

Søgemuligheder i fremtiden

Søgmaskiner på Internettet

Robot, bibliotekar eller begge dele

Ved Jakob Burkard

"We are drowning in information but starved for knowledge" --John Naisbitt

Introduktion

Internettet er nu flyttet ind på bibliotekerne og er på vej til at blive endnu et værktøj i søgen efter information. Dermed er søgeværktøjer taget i brug, som er tilgængelige for fri afbenyttelse. Men hvor gode er søgemulighederne og hvad dækker de kendteste søgemaskiner som Yahoo, Lycos og AltaVista. Denne artikel giver nogle råd om, hvordan du kan evaluere søgemaskiner, og hvad det er der søges i, når du søger på Internettet

Sådan evaluerer du søgemaskiner

Hvordan fungerer en søgemaskine i praksis? Det er vigtigt at være opmærksom på at en søgemaskine på Internettet ikke søger hele Internettet i det øjeblik der trykkes på "Søg"-knappen. Det foregår med andre ord ikke i det, der i fagsprog kaldes for "real-time", men i stedet søges en stor database, dannet på baggrund af de ressourcer, som er på Internettet. Derfor er indholdet og hvordan databaserne er dannet relevante faktorer, når søgemaskinen skal evalueres [Se reference [4](#)].

Når Internet-søgmaskinerne skal opbygges er der to primære problemer forbundet med at søge og indeksere de mange Internet-ressourcer [Se reference [1](#)]. For det første er der for meget information på Internettet. Der er World Wide Web med mere end 22 millioner hjemmesider - ifølge Altavista søgemaskinen. Det er alle hjemmesider, man gennem Netscape eller Explorer kan få adgang til. Dertil skal der lægges de andre former for arkivering og kommunikation som foregår over Internettet. Det være sig Gopher, FTP-steder samt mindst 50.000 emnespecifikke nyhedsgrupper og elektroniske konferencer, hvor der hver dag postes millioner af "beskeder". Meget af denne kommunikationsform arkiveres forskellige steder på Nettet. Den anden faktor, som gør det vanskeligt at søge Internettet er den meget ringe organisering af alle disse ressourcer og det at de ikke er samlet i noget centralt katalog.

Flere biblioteker tilbyder idag lister over søgemaskiner på deres hjemmesider, og derved åbnes der op for at bibliotekarerne også kan bruge disse søgemaskiner som referenceværktøj, vejvisere, leksika eller hvilken rolle søgemaskinen passer ind i, i forhold til den konkrete opgave der skal løses. Det er vigtigt at være opmærksom på, at søgemaskiner hovedsagligt er baseret på automatisk indeksering, hvor antallet af ord i søgestrængen der matcher ord i det bestemte dokument, er den faktor der relevansbedømmer dokumentet og viser resultatet på skærmen. Denne ranking af relevante dokumenter kan i Internettets store informationsmængder virke meget anvendelig, men oftest har det den modsatte effekt, når et par tusinde dokumenter relevansbedømmes ud fra nogle ikke nærmere kendte kriterier.

For at kunne anvende søgemaskiner og specialisere sig i få udvalgte, er det nødvendigt at forstå den bagvedliggende struktur og hvad der reelt søges på. Er et godt udgangspunkt til vurdering af søgemaskiner, er de metoder, hvormed man generelt vurderer andre former for elektroniske databaser. Det er samtidig en logisk følge ikke at ensrette tankegangen på søgningen, som man gør

når man bruger kontrollerede databaser. Dokumenterne i søgemaskinerne er ikke repræsentationer for dokumenter, men de fleste gange dokumentet selv, som er indekseret uden anden intellektuel bearbejdning.

På Nettet findes der efterhånden mulighed for et selvstudie i, hvordan man bruger søgemaskinerne [Se reference [2](#), [3](#), [4](#), [5](#)].

Indhold af data

Indhold dækker generelt over hvad søgemaskinen har akkumuleret af data. Der vil sige, at følgende kriterier skal evalueres, for at vurdere søgemaskinens anvendelighed:

1. **Størrelse:** Hvor mange enkeltstående dokumenter er indekseret. Hermed menes antallet af Web-sider, postede "breve" på nyhedsgrupper m.v. Man kan også vurdere hvor mange protokoller, der indekseres (http, ftp m.v.) Antallet af dokumenter i søgemaskinen kan variere fra nogle få tusinde til millioner af dokumenter.
2. **Dybde:** Dybde i indeksering af Internetressourcerne har stor betydning for genfindingsmulighederne. Indekseringer i søgemaskiner kan variere fra den helt enkle, hvor blot hjemmesidens titler (den der ses i headeren, og level 1+2 overskrifter i dokumentet) er automatisk indekseret, og til den fuldstændige, der indekserer alle ord i dokumentet samt en intellektuel behandling af de enkelte dokumenter. F.eks. indekseres typisk:
 - Kun titlen
 - Fuld tekst
 - URL'en på dokumentet
 - Links i Web-dokumentet
 - Kombination
3. **Accessionsfrekvens:** Hvor ofte afsøges Nettet for ny information, der opdateres i databasen? Dagligt, månedligt eller?
4. **Kassationsfrekvens:** Hvor tit renses der ud i databasen. Dagligt, månedligt eller? For mange "døde" links udgør et irritationsmoment, når det helt rigtige link ikke virker.
5. **Opbygning og vedligeholdelsesmetode:** Opbygges og vedligeholdes søgemaskinen af mennesker som manuelt tilføjer poster til databasen (som f.eks. Yahoo), eller benyttes "robotter" som afsøger Nettet og returnerer sider, der skal tilføjes.
6. **Emnefokus:** Er søgemaskinen opbygget med et bestemt fokus, f.eks. at dække virksomhedssites i USA (som Newspaper), e-mail adresser på personer (som Teledanmark i Danmark), software steder (som ArchiePlex) m.v. Med andre ord spiller sprog og geografi stadig en rolle på nettet.

Brugervenlighed

Brugervenlighed er mht. betjening af Web-browseren blevet relativt standardiseret med Netscape, omend Microsoft Explorer vinder markedsandele. Samtidig er udformningen af søgefladen ikke længere bundet til det program, som det er en del af, men kan antage lige den form, som designeren måtte ønske. Der er stadig de basale faktorer, som påvirker tilgængeligheden af dataene.

1. **Brugerflade:** Hvor let er brugerfladen at betjene. Bokse, felter og logiske valgmuligheder.
2. **Hastighed:** Hvor hurtigt bliver søgningen foretaget og hvad er svartiden på søgemaskinen?

Hvor hurtig er den web-server søgemaskinen kører på?

3. **Søgemekanisme:** Hvilke søgemuligheder har brugeren af databasen.

- Bruges der f.eks. "Fuzzy og", hvilket returnerer alle dokumenter indeholdende et eller flere af søgeordene. I hvilken grad vurderes dokumenterne relevante, og hvordan?
- Er alle boolske muligheder til stede?
- Er frase-søgning mulig? Emneord?
- Trunkering. Automatisk eller selvvalg?
- Phonetisk søgemulighed.
- Exact match - skal alle termer i søgeprofilen modsvares af termer i dokumentet?
- Maksimum hits der vises

4. **Præsentation af fund:**

- Hvordan præsenteres fundene? Som Yahoo, hvor emnestrukturen som links vises i fundene, eller som rankede poster med tekst fra begyndelsen af dokumentet og link videre til originalen. Eller en mere fyldig dokumentbeskrivelse. Hvordan rankes fundene - efter termer i hele dokumentet eller efter emneord, og hvordan præsenteres dette?
- Er der mulighed for yderlig forbedring af søgeresultatet?
- Værdimærkning/kvalitetsvurdering /autoritetskontrol. Hvem vurderer kvaliteten, designet, organiseringen m.v.
- Links videre til andre søgesteder/umiddelbar mulighed for søgning i andre søgemaskiner.

5. **Hjælp:** Hvor gode hjælpefunktioner er der. Kontekstafhængige? Er der anvendelige hjælpefunktioner. Er eksemplerne forståelige.

6. **Pris:** Koster det noget at søge, at få vist resultatet m.v.

De fleste af søgemaskinerne på WWW dækker primært WWW-sider, men enkelte andre "protokoller" dækkes også. Skema 1 viser hvad et udvalg af søgemaskiner dækker og hvilke søgemuligheder de har. Ikke alle søgemaskiner er lige lette at gå til.

Et eksempel på en søgemaskine, som kan være vanskelig at bruge er Altavista. Mulighederne i Advanced Search er så avancerede, at den er vanskelig for ikke professionelle søgere. Samtidig åbner den op for virkeligt stærke søgemuligheder, hvis man forstår syntaksen. F.eks. "Host:dk"

	Andre "protokoller"	Automatisk "eller"	Brug af +/-	Bruger "eller", "og", "ikke"	Specificering af felter til søgning	Trunkering	Frase	Nærhedsso
AltaVista	Usenet	Ja	Ja	I avanceret søgning	Ja	"*"	".."	NEAR
Exite	Usenet	1)	Ja	Ja	Nej	Nej	Nej	Nej
Galaxy	Telnet Gopher	Ja		Ja	Ja	Højre "*"	Nej	Nej
Harvest	Nej	Nej		"og" "eller"	Ja	Ja	Ja	Ja
	Usenet							

Infoseek	E-mail	Ja	Ja	Nej	Nej	Nej	".."	[..]
Inktomi (HotBot)	Nej	1)	Ja	Nej	Nej	Ser bort fra alm. endelser	Nej	Nej
Lycos	Gopher FTP	Ja	Nej	Ja	Nej	Behandler søgetermer som ordstreng	Nej	Nej
Open Text	Usenet	Nej	Nej	Ja	Ja	Nej	Pr. definition en frase	NEAR
Web- crawler	Nej	Ja	Nej	Ja	Nej	Nej	".."	NEAR
Yahoo	Nej	Nej		"og" "eller"	Ja	Højre "*"	Nej	Nej

Skema 1 : Skema samarbejdet på baggrund af referencerne: [6, 7, 8]. Alle søgemaskiner har Web-sider indekseret pr. definition. Frase betyder her, at ord skal være ved siden af hinanden. Nærhedsoperator sikrer at ordene er i nærheden af hinanden, f.eks. i samme sætning.

1) Exite og Inktomi registrerer automatisk mere end et ord, og ranker efter hvor mange af søgetermerne der optræder i de fremfundne dokumenter.

+/- giver mulighed for, at termer der skal være til stede kan "bindes" med "+". Termer der ikke skal være til stede kan "smides væk" med "-"

Søgmaskinernes fremtid på biblioteket

Søgmaskinerne bliver efterhånden en del af de almindelige referenceværktøjer, som også brugerne af bibliotekerne kan forvente at blive betjent med.

Stort set alle kan bruge en database uden det store forhåndskendskab- ville nogle hævde. Jeg er tilbøjelig til at give dem ret. Men at bruge en kommerciel database eller gratis (for the time being) søgemaskine på Internettet professionelt, kræver et kendskab til det der ligger bag databasen - og i den.

Jeg håber at denne artikel har givet et overblik over, hvad der kan forventes til søgemuligheder i søgemaskiner på Internettet, og hvad det er værd at have øjnene åbne for, når man går i gang med sin søgning og valg af base. Jeg tror det er centralt at bibliotekarerne kender disse værktøjer - også på det meget avancerede niveau.

Anvendt litteratur

1. Special report: Find it on the net. - PC World, january 1996 by R. Scoville. - <http://www.pcworld.com/reprints/lycos.htm>
2. Searching the Internet: Criteria for evaluating search engines. - By Bruce Palmer. - <http://jan.ucc.nau.edu/~bwp2/isearch3.html>
3. Att värderare sökmaskiner. - By Jörgen Eriksson. - <http://www.ub2.lu.se/~jorgen/Undervisning/folkebibliotek/vaerdering.html>

4. Tips for evaluating search engines. - <http://diogenes.baylor.edu/WWWproviders/Library/BeyondLib/SrchEngEval.html>
5. Evaluating Internet sources: suggested criteria for evaluation of search engines. - <http://www.bowdoin.edu/dept/library/internet/eval/index.html#web>
6. Web search tool features. - <http://www.unn.ac.uk/features.htm>
7. World Wide Web searching tools - an evaluation. - Ian Winship. - <http://www.bubl.bath.ac.uk/BUBL/IWinship.html>
8. Web matrix: Overview matrix: Your Internet service shopping list. - <http://www.sils.umich.edu/~fprefect/matrix/overview.html>

Relateret litteratur

1. Internet search services. - By T. Koch. - <http://www.ub2.lu.se/tk/demos/DO9603-meng.html>
2. Internet ressource evaluation: A discussion of review sites. - By K.W. Wagner. - <http://www.wilpaterson.edu/home/staff/kwagner/eval.htm>
3. Internet search tool details. - <http://sunsite.berkeley.edu/Help/searchdetails.html>
4. The search engine that could. - By N. Randall. - <http://www.zdnet.com/pccomp/>
5. Robots in the Web: threat or treat?. - By M. Koster. - <http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html>

FABITA | [Konferencen Søgmuligheder i fremtiden.](#)

Sidst opdateret 2.12.96. Informationer på disse sider må benyttes såfremt kilden angives.