

Den osynliga webben

KONSTEN ATT HITTA DET DU INTE VISSTE FANNS

På bara några få år har internet blivit så stort att det är omöjligt att indexera alla sidor. De flesta har säkert upplevt frustrationen över att inte hitta uppgifter som man vet måste finnas därute någonstans, eller känslan av att fakta man hittat bara är skrap på ytan. För att hitta på den gömda delen av webben krävs kunskaper, och i denna artikel får du lära dig många nyttiga tips.

Omkring 85 procent av alla surfare använder söktjänster när de ska söka information på internet, enligt studier, och flera av söktjänsterna hör ständigt till de tio mest besökta webbplatserna på nätet. Frågan är hur många av dessa användare av söktjänster som egentligen vet vad de söker på.

Faktum är att varje söktjänst bara täcker en liten del av internet, eller till och med en bråkdel. Dessutom klarar söktjänsterna av tekniska skäl inte av att indexera en stor del av webben. Man kan därför påstå att en stor del av internet är osynlig. Hur kommer sig då detta och hur hittar man "den osynliga webben"?

EN SÖKMOTORS INDEX

Låt oss ta en av webbens mest populära söktjänster som exempel, nämligen Yahoo. Yahoo är en så kallad hybrid, det vill säga att det både är en webbkatalog och en sökmotor. Yahoos stora popularitet beror till stor del på att den är en enkel ämneskatalog där man letar sig fram i en släktträdsliknande struktur för att hitta just det specifika ämne man själv är intresserad av.

UTESLUTEN INFORMATION

Det finns faktiskt information som söktjänsterna utsluter av helt andra skäl än rent tekniska. I en intervju i L A Times i juli förra året påstod Kris Carpenter från söktjänsten Excite att de ignorerar en stor del av webben. Av hänsyn till besökarna helt enkelt.

Det finns även information som webbspindeln faktiskt klarar av att indexera men struntar i.

Extremt långa dokument indexerar ibland inte alls av utrymmesskäl.

Vissa ord indexerar över huvud taget inte. Oftast är det vanliga ord som "och" och "the", vilket man kan förstå. Men Greg Notess berättade vid en föreläsning att till och med ett så väsentligt ord som "online" inte indexerades av en populär söktjänst. Först efter att det påtalats gjorde söktjänsten ordet sökbar.

Vad man under respektive ämnesområde får som svar är ett antal utvalda länkar som hänvisar till intressanta webbsidor. Att dessa rekommenderade länkar bara utgör en mycket, mycket liten del av all information på internet är inte svårt att inse. Men hur är det då med sökmotorn?

Yahoo har också som sagt en så kallad sökmotor som söker i ett index som består av en mängd webbsidor insamlade från internet med hjälp av en så kallad webbspindel. Detta index är inköpt av ett företag som specialiserat sig på att bygga upp stora webbindex, nämligen Inktomi. Även Snap, HotBot och GoTo har Inktomi som index till sina sökmotorer. Vad som är okänt för många är att detta index liksom alla andra sökmotorers index bara täcker en liten del av all information på internet. Och vill man vara mer exakt så täcker de mycket mindre än så. Det är den enkla sanningen. Hur kommer sig då detta?

I juli förra året publicerades i den vetenskapliga tidskriften Nature en artikel av två forskare från NEC Research Institute – Steve Lawrence och C. Lee Giles – som fick visst genomslag i media, men betydligt större genomslag bland dem som jobbar professionellt med att producera, utvärdera och använda sökmotorer.

"De flesta söktjänsterna täcker bara en liten procent av alla 800 miljoner webbsidor som är åtkomliga offentligt", skrev New York Times. "Ingen av de elva ledande söktjänsterna på internet täcker mer än en sextondel av de 800 miljoner webbsidor som den offentligt åtkomliga webben uppskattas bestå av", skrev bland annat Boston Globe.

NORTHERN LIGHT TÄCKER MEST

Lawrence och Giles studie utvärderade helt enkelt hur stor del av internet som respektive sökmotor egentligen täckte. Och siffrorna var

graverande. Den söktjänst som täckte mest täckte bara 16 procent. Kan du gissa vilken det var? Det var varken Yahoo eller AltaVista, utan Northern Light, som är mer känd bland professionella användare än bland allmänheten.

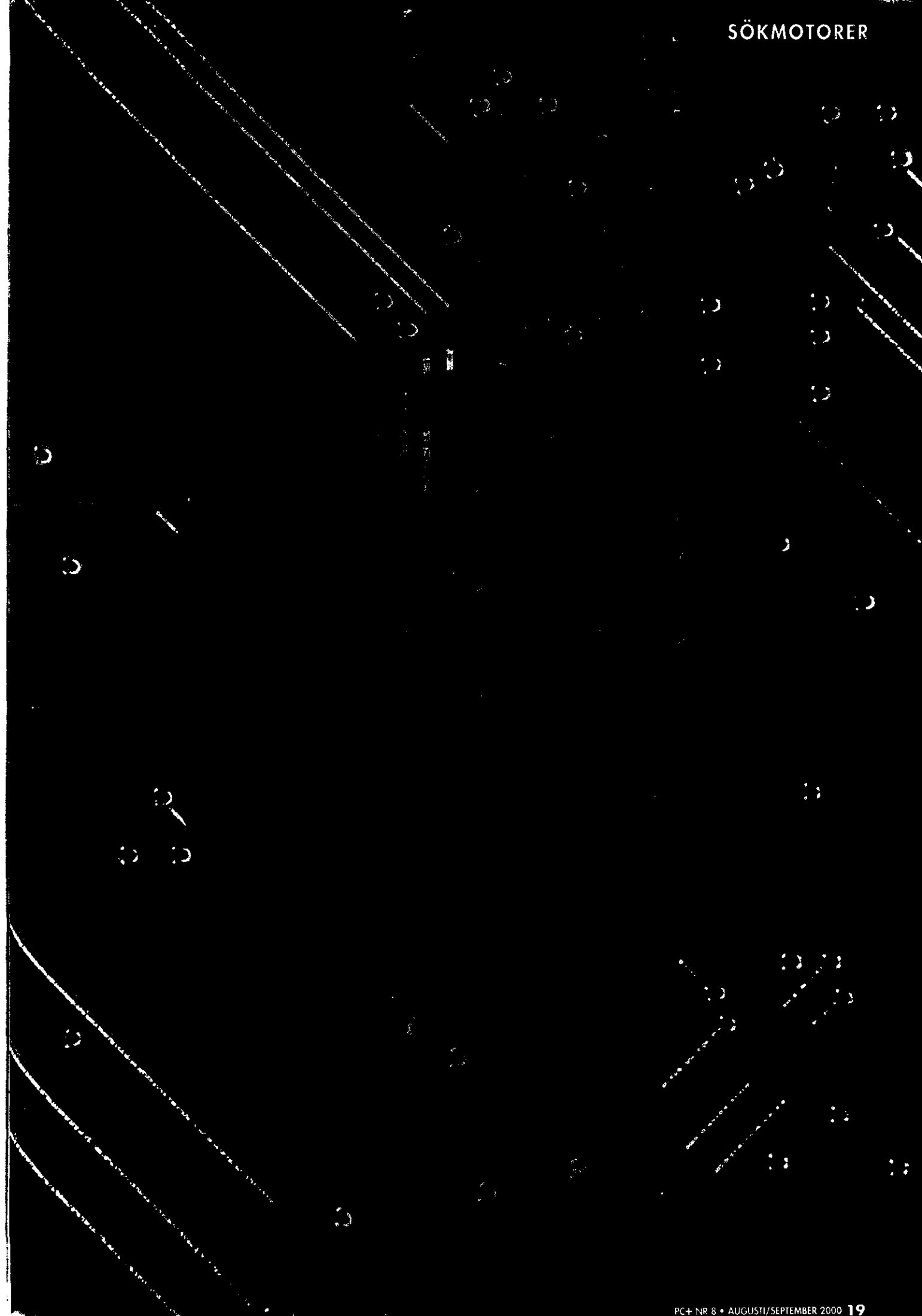
Lawrence och Giles studie byggde på 1 050 sökningar genomförda av anställda på NEC Research Institute under perioden 2–28 februari förra året. Metoden var att slumpmässigt samla in IP-adresser och därefter utsluta servrar bakom brandväggar, de som kräver rättigheter och de som saknade innehåll. Sedan lät man en webbspindel försöka fånga upp alla webbsidor på 2 500 slumpmässigt utvalda servrar.

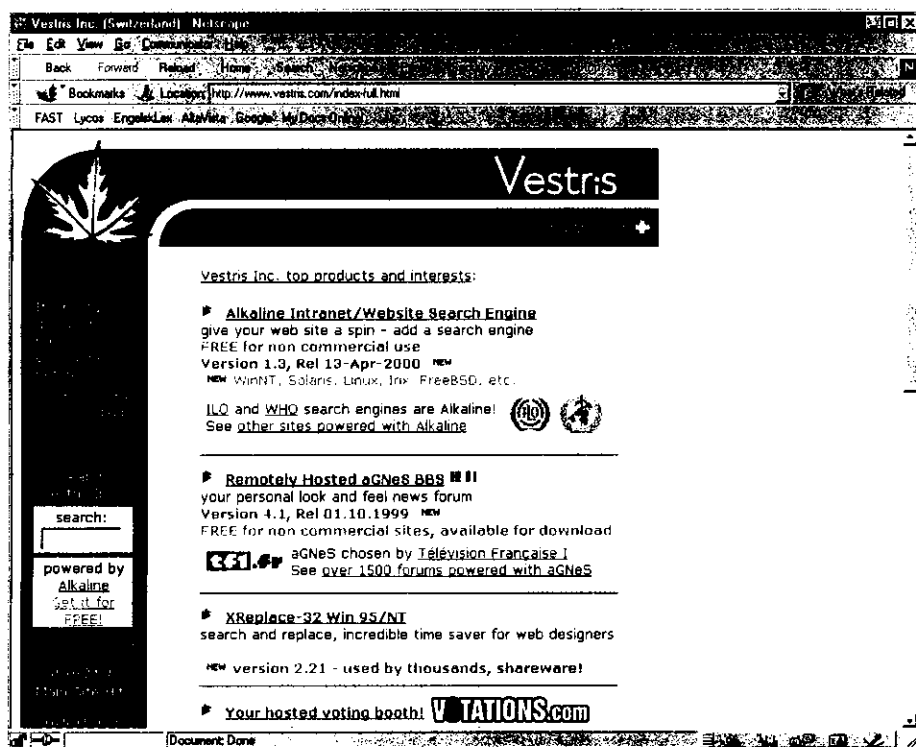
Lawrence och Giles kom också fram till att överlappningen mellan olika söktjänster var relativt låg. Det vill säga, antalet unika dokument i respektive söktjänst var betydande. Några närmare detaljer om hur de kommit fram till den slutsatsen finns tyvärr inte i studien.

Däremot kan man hitta siffror som bekräftar detta hos Greg Notess, som står bakom webbplatsen Searchengine showdown.com. I hans senaste analys från 21 februari i år visar det sig att 37 procent av 298 unika dokument bara hittades av en enda söktjänst bland 14 utvärderade.

Vad kan man dra för slutsats av liten överlappning och dålig täckning bland sökmotorerna? Man bör inte vara trogen en enda söktjänst (och inte två heller) om man vill göra mer uttömmande genomsökningar av nätet. Minst tre-fyra olika söktjänster bör man växla mellan, annars förblir en mycket stor del av webben osynlig.

Metasöktjänster då? Om man använder en metasöktjänst som söker i flera sökmotorer, och ibland webbkataloger, samtidigt? Tid sparar man säkert, men man bör vara medveten om vad för begränsningar





Webbspindeln Alkaline klarar av att indexera pdf-filer.

det innebär (se faktarutan Varning för metasöktjänster).

FAST

Hur har då sökmotorerna angripit detta problem med dålig täckning? Redan i maj förra året, precis innan Lawrence och Giles studie publicerades, lanserade det norska företaget Fast Search and Transfer en sökmotor på webbadressen www.alltheweb.com. Deras affärsstrategi är densamma som Inktomis, det vill säga att bygga upp ett index som man sedan kan sälja till andra webbtjänster. Och inte vilket

index som helst. FAST startade med ambitionen att bli det största indexet och deras intåg i världen av söktjänster drog på allvar igång det som brukar kallas "the size war", webbtäckningskriget på svenska.

Och det började bra för FAST. Om vi tittar på Greg Notess mätningar över webbtäckning gjorda 4 till 5 augusti låg FAST som klar etta. I den senaste mätningen från 21 februari ligger de också som klar etta. Endast i novembermätningen hade FAST halkat till andra plats marginellt passerad av Northern Light.

VARNING FÖR METASÖKTJÄNSTER

Många väljer att, istället för att söka i flera söktjänster var för sig, ta genvägen att använda en metasöktjänst. Det finns några saker man bör ha i åtanke då:

- 1) Metasöktjänster presenterar vanligtvis bara de tio första träffarna från respektive söktjänst.
- 2) De flesta metasöktjänster söker inte i söktjänsten Northern Light, som är en av nätets riktigt stora söktjänster.
- 3) Man missar de kringtjänster och avancerade sökmöjligheter som respektive söktjänst erbjuder.

Vad är faran med påståendet? Jo, varje söktjänst har sina olika sätt att ranka sökresultatet. En söktjänst med sämre sätt att ranka kan presentera fler ovidkommande svar bland de tio första träffarna. Med andra ord fungerar enligt min mening, eftersom det naturligtvis ligger en viss subjektiv bedömning i vilken rankingmetod man själv tycker fungerar bäst, Google bättre i metasöktjänster än många andra. I en del metasöktjänster finns faktiskt valmöjligheten att se upp till 50 träffar från respektive söktjänst.

Faran i påståendet två då? Eftersom Northern Light hör till de söktjänster som täcker störst del av webben är det synd om man inte använder den när man vill göra mer uttömmande sökningar. I Greg Notess "Unique hits report" från 21 februari i år visar det sig dessutom att av 110 unika dokument i 13 utvärderade sökmotorer fanns tio av dem i Northern Light. Meteor är den enda metasöktjänsten mig veterligen som även söker i Northern Light.

Till sist – faran i påståendet tre? Många av de avancerade kringtjänster som finns i många sökmotorer, till exempel NEAR-operatören i AltaVista avancerad sökning, går ej att använda i metasöktjänster, vilket kan vara en nackdel om du ska söka precis och uttömmande.

Vad är då argumentet för att ha störst index? Förutom att det finns ett stort intresse på marknaden är det av stor vikt att ha det största indexet som täcker den största delen av webben för att öka möjligheten att hitta unika dokument. Men det finns hakar i resonemanget och det finns även hakar i Lawrence och Giles studie.

500 MILJONER SIDOR I NYTT INDEX

Vid årets konferens, Search Engines Meetings som hålls i Boston i början av april varje år, hade man bland annat en paneldebatt där representanter för de stora sökmotorerna deltog. Eric Brewer från Inktomis berättade om Inktomis nya index GEN3, som består av 500 miljoner webbsidor. I skrivande stund finns inga tecken på att någon av Inktomis partners har bytt till GEN3.

Några veckor senare lanserar så AltaVista både ett nytt index på 350 miljoner och den nya söktjänsten Raging.com. Med detta index tar AltaVista platsen som god tvåa i webbtäckningskriget, tätt följd av FAST med 340 miljoner sidor. Detta bygger på den statistik som är inrapporterad av respektive söktjänst, och som naturligtvis måste tas med en nypa salt.

För att se de senaste siffrorna på sökmotorernas storlek kan man titta in på Danny Sullivans webbplats Searchenginewatch.com. Under rubriken "Reviews, Ratings & Tests" kan man hitta "Search engine sizes".

Helt klart är att vi bara ser början på webbtäckningskriget. Trots dessa ökning av de stora indexen består vissa problem för sökmotorerna. Det finns nämligen information på webben som webbspindlarna inte klarar av att indexera.

Här är några exempel:

1) Information endast åtkomlig genom sökförmulär. Eftersom webbspindlar använder hyperlänkar för att hitta webbsidor att indexera innebär det att de sidor som inte har en precis, publicerad länk till sig inte kommer att bli indexerad av webbspindeln. Däremot kan man naturligtvis lägga till en webbsida manuellt.

2) "Olänkade" sidor. En webbsida som inte har någon länk till sig kan inte hittas av webbspindeln och därför inte indexeras. En "olänkad" sida måste läggas till indexet manuellt. Å ena sidan kan man tycka att en sida som inte har någon länk till sig har ett lågt värde (i varje fall om

man är Google-fantast). Å andra sidan kan man undra hur stor chans det är att någon upptäcker sidan och väljer att länka till den om de inte hittar den i någon söktjänst.

3) Webbssidor bakom lösenord. Webbspindlarna kan inte indexera sidor gömda bakom lösenord. Många lösenord finns ju där för att det är en betaltjänst, men det är inte alls ovanligt att webbplatser kräver lösenord trots att all information bakom är helt gratis eller åtminstone delvis gratis. Ett sådant exempel är www.biomednet.com.

4) Webbssidor som en webmaster inte vill ska indexeras av en sökmotor. Robots Exclusion Standard är en standard för att ange vilka sidor på en webbplats som webbspindeln ska hoppa över. Med hjälp av textfilen robots.txt, som placeras i roten, kan webmastern skriva in vilka sidor och eventuella sökmotorer som ska undantas.

5) Dåligt skapade sidor med ramar kan webbspindeln ha svårt med, utan att jag här går in på hur det tekniskt kan ske.

6) Udda dokumenttyper såsom pdf, flash, word, wordperfect. En del elektroniska tidskrifter som finns gratis publicerar sina artiklar i enbart pdf-format. Till exempel Epox magazine på EpoX.nu.

UTMANINGAR FÖR SÖKMOTORERNA

Det är just här de stora utmaningarna ligger för sökmotorerna om de vill nå en verklig täckning – inte bara att öka sitt index med dagens tekniska lösningar som utesluter viss information, utan att hitta de där nya teknikerna som klarar av den alltmer dynamiska webben. Detta poängterade Knut Magne Risvik från FAST som också pratade vid tidigare nämnda konferens i Boston.

Det är med denna insikt om webbspindlarnas tekniska brister som den verkliga haken i Lawrence och Giles studie blir tydlig. Hur definierar de internet i sin studie?

Webben uppskattar de till 800 miljoner sidor. Det vill säga den delen av webben som de kallar "publicly indexable", offentligt indexeringsbar. Utan att gå in med pekpinna och titta på den exakta betydelsen i begreppet kan vi konstatera att de menar den delen av webben som är möjlig att indexera av de allmänt kända sökmotorerna med de tekniker som råder idag. Det innebär att alla sidor som inte kan indexeras på grund av de tekniska begränsningar som vi tog upp i punkterna ovan inte

HITTA DEN OSYNLIGA WEBBEN

Här är de bästa tipsen för att hitta dokument som inte är indexerade i de vanliga sökmotorerna.

DIRECT SEARCH

Ett måste för den professionella sökaren. Något grönig och designen är inte den mest användarvänliga, men innehållet är av högsta kvalitet. I motsats till allt för många andra webbplatser är det innehållet som lyfts fram och ytan som får stå tillbaka. Webbplatsen underhålls av bibliotekarien Gary Price. Direct Search innehåller kommenterade länkar till över 1000 sökbara databaser på nätet.

gwis2.circ.gwu.edu/~gprice/direct.htm

INVISIBLEWEB

En webbkatalog som kallar sig databasen över databaser. Den täcker cirka 10 000 olika källor uppdelade i 800 olika kategorier som innehåller information som sökmotorerna inte klarar av att indexera. Här är det både yta och innehåll, men så är det inget enmansjobb heller. Här jobbar ett helt team av redaktörer som skriver referat till varje utvald länk. Invisibleweb.com är den hetaste tjänsten vad gäller täckning av den osynliga webben. Genom Intellisearchs metasökprogram Bulls Eye Pro kan man söka direkt i Invisibleweb.com. En trevlig funktion är att man kan söka direkt i en databas via Invisiblewebs gränssnitt.

www.invisibleweb.com

THEBIGHUB

Den tidigare iSleuth har nu fått namnet The Big Hub. Här kan man både göra metasökningar i andra söktjänster och söka i en webbkatalog som innehåller mer än 1 500 olika ämnesinriktade databaser. Valet av söktjänster i metasökning-funktionen kan ju diskuteras, men katalogen över databaser är en guldgruva. Precis som i Invisibleweb går det att söka direkt via The Big Hubs gränssnitt utan att behöva gå in på den rekommenderade webbplatsen och göra sökningen.

www.thebighub.com

ALLT.COM

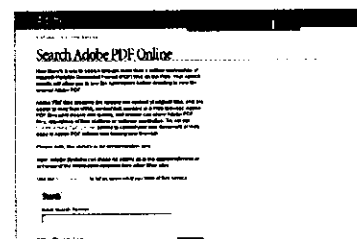
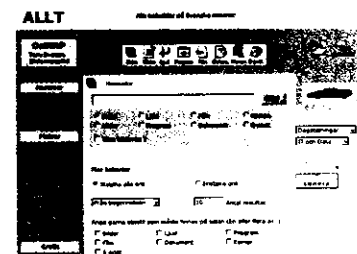
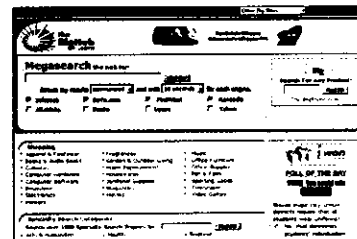
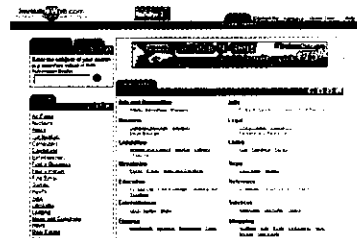
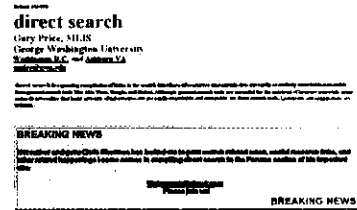
Allt.com är en svensk söktjänst som har några intressanta funktioner som gör det möjligt att söka udda dokumentformat, till exempel pdf- eller power point-filer. Sökmotorn indexerar inte hela pdf-dokument men däremot informationen på länken som pekar till en viss pdf-fil. Det går även att söka på både packade filer och programfiler av olika format.

www.allt.com

SEARCH ADOBE PDF ONLINE

Bakom det allmer populära formatet pdf står Adobe Systems. Adobe har själva en webbtjänst där man kan söka efter pdf-dokument. I motsats till Allt.com har man indexerat delar av innehållet i mer än en miljon pdf-dokument. Om siffran på över en miljon dokument stämmer kan ni ju tänka er hur stor hela den osynliga webben kan tänkas vara.

searchpdf.adobe.com/



Tabell 1: Statistik över söktjänsternas täckning och uppdatering

Söktjänst	Northern Light	Snap	AltaVista	HotBot	MSN	Infoseek	Google	Google	Yahoo	Excite	Lycos	EuroSeek	Genomsnitt
Täckning i förhållande till kombinerad täckning*	38.3	37.1	37.1	27.1	20.3	19.2	18.6	18.6	17.6	13.5	5.9	5.2	-
Täckning i förhållande till uppskattad storlek av webben	16.0	15.5	15.5	11.3	8.5	8.0	7.8	7.8	7.4	5.6	2.5	2.2	-
Döda länkar i %	9.8	2.8	6.7	2.2	2.6	5.5	7.0	7.0	2.9	2.7	14.0	2.6	5.3
Genomsnittlig tid innan uppdatering av nya sidor	141	240	166	192	194	148	-	-	235	206	174	-	186
Median-tid innan uppdatering av nya sidor	84	91	33	51	57	60	-	-	76	47	174	-	57

Så bra täcker de olika sökmotorerna webben. Källa: Lawrence, S. & Giles, C. L. "Accessibility of information on the web" in Nature 400, 107-109 (2000).

*Täckning i procent av alla söktjänsternas gemensamma täckning av webben. Den sammanlagda kombinerade täckningen hos ovan angivna söktjänster var 42 % av webben, det vill säga 335 miljoner sidor.

finns med i Lawrence och Giles uppskattning av webbens storlek.

När det pratas om täckning av webben bör man göra klart för sig om det är beräknat på den indexeringsbara webben eller den åtkomliga webben.

FÖRDUUBLAD WEBB PÅ ÅTTA MÅNADER

Nu har ju flera söktjänster avisert och genomfört betydande ökning av sina index det senaste året. Om man då betänker att sedan november 1995 har webben fördubblat sig nästan var åttonde månad, och att den indexeringsbara webben enligt en studie genomförd i början av januari detta år av Steve Lawrence i samarbete med Inktomi uppskattas till över en miljard sidor, så är ett index på 500 miljoner knapp hälften av den indexeringsbara webben. Och då utan att man räknar med tillväxten av webbsidor sedan januari i år.

Några mer exakta uppgifter om hur stor den åtkomliga webben är, grundade på någon studie och inte bara på spekulationer, finns inte mig veterligen. Vissa uppskattar att den är dubbelt så stor som den indexeringsbara webben och andra att den bara utgör en fjärdedel. Men man kan konstatera att den den osynliga webben växer stadigt då både pdf-dokument och databaslösningar ökar i popularitet.

För att återgå till Lawrence och Giles studie finns det fler hakar. En icke oviktig aspekt är antalet döda länkar som redovisas – 9,8 procent av Northern Lights index rapporterades be-

stå av döda länkar, vilket inte tagits i beräkning när man räknat ut täckningen på 16 procent. I själva verket täcker Northern Light 14,4 procent om man tar detta i beaktande.

Med döda länkar exkluderat skulle Snap täcka mer av webben än Northern Light, nämligen 15 procent.

UPPDATERING

En annan intressant aspekt är uppdateringen. När sökmotorerna indexerar webben kopierar de alltså endast en del av den då existerande webben. Men om en länk dör, till exempel för att skaparen valt att ta bort den, så blir indexet inaktuellt. Eller låt säga att informationen på en webbsida har förändrats. Detta innebär att endast en del av indexet verkligen speglar den "aktuella webben". Den andra delen får mer karaktären av ett pseudo-arkiv, som är en bild över hur webben såg ut under en viss redan passerad tid.

Förutom nämnda tekniska problem för sökmotorerna att hantera den alltmer dynamiska webben så är det just detta med att ha fräscha, uppdaterade index som är en av de stora utmaningarna för de ledande sökmotorerna.

I Lawrence och Giles studie finns även uppskattningar av hur lång tid respektive sökmotor tog på sig för att hitta nya dokument, både ett medelvärde och ett medianvärde.

Här kan man utläsa att Northern Light i genomsnitt var snabbast på att uppdatera, nämligen 141 dagar. Titlar man på medianvärdet istället så

ligger AltaVista bäst till med 33 dagar i jämförelse med Northern Lights 84. Den slutsats man kan dra är att det faktiskt ofta tar flera månader innan nya sidor indexerats och att de stora skillnaderna mellan medianvärde och medelvärde kan betyda att vissa webbplatser medvetet indexerats oftare än andra.

Den som någon gång har lagt upp egna webbsidor och sedan försökt kolla upp när sidorna finns i sökmotorernas index har säkert stött på denna problematik. Det beror oftast inte på den mänskliga faktorn att man inte klarar av att hitta sidorna i en sökmotor, utan oftast kan man skylla på den dåliga förmågan hos sökmotorerna att snabbt hitta och indexera nya dokument. Eller kanske saknar man helt enkelt en extern länk till sina sidor.

Vad kan vi då konstatera efter alla dessa analyser?

- Att försöka uppskatta antalet webbsidor på webben är tyvärr som att försöka räkna regndroppar i spöregn.

- Söktjänsterna täcker bara en mycket liten del av webben, speciellt om vi pratar om den åtkomliga webben.

- Överlappningen mellan söktjänsterna är låg vilket innebär att många unika dokument finns i respektive söktjänst.

- Inte ens hälften av den indexeringsbara webben täcktes tillsammans av alla de utvärderade söktjänsterna i Lawrence och Giles studie.

- Söktjänsterna har problem med att hålla sina index uppdaterade.

- Söktjänsterna har av tekniska orsaker svårt att indexera stora delar av den allt mer dynamiska webben.

Även om myten att allt finns på internet måste krossas är det ändå så att hur man än söker kan man aldrig vara riktigt säker på att informationen inte finns någonstans därute på nätet. Kan det låta mer utmanande?

ALTAVISTA TAR OCKSÅ TIME-OUT IBLAND

I ambitionen att presentera ett snabbt svar från en sökning kan AltaVista stoppa sökprocessen och presentera ett ofullständigt svar. Man kan alltså inte vara riktigt säker på att alla möjliga resultat finns med. Om man repeterar samma sökning flera gånger i rad kan det ge fler svar. Om man får fyra miljoner svar eller fem miljoner svar har väl knappast någon betydelse, men om man första gången får 17 svar på en specifik fråga och kort därefter får 500 svar så kan det ha stor betydelse.

Denna ofullständighet i en sökmotor och många andra kan du hitta på [searchengineshowdown.com](#) under rubriken "Inconsistencies".