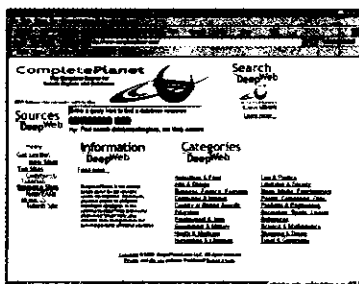


Webbtäckningskriget –är störst verkligen bäst?

Med tanke på att den populäraste aktiviteten på nätet efter e-post är söktjänster, och att det beräknas finnas drygt 350 miljoner Internetanvändare runtom i världen, är det inte svårt att inse att söktjänster är "big business". Men hur mycket klarar sökmotorerna? Och vilken är bäst?

Även om störst inte behöver betyda bäst är det ändå ett faktum att vill man vara en ledande sökmotor på Internet så kan man inte ignorera betydelsen av stora index. Index är den samling av webbsidor som respektive sökmotors webbspindel hämtat hem från Internet



Complete Planet

Completeplanet.com är ännu en värdefull söktjänst för att hitta information på den osynliga webben, eller den djupa webben som grundaren Bright Planet väljer att kalla den. Vad de syftar på är den delen av den osynliga webben som gömmer sig i databaser som genererar webbsidor dynamiskt. Med nästan 18.000 databassidor i sin webbkatalog är de en stark konkurrent till bland annat Invisibleweb.

Men i Complete Planet sker inte ämnesindelningen i webbkatalogen av den mänskliga handen utan av ett program. Däremot kommenteras en del utvalda databassidor av användare eller någon redaktör. Varje sökning ger alltid en beräkning av relevans, popularitet (bland besökarna) och länkning (från andra webbsidor).

och indexerat. Helt enkelt den informationsmängd som vi söker på i en sökmotor.

– Marknaden efterfrågar stora index, menade Eric Brewer från Inktomi när han talade på Search Engines Meetings i Boston i april i år.

Stora index ökar framför allt möjligheten vid en sökning att hitta unika dokument. Det vill säga dokument som inte går att finna i någon annan sökmotor. Men stora index ställer också högre krav på bra rankingmetoder. Med ranking menas i vilken turordning träffarna i ett sökresultat presenteras. Söker man till exempel på ordet "bilar" i en stor söktjänst kan man vara tämligen säker på att antalet träffar blir minst sexsiffrigt. Och eftersom de flesta användarna inte tittar igenom fler än de 10-20 först presenterade länkarna så förstår man hur viktigt det är att rankas högt.

Även om marknaden efterfrågar stora index lär inte den söktjänst som lågprioriterar förbättrade rankingmetoder och högprioriterar stor täckning av webben att överleva på lång sikt.

Tre miljoner nya sidor om dagen

Upprinnelsen till fokuserandet på stora index kan man söka ända tillbaka till december 1995, när AltaVista lanserades och visade upp ett index på 20 miljoner webbsidor. Vilket innebar tio gånger fler sidor än närmsta konkurrenten.

Dessutom hävdade AltaVista att de indexerade tre miljoner nya sidor per dag. I och med AltaVistas intåg var inget sig längre likt och det blev ingen lugn jul för konkurrenterna, som fick se

sig omsprungna med hästlängder.

I maj 1996 såg söktjänsten Hotbot för första gången dagens ljus och blev genast AltaVistas huvudkonkurrent genom att hävda att de kunde indexera tio miljoner nya webbsidor per dag.

Det som brukar kallas "the Size War", på svenska "webbtäckningskriget", var ett faktum.

Eftersom Internet var inne i en period av exponentiell tillväxt positionerade sig många söktjänster som den sökmotor som kunde erbjuda hela webben i sitt index. Och webbären gick.

Inte förrän i juli 1999 skulle sanningen bakom alla fina ord uppdagas, när Lawrence och Giles omtalade rapport i den vetenskapliga tidskriften Nature visade att den sökmotor som täckte störst del av internet endast täckte 16 procent av webben.

Men efter ytterligare analyser av deras resultat så visade det sig finnas en hake. Täckningen var bara beräknad på den indexeringsbara webben – det vill säga den delen av webben som dagens sökmotorer klarar av att indexera och alltså inte den hela åtkomliga webben. Det finns nämligen mängder av information och dokumentformat som inte dagens webbrobotar klarar av att indexera. (Mer om denna problematik kan man läsa i artikeln "Den osynliga webben", publicerad i PC+ nr 8 -2000).

Sökmotorernas brister i täckningen blev plötsligt mer allmänt kända och det verkade som eld i baken på de stora söktjänsterna.

Rekordsnabbt FAST

En söktjänst som inte fanns med i undersökningen var norsk och introducerades i maj 1999 av FAST transfer and



search på webbadressen www.alltheweb.com. Med samma affärsstrategi som Inktomi, det vill säga att sälja sitt index till andra söktjänster, gjorde FAST ett rekordsnabbt intåg på marknaden och gjorde tillsammans med Lawrence och Giles studie att webbtäckningskriget tog ny fart.

Redan i augusti samma år rapporterade FAST det största indexet på 200 miljoner dokument. Ett storlekstest samma månad av Greg Notess på Searchengineshowdown.com visade på ett ännu större index: 213 miljoner dokument.

Belöningen lät heller inte vänta på sig alltför länge. I januari i år valde Lycos FAST:s index för sin avancerade sökning. Mer känd under namnet Lycos Pro. Om kampen före FAST:s intåg stött

mellan AltaVista och Inktomi så skulle den nu stå mellan AltaVista, FAST och Northern Light. I januari i år var FAST först med att passera 300-miljonersvallen och i maj replikerade AltaVista med 350 miljoner. Månaden efter bröt Google 500-miljonersvallen. Webtop visade att de ville vara med i leken och presenterade ett index på 500 miljoner i juni.

Men redan i april hade Inktomi gått ut med ett pressmeddelande om att också de nu hade ett index på 500 miljoner. För första gången på över ett år var man plötsligt störst. Enligt sina egna ord vill säga.

Indexet brukar kallas GEN3 eftersom det är baserat på en patenterad sökarkitektur med samma namn. Tester visade däremot att GEN3-indexet inte alls var

lanserat hos de söktjänster som använde Inktomis index.

Censur

Men vad Inktomi egentligen sysslar med är, för att använda konkurrenten FAST:s ord, "en subtil form av censur". Inktomi har inte lagt till nya dokument i det gamla indexet på 110 miljoner för att nå upp till 500. I stället har de skapat ett andra index parallellt som arbetar tillsammans med det gamla. Om en fråga inte blir besvarad tillfredställande, först då går den vidare till det nyare indexet. Smart eller fult? Inktomi skyl- ler ifrån sig och ger ett förklarande exempel: "Om någon slår in ordet "bilar" så vill inte den personen att man söker igenom 500 miljoner sidor." Men detta tänkesätt att använda små index

vid breda sökningar behöver inte på något vis innebära att man får bättre svar bara för att man får färre svar. Det handlar om undanhållande av information som hade kunnat vara relevant för en viss frågeställning. Däremot sparar Inktomi pengar genom att slippa skaffa den hårdvara som skulle behövas om alla sökningar gjordes i hela GEN3-indexet. Kritikerna hävdar att det inte är rent spel att inte tydliggöra den strategin för besökarna.

Fem söktjänster positionerade sig själva i juni med index på minst 500 miljoner sidor: Snap, Hotbot och iWon (alla med Inktomis GEN3-index och med alla tidigare nämnda tricks det innebär) samt Google och Webtop.

I Searchengineshowdowns storlekstest från juli använder iWon drygt 70 procent av GEN-indexet, vilket gör den till den bäst täckande sökmotorn med 356 miljoner dokument. Tätt efter kommer Google med 355 miljoner dokument, AltaVista med 331 miljoner och FAST med 327 miljoner dokument.

Men baserar man det på sökmotorernas egna inrapporterade siffror är Google störst med sina 560 miljoner sidor.

Hur stor är webben?

Hur stor är webben då? NEC Research Institute i samarbete med Inktomi uppskattade den indexeringsbara webben till en miljard sidor i januari – 200 miljoner fler än i Lawrence och Giles mätning ett år tidigare.

Men den åtkomliga webben då? I juni presenterade Cyveillance.com att den åtkomliga webben består av 2,1 miljarder unika sidor. Dessutom ökar webben med mer än sju miljoner sidor varje dag. Och enligt deras prognoser kommer webben att ha fördubblats redan i början av nästa år.

Men det har gjorts andra mätningar också.

I juli kom Bright Planet med en studie som till och med påstår att den djupa webben, som de väljer att kalla den osynliga webben, innehåller nästan ofattbara 550 miljarder individuella dokument. Vad de egentligen syftar på är all information som gömmer sig på webbplatser som använder databaslösningar som genererar webbsidor dynamiskt.

De hävdar både att det är den snabbast växande kategorin av information på webben och att innehållet är åtminstone 1000-2000 gånger större än det

Relativ storlek på sökmotorernas index – ett test

500 miljoners-klubben			
	remjohan	nördvarning	fläcktyfus
Iwon(Inktomi)	3	7	11
Hotbot(Inktomi)	3	4	6
Snap(Inktomi)	7	15	29
Google	14	27	69
Webtop	0	2	1

Närmsta konkurrenterna			
	remjohan	nördvarning	fläcktyfus
Fast (alltheweb.com)	5	20	32
AltaVista	1	18	40
Northern Light	8	24	42

Sökningen utfördes den 4 september och gjordes på orden "nördvarning", "fläcktyfus" och "remjohan", alla ganska ovanliga ord. Det är ett förenklat storlekstest, men ger ändå en uppfattning om hur söktjänsternas inrapporterade siffror stämmer med verkligheten. Google är i särklass störst med flest träffar på alla tre sökorden. Northern Light har näst flest träffar på alla tre sökorden. Snap täcker bäst av Inktomi-indexen, men mycket anmärkningsvärda är Webtops siffror. Observera att klustrade träffar är inräknade.

som finns i den indexeringsbara webben. Detta påstående ligger det antagligen mycket sanning i med tanke på att mycket av den information som döljer sig i databaser har ett format som är betydligt lättare att göra sökningar på än vad som är fallet med vanliga HTML-sidor i en sökmotor. Sedan kan det ändå vara mycket skiftande kvalitet på själva innehållet – allt från bibliotekskataloger till stora databaser.

Om webben också, som Bright Planet påstår, skulle vara omkring 500 gånger större än vad de vanliga sökmotorerna kan indexera, skulle det innebära att det finns över 500 miljarder sidor med information på webben.

Högst några promille av dessa är i så fall tillgängliga via de vanliga sökmotorerna.

500 miljarder sidor i 600 databaser

Bright Planet erbjuder en nedladdningsbar söktjänst under namnet Lexibot som kan söka mot 600 databaser samtidigt, vilket påstås innebära 500 miljarder webbsidor. Bright Planet har inga ambitioner med Lexibot att skapa en söktjänst på nätet, utan de tänker sälja tekniken till andra. Kanske kan tekniken vara intressant för den verti-

kala portalmarknaden, det vill säga de specialiserade sökmotorerna.

Även om det pratas mycket om den dolda informationen i databaser så får man inte glömma alla dokumentformat som Word-filer, PDF-filer, Flash och alla olänkade sidor som lever sitt liv i det tysta på den osynliga webben då webb-spindlarna inte klarar av att indexera dem. Ämneskategoriserade databaser kan vi trots allt hitta genom förträffliga webbtjänster som Invisibleweb.com, Direct Search och även Bright Planets egen Completeplanet.com.

För att slutligen återknyta till frågan: är störst verkligen bäst? Det går inte att sticka under stol med att den söktjänst som har mest vind i seglen för tillfället är Google. I septembernumret av PCworld valdes den till bästa sökmotorn och där nämns inget om deras stora index. Som PCworld själva motiverar valet: "Google lovar att presentera webbens mest relevanta resultat – och de levererar det".

Prova själva att söka efter det tidigare nämnda webbtjänsten Direct Search i några stora söktjänster. Gå sedan över till Google och gör samma sökning så förstår ni varför de är så heta.

Lars Iselid