryckt i sina ursprungliga sökfrågor och 30 sökfrågor formulerades. I 26 fall återvanns minst en relevant artikel, vilket gjorde att fyra ströks ur studien. Vidare fanns det endast 18 sökfrågor med termer för de båda expansionsstrategierna.

För att förbättra återvinningsresultatet studerades två expansionsformer och man behöll den logiska strukturen från den initiala sökning. Sökstrategin var enligt följande: Först utvidgades den initiala sökning med hjälp av SYN och sedan med både SYN och RT. Journalisterna bedömde sedan de återvunna dokumenten som antingen relevanta, kanske relevanta eller inte relevanta. Endast de dokument som tillhörde den förstnämnda gruppen ansågs relevanta och togs med i beräkningarna av precision och relativ recall, vilket beräknades för varje sökstrategi. Författarna satte den relativa recallen till 100 % (den maximala relativa recallen för alla sökningarna) vid sökning med både SYN och RT och man jämförde de två andra sökstrategiernas relativa recallnivåer med denna.

⁷En operationell miljö syftar till den verkliga miljön, i vilket ett system används. Motsatsen är en testmiljö som är konstruerat specifikt för utförandet av experimentella undersökningar.

⁸I denna uppsats används den engelska benämningen *topic* för att beskriva ett informationsbehov, konkret uttryckt med naturligt språk. Ett topic konkretiseras i en sökfrågeformulering då det delas in i facetter efter ämnesinnehåll.

⁹Kristensen, Jaana & Järvelin, Kalervo 1990. The effectiveness of a searching thesaurus in free-text searching in a full-text database, *International Classification*, vol. 17, nr. 2, s. 77-84.

¹⁰Relativ recall kan användas då det totala antalet relevanta dokument för ett visst topic inte är känt i en dokumentinsamling. Målet med detta är att få en uppfattning av hur många relevanta dokument som kan förväntas finnas för det givna topicet. (Om recall se kap. 6.3.)

Resultatet av undersökningen visade att för de 26 sökfrågorna ökade den genomsnittliga relativa recallen från 45,00 % till 82,00 % med användningen av SYN och genom ytterligare expansion med RT från 82,00 % till 100,00 %. Den genomsnittliga precisionen minskade däremot från 50,60 % vid den initiala sökningen till 40,60 % vid användningen av SYN, samt till 33,10 % vid användningen av SYN och RT. Den relativa recallen förbättrades sålunda medan precisionen minskade något.

Sökfrågestruktur	Precision	Relativ recall
Initial sökning	50,60 %	45,00 %
SYN	40,60 %	82,00 %
SYN och RT	33,10 %	100,00 %

Tabell 1: Värdena för precision och relativ recall för de 26 sökfrågorna.

De 18 sökfrågor som hade termer för de båda expansionsstrategierna visade liknande resultat. Här ökade den relativa recallen från 44,40 % vid den initiala sökningen till 76,80 % vid sökningen med SYN, samt till 100,00 % vid den sista expansionen. Vidare såg trenden likadan ut vid precisionen som minskade från 48,50 % vid den initiala sökningen till 40,20 % vid sökningen med SYN och vidare till 35,00 % vid sökningen med både SYN och RT. Ungefär lika många nya dokument återvanns vid varje strategi, men antalet relevanta dokument minskades från den initiala sökningen till sökningen med SYN och RT.

Sökfrågestruktur	Precision	Relativ Recall
Initial sökning	48,50 %	44,40 %
SYN	40,20 %	76,80 %
SYN och RT	35,00 %	100,00 %

Tabell 2: Värdena för precision och relativ recall för de 18 sökfrågorna med termer för de båda expansionsstrategierna.

Den tesaurus som skapades för undersökningen förbättrade sökningarnas relativa recallnivåer genom att användaren fick ytterligare termer, SYN och andra RT för de ursprungliga söktermerna. Användningen av tesaurusen var enligt författaren en bra strategi när det gällde fritextsökning¹¹ i en fulltextdatabas och andra intressanta områden för efterföljande forskning omnämndes. Exempel på detta var hierarkisk expansion för att uppnå en så bred täckning av termer som möjligt vid sökning.

En uppföljning av den ovan nämnda undersökningen publicerades av Kristensen i artikeln *Expanding End-Users' s Query Statements for Free Text Searching with a Search-Aid Thesaurus*.¹² Precis som i föregående fall skapades här en tesaurus i studiesyfte och studien utfördes i en operationell fulltextdatabas med 227 000 finska tidningsartiklar, också den tillhörande BASIS. Tesaurusen, denna gång kallad "search aid tesaurus"¹³, konstruerades med likvärdiga, associativa samt hierarkiska relationer och ett antal sökningar kördes med hjälp av denna mot databasen. Olika relationstyper från tesaurusen kombinerades en efter en med termerna från den initiala sökningen. Den första expansionen utfördes med SYN och följdes sedan av en expansion med RT samt därefter en sökning med NT. Till sist utfördes en sökning med alla de fem strategierna inkluderade.

¹¹Fritextsökning innebär att söktermerna hämtas fritt från tillgängliga källor (exempelvis från eget ordförråd eller ordlista).

¹²Kristensen, Jaana 1993. Expanding end-user's query statements for free text searching with a search-aid thesaurus, *Information Processing & Management*, vol. 29, no. 6, s. 733-744.

¹³Vi har tolkat det som att denna tesaurus är av samma sort som den tesaurus som skapades i den föregående undersökningen.

Genomsnittligt visade sig den initiala sökningen ge en relativ recallnivå på 50,30 %. Expansionen med SYN 67,20 %, RT 67,30 %, NT 60,00 % och slutligen den unionsökningen som gav 100,00%. Precisionsnivåerna visade relativt små skillnader men sökningen med NT gav ett bättre resultat i förhållande till flera av de andra. Den initiala sökningen gav bäst precisionvärdet 64,40 %, följd av sökningen med NT 59,10 %, sökningen med RT 57,40 %, sökningen med SYN 55,20 % samt till sist unionsökningen 49,50 %.

Sökfrågestruktur	Precision	Relativ recall
Initial sökning	64,40 %	50,30 %
SYN	55,20 %	67,20 %
RT	57,40 %	67,30 %
NT	59,10 %	60,00 %
Unionsökning	49,50 %	100,00 %

Tabell 3: Värdena för precision och relativ recall i undersökningen.

Sammanfattningsvis visade undersökningen att kombinationen av den initiala sökningen med SYN respektive RT återvann flest relevanta artiklar, men skillnaderna mellan de tre expansionsformerna (SYN, RT och NT) var enligt författaren, inte statistiskt säkerställda. På det hela taget stämde undersökningen överens med den tidigare undersökningen (av Kristensen och Järvelin) och tesaurusen förbättrade de relativa recallnivåerna betydelsefullt medan precisionen minskade och skillnaderna var små men ändå betydelsefulla mellan de olika sökstrategierna.

En testtesaurus i INQUERY retrieval system

*The impact of query structure and query expansion on retrieval.*¹⁴

Ytterligare en studie av Järvelin och Kekäläinen¹⁵ analyserade hur återvinningsresultatet skiljde sig åt då sökfrågor formulerades på olika sätt i ett probabilistiskt system. Följande variablers inverkan på återvinningsresultatet testades: Antal sökkoncept,¹⁶ antal söksträngar som representerade konceptet, query expansion med olika semantiska relationer och slutligen hur sökfrågestruktur och termers vikter påverkar sökresultatets ranking.

Testerna utfördes i INQUERY retrieval system (version 3.1) i en databas med finska tidningsartiklar och en testtesaurus skapades för studien. 30 topic användes i studien och 35 sökfrågor skapades för var och ett av dessa. Sökfrågorna konstruerades med totalt åtta olika sökfrågestrukturer. Tre av sökfrågestrukturerna var så kallade svaga strukturer vilket innebar att sökfrågorna endast hade en operator som band samman termerna och att huvudtermerna inte hade någon relation med varandra. Resterande fem sökfrågestrukturer hade en stark struktur vilket innebar att flera operatörer användes, samt att det fanns olika relationer mellan termerna och strukturen baserades på facetter¹⁷ till huvudtermerna. Själva sökningarna utfördes på fem olika sätt: Inledningsvis en initial sökning, som därefter expanderades med först SYN, sedan NT och därefter RT. Slutligen utfördes en unionsökning med alla termer som använts vid föregående expansioner.

¹⁴Kekäläinen, Jaana & Järvelin, Kalervo 1998. The impact of query structure and query expansion on retrieval performance. *SIGIR ninety-eight: proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 24-28, s. 130-137.

¹⁵Tidigare Kristensen.

¹⁶Ett topic består av olika koncept vilka representerar det huvudsakliga ämnesinnehållet. Koncepten representeras vid formuleringen av en sökfråga av de söktermer som väljs. (Se mer ingående i kap. 7.)

¹⁷En facett är en gruppering termer med inbördes relationer.

Studiens resultat visade att effekten av query expansion berodde på sökfrågans struktur. De expansioner som gav bäst precision var de med NT, SYN och associativa koncept. Den sökfrågestruktur med störst förbättring av återvinningen var unionsökningen medan de booleska sökningarna i kombination med svaga strukturer hade de sämsta effekterna på återvinningsresultaten. Ytterligare resultat som lades fram av artikelförfattarna var att en omfattande query expansion kräver en stark sökfrågestruktur, medan en enkel sökning klarar sig utan en stark struktur. Vidare fann de att tesaurusen inte var till någon större hjälp vid semantisk query expansion, den var mest behjälplig vid unionsökningen där alla termer var inkluderade. Detta resultat skiljer sig från deras tidigare studier. I Kristensen och Järvelin¹⁸ samt Kristensen¹⁹ framkom det som bekant att tesaurusen kan vara till hjälp vid expanderings. Värt att notera är att sökningarna i de olika studierna utfördes med hjälp av olika sorters tesaurusar.

ProQuest Controlled Vocabulary i ABI/Inform database²⁰

Greenberg publicerade i artikeln *Query Expansion via Lexical-Semantic Relationship*²¹ en studie där det undersöktes om deskriptorer i form av SYN, delvis SYN, NT, RT, och BT har en positiv inverkan som expansionstermer vid automatisk query expansion. Studien genomfördes i det operationella systemet the ABI/Inform database, tillgängligt via DIALOG. Testpersonerna bestod av användare av systemet som var studenter på Katz Graduate School of Business. Dessa blev ombudade att formulera ett topic på naturligt språk inom ämnesområdet affärer/ekonomi och sedan översätta det till en sökfråga enligt den booleska logiken. Därmed var både systemet och användarna samt deras topic hämtade från en verklig situation. Vad studenterna inte visste var forskarens avsikt att endast undersöka de sökfrågor som hade minst en term som matchade testtesaurusen, the *ProQuest® Controlled Vocabulary*. Greenberg hämtade manuellt termer till sökfrågorna ur tesaurusen, en process som kunde simulera automatisk query expansion. Av studenternas sökfrågor valdes 42 och en initial sökfråga, som bestod av de utvalda termerna från tesaurusen, expanderades i fyra steg: 1. SYN och partiell SYN, 2. NT, 3. BT och 4. RT.

Studenterna som formulerade de aktuella sökfrågorna fick även relevansbedöma de återgivna dokumenten utifrån hur de motsvarade informationsbehovets ämnesmässiga innehåll. Denna bedömning skedde på en tregradig skala från relevant, delvis relevant till inte relevant. Dokumenten rankades och Document Cutoff Value (DCV)²² var satt till 15.

Sökfrågornas återvinningseffektivitet mättes sedan med relativ recall, samt precision. Både relativ recall samt precision beräknades för den initiala sökningen och dessa värden jämfördes sedan med de expanderade sökstrategierna. Kort om resultatet är att den initiala sökningen gav den relativa recollen 26,60 % och precision 79,40 %. Den genomsnittliga ökningen för relativ recall, samt precision beräknades sedan utefter detta och resultaten var enligt följande: SYN: Precisionen 76,60 % och relativ recall 14,30%, NT: Relativ recall 20,70 % och precision 73,30 %, RT: Relativ recall 23,40 % och precision 54,40 %, BT: Relativ recall 23,30 % och precision 59,50 %.

¹⁸Kristensen & Järvelin 1990.

¹⁹Kristensen, 1993.

²⁰ABI/Inform database innehåller både fulltext och bibliografiska referenser. Det framgår inte av artikeln vilket eller vilka som använts i studien.

²¹Greenberg, 2001.

²²Anger antalet dokument i dokumentlistan som inkluderas i undersökningen.

Sökfrågestruktur	Precision	Relativ recall
Initial sökning	79,40 %	26,60 %
SYN	76,60 %	14,30 %
NT	73,30 %	20,70 %
RT	54,40 %	23,40 %
BT	59,50 %	23,30 %

Tabell 4: Värdena för precision och relativ recall i undersökningen.

Wordnet som termkälla i fulltextdatabas

Voorhees beskriver i *On Expanding Query Vectors with Lexically Related Words*²³ användningen av lexikala relationer och dess effektivitet vid expansion av sökfrågor, utförd i testkollektionen TREC. Till hjälp användes WordNet, ett manuellt konstruerat lexikalt system som utvecklats på Cognitive Science Laboratory vid Princeton universitetet. WordNet innehåller uppsättningar av synonymer, så kallade synset, vilka organiseras efter synonymernas lexikala relationer. Ett synset organiseras därefter i set av cirka tio hierarkier.

Fyra expansionsformer undersöktes i studien: 1. Expansion med endast SYN. 2. Expansion med SYN och en hierarkisk underordnad relation. 3. Expansion med både under- och överordnade hierarkiska relationer. 4. Expansion med SYN och direkt relaterade synset.

Studiens resultat visade att den expansionsteknik som använts inte gav några nämnvärda fördelar i TREC miljön. De fyra expansionsteknikerna genererade med andra ord inte någon förbättrad informationsåtervinning. En orsak kan enligt artikelförfattaren vara att de topic som togs från TREC kollektionen var detaljerat beskrivna och specifika termer förekom redan i de initiala sökfrågorna. I de flesta fallen tillkom nya termer i expansionerna, men de viktigaste termerna hade redan förekommit i den initiala sökfrågan. (En annan orsak kan vara att WordNet inte var lämplig för studien, men artikelförfattaren misstänker också samtidigt att ingen annan källa skulle passa.)

²³Voorhees, Ellen M, 1994. On expanding query vectors with lexically related words, *NIST Special publication 500-215: The Second Text REtrieval Conference (TREC 2)*. http://trec.nist.gov/pubs/trec2/t2_proceedings.html [2006-05-08].

5 Databaser och databasvärdar

Databaser började utvecklas under andra hälften av 1960-talet efter datateknologins genombrott.²⁴ Termen databas användes för första gången enligt Tamiko Matsumura på 1950-talet för att syfta till en samling information om amerikanska militärbaser. När sedan den numera etablerade databasen PubMed²⁵ och andra stora vetenskapliga tidskrifter, dagstidningar samt nationalencyklopedier började göra sökningar i sina databaser möjliga ökade användningen av termen ytterligare. Inledningsvis var marknaden för databaser relativt liten och eftersom det endast fanns ett fåtal kunde de hålla en viss standard med avseende på kvalitet. När Internet fick sitt genombrott, med allt var det innebar, ökade produktionen av databaser markant²⁶ vilket medförde att det inte längre bara var professionella informationsförsörjare som kunde skapa en databas och tillgängliggöra den via www, utan även privatpersoner och företag. Detta har gjort databasen till en av vår tids största hjälpmedel för informationsförmedling och termen används flitigt.

Det finns en rad definitioner av begreppet databas. Enligt ODLIS - Online Dictionary for Library and Information Science definieras en databas: "A large, regularly updated file of digitized information (bibliographic records, abstracts, full-text documents, directory entries, images, statistics, etc.) related to a specific subject or field, consisting of records of uniform format organized for ease and speed of search and retrieval and managed with the aid of database management system (DBMS) software."²⁷

När man talar om databaser tillgänggjorda via Internet och informationsåtervinningssystem online definierar man en databas som en samling förteckningar/dokument som är läsbara i maskinell form och som gjorts sökbara från såväl närliggande som avlägsna dataterminaler.²⁸

Precis som det finns varierande definitioner av begreppet databas finns det också varianter av databaser. Enligt Chowdhury bör man klassificera dessa efter datatyp samt i vissa fall efter ämnesinnehåll. Författaren framhåller främst två sorters databaser; de så kallade källdatabaserna (source databases) samt referensdatabaserna (reference databases), som kan delas in i underkategorier. Källdatabaserna innefattar bland annat fulltextdatabaser och numeriska databaser, vilka innehåller numerisk information i form av exempelvis statistik. En källdatabas skall förse användaren med den faktiska informationen medan referensdatabaserna hänvisar användaren till informationens källa, exempelvis en person eller ett dokument. Referensdatabaser inkluderar katalogdatabaser samt bibliografiska databaser²⁹ och exempel på bibliografisk information är abstrakt som beskriver dokumentets innehåll, författarnamn eller klassificeringsnummer.³⁰

Online databaser delas in i bibliografiska (referensdatabaser), numeriska, samt konceptuella databaser (de två sista tillhör källdatabaserna). Den konceptuella databasen består av ämnesspecifik information om exempelvis produkter.³¹ Sociological Abstract är en databas som innehåller referenser till dokument inom ämnesområdet sociologi och är tillgänglig online. Detta gör den till

²⁴Chowdhury, G.G. 1999, *Introduction to modern information retrieval*. s. 12f.

²⁵Dåvarande MEDLARS.

²⁶Matsumura, Tamiko 2004. *Venturing into a new area-database evaluation*, s. 1. www.dpc.or.jp/english/SA_2004.pdf [2006-05-19].

²⁷Reitz, Joan M. *ODLIS - Online Dictionary for Library and Information Science*, http://lu.com/odlis/odlis_d.cfm [2006-06-13].

²⁸Convey, John 1992. *Online Information Retrieval: An introductory manual to principles and practice*, s. 8.

²⁹Chowdhury 1999, s. 14f.

³⁰Convey 1992, s. 8.

³¹Chamis, Alice Yanosko 1991. *Vocabulary Control and Search Strategies in Online Searching*, s. 6ff.

en del av den förstnämnda gruppen, den bibliografiska online databasen *och* tillhör den tidigare nämnda referensdatabasen. Vi anser att Chowdhury's definition av en databas passar bra i vår studie: "[...] an organised collection of related sets of data that can be accessed by more than one user by simple means and can be searched to reveal those that touch upon a particular need."³² Detta eftersom databasen i vår undersökning är tillgänglig online och kan nås av många på ett enkelt sätt och är till för att tillgodose informationsbehov inom ett specifikt ämnesområde.

En databas kan tillgängliggöras av skaparen själv, eller genom en så kallad databasvärd³³ som samlar och distribuerar databaser till olika användare. Databasvärden kan sägas vara en slags intermediär mellan databasskaparen och den slutgiltiga användaren, eftersom värden bland annat bestämmer gränssnittet som användaren sedan rättar sig efter.

5.1 Sociological Abstracts

För att beskriva den databas som vi utför vår undersökning i kan följande fakta nämnas: Sociological Abstracts tillgängliggörs av databasvärden CSA.³⁴ Det ämnesområde den skall täcka är sociologi och relaterade ämnen likt beteendevetenskap. Databasen är bibliografisk och innehåller därmed referenser om var den sökta informationen kan hittas, samt abstract från dokument. Referenserna som är indexerade innefattar information om följande dokument: Huvudsakligen artiklar från tidskrifter men även referenser till avhandlingar och konferansskrifter, referenser till böcker, samt till speciella kapitel ur böcker. Dokument från sammanlagt 30 olika språk är indexerade, varav svenska är ett av dem. Om ett dokument är skrivit på ett annat språk än engelska återges titeln trots detta på engelska och i de flesta fall finns även ett abstract på engelska. (Önskas endast artiklar på engelska finns funktionen English Only).

Då en sökning körs i Sociological Abstracts söks det även i två andra databaser parallellt. Även dessa tillhör CSA och den ena innehåller information från 650 000 webbsidor som CSA valt ut för att även ämnesrelaterade sidor på *www* skall komma med vid sökning i Sociological Abstracts, vilka uppdateras månadsvis. Dokumenten från sökningarna bland webbsidorna presenteras inte i den ordinarie dokumentlistan, utan är tillgängliga via en länk vid dokumentlistan. Den andra databasen, Recent References, innehåller de senaste publicerade verken, som ännu inte har hunnit bli indexerade av CSA, och dagligen uppdateras indexeringen av citeringar från 110 tidskrifter. Träffarna i denna databas inkluderas i den ordinarie dokumentlistan och de uppmärksammas särskilt genom en länk till en separat dokumentlista. (Våra sökningar har endast träffar ur Sociological Abstracts.) Sökmöjligheterna som finns är Quick Search, Advanced Search och Command Search. Sökspråket som används är booleskt³⁵ och sökning kan därför ske med operatorerna AND, OR, NOT och andra vanligt förekommande sökfunktioner som påverkar sökningen likt NEAR. Vid användning av NEAR måste söktermerna förekomma i dokumentet med en radie om maximalt 10 ord, det vill säga högst tio ord får förekomma mellan söktermerna för att ett dokument skall återvinnas av systemet.

Databasen har två rankningsfunktioner, en som relevansrankar och en som rankar efter senast publicerade. Den sistnämnda är bra om man vill ha fram den absolut senast publicerade forskningen, men denna aspekt kan även täckas in av en funktion där användaren väljer vilket årtal

³²Chowdhury 1999, s. 14f.

³³Chamis 1991, s. 2.

³⁴Tillgång till databasen genom CSA finns för studenter och personal på Högskolan i Borås på <http://www.hb.se/bib/> och för studenter och personal på Chalmers via <http://www.lib.chalmers.se/> I vår utvärdering har vi använt tillgången via Chalmers eftersom den sker på uppdrag av dem. För att nå information om själva databasen: Gå in på Sociological Abstract och klicka på Pdf eller Word-doc, under Quick Links.

³⁵Mer om det booleska sökspråket i kapitel 6.1.1.

dokumentet skall vara publicerade. Relevansrankningen, som vi använder oss av i denna undersökning, tar hänsyn till hur relevanta dokumenten är med hänsyn till sökfrågan. Relevansen bestäms utifrån de åtta första deskriptorerna i deskriptorfältet och de dokument som innehåller sökfrågans termer i deskriptorfälten rankas högst. Även indexeringen av databasens innehåll har utförts med hjälp av dess thesaurus.³⁶

Ytterligare funktioner som är bra att känna till vid sökning i databasen är Cited Reference och Cross Reference. Cited Reference innebär att vid vissa artiklar finns i referenslistan en uppgift om hur många ytterligare dokument artikeln förekommer i Sociological Abstracts. Användaren kan klicka på denna uppgift och komma till de aktuella dokumenten där de förekommer som referens. Cross Reference innebär som termen indikerar att olika artikelförfattare refererar till varandra, och användaren kan klicka sig från den ena till den andra. Databasens gränssnitt innehåller tydliga instruktioner och är anpassat så att både en van och ovan användare skall kunna utnyttja den maximalt.

5.1.1 Databasen i sitt sammanhang: Chalmers bibliotek

Eftersom studien i denna uppsats utförs i en databas tillgänglig via Chalmers bibliotek vill vi i detta avsnitt uppmärksamma Sociological Abstracts roll och funktion i det sammanhang där den ingår.

Biblioteket är en självständig del av Chalmers tekniska högskola där forskning sker inom institutionerna:

- Arkitektur
- Bygg- och miljöteknik
- Data- och informationsteknik
- Energi och miljö
- Fundamental fysik
- Kemi- och bioteknik
- Matematiska vetenskaper
- Material- och tillverkningsteknik
- Mikroteknologi och nanovetenskap
- Produkt och produktionsutveckling
- Radio och rymdvetenskap
- Signaler och system
- Sjöfart och marinteknik
- Teknikens ekonomi och organisation
- Teknisk fysik
- Tillämpad mekanik

Biblioteket består av huvudbiblioteket samt arkitekturbiblioteket på campus Johanneberg och Lindholmenbiblioteket på campus Lindholmen och spelar en central roll för informationsförmedlingen till högskolans studenter, forskare, lärare samt övrig personal. Förutom för studenter och anställda inom institutionerna och högskolans övriga avdelningar är bibliotekets resurser även tillgängliga för informationssökande utifrån och de satsat på att vara tillgängliga både digitalt och fysiskt. Bland annat finns 143 databaser.³⁷

³⁶Denna information kommer från Lin Kwaun, CSA. Mer detaljerad information om algoritmen för rankningen lämnas inte ut eftersom detta anses vara affärshemlighet.

³⁷Detta enligt 2004 års siffror, Chalmers bibliotek, <http://www.lib.chalmers.se/bibl/Biblsiffror.xml> [2006-02-12].

Forskning och undervisning på Chalmers förknippas mest med tekniska ämnen och de ämnen som finns representerade i Sociological Abstracts tillhör inte de mest traditionella forsknings- och undervisningsområdena på högskolan. Vid samtal med bibliotekarierna på Chalmers framkommer det dock att det på senare år uppstått ett informationsbehov även för ämnen som inte enbart är tekniska. Detta då högskolan erbjuder nya utbildningar där teknik och mer samhällsvetenskapliga ämnen samspelar, likt industriell ledarskap med teknik och management. Det är i detta sammanhang som databasen kan tänkas ha en funktion för Chalmers.

6 IR

IR började växa fram som forskningsområde under slutet av 1950-talet och det primära var då att indexera text och göra textsamlingar sökbara efter användbar information. Några av de forskningsinriktningar som idag finns inom IR är: Modellering, dokumentklassifikation, systemarkitektur, användargränssnitt, informationsvisualisering, filtrering och språk.³⁸ IR betraktades länge som ett intressant område enbart för bibliotekarier och informationsvetare, men då www etablerades ökade möjligheten att tillgängliggöra informationssystem för ett större klientel, vilket ledde till ett ökat intresse för IR. Nya problem introducerades inom IR-forskningen, som exempelvis hur all information som publiceras på www skall göras lättillgänglig för användaren.³⁹

Inom de ovan nämnda forskningsinriktningarna studeras bland annat representation, lagring, organisering samt tillgång till information som finns i ett informationssystem.⁴⁰ Detta innefattar klassifikation, katalogisering och ämnesindexering.⁴¹ Att dessa är väl genomförda är en förutsättning för att en informationssökning i ett system effektivt skall kunna tillgängliggöra information för användaren.⁴² Därmed introduceras en av de problematiska faktorerna inom IR, nämligen användaren av informationssystemet och dennas informationsbehov. Användaren måste först översätta sitt informationsbehov till en sökfråga.⁴³ Översättningen kan vara problematisk, eftersom användaren kanske inte kan definiera och formulera sitt informationsbehov på rätt sätt för det aktuella informationssystemet. Målet för det ideala systemet är att återge information som är relevant för användarens informationsbehov och då denna inte kan uttrycka sitt behov korrekt riskerar han eller hon att få irrelevanta dokument eller missa relevanta dokument.

Relevansbedömningen, det vill säga den process då det avgörs vilka dokument som är relevanta, är ett annat centralt problem inom IR. Detta beslut beror på rankningsalgoritmen, som bestämmer i vilken ordning dokumenten skall återges efter en sökning.⁴⁴ För att få ett bra återvinningsresultat är det därmed viktigt att utveckla och använda effektiva algoritmer för att få bra träffar där relevanta dokument kommer högt upp i träfflistan.

Gällande relevansbedömning är det viktigt att skilja mellan systemets relevansbedömning och användarens relevansbedömning. Systemet relevansbedömer endast utifrån den information som lämnas av användaren, det vill säga sökfrågan. Denna matchas sedan mot dokumentens indextermer och är de relevanta återvinns de, vilket avgörs av den aktuella IR-modell som systemet är uppbyggt efter. Användarens relevansbedömning grundar sig på hur relevanta de återvunna dokumenten är för användarens informationsbehov.

6.1 IR-modeller

I detta avsnitt vill vi ge inblick i modeller för hur information återvinns i ett IR-system och visa vilka olika premisser som kan finnas då en sökfråga körs mot ett informationssystem. En IR-modell är en beskrivning för hur dokument skall återges i ett informationssystem, som har till huvuduppgift

³⁸Baeza-Yates & Ribeiro-Neto 1999, *Modern information retrieval*, s. 2.

³⁹Baeza-Yates & Ribeiro-Neto 1999, s. 2f.

⁴⁰Baeza-Yates & Ribeiro-Neto 1999, s. 1.

⁴¹Lancaster, Wilfrid. F. 1979. *Information retrieval systems: Characteristics, testing and evaluation*, s. 9.

⁴²Baeza-Yates & Ribeiro-Neto 1999, s.1.

⁴³Baeza-Yates & Ribeiro-Neto 1999, s. 1.

⁴⁴Baeza-Yates & Ribeiro-Neto 1999, s. 19.

att återge relevanta dokument. Hur dessa dokument skall återges beror på vilken IR-modell systemet bygger på. IR-modellen bygger i sin tur på de fundamentala premisser som algoritmen arbetar efter, det vill säga vilken rankning som används då dokumenten i det aktuella informationssystemet skall relevansbedömas.⁴⁵

Det finns ett antal olika modeller för informationsåtervinning och två vanligt förekommande har valts ut för att exemplifiera hur dessa kan återvinna relevanta dokument. Modellerna som beskrivs är den booleska modellen och vektormodellen,⁴⁶ vilka återvinner information på två olika sätt.

6.1.1 Booleska modellen

Den booleska modellen anses vara en enkel modell och användes bland annat av flera av de tidiga kommersiella återvinningsystemen.⁴⁷ Idén med modellen är att sökfrågan formuleras av användaren till ett logiskt uttryck innehållande operatorer som NOT, AND och OR. Relevansen för ett dokument bedöms sedan av det aktuella systemet genom att termerna i sökfrågan tilldelas sanningsvärdet sant eller falskt, beroende på om de förekommer i dokumenten eller inte. Är en sökfråga sann för ett dokument betraktas det som relevant och tvärtom för falskt.

En stor nackdel vid sökning i ett booleskt system är att användare kan finna det svårt att formulera frågor enligt den booleska logiken vilket kan leda till att användarens informationsbehov inte besvaras som förväntat, eftersom sökfrågorna inte formulerats korrekt. Detta då modellen som beskrivits ovan förutsätter exakt matchning, genom att termen antingen förekommer eller inte.⁴⁸

För att exemplifiera en sökning med den booleska modellen och de booleska operatorerna används följande topic: *Ett recept på en tårta med bär, men inte grädde.*

De booleska operatorerna fungerar enligt följande.⁴⁹ AND: Används för att tala om att varje facett, i det här fallet tårta och bär, måste vara närvarande i ett återvunnet dokument för att betraktas som relevant. Denna operator begränsar sökningen.

tårta AND bär

Dokumentet som återges skall nu innehålla information om tårta och bär.

Eftersom det finns många bärsorter kan användaren specificera vilka sorter som är intressanta för topicet och genom att använda operatören OR expandera sökningen. OR används, som i nedanstående exempel inom en facett, ofta när användaren är osäker på vilken synonymterm som är bäst att använda. För att ett dokument skall återvinnas måste något eller alla av orden bär, blåbär, hallon eller björnbär förekomma i kombination med tårta. Denna operator ökar ofta antalet återvunna dokument.

tårta AND (bär OR blåbär OR hallon OR björnbär)

Dokumentet som återges skall nu innehålla information om tårta tillsammans med någon av bärsorterna, eller alla tillsammans. Parentesen kring OR gör att termerna inom parentesen behandlas

⁴⁵Baeza-Yates & Riberio-Neto 1999, s. 23.

⁴⁶Baeza-Yates & Riberio-Neto 1999, s. 25-30.

⁴⁷Baeza-Yates & Riberio-Neto 1999, s. 25.

⁴⁸Baeza-Yates & Riberio-Neto 1999, s. 26f.

⁴⁹Large, Andrew 2001. *Information Seeking in the Online Age: Principles and Practices*, s. 148ff.

innan AND, som annars skulle ha behandlats först.

Ytterligare en expansion sker genom operatoren NOT. Denna operator talar om vilken eller vilka termer som *inte* får förekomma för att ett dokument skall vara relevant för sökfrågan. Eftersom operatoren utesluter vissa dokument finns en risk att även de dokument som utesluts kan innehålla relevant information. Exempelvis kan ett dokument innehålla recept med tårter med antingen grädde eller bär. I exemplet nedan har NOT grädde satts inom parentes för att göra exemplet extra tydligt, då NOT grädde måste ha en sammanbindande operator framför sig. (NOT i samband med en term behandlas som en egen term.)

tårta AND (bär OR blåbär OR hallon OR björnbär) AND (NOT grädde)

Dokumenterna som återges skall nu innehålla information om tårter och någon av de olika bärsorterna, men inte innehålla termen grädde. I våra sökningar i Sociological Abstracts används som bekant den booleska söklogiken och systemet har en boolesk matchning.

6.1.2 Vektormodellen

I vektormodellen beskrivs dokument och sökfrågor som vektorer. En komponent i en dokument- eller sökfrågevektor kallas för termvikt och motsvarar hur relevant en indexterm är för dokumentet eller sökfrågan. Dessa vektorer kan användas för återvinning och rankning genom att ett likhetsmått beräknas mellan sökfråge- och dokumentvektorerna. Det likhetsmått som ofta används i vektormodellen är cosinusmättet, som bygger på vinkeln mellan vektorerna.⁵⁰ I exemplet nedan visualiseras detta genom att cosinus går från noll till ett och då cosinus nästan är ett är det en bra matchning mellan sökfrågan och dokumentet. Ett dokument som inte matchar sökfrågan närmar sig noll.⁵¹



Figur 1: Visar hur vinkeln mellan sökfråge- och dokumentvektorn blir större ju mindre likt ett dokument är för en sökfråga.

Till skillnad från den booleska modellen kan vektormodellen även klara partiell matchning, det vill säga gradvis matchning och modellen återvinner dokument som helt eller delvis motsvarar sökfrågans innehåll. Likhetsmättet kan användas som rankningsfunktion, då dokument med större likhet med sökfrågan kommer före i träfflistan, vilket kan ge ett precist svar med hänsyn till användarens topic.⁵²

Modellen anses vara en av de bättre då den är snabb och enkel och har fördelen att termviktningen ger hög återvinningseffektivitet. Idag är den en av de mest använda, men nackdelen är att det är svårt att förbättra de rankade sökresultaten utan query expansion eller relevansfeedback.⁵³

⁵⁰Baeza-Yates & Riberio-Neto 1999, s.27ff.

⁵¹Ackerman, Rich, 5263 - *Vector Model Information Retrieval*, <http://www.hray.com/5264/math.htm> [2006-05-11].

⁵²Baeza-Yates & Riberio-Neto 1999, s. 27.

⁵³Baeza-Yates & Riberio-Neto 1999, s. 30.

Samma topic som i det booleska exemplet används nedan. *Ett recept på en tårta med bär, men inte grädde.*

I stället för de booleska operatorerna anges i en lista endast de söktermer som användaren vill skall finnas med i de dokument som återges. Därmed kan användaren inte som i den booleska sökningen uttryckligen exkludera ett ord som inte får förekomma i dokumenten och grädde utelämnas sålunda eftersom termen inte är önskad.

tårta bär blåbär hallon björnbär

Förutsatt att dokumenten rankas bör nu de dokument som placeras högst upp i träfflistan innehålla flest termer från sökfrågan.

Relevansfeedbacken är till stor hjälp då systemet skall återge de dokument som antas vara relevanta. Detta eftersom användaren kan specificera både relevanta och irrelevanta dokument. Sökfrågans vektor ändrar riktning efter den relevansfeedback som systemet fått och rör sig mot de dokument som angivits som relevanta och i från dem som angivits som irrelevanta.

6.2 Relevansbedömning

I föregående avsnitt behandlades relevans huvudsakligen utifrån systemets relevansbedömning medan vi här redogör för relevansbedömning utifrån användarens perspektiv.

Relevansen av ett dokument innebär kortfattat att man ser till hur relevant ett dokument är för ett visst informationsbehov. Därmed är relevansbedömningen en viktig del av informationssökningsprocessen, men termen är omdiskuterad och långt ifrån exakt.⁵⁴ Inom IR finns olika sätt att se på relevansbedömning och de två mest återkommande i forskningslitteraturen är begreppen pertinence och relevance.

Begreppet relevance syftar i dessa sammanhang till förhållandet mellan ett dokument och det uttryckta informationsbehovet. Relevansen bedöms utifrån dokumentets ämnesmässiga innehåll i förhållande till sökfrågan och det är denna relevansbedömning vi använder oss av i vår undersökning. Bedömningen bör ske av mer än en person, oberoende av varandra för att få en så objektiv bedömning som möjligt. Relevansen är i detta sammanhang en fråga om hur olika personer bedömt graden av överensstämmelse mellan ett dokument och sökfrågan som genererat listan med dokumentet. Lancaster skriver följande om relevance: ”In fact, rather than saying that a document is relevant to a requests, it would be better to say that the document has been judged by a particular individual or group of individuals.”⁵⁵

Pertinence fokuserar på förhållandet mellan det framtagna dokumentet och den som har informationsbehovet. Beslutet om huruvida det är relevant avgörs av den informationssökande då det är endast denna som kan avgöra detta. Relevansen ändras allt eftersom informationssökarens behov ändras. Detta sammanfattas i följande citat av Lancaster: ”Clearly only the requester can decide whether or not a particular document contributes to the satisfaction of his information need, since only he knows what his need is.”⁵⁶

⁵⁴Lancaster 1979, s. 256.

⁵⁵Lancaster 1979, s. 261f.

⁵⁶Lancaster 1979, s.263ff.

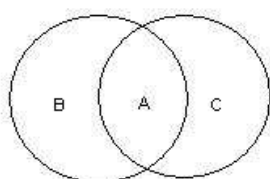
6.3 Precision och Recall

Två standardmått vid utvärdering av en sökfrågas återvinning är precision och recall.

Recall anger hur stor del av databasens relevanta dokument som återvunnits. Detta görs genom att dividera de framtagna relevanta dokumenten med det totala antalet relevanta dokument som finns i databasen. Måttet svarar på frågan: "Hur stor del av de relevanta dokumenten har återvunnits?"

Precision dividerar de relevanta återvunna dokumenten med sökfrågans alla återvunna dokument för att se i hur hög grad systemet återvunnit relevanta dokument. Måttet svarar på frågan: "Hur stor andel av de återvunna dokumenten är relevanta?"

I exemplet nedan är A=totala antalet relevanta återvunna dokument, B=totala antalet relevanta dokument som finns i databasen och C=totala antalet dokument som återvunnits.



Figur 2: Antal dokument vid återvinning för att visualisera precision och recall.⁵⁷

$$\text{Recall} = \frac{A}{A + C} \quad \text{Precision} = \frac{A}{A + B}$$

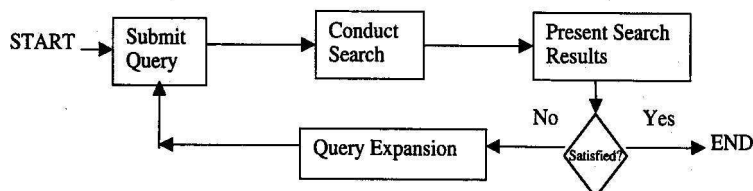
Effektiviteten av återvinning är som bäst då precision och recall är på någorlunda samma nivå och då så hög som möjligt.⁵⁸

⁵⁷Efter idé av: Jizba, Richard. *Measuring search effectiveness*, Creighton University, <http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html> [2006-05-19].

⁵⁸Chamis 1991, s.13.

7 Query Expansion

Query expansion är en återvinningsstrategi som innebär att användaren vid en sökning kompletterar sin sökfråga med alternativa termer vilka överensstämmer med de ursprungliga söktermernas betydelse för att förbättra återvinningsresultatet. Med query expansion försöker användaren förbättra den initiala sökfrågan och ett problem inom ämnesområdet är att utse de termer som skall användas i sökprocessen. Metoden kan användas oberoende av informationssystemets återvinningsteknik.⁵⁹



Figur 3: Modellen illustrerar query expansion i en sökprocess.⁶⁰

Vid en traditionell sökning (exempelvis en boolesk sökning) bör användaren bryta ner informationsförfrågan i dess beståndsdelar, det vill säga uppmärksamma de huvudkoncept som informationsbehovet består av samt välja termer som representerar dessa. Den sökande bör tänka på hur de valda termerna stämmer överens med termerna som används för att representera dokumenten. Ibland är en term kanske inte tillräcklig för att representera ett koncepts innehåll och behöver kompletteras med andra termer för att innehållet skall bli korrekt och fullständigt representerat.⁶¹

Query expansion kan delas in i manuell, automatisk och interaktiv query expansion. Två viktiga faktorer som bör uppmärksammas vid användning av någon av dessa är källan som används för urvalet av termer samt metoden för att välja ut termerna till expansionen. Källorna för urvalet av termer till expansionen delas in i kollektionsberoende samt kollektionsoberoende källor. En kollektionsberoende källa är som namnet antyder bunden till den kollektion som den är konstruerad för, medan en kollektionsoberoende källa inte är bunden till en speciell samling. I vår undersökning använder vi en domänspecifik tesaurus som tillhör de kollektionsoberoende källorna.⁶² Nästa sida visar en schematisk bild över hur query expansion kan delas in i underavdelningar. De tre olika indelningarna av query expansion, samt källorna för urvalet av termerna blir tydliga.

Baeza-Yates och Ribeiro-Neto har delat in insamlingen av information, det vill säga termer för expansion i tre kategorier: Informationsfeedbacken från användaren, information som kan inhämtas genom dokument som redan återvunnits och information som är möjlig att inhämta från hela dokumentkollektionen.⁶³

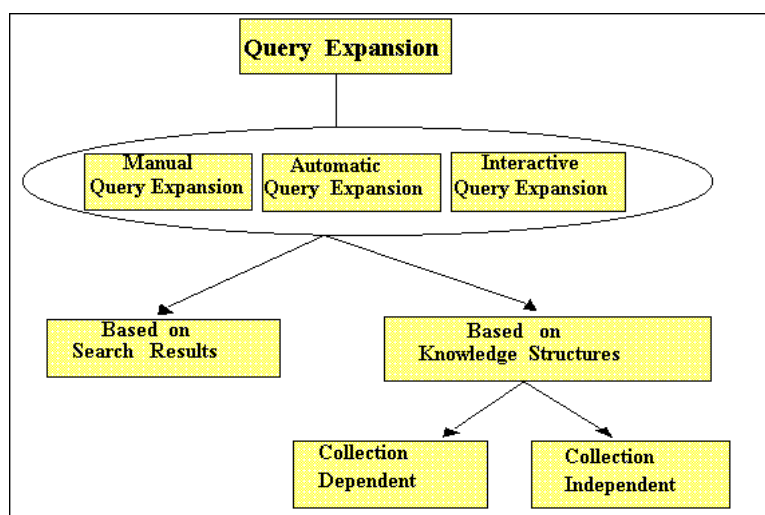
⁵⁹Efthimiadis, Efthimis N. 1996. Query expansion, *Annual Review of Information Science and Technology*, vol. 31, s. 122.

⁶⁰Chu, Heting 2005. *Information Representation and Retrieval in the Digital Age*. s. 67. (Med tillstånd att användas av Chu.)

⁶¹Efthimiadis, 1996, s. 121.

⁶²Efthimiadis, 1996, s. 122ff.

⁶³Baeza-Yates & Ribeiro-Neto 1999, s. 117.



Figur 4: Modellen visar query expansion samt vad valet av termkällor kan baseras på, samt indelningen av beroende och oberoende källor.⁶⁴

7.1 Automatisk Query Expansion

Detta är en expansionsteknik som ofta ingår i informationssystem som bygger på termers viktning samt association med varandra. Efthimiadis anser att expansionsprocessen ofta är svår att detaljerat beskriva eftersom den, som han uttrycker det, är gömd i det system i vilket det ingår.⁶⁵

Metoden kan bland annat fungera så att ett sökresultats överst placerade dokument (de högst rankade) förmodas vara relevanta och inkluderas i en omformulering av söksträngen.⁶⁶ Användaren är inte själv delaktig i valet av termer, utan informationssystemet utser på egen hand termer från förutbestämda fält hos bibliografiska poster i samlingen för att sedan vikta och relevansranka dessa (exempelvis, som i denna undersökning, från deskriptorfältet).⁶⁷ Användaren är sålunda inte direkt involverad i omformuleringsprocessen och söksträngen ändras bitvis av systemet vid varje sökomgång. Enligt Chu är denna query expansion egentligen inte en utökning av en sökning (som den engelska termen expansion indikerar) utan snarare en modifiering av den eftersom grundsökningen inte nödvändigtvis expanderas utan istället modifieras vid varje söktillfälle.⁶⁸

7.2 Interaktiv Query Expansion

Denna expansionsteknik bygger på att två parter samverkar när termerna för sökningen skall väljas ut. Den ena parten är, likt vid automatisk query expansion, själva systemet och den andra är informationssökaren, det vill säga användaren. Systemet kan vara konstruerat så att det väljer ut och rankar termer bland förutbestämda fält hos de bibliografiska posterna. Detta för att sedan presentera termen eller termerna för användaren, vilken i sin tur väljer dem som passar. Användaren fattar sålunda de slutgiltiga besluten grundade på vad systemet bedömt vara bra termer. Källorna som används för urvalet av termerna kan exempelvis vara dokument från tidigare sökresultat, precis som

⁶⁴Efthimiadis, 1996, s. 122ff. (Med tillstånd att användas av Efthimiadis.)

⁶⁵Efthimiadis 1996, s. 143.

⁶⁶Chu 2005, s. 67.

⁶⁷Efthimiadis 1996, s. 156.

⁶⁸Chu 2005, s. 67.

vid automatisk query expansion.⁶⁹

Förutom att systemet utser användbara termer för expansionsprocessen kan det även plocka fram hela dokument som anses vara relevanta för sökningen. Dessa relevansbedöms sedan av användaren och utifrån denna bedömning utför systemet sedan nya sökningar. Processens huvudsakliga syfte är att identifiera viktiga termer eller uttryck från dokumenten, vilka sedan kan användas för den utvidgade sökningen.⁷⁰

7.3 Manuell Query Expansion

Manuell query expansion förknippas oftast med booleskt sammansatta sökningar och det är endast användaren av systemet som utser termer till expansionen, utan hjälp av systemet. För att utse dessa termer kan användaren bland annat ta hjälp av tesaursar, tidigare återvunna dokument samt olika index och ordlistor med mera, men även hämta termerna från sitt eget ordförråd. Vid manuell query expansion är sökstrategin viktig och användaren bör enligt Efthimiadis ha förståelse för den matchningsteknik som används i systemet samt kunskap om indexeringsvokabuläret. Vidare bör användaren vara insatt i det informationsbehov som ligger bakom sökningen.⁷¹ Detta kan tolkas som att användaren bör ta reda på så mycket som möjligt om den matchningsteknik som används i det aktuella systemet för att kunna uppnå bästa möjliga sökresultat.

En viktig del vid formuleringen av sökfrågan är att dela in informationsbehovet i rätt facetter, det vill säga beståndsdelar som bäst matchar informationsbehovet. Därefter kan användaren bestämma metod för sökningen och välja mellan olika sökstrategier. Beroende på syftet med informationssökningen kan valet av sökstrategier variera. En grov indelning är så kallad direkt sökning där användaren redan har viss kännedom om det som eftersöks, exempelvis ett författarnamn. Motsatsen till detta är så kallad browsing som kan användas vid sökning för att skapa sig mer specifik kunskap om ett ämne som användaren inte är speciellt insatt i.⁷²

I vår studie används Building Blocks, en ofta använd metod vid boolesk sökning⁷³ som går till på följande sätt: Först delar användaren in informationsbehovet i olika koncept eller facetter. Dessa delas ytterligare in i mindre beståndsdelar och grupper av termer med inbördes relationer (exempelvis SYN och/eller NT) bildas vilka fogas samman med den booleska operatör OR. Varje underavdelning sätts sedan samman med hjälp av booleska operatör AND. Sökstrategin gör det enkelt för användaren att få en översikt av söklogiken för en eventuell senare tillbakablick.⁷⁴ Det är denna strategi som följts vid utförningen av sökningarna i denna undersökning och på nästa sida visas en schematisk beskrivning av metoden.

⁶⁹Efthimiadis 1996, s. 156f.

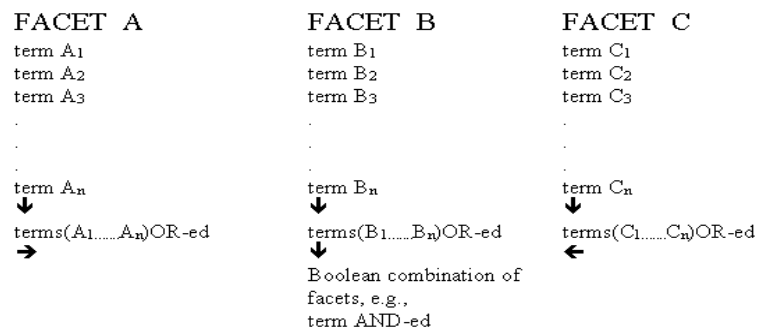
⁷⁰Baeza-Yates & Ribeiro-Neto 1999, s. 118.

⁷¹Efthimiadis 1996, s. 126.

⁷²Rowley, Jennifer & Farrow, John 2000. *Organizing knowledge: An introduction to managing access to information*, s. 104.

⁷³Kekäläinen & Järvelin 1998, s. 131.

⁷⁴Efthimiadis 1996, s. 127f.



Figur 5: Modellen illustrerar Building Blocks: Ett informationsbehov delas in i facetter, vilka kan bindas samman med exempelvis booleska operatorer.⁷⁵

⁷⁵Efthimiadis 1996, s. 127f. (Med tillstånd att användas av Efthimiadis.)

8 Kontrollerad vokabulär

Kontrollerad vokabulär är en uppsättning termer som återfinns i en kontrollerad och standardiserad lista, likt en tesaurus. Dessa bör användas vid indexering och sökning efter dokument för att få bästa resultat och för att kringgå de nackdelar som finns med okontrollerad vokabulär⁷⁶. Ett exempel på problem som kan uppstå vid sökning eller användning av okontrollerad vokabulär är fel sammansättningar av ord, fel relationer mellan ord och ”homografproblem”.⁷⁷ Mer specifikt är syftet med att skapa kontrollerad vokabulär främst att standardisera termerna, att öka konsistensen vid indexeringen, samt att få en förståelse för relationerna mellan termerna.⁷⁸ De redskap som användas vid indexering av kontrollerad vokabulär är bland annat ämnesordslistor och tesaurusar.

En ämnesordlista är en lista över ämnesrelaterade termer som ursprungligen skapats för att förbereda ämnesingångar (entries/headings) och därmed är termerna relativt breda. Enligt Chowdury innehåller dock dagens ämnesordslistor från Library of Congress även information om vilken relation termerna har, likt i en tesaurus.⁷⁹ En tesaurus innehåller både en lista över alla kontrollerade termer, samt visar relationerna/släktskapen mellan termerna. En officiell definition av termen tesaurus är enligt Chamis (1991): ”a controlled and dynamic documentary language containing semantically and generically related terms, which comprehensively covers a specific domain of knowledge.”⁸⁰ I en tesaurus framgår tydligt vilka termer som skall användas vid indexering samt sökning i en databas. Termerna i en tesaurus är antingen föredragna (preferred) och visar de termer som skall användas vid indexering samt sökning i databasen, eller inte föredragna (non-preferred), vilka fungerar som ingångar till de föredragna termerna.⁸¹ Vilken term som bör väljas ur tesaurusen beror på benämningarna som visar vilken form av relation termerna har till varandra.

Det finns tesaurusar för olika ämnesområden och inom ett givet ämne finns det vanligtvis ett fackspråk, vilket gör det lättare att välja enhetliga termer till indexeringen. Då det primära syftet med kontrollerad vokabulär är, som nämndes ovan, att standardisera termer är användningen av fackspråk en fördel för att öka konsistensen. Förutom den klassiska tesaurusen som exemplifieras på följande sida finns det en rad sorter som är utvecklade efter specifika behov och innehåll. (En grov indelning är de tidigare nämnda kollektionsberoende samt kollektionsoberoende tesaurusarna.⁸²)

Här följer en genomgång av de olika relationerna som kan förekomma i en tesaurus.

En tesaurus innehåller deskriptorer, det vill säga ord eller fraser som kan representera ett koncept. Dessa deskriptorer har hierarkiska relationer beskrivna i form av korsreferenser där det framgår vilka ord som skall användas. En NT är underordnad den deskriptor som den har sin referens till. I vårt exempel är *agricultural technology* begreppsmässigt underordnad *technology*, vilket visar att det är en mer specifik term. BT är överordnad den term som den har sin referens till. Deskriptorn *Science and technology* är överordnad (BT) *technology* medan *agricultural technology* är underordnad (NT) *technology*. Related Terms (RT) visar att två deskriptorer är släkt på det innehållsmässiga

⁷⁶Okontrollerad vokabulär kallas även naturligt språk.

⁷⁷Chamis 1991, s. 22.

En homograf är ett ord som stavas likadant som ett annat, men som har ett annorlunda uttal samt ofta en annan betydelse. (NE)

⁷⁸Chamis 1991, s. 14.

⁷⁹Chowdhury 1999, s. 122.

⁸⁰Chamis 1991, s. 16.

⁸¹Chowdhury 1999, s. 125f.

⁸²Efthimiadis 1996, s. 122f.

planet. I exemplet ser vi att *scientific research* är relaterad till *technology*, vilket betyder att de är besläktade innehållsmässigt. Use For (som här efter förkortas UF) kan antingen vara kvasi-synonym⁸³ eller vanlig synonym och termen som står efter UF ersätts av en annan term. Exempelvis kan det stå *technology* UF *applied sciences* vilket innebär att man bör använda *technology* istället för *applied science*.

Exempel på utdrag ur Sociological Abstracts tesaurus:

Technology

UF

Applied Sciences

Hydraulic

NT

Agricultural Technology

Appropriate Technologies

Automation

Biotechnology

BT

Science and technology

RT

Labor Process

Research Applications

Resources

Science

Scientific Knowledge

Scientific Research

Scientific Technological Revolution

Chamis (1991) framhäver att kontrollerad vokabulär påverkar precision och recall. Detta eftersom möjligheten för en användare att hitta specifik information beror på de specifika termerna som används i den kontrollerade vokabulären.⁸⁴

⁸³En term som är synonym med en annan, men som inte alltid går att bytas ut mot den term den är synonym med.

⁸⁴Chamis 1991, s. 14f.

9 Metod

I detta kapitel redogör vi för praktiskt tillvägagångssätt för undersökningen samt för de metoder som användes för att analysera de data som tagits fram.

Vår undersökning är av kvantitativ art och grundar sig på en experimentell studie där värdena är det väsentliga och är förutsättningen för analysen. Eftersom vi inledningsvis inte kände till någonting om Sociological Abstracts började vi med att undersöka vilka aspekter som är viktiga för sökningen i denna specifika databas. Som nämndes i kapitlet om query expansion, kan det vara bra att ha kunskap om systemet för att kunna göra bästa möjliga sökningar.⁸⁵ Detta gjorde vi genom att ”surfa runt” på CSAs webbsida för att samla information om databasvärden, samt sätta oss in de premisser och valmöjligheter som finns vid sökning i databasen.

- På vilka sätt kan användarna söka i databasen? Exempelvis snabbsökning (Quick Search), avancerad sökning (Advanced Search), kommandosökning (Command Search). Vilka fält är sökbara och hur kan man avgränsa sina sökresultat och så vidare.

- Vid formulering av sökfrågor: Vilket sökspråk bör användas och vilken matchningsteknik ligger bakom återvinningen?

- Vid återvinningen: Rankar databasen dokumenten, och i så fall utifrån vilka relevanskriterier?

En del av den information som vi fick efter de ovan nämnda iakttagelserna presenterades i kapitlet om Sociological Abstracts. Mycket visade sig vara av stor vikt för oss när det gällde bland annat att fatta rätt beslut om fortsatt utvärderingsmetod, som exempelvis val av sökstrategi.

9.1 Formulering av topic

Vi ville knyta an till reella informationsbehov som kan tänkas finnas för databasens användare på Chalmers tekniska högskola och konstruerade topic i samråd med bibliotekarierna på Chalmers bibliotek eftersom de är insatta i vilka sökfrågor som användarna kan tänkas rikta till databasen. Utöver detta tog vi även hjälp av Chalmers egen databas för publikationer, såsom examensarbeten och avhandlingar för att skapa oss en bild av vad ett eventuellt informationsbehov skulle kunna vara.

När ett topic formuleras används naturligt språk för att uttrycka ett informationsbehov och man utgår sedan från detta topic då sökfrågorna formuleras.⁸⁶ Vi konstruerade 15 topic och DCV sattes till 15, vilket innebär att återvinningsresultatet skulle bestå av minst 15 träffar.

⁸⁵Efthimiadis 1996, s. 126.

⁸⁶Harman, Donna. Overview of the third text retrieval conference (TREC-3), *NIST Special Publication 500-226: Overview of the Third Text Retrieval Conference (TREC-3)*, s.3, http://trec.nist.gov/pubs/trec3/t3_proceedings.html [2006-03-09].

9.2 Sökstrategi

Vid planeringen av sökstrategierna tog vi hjälp av tidigare forskning. Ett exempel på en studie där sökningarna utfördes med samma upplägg som i denna undersökning är tidigare nämnda undersökning av Kristensen och Järvelin (1990). Upplägget inkluderar en initial sökning, en expansion av denna samt slutligen en expansion med alla termer inkluderade. Deras expansionstermer består dock av andra relationer än de vi har valt.

Vid val av strategi togs det även hänsyn till den aktuella databasen som undersökningen utfördes i och alla sökningar skedde mot deskriptorfältet. Eftersom rankningen sker efter deskriptorernas förekomst i deskriptorfältet utfördes sökningarna med hjälp av deskriptorer. Dessutom har databasens dokument indexerats med hjälp av tesaurusen⁸⁷ och med denna vetskap fann vi det naturligt att använda oss av deskriptorerna i våra sökningar. Då vi inte hade möjlighet att relevansbedöma samtliga återvunna dokument använde vi oss av funktionen Relevance Rank. Denna funktion gör så att de för systemet mest relevanta dokumenten placeras överst i dokumentlistan

I denna undersökning utfördes manuell query expansion som skulle kunna simulera automatisk query expansion. Denna tänker vi oss utförs enligt följande: Först sker en initial sökning, där användaren väljer två eller tre deskriptorer som representerar topicets huvudkoncept. I den första expansionen hämtar systemet alla NT som finns under respektive deskriptor från s1. Alla NT binds samman med operatoren OR, samt att koncepten binds samman med operatoren AND. I den tredje expansionen hämtas även alla BT till deskriptorerna från s1 och sökfrågan utförs med samma kombination av operatörer, det vill säga koncepten binds samman med AND och de hierarkiska termerna till deskriptorerna från s1 med operatoren OR.

Exempel på sökstrategierna är:

Sökning 1.

Innehåller minst två deskriptorer som hämtas ur tesauren.

Exempel: DE="worker attitudes" AND DE=innovations

Sökning 2.

Deskriptorerna från s1 expanderas med alla NT, minst en.

Exempel: DE="worker attitudes" AND DE=(innovations OR "technological innovations")

Sökning 3.

Deskriptorerna från s1 expanderas här med alla NT och BT, minst en av varje.

Exempel: DE=("worker attitudes" OR attitudes) AND DE=(innovations OR "technological innovations")

9.3 Strategi för relevansbedömning

När sökningen väl var utförd relevansbedömdes de 15 översta dokumenten i träfflistan. Vår strategi för detta var att ta fasta på det innehåll som topicet representerade. För att undersöka om de träffar vi fick överensstämde med topicet och därmed var relevanta, studerade vi den bibliografiska referensen och bedömde dess ämnesmässiga innehåll. Vi bedömde ingen gradvis relevans, utan

⁸⁷Denna information kommer från Lin Kwaun, CSA.

dokumentet ansågs antingen vara relevant eller inte, det vill säga 1 eller 0. Abstract, samt titel studerades ingående och om inte topicets ämnesmässiga innehåll representerades ansåg vi inte träffen vara relevant.

Översikt av tillvägagångssättet vid relevansbedömning:

1. Se till topicets ämnesmässiga innehåll.
2. Läser främst abstractet, men även titeln, och bedömer om de stämmer överens med topicets ämnesmässiga innehåll.

Relevansbedömningen utfördes av uppsatsförfattarna oberoende av varandra för samtliga topic, enligt relevansmetoden relevance som beskrevs i kapitel 6.2. Detta för att minska subjektiva bedömningar i den mån det går och för att få en god uppfattning huruvida träffen var relevant eller inte. Till hjälp formulerades relevanskriterier för varje topic vilka måste uppfyllas för att dokumenten skulle betraktas som relevanta. Ett exempel på detta är:

Topic nummer 1: The reception of innovations on workplaces.

Relevanskriterier: Dokumenten skall beröra mottagningen/reaktionen bland anställda av en uppfinning/nyhet på en arbetsplats.

9.4 Beräkning av precision och överlappning

I detta avsnitt beskrivs de mått som används för att beräkna återvinningseffektiviteten och överlappningen.

9.4.1 GPRD

För att mäta återvinningseffektiviteten av relevanta dokument för de tre sökstrategierna beräknades den genomsnittliga precisionen vid observerade relevanta dokument (GPRD). Av de 15 högst rankade dokumenten i dokumentlistan mätte vi precisionen vid varje relevant dokument och beräknade sedan ett genomsnittligt värde för varje sökstrategi och topic (vilket är GPRD). Ett genomsnittligt värde för varje sökstrategi beräknades därefter över de 15 topicen. I exemplet nedan representerar r ett relevant dokument i dokumentlistan och r är därmed ett visst dokument på en viss plats som precisionen beräknas för.

Vi konstruerar ett exempel för beräkning av GPRD: Vid en informationssökning med en av de tre sökstrategierna där DCV är satt till 15 återvinns tre relevanta dokument. Det första relevanta dokumentet har position två i dokumentlistan, det andra position åtta i dokumentlistan och det tredje position tolv. Precisionen vid observerade relevanta dokument beräknas genom att dividera de hittills kända relevanta dokumentens antal med det aktuella dokumentets position i dokumentlistan.⁸⁸ Detta blir tydligt om man studerar tabellen nedan, där de 15 olika posterna motsvarar de 15 översta träffarna i dokumentlistan.

⁸⁸Baeza-Yates & Riberio-Neto 1999, s. 76. Vårt exempel är inspirerat av författarnas exempel på denna sida.

Precision vid observerade relevanta dokument.	
1. 0,00 %	9. 0,00 %
2. $1/2=0,5=50,00\%$ <i>r</i>	10. 0,00 %
3. 0,00 %	11. 0,00 %
4. 0,00 %	12. $3/12=0,25=25,00\%$ <i>r</i>
5. 0,00 %	13. 0,00 %
6. 0,00 %	14. 0,00 %
7. 0,00 %	15. 0,00 %
8. $2/8=0,25=25,00\%$ <i>r</i>	

Tabell 5 : Visar relevanta dokument vid punkten *r*.

Då precisionsvärdet beräknats vid varje observerat relevant dokument i det aktuella topicet beräknas ett genomsnittligt precisionsvärde (GPRD) för den aktuella sökstrategin och topicet.

I exemplet ovan skulle GPRD då bli $(0,5+0,25+0,25)/3=0,3333$. Beräknat i % är GPRD då 33,33%.

Därefter beräknas även ett genomsnittligt värde av GPRD för varje sökstrategi, över de 15 topicen.

9.4.2 Jaccards index

Jaccards index är ett likhetsmått, som inom IR kan användas för att mäta likheten mellan återvunna dokumentlistor. Detta mått har ursprungligen skapats av biologer för att till exempel mäta mångfalden av särskilda djur- eller växtarter inom två samhällen samt se till likheter mellan dem.⁸⁹ Metoden är därmed inte ämnesbunden utan kan användas som likhetsmått inom varierande vetenskaper. Eftersom en av frågeställningarna i denna uppsats avser att besvara huruvida de olika sökstrategierna återvinner samma relevanta dokument och därmed överlappar varandra, används måttet för att beräkna eventuell likhet.

Tillvägagångssättet för att beräkna likheten kan beskrivas enligt följande: I nedanstående formel betecknar X mängden av återgivna relevanta dokument för en sökstrategi och Y betecknar återgivna relevanta dokument för en annan sökstrategi. Vid beräkning av Jaccards index delas antalet dokument i både X och Y (snittet) med antalet dokument som förekommer i antingen X eller Y (unionen).

$$\frac{|X \cap Y|}{|X \cup Y|}$$

För att konkretisera detta tänker vi oss följande exempel: I första sökningen finns det fem relevanta dokument och i andra sökningen finns det tre relevanta, dessa tre återkommer i båda sökningarna. Snittet blir därmed tre. Vid uträkning av unionen tas varje dokument endast med en gång och den blir därmed i detta fall fem. Som exemplet nedan visar skulle resultatet bli 0.6 vilket visar en överlappning på 60,00 %.

$$\frac{3}{5} = 0.6$$

⁸⁹Meyer, John R. *So what's the big deal?*, NC State University, <http://www.cals.ncsu.edu/course/ent525/soil/gcextend.html> [2006-03-11].

Värdet som beräknas står för hur lika de båda sökstrategierna återvinner dokument, det vill säga vilken överlappning de båda sökstrategierna har. Likhetsmättet får ett värde mellan noll och ett. Är det en hög matchning mellan X och Y blir det ett högre värde, vilket indikerar att de återvinner till stor del samma relevanta dokument. Ett lägre värde indikerar däremot att de båda sökstrategierna återvinner olika relevanta dokument, det vill säga de kan komplettera varandra.

9.5 Sammanställning av resultat

Överlappningen mellan sökstrategierna beräknades, som beskrevs i föregående kapitel, med Jaccards index. Först beräknades överlappningen parvis för att vi sedan skulle kunna jämföra s1, s2 och s3 sinsemellan.

För att beräkna prestandan för varje sökstrategi använde vi oss av GPRD enligt den metod som nämndes i 9.4.1.

Vi valde att i presentationsavsnittet redovisa vårt framtagna material med hjälp av diagram och tabeller. Det finns inom metodlitteraturen en mängd förhållningssätt som bör följas vid valet av presentationsform samt utförandet av presentationen. Detta för att utläsandet av materialet skall vara så klart som möjligt samt inte ge upphov till eventuella misstolkningar. Vi har valt att använda oss av en kombination av diagram och tabeller eftersom det material vi kommer att redovisa består av siffror. Diagrammet ger en översikt medan tabellen också mer i detalj visar undersökningens olika siffervärden.⁹⁰

9.6 Metodologiska problem

Eftersom vi begränsade DCV till 15 behövde vi minst 15 dokument för att kunna relevansbedöma våra sökresultat. Problem som uppstod då vi började söka i databasen var att det vid ett flertal sökningar endast återvanns ett fåtal dokument. Detta skulle kunna ha att göra med topicets karaktär i förhållande till databasens ämnesinnehåll. För att få ett bra underlag för undersökningen, det vill säga åtminstone 15 träffar per sökning, omformulerade vi vissa topic för att inte behöva ändra strategi och exkludera sökningar med teknisk karaktär.

⁹⁰Eggeby, Eva & Söderberg, Johan 1999. *Kvantitativa metoder: För samhällsvetare och humanister*, s. 46ff.

10 Resultat och analys

I detta kapitel redovisas de resultat som undersökningen visat på, visualiserade med hjälp av tabeller och diagram. Resultaten analyseras och vi uppmärksammar även den aktuella databasens sökpremissor. Vid mätningarna av överlappningen samt återvinningseffektiviteten utfördes beräkningarna med två decimalers noggrannhet.

10.1 Analys av sökningen i databasen

I undersökningens inledande skede försökte vi klargöra systemets premisser för sökning och en viktig aspekt för återvinningsresultatet är hur systemet behandlar söktermer, det vill säga om den söker på fraser eller ord.

Vid iakttagelser av deskriptorfält hos de dokument som återvanns kan man se att markering av termer i deskriptorfältet ter sig olika vid sökning med respektive utan citationstecken. Om sökningen $DE="history\ of\ social\ work"$ utförs med citationstecken är det endast denna deskriptor som i sin helhet markeras i det återvunna dokumentets deskriptorfält. Körs en sökning däremot utan citationstecken iakttog vi att även komponenter, det vill säga delar som ingår i andra deskriptorer markerades i deskriptorfältet. Exempel på detta är *federal republic of germany* som inkluderades i sökresultatet eftersom ordet *of* är en komponent av deskriptorn och markerades, vilket även gäller *history* i *womens history*. Gemensamt för sökning med respektive utan citationstecken var dock att de återgav exakt samma resultat, det vill säga samma dokument återgavs och placerades på exakt samma plats i dokumentlistan. Vi antar därför att användningen av citationstecken i detta fall inte påverkade återvinningsresultatet men det finns andra exempel då så var fallet. Ett sådant exempel är då ordet *and* är en komponent i en deskriptor och citationstecken måste användas. I sökningen $DE="science\ and\ technology"$ med citationstecken, återgavs 898 träffar. Om sökningen inte utfördes med citationstecken gav den däremot 4685 träffar eftersom *science* och *technology* behandlades var för sig.

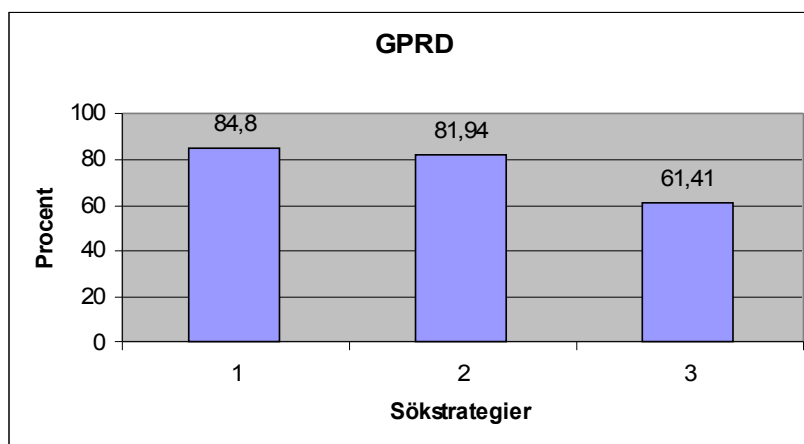
I och med ovan nämnda iakttagelser av behandlingen av ord respektive fraser ville vi vara konsekventa i sökformuleringarna, samt försäkra oss om att det endast var våra valda deskriptorer som helhet och inte komponenter av dessa som matchning skedde efter. Vi valde därför att använda oss av citationstecken då deskriptorerna bestod av fler än ett ord.

För att undersöka vilken modell för återvinning som systemet arbetar efter kontrollerade vi om samtliga facetter i sökningen som bands samman med AND inkluderades i samtliga dokument. Så är fallet och därmed drog vi slutsatsen att det är en boolesk matchning som ligger bakom systemets återvinning. Hade det däremot varit partiell matchning borde dokument som endast delvis matchade sökfrågans termer tagits med i dokumentlistan. Att systemet arbetade efter binär matchning kan exemplifieras genom topic nummer 14. De två deskriptorerna i s1 gav tillsammans 15 dokument, men var för sig 3052, respektive 250 dokument. Hade matchningen varit partiell borde minst 3287 dokument ha återvunnits. Att systemet kan relevansranka de återgivna dokumenten tyder dock på att systemet inte enbart är booleskt. Rankningen skulle exempelvis kunna baseras på viktning av termerna som i exempelvis vektormodellen.

10.2 Analys av GPRD

Det övergripande resultatet visar att s1 och s2 skiljer sig marginellt då det gäller den genomsnittliga

precisionen vid observerade relevanta dokument. Skillnaden mellan s1 och s3 är däremot tydligare då GPRD vid s3 minskar med cirka en fjärdedel i förhållande till s1. I figur 6 nedan visualiseras detta, hur den genomsnittliga precisionen minskar i fallande ordning från 84,80 % till 81,94 % och till sist 61,41 % vid s3.



Figur 6: Medelvärdet för GPRD för s1, s2 och s3 över de 15 topicen.

I tabell 6 nedan presenteras de genomsnittliga precisionsvärdena vid observerade relevanta dokument per sökstrategi och topic. Om vi ser till hela expansionsprocessen skedde en minskning av GPRD från s1 till s3 i tolv av 15 topics, medan den ökade vid två topics och var oförändrad vid ett topic.

Topic	s1	S2	s3
1.	69,56 %	69,56 %	89,44 %
2.	97,87 %	91,03 %	23,72 %
3.	93,92 %	93,92 %	72,94 %
4.	78,34 %	78,34 %	42,02 %
5.	88,35 %	91,71 %	91,71 %
6.	93,79 %	80,56 %	80,56 %
7.	78,13 %	78,13 %	0,00 %
8.	61,34 %	37,26 %	37,26 %
9.	100,00 %	100,00 %	0,00 %
10.	100,00 %	100,00 %	89,39 %
11.	69,03 %	69,03 %	64,00 %
12.	100,00 %	100,00 %	100,00 %
13.	54,67 %	52,49 %	53,63 %
14.	100,00 %	100,00 %	91,57 %
15.	87,01 %	87,01 %	84,86 %
Summa	1272,01	1229,04	921,10
Medelvärde	84,80 %	81,94 %	61,41 %

Tabell 6: Visar GPRD för varje topic och sökstrategi.

Att s1 i majoriteten av topicen visar hög GPRD är i denna undersökning troligtvis ett resultat av att en tesaurus användes för sökningarna och att vi redan i denna sökning har tillräckligt specifika deskriptorer, samt att det endast är två till tre termer som matchas mot dokumenten i databasen. Då matchningen sker efter få deskriptorer, som valts för att de representerade topicets huvudkoncept, är sannolikheten stor att de återvunna dokumenten är relevanta för topicet.

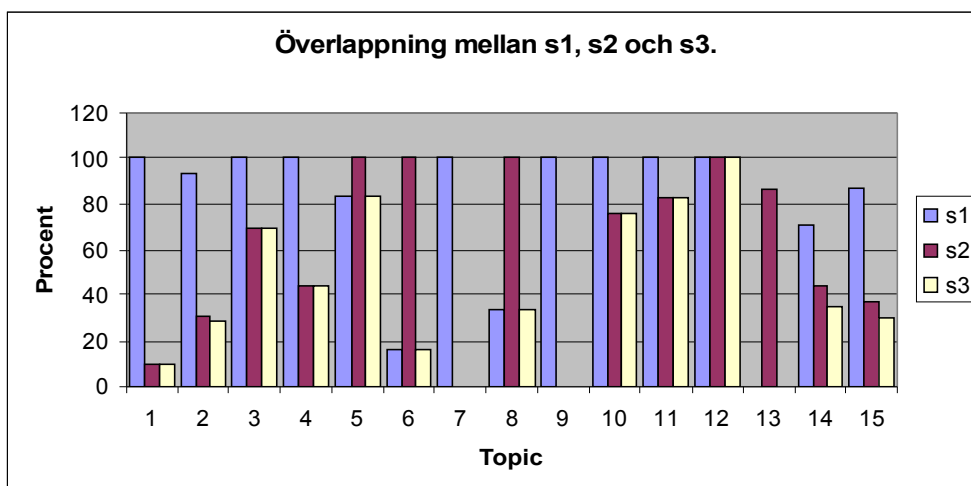
Medelvärde för skillnaden mellan s1 och s2 är marginell gällande GPRD. I s2 tillkommer alla NT för deskriptorerna i den initiala sökningen vilka som mest uppgår till 17 NT. Därmed kan även deskriptorer som inte matchar topicets ämnesmässiga innehåll, men som är relevanta enligt systemets relevansbedömning⁹¹, inkluderas. Risken ökar då för att irrelevanta dokument, utifrån topicets relevanskriterier, återvinns. I denna undersökning har dock tio av tolv topic samma GPRD i s1 som s2, vilket ytterligare styrker teorin om att deskriptorerna i s1 redan är tillräckligt specifika för topicet och att NT därmed inte fyller sin funktion att specificera sökningen.

I s3 tillkommer även BT och sökningen blir mer omfattande, vilket i vår undersökning resulterar i att medelvärdet för GPRD minskar ytterligare. Eftersom både NT och BT förekommer i samma sökning ökar risken ytterligare att irrelevanta dokument kan påverka att GPRD minskar. Dokumenten i dokumentlistan rankas som bekant efter de åtta första deskriptorerna i deskriptorfältet och därför kan en omfattande sökning med många deskriptorer i sökfrågan resultera i att för topicet irrelevanta dokument återvinns. Dokument som är relevanta för topicet riskerar samtidigt att placeras längre ner i dokumentlistan. Ett exempel på en expansion som inkluderar, för topicet irrelevanta termer, finns i topic nummer fem där s1 innefattar termerna *agriculture* och *technology*. I expansionen med NT inkluderas bland andra "*space technology*" och "*medical technology*" eftersom de är NT till *technology*, trots att de inte är relevanta för topicets ämnesmässiga innehåll.

10.3 Analys av överlappningen

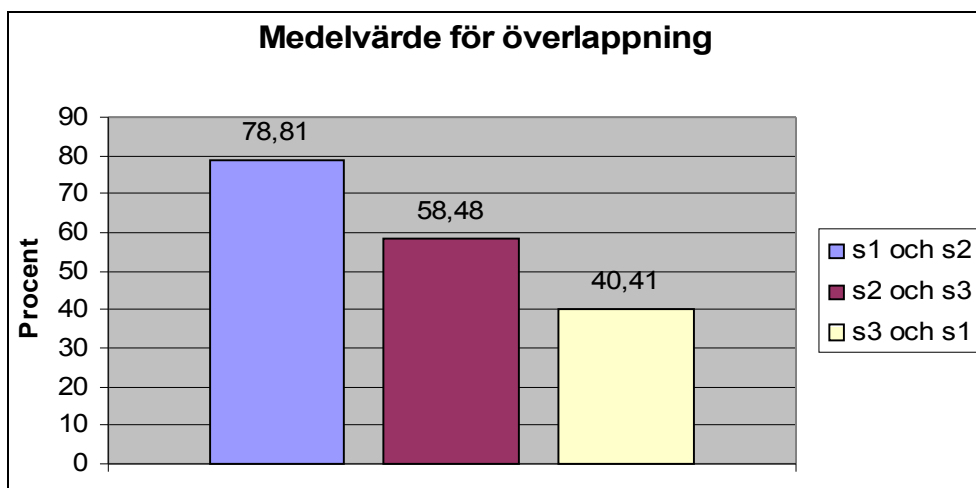
Överlappningen mellan de tre sökstrategierna visar en stor variation mellan olika topic. Som visualiseras i figur 7 har topic nummer tolv en överlappning på 100,00 % mellan alla tre par av sökstrategier. Detta innebär att exakt samma relevanta dokument har återgivits i de tre sökningarna, men det finns även exempel där överlappningen mellan två sökstrategier är 0,00 %, som i topic nummer nio, enligt figur 7. I de fall där ingen överlappning sker kompletterar sökstrategierna varandra och nya relevanta dokument för topicet återvinns. Motsatsen gäller för de topic med hög överlappning där få nya relevanta dokument tillkommer, vilket kännetecknas av de höga stolparna i diagrammet.

⁹¹Se kapitel 6.



Figur 7: Visar överlappningen mellan de tre sökstrategierna för varje topic.

Figur 8 visualiserar den genomsnittliga överlappningen mellan s1 och s2, s2 och s3 samt s3 och s1. Här framgår det att överlappningen minskar i relativt jämna intervaller från 78,81 % mellan sökstrategi s1 och s2 till 58,48 % mellan sökstrategi s2 och s3 och slutligen 40,41 % mellan sökstrategi s3 och s1.



Figur 8: Visar den genomsnittliga överlappningen parvis för de tre sökstrategierna.

I tabell 7 på följande sida framgår det att överlappningen i tio fall av 15 är störst mellan s1 och s2, medan överlappningen mellan strategierna i topic nummer tolv är fullständig. I fyra av 15 sökningar har s2 och s3 störst överlappning medan s1 och s3 inte har den största överlappningen i någon sökning.

Topic	s1 och s2	s2 och s3	s3 och s1
1.	100,00 %	10,00 %	10,00 %
2.	92,86 %	30,77 %	28,57 %
3.	100,00 %	69,23 %	69,23 %
4.	100,00 %	44,44 %	44,44 %
5.	83,33 %	100,00 %	83,33 %
6.	15,38 %	100,00 %	15,38 %
7.	100,00 %	0,00 %	0,00 %
8.	33,33 %	100,00 %	33,33 %
9.	100,00 %	0,00 %	0,00 %
10.	100,00 %	75,00 %	75,00 %
11.	100,00 %	81,82 %	81,82 %
12.	100,00 %	100,00 %	100,00 %
13.	0,00 %	85,71 %	0,00 %
14.	70,59 %	43,37 %	35,00 %
15.	86,67 %	36,84 %	30,00 %
Summa	1182,16	877,18	606,10
Medelvärde	78,81 %	58,48 %	40,41 %

Tabell 7: Visar överlappningen för sökstrategierna parvis vid varje topic.

Ett exempel på ett topic där ingen överlappning sker i två av de parvis jämförda sökstrategierna är topic nummer tretton. Mellan s1 och s2 samt s3 och s1 är överlappningen 0,00 % vilket innebär att inte några gemensamma dokument har återvunnits. De söktermer som användes var i s1 *health care services* samt *medical technology* och expansionstermerna tillförde i detta fall nya relevanta dokument. I det tredje fallet, mellan s2 och s3 var överlappningen däremot 85,71 %. Eftersom även GPRD ökade från s2 till s3 tyder det på att expansionen av BT i detta fall resulterade i att BT gav nya dokument och att några av dessa var relevanta. Detta visar att BT inte endast tillför ett topic för omfångsrika termer, vilket nämndes i 10.2.

11 Slutsatser och slutdiskussion

Här uppmärksammas våra frågeställningar var och en för sig och vi diskuterar de resultat som framkommit i undersökningen.

Syftet med denna uppsats är att undersöka hur olika sökstrategier presterar när det gäller återvinningseffektivitet. För att göra detta jämförs tre olika sökstrategier (s1, s2 och s3) och dess prestationer relativt varandra i fråga om återvinning av relevanta dokument. Vi vill även undersöka i vilken utsträckning samma relevanta dokument återvinns med de tre strategierna och se vilka eventuella mönster som kan utläsas.

Den första frågeställning är: Vilken av sökstrategierna s1, s2 och s3 har högst genomsnittlig precision vid observerade relevanta dokument (GPRD)?

I denna studie har s1 högst GPRD, det vill säga den högsta återvinningseffektiviteten av de tre sökstrategierna. En huvudsaklig faktor som vi tror bidrar till detta är att vi väljer söktermerna ur en tesaurus, vilket medför att deskriptorerna som används i s1 redan är tillräckligt specifika för topicets ämnesmässiga innehåll. Denna analys av resultatet påminner om den diskussion Voorhees för i sin artikel⁹² men då undersökningarna skiljer sig åt bland annat genom val av termkälla samt expansionsstrategi var det inte förutsägbart.

I s2 är den genomsnittliga värdet för GPRD endast aningen lägre än s1 (84,80 % jämfört med 81,94%). För s3 är däremot medelvärdet för GPRD 61,41 % vilket visar en tydligare minskning i förhållande till s1. Gemensamt för s2 och s3 är att de båda expanderar med nya termer som inte alltid behöver vara relevanta för topicets ämnesmässiga innehåll. Konsekvenserna av att en helt ny deskriptor inkluderas i sökningen kan resultera i att dokument som inte motsvarar topicets relevanskriterier återvinns, vilket i sin tur minskar värdet av medelvärdet. Det ideala är dock att även de nytillkomna deskriptorerna är relevanta för topicet, samt att nya relevanta dokument tillkommer. Vid majoriteten av topicen ger expansion med hjälp av tesaurerna sålunda inte någon förbättring av återvinningsresultatet, men visar bra resultat redan i s1.

Den andra frågeställning är: I vilken utsträckning återvinns samma relevanta dokument, med hjälp av de olika sökstrategierna?

Hur de olika sökstrategierna genererar samma dokument varierar från topic till topic. I tio fall av 15 är överlappningen högst mellan s1 och s2, medan överlappningen mellan strategierna i ett topic är fullständig. I fyra av 15 sökningar har s2 och s3 högst överlappning medan s1 och s3 inte uppnår högst överlappning i någon sökning. Att överlappningen är hög mellan s1 och s2 visar att inte många nya relevanta dokument återvinns vid expansion med NT, vilket även GPRD visar då den sjunker vid expansionerna. Eftersom det endast är relevanta dokument i träfflistan som det tas hänsyn till vid beräkning av Jaccards index och GPRD följs dessa två kurvor åt.

Grundat på ovannämnda resultat drar vi slutsatsen att valda metod som användes i undersökningen inte är till fördel då målet är att öka GPRD med hjälp av query expansion. Genomsnittligt återvinns vid expansion färre relevanta dokument och eftersom det endast är de relevanta dokumenten som det tas hänsyn till vid beräkning av Jaccards index minskar även detta värde. Att detta värde är lågt borde teoretiskt sett vara positivt eftersom sökstrategierna kompletterar varandra, men vi tror inte att så är fallet i denna undersökning eftersom precisionen även minskar i s3.

⁹²Voorhees, 1994.

Resultaten gör att vi vill ge följande råd till dem som utför sökningar i Sociological Abstracts med CSA som databasvärd: Sökning med tesaurus är till fördel då sökningar som liknar s1 i strukturen utförs, eftersom användaren har möjligheten att hitta exakta termer direkt. Användaren behöver sålunda inte utföra efterföljande expansioner för att återvinna resultat med hög precision förutsatt att denna är tillräckligt insatt i sitt informationsbehov och vet vilka deskriptorer som bör väljas. Vi kan inte yttra oss om hur query expansion med hjälp av tesaurusen skulle påverka återvinningseffektiviteten om termerna hade handplockats ur denna. Detta eftersom resultaten skulle kunna ha blivit annorlunda om deskriptorerna till expansionen valts ut endast efter topicets ämnesmässiga innehåll.

11.1 Faktorer som kan ha påverkat resultatet

I denna undersökning är DCV satt till 15. Gränsen för DCV kan ha påverkat det slutgiltiga resultatet eftersom ett högre värde för DCV skulle kunna ha bidragit till högre GPRD. Detta då relevanta dokument även kan placeras längre ner i dokumentlistan.

Vid formuleringen av sökfrågorna visade det sig att vår ämneskunskap för ett topic i vissa fall var svagt. Detta ser vi dock inte som ett problem, eftersom vi för att motverka felaktigheter fick expertkunskap från bibliotekarier och andra anställda på Chalmers tekniska högskola.

11.2 Förslag till vidare studier

Då vi i denna studie har utfört en simulering av automatisk query expansion skulle det även vara intressant att se hur resultaten skiljer sig åt då expansionstermer handplockas ur tesaurusen. En grupp testpersoner med expertkunskap inom ett specifikt ämnesområde skulle kunna välja ut de deskriptorer ur tesaurusen som passar ett utvalt topic. Därmed kan man testa om expansionen blir effektivare då deskriptorer som riskerar vara irrelevanta för topicen inte inkluderas.

12 Sammanfattning

Denna uppsats ansluter sig till forskningsområdet query expansion. Inom detta forskningsområde undersöks hur en expanderad sökning kan förbättra återvinningsresultatet. En aspekt som kan påverka återvinningsresultatet är användningen av kontrollerad vokabulär som termkälla vid query expansion, samt de relationer som termerna i sökningarna har.

Uppsatsens syfte var att undersöka hur olika sökstrategier presterar när det gäller återvinningseffektivitet. För att göra detta jämfördes tre olika sökstrategier (s1, s2 och s3) och dess prestationer relativt varandra i fråga om återvinning av relevanta dokument. Vi ville även undersöka i vilken utsträckning samma relevanta dokument återvanns med de tre strategierna och se vilka eventuella mönster som kunde utläsas. De frågeställningar som formulerades för att uppnå syftet var:

- 1. Vilken av sökstrategierna s1, s2 och s3 har högst genomsnittlig precision vid observerade relevanta dokument (GPRD)?
- I vilken utsträckning återvinns samma relevanta dokument, med hjälp av de olika sökstrategierna?

En undersökning utfördes i den bibliografiska databasen Sociological Abstracts där s1, s2 och s3 jämfördes med varandra. Topic konstruerades utifrån fiktiva behov som kan tänkas finnas för studenter vid Chalmers tekniska högskola. Anledningen till detta är att bibliotekarierna på högskolans bibliotek har statistik på att databasen används vilket vi fann intressant eftersom Chalmers är en teknisk högskola och databasen har en sociologisk och beteendevetenskaplig inriktning.

Det metodologiska tillvägagångssättet gick till enligt följande: 15 topic med teknisk och sociologisk karaktär i kombination konstruerades. Söktermerna för dessa hämtades ur databasens tesaurus och strategin för query expansion var s1: En sökfråga med minst två deskriptorer som valdes ur tesaurusen. S2: En sökfråga där deskriptorerna från s1 expanderades med alla NT, minst en. S3: En sökfråga där deskriptorerna från s1 expanderades med alla NT och alla BT, minst en av varje. Strategin simulerade automatisk query expansion som arbetade efter principen att inkludera samtliga NT respektive BT för varje deskriptor från s1. Antalet dokument som beaktades vid varje sökning, det vill säga Document Cutoff Value (DCV), sattes till 15. Relevansbedömningen utfördes av båda uppsatsförfattarna oberoende av varandra och skedde efter relevanskriterier som konstruerades för varje topic. Sökstrategiernas återvinningseffektivitet mättes utifrån genomsnittlig precision vid observerade relevanta dokument (GPRD) och överlappningen av relevanta dokument i de tre sökningarna beräknades med likhetsmättet Jaccards index. Överlappningen beräknades per topic och genomsnittlig precision vid varje observerat relevant dokument och ett medelvärde beräknades slutligen för de båda måtten.

Resultaten av undersökningen visade att s1 presterade bäst gällande medelvärdet för GPRD (84,80 %) medan medelvärdet minskade marginellt vid s2 (81,94 %) och s3 presterade sämst (61,41 %). Att s1 hade högst medelvärde beror med stor sannolikhet på användningen av tesaurusen vilket kan ha bidragit till att det redan i s1 fanns tillräckligt specifika deskriptorer för att återvinna relevanta dokument för topicets relevanskriterier. Eftersom alla NT och BT inkluderades i s2 och s3 fanns risken att även irrelevanta dokument för topicets ämnesmässiga innehåll inkluderades i dokumentlistan, vilket kan ha bidragit till minskad genomsnittlig GPRD vid expansion.

Överlappningen mellan de tre sökstrategierna visade en stor variation mellan olika topic. Ett av dessa hade en fullständig överlappning, det vill säga 100,00 %, medan värdet för topicet med lägst överlappning var 0,00 %. I tio fall av 15 var överlappningen högst mellan s1 och s2. Medelvärdet för överlappningen mellan s1 och s2 var 78,81 %, mellan s2 och s3 58,48 % och mellan s3 och s1 40,41 %. Att överlappningen var hög mellan s1 och s2 visade att inte många nya relevanta dokument återvanns vid expansion med NT, vilket även GPRD visade då den minskade vid expansionerna. Tendensen att värdet minskade vid expansionerna gäller även vid överlappningen, där majoriteten av topicen visade en minskning av överlappningen mellan s1 och s2 till s3 och s1.

Grundat på svaren till frågeställningarna drar vi slutsatsen att valda metod inte är till fördel om målet skulle vara att öka GPRD med hjälp av query expansion. Genomsnittligt återvinns vid expansion färre relevanta dokument och eftersom det endast är de relevanta dokumenten som det tas hänsyn till vid beräkning av Jaccards index minskar även detta värde. Att detta värde är lågt borde teoretiskt sett vara positivt eftersom sökstrategierna kompletterar varandra, men vi tror inte att så är fallet i denna undersökning eftersom precisionen även minskar i s3.

Källförteckning

- Ackerman, Rich, 5263 - *Vector model information retrieval*, <http://www.hray.com/5264/math.htm> [2006-05-11].
- Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. Harlow: Addison-Wesley.
- Chalmers Bibliotek, <http://www.lib.chalmers.se/bibl/Biblsiffror.xml> [2006-02-12].
- Chamis, Alice Yanosko (1991). *Vocabulary control and search strategies in online searching*. New York: Greenwood Press.
- Chowdhury, G.G. (1999). *Introduction to modern information retrieval*. London: Library Association Publishing.
- Chu, Heting (2005). *Information representation and retrieval in the digital age*. Medford, New Jersey: Information Today.
- Convey, John (1992). *Online information retrieval: An introductory manual to principles and practice*. London: Bingley.
- CSA, <http://www.csa.com/> [2006-04-05].
- Efthimiadis, Efthimis N. (1996). Query Expansion, ingår i Williams, Martha E., ed., *Annual Review of Information Science and Technology (ARIST)*, vol. 31, s. 121-187. Medford, New Jersey: Information Today.
- Finns även tillgänglig via http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html#tth_sEc4 [2006-04-01].
- Eggeby, Eva & Söderberg, Johan (1999). *Kvantitativa metoder: För samhällsvetare och humanister*. Lund: Studentlitteratur.
- Greenberg, Jane (2001). Automatic query expansion via lexical-semantic relationships, *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 52, no. 5, s. 402-415.
- Harman, Donna. Overview of the third text retrieval conference (TREC-3), *NIST Special Publication 500-226: Overview of the Third Text Retrieval Conference (TREC-3)*, http://trec.nist.gov/pubs/trec3/t3_proceedings.html [2006-03-09].
- Jizba, Richard. *Measuring search effectiveness*, Creighton University, <http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html> [2006-05-19].
- Kekäläinen, Jaana & Järvelin, Kalervo (1998). The impact of query structure and query expansion on retrieval performance, *SIGIR ninety-eight: proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 24-28, W. Bruce Croft ed., s. 130-137.
- Kristensen, Jaana (1993). Expanding end-users' query statements for free text searching with a

- search-aid thesaurus, *Information Processing & Management*, vol. 29, no. 6, s. 733-744.
- Kristensen, Jaana & Järvelin, Kalervo (1990). The effectiveness of a searching thesaurus in free-text searching in a full-text database, *International Classification*, vol. 17, no. 2, s. 77-84.
- Lancaster, Wilfrid. F. (1979). *Information retrieval systems: Characteristics, testing and evaluation*. New York: Wiley.
- Large, Andrew, (2001). *Information seeking in the online age: Principles and practice*. München: Saur.
- Matsura, Tamiko (2004). *Venturing into a new area-database evaluation*, www.dpc.or.jp/english/SA_2004.pdf [2006-05-19].
- Meyer, John R. *So what's the big deal?*, NC State University, <http://www.cals.ncsu.edu/course/ent525/soil/gcextend.html>, [2006-03-11].
- NE, "homograf", http://www.ne.se.lib.costello.pub.hb.se/jsp/notice_board.jsp?i_type=1 [2006-04-20].
- Reitz, Joan M. *ODLIS - Online Dictionary for Library and Information Science*, http://lu.com/odlis/odlis_d.cfm [2006-06-13].
- Rowley, Jennifer & Farrow, John (2000). *Organizing knowledge: An introduction to managing access to information*. Burlington, Vermont: Gower.
- Van Rijsbergen, C. J. *Information retrieval*, <http://www.dcs.gla.ac.uk/Keith/Preface.html> [2006-03-01].
- Voorhees, Ellen M. (1994). On expanding query vectors with lexically related words, *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC 2)*, http://trec.nist.gov/pubs/trec2/t2_proceedings.html [2006-05-08].
CSA, <http://www.csa.com/> [2006-04-05].
- Zinovjeva, Natalia. *Utvinning av information om ord och ordkombinationer från korpusar*, <http://stp.ling.uu.se/~lars/stp/kurser/fkorpus/synopsis/tema05.html> [2006-05-19].

Appendix

1. The reception of innovations in workplaces.

- 1.DE= "worker attitudes" AND DE=innovations
- 2.DE= "worker attitudes" AND DE=(innovations OR "technological innovations")
- 3.DE=("worker attitudes" OR attitudes) AND DE=(innovations OR "technological innovations")

Relevanskriterier: Dokumenten skall beröra mottagningen/reaktionen bland anställda av en uppfinning/nyhet på en arbetsplats.

2. How can the managerial body handle disagreements in an organization?

- 1.DE="management styles" AND DE=conflict
2. DE="management styles" AND DE=(conflict OR "cultural conflict" OR disputes OR "family conflict" OR "ideological struggle" OR "international conflict" OR "interpersonal conflict" OR "role conflict" OR "social conflict")
- 3.DE=("management styles" OR styles) AND DE=(conflict OR "cultural conflict" OR disputes OR "family conflict" OR "ideological struggle" OR "international conflict" OR "interpersonal conflict" OR "role conflict" OR "social conflict" OR interaction)

Relevanskriterier: Dokumenten skall beröra ledningsgruppens/en ledares hantering av konflikter i en organisation.

3. The influence of stress on job performance.

- 1.DE=stress AND DE="job performance"
- 2.DE=(stress OR "occupational stress" OR "psychological stress" OR trauma) AND DE="job performance"
- 3.DE=(stress OR "occupational stress" OR "psychological stress" OR trauma) AND DE=("job performance" OR performance)

Relevanskriterier: Dokumenten skall beröra sambandet mellan stress och arbetsresultat.

4. The effects of automation on the results of production.

- 1.DE=automation AND DE=productivity
- 2.DE=(automation OR "industrial automation" OR "office automation") AND DE=(productivity OR "labor productivity")
- 3.DE=(automation OR "industrial automation" OR "office automation" OR technology) AND DE=(productivity OR "labor productivity")

Relevanskriterier: Dokumenten skall beröra automatisering och dess inverkan på produktiviteten.

5. The use of technological tools in an agricultural environment.

- 1.DE=agriculture AND DE=technology
- 2.DE=(agriculture OR "animal husbandry" OR "part time farming") AND DE=(technology OR "agricultural technology" OR "appropriate technologies" OR automation OR biotechnology OR cybernetics OR "electronic technology" OR engineering OR "information technology" OR "medical technology" OR "metallurgical technology" OR "space technology")
- 3.DE=(agriculture OR "animal husbandry" OR "part time farming") AND DE=(technology OR "agricultural technology" OR "appropriate technologies" OR automation OR biotechnology OR cybernetics OR "electronic technology" OR engineering OR "information technology" OR "medical technology" OR "metallurgical technology" OR "space technology" OR "science and technology")

Relevanskriterier: Dokumenten skall beröra tekniska redskap inom jordbruket.

6. The access to technology in the scientific context in underdeveloped countries.

- 1.DE="science and technology" AND DE="developing countries"
- 2.DE=("science and technology" OR science OR technology) AND DE="developing countries"
- 3.DE=("science and technology" OR science OR technology) AND DE=("developing countries" OR countries)

Relevanskriterier: Dokumenten skall beröra tillgången till teknik vid vetenskaplig forskning i fattiga länder.

7. Recycling energy for a better environment.

- 1.DE="sustainable development" AND DE=energy
- 2.DE="sustainable development" AND DE=(energy OR "nuclear energy" OR radiation OR "solar energy")
- 3.DE=("sustainable development" OR development) AND DE=(energy OR "nuclear energy" OR radiation OR "solar energy")

Relevanskriterier: Dokumenten skall beröra miljövänlig energi.

8. What kinds of technical aid exist to support different handicaps?

1. DE=technology AND DE=handicapped
2. DE=(technology OR "agricultural technology" OR "appropriate technologies" OR automation OR biotechnology OR cybernetics OR "electronic technology" OR engineering OR "information technology" OR "medical technology" OR "metallurgical technology" OR "space technology") AND DE=(handicapped OR blind OR "congenitally handicapped" OR deaf OR "mentally retarded" OR "physically handicapped")
- 3.DE=(technology OR "agricultural technology" OR "appropriate technologies" OR automation OR biotechnology OR cybernetics OR "electronic technology" OR engineering OR "information technology" OR "medical technology" OR "metallurgical technology" OR "space technology" OR "science and technology") AND DE=(handicapped OR blind OR "congenitally

handicapped" OR deaf OR "mentally retarded" OR "physically handicapped")

Relevanskriterier: Dokumenten skall beröra tekniska hjälpmedel för olika typer av handikapp.

9. How work environment affects people's wellbeing.

1.DE="work environment" AND DE=health

2.DE="work environment" AND DE=(health OR "mental health" OR "occupational safety" and health OR "public health")

3.DE=("work environment" OR environment) AND DE=(health OR "mental health" OR "occupational safety and health" OR "public health")

Relevanskriterier: Dokumenten skall beröra sambandet mellan arbetsmiljö och hälsa.

10. City planning in a historical perspective.

1.DE=history AND DE=city planning

2.DE=(history OR "art history" OR chronologies OR "economic history" OR "history of social work" OR "history of sociology" OR "intellectual history" OR "oral history" OR "political history" OR "psycho history" OR "social history" OR "womens history") AND DE="city planning"

3.DE=(history OR "art history" OR chronologies OR "economic history" OR "history of social work" OR "history of sociology" OR "intellectual history" OR "oral history" OR "political history" OR "psycho history" OR "social history" OR "womens history" OR "social science") AND DE=("city planning" OR "local planning")

Relevanskriterier: Dokumenten skall beröra planering av städer i ett historiskt perspektiv.

11. Architecture in different city environments.

1.DE=architecture AND DE=cities

2.DE=architecture AND DE=(cities OR "central cities" OR "global cities")

3.DE=(architecture OR "fine arts") AND DE=(cities OR "central cities" OR "global cities" OR communities)

Relevanskriterier: Dokumenten skall beröra arkitektur i stadsmiljö.

12. Methods to measure poverty.

1.DE=measurement AND DE=data AND DE=poverity

2.DE=(measurement OR "interval measurement" OR "nominal measurement" OR "ordinal measurement") AND DE=(data OR "aggregate data" OR "categorical data" OR "data banks" OR "panel data" OR scores) AND DE=(poverity OR "child poverity" OR "rural poverity" OR "urban poverity")

3.DE=(measurement OR "interval measurement" OR "nominal measurement" OR "ordinal measurement") AND DE=(data OR "aggregate data" OR "categorical data" OR "data banks"

OR "panel data" OR scores OR information) AND DE=(poverty OR "child poverty" OR "rural poverty" OR "urban poverty" OR "economic conditions")

Relevanskriterier: Dokumenten skall beröra metoder för att mäta fattigdom.

13. Technology to support medical treatment.

1.DE="health care services" AND DE="medical technology"

2.DE=("health care services" OR "dental care" OR "emergency medical services" OR "home health care" OR "long term care" OR "managed care services" OR "mental health services" OR "palliative care" OR "primary health care" OR "womens health care") AND DE=("medical technology" OR "reproductive technologies")

3.DE=("health care services" OR "dental care" OR "emergency medical services" OR "home health care" OR "long term care" OR "managed care services" OR "mental health services" OR "palliative care" OR "primary health care" OR "womens health care" OR "human services") AND DE=("medical technology" OR "reproductive technologies" OR technology)

Relevanskriterier: Dokumenten skall beröra tekniska hjälpmedel för medicinsk behandling (sjukvård, tandvård, äldreomsorg etc.).

14. Human interaction in virtual reality.

1.DE="interpersonal relations" AND DE="virtual reality"

2.DE=("interpersonal relations" OR "client relations" OR dating OR "family relations" OR friendship OR "homosexual relationships" OR "intergenerational relations" OR "marital relations" OR mentoring OR "opposite sex relations" OR "peer relations" OR "researcher subject relations" OR "student teacher relationship" OR "superior subordinate relationship" OR "victim offender relations") AND DE="virtual reality"

3.DE=("interpersonal relations" OR "client relations" OR dating OR "family relations" OR friendship OR "homosexual relationships" OR "intergenerational relations" OR "marital relations" OR mentoring OR "opposite sex relations" OR "peer relations" OR "researcher subject relations" OR "student teacher relationship" OR "superior subordinate relationship" OR "victim offender relations" OR relations) AND DE="virtual reality"

Relevanskriterier: Dokumenten skall beröra interaktionen mellan människor i virtuella världar.

15. The use of Artificial Intelligence in communication technology.

1. DE="information technology" AND DE="artificial intelligence"

2. DE="information technology" AND DE=("artificial intelligence" OR "expert systems")

3. DE=("information technology" OR technology) AND DE=("artificial intelligence" OR "expert systems")

Relevanskriterier: Dokumenten skall beröra användning av artificiell intelligens inom informations- och/eller kommunikationsteknologi (internet, mobiltelefoner, nätverk etc.).