Author(s):	Järvelin, Kalervo; Sormunen, Eero
Title:	Informationslagring och -återvinning
Main work:	Introduktion till informationsvetenskapen
Editor(s):	Mäkinen, Ilkka; Sandqvist, Katja
Year:	2003
Pages:	103-133
ISBN:	951-44-5626-2
Publisher:	Tampere University Press
Discipline:	Computer and information sciences
Item Type:	Article in Compiled Work
Language:	sv
URN:	URN:NBN:fi:uta-201412102390

INFORMATIONSLAGRING OCH -ÅTERVINNING

Kalervo Järvelin och Eero Sormunen

Ordet *informationsåtervinning* har flera betydelser. Det kan å ena sidan syfta på (i) ett läroämne eller forskningsområde som befinner sig någonstans vid gränsen mellan informationsvetenskap och datavetenskap. Å andra sidan kan informationsåtervinning för en professionell informatiker innebära (ii) ett informationssökningsuppdrag, (iii) den till uppdraget anknutna *sökprocessen* med dess olika delmoment, (iv) *resultatet* av sökprocessen, samt (v) den arbetsinsats som omfattar all aktivitet från och med att informationssökningsuppdraget mottagits till och med att sökresultatet överlämnats. I det här kapitlet kommer termen informationsåtervinning att syfta på dels (i) läroämnet och forskningsgrenen, dels (iii) sökprocessen. Då sökprocessen avses används även den förenklade formen *sökning*.

Kapitlets tema åskådliggörs med ett exempel på informationssökning i Internet-miljö, men samma fenomen påträffas också i andra informationsmiljöer. Att använda Internet och i synnerhet World Wide Web (WWW) och dess tjänster som exempel är dock befogat, eftersom de flesta läsare torde ha en viss uppfattning om WWW som informationssökningsmiljö.

Exemplet webbsökning

WWW har med sitt utbud av tjänster gjort informationssökning möjlig för oss alla. För att använda en sliten fras är all information i världen tillgänglig för webbanvändaren genom en enkel knapptryckning. Var och en som använder WWW-söktjänster märker dock snabbt att det nog inte alltid är så okomplicerat som talesättet låter förstå.

Låt oss anta att vår vän Per befinner sig i följande situation. Hans engelske vän Edward har efter att ha besökt Finland blivit förtjust i att bada bastu och bombarderar

nu vår vän med frågor om hur han borde gå till väga för att bygga en bastu i sin höghuslägenhet i London. Per har visserligen erfarenhet av bastubygge, men anser det klokast att skaffa ett "verktyg", dvs. engelskspråkiga bastubyggnadsinstruktioner. Sådana kan han säkert hitta på WWW. För att uttrycka det med informationsvetenskapliga termer har Edward ett *informationsbehov (information need)* som Per strävar efter att tillfredsställa genom att skaffa lämplig information.

Per har använt sig av WWW-tjänster förut och vet att de erbjuder åtskilliga alternativa sätt att söka information. En möjlighet vore att utgående från sin egen hemsida bläddra (surfa, navigera) i hypertextstrukturerna på webben, dvs. utnyttja *länkar* mellan olika dokument. Alternativet tilltalar dock inte Per, eftersom han inte kan komma på en enda *länksamling* eller *portal (directory service)* som har anknytning till ämnet och därför skulle gå att använda som utgångspunkt för bläddrandet. Därför förlitar han sig på en *söktjänst (search service)* i stället och väljer den han bäst känner till, nämligen internationella *AltaVista.*

AltaVista är en söktjänst bland många, av vilka de största erbjuder sökmöjlighet till tiotals miljoner WWW-dokument. Även om tjänsternas egenskaper varierar ifråga om täckning, återvinningskapacitet osv. grundar de sig alla på tre huvudfunktioner:

- *Kartläggning och uppsamling av material*. Sökrobotar (harvesting robots) plockar upp dokument som lagts ut på webben. Samtidigt utvidgar de automatiskt sitt område genom att också följa länkarna mellan dokumenten.
- *Uppbyggnad av databaser*. De uppsamlade dokumentens adress- och referensuppgifter lagras i en databas (data base) och orden i dokumenten lagras i databasens index.
- *Söktjänst*. Användningen av databaser baserar sig på ett informationsåtervinningssystem (information retrieval system) med vars hjälp man kan rikta sökfrågor (queries) till databasen. Vilken teckensträng som helst som ingår i ett webbdokument kan användas som *söknyckel* (search key).

I textrutan "Informationsåtervinningssystem" behandlas principerna för den tekniska tillämpningen av söktjänster, medan rutan "Lagrings- och återvinningsprocessen" redogör för söktjänstens produktions- och användningsprocess. I avsnittet "Webbsökning" återgår vi till redskap för WWW-sökning.

Per skriver *sökordet* "sauna" på AltaVistas sökformulär och begär engelskspråkiga dokument som svar. Ett ögonblick senare svarar AltaVista att drygt 90 000 dokument

91246 matches were found.

1. Sauna
[URL: www.cedunet.com/koulut/nakkya/saunae.html]
Sauna and the Finnish. Sauna is a piece of Finnish history. Through generations the traditions are retained. In the real Finnish sauna people sweat. In...
Last modified 6-Apr-97 - page size 704 bytes - in English [Translate]

2. Sport, Sauna, Lifestyle [deutsch]
[URL: medsun08.uni-muenster.de/~bultel/menu/style.html]
Sport, Sauna, Lifestyle. Papermag. Sport. Finische Sauna. Videotext. Skigebiet Davos. Ski Central. Ski in, Ski out. Skimaps. Home | Services | Life | Job..
Last modified 25-Jun-97 - page size 2K - in English [Translate]

3. sauna.net
[URL: sauna.net/]
Sauna.net is a place for everything cool, lame and in-between that either has something to do with saunas or nothing to do with them whatsoever. The real..
Last modified 18-Dec-97 - page size 1002 bytes - in English [Translate]

4. Sauna Room
[URL: lei.net/sprawl/42123]
Sauna Room. As you enter through the glass doors you are greeted by a blast of warm moist air. Steam issues from the middle to the room where a small pool.
Last modified 30-Mar-98 - page size 1K - in English [Translate]

5. Statemachine | Reviews and interviews | Happy Endings (Sauna)
[URL: www.statemachine.com/7_review/hap_saun.htm]
Statemachine Happy Endings Grade: 4 (out of 5) I've had many varying opinions about Statemachine. The debut single "Hologram" which I thought was quite...
Last modified 24-Feb-98 - page size 6K - in English [Translate]

6. Apex Sauna & Wine Cellars Link Page
[URL: www.apex-sauna-wine.com/links.html]
Sauna & Wine Related Links Page. Below is a list of links relating to sauna, wine or other related items. When you add, you will be automatically...
Last modified 6-Mar-98 - page size 5K - in English [Translate]

7. Sauna
[URL: www.bemberg.de/english/edampf.htm]
Bemberg Saunabau. D-74336 Brackenheim. . steambath. The price range of DM 15.000 for genuine steam baths is for many customers a constraint on their..
Last modified 8-Sep-97 - page size 2K - in English [Translate]

8. The Sauna Site
[URL: www.saunasite.com/text/off-l_1.htm]
Free Trade Ruins the World Economy ? Free trade has been one of the main goals in world politics since the second world war. The common belief is that...
Last modified 17-Sep-97 - page size 11K - in English [Translate]

9. AM-FINN SAUNA COMPANY Greensboro, NC USA
[URL: www.am-finnsauna.com/]
commercial and private sales of am finn sauna steam bath
Last modified 11-Feb-98 - page size 2K - in English [Translate]

10. The Sauna Site
[URL: www.saunasite.com/text/l-height.htm]
THE HEIGHT OF LAUDE There is a common misunderstanding, a common instruction, "the law of löyly," that says that the lower laude must be...
Last modified 14-May-97 - page size 5K - in English [Translate]

Pages: **1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 [>>]

word count: sauna: 245773

Bild 1
AltaVistas sökformulär och början av resultatlistan, länkarna 1–10

har hittats på sökordet "sauna". Per börjar undersöka sitt sökresultat på *resultatlistan* (bild 1). Allt som allt går han igenom de 20 första dokumentlänkarna.

Pers primära avsikt med genomgången är att identifiera sådana länkar som uppenbarligen erbjuder instruktioner för bastubyggnad. Han följer i mån av möjlighet varje länk till dess slut, ett *WWW-dokument*. Om sidan innehåller bastubyggnads-instruktioner skriver han ut den på papper, lagrar den som en fil på sin dators hårdskiva och/eller lagrar adressen som ett ***bokmärke*** (book mark) på sin webbläsare.

När Per ögnar igenom WWW-dokumenten och evaluerar dem med tanke på bastubyggnadsprojektet gör han en *relevansutvärdering*. Pers relevansutvärdering går ut på att jämföra webbdokumenten med Edwards informationsbehov inför byggnadsprojektet. Om ett dokument enligt Pers bedömning innehåller användbar information om bastubyggnad är det *relevant*, i annat fall är det *irrelevant*.

Per gör bland annat följande iakttagelser av sökresultatet: den första länken visar sig leda till en finsk skolpojkes föredrag om bastur och verkar onyttig, den andra leder till en tysk students hemsida och den tredje till presentationssidorna för en tysk badinrättning. Den fjärde länken verkar användbar, eftersom den leder till hemsidan för ett amerikanskt företag som marknadsför bastubyggsatser. Där kan man bl.a. hitta ritningar till olika bastumodeller. Nästa intressanta länk är den elfte länken, som består av ett amerikanskt företags basturelaterade länksamling. Denna leder bl.a. tillbaka till Finland, nämligen till Tekniska högskolans informationssidor för bastubyggare.

När Per sammanfattar de 20 första länkarna på resultatlistan kan han konstatera att:

- tre länkar leder till ytterst relevanta informationskällor,
- två länkar leder till i någon mån relevanta dokument och
- de återstående 15 länkarna är oanvändbara.

De relevanta länkarna placerar sig på resultatlistan enligt följande (bild 2):

Länkens nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevans	-	-	-	+	-	-	-	-	-	-	+	+	-	-	+	-	+	-	-	-

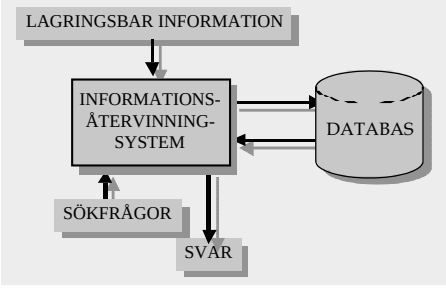
Bild 2
Relevanskontroll för Pers sökresultat

Har Pers sökning lyckats? Frågan är svår att besvara uttömmande av många olika skäl. Man kan konstatera att söktjänsten inte fungerade idealiskt, eftersom största delen av länkarna var oanvändbara. Å andra sidan är det föga ansträngande att bläddra igenom ett tjugotal länkar och skriva ut några intressanta sidor, ifall informationen kan hjälpa Edward att lösa sina problem med bastubygget. Detta kan vi dock inte avgöra, eftersom vi själva inte känner till vare sig Edwards individuella informationsbehov eller Pers förmåga att tolka dem.

Utifrån sin erfarenhet av bastubygge inser Per säkert vilken information man *i allmänhet* behöver då man skall bygga en bastu. Däremot känner han nödvändigtvis inte till Edwards informationsbehov i detalj: har Edward förmåga att utnyttja tekniska instruktioner; vilka är de lokala förutsättningarna för bastubyggnad i London och inom Edwards bostadsbolag? Pers relevansutvärdering kan endast ange vilka länkar som förmedlar *potentiellt* användbar information om bastubygge.

Informationsåtervinningssystem

Informationsåtervinningssystem är ett mycket äldre fenomen än Internet och WWW. Traditionella tillämpningar är t.ex. bibliotekens datakataloger samt referens- och textdatabaser som utvecklats för att underlätta informationssökningen inom olika ämnesområden. Nyare tillämpningar är t.ex. digitala bild-, video- och ljuddatabaser.



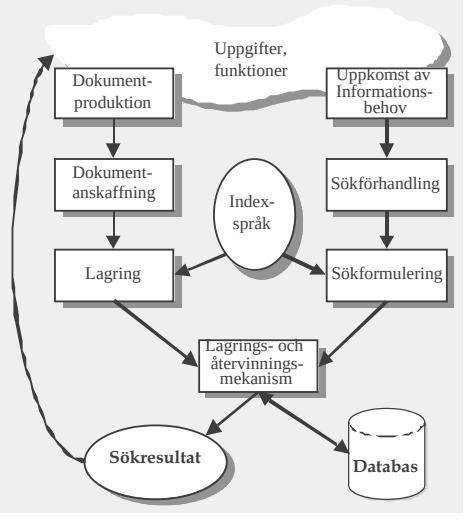
Bilden bredvid visar på vilket sätt informationen som skall lagras, frågorna sök som uttrycker informationsbehov samt svaren på dessa (=*sökresultaten)* anknyter till informationsåtervinningssystemet. Att lagra ny information i en databas samt aktualisera tidigare inmatade data kallas att *uppdatera*. Informationsbehovet formuleras som *sökfrågor* på ett *sökspråk,* vilka informationsåtervinningssystemet behandlar och ger svar på.

Systemet för lagring och återvinning av information (*information storage and retrieval system)* är ett system för informationshantering. Den förkortade benämningen *informations(återvinnings)system* används ofta i stället för hela namnet och täcker också lagringsfunktionen, ifall inte annat uppges.

Användningen av informationssystemet resulterar i en sökprocess vilken analyseras i textrutan "Lagrings- och återvinningsprocessen".

Lagrings- och återvinningsprocessen

Bilden nedan åskådliggör informationens lagrings- och återvinningsprocess. Producenter skriver och publicerar dokument (inkl. s.k. inofficiell webbpublicering). Databasproducenterna utför lagringsarbetet, dvs. skapar databaser av dokumenten eller producerar de bibliografiska uppgifter som skall beskriva dokumenten i databasen. Söktjänsterna ställer databaserna till förfogande, dvs. anskaffar eller producerar själva databaser som de tillhandahåller för användare. Telekommunikationstjänster möjliggör långdistansanvändning av databaser. Förmedlare (informationsexperter såsom bibliotekarier och informatiker) utför sökningar på användarnas vägnar. Användare som vill få tag på intressanta dokument med hjälp av databaserna söker antingen själva i dem eller ber en förmedlare utföra sökningen för deras del. Nuförtiden är användaren, i synnerhet i Internet-miljö, ofta en slutanvändare som tyr sig till en förmedlare bara i speciella situationer.



I det här kapitlet granskas bl.a. lagrings- och återvinningsmekanismer, indexspråk och framför allt informationssökning på naturligt språk samt evaluering av informationsåtervinningssystem utgående från begrepp som informationsbehov och relevans.

Relevans och evaluering av sökresultat

En jämförelse av Pers relevansutvärdering av sökresultatet med Edwards informationsbehov leder till mycket intressanta frågor: Vad menas med relevans och hur mäter man värdet på en sökning, eller mer begränsat en sökfråga? Dessa principiella frågor är viktiga vid *evalueringen av informationsåtervinningssystem* samt för *forskningen om informationsåtervinning*.

Begreppet relevans

Avsikten med informationsåtervinning är att få tag på *relevant information*. Relevans kan definieras utgående från två olika aspekter: *ämnesrelevans* och *användarrelevans* (bild 3). Ämnesrelevans innebär att dokumentet behandlar det ämne som definierats i sökfrågan (exempelvis bastubyggnad eller digital bildbehandling). En ämnesrelevansutvärdering kan t.ex. utföras av en grupp sakkunniga som ställer sökfrågan i relation till varje enskilt dokument. Det är lättare att lösa problem som rör utveckling och evaluering av informationssystem om man kan bortse från användar- och situationsbundna faktorer. I vårt exempel hade Per som sakkunnig förutsättningar att göra en ämnesrelevansutvärdering av sökresultatet.

Användarrelevansen beaktar förutom dokumentets ämne faktorer förbundna med användaren – hans eller hennes bedömning av dokumentens användbarhet. Faktorer som påverkar användarens bedömning är bl.a. sökuppgiftens natur, dokumentens språk och utseende, användarens förtrogenhet med dokumenten osv. Edward i vårt exempel kunde kanske betrakta amerikanska designinstruktioner som oanvändbara, eftersom de för honom representerar dålig smak. Hans uppfattning kunde också förändras dynamiskt: ett dokument som till en början uppfattats som relevant kunde klassificeras som oanvändbart efter att ett väsentligt bättre dokument påträffats. Kriterier av det här slaget kan det vara svårt för Per att ta i beaktande.

Ett informationsbehov definieras ofta som en användares upplevelse av oklarhet (osäkerhet) inför en situation eller omgivning. Användaren har en uppfattning om nyttan av informationen för att kunna behärska situationen (*sense making*). Att behärska situationen förutsätter insikt i den rådande situationen samt i tidigare och kommande situationer. Relevansutvärderingar är således situationsbundna och dynamiska. Användaren gör en personlig bedömning av relationen mellan den egna situationen och den information som står till förfogande. Utgående från detta kan relevans definieras på följande sätt:

> *Med relevans avses informationens uppskattade användbarhet i en särskild situation med hänsyn till användarens syften, värderingar och förväntningar.*

Under årens lopp har uppmärksamheten inom forskningen om informationsåtervinning i allt högre grad riktats från ämnesrelevans mot användarrelevans. Detta betyder dock inte att man inom forskningen nödvändigtvis alltid borde eller ens skulle kunna prioritera användarrelevans framför ämnesrelevans vid utvärdering av sökresultat. I många systemcentrerade undersökningar bildar ämnesrelevansen en tillräcklig och lättare realiserbar grund för evaluering av informationssystem. Valet av metod är beroende av det problem som skall undersökas.

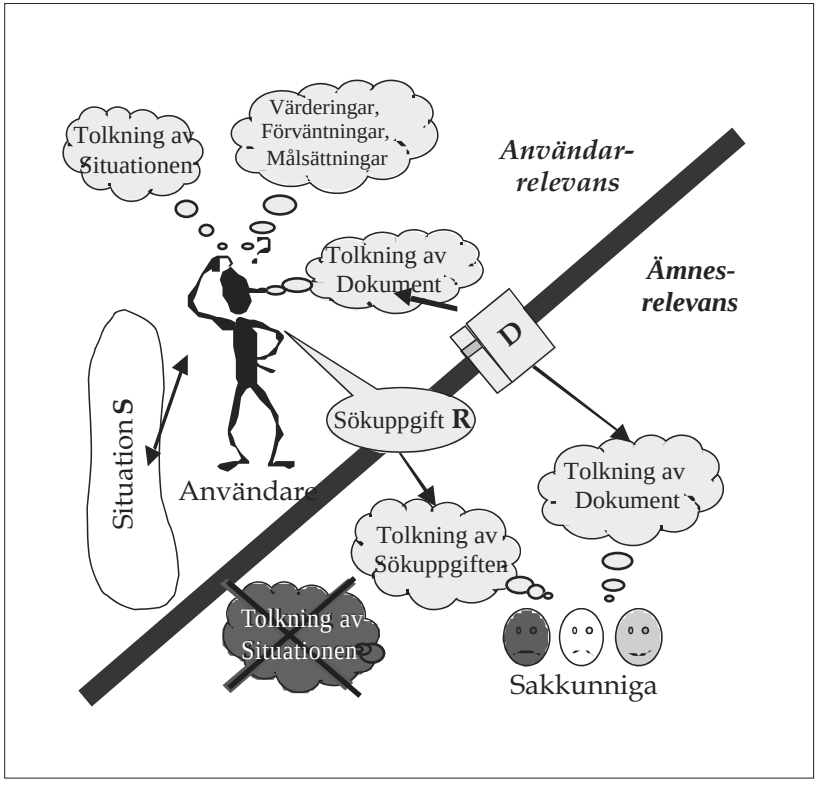


Bild 3
Ämnes- och användarrelevans

Evaluering av sökresultatet: återvinning och precision

För evaluering av sökresultat behövs utvärderingskriterier, mått enligt vilka sökresultatet kan jämföras med andra sökresultat. De vanligaste och mest använda kriterierna är *återvinning* och *precision.*

Sökresultatet indelar databasens dokument i två grupper: återfunna och icke-återfunna. Å andra sidan ger en relevansutvärdering av dokumenten i princip också upphov till en tudelning i relevanta och irrelevanta dokument. Vanligtvis är det dock bara möjligt att göra en relevansutvärdering av återfunna dokument. Förhållandet åskådliggörs i tabell 1 respektive bild 4.

Tabell 1
Definiering av återvinning och precision

	Relevansutvärdering		
Sökresultat	**Relevant**	**Irrelevant**	**Summa**
Återfunnen	**a** fullträffar	**b** brus	**a + b** återfunna
Icke-återfunnen	**c** bortglömda	**d** avvisade	**c + d** icke-återfunna
Summa	**a + c** relevanta	**b + d** irrelevanta	**a + b + c + d** databasen

Med sökresultatets *precision* avses förhållandet mellan antalet fullträffar och det totala antalet återfunna dokument, dvs. precisionen = a / (a + b). Precision uttrycks i allmänhet med ett decimaltal mellan 0 och 1 eller med ett procenttal mellan 0 och 100. Precisionen anger andelen relevanta dokument i sökresultatet.

Om samtliga av de 5 relevanta återfunna WWW-dokumenten i Pers bastubyggnadssökning beaktas blir precisionen för sökresultatet 5/20, dvs. 25%, med avseende på

de 20 dokumentlänkar som Per bläddrat igenom. I fall vi enbart godkänner de tre särskilt användbara dokumenten (som innehåller bastubyggnadsritningar) blir precisionen 3/20, dvs. 17%.

Med sökresultatets *återvinning* (recall) avses förhållandet mellan antalet fullträffar i sökresultatet och det totala antalet relevanta dokument i databasen, dvs. återvinningen = a / (a + c). Återvinning uttrycks oftast med ett decimaltal mellan 0 och 1 eller med ett procenttal mellan 0 och 100. Återvinningen anger hur stor andel av de relevanta dokumenten i databasen man lyckats återvinna med hjälp av sökfrågan.

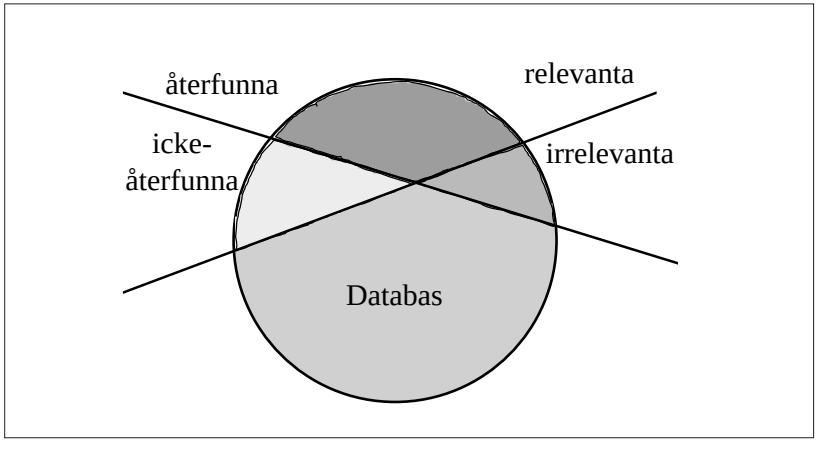


Bild 4
Relevanta och återfunna dokument

Att räkna ut återvinningen för Pers bastubyggnadssökning i databasen AltaVista är inte möjligt, såvida man inte på ett eller annat sätt tar reda på det totala antalet relevanta dokument i databasen. I praktiken är detta en i det närmaste omöjlig uppgift. I vårt exempel meddelade AltaVista att sammanlagt över 90 000 engelskspråkiga dokument innehållande ordet "sauna" påträffats. Att bläddra igenom den länksamlingen skulle uppskattningsvis kräva en månads oavbrutet arbete, om man utgår ifrån att en länk kontrolleras på ungefär 30 sekunder. Och även då kunde något relevant dokument åtminstone i princip förbli oupptäckt. Det språkidentifierande programmet

har kanske gjort ett fel, eller så existerar det ett relevant dokument där ordet "sauna" inte uppträder (bastutillverkaren har kanske använt enbart produktnamnet, "Finnsauna").

Vid beräkning av återvinning får man i praktiken oftast nöja sig med att granska sökresultatets *relativa återvinning* i relation till resultatet av andra sökningar eller till en samling relevanta dokument som erhållits på annat sätt, *återvinningsbasen (recall base)*. Man kan t.ex. ställa ett antal olika sökfrågor utifrån samma sökuppgift, skriva ut ett givet antal dokument i varje sökresultat (t.ex. de 20, 50, 100 eller 200 första), identifiera de relevanta dokumenten och på så sätt sammanställa återvinningsbasen. Resultatet av en enskild sökning kan således värderas i relation till resultatet av alla sökningar sammanlagt.

I Pers fall vet vi inte hur många länkar till dokument om bastubygge databasen AltaVista innehåller. I fall man med hjälp av olika metoder kunde få tag på sammanlagt 50 länkar skulle återvinningen för Pers sökresultat bli 5/50, dvs. 10 %. I AltaVista grundar sig jämförelsen av sökfrågan och dokumenten på sökfrågans och dokumentens *partiella överensstämmelse* och sökresultatets *relevansranking*. På grund av detta koncentreras de relevanta dokumenten till början av resultatlistan. Traditionella, s.k. booleska informationsåtervinningssystem grundar sig på *fullständig överensstämmelse* mellan dokument och sökfråga, och de relevanta dokumenten splittras slumpmässigt i sökresultatet (se "Överensstämmelse, sökfrågor och relevansranking").

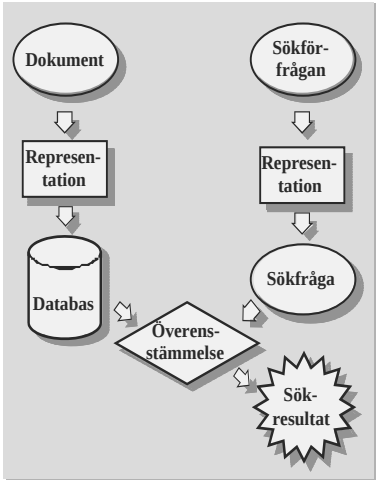
Tillsammans utgör återvinning och precision konkreta mått på hur väl en sökning lyckats. De beskriver å ena sidan den informationsmängd (i förhållande till den maximala mängden tillgänglig information) användaren erhållit, å andra sidan det arbete han/hon är tvungen att utföra för att urskilja de relevanta dokumenten i sökresultatet.

För 20 återfunna dokument var återvinningen 10% och precisionen 25% i Pers sökresultat. Per är kanske nöjd med resultatet, men det är ofta möjligt att erhålla ett ännu bättre sökresultat. Ett maximalt resultat vore en återvinning och en precision på vardera 100%, vilket skulle förutsätta att samtliga 50 relevanta länkar placerade sig bland de 50 första länkarna på resultatlistan. Att uppnå ett maximalt resultat är möjligt endast i mycket speciella och avgränsade sökuppgifter.

Företeelserna återvinning och precision har visat sig stå i motsatsförhållande till varandra: när återvinningen förbättras försämras precisionen och vice versa (se bild 5). För att uppnå en hundraprocentig återvinning i samband med en "normal" sökuppgift krävs en genomgång av en mycket lång resultatlista. Härvid sjunker

Överensstämmelse, sökfrågor och relevansranking

Kärnan i ett informationsåtervinningssystem utgörs av en matching-algoritm, som är en formell metod att beräkna överensstämmelsen mellan en sökfrågas och ett dokuments representationer. På basis av likheten bestämmer informationsåtervinningssystemet om ett enskilt dokument skall ingå i resultatlistan eller inte. Därtill fastställs eventuellt dokumentets ordningsnummer på resultatlistan. I bilden nedan åskådliggörs matching-metoden, som går ut på att dokumentens representationer i databasen jämförs med sökfrågans representation. Dokumentens alternativa representationssätt i databasen behandlas i avsnittet "Intellektuell och automatisk indexering".



Fullständig överensstämmelse

Matching-metoder som baserar sig på traditionell boolesk logik förutsätter att ett dokument uppfyller en sökfrågas logiska villkor: dokumentet bör innehålla någon av de kombinationer av söknycklar som sökfrågan tillåter. Sökfrågan "bastu OCH (design ELLER byggnad ELLER ritningar)" förutsätter att ordet "bastu" och därtill minst ett av orden "design", "byggnad" eller "ritningar" ingår i dokumentet. Med på resultatlistan kommer samtliga dokument som uppfyller de logiska villkoren. De är inte uppställda i någon inbördes ordning (relevansordning) baserad på likhet mellan dokument och sökfråga.

Partiell överensstämmelse

Många WWW-söktjänster använder sig av partiell överensstämmelse, vilket innebär att ett dokument inte nödvändigtvis behöver innehålla någon särskild kombination av söknycklar för att överensstämma med sökfrågan. De enklaste sökfrågorna är bara listor på söknycklar, t.ex. "(bastu design byggnad ritningar)". När dokumenten indexeras ges orden de innehåller en dokumentspecifik vikt som återspeglar ordets statistiska representativitet i beskrivningen av dokumentet. På motsvarande sätt kan man lägga vikt vid söknycklarna i en sökfråga. Överensstämmelsen fastställs genom att ett jämförelsetal för dokumentet beräknas utgående från vikten hos de gemensamma orden i dokumentet och sökfrågan. Med på resultatlistan kommer det antal träffar som användaren begärt av de på basis av jämförelsetalet mest relevanta dokumenten. Vid indexering av dokument och beräkning av jämförelsetal kan många olika statistisk-matematiska metoder användas. Dessa ges dock inte något utrymme i denna bok.

Relevansranking

Relevansranking innebär att sökresultatet sorteras i fallande ordning på basis av dokumentens jämförelsetal. De dokument som bäst överensstämmer med sökfrågan placerar sig sålunda i början av resultatlistan. Ofta är dessa dokument också i genomsnitt mer relevanta än de som befinner sig lägre ner på listan.

precisionen vanligtvis till mycket nära 0% i slutet av listan. En modifiering av sökfrågan eller en tillämpning av en högre utvecklad matching-algoritm kan höja resultatkurvan, men det grundläggande problemet kvarstår. Ett maximalt resultat kan inte uppnås i samband med alla sökuppgifter.

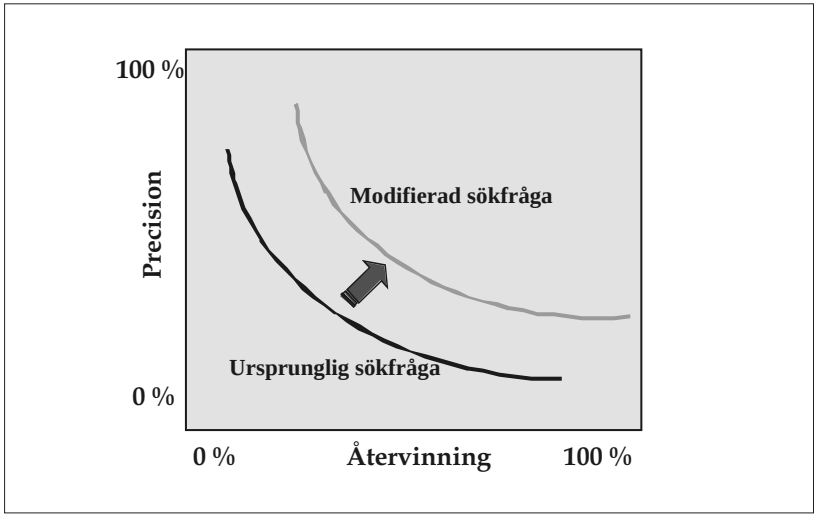


Bild 5
Evaluering av sökresultatet med avseende på återvinning och precision

Motsatsförhållandet mellan återvinning och precision samt svårigheten att uppnå ett idealiskt sökresultat beror på egenskaper hos matching-algoritmer och naturligt språk. Något förenklat kunde man säga att matching-algoritmerna inte klarar av att särskilja betydelser i texten. Vid matchningen jämförs endast förekomsten av teckensträngar ("ord") och kombinationer av dessa i dokument och sökfrågor. Till exempel ger sökfrågan "sauna AND building" träffar både för texten "The *sauna* of the hotel is in the Main *Building*..." (reklamtext för ett hotell) och för texten "This document contains detailed guidelines for *building* a *sauna*..." (bastubyggnads-instruktioner).

Nivåprincipen och naturligt språk

Nivåprincipen

Dokument och informationsbehov kan granskas på tre nivåer (bild 6). För det första är dokumenten *teckensträngar*. Förutom skrivtecken kan de innehålla multimediamaterial. Med hjälp av skrivtecken framförs *uttryck på naturligt språk*. Uttrycken på naturligt språk representerar i sin tur dokumentets *begreppsliga innehåll*. På motsvarande sätt har informationsbehovet ett begreppsligt innehåll som kan uttryckas på naturligt språk med hjälp av skrivtecken. Dokument och informationsbehov kan således granskas på följande nivåer: begrepps-, uttrycks- och teckensträngsnivå. Dessa nivåer är i princip alltid närvarande då människor granskar dokument och sökfrågor som representationer av informationsbehov. Så här förhåller det sig trots att man i informationssökningssituationer inte alltid medvetet funderar på begreppsligt innehåll eller alternativa uttryckssätt.

När en producent skapar ett dokument är det för att förmedla ett begreppsligt innehåll till läsarna. Detta begreppsliga innehåll kan struktureras på många olika sätt och ur många olika perspektiv. För att det begreppsliga innehållet skall kunna förmedlas till läsarna uttrycker producenten det med hjälp av naturligt språk (och multimedia), eftersom det inte går att förmedla ett begreppsligt innehåll direkt till andra människor. Producenten kan uttrycka sina tankar på många olika sätt, men torde välja det som han anser lämpligast för läsarna. Uttrycken klär han i skrift (och multimediaframställning). På motsvarande sätt tolkar läsarna dokumentets skrivtecken som ord och orden på ett subjektivt sätt som begreppsligt innehåll. Åtskilliga vardagssituationer visar att läsarens tolkning inte alltid överensstämmer med producentens.

Producenten presenterar kanske sitt dokument som ett WWW-dokument (HTML), som kan lagras och överföras i form av teckensträngar till läsarnas förfogande. I datanät och informationsåtervinningssystem lagras, behandlas och överförs dessa fysiska representationer av dokument, dvs. teckensträngar.

Om användaren vill få tag på ett flertal dokument som motsvarar hans informationsbehov bör han beakta följande:

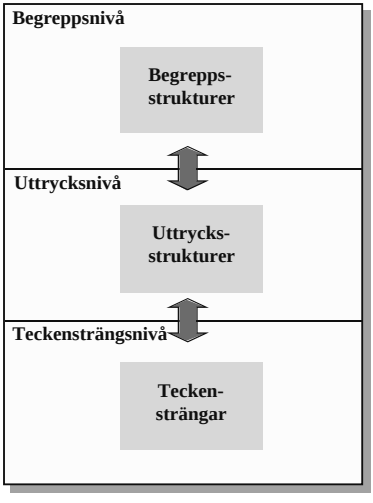


Bild 6
Nivåprincipen

- olika producenters eventuellt olika begreppsliga perspektiv i eventuellt relevanta dokument
- olika producenters eventuellt olika uttryck för använda begrepp samt
- datorns olika sätt att behandla uttryck i form av teckensträngar.

Därtill bör användaren ta hänsyn till alla väsentliga drag i sitt informationsbehov samt de begreppsliga uttrycken för dessa. Eftersom det i Pers fall är fråga om att bygga en bastu kan informationsbehovet uttryckas med minst två begrepp: [bastu] och [byggnad]. I fall man bara använder endera begreppet leder det lätt till att man får tag på irrelevanta dokument som saknar någon viktig aspekt av informationsbehovet. När Per sökte dokument om bastubyggnad med enbart söknyckeln "bastu" utelämnade han byggnadsaspekten. Detta återspeglas i sökresultatet i form av dokument där det berättas om bastur, till exempel om hur de används, men inte om hur man bygger dem. Per borde också fundera på vilka uttryck han kunde använda för begreppen bastu och byggnad.

För att kunna utföra sin sökuppgift bör användaren formulera en sökfråga med hjälp av informationssystemets sökspråk. Per använde AltaVista. Hans sökning gick att genomföra på grund av att enskilda ord i dokumenten, såsom "bastu", är sökbara i AltaVista. Om han velat inkludera begreppet byggnad (på engelska) i sin sökfråga hade den kunnat formuleras som t.ex. "sauna building". Mer komplicerade formuleringar på sökspråket behandlas senare.

Datorbaserad informationssökning sker i praktiken alltid på teckensträngsnivå, och därför är denna nivå alltid närvarande vid informationsåtervinning. Skillnaden mellan uttrycksnivå och teckensträngsnivå kan t.ex. åskådliggöras med teckensträngarna "bok", "boken" och "bokstav". Datorn uppfattar dem som tre olika teckensträngar i vilka de tre första bokstäverna är identiska. Läsaren torde uppfatta de två första teckensträngarna som samma begrepp, dvs. som två olika böjningsformer av ordet "bok", och den sista som ett ord, "bokstav", som refererar till ett helt annat begrepp. Läsaren kan uppfatta det begreppsliga sambandet mellan orden "bok" och "bokstav".

Datorer behandlar enbart teckensträngar, även om de kan programmeras att fungera *som om* de i ett begränsat avseende kunde förstå naturligt språk. Med hjälp av ett språkprogram och en ordlista skulle datorn kunna identifiera teckensträngarna "bok" och "boken" som olika former av ordet "bok" och teckensträngen "bokstav" som en form av ordet "bokstav". Däremot klarar datorn inte av en tolkning som kräver sunt förnuft: en ***bok*** innehåller sidor med tryckta ***bokstäver***. Informationsåtervinningssystemet fäster enbart uppmärksamhet vid de olika teckensträngarnas (i bästa fall även uttryckens som de ingår i) förekomst i sökfrågor och dokument, samt lägger särskilt stor vikt vid deras inbördes position och antal eller andra statistiska eller logiska egenskaper.

Det naturliga språkets egenskaper

Ett naturligt språk är en social konstruktion som genomgår ständiga förändringar och delvis är gemensam för medlemmar av samma kultur. Det är mycket omväxlande, flexibelt och rikt på uttryck. I samhället uppträder ett flertal olika språkliga subkulturer (diskurser) – precis som det uppstår subkulturer även på andra grunder. Dessa subkulturer tar sig uttryck i variationer inom såväl ordförråd och satsbyggnad som bakomliggande begrepp (se textrutan "Informationsåtervinning och naturligt språk: problem"). (Järvelin 1995). Dessa skillnader avspeglas såväl i innehållet på de

dokument som produceras inom de olika subkulturerna som i sätten att uttrycka informationsbehov.

Informationsåtervinning och naturligt språk: problem

Ambiguitet. Ambiguitet är en av språkets grundläggande egenskaper och innebär att ett stort antal uttryck bildas med ett mindre antal element. Det förekommer på både ord- och satsnivå. Till exempel är orden "anden" (fågeln) och "anden" ('övernaturligt väsen') homografer vars betydelse är beroende av sammanhanget. Varje läsare bidrar med sin egen tolkning och därmed ger varje läsning upphov till en ny tolkning.

Homonymi. Med homonymi avses identiskt uttal och/eller identisk stavning för två skilda ord. Homonymer är t.ex. "färga" och "färja" samt "får" (av verbet "få") och "får" (djuret). Homografi (identisk stavning) är en underkategori av homonymi. Polysemi är namnet på företeelsen då ett ord har två eller flera närliggande betydelsevarianter, t.ex. "blad" (bokens) och "blad" (blommans).

Synonymi. Samma begrepp kan uttryckas med många olika ord. Ett synonymt förhållande är ingalunda alltid exakt. Begreppen kan också bilda en familjeliknande grupp: de sammanfogas till en sammanhängande fläta vars ändar dock befinner sig långt ifrån varandra betydelsemässigt sett.

Sammansatta ord och ordfogningar. I sammansatta ord anger efterleden ofta huvudgrupp och förleden/-lederna undergrupp. Att identifiera efterleden är ofta viktigt. Språkbruket uppvisar också variationer mellan sammansatta ord och ordfogningar, t.ex. används både "rödvin" och "rött vin".

Prefix och suffix. Naturliga språk har många olika betydelsebärande prefix och suffix. Deras användning och betydelse varierar från språk till språk. Exempel på prefix och suffix i svenskan: för-, in-, o-, ... -nad, -het, -ig...

Ordböjning. I de flesta språk skiljer sig plural- och singularformerna från varandra. Därtill kan orden ha olika genus och böjs då efter detta. Kasusformer uttrycks antingen med böjningsändelser eller med hjälp av pre- eller postpositioner.

Avledningar. Med hjälp av avledningsändelser bildas nya ord utgående från gamla. Avledningar till ordet "grav" är t.ex. "begrava" och "begravning". Avledningens grundord framgår då man avlägsnar avledningsändelserna.

Det naturliga språkets egenskaper har en betydande inverkan på informationsåtervinningen. De bör beaktas såväl vid lagring av dokument och planering av sökstrategier som vid evaluering av sökresultat. Karakteristiska problem för t.ex. finskan och svenskan är (Alkula & Honkela 1992; Hedlund et al. 2000):

- Ordens och ordstammarnas böjning
- Sammansatta ord och fraser
- Ambiguitet, i synnerhet böjningshomografi (“anden”)
- Avledningar

Formulering av sökfrågor

I sökfrågan uttrycks ett informationsbehov på informationssystemets sökspråk. Valet av söknycklar, sökfrågans innehåll, sker efter att man övervägt olika begrepp och uttryck. Sökfrågans struktur är beroende av om man åtskiljer uttrycken för de olika begreppen i sökuppgiften strukturellt, om man åtskiljer uttrycken för en fras från de andra uttrycken samt om man betraktar de olika uttrycken som likvärdiga. Sökfrågans innehåll struktureras med hjälp av sökspråkets *operatorer.*

Vi har ovan konstaterat att Pers bastubyggnadssökning var rent ut sagt enkel och uppenbarligen inte ledde till bästa möjliga resultat. Det har småningom blivit klart att Per borde överväga:

- Fler begrepp: kunde man lägga till nya begrepp i bastubyggnadssökningen?
- Fler uttryck: kunde man uttrycka de begrepp som redan använts på ett annat sätt? Är det sannolikt att just dessa uttryck använts i relevanta dokument?
- Lämpliga sökoperatorer: räcker det med att bara räkna upp ord, eller skulle det löna sig att bilda fraser av orden eller åtskilja alternativa uttryck för samma begrepp med OR-operatorer samt olika begrepp från varandra med AND-operatorer?

Efter att ha funderat på ovanstående saker formulerade Per följande sökning: “sauna AND (drawing* OR construction* OR insulation*)”. I detta sammanhang är asterisken (*) ett s.k. *jokertecken (wild card)*, med vars hjälp alla ord som börjar på “drawing”, “construction” eller “insulation” kan inkluderas i sökningen. Efter att Per granskat sökresultatet kunde han konstatera att dokumenten 1, 2, 4, 8, 10 – 14 samt 19 var användbara. Resten var antingen oanvändbara eller svåra att bedöma. Detta åskådliggörs i bild 7.

Sålunda blev resultatets precision för 20 återfunna dokument 50%, dvs. precisionen steg med 25 procentenheter. Återvinningen höjdes också till 20 procent, dvs.

Länkens nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevans	+	+	-	+	-	-	-	+	-	+	+	+	+	+	-	-	-	-	+	-

Bild 7
Relevanskontroll för Pers nya sökning

fördubblades. Med synnerligen enkla åtgärder förbättrades resultatet väsentligt, i synnerhet vad de första återfunna dokumenten beträffade.

Intellektuell och automatisk indexering

Den mest centrala frågan vid informationslagringsprocessen är med hjälp av vilka metoder informationen i databasens index produceras. Till exempel i söktjänsten AltaVista lagras samtliga teckensträngar i ett WWW-dokument i databasens register. Det är således fråga om *fulltextindexering (full-text indexing)*, vilket innebär att dokumentet kan återvinnas utifrån vilket ord som helst som ingår i det. Fulltextindexering är en form av *automatisk indexering (automatic indexing)*.

Traditionell fulltextindexering tillämpas i booleska informationsåtervinningssystem och går till på följande sätt: (1) Indexeringsprogrammet identifierar olika teckensträngar i dokumentet. (2) För varje teckensträng bifogas adressuppgifter, såsom dokumentets identifierare, s.k. fältuppgift (förekommer teckensträngen i titeln, abstraktet, brödtexten, upphovsuppgiften osv.) samt position inom fältet (t.ex. tredje teckensträngen). (3) Teckensträngarna med adressuppgifter sorteras alfabetiskt till ett index, dvs. en *inverterad fil (inverted file)*.

Vid sökning i en fulltextindexerad databas är det möjligt att ställa mångsidiga krav på sökordens placering. Man kan t.ex. ange i vilken del av dokumentet sökordet bör förekomma (fältsökningar) samt vilken inbördes position och ordning i dokumentet sökorden bör ha (sökningar med närhetsoperatorer, frassökningar). En svaghet med traditionell fulltextindexering är att varje ord representerar dokumentet med exakt lika stor vikt (1 – förekommer; 0 – förekommer ej).

Viktkoefficient(fulltext)indexeringsmetoderna utgår ifrån att en del ord är viktigare än andra vid beskrivningen av dokumentets innehåll. Ett ords betydelsefullhet värderas ofta ur två aspekter: *termfrekvens (term frequency)* och *inverterad*

dokumentfrekvens (inverted document frequency). Termfrekvensen anger kort och gott hur många gånger ordet förekommer i dokumentet. Den inverterade dokumentfrekvensen står i proportion till ordets sällsynthet i hela databasen. *Vikten (weight)* för ett enskilt ord i ett dokument kan beräknas t.ex. med formeln

w_i = termfrekvensen x den inverterade dokumentfrekvensen

När man beräknar ett ords vikt kan formeln lämpligen formuleras så att resultatet alltid placerar sig mellan 0 och 1. Fulltextindexering baserad på viktkoefficienter har utvecklats som en del av forskningen kring algoritmer för partiell överensstämmelse, och relevansranking grundar sig på tillämpning av viktkoefficienter.

Styrkan hos de automatiska indexeringsmetoderna ligger i att de är mycket uttömmande, dvs. ett dokument är sökbart in till minsta detalj. Indexeringen är en snabb automatisk process med vars hjälp stora mängder dokument kan behandlas till låga kostnader. Ett problem är dock det mekaniska: indexeringen grundar sig på identifikation av teckensträngar i dokument samt värdering av teckensträngarna på statistiska grunder. Någon hänsyn till textens betydelser tas däremot inte.

Intellektuell indexering baserar sig på en människas innehållsanalys av ett dokument och val av termer för att beskriva det. Två nivåer av intellektuell indexering kan särskiljas. *Katalogisering (cataloguing),* dvs. beskrivning av ett dokuments yttre egenskaper, hänför sig till registrering av formella uppgifter med vars hjälp dokumentet kan identifieras (t.ex. upphovsman, titel, förläggare, tryckort, tryckår osv.). *Innehållsbeskrivningen* riktar sig till dokumentets innehåll och görs antingen på naturligt språk eller på ett specifikt dokumentationsspråk. Nedan koncentrerar vi oss på innehållsbeskrivning.

Innehållsbeskrivning på naturligt språk *(indexering med fri vokabulär)* innebär att indexeraren ur dokumentet eller sitt eget minne väljer ut de nyckelord eller -fraser han/hon anser bäst beskriva dokumentets innehåll. Fördelarna med denna typ av indexering är få, eftersom man stöter på samma problem som vid automatisk indexering vad gäller det naturliga språkets mångtydighet. Indexering med fri vokabulär begränsar sig närmast till material som inte innehåller text i elektronisk form (bl.a. tryckta publikationer, audiovisuellt material).

Formuleringen av *abstrakt* och den automatiska indexeringen av dem påminner om indexering med fri vokabulär, men är ett naturligare sätt att beskriva dokument. Fördelen med abstrakt är att användaren kan utföra sökningar på orden i dem och därtill bilda sig en uppfattning om dokumentets innehåll.

Ett dokumentationsspråk strävar efter att vara ett specialspråk för innehållsbeskrivning av dokument, fritt från mångtydighet. *Klassifikationsscheman* och *ämnesordlistor* bygger på dokumentationsspråk. Klassifikationsscheman används bl.a. för att beskriva biblioteksmaterial. De eftersträvar ofta universalitet, dvs. att omfatta samtliga ämnesområden inom vilka det publiceras dokument. Många universitetsbibliotek använder sig t.ex. av Universella decimalklassifikationen (UDK). Att analysera material utgående från ett klassifikationsschema kallas att ***klassificera***. Klasserna betecknas i regel med alfanumeriska koder, klassifikationskoder (böcker om datateknik placeras t.ex. i klassen "681.3").

Ämnesordlistor innehåller *ämnesord,* som påminner om naturligt språk. Man strävar efter att göra ämnesordens betydelse så entydig som möjligt i enlighet med dokumentationsspråket. En *tesaurus* är en ämnesordlista i vilken förhållandet mellan de olika ämnesorden är angivet på ett standardiserat sätt. Tesaurusar är i allmänhet ämnesspecifika, men även allmänna tesaurusar förekommer, som t.ex. Allmän tesaurus på svenska (ALLÄRS). Innehållsbeskrivning utgående från en ämnesordlista kallas *indexering med kontrollerad vokabulär*.

En tesaurus består av en alfabetisk och en systematisk del. I den alfabetiska delen presenteras mikrostrukturen för varje enskilt ämnesord, bl.a. hierarkiskt överordnade och underordnade termer samt övriga besläktade termer. Den systematiska delen påminner om ett klassifikationsschema, eftersom den presenterar tesaurusens hierarkiska struktur i sin helhet. Denna kan stödas av ett kodsystem med vars hjälp man lätt kan utföra sökningar på ämnesord från breda underhierarkier, t.ex. alla ämnesord som hänför sig till datateknik. Klassifikationsscheman och ämnesordlistor (speciellt tesaurusar) skiljer sig således från varandra närmast vad det yttre beträffar. Bägge är de i bästa fall systematiska och hierarkiska beskrivningar av begreppsapparaten inom ett specifikt ämnesområde.

Forskarna i informationsåtervinning förde under 1960- och 1970-talen en intensiv debatt om fördelarna med automatiska respektive intellektuella indexeringsmetoder. Slutsatsen av debatten torde vara att meningsfulla tillämpningsområden står att finna för bägge metoder:

1) *Automatisk indexering* innebär synnerligen uttömmande och detaljerad innehållsbeskrivning på kort tid och till låga kostnader. De automatiska indexeringsmetoderna är universella och kan tillämpas oberoende av ämnesområde eller dokumenttyp. Den mångtydighet och de tolkningsproblem som förknippas med

naturliga språk utgör dock en begränsning. Automatisk fulltextindexering är mest resultatrik och fördelaktig för användaren i bl.a. följande situationer:

a) Tillräckliga resurser eller redskap för intellektuell indexering saknas.
b) Tesaurusen innehåller inte ämnesord av en viss typ (till exempel sökningar på namn).
c) Tesaurusens vokabulär är på en alltför allmän nivå (specifika sökningar).
d) Tesaurusens vokabulär är föråldrad (sökningar gällande nya fenomen).

2) Styrkan hos den *intellektuella indexeringen* ligger i tesaurusar vars innehåll och struktur skräddarsytts för att tillämpas som begreppsmodell för en viss typ av informationssökning. Viktiga begrepp kan presenteras i tesaurusen med användarnas egen terminologi, oavsett vilka benämningar dokumentens upphovsmän använt. Indexeringen grundar sig emellertid på indexerarens förmåga att tolka textens betydelser, skilja mellan väsentligheter och oväsentligheter samt tillämpa tesaurusens begreppsmodell. Intellektuell indexering och tesauruskonstruktion är kostsamma företeelser, vilket innebär att möjligheterna att tillämpa metoderna alltid är begränsade. Intellektuell indexering förbättrar sökresultatet i bl.a. följande situationer:

a) Dokumenten innehåller knappt med text (samlingar av bilder, ljudupptagningar eller videoband).
b) Ämnet för sökningen är omfattande (till exempel ett helt vetenskapligt eller tekniskt område).
c) Sökning i flerspråkigt material (katalogen för ett vetenskapligt bibliotek).
d) Informationsbehovsaspekten framgår inte av texten (textens karaktär: vetenskaplig – populär).

Informationsåtervinningens forsknings- och tillämpningsområden

Den textbaserade sökningen, dvs. en sökfråga i textform som jämförs med dokument i textform, är den mest traditionella formen av informationsåtervinning. Exemplet ovan koncentrerade sig på principerna för just textbaserad sökning, även om Pers bastubyggnadssökning riktade sig till WWW, där dokumenten ofta innehåller multimedia. För det mesta utgår även multimediasökning från text, dvs. verbala

söknycklar. Så här förhåller det sig dock inte alltid. Nedan kommer vi att granska andra tillämpningsområden för informationsåtervinning. I de flesta fallen spelar den textbaserade sökningen en huvudroll, men andra möjligheter presenteras också.

Mellanspråklig informationsåtervinning

Inom mellanspråklig informationsåtervinning (Cross-Lingual Information Retrieval) strävar man efter att utveckla hjälpmedel för att behärska flerspråkighet i sökprocesser. Ett flertal dokumentsamlingar är flerspråkiga. Till exempel WWW innehåller dokument på tiotals språk. Ur användarens synvinkel vore det enklast om hans sökfråga automatiskt översattes till samtliga språk i databasen och de med sökfrågan överensstämmande dokumenten tillbaka till hans modersmål. Åtminstone än så länge klarar maskinöversättningsmetoderna inte av att översätta obegränsade mängder naturligt språk till dugliga dokument. Att översätta sökfrågor är betydligt enklare. Dessutom klarar användaren kanske av att läsa dokument på flera olika språk, även om det är enklast att formulera sökfrågor på ett enda språk, speciellt det egna modersmålet.

Sökfrågor kan översättas med hjälp av ordböcker (Ballesteros & Croft 1998; Oard & Dorr 1996; Pirkola 1998; Pirkola & al 2001) eller via en dokumentsamling bestående av parallella dokument som behandlar samma ämnen på de behövliga språken (Oard & Dorr 1996; Sheridan & Ballerini 1996). I det senare fallet erhålls lämpliga söknycklar på målspråket genom att man först söker dokument som överensstämmer med frågan på källspråket. Därefter identifieras motsvarande dokument på målspråket och slutligen plockas de statistiskt sett viktiga söknycklarna ur dessa dokument. Med de högst utvecklade översättnings- och sökmetoderna återvinns dokument på målspråket med ett nästan lika gott resultat som om man sökte direkt på detta språk.

Talbaserad informationsåtervinning

Framför allt inom radio- och televisionsverksamheten uppstår stora elektroniska arkiv bestående av talat naturligt språk. I de fall lagringen sker med hjälp av digitala metoder kan dokumenten återvinnas utifrån det talade innehållet. Härvid jämförs en sökfråga i talad form med utvalda fonetiska drag i taldokumenten (Glavitsh & Schäuble 1992; Sparck-Jones 1997).

Tal och text kan i princip kombineras på många olika sätt beroende på databasens natur och dokumentens användningssätt. Under en resa kunde man till exempel med hjälp av en mobiltelefon begära nyheter (eller andra ljuddokument) om ett intressant ämne. Om den använda databasen är en textdatabas kan sökfrågan i talad form transformeras till skrift och en vanlig textsökning utföras. De återfunna dokumenten kan i sin tur omvandlas till syntetiskt tal antingen i servern eller i mobiltelefonen. Att återvinna taldokument med hjälp av sökfrågor i textform är också möjligt (Glavitsh & Schäuble 1992). Inom en snar framtid kommer man förmodligen att kunna formulera sökfrågor i talad form för all slags multimediasökning och ta emot resultatet med en multimediatelefon.

Återvinning av musik

Inom traditionell musikbiblioteksverksamhet katalogiseras och klassificeras musikdokument (grammofonskivor, ljudkassetter, kompositioner) med hjälp av text, och dessa beskrivningar kan återvinnas med textbaserade sökmetoder. Att återvinna egentlig musik direkt är däremot inte möjligt. För musik har emellertid etablerats ett digitalt framställningssätt, MIDI-filrepresentationer, utgående från vilket det vore möjligt att utveckla sökmetoder för direkt återvinning av kompositioner. Härvid kunde även sökfrågorna utgöra musik. En dylik sökmöjlighet kan behövas av många olika orsaker, bl.a. följande:

- användarens biografiska uppgifter om en komposition kan vara bristfälliga
- han/hon känner kanske till bara en del av kompositionens melodi
- övervakning av upphovsrätten: man vill kartlägga användningen av en viss melodislinga i de olika verken i databasen.

Utgående från MIDI-filrepresentationen kan man bygga upp en förenklad representation av en komposition, vilken beskriver melodin (som torde vara det drag i en komposition man vanligen minns lättast) samt:

- tillåter sökfrågor i form av musik
- förenklar noternas och tolkningens komplicerade MIDI-representation
- tillåter dunkel sökning, dvs. sökfrågor som musikaliskt sett är "något ditåt", t.ex. delvis felaktiga med avseende på tonhöjd, tonlängd eller tonföljd. (Lemström 1998; McNab & al, 1997)

Webbsökning

Med webbsökning avses informationssökning i WWW-miljö. I början av kapitlet behandlades redan några fenomen, redskap och problem som förknippas med webbsökning. WWW karakteriseras av *fragmentariskt producerad, global och slumpmässigt sammanlänkad hypertext* där dokumenten innehåller olika former av multimedia. I och med WWW har användarna av informationssystem ökat i antal och webbsökning håller på att bli en färdighet jämförbar med läskunnighet.

Till följd av den decentraliserade dokumentproduktionen och organiseringen i hypertextstrukturer är huvudproblemet för söktjänsternas och portalernas producenter att ta reda på och hålla sig ajour med vilka resurser som finns på webben samt bestämma vilka av dessa som skall ingå i tjänstens databas. Skillnaderna mellan olika tjänster kan karakteriseras på följande sätt (Pfaffenberger 1996):

1) *Söktjänster* använder sig av autonomiskt fungerande sökrobotar, vilka genom att följa länkstrukturerna på webben samlar upp antingen enbart textdokument eller därtill även bilder och andra multimediadokument. Därefter byggs databasen upp på ungefär samma sätt som i traditionella informationsåtervinningssystem. Söktjänsterna kännetecknas av att:
   a) Perspektivet är allmänt, dvs. god täckning eftersträvas på antingen global eller lokal nivå (hela världen – Norden – Finland). Vid valet av material sker ingen utgallring med avseende på innehåll, kvalitet, ämnesområde osv.
   b) Materialet är mycket heterogent.
   c) Materialet kan återvinnas med hjälp av sökmetoder baserade på antingen sökalgoritmer för partiell överensstämmelse eller kombinationer av dylika och boolesk sökning.
   d) Med hjälp av en sökfråga kan man återvinna enskilda intressanta dokument eller få tips om mer omfattande helheter av webbsidor som man kan bläddra igenom.

2) *Portaler* är ofta inriktade på kartläggning av mer omfattande resurser än enskilda dokument, t.ex. på sökning av tillhandahållare av tjänster eller mer omfattande helheter av webbsidor med hjälp av manuella metoder. En portal kan vara en med begränsade resurser uppbyggd *länksamling,* som upprätthålls av en aktiv användare, eller en omfattande universell tjänst såsom *Yahoo!*. Utmärkande för portalen är:

a) För det mesta är portalerna specialiserade med avseende på ämnesområde och målgrupp. Vid valet av material kan urvalsprinciper beträffande innehåll och kvalitet tillämpas.
b) Materialet är intellektuellt indexerat och organiserat i hierarkiska menystrukturer. Sökfrågor kan ofta riktas även till menyn.
c) Det är i allmänhet lättare att få tag på mer omfattande materialenheter via en portal än med hjälp av en söktjänst.

3) Avsikten med *kombinationssöktjänster* är att tillhandahålla såväl en allmän söktjänst som en på intellektuellt urval baserad portal som en enda helhet.

4) *Metasöktjänster* är förmedlingstjänster som skickar användarens sökfråga till ett antal olika söktjänster och portaler samtidigt och förmedlar resultatet antingen som sådant eller t.ex. befriat från duplikat.

Webbsökning har hittills utforskats ganska knappt, ur såväl användarens som tjänsternas perspektiv. Förutom nättjänsternas snabba utveckling utgör även informationsmiljöns heterogenitet ett problem. Forskaren klarar inte av att kontrollera webbresurserna, och det ständigt föränderliga forskningsobjektet förorsakar metodiska problem. Å andra sidan skiljer sig webbmiljön tydligt från traditionella informationsåtervinningssystem i det avseendet att användaren under en webbsökning inte nödvändigtvis är verksam i bara en sökmiljö (t.ex. en söktjänst). Efter att ha utfört sökningen kan användaren bläddra igenom dokumenten eller navigera mellan dem. Att mäta resultatet är också problematiskt, eftersom man bör klara av att jämföra informationsbärare av olika karaktär: dokument eller mer omfattande informationsresurser.

Återvinning av bilder och videomaterial

Bild- och videoåtervinning är ett område inom vilket det är meningsfullt att tillämpa verbala söknycklar parallellt med indexerings- och sökmetoder som hänför sig till själva det visuella objektet. I bild 8 ser vi ett exempel på ett nyhetsfotografi med tillhörande dokumentbeskrivning (i textform). I varje nummer av en större dagstidning publiceras tiotals fotografier som fångats upp ur den dagliga strömmen av nyhetsbilder eller plockats ur tidningens bildarkiv. Bildbyrån har förutom själva bilden även producerat textuppgifterna, vilka tidningsredaktören också behöver då han väljer ut bilder.

NAMN: j291341k.jpg
HBG02 ; A Mercedes A-Class car bounces dangerously during a slalom test carried out by the German car magazine "Autobild" in Lemwerder near Bremen, northern Germany, 27 October. An example of the "Baby Benz" already rolled over while executing a tight bend during technical tests in Sweden earlier this month. Mercedes Bernz said 29 October in a press conference that it would put a new steering mechanism and new tires on its A-Class car following these reported failures in crucial security tests. DPA/AUTOBILD /
EPA GERMANY-A-CLASS MERCEDES GERMANY DPA
FRI INDEXERING: mercedes-benz test, provkörning, baby-benz, a-klassen, testkörning
FOTOGRAFERINGSDATUM: 1997–10–27
FOTOGRAF: AUTOBILD
BILDKÄLLA: DPA
STAD: HAMBURG
URSPRUNGSLAND: GERMANY
PUBLIKATION: IL
PUBLICERINGSDATUM: 1997–11–10

Bild 8
Exempel på en nyhetsbild med dokumentbeskrivning
(Källa: Aamulehtis arkiv/Lehtikuva)

En digital bilddatabas kan mycket enkelt indexeras utgående från dokumentbeskrivningar i textform. Det gamla ordspråket "En bild säger mer än tusen ord" illustrerar på ett träffande sätt problemet med textindexering. En bild innehåller alltid fler detaljer och tolkningsmöjligheter än det är möjligt att producera med hjälp av verbal indexering. Det är också svårt att hitta språkliga uttryck som överensstämmer med de visuella intrycken. Många av bildens egenskaper lönar det sig bäst att kontrollera genom att titta på bilden.

Som redskap för visuell sökning har utvecklats *egenskapsbaserade (feature based, content based) indexerings- och sökmetoder* som baserar sig på digital bildbehandling. De grundar sig på en jämförelse av överensstämmelsen mellan en modellbild som utgör sökfråga och bilderna i databasen. Som sökresultat erhålls bilder ur databasen ordnade enligt visuell likhet. Beräknandet av den visuella likheten sker med hjälp av *egenskapsvektorer (feature vectors)*. Egenskapsvektorerna mäter

t.ex. bildens färginnehåll, egenskaper i ytstrukturen (texture) samt identifierbara former.

De egenskapsbaserade algoritmerna klarar för närvarande endast av att identifiera drag i bilden på en mycket låg abstraktionsnivå. Att identifiera olika objekttyper – för att inte tala om att återvinna bilder på en namngiven person utgående från ett modellfotografi – är fortfarande en övermäktig uppgift. Enligt algoritmen kan en katt se likadan ut som en hund eller en kudde, medan en annan katt i en annan ställning i sin tur kan påminna om en ko eller en stol. Algoritmerna förefaller dock användbara t.ex. vid sortering av resultatet av en textsökning i visuellt likartade grupper. Detta är en intressant möjlighet, eftersom användarna behöver effektivare redskap för bläddring. Bläddringsmöjligheterna är en väsentlig del av bildåtervinningssystemet, eftersom direkt observationsmöjlighet av bilder är nödvändig och samtidigt det effektivaste sättet att identifiera lämpliga bilder (Markkula & Sormunen 1998).

Digital bildbehandlingsteknik har tillämpats för återvinning av videomaterial, bl.a. för att identifiera tagningar (shots) och rörelser, samt för att identifiera de bildrutor (frames) som en tagning består av. På bildrutenivå blir videoåtervinningen åter en fråga om egenskapsbaserad indexering och återvinning. Å andra sidan innehåller en video i regel även textbeskrivningar samt olika former av ljud: tal, musik, ljudeffekter och bakgrundsljud (brus). En digital video är således ett material på vilket man i princip parallellt kan tillämpa samtliga indexerings- och sökmetoder som utvecklats för olika typer av multimedia (Maybury 1997).

Mångsidig informationsåtervinning

Att vidareutveckla informationsåtervinningen kräver kunskap inom många olika områden (bild 9). Naturligt språk (skrivet eller talat) innebär en utmaning för såväl användaren och hans/hennes fantasi som för systemen och deras kapacitet. Dokumenten i sig kan förmedla mångsidig information avsedd för många olika ändamål. Detta inverkar på innehållet, utseendet och strukturen, samt i sista hand även på resultatet av sökningen. Informationstekniken har genererat en brokig mängd metoder för representation och jämförelse av sökfrågor och dokument (i form av text och andra medier). Olika informationsmiljöer, såsom ett öppet nätverk, organisationers interna databaser eller databaser inom specialområden, skiljer sig från varandra tekniskt sett, vilket innebär att dokumenten väljs ut och skyddas på olika sätt. Informationsmiljön reglerar valet av dokument samt påverkar dokumentens representations- och

användningssätt. I olika arbetsmiljöer verkar också synnerligen heterogena användargrupper, vars informationsbehov, informationsbeteende och informationssökningsfärdigheter varierar. Alla dessa omständigheter bör beaktas då man utvecklar informationsåtervinningen. Därför erbjuder området också möjligheter för såväl humanister och samhällsvetare som för personer som intresserar sig för informationsteknik – förutsatt att de också är intresserade av att samarbeta med de övriga grupperna.

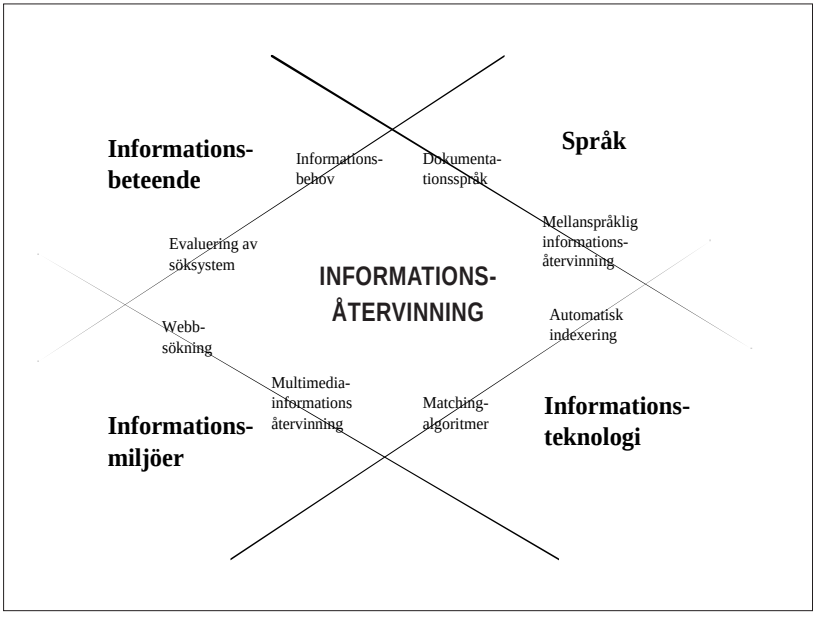


Bild 9
Informationsåtervinningens många förgreningar

Litteratur

Alkula, R. & Honkela, T. (1992). Tekstin tallennus- ja hakumenetelmien kehittäminen suomen kielen tulkintaohjelmien avulla. Espoo: (Valtion teknillinen tutkimuskeskus, Julkaisuja 765.)

Ballesteros, L. & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In: Croft, W. B. & Moffat, A. & van Rijsbergen, C.J. & Wilkinson, R. & Zobel, J. (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '98), Melbourne, Australia, August 24–28, 1998. New York, NY: ACM Press. S. 64–71.

Buckland, M. (1991). Information and information systems. New York, NY: Praeger.

Glavitsh, U. & Schäuble, P. (1992). A system for retrieving speech documents. I: Belkin, N. & Ingwersen, P. & Mark Pejtersen, A.-L. (eds.), Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special Issue of SIGIR Forum, June 21–24, 1992. S.168–176.

Hedlund, T., Pirkola, A. & Järvelin, K. (2000). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. – Information Processing and Management, 37(1):147–161.

Ingwersen, P. (1992). Information retrieval interaction. London, Taylor Graham.

Jones, G.J.F. & al. (1996). Retrieving spoken documents by combining multiple index sources. I: Frei, H.P. & al. (Eds.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '96), Zürich, Switzerland, August 18–22, 1996. New York, NY: ACM, 1996. S. 30–38.

Järvelin, K. (1995). Tekstitiedonhaku tietokannoista. Espoo: Suomen ATK-kustannus. (Asiantuntija-sarja: Tiedonhaku.)

Kekäläinen, J. & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. I: Croft, W. B. & Moffat, A. & van Rijsbergen, C.J. & Wilkinson, R. & Zobel, J. (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '98), Melbourne, Australia, August 24–28, 1998. New York, NY: ACM Press. S. 130–137.

Lemström, K. & Laine, P. (1998). Musical information retrieval using musical parameters. I: Proceedings of the 1998 International Computer Music Conference (ICMC 98), Ann Arbor, USA, October 1–6, 1998. S. 341–348.

Markkula, M. & Sormunen, E. (1998). Searching for photos – journalistic practices in pictorial IR. I: Eakins, J.P., et al. (Eds), The Challenge of Image Retrieval: A Workshop and Symposium on Image Retrieval. University of Northumbria at Newcastle, Newcastle upon Tyne, 5–6 Feb 1998.

Maybury, M.T. (1997). Intelligent multimedia information retrieval. Menlo Park, AAAI Press/ The MIT Press.

McNab, R. J., Smith, L. A., Bainbridge, D. & Witten, I. H. (1997). The New Zealand Digital Library MELody inDEX. D-Lib Magazine, May 1997. URL: http://www5.cnri.reston.va.us/dlib/may97/meldex/05witten.html

Oard, D. & Dorr, B. (1996). A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Pfaffenberger, B. (1996). Web search strategies. MIS Press.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. I: Croft, W. B. & Moffat, A. & van Rijsbergen, C.J. & Wilkinson, R. & Zobel, J. (Eds.), Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '98), Melbourne, Australia, August 24–28, 1998. New York, NY: ACM Press. S. 55–63.

Pirkola, A. & Hedlund, T. & Keskustalo, H. & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. – Information Retrieval 4(3/4): 209–230.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MA: Addison-Wesley.

Sheridan, P. & Ballerini, J. (1996). Experiments in multilingual information retrieval using SPIDER system. I: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland. S. 58–65.