MÖTESPLATS INFÖR FRAMTIDEN

Borås 8-9 oktober 2003

Lars Jonsson, Chalmers tekniska högskolas bibliotek

Söktjänster för akademiskt bruk
En utvärdering av Google och Argos med frågor från en akademisk ämnesdisciplin

Bakgrund

Som bakgrund till denna magisteruppsats i Biblioteks- och informationsvetenskap fanns ett intresse av att genomföra en undersökning som på något sätt kunde mäta söktjänster på webbens effektivitet i ett vetenskapligt sammanhang. Jag såg det även som intressant att göra en undersökning av ett icke-naturvetenskapligt ämne, då vetenskap på webben verkar domineras av just sådana. Valet föll på ämnet Antikens kultur och samhällsliv, och söktjänsterna Google och Argos. Antikens kultur och samhällsliv valdes för att jag själv har en viss ämneskunskap, som jag ansåg som nödvändig i arbetet med relevansbedömningarna. Söktjänsten Google valdes för att representera en stor, generell söktjänst, och Argos valdes för att den är ämnesspecifik med inriktning mot klassisk arkeologi.

Syfte och frågeställningar

Syftet med arbetet är att mäta och jämföra återvinningseffektiviteten hos två frågebaserade söktjänster, Google och Argos, med frågor från ett avgränsat akademiskt ämnesområde som utgångspunkt. Det ligger därmed alltså även i mitt intresse att undersöka i vilken utsträckning dessa typer av söktjänster klarar av att tillgodose ett informationsbehov grundat i en utbildningsmässig kontext.

De frågeställningar ja valt att utgå ifrån är:

- Vilken återvinningseffektivitet uppvisar söktjänsterna Google och Argos, med avseende på måttet *first twenty precision*?
- Hur påverkar olika definitioner av relevans resultaten för *first twenty precision*?
- Föreligger det då någon signifikant skillnad mellan resultaten om dessa statistiskt generaliseras?

Information Retrieval och begreppet relevans

Arbetet är gjort inom ramen för Information Retrieval (IR) som forskningsområde. Generellt sett kan man säga att IR handlar om representation, lagring, organisation av och åtkomst till information. Inom IR-forskningen har det har genomförts utvärderingsstudier sedan början av 1950-talet och de vanligaste måtten vid undersökningar av detta slag är r*ecall* – andelen av de relevanta dokumenten som återvunnits, och *precision* – andelen återvunna dokument som är relevanta. Man måste alltså känna till andelen relevanta dokument i samlingen. Eftersom relevans är ett sådant centralt begrepp inom IR-forskingen har jag valt att ge det ett eget avsnitt i uppsatsen, där jag utgår från att relevans är ett mått på effektiviteten av kontakten mellan en källa och en destination i en kommunikationsprocess, där det existerar flera nivåer av relevans. Mer praktiskt innebär det att relevans i uppsatsen betraktas som ett mått på till vilken nivå ett dokument tillfredsställer ett informationsbehov utifrån ett antal på förhand definierade kriterier.

Det har genomförts ett flertal utvärderingar av söktjänster på webben. Dessa skiljer sig dock på en viktig punkt från traditionella utvärderingar eftersom man här inte kan känna till antalet relevanta dokument. Som en lösning på detta har det utvecklats alternativa mått.

Metod

Jag arbetet valt att utgå från verkliga informationsbehov hämtade från Antiken på Internet (Antiken på Internet 2001) – en frågespalt skapad och underhållen av Klassiska institutionen, Antikens kultur och samhällsliv vid Göteborgs universitet. Utifrån dessa informationsbehov har sedan 30 frågor till söktjänsterna utformats.

Jag har utgått från en tidigare studie där man använt 6 olika relevanskategorier. Följande kategorier med respektive kriterier användes vid relevansbedömningen:

- **Dubbletter** - dokument med i grunden samma URL som ett dokument vilket förekom tidigare i resultatlistan hamnade i denna kategori, oavsett om det var relevant eller inte. Spegelsidor betraktades inte som dubbletter.
- **Inaktiva länkar** - 404 fel - att servern kontaktades men man kom inte fram, att tillträde till sidan var förbjudet eller att sidan flyttat, samt 603 fel - att servern inte svarade. Vid 603 fel och fel p.g.a. att tillträde var förbjudet testades dessa igen vid senare tillfälle.
- **Kategori noll** – dokumentet var icke-relevant då det inte behandlade någon aspekt av ämnet för sökformuleringen. Innehöll inte någon av söktermerna.
- **Kategori ett** – dokument som tekniskt sett tillfredställde sökformuleringen, eller innehöll söktermerna, men som inte var relevant för att frågans ämne inte behandlades alt. att ämnet behandlades för kortfattat. T.ex. om det vid en sökformulering med termerna *”doric”*, *”temples”* och *”athens”* återvanns dokument som behandlade doriska tempel i Italien, hamnade detta dokument i denna kategori.
- **Kategori två** – relevant för sökformuleringen och relevant för åtminstone någon del av informationsbehovet. Dokumenten potentiellt användbara för vissa användare. T.ex. om det vid en sökformulering som behandlade lediga tjänster vid utgrävningar, återvanns dokument som behandlade just detta, men som inte var uppdaterade, hamnade dessa dokument i denna kategori. Till denna kategori räknades även dokument med länkar till högst relevanta dokument av kategori tre-slag.
- **Kategori tre** – relevant för ett vitt antal möjliga aspekter av informationsbehovet, alltså att vilken användare som helst som ställt frågan skulle bedöma dokumentet som relevant. T.ex. en genomgående och utförlig behandling av ämnet för sökformuleringen.

Alla dokument som återvanns placerades sedan i någon av dessa kategorier. Bedömningarna var binära, antingen hamnade det i en kategori eller inte, och därmed utgör inte kategorierna olika intervall på en skala, utan istället olika definitioner av relevans.

Jag har även valt effektivitetsmåttet *first twenty precision* från en tidigare studie. Måttet mäter hur bra söktjänsterna är på att återvinna dokument bland de 20 första träffarna i en resultatlista. Måttet tar även hänsyn till var, resultatlistan bland de 20 första träffarna, ett återvunnet dokument hamnar. Detta är också ett argument för att använda detta mått, då en söktjänst som kan presentera relevanta resultat högt upp i sina resultatlistor är att föredra, och i en utvärdering därmed också bör premieras för detta.
Med anledning av att man kan definiera vad som är ett relevant och vad som är ett icke-relevant dokument på en mängd olika sätt, har jag här, för att kunna mäta och jämföra

återvinningseffektivitet vid olika definitioner av relevans, valt att genomföra fem olika tester, med fem olika definitioner av relevans.

De första tre testerna visar söktjänsternas prestanda vid olika trösklar för *precision*. Här bestraffades även söktjänsterna för eventuella dubbletter och inaktiva länkar genom att dessa sänkte summan av täljaren men inte nämnaren. Testen kan beskrivas på följande vis:

- Det första testet utgick från en låg tröskelnivå för *precision* genom att tilldela 1 till samtliga dokument som återfanns i kategorierna ett, två och tre och på så vis illustreras hur väl söktjänsterna återvann dokument som på ett minimalt sätt tillfredställde informationsbehovet.
- Det andra testet utgick från en moderat tröskelnivå genom att tilldela 1 till samtliga dokument som återfanns i kategorierna två och tre. På så vis illustreras hur väl potentiellt användbara dokument återvanns.
- Det tredje testet utgick från en hög tröskelnivå genom att endast tilldela 1 till de dokument som återfanns i kategori tre och på så vis illustreras hur väl ytterst relevanta dokument återvanns.

I det fjärde och femte testet eliminerades dubbletter och inaktiva länkar från resultatlistorna samtidigt som jag behandlade återstoden som en resultatlista bestående av färre än 20 återvunna dokument. Om en resultatlista från en söktjänst t.ex. bestod av tre dubbletter bland de 20 första träffarna, togs dubbletterna bort och nämnaren beräknades som om resultatlistan hade bestått av 17 träffar. På detta vis bestraffades inte söktjänsterna för eventuella dubbletter och inaktiva länkar i resultatlistorna eftersom både täljarens och nämnarens summa sänktes. Testerna kan mot denna bakgrund beskrivas på följande sätt:

- Det fjärde testet utgick från en låg tröskelnivå för *precision* genom att på samma sätt som det första testet tilldela 1 till dokument som återfanns i kategorierna ett, två och tre.
- Det femte testet utgick från en moderat tröskelnivå genom att på samma vis som det andra testet tilldela 1 till dokument som återfanns i kategorierna två och tre.

Resultat

Resultaten har även signifikanstestats för att undersöka om skillnaden mellan resultaten mellan de båda söktjänsterna beror på att de finns en verklig skillnad eller om det enbart kan förklaras med slumpen.

Google presterade signifikant bättre än Argos i samtliga 5 tester. En av förklaringarna till Argos sämre resultat var de många dubbletter och inaktiva länkar som söktjänsten återvann. Ytterligare en förklaring ligger i Argos begränsade sökfunktioner, samt att den inte alltid återvann 20 dokument. Googles rankingalgoritm, som bygger på hur många länkar som leder till en sida, kunde ibland innebära en begränsning, då populära sidor hamnade högre i resultatlistorna än sådana som var mer relevanta.

Slutsatser

Man kan således, med avseende på *first twenty precison* som mått, konstatera att Google uppnådde ett bättre resultat än Argos i samtliga fem tester. Vidare kan man tydligt se att olika definitioner av relevans också påverkar resultaten för *first twenty precison*. Ju högre tröskel

för relevans desto lägre blir siffrorna för resultaten. Man kan även se att Argos påverkades av detta i större utsträckning än vad Google gjorde. Man kan slutligen också konstatera, att utifrån de resultat som uppnåtts, är Google bättre än Argos på att återvinna information av vetenskaplig karaktär från den akademiska ämnesdisciplinen Antikens kultur och samhällsliv. Därmed är även Google mer lämpad än Argos att användas i denna utbildningsmässiga kontext. Påståendet att söktjänster som Google inte lämpar sig för akademiskt bruk kan alltså med denna studie som utgångspunkt förkastas. Denna studie bär dock på sina begränsningar och det skulle därför också vara av intresse att källkritiskt undersöka den återvunna informationen.