BOK OG BIBLIOTEK

Men trenger vi det?

Om søkemaskiner med automatisk klassifisering (clustering)

Av Even Flood, *førstebibliotekar*

Når jeg ser en forestilling med tryllekunstnere blir jeg alltid imponert over hva de får til å gjøre. Og spørsmålene kommer: Hvor hadde de egentlig den kaninen de trakk opp av hatten? Hvor kom de fra, de duene som flyr ut av ermet? Hvordan satte de sammen assistenten igjen etter å ha saget over henne? Hvordan gjør de det? Og en snikende følelse, som undertrykkes raskt, kommer også: Hva er det godt for?

Lignende følelser har jeg når jeg prøver ut en del søkemaskiner som er kommet de siste årene som forsøker med automatisk indeksering, også kalt clustering. Imponert over at det i det hele tatt går an, og hvordan gjør de det? (Og det evige spørsmål, har vi et godt norsk ord for dette? Trenger vi det? Jeg har forsøkt å bruke ordet gruppering her)

Når vi søker på internett er det to kriterier for å evaluere treffene. Det første er, er dette relevant? Handler det i det hele tatt om det jeg er ute etter? Det andre er kvalitet: Kan jeg stole på kilden, er siden oppdatert osv. Kvaliteten er det svært vanskelig for søkemaskinene å gjøre noe med, men den første, relevansvurderingen forsøker mange å automatisere. Resultatet er søkemaskiner som deler svarene inn i grupper eller clustere hvor referanser som hører sammen havner. Og det er her sammenlikningen med tryllekunstnere kommer inn, man er imponert over det man ser og at det er mulig, men er det også praktisk?

En av de første søkemaskinene som gjorde dette var Northern Light, nå dessverre lagt ned. De var en stund meget store og lå an til å kunne konkurrere med Google før de etter et par års voldsom satsing plutselig sluttet. Gruppering av resultat fungerte ofte overraskende bra. For eksempel en søking etter rennesanse musikk eller renaissance musicpå engelsk ga resultat hvor svarene var inndelt etter forskjellige komponister.

Metodene for å gjøre denne grupperingen av resultatene er mange og kompliserte, det er skrevet svært mye om det. En liten smakebit får man ved å søke i Google Scholar eller Scirus på ordene ”search” ”engines” og ”clustering”. Langt mer omfattende kan man nok få ved å søke i INSPEC på dette. Imidlertid er hensikten her ikke å gå inn i teoriene, men å se på resultatene på de søkemaskinene som har innført dette.

Mange søkemaskiner har innført dette nå. Søkemaskinene er enten relativt små eller det er metasøkemaskiner som ikke har egne databaser, men søker i flere andre søkemaskiner og bearbeider resultatene fra disse. Relativt små vil si sammenlignet med Google og de andre store (Yahoo!, MSN), men Exalead med rundt en milliard vevsider er ikke direkte liten heller.

Clustering fungerer best, til og med meget bra, når man har dokumenter som hører til

innen et begrenset fagområde og med lik struktur. Derfor kan man vente mye bedre resultater på mindre tjenester med begrenset antall dokumenter enn i den store kaotiske samling av ulike typer informasjon som verdensveven inneholder. Jeg gjorde en del tester på maskinene og oppdaget at de kan greie å skille ord som har flere betydninger i forskjellige grupper. En test jeg gjorde på alle var å søke på ordet ”marburg”, her greide søkemaskinene greit å skille informasjon om byen Marburg i Tyskland og universitetet der fra artikler om sykdommen som er oppkalt etter byen, og som nå herjer i Angola.

Vivisimo og Clusty

Så til de enkelte søkemaskinene som bruker clustering. Den mest omtalte er Vivisimo, vivisimo.com. Vivismo er firma som primært er på veven for å markedsføre programvaren sin som de har utviklet for å kunne gjøre en automatisk indeksering og inndeling. De har lagen en oversikt om hva de gjør, http://vivisimo.com/html/whyclustering. Vivisimo blir brukt av Institute of Physics for å indeksere New Journal of Physics, et open access tidsskrift. De har også lagt ut Vivisimo som egen søkemaskin, men primært sender de folk til avleggeren Clusty, www.clusty.com som gir de samme resultatene, men har flere tilleggstjenester i utskriftene og også søker gjennom nyheter. Clusty er en metasøkemaskin, de har ikke noen egen webdatabase. De søker i basene til Wisenut, Open Directory, MSN, GigaBlast, Lycos, Ask Jeeves og Looksmart blant andre (men ikke Google eller Yahoo!) for så å bearbeide dem videre i clustere. Det fungerer overraskende bra. Et eksempel jeg prøvde var å søke på søtningsmidlet aspartam som blant annet finnes i lettbrus og utallig andre produkter som skal gjøre livet lett for oss. Det foregår en intens kampanje mot dette midlet på nettet for å overbevise oss om at det er livsfarlig og roten til alt ondt. Så hvordan vil Clusty behandle noe så kontroversielt? Utrolig bra, den greide å samle en rekke sider under headingen ” Safety Of Aspartame” som alle var rapporter om at giftvirkningene er, for å si det mildt, overdrevet. Poenget her er ikke om Aspartam er sikkert eller ikke, selv om jeg har sterke synspunkter i den saken. Poenget er at Clusty faktisk greide å samle de samme sidene av kontroverset i en egen gruppe, til tross for at ord som safe og lignende forekommer i anti-sidene, kombinasjonen ”not safe” brukes mye av disse. Clusty har også en del andre finesser som det ikke er plass til å diskutere her, men grupperingene kan fungere utrolig bra.

Exalead og andre

Den franske søkemaskinen Exalead http://www.Exalead.com lager også grupper – eller noe som ligner veldig. De har som nevnt sin egen database, på noe over en milliard vevsider. Franskmennene satser en del på bli med på dette markedet og dette er en spennende nykommer. I tillegg til grupperingene søker den (i likhet med mange av de andre) også i katalogen til Open Directory, så når man søker blir det tre typer treff: Gruppene øverst til venstre, ODP treffene under og den vanlige trefflisten dominerer resten av bildet. Exalead godtar, som den eneste av de store søkemaskinene at man trunkerer. Veldig greit at man for eksempel kan skrive mullig* soup når man har glemt hvordan mulligatawny staves. Sammen med Clusty er Exalead mine favoritter av disse søkemaskinene

Killerinfo, http://www.killerinfo.com/ er en annen metasøkemaskin som søker i Google, Yahoo, MSN, FAST, Lycos, Alta Vista, Netscape, Wisenut , AOL , Overture, Looksmart og Open Directory. En del av disse overlapper en del, for eksempel er AOL basert på Google. Her er alle de største søkemaskinene brukt.

Mooter, http://www.mooter.com er australsk søkemaskin, og algoritmene for grupperingene virker litt mer primitiv enn i de andre, blant annet grupperer den sider

med samme ekstra ord i egen gruppe. Det resulterte at en søking på "marburg" ga en gruppe som bare hadde tittelen "die". Siden Marburg sykdommen er oftest dødelig var ikke dette ikke helt uvente. Men en rask sjekk viste at dette var tyskspråklige sider, og gruppen var dannet ut fra artikkelen "die" som forekommer ofte på tysk.

Iboogie, http://www.iboogie.tv/ er enda en søkemaskin som lager grupper. Den er ikke ueffen. Pussig nok er det et mysterium hvem som står bak og hvor den er henne, det har jeg ikke greid å finne.

(Se også Søkemaskiner er ikke alltid enkle)

Mer Lesning

Det er en del gode artikler om dette på veven.
Cluste: Reducing Information Overkill av Chris Sherman & Gary Price, http://searchenginewatch.com/searchday/article.php/3415071

Gary Price har også en del om det på sin blog, Blog, http://blog.searchenginewatch.com/blog/050111-115841

Google jobber med dette også, selvfølgelig, men de har ikke noe produkt på luften ennå. En kort omtale av en presentasjon fra Google: http://battellemedia.com/archives/000960.php

Et gratisprogram som søker i flere søkemaskiner og lager gruppene lokalt på egen maskin er Topgist, http://www.topgist.com. Den fungerer bare sammen med Internet Explorer, og jeg har ikke fått prøvet den ut.

En god, men foreldet artikkel er Tara Calishain: Clustering With Search Engines, http://www.llrx.com/features/clusteringsearch.htm fra juni 2002. Ellers vil et søk i de kommersielle basene gi en masse treff om dette og som nevnt gir et søk i Google Scholar på ordene search engines clustering gode resultater for dem som vil sette seg inn i teorien.

 Sist oppdatert: 28.09.05