

MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP  
VID BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN  
2002:17

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

LARS JONSSON

© Lars Jonsson

Mångfaldigande och spridande av innehållet i denna uppsats  
– helt eller delvis – är förbjudet utan medgivande av författaren/författarna.

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

Engelsk titel: Search Engines for academic use: An evaluation of Google and Argos with queries from a academic discipline

Författare: Lars Jonsson

Färdigställt: 2002

Handledare: Anders Stenström, Kollegium 2

Abstract: The purpose of this MSc thesis is to examine the retrieval effectiveness of two Web search engines with queries from the academic discipline of Classical Studies. The two search engines are chosen to represent two different types – Google as a broad general search engine and Argos as a specialised search engine for the subject of Classical Studies. The search engines are compared for precision among the first twenty results returned for thirty queries. In order to avoid bias in the study, the queries are based on real users information needs. Due to the subjective character of the concept of relevance, the study is performed with five different experiments with five different definitions of relevance. The formula for calculating the metrics measures precision with weights for ranking effectiveness. Wilcoxon's signed rank test is used to control if there are any significant differences between the results. Analysis shows that Google is the top service of the two, since it performs better than Argos in all of the five experiments. The signed rank test also shows that there are significant differences between all of the results. One possible reason for Argos lower results is the many duplicate links that it returned. Other explanations discussed are Argos limited number of search facilities in comparison with Google, and that the ranking algorithm Argos uses is not advanced enough for Information Retrieval in this context.

Nyckelord: www, Internet, söktjänster, IR, utvärdering

<b><u>1 INLEDNING</u></b> .....	<b>4</b>
<u>1.1 Problem och ämnesrelevans</u> .....	4
<u>1.2 Avgränsningar</u> .....	5
<u>1.3 Syfte</u> .....	5
<u>1.4 Frågeställningar</u> .....	5
<u>1.5 Disposition</u> .....	6
<b><u>2 TEORI OCH BAKGRUND</u></b> .....	<b>6</b>
<u>2.1 IR och IR-system</u> .....	6
<u>2.2 IR-modeller</u> .....	8
<u>2.2.1 Booleska modellen</u> .....	8
<u>2.2.2 Vektormodellen</u> .....	8
<u>2.2.3 Probabilistiska modellen</u> .....	9
<u>2.3 Utvärdering av IR-system</u> .....	9
<u>2.3.1 Recall och precision</u> .....	9
<u>2.3.2 Alternativa mått</u> .....	10
<u>2.4 Utvärderingsstudier av återvinningseffektivitet</u> .....	10
<u>2.4.1 Cranfield</u> .....	11
<u>2.4.2 STAIRS</u> .....	11
<u>2.4.3 TREC</u> .....	11
<u>2.5 Relevansbegreppet</u> .....	12
<u>2.6 IR och webben</u> .....	15
<u>2.7 Söktjänster</u> .....	16
<u>2.7.1 Typer av söktjänster</u> .....	17
<u>2.7.2 Indexeringsmetoder</u> .....	18
<u>2.7.3 Återvinnings- och rankningsmetoder</u> .....	20
<u>2.7.4 Söktjänsternas och webbens storlek</u> .....	21
<b><u>3 TIDIGARE FORSKNING</u></b> .....	<b>21</b>
<u>3.1 Utvärderingsstudier</u> .....	22
<u>3.1.1 Sammanfattning utvärderingsstudier</u> .....	28
<b><u>4 METOD</u></b> .....	<b>30</b>
<u>4.1 Ämnesval</u> .....	30
<u>4.2 Val av söktjänster</u> .....	30
<u>4.2.1 Google</u> .....	31
<u>4.2.2 Argos</u> .....	32
<u>4.3 Val av informationsbehov och utformning av sökformuleringar</u> .....	33
<u>4.3.1 Informationsbehov</u> .....	33
<u>4.3.2 Utformning av sökformuleringar</u> .....	34
<u>4.4 Utvärderingskriterier</u> .....	37
<u>4.4.1 Relevanskategorier</u> .....	37
<u>4.4.2 First twenty precision som effektivitetsmått</u> .....	38
<u>4.4.3 Tester vid olika definitioner av relevans</u> .....	39
<u>4.5 Signifikanstest</u> .....	40
<u>4.6 Praktiska aspekter</u> .....	40
<b><u>5 RESULTAT OCH DISKUSSION</u></b> .....	<b>41</b>
<u>5.1 Resultat</u> .....	41

<u>5.1.1 Fördelning av resultat – relevanta kategorier</u> .....	41
<u>5.1.3 <i>First twenty precision</i> vid olika definitioner av relevans</u> .....	44
<u>5.2 Resultat för signifikanstest</u> .....	45
<u>5.3 Diskussion</u> .....	45
<u>5.3.1 Slutsatser</u> .....	47
<u>6 SAMMANFATTNING</u> .....	48
<u>KÄLLFÖRTECKNING</u> .....	50
<u>BILAGA 1</u> .....	54

# 1 Inledning

*"It is fair to say that Internet-based information retrieval would collapse if search engines were not available; without search engines, searchers would be about as successful negotiating the Internet as someone trying to look up a phone number in an unsorted Manhattan phone book."*(Gordon & Pathak 1999, s. 142)

Den snabba utvecklingen av Internet och då särskilt dess yngre del *World Wide Web*<sup>1</sup>, har under de senaste åren inneburit att människors vardagliga, yrkesmässiga och utbildningsmässiga tillgång och förhållande till information förändrats på ett radikalt sätt. Trots att en stor mängd information blivit tillgänglig elektroniskt, har det samtidigt uppkommit problem när det gäller att få tag i den information man söker. Problemen kan primärt ses som en konsekvens av den stora mängden information, men även som konsekvenser av webbens heterogenitet och dess inkonsekvens (Oppenheim, Morris & McKnight 2000, s 190-211).

Det finns olika metoder för att lokalisera information på webben. Det här arbetet kommer ta sin utgångspunkt i den metod som går ut på att man formulerar en fråga hos en söktjänst, som sedan återvinner information som förhoppningsvis motsvarar behovet. Ungefär 85 % av webbanvändarna använder söktjänster för att lokalisera information och flertalet söktjänster ligger bland de tio vanligast utnyttjade sidorna på webben (Lawrence & Giles 1999 s. 107). Detta understryker också söktjänsternas stora betydelse när det gäller informationsåtervinning på webben.

Internet och webbens utveckling har även inneburit nya utmaningar för *Information Retrieval* (IR) som forskningsområde. Traditionella IR-system utformade för att indexera en statisk dokumentsamling har i och med webbens dynamiska karaktär konfronterats med stora problem, och så är även fallet med de traditionella utvärderingsteknikerna. Det har gjorts ett flertal utvärderingar av söktjänster på webben. De flesta påminner om varandra och tar sin utgångspunkt i traditionell utvärdering av IR-system, där man främst använt sig av *precision* som mått, men samtidigt också försökt införa alternativa mått anpassade till webben.

Liksom fallet med övrig information har också mängden vetenskaplig information på webben ökat, även om den dock utgör en relativt liten del. Lawrence & Giles (1999, s. 107) beräknade att ca 6 % av webbservrarna utgjordes av vetenskapligt eller utbildningsmässigt material.<sup>2</sup> Det har framförts att kommersiella söktjänster producerar ytterst lite material av värde för akademiskt bruk (The Chronicle of Higher Education 1996). Med det som utgångspunkt har jag i detta arbete valt att utvärdera en generell frågebaserad söktjänst, Google, och en ämnesspecifik söktjänst, Argos, effektivitet när det gäller att återvinna information av vetenskaplig karaktär från ett avgränsat akademiskt ämnesområde.

## 1.1 Problem och ämnesrelevans

Problematiken har till viss del redan berörts, men jag har själv under min utbildning upplevt det som om informationssökningar på webben många gånger resulterat i en stor mängd icke-relevant material. Detta både när det gäller sökningar gjorda med generella söktjänster, så väl

---

<sup>1</sup> Även www, i detta arbete fortsättningsvis refererat till som webben.

<sup>2</sup> Definierat som universitets-, college- och forskningsservrar.

som med ämnesspecifika sådana. Med tanke på att webben från början var avsedd för vetenskaplig kommunikation (se kap. 2.7), men att den med tiden blivit dominerad av andra områden, skulle jag också finna det intressant att i nuläget genomföra en undersökning som på något sätt kan mäta söktjänsters effektivitet i en sådan kontext. Vetenskap på webben ger intrycket av att vara synonymt med naturvetenskap och då främst med medicin.<sup>3</sup> Även om så inte är fallet, skulle det enligt min uppfattning därför också vara intressant med en undersökning av ett icke-naturvetenskapligt ämne. Med denna bakgrund är det även min övertygelse att en sådan undersökning skulle vara väl motiverad och möjlig att genomföra inom ramen för IR, vilket även gör ämnet relevant för biblioteks- och informationsvetenskap som forskningsområde.

## 1.2 Avgränsningar

Jag har valt att begränsa mig till ett informationsbehov grundat i den akademiska ämnesdisciplinen Antikens kultur och samhällsliv (se kap. 4.1), dels med anledning av arbetets förväntade storlek och omfattning och dels med anledning av att jag besitter en ämneskunskap i det berörda ämnet. Denna ämneskunskap ser jag som en nödvändighet, då det i utvärderingsarbetet krävs någon form av relevansbedömningar från min sida. Fokus kommer, med utgångspunkt från frågorna, inte ligga på vilken eventuell relevans det återvunna materialet skulle ha för akademisk forskning i det berörda ämnet, utan snarare vilken relevans det skulle ha för en mer grundläggande utbildningsnivå.

## 1.3 Syfte

Syftet med detta arbete 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.

## 1.4 Frågeställningar

- Vilken återvinningseffektivitet uppvisar söktjänsterna Google och Argos, med avseende på måttet *first twenty precision*?<sup>4</sup>
- Hur påverkar olika definitioner av relevans resultaten för *first twenty precision*?
- Föreligger det då någon signifikant skillnad mellan resultaten om dessa statistiskt generaliseras?

---

<sup>3</sup> Utan att ha gjort någon djupare efterforskning, kan man dock t.ex. vid en konsultation av Search Engine Watch lista över specialiserade sökmotorer se att medicin har fått en egen kategori där nio sökmotorer listas. I kategorin vetenskapliga sökmotorer listas elva sökmotorer, vilka domineras av naturvetenskap och där endast två av dessa har med samhällsvetenskapliga kategorier, medan ingen har med någon humanistisk ämneskategori. (Search Engine Watch 2001c)

<sup>4</sup> För beskrivning av det aktuella måttet se kap. 4.4.2

## 1.5 Disposition

- Kapitel 1 – Inledning med en kort bakgrund följt av en presentation av avgränsningar, syfte och frågeställningar.
- Kapitel 2 – Teori och bakgrund. En genomgång av IR, IR-system, IR-modeller och olika aspekter av detta, så som mått, relevans, utvärdering, söktjänster etc. Detta för att förankra mitt arbete i en teoretisk grund samt ge en bakgrundsbild av forskningsområdet.
- Kapitel 3 – Tidigare forskning. En genomgång av sju tidigare utvärderingsstudier av söktjänster på webben. Detta för att knyta mitt eget arbete till en forskningstradition, samt därmed också belysa vad som gjorts tidigare.
- Kapitel 4 – Metod. En presentation av min egen metod. En genomgång av vilka söktjänster som valts och varför, hur sökformuleringar konstruerats, vilket mått som använts samt mitt förhållningssätt till relevans.
- Kapitel 5 - Resultat och diskussion. Med frågeställningen som utgångspunkt, en presentation av de resultat som söktjänsterna uppnått med avseende på *first twenty precision*, samt en diskussion kring detta och vilka slutsatser man kan dra.
- Kapitel 6 – Sammanfattning av arbetet som helhet.

## 2 Teori och bakgrund

### 2.1 IR och IR-system

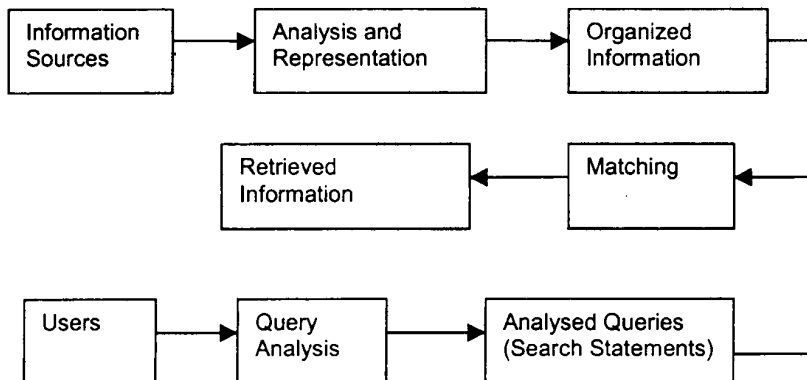
Detta avsnitt syftar till att, utifrån en begränsad mängd grundlitteratur, ge en bakgrund och en övergripande bild av IR och IR-system. Detta betyder samtidigt också att alla aspekter av IR inte kommer att täckas in.

Enligt Baeza-Yates & Ribeiro-Neto (1999, s. 1) handlar IR om representation, lagring, organisation av och åtkomst till information. Ett IR-system som företeelse kan med utgångspunkt från dess mest grundläggande terminologiska betydelse betraktas som självbeskrivande – ett system som lagrar och återvinner information. Företeelsen förutsätter att det finns vissa dokument som innehåller information, som blivit organiserad på ett sådant sätt att en smidig återvinning är möjlig (Chowdhury 1999b, s. 1). Att återvinna samtliga dokument som är relevanta i förhållande till en användares fråga är i själva verket den primära målsättningen med ett IR-system (Baeza-Yates & Ribeiro-Neto 1999, s. 2). En fråga (*query*) kan i detta sammanhang enligt Baeza-Yates & Ribeiro-Neto (1999, s 449) definieras som en formell representation av ett informationsbehov.

Chowdhury (1999b, s. 2) menar att man kan tala om tre olika områden som konstituerar ett IR-system:

- (1) Föremål som innehåller information.
- (2) Användares frågor.
- (3) Matchning av dessa frågor med en dokumentdatabas.

**Figur 1. IR-system.**



Källa: Chowdhury (1999b, s. 4)

Figur 1 illustrerar schematiskt gången i ett IR-system. Chowdhury (1999b, s. 4-5) skiljer vidare mellan två kategorier av IR-system – *in-house* och *online*. Den första kategorin kännetecknas av att den är framtaget för att fungera inom t.ex. ett specifikt bibliotek och därmed endast åtkomligt för användare inom denna specifika organisation. Den andra kategorin av IR-system som här åsyftas är utvecklade för att möjliggöra åtkomst till flera databaser och för en stor variation av användare. Författaren poängterar också att Internet och webben flyttat fram gränserna för denna typ av IR-system och gjort information tillgänglig för nästan alla förutsatt att de har den utrustning som krävs. Det är också den sistnämnda kategorin av IR-system som detta arbete kommer att behandla. I detta sammanhang bör det även nämnas att man skiljer mellan experimentella – system/test i laboratoriemiljö och operationella – system/test i verklig miljö (Tague-Sutcliffe 1997, s. 206). Utifrån denna definition kommer mitt arbete vara av operationell karaktär.

Merparten av de tidiga IR-systemen var utvecklade för bibliografiska databaser, d.v.s. en databas som ger tillgång till dokumentsurrogat, vilka utgörs av t.ex. abstrakt och titel, av de verkliga dokumenten. Med tiden har detta dock utvecklats till att man idag vanligen söker i fulltextdatabaser, där varje betydelsebärande ord är sökbart (Sparck Jones & Willett 1997, s. 1).

En av de viktigaste funktionerna i ett IR-system är alltså att matcha innehållet i dokumenten med användares frågor. För att möjliggöra detta krävs det surrogat för varje dokument. Det är dessa surrogat man skapar vid indexering. Indexeringen kan ske manuellt eller automatiskt och ett index består av ett antal utvalda termer med tillhörande adresser. Vid automatisk indexering skapas en inverterad fil, en typ av index och som utgörs av två delar. (1) en lista av alla distinkta ord i dokumentsamlingen och (2) för varje ord i listan, en lista av pekare till de dokument som innehåller ordet (Chowdhury 1999b, s. 92).

För att effektivt kunna tillfredsställa en användares informationsbehov måste ett IR-system på något vis tolka innehållet av informationsföremålen i en samling och ranka dem med hänsyn till vilken grad av relevans de har i förhållande till användarens fråga (Baeza-Yates & Ribeiro-Neto 1999, s. 2). I detta sammanhang kan man även skilja mellan återvinning av data och återvinning av information. Vid återvinning av data är man intresserad av fakta och vid återvinning av information är man intresserad av ett ämnesområde.

## 2.2 IR-modeller

Följande avsnitt bygger, om inte annat anges, på Baeza-Yates & Ribeiro-Neto (1999). Att förutsäga vilka dokument som är relevanta och vilka som inte är det är ett centralt problem när det gäller IR-system. Beslut av den karaktären är oftast beroende av någon form av rankningsalgoritm vars funktion är att bringa ordning bland de återvunna dokumenten och placera de som anses mer relevanta högst upp i denna ordning. Således är dessa rankning algoritmer<sup>5</sup> också att betrakta som själva kärnan i ett IR-system (Baeza-Yates & Ribeiro-Neto 1999, s.19).

Baeza-Yates & Ribeiro-Neto (1999, s. 23) menar att det är rankningsalgoritmernas förutsättningar som bestämmer IR-modellen. De väljer att definiera en IR-modell som en kvadrupel – ett objekt med fyra komponenter -  $\mathbf{D}$ ,  $\mathbf{Q}$ ,  $F$ ,  $R(q_i, d_j)$  , där:

- (1)  $\mathbf{D}$  är en mängd av dokumentrepresentationer.
- (2)  $\mathbf{Q}$  är en mängd av representationer av användares informationsbehov.
- (3)  $F$  är ett ramverk för modellering av dokumentrepresentationer, frågor och deras relationer.
- (4)  $R(q_i, d_j)$  är en rankningsfunktion som associerar ett reellt tal med en fråga  $q_i$  i  $\mathbf{Q}$  och en dokumentrepresentation  $d_j$  i  $\mathbf{D}$ .

Jag kommer här endast att kortfattat gå igenom tre modeller, de som författarna benämner de klassiska modellerna – booleska, vektor och probabilistiska. Syftet med denna genomgång är att ge en bakgrund till hur sökmotorer fungerar, då de flesta av dessa är baserade på någon form av den booleska modellen eller vektormodellen (se kap. 2.7.3). Denna genomgång därmed också relativt informell till sin karaktär.<sup>6</sup>

### 2.2.1 Booleska modellen

Den booleska modellen baseras på mängdlära, eller satslogik. Modellen fick tidigt ett stort genomslag. Dokumenten representeras här av binära termvektorer alt. av en mängd termer. Användarens informationsbehov måste uttryckas i sökfrågor där man använder *boolean expressions*, d.v.s. operatorerna AND, OR, NOT. Enligt Baeza-Yates & Ribeiro-Neto (1999, s. 26) bär dock den booleska modellen på en mängd problem. Det första problemet har sin grund i att modellen bygger på binära beslut – att ett dokument antingen är relevant eller icke-relevant. Det andra problemet som författarna tar upp, har att göra med att *boolean expressions* har en precis semantik, vilket innebär att det kan vara svårt att formulera ett informationsbehov till en fråga. Trots dessa nackdelar är den booleska modellen fortfarande den dominerande modellen i kommersiella dokumentdatabaser. Fördelarna med modellen är att formalismen bakom den är klar samt dess enkelhet.

### 2.2.2 Vektormodellen

Vektormodellen baseras på algebra där informationsbehov och dokument representeras av termvektorer. Graden av likhet mellan fråga och dokument mäts genom att likheten mellan respektive vektor mäts. Termer i dokument och frågor får numeriska, icke-binära vikter, vilka

---

<sup>5</sup> En algoritm kan i detta sammanhang definieras som den metod ett IR-system använder för att lösa ett problem (Korfhage 1997, s. 313).

<sup>6</sup> För den som är intresserad av en mer formell och matematisk framställning av dessa tre modeller, se exempelvis Baeza-Yates & Ribeiro-Neto (1999, kap. 2)

är avsedda att spegla termernas betydelse i dokument och frågor. Två viktiga faktorer när termer skall tilldelas vikter är *tf factor* och *idf factor*. Med *tf factor* avses här antalet förekomster av en term i ett dokument, vilket ger ett mått på hur väl denna term beskriver dokumentets innehåll. Med *idf factor* avses den inversa frekvensen av en term i en dokumentsamling. Anledningen till att man är intresserad av den sistnämnda faktorn är att termer som förekommer i många dokument inte är särskilt användbara för att skilja relevanta dokument från icke-relevanta sådana. En sökterm som alltså har en hög frekvens i ett specifikt dokument samtidigt som söktermen inte förekommer i några andra dokument i samlingen bör vara beskrivande och ytterst användbar vid återvinning för det specifika dokumentet där termen förekommer ofta. Genom att sortera de återvunna dokumenten i fallande ordning utifrån graden av likhet, tar vektormodellen också hänsyn till dokument som matchar frågan endast partiellt. På så vis tillåter modellen alltså återvinning av dokument som approximerar frågevillkoren. Termviktningen förbättrar slutligen också återvinningseffektiviteten. (Baeza-Yates & Ribeiro-Neto 1999, s. 27-29)

### 2.2.3 Probabilistiska modellen

Den probabilistiska modellen är enligt Baeza-Yates & Ribeiro-Neto (1999, s. 31) baserad på antagandet att givet en fråga  $q$  och ett dokument  $d_j$  i en samling, försöker den probabilistiska modellen uppskatta sannolikheten att användaren finner dokumentet  $d_j$  relevant. Modellen utgår ifrån att denna sannolikhet för relevans endast är avhängig frågan och dokumentet. Vidare utgår modellen från att det existerar en delmängd av samtliga dokument som användaren föredrar som svarsmängden för frågan  $q$ . En sådan ideal svarsmängd betecknas  $R$  och borde maximera den övergripande sannolikheten av relevans för användaren. Dokument i mängden  $R$  förutsätts vara relevanta för användaren, medan dokument som inte förekommer i denna samling förutsätts vara icke-relevanta.

Modellen försöker alltså estimeras sannolikheten för att ett dokument skall vara relevant för en fråga. Fördelarna med denna modell är att dokumenten rankas i fallande ordning efter dess sannolikhet för relevans, användaren slipper formulera frågorna med *boolean expressions* samt att modellen tar hänsyn till återvinningens osäkerhet.

## 2.3 Utvärdering av IR-system

Innan ett IR-system slutligen implementeras utförs enligt Baeza-Yates & Ribeiro-Neto (1999, s. 73) vanligtvis en utvärdering av systemet, bl.a. bör man undersöka hur precis svarsmängden är. Denna typ av utvärdering refereras till som *retrieval performance evaluation*, eller översatt till svenska, utvärdering av återvinningseffektivitet.

Utvärdering av ett IR-systems återvinningseffektivitet är vanligen baserad på en testkollektion. En sådan testkollektion består i sin tur av tre delar:

- (1) En samling dokument.
- (2) En samling informationsförfrågningar.
- (3) För varje informationsförfrågan en mängd relevanta dokument i kollektionen.

### 2.3.1 Recall och precision

De två vanligaste måtten vid utvärdering av återvinningseffektivitet är *recall* – andelen av de relevanta dokumenten som återvunnits, samt *precision* – andelen återvunna dokument som är

relevanta. Om man låter  $|R|$  vara antalet relevanta dokument,  $|A|$  lika med svarsmängden och  $|Ra|$  lika med antalet relevanta dokument i svarsmängden, får man följande ekvationer:

$$- \text{ Recall} = \frac{|Ra|}{|R|}$$

$$- \text{ Precision} = \frac{|Ra|}{|A|}$$

*Recall* och *precision* definierade som ovan, förutsätter att alla dokumenten i svarsmängden har undersökts. I en verklig situation undersöker användaren listor där dokumenten är rankade utifrån dess relevans för frågan, på så vis varierar också *recall* och *precision* beroende på hur många dokument användaren undersökt i denna lista. Med anledning av detta är det därför nödvändigt att konstruera *recall/precision*-kurvor för att på ett lämpligt sätt undersöka återvinnings effektiviteten och på så vis ta fram en genomsnittlig värde för måtten. Ytterligare ett sätt att på detta vis mäta återvinnings effektivitet är att beräkna *recall* och *precision* vid givna DCV:s (*document cutoff values*), när man t.ex. undersökt 5, 10, 15 o.s.v. av de återvunna dokumenten (Baeza-Yates & Ribeiro-Neto 1999, s. 73-78).

*Recall* och *precision* som mått har också sina begränsningar. Så är t.ex. fallet vid mycket stora testkollektioner, eller som med webben, där antalet relevanta dokument för en fråga kan vara okänt, vilket i sin tur får till följd att det egentligen är omöjligt att uttala sig om *recall*. Vidare är de båda måtten besläktade med varandra och mäter olika aspekter av samma sak. Det vore därför i många situationer mer lämpligt att använda ett mått som kombinerar de båda måtten (Baeza-Yates & Ribeiro-Neto 1999, s. 80).

### 2.3.2 Alternativa mått

För att till viss del kunna kringgå de problemen som *recall* och *precision* bär på, har det tagits fram ett stort antal alternativa mått, där vissa fått större genomslag och andra mindre. Vissa bygger på liknande principer som *recall* och *precision*, medan andra är användarorienterade.

De alternativa mått vilka är användbara och relevanta att ta upp för detta arbete kommer inte att belysas här, utan fortlöpande längre fram i arbetet, och då först och främst i avsnittet som behandlar tidigare utvärderingsstudier på webben (se kap. 3.1). Detta dels för att undvika upprepningar, och dels för att jag ser det som mer passande att presentera dem i de sammanhang där de använts.

## 2.4 Utvärderingsstudier av återvinnings effektivitet

Det har genomförts en stor mängd utvärderingsstudier av återvinnings effektivitet av IR-system över åren. Chowdhury (1999b, s.215) menar att man historiskt och generellt sett kan urskilja två olika intresseområden beträffande studier av detta slag. Det första området, vilket präglade studierna under de första decennierna (se kap. 2.4.1), hade sin fokus på utvärdering av indexeringssystem och på så vis ville man undersöka vad som skedde i ett IR-system. Det andra intresseområdet, vilket präglat senare studier, utmärks snarare av en strävan att förstå varför det som skedde sker. Detta genom att undersöka olika aspekter, så som söktekniker, ranking, termviktning, kostnadseffektivitet etc. För att ge en överblick presenteras i följande

avsnitt ett urval av ett antal utvärderingsstudier som styrt och påverkat IR-forskningen i hög grad. Liksom fallet med vissa av de alternativa måtten, kommer utvärderingsstudier gjorda på webben behandlas i en separat del senare i arbetet.

### 2.4.1 Cranfield

Den första utförliga utvärderingen av IR-system genomfördes i Cranfield, England, vid slutet av 1950-talet och har fått namnet Cranfield I. En andra serie av studier – Cranefield II, genomfördes på 60-talet och det är även den modell som senare kommit att utgöra typexemplet för denna form av experimentella utvärderingar av IR-system. Cranfieldmodellen kan sedan dess också sägas ha fungerat som ett paradigm inom IR-forskning och utvärdering (Harter & Hert 1997, s. 7-8).

Under Cranfield II testade man den relativa återvinningseffektiviteten hos 33 olika indexeringspråk. De grundläggande komponenterna av testerna var (1) en testkollektion av dokument, (2) ett antal frågor och (3) ett antal relevansbedömningar. Testkollektionen bestod av 1400 dokument vilka alla behandlade olika aeronautiska aspekter. Frågorna var 331 till antalet och utformade i samverkan med författare av utvalde artiklar. Man utförde binära relevansbedömningar genom att varje dokument bedömdes relevant eller icke-relevant i förhållande till varje fråga. De mått som man primärt använde sig av var *recall* och *precision*, men även *fallout* – andelen icke-relevanta dokument som återvunnits (Harter & Hert 1997, s. 8).

### 2.4.2 STAIRS

Det var först under 1980-talet som det, med anledning av teknikutvecklingen – lagring etc., blev möjligt att genomföra storskaliga tester. STAIRS (*Storage And Information Retrieval System*) -projektet som genomfördes på IBM:s STAIRS-system, är ett exempel på den typen av test. Detta projekt var även den första genomgripande utvärderingen av ett fulltextsystem. Databasen man genomförde testerna på bestod av ungefär 350 000 sidor av onlineinformation. STAIRS-projektet är fortfarande en av de få storskaliga studierna som gjorts på ett kommersiellt operationellt system, vilket också gör projektet unikt och värt att ta upp. (Harter & Hert 1997, s. 23) Syftet med utvärderingen var att bedöma hur väl ett system kunde återvinna samtliga relevanta dokument till en given fråga. Måtten som användes var *recall* och *precision* (Chowdhury 1999b, s. 226).

### 2.4.3 TREC

TREC (*Text REtrieval Conference*) är ett omfattande forskningsprojekt inom IR-utvärdering, organiserat av NIST (*National Institute of Standards and Technology*) och sponsrat av ARPA (*Advanced Research Projects Agency*). Resultatet från det första mötet publicerades 1993 och sedan dess har möten hållits fortlöpande varje år.

TREC har enligt Smeaton & Harman (1997, s. 171) fyra primära målsättningar:

- (1) Att uppmuntra forskning kring textåtervinning i storskaliga tester.
- (2) Att öka kommunikationen mellan forskare inom industrin, den akademiska världen och från statligt håll.
- (3) Att påskynda förflyttningen av teknologi från forskningslaboratorier till kommersiella produkter.

(4) Att öka tillgången av lämpliga utvärderingstekniker.

Även om TREC-testerna i många avseende skiljer sig från tidigare IR-experiment, är de enligt Harter & Hert (1997, s. 24) i grunden utformade som Cranfieldtesterna.

TREC-testerna är centrerade kring två uppgifter – *routing* och *ad hoc*. *Routing* utgår från att samma frågor ställs, men att nya dokument ideligen genomsöks. Vid *routing*-uppgiften känner man till ämne och relevanta dokument samt att ny data används för testning. *Ad hoc*-uppgiften utgörs av att nya frågor ideligen ställs till en statisk samling av data.

Frågorna är konstruerade för att representera användares informationsbehov och är utformade på tre olika sätt: automatiskt, manuellt eller med *relevance feedback* – en process där användaren markerar de relevanta dokumenten, av vilka man sedan väljer ut termer och uttryck för fortsatt sökning. Dokumentsamlingarna består i huvudsak av fulltextdokument vilka är hämtade från flertalet olika källor. Relevansbedömningarna utförs genom *pooling* metoden. Detta innebär att dokument som troligen är relevanta alstras genom att samla in de översta 100 eller 200 av antalet återvunna dokument hos ett system för en specifik fråga, och sedan föra samman dessa för bedömning. De insamlade dokumenten för varje fråga relevansbedöms sedan på en binär skala av en bedömare. Man använder sig av traditionella mått som *recall*, *precision* och *fallout* för att presentera resultaten (Harter & Hert 1997, s. 24).

TREC har kritiserats på flera punkter. Med anledning av att testerna vilar på samma grund som Cranfieldtesterna, har även liknande kritik, då främst rörande relevans, använts mot TREC som mot dess föregångare. Harter & Hert (1997, s. 27) sammanfattar kritiken i följande punkter:

- Orealistiska relevansbedömningar i förhållandet till verkliga användare.
- Onaturliga dokumentsamlingar och frågor.
- Avsaknaden av slutanvändare eller verkliga användare.
- Användandet av *recall* fungerar inte tillfredställande i stora databaser som TREC:s.
- Användandet av *recall/precision*- och *recall/fallout*-kurvor skildrar inte systemens verkliga effektivitet.

## 2.5 Relevansbegreppet

Som man kan se i tidigare avsnitt av arbetet är relevans ett ständigt återkommande begrepp inom IR och IR-forskningen. Vilket redan poängterats är det också ett IR-systems huvuduppgift att återvinna information relevant för en användare. Jag ser det med anledning av detta därför också som nödvändigt att lyfta ut begreppet relevans, och här behandla det i en separat del där jag belyser olika aspekter och tolkningar av det. Utifrån den litteratur jag tagit del av kan man så här inledningsvis redan göra konstaterandet att det verkar råda en allmän enighet om att relevans är ett centralt begrepp, samt att det saknas en gemensam ram för hur man bör förhålla sig till relevans inom IR-området. Jag tänker i detta avsnitt utgå från Saracevic (1997) och Mizzaro (1997). Två artiklar som ger en översikt på vad som skrivits om relevans inom informationsvetenskap och samtidigt också försöker konstruera ramverk kring begreppet.

Saracevic (1997, s. 143) menar att informationsvetenskap är det tredje ämnesområdet, vid sidan av logiken och filosofin, som kommit att handskas med begreppet relevans. Enligt

honom var S. C. Bradford, under 1930-talet, den förste att använda relevansbegreppet i en informationsvetenskaplig kontext.

Ett grundläggande problem med relevans är att man kan betrakta det ur en mängd olika perspektiv, att begreppet till sin karaktär alltså är relativt. Det har med anledning av detta även presenterats förslag på alternativa namn, som t.ex. *usefulness*, *appropriateness* och *utility*, för att på så vis definiera relevans. Problemet med dessa definitioner av parafraskaraktär är enligt Saracevic (1997, s. 150) att de endast byter ut en odefinierad term mot en annan, och att det samtidigt inte heller löser de problem som föreligger.

För att kunna konstruera en ram i vilken man kan placera de olika synsätten på relevans måste man enligt Saracevic (1997, s. 147) ta företeelsen kommunikation i beaktande. Där kommunikation är en process där information överförs från ett objekt till ett annat. Det första objektet kan kallas källa och det andra destination. Således skriver Saracevic att man kan betrakta:

*”relevance as a measure of the effectiveness of a contact between a source and a destination in a communication process”* (Saracevic 1997, s. 147)

Inom informationsvetenskapen behandlas information och kommunikation i en kunskapskontext, och man använder sig av uttrycket kommunikation av kunskap.<sup>7</sup> Saracevic fortsätter därmed sitt resonemang med utgångspunkt från ovan nämnda och gör följande konstaterande:

*”Communication of knowledge is effective when and if information that is transmitted from one file results in changes in another. Relevance is the measure of these changes.”* (Saracevic 1997, s. 147)

Man kan därmed utifrån Saracevic sammanfatta relevans, ur dess mest fundamentala betydelse i en informationsvetenskaplig kontext, som ett mått på effektiviteten av kontakten mellan en källa och en destination i en kommunikationsprocess. Det har utvecklats en rad informationssystem för att möjliggöra denna kommunikation, system som alla alltså har det gemensamt att de implicit bygger på någon form av tolkning av relevans (Saracevic 1997, s. 148).

Merparten av arbetena om relevans inom informationsvetenskap har enligt Saracevic (1997, s. 148) strävat efter att fastställa: (1) Vilka faktorer eller element som ingår i föreställningen om relevans, och (2) vilken relation specificerar föreställningen om relevans. Han menar vidare att det därför också är i dessa strävanden det ramverk inom vilket man kan placera in begreppet relevans, står att finna.

Mizzaro (1997, s. 814) utgår även han från att det existerar olika typer av relevans, samt att det är allmänt accepterat att relevans är en relation mellan två enheter (*entities*) från två grupper.

---

<sup>7</sup> Saracevic poängterar kunskapsbegreppets mångtydighet, och väljer att för informationsvetenskap använda Bells definition: *”Knowledge is a set of organized statements of facts or ideas, presenting a reasoned judgement or an experimental result, which is transmitted to others through some communication medium in some systematic form.”* (Bell 1973, s. 175) Inom IR kan kunskap definieras något annorlunda, t.ex. Korfhage (1997, s. 325): *”Information integrated to form a large, coherent view of a portion of reality.”*

I den första gruppen finns en av följande tre enheter:

- (1) Dokument.
- (2) Surrogat - en representation av ett dokument.
- (3) Information - det användaren får från dokumentet.

Den andra gruppen utgörs av en av följande fyra enheter:

- (1) Problem - det som användaren står inför och behöver information för att lösa.
- (2) Informationsbehov - här betraktat som en, hos användaren tänkt, representation av problemet, ännu inte formulerad.
- (3) Uppmaning (*request*) - en representation av informationsbehovet på ett "mänskligt" naturligt språk.
- (4) Fråga (*query*) - en fråga formulerad på ett språk anpassat till ett IR-system.

Vidare menar Mizzaro att de ovan nämnda enheterna kan brytas ned i följande tre beståndsdelar:

- (1) Ämne (*topic*) - det ämnesområde som användaren är intresserad av, t.ex. doriska tempel i Grekland.
- (2) Uppgift (*task*) - det användningsområde de återvunna dokumenten kommer att falla in under, t.ex. skriva en magisteruppsats.
- (3) Kontext - det som innefattar allt utöver ämne och uppgift, t.ex. dokument som användaren redan känner till, tid som sökningen får ta etc.

Även Mizzaro (1997, s. 815) belyser begreppets subjektivitet och poängterar att man idag inte betraktar relevans som en dikotomi mellan ja/nej beslut, utan att man snarare bör se det som olika nivåer av relevans.

Liksom Saracevic framhåller han begreppet *pertinence* – relationen mellan ett dokument och en fråga bedömt av användaren, som viktigt, och att man i detta sammanhang betraktar relevans som en utomståendes bedömning av samma relation. (Mizzaro 1997, s. 820) Saracevic (1997, s. 153) menar att skillnaden ursprungligen kommer ur distinktionen mellan fråga och informationsbehov.

Med ovanstående som en teoretisk bakgrund inser man snart att relevansbedömning är ett arbete präglad av många aspekter, att risken för arbiträra bedömningar är stor och att bedömningarna är avhängiga det sätt man väljer att se på relevans. I avsnittet av detta arbete som går igenom tidigare utvärderingsstudier av söktjänster (se kap. 3.1) är tanken att läsaren ska få en mer praktisk inblick i hur man gjort relevansbedömningar i liknande studier som denna, samt att detta även ska ge en bakgrund till hur jag själv rent praktiskt kommer att genomföra mina relevansbedömningar senare i arbetet (se kap. 4.4).

Teoretiskt kommer jag att ta avstamp i föreställningen 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 detta att relevans betraktas som ett mått på till vilken nivå ett dokument tillfredsställer ett informationsbehov utifrån ett antal på förhand definierade kriterier.

## 2.6 IR och webben

Parallellt med utvecklingen och tillväxten av Internet och webben, har även intresset från IR-världen ökat beträffande denna form av lagring och återvinning av information. Utifrån de senast publicerade forskningsrapporterna<sup>8</sup> som på något sätt berör IR på Internet och webben, delar Chowdhury (1999a, s. 210) in denna forskning i ett antal bredare kategorier:

- Sökmotorer.
- Utvärdering av återvinning.
- Tillförlitlighet av information på webben.
- Användargränssnitt.
- Användarstudier.
- Organisation av information på webben.
- Kontrollerad vokabulär.
- Sökeffektivitet på webben.
- Intelligent agenter.
- Webben kontra traditionella databaser.

Det här arbetet kommer i huvudsak att sälla sig till de två första kategorierna. När de gäller utvärdering av sökmotorer/söktjänster så har de flesta studierna utförts på frågebaserade<sup>9</sup> sådana (Oppenheim et al. 2000, s. 192). Inte heller på denna punkt kommer detta arbetet utgöra något avvikande exempel.

Gordon & Pathak (1999, s. 145) menar att det går att urskilja två typer av utvärderingar av söktjänster – *testimonials* och *shootouts*. Den första av de två, *testimonials*, kan generellt sett sägas representeras av de tester som utförs av populärtidskrifter och datorindustrin med avseende på snabbhet, användarvänlighet, design etc. Denna typ av studier ger möjligen, enligt författarna, användbar information, men de ger endast indirekta indikationer på vilken sökmotor som effektivast återinner relevanta webbdokument. Den senare typen, *shootouts*, representeras av typiska IR-utvärderingar, där man jämför olika återvinningsalgoritmer.

I detta arbete kommer jag att utesluta *testimonials*, för att istället fokusera på *shootouts*, då det är denna typ av utvärderingar som är aktuella i detta sammanhang.

De flesta studier i syfte att mäta återvinningseffektivitet, som genomförts i detta sammanhang, kan sorteras under det paradig som stammar ur cranfieldmodellen. Det stora problemet med webben, vilket redan poängterats, är den enorma mängden heterogen information samt dess dynamiska karaktär. Detta problem får därmed givetvis också konsekvenser när det gäller studier företagna inom ramen för IR. Det är t.ex. omöjligt att beräkna antalet relevanta dokument i förhållande till en specifik fråga och följderna av detta blir att det inte går att beräkna *recall*. Landoni & Bell (2000, s. 125) menar att det, problemen till trots, ändå är tydligt att de ackumulerade erfarenheterna från cranfieldtypen av utvärderingar är användbara för att bringa ordning. Alltså att denna modell ändå är lämplig att utgå ifrån.

Oppenheim et al. (2000, s. 194) presenterar fyra metoder som i olika utsträckning använts för att närma sig ovan nämnda problem:

---

<sup>8</sup> Sökningen gjordes i databasen LISA med avseende på tidskrifter inom biblioteks- och informationsvetenskap.

<sup>9</sup> För definition av frågebaserade söktjänster se kap. 2.7.1

- (1) Genomföra en cranfieldtyp av studie med ett strikt definierat ämne, där man känner till de relevanta dokumenten.
- (2) Genomföra en cranfieldtyp av studie, men att i stället för *recall* använda sig av relativ *recall*.<sup>10</sup>
- (3) Hitta något sätt att uppskatta antalet möjligen relevanta dokument med hjälp av någon statistiskt giltig metod.
- (4) Undvika *recall* som mått helt och hållet, trots att det är ett viktigt mått.

Dong & Su (1997, s. 77) skriver att de speciella funktionerna hos söktjänster, så som relevansrankning, hyperlänkar etc., gör att måtten man använder för söktjänster till viss del skiljer sig från traditionella sådana. Landoni & Bell (2000, s. 126) lyfter fram samma problematik och hävdar att man bör ta hänsyn till följande faktorer som karakteriserar webbdokument:

- Dubbletter – sidor med samma innehåll och samma URL.
- Spegelsidor – sidor med samma innehåll, men olika URL.
- Inaktiva hyperlänkar.
- Sidor där endast hyperlänkar returneras.
- Språk.

Landoni & Bell (2000, s. 126-127) fortsätter, med utgångspunkt från detta, sitt resonemang med att lista de mått och olika aspekter vilka de anser lämpliga att ta med i en utvärdering av söktjänster:

- *Precision*.
- *Relative recall*.
- *Relevance-ranking* – att dokument tilldelas vikter efter var i resultatlistan de återfinns.
- *Coverage* – hur mycket som indexerats av söktjänsterna.
- *Directory assessment* – kvalitativ bedömning av hur resultaten presenteras.
- *Noise* – inaktiva länkar, dubbletter, spegelsidor
- *Accessibility* – dokumentens tillgänglighet, hur ofta man får felmeddelanden etc.

## 2.7 Söktjänster

Det var vid slutet av 1980-talet som Internet, och dess funktion som forsknings- och kommunikationsverktyg, blev tillgängligt för en bredare användargrupp. Utvecklingen tog sin början i forskningsvärlden, fortsatte i affärsvärlden och slutligen i samhället som helhet. Det har utvecklats ett antal program och verktyg för att möjliggöra informationsökning på Internet. Det första programmet som på allvar gjorde det möjligt att söka på Internet var WAIS (*Wide Area Information Server*) (Schatz 1997, s. 332). WAIS-sidor för allmänheten bestod av dokumentsamlingar möjliga att söka genom ämneskataloger. Sökningen resulterade i en lista av filer i huvudsak rankade efter termförekomst. I början av 1990-talet introducerades programmet Gopher och dess sökindex Veronica som det första navigeringsverktyget för multimedia, vilket ytterligare ökade allmänhetens tillgång till internet (Chowdhury 1999b, s. 397; Schwartz 1998, s. 974).

1991 lanserades den första, för allmänheten tillgängliga, formen av webben bl.a. bestående av en alfabetisk ämnesordlista med hyperlänkar till de sidor som kom att utgöra www. Systemet

---

<sup>10</sup> För beskrivning av relativ *recall* se kap. 3.1 och studien som Clarke & Willett (1997) genomförde.

som togs fram av CERN – europeiska organisationen för atomforskning, var från början avsett för att underlätta informationssökning för forskare. 1993 släppte NCSA (*National Center for Supercomputing Applications*) programmet Mosaic för samtliga datorplattformar - Windows, Macintosh och UNIX. Detta blev också startskottet för den enorma tillväxt av webben som skett sedan dess (Chowdhury 1999b, s. 397; Zeizig & Lattermann 1996, s. 318). I takt med att antalet http-resurser ökade, så ökade även behovet av faciliteter för informationsåtervinning på webben. Det var mot den bakgrunden som de första webbaserade söktjänsterna, så som vi känner dem, introducerades 1994. De flesta var från början forskningsprojekt. De som överlevde kom med tiden att antingen bli uppköpta av företag, finansierade med reklam eller kapitalinvesteringar, alternativt stöttade med hjälp av någon form av forskningsinitiativ. Söktjänsterna präglas av en komplexitet och mångfald och speglar på så vis också webbens dynamiska karaktär av idag (Schwarz 1998, s. 974).

### 2.7.1 Typer av söktjänster

Det finns enligt Schwarz (1998, s. 974) i huvudsak två typer av söktjänster, de som är uppbyggda kring en hierarkisk ämnesordlista och de som är frågebaserade. Båda typerna upprätthåller databaser vilka innehåller representationer av webbsidor. De hierarkiska ämneskatalogerna, där Yahoo kanske är det mest kända exemplet, presenterar länkar genom hierarkiskt systematiserade ämneskategorier. De frågebaserade söktjänsterna använder sig av algoritmer baserade på sökfrågor för att söka upp dokument som motsvarar dessa frågor. Exempel på sådana är Google, Alta Vista och Excite. De hierarkiska ämneskatalogerna erbjuder även ofta frågebaserade tjänster, då med möjligheten att söka olika kategorier och titlar som ingår i dessa kategorier. På liknande vis erbjuder de frågebaserade söktjänsterna ofta någon form av möjlighet att bläddra (*browse*) i olika kategorier.

Oppenheim et al. (2000, s. 191-193) väljer att dela upp söktjänsterna i fyra olika typer - *robot-driven*, *directory-based*, *meta-search engine* och *software tools*. De första två motsvaras av de ovan nämnda frågebaserade söktjänsterna och de hierarkiska ämneskatalogerna. En *meta-search engine* konsulterar vid en sökfråga flera frågebaserade söktjänsterna samtidigt och bygger därmed alltså på samma princip som dessa. *Software tools* påminner i hög grad om *meta-search engines* med den skillnaden att dessa måste installeras i användarens dator samt att de ofta är spridprogram man får betala för. Fördelarna är att det bl.a. är möjligt att utesluta dubletter och inaktiva hyperlänkar.

Gudivada et al. (1997, s. 62-64) väljer att beskriva söktjänsterna på ytterligare ett annat vis, och delar in metoderna för automatisk återvinning av information på webben i två klasser - *search tools* och *search services*. Med *search tools* avser man faciliteter som använder sig av en robot för indexering. Man väljer i sin tur att dela in denna klass i två typer, där den första utgörs av de som Schwarz kallar frågebaserade söktjänster och den andra av de som samma författare kallar hierarkiska ämneskataloger. Med *search services* avser man samma sak som Oppenheim, Morris & McKnight gör med *meta-search engine*, alltså en sökmotor som möjliggör sökning hos flera frågebaserade söktjänster samtidigt.

I sin genomgång av söktjänster väljer Lawrence (2000, s. 27) att beskriva vad han kallar för *specialized search engines*, vilka fungera på liknande sätt som vanliga söktjänster på webben, men med skillnaden att dessa endast indexerar material från ett specifikt område. Argos är ett exempel på en sådan söktjänst, vilka jag i mitt arbete valt att kalla för ämnesspecifika söktjänster.

De olika namnen och indelningarna ovan visar väl den inkonsekvens och oenighet som råder när det gäller att beskriva söktjänster som företeelse. För att försöka var konsekvent i detta arbete har jag studerat Svenska datatermgruppens terminologi. Svenska datatermgruppen menar man att man bör använda termen söktjänst när det är själva tjänsten som avses och sökmotor när det är själv programvaran som avses (Svenska datatermgruppen 2001). Detta arbete kommer att fokusera på det som ovan bl.a. kallas frågebaserade sökmotorer. Fortsättningsvis kommer jag för detta genomgående använda mig av termen söktjänst alt. sökmotor, beroende på om det är tjänsten eller programvaran som avses.

Gordon & Pathak (1999, s. 142-143) menar att söktjänster tillhandahåller tre funktioner:

- (1) De indexerar ett antal webbsidor som utgör en samling från vilka användare kan återvinna information.
- (2) De strävar efter att presenterar dessa webbsidor på ett sätt som skildrar deras innehåll.
- (3) De tillåter användarna formulera frågor och använder sig sedan av återvinningsalgoritmer för att hitta de mest relevanta dokumenten i förhållande till sökfrågan i samlingen.

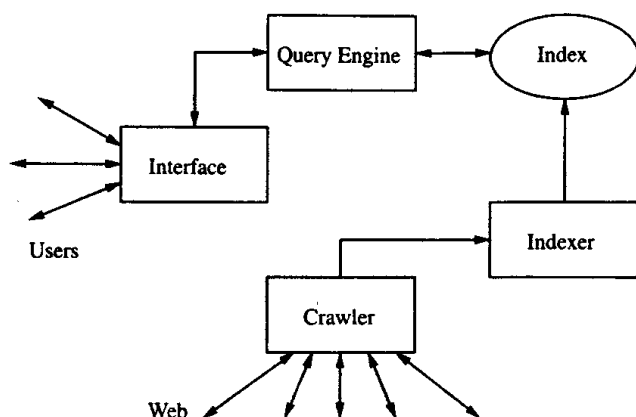
### 2.7.2 Indexeringsmetoder

Det finns enligt Clarke (2000, s. 82) två huvudsakliga metoder för söktjänsterna att indexera webbsidor. Antingen kan de bli registrerade av skaparen av webbsidan via ”*add URL*” alternativt ”*submit your URL*” till en söktjänst, eller så blir de automatiskt insamlade och indexerade. Automatisk indexering av information kan, med anledning den inbyggda tvetydigheten i alla naturliga språk, enligt Savoy & Picard (2001, s. 544) betraktas som ett ytterst komplext problem. Clarke & Willett (1997, s. 184) menar att sökmotorerna består av två komponenter, dels en robot<sup>11</sup> och dels ett textåtervinningsprogram. Roboten färdas över Internet via hyperlänkar från ett dokument till ett annat och samlar in information. Genom att indexera ord och fraser, var dessa befinner sig i texten, hur ofta en term förekommer samt genom att utesluta stoppord försöker sökmotorerna med olika algoritmer fånga dokumentens innehåll. Den insamlade datan, som kan ses som förkortade representationer av dokumenten, där bl.a. URL-adressen lagras, liksom i ett konventionellt online-system i en databas bestående av inverterade filer. Vid en fråga söker textåtervinningsprogrammet databasen för att sedan presentera dokument som motsvarar frågan (Clarke & Willett 1997, s 184; Gordon & Pathak 1999, s. 143). Baeza-Yates & Ribeiro-Neto (1999, s. 373-374) väljer att kalla detta en centraliserad arkitektur (se fig. 2). Vidare menar de att det ligger en avgörande skillnad mellan ett konventionellt IR-system och webben med anledning av att på webben måste alla frågor besvaras utan tillgång till den verkliga texten.

---

<sup>11</sup> Benämns i litteraturen även som *crawler*, *spider*, *wanderer* och *worm*.

**Figur 2. Centraliserad arkitektur.**



Källa: Baeza-Yates & Ribeiro-Neto (1999, s. 374)

De flesta söktjänsterna vill dock göra gällande att deras index är av fulltextkaraktär. Detta innebär att dokumenten är representerade på ett sådant sätt att de återvinns om en sökfråga innehåller ord eller fraser som finns på den webbsida de representerar. Detta stämmer dock inte helt enligt Clarke (2000, s. 84).

Webbsidor som innehåller bildmappar och ramar<sup>12</sup> indexeras inte fullständigt. Det samma gäller för sidor som är lösenordsskyddade och för PDF-filer. Vissa söktjänster tillhandahåller idag sökning av PDF-filer, bl. a. Google (Clark 2000, s. 85; Gordon & Pathak 1999, s. 143; Sullivan 2001). Clarke (2000, s. 85) poängterar även problemet med att sökmotorerna inte indexerar samtliga webbsidor utan endast startsidan på en webbplats. Hon menar att detta även är en starkt bidragande orsak till den låga frekvensen av överlappningar i sökmotorernas resultatlistor sinsemellan. Med anledning av att indexeringen sker genom att roboten följer hyperlänkar, är chansen större att en webbsida med många hyperlänkar blir indexerad, än att en med få blir det. Vissa sökmotorer, som t.ex. Google och DirectHit, rankar på grundval av hyperlänkar uttryckligen populära sidor högre. Detta får enligt Lawrence & Giles (1999, s. 109) till följd att populära sidor blir mer populära, medan nya sidor med få länkar får det svårt att synas i söktjänsternas listor. Vilket i sin tur leder till att en av de stora fördelarna med webben – att man har lika tillgång till alla sidor, är på väg att gå förlorad oavsett mångfald.

De flesta söktjänsterna bestraffar s.k. *spamming* eller *spamindexing*. *Spamming* eller *spamindexing* innebär att man försöker ge en webbsida en hög ranking genom att upprepa nyckelord (som dessutom kan vara irrelevanta för sidans egentliga innehåll), gömda nyckelord (genom att använda liten storlek på texten, eller samma färg på text och bakgrund) och att URL adressen upprepas i samma webbsida.

Det som enligt Clarke (2000, s. 86) slutligen kan påverka en söktjänsts index av webbsidor är om en webbansvarig bestämmer sig för att stänga ute robotar från specifika sidor. Slutsatsen man kan dra utifrån dessa givna förutsättningar är alltså att sökmotorerna inte indexerar allt på webben. Det finns inte heller enligt Clarke någon sökmotor som klara av att indexera hela webben.

---

<sup>12</sup> Indelning av webbsidan i flera rutor där varje ruta består av ett separat HTML-dokument.

Många sökmotorer utesluter även den information som finns i metafälten hos webbsidan (Gordon & Pathak 1999, s. 143). Detta är kanske också förklaringen till det ytterst begränsade användandet av metataggar. Lawrence & Giles (1999, s. 108) beräknade i sin undersökning förekomsten av HTML metataggar till 34,2 %. Inte heller mer komplexa metastandarder, så som XML (*eXtensible Markup Language*) och Dublin core verkar ha fått något större genomslag. I samma undersökning beräknade de förekomsten av Dublin core metadata till 0,3 %.

Inaktiva länkar, webbsidor som har flyttat och inaktuell information är stora irritationsmoment för webbanvändare när de stöter på dem. Det ligger därför i allas intresse att söktjänsternas index innehåller ett uppdaterat material. Clarke (2000, s. 85) skriver att frekvensen med vilken robotarna besöker en webbsida varierar från en dag till ett antal månader. I Lawrence & Giles (1999, s. 109) undersökning beräknade man medelåldern på matchande webbsidor till 186 dagar, och medianen till 57 dagar.

### 2.7.3 Återvinnings- och rankningsmetoder

”For a typical search, hundreds of thousands of information resources will contain keywords that have been included within a user’s query. It is infeasible for the user to evaluate the relevancy of all these resources manually. Therefore the success of a search is always bottlenecked by relevancy ranking techniques.” (Back 2000, s. 30)

Om inte annat anges bygger detta avsnitt på Clark (2000, s. 86-91) och syftar till att ge en generell överblick av hur söktjänsterna återvinner och rankar information. En mer detaljerad genomgång av vilka funktioner söktjänsterna i denna undersökning, Google och Argos, stöder exakt, följer senare i arbetet. (se kap. 4.2) Funktioner som gör det möjligt för användare av sökmotorer att avgränsa sina informationssökningar är av stor vikt eftersom de söker bland en så stor mängd information. Under de senaste åren har söktjänsterna förbättrat dessa möjligheter och på så vis också ökat möjligheterna att minska mängden brus. När det gäller textåtervinning stöder de flesta söktjänsterna frassökning, närhetsökning, fältsökning, boolesk sökning samt trunkering. Vissa tillhandahåller även möjligheten att begränsa sökningen med avseende på datum.

Metoderna som de enskilda söktjänsterna använder sig av varierar. Vid frassökning stöds vanligen funktionen att använda citationstecken före och efter det som användaren vill formulera som en fras. Vanligen används plustecken före en term som måste vara med och på samma vis ett minustecken före en term som inte får vara med. Det har också skett en utveckling av att styra avgränsningarna med hjälp av menyer och boxar. Det har också blivit vanligt med sökning i naturligt språk till förmån framför boolesk operatorer, vilket underlättar för användaren, eftersom denne då inte behöver vara insatt i varje enskild sökmotors funktioner. Det finns visserligen ett problem i att sökmotorn kan misstolka den verkliga betydelsen av *and* och *or* med den booleska logikens AND och OR.

Sökmotorer är primärt framtagna för textåtervinning, under de senaste åren har dock de flesta söktjänsterna utvecklat möjligheterna att återvinna bilder på webben. I avsnittet om indexeringsmetoder nämndes att sökmotorerna inte alltid indexerar bilder, vilket alltså får till följd att de inte heller återvinns. Även om det förekommit experiment med syftet att återvinna bilder genom frågor om bildinnehåll, så krävs det fortfarande, för att en sådan återvinning skall vara möjlig, att bilderna är beskrivna med text. Den vanligaste metoden som användas av sökmotorerna är att de söker upp de sidor som innehåller HTML-taggar IMG SRC och HREF och som bär på ändelserna .GIF eller .JPG.

Hur sökmotorerna rankar sökresultaten har berörts något tidigare, då med anledning av att populära sidor rankades högre av vissa sökmotorer. Detta är av stor betydelse eftersom de flesta användare endast har tid och tålamod att granska de första tio eller tjugo sidorna av de returnerade svaren (Sullivan 1998). Exakt hur sökmotorerna rankar informationen kan man inte säga, eftersom dessa metoder är företagshemligheter. Man kan dock sluta sig till att de flesta utgår ifrån i princip samma metoder. Baeza-Yates & Ribeiro-Neto (1999, s. 380) menar att de flesta sökmotorerna använder sig av varianter på den booleska modellen eller vektormodellen, samt de redan nämnda metoderna baserade på hyperlänkar. De påpekar återigen problemet att rankning liksom sökning utförs utan tillgång till den verkliga texten.

#### 2.7.4 Söktjänsternas och webbens storlek

Med anledning av webbens karaktär är det svårt att göra några exakta uttalanden om dess storlek, 1999 beräknade Lawrence & Giles (1999 s.107) webbens storlek till 800 000 000 sidor. En siffra som förmodligen redan är inaktuell, med tanke på att man på Google hävdar att man indexerat 1 610 476 000 webbsidor (Google 2001a). Söktjänsterna själva använder sig självklart av statistik av detta slag för att marknadsföra sig, och frågan är i vilken grad man kan förlita sig på sådan information i ett vetenskapligt sammanhang. Mer information om söktjänsternas storlek etc. kan man finna hos Search Engine Watch (Search Engine Watch 2001a).

Sättet att mäta webbens storlek varierar självklart också. Det finns ett flertal andra webbplatser som tillhandahåller information om webbens och söktjänsternas storlek. Exempel på tjänst som presenterar statistik om webbens storlek är Netsizer som, när detta arbete skrivs, mäter webbens storlek till bestående av 127 333 483 webbvårdar (Telecordia Technologies 2001). Vid en studie företagen av School of Information Management and Systems vid University of California at Berkeley, menar man att webben består av två delar – en ytlig del (synlig del) och en djupdel (osynlig del). Den ytliga delen, som man vanligen förknippar med webben, består enligt denna studie av ungefär 2 500 000 000 dokument (Lyman & Varian 2000).

### 3 Tidigare forskning

Med tidigare forskning avser jag i detta arbete utvärderingsstudier av återvinningseffektivitet hos söktjänster på webben. Cranfieldtesterna, TREC och STAIRS-projektet kan givetvis också till viss del betraktas som tidigare forskning, jag har dock valt att redogöra för dessa tidigare i arbetet (se kap. 2.4). Jag har även berört den tidigare forskningen generellt i föregående kapitel (se kap. 2.6), men i detta kapitel kommer jag istället fokusera på de verkliga studierna.

Det har genomförts ett stort antal utvärderingsstudier på webben, vilket också är orsaken till att endast ett urval presenteras här. Jag har som tidigare nämnts valt bort *testimonials*, samt även studier som till sin karaktär är vetenskapliga, men utförda utanför ramen för IR.<sup>13</sup> Syftet

---

<sup>13</sup> Utan att närmare gå in på detta, har det genomförts ett antal utvärderingsstudier i syfte att från söktjänster återvinna information från en vetenskaplig disciplin och då naturvetenskapliga sådana. De som jag studerat har helt saknat ett IR-vetenskapligt närmande – teori, mått, begrepp etc., vilket också är anledning till att jag valt att bortse från denna typ av studier. Exempel på sådana studier är: Lebedev (1997) som undersökte hur väl söktjänster på webben återvann vetenskapligt material med avseende på kemi och fysik. Schimrich (1996) undersökte hur väl vetenskaplig information från ämnesområdet geovetenskap återvanns av söktjänster på webben.

här är att ge en metodologisk bakgrund, att lyfta fram hur man i tidigare studier förhållit sig till vilka söktjänster som skall undersökas, vilka och hur många frågor som skall användas, hur man skall förhålla sig till relevans samt vilka mått som använts. Kanske bör man inte lägga någon större vikt vid de exakta resultat som uppnåtts, utan snarare vid just de metoder som använts. Jag har utgått från två av de fyra typer som Oppenheim et al. (2000, s. 194) presenterar, de som använder relativ *recall* samt de som undviker *recall* som mått (se kap. 2.6). Anledningen till att jag valt ut dessa är att min egen undersökning i många hänseenden bygger på en liknande metodologi i hur frågor utformats, hur många frågor som använts och hur relevansbedömningar genomförts (se kap. 4).

### 3.1 Utvärderingsstudier

I följande avsnitt kommer jag att redogöra för sju utvärderingsstudier av återvinningseffektivitet på webben.

**Leighton** (1995) var en av de första att genomföra en utvärderingsstudie av återvinningseffektivitet på webben. Han jämförde fyra söktjänster - Infoseek, Lycos, Webcrawler och WWWorm. Testet kan sägas utgöras av två delar, dels ett relevansexperiment och dels ett responseexperiment.

Med anledning av söktjänsternas olika sökfunktioner – syntax etc., valde författaren att utifrån de givna funktionerna försöka optimera frågorna till varje söktjänst. Sammanlagt rörde det sig om åtta frågor som var kombinerade på ett sådant sätt att de skulle representera svårt och enkelt att finna.

För att mäta effektiviteten valdes *precision* som mått samtidigt som *recall* valdes bort. Han valde även att göra två *precision*-kurvor – en för samtliga dokument som återvunnits och en för de första tio. Frågorna relevansbedömdes på en femgradig skala, där fyra eller fem innebar att dokumentet klassades som relevant. Slutligen utvecklades följande kategorier av mått:

- Icke-relevanta eller dubletter av samtliga träffar i förhållande till topp tio.
- Totala antalet relevanta träffar, där man inte räknade med dubletter.
- Totala antalet relevanta träffar, där man inte räknade med dubletter, i förhållande till samtliga träffar.
- Antalet relevanta träffar i topp tio, där man inte räknade med dubletter.

Författarens slutsatser var att WWWorm fick ett sämre resultat än de övriga tre, vilka samtliga, med små variationer, uppnådde liknande resultat. Med en så liten mängd frågor och undersökta dokument tillsammans med så små variationer i resultaten såg författaren det som onödigt att pröva den statistiska betydelsen.

Tanken med responseexperimentet var att mäta responstiden hos söktjänsterna dels när många användare var i kontakt med dem och dels under lugnare perioder. I slutändan fick författaren dock nöja sig med att mäta responstid under endast lugnare perioder, d.v.s. lördagar och söndagar. Här hade Webcrawler och Infoseek de bästa tiderna medan WWWorm aldrig hade en tid under 70 sekunder.

Denna studie är kanske främst intressant i sin egenskap som en av de tidigast utförda utvärderingarna av återvinningseffektivitet på webben.

**Chu & Rosenthal** (1996) utförde en studie där man jämförde och utvärderade söktjänsterna Alta Vista, Excite och Lycos, med en målsättning att finna en gångbar metod för studier av detta slag.

Söktjänsterna valdes för att de enligt författarna representerade en bredd. Av de tio frågorna som användes var nio verkliga frågor hämtade från biblioteket vid Long Island University. Den tionde frågan konstruerades av författarna till arbetet i syfte att testa fältsökning.

Man jämförde söktjänsterna med utgångspunkt från olika frågesyntax, så som boolesk logik, trunkering, fältsökning och ord/frassökning. Detta med anledning av att söktjänsterna stödde olika sådana. Vidare utvärderade man söktjänsternas prestanda med utgångspunkt från *precision* och responstid. *Recall* valde man att bortse från därför att man ansåg det omöjligt att beräkna detta mått.

Relevansen av de tio först återvunna dokumenten bedömdes separat av de båda författarna på en tregradig skala – 1 för relevant, 0,5 för lite relevant och 0 för icke-relevant. Relevansbedömningen skedde enbart utifrån information hos de respektive söktjänsternas träfflistor.

Alta Vista fick det högsta värdet (0,78) följt av Lycos (0,55) och Excite (0,45). Av de 250 dokument som återvanns fanns det få dubletter söktjänsterna sinsemellan, vilket enligt författarna möjligen kunde tala för att de i stor utsträckning indexerar olika delar av webben.

Studien avslutas med att man formulerar en utvärderingsmetod för söktjänster baserad på följande fem delar: *composition of web indexes*, *search capability*, *retrieval performance*, *output option* och *user effort*.

**Ding & Marchionini** (1996) jämförde de tre söktjänsterna Infoseek, Lycos och OpenText med utgångspunkt från funktioner och effektivitet. Tjänsterna valdes med anledning av dess olika gränssnitt och rankningsmetoder och att de därmed också skulle utgöra ett representativt urval av frågebaserade söktjänster. Med olika typer av frågor ville man finna positiva likheter som kunde ligga till grund för nästa generation av söktjänster.

Tre av de fem frågorna som användes valdes slumpvis från övningsfrågor till sökning i DIALOG för undervisning i informationsvetenskap. De övriga två frågorna formulerades utifrån personliga intressen.

För att uppnå bästa resultat av sökningarna anpassade man frågesyntax till söktjänsternas olika funktioner och gränssnitt.

De kriterier av effektivitet man var intresserade av var: *Precision*, dubletter, hyperlänkars validitet och överlappningar mellan söktjänsterna. Man begränsade sig till att undersöka de tjugo första träffarna. Samma fråga ställdes med tjugo minuters mellanrum till varje söktjänst för att på så vis minska risken för att databaserna vuxit.

Relevansbedömningarna gjordes av författarna med utgångspunkt från en sexgradig skala, anpassad för varje enskild fråga, där 5 var lika med mycket relevant och 0 lika med att frågans ämne överhuvudtaget inte behandlades. Man valde även att gå till det verkliga dokumentet för

att genomföra sina relevansbedömningar. Måtten som användes var *precision*, *saliency* och *relevance concentration*.

*Precision* delades in tre typer:

- (1) 1a – antalet dokument som fick 3, 4 eller 5 i poäng hos de tjugo första träffarna hos varje söktjänst.
- (2) 1b - antalet dokument som fick 4 eller 5 i poäng hos de tjugo första träffarna hos varje söktjänst.
- (3) 2 – antalet dokument som fick 3, 4, eller 5 i poäng i en sammanslagning av tre uppsättningar av de tjugo första träffarna från de tre olika söktjänsterna.

*Saliency* definieras som summan av poängen från samtliga tjugo träffar för varje söktjänst i förhållande till summan av poängen från samtliga söktjänster. *Relevance concentration* definieras som antalet dokument med 4 eller 5 i poäng bland de tio första träffarna delat med dokument med 4 eller 5 i poäng bland de tjugo första träffarna.

De resultat man kom fram till var att Lycos och OpenText överträffade Infoseek när det gällde *precision* och *saliency* och relevanskoncentration. *Precision* för samtliga tre söktjänster låg på under 55%, vilket författarna tycker är lågt med tanke på att endast de tjugo första träffarna undersöktes. Man fick också förvånansvärt få överlappningar, vilket även andra undersökningar pekat på.

Man lyfte själva fram följande brister i sin undersökning:

- Endast fem frågor användes.
- Endast de tjugo första svaren undersöktes.
- Subjektiv relevansbedömning.
- Responstid och tillgänglighet undersöktes inte.
- Måtten var *exploratory*.<sup>14</sup>
- Söktjänsternas ständiga utveckling.

Man efterfrågade slutligen större möjligheter för användaren att styra sökprocessen.

Nicholson (2000) genomförde en replikation av Ding & Marchioninis studie för att utforska problemet med webbens dynamiska karaktär. Man kom fram till helt skilda resultat i jämförelse med det ursprungliga arbetet, och poängterar att det är just webbens dynamiska karaktär som är orsaken till detta.

**Tomaiuolo & Packer** (1996) menade att man vid tidigare utvärderingar av detta slag använt för få frågor och undersökte själva de frågebaserade söktjänsterna Magellan, Point, Lycos, Infoseek och Alta Vista med utgångspunkt från 200 frågor. Magellan och Point skiljde sig något från de övriga så till vida att deras material var förgranskat.

Frågorna togs bl.a. från verkliga frågor som ställts vid Connecticut State University Library, men ämnen för frågor togs även från Reader's Guide to Periodicals samt andra ämnen som kunde vara relevanta för grundutbildningsstudenter.

---

<sup>14</sup> Med *exploratory* avser författarna att måtten som användes inte är förankrade i någon, sedan tidigare fastställd metod, utan utforskande på det viset att det är svårt att använda måtten i jämförelse med t.ex. andra studier.

Innan sökningarna utfördes studerades varje söktjänsts söktips och FAQs för att på så vis kunna optimera varje fråga. Man strävade efter hög *precision* snarare än hög *recall*, möjligen underförstått att det är omöjligt att räkna ut *recall*. Att vissa söktjänster erbjuder fler och mer avancerade funktioner innebär självklart att frågorna kan bli mer precisa och därmed också att man kan återvinna mer relevanta dokument.

Man relevansbedömde de tio första träffarna för varje fråga, i första hand utifrån den beskrivning som gavs i söktjänsternas resultatlistor och om relevansen inte kunde bedömas utifrån dessa gick man till de verkliga dokumenten. Dubbletter och spegelsidor räknades inte som separata relevanta träffar.

Magellan och Point gav inte alltid tio träffar, så man fick för dessa räkna ut ett genomsnittligt antal relevanta träffar. Överlag presenterade dessa söktjänster med granskat material färre träffar och var därmed också till mindre hjälp än de traditionella söktjänsterna som presenterade fler relevanta träffar.

Även här förvånades författarna av det låga antalet överlappningar i resultaten söktjänsterna sinsemellan.

Författarna förespråkar slutligen någon form av ”*keyword boxes*” som skaparna av webbsidorna kan fylla i. Detta låter mycket som de redan existerande metastandarderna, fast det i detta fall skulle röra sig om nyckelord som användaren ser direkt på sidan.

**Leighton & Srivastava** (1997) jämförde söktjänsterna Alta Vista, Excite, Hotbot, Infoseek och Lycos. Dessa fem valdes därför att de rekommenderats för återvinning av relevant information från tidigare undersökningar.

Utvecklandet av en testbädd utgörs enligt författarna av två steg: (1) välja ämnen för vilka man vill utföra sökningar, och (2) välja exakt vilka uttryck som skall användas vid varje fråga. Ämnena som låg till grund för de femton frågorna som användes i denna undersökning var baserade på verkliga frågor tagna från ett universitetsbiblioteks referensdisk. Frågornas exakta utformning var enligt författarna kanske den svagaste punkten i arbetet. Man valde att dela in frågorna i tre kategorier och antal – strukturerade (7), enkla (7) och personnamn (1). Strukturerade, d.v.s. man använde operatörer, valdes när ämnet var öppet för många tolkningar. De enkla, formulerade i naturligt språk, valdes med anledning av att de flesta användare vanligen inte använder operatörer.

Frågorna ställdes samma dag, de flesta inom en trettiominutersperiod, för att minimera risken för att databaserna ändrat sig. Man utvecklade även en metod för att dölja från vilken söktjänst en sida återvunnits, för att på så vis eftersträva en objektivitet i bedömningarna. Sidor som gav ”*server not responding*” till svar undersöktes senare vid ett flertal tillfällen.

Följande relevanskategorier, togs fram:

- Dubbletter, inaktiva länkar.
- Icke-relevanta dokument.
- Tekniskt sett relevanta dokument.
- Potentiellt användbara dokument.
- Högst sannolikt användbara dokument.

Det effektivitetsmått man valde att använda var ”*first twenty precision*”, d.v.s. antalet relevanta dokument bland de tjugo först återvunna, där träffar högt upp i listan gavs ett högre värde. Man följde Ding & Marchionini (1996), med den skillnaden att man ökade antalet undersökta dokument från tio till tjugo.

För att kunna analysera och jämföra ”*first twenty precision*” utvecklades en formel där resultatet för varje fråga tilldelas ett nummer mellan 0 och 1. Vidare delades de tjugo första träffarna upp i tre grupper vilka i sin tur tilldelades olika vikter – varje relevant träff bland de tre första fick 20, fyra till tio 17 och de tio sista 10. Detta alltså för att ge dokument högre upp i resultatlistan ett högre värde.

Eftersom det är möjligt att definiera vad som är ett relevant dokument på flera olika sätt valde man att utforma fem test för att pröva utfallet vid olika definitioner av relevans. Dessa olika definitioner var baserade på de relevanskategorier man tagit fram. Detta visade sig också få avgörande skillnader i resultaten, vilket kan vara intressant att ta upp. När en sida bedömdes som relevant när den tekniskt sett tillfredställde en sökning uppnåddes ett medianvärde på 0,81 med en toppnotering på 0,93 (Excite). Om definitionen var striktare och endast de potentiellt användbara sidorna bedömdes som relevanta uppnåddes ett medianvärde på 0,39 med en toppnotering på 0,51 (Infoseek). Om man slutligen endast bedömde högst sannolikt användbara sidor som relevanta hamnade medianen på 0,06 med en topp på 0,10 (Infoseek). Vidare genomfördes tester föra att se hur söktjänsternas resultat påverkades om dubletter och inaktiva länkar bestraffades alt. inte bestraffades.

Samma författare, Leighton & Srivastava (1999), ville granska de metoder som man använt och genomförde därför ytterligare en undersökning. Studien genomfördes delvis också p.g.a. söktjänsternas kontinuerliga förändring. Författarna fäster ingen större vikt vid förändringen av resultat utan poängterar att för att bättre första resultaten skulle man behöva känna till exakt vilka rankingsmetoder som de respektive söktjänsterna använder sig av.

**Clarke & Willett** (1997) valde att jämföra återvinningseffektiviteten hos de tre söktjänsterna Alta Vista, Excite och Lycos.

Med anledning av att de flesta tidigare studierna undvikit *recall*, ville man här närma sig detta problem genom att utveckla en metod för att ändå beräkna detta mått. Metoden man använde tog sin utgångspunkt i de, för IR-forskningen ända sedan cranefieldtesterna, traditionella komponenterna – ett antal dokument, ett antal frågor och ett antal relevansbedömningar för varje fråga.

Med anledning av webbens karaktär innebar metoden att det *recall* man räknade ut var relativt snarare än absolut till sin natur. Man använde sig i detta fall av den *pooling* metod som tagits fram vid TREC (se kap. 2.4.3). Vilket i detta fall innebar att man samlade alla relevanta dokument som samtliga söktjänster återvunnit för att på så vis beräkna, eller snarare uppskatta, det totala antalet relevanta dokument. En avgörande skillnad här var att de olika söktjänsterna hade sina egna databaser indexerade efter egna metoder, till skillnad från TREC där databasen är gemensam för alla system som testas.

Man använde sig av trettio frågor som konstruerats utifrån ämnen för forskning och uppsatser vid Department of Information Studies vid University of Sheffield, samt utifrån ämnen som var av personligt intresse för författarna. Det var även dessa som utförde relevansbedömningarna.

När det gällde frågesyntax vid sökningarna valde man att utgå från minsta gemensamma nämnare hos söktjänsterna.

Man relevansbedömde de tio första träffarna på en tregradig skala – 1 för relevant, 0,5 för delvis relevant och 0 för icke-relevant.

Utifrån tidigare studier arbetade man även fram följande relevanskriterier:

- En sida som matchade ämnet för frågan mycket väl fick 1 poäng.
- En sida som innehöll ett antal hyperlänkar som var användbara snarare än själva sidan fick 0,5 poäng.
- Dubbletter fick 0 poäng.
- Spegelsidor fick 0 poäng.
- Fick man ”*file not found*”, eller liknande, till svar för en specifik URL, tydde detta på att indexet inte var uppdaterat och sidan fick 0 poäng.
- Fick man meddelandet ”*There was no response. The server could be down or is not responding*”, för en specifik URL, kontrollerades sidan senare. Fick man samma svar igen fick den 0 poäng.
- Sidor på andra språk än engelska ersattes av den näst följande sidan på engelska som återvunnits.
- Poolen av relevanta dokument skapades genom att slå samman resultaten från de tre söktjänsterna.
- Överlappningar mellan söktjänsterna noterades och en andra sökning genomfördes för de relevanta sidor som inte återvunnits av samtliga söktjänster.

Frågorna för denna sistnämnda andra sökning konstruerades genom att ta text från en relevant sida som återvunnits av en eller två söktjänster. Hittade man då inte sidan drogs slutsatsen att sidan inte var tillgänglig för återvinning av den berörda söktjänsten.

De effektivitetsparametrar man beräknade var *precision*, relativ *recall* och *coverage*. *Coverage* definieras här som det totala antalet relevanta sidor som en söktjänst återvunnit delat med det totala antalet relevanta sidor som samtliga tre söktjänster återvunnit.

Man beräknade medelvärden för de tre måtten samt att man testade betydelsen av de eventuella skillnaderna i resultaten som fanns. Detta genom Friedmans signifikanstest.

Resultaten man fick fram var att Excite hade bästa *recall* 0,66 följt av Lycos 0,57 och AltaVista 0,56. När det gäller *precision* fick Alta Vista 0,46 följt av Excite 0,34 och Lycos 0,25. Slutligen *coverage* där Alta Vista fick 0,81, Excite 0,55 och Lycos 0,38.

Till skillnad från många andra studier fick man här en relativt hög grad av överlappningar, t.ex. var 104 av de totalt 353 relevanta sidorna man fick till svar från Alta Vista även med som svar från Excite.

Avslutningsvis poängterar författarna bristerna med att använda så få frågor, dessutom från samma ämne, men man menar dock metoden kanske är viktigare än resultatet i detta fall. Att den *pooling*-metod man använt sig av för att beräkna *recall* visat sig vara gångbar när det gäller utvärderingsstudier företagna på webben.

**Gordon & Pathak** (1999) mätte effektiviteten hos sju frågebaserade söktjänster – Alta Vista, Excite, Infoseek, OpenText, HotBot, Lycos och Magellan, samt en ämneshierarkisk söktjänst – Yahoo. Dessa valdes med anledning av att de ansågs representera de viktigaste söktjänsterna samtidigt som de tillsammans inkorporerade de flesta indexeringsmetoder med robot.

Fakultetsmedlemmar vid University of Michigan Business School fick fylla i formulär som var utformade för att ta fasta på informationsbehov. Dessa informationsbehov låg sedan till grund för de sökningar och sökfrågor som utfördes och konstruerades av vana informationssökare. Det var sedan dessas uppgift att utifrån ett informationsbehov optimera, bl.a. genom att pröva sig fram, varje fråga för varje enskild söktjänst. Sammanlagt rörde det sig om 33 frågor.

Man relevansbedömde de tjugo första träffarna. Vilket utfördes av de som utformat informationsbehoven. Genom att utgå från verkliga informationsbehov och sedan låta verkliga användare relevansbedöma det återvunna materialet, fick studien enligt författarna en autenticitet som de saknat i flertalet tidigare studier. Vid relevansbedömningarna användes följande nivåer av relevans:

- Högst relevant.
- Till viss del relevant.
- Till viss del irrelevant.
- Högst irrelevant.

I sin effektivitetsundersökning var författarna intresserade av två frågor. (1) Hur effektiv är en söktjänst när det gäller att återvinna endast relevanta sidor? (2) Är en enskild söktjänst kapabel till att finna de flesta av de existerande relevanta sidorna?

De mått man använde var *precision* och relativ *recall* hos de tjugo första träffarna, man beräknade även DCV:s vid flera nivåer.

Med anledning av precision kan nämnas att man bl.a. kom fram till att söktjänsterna var dåliga på att presentera de högst relevanta sidorna först, utan tenderade istället att sprida dessa över de tjugo första träffarna i resultatlistorna.

Överlag så presterade Alta Vista och OpenText bäst, medan Yahoo nådde de sämsta resultaten.

Precis som vid de flesta tidigare studier förvånades man över den låga nivån av överlappningar mellan söktjänsterna.

### **3.1.1 Sammanfattning utvärderingsstudier**

I följande stycke har jag valt att komprimera de olika aspekterna av de olika metoderna som använts vid tidigare utvärderingsstudier. Detta för överskådlighet samt för att underlätta arbetet med utformningen av min egen metod.

Val av söktjänster:

- Söktjänsterna representerade en bredd.

- Söktjänsternas olika rankingmetoder och gränssnitt gjorde dem till ett representativt urval.
- Ville undersöka både förgranskade och icke-förgranskade söktjänster.
- Söktjänsterna hade bedömts som bra på att återvinna relevant material i tidigare undersökningar.
- De ansågs representera de viktigaste söktjänsterna.

Frågornas ursprung, informationsbehov:

- Referenser av frågor i bibliotek.
- Övningar i Dialog.
- Ämne relevanta för grundutbildningsstudenter.
- Ämnen för forskning och uppsatser i informationsvetenskap.
- Informationsbehov framtagna med hjälp av att användare fick fylla i formulär.
- Personliga intressen.

Frågornas utformning:

- Optimerade utifrån sökfunktioner och gränssnitt hos söktjänsterna, bl.a. genom att studera FAQ och söktips.
- Jämförde olika frågesyntax utifrån funktioner, så som boolesk logik, fältsökning, trunkering och ord/frassökning.
- Strukturerade, enkla och personnamn.
- Minsta gemensamma nämnare hos söktjänsternas funktioner.

**Tabell 1. Antal söktjänster och frågor.**

Undersökning	Antal söktjänster	Antal frågor
Leighton (1995)	4	8
Chu & Rosenthal (1996)	3	10
Ding & Marchionini (1996)	3	5
Tomaiuolo & Packer (1996)	5	200
Leighton & Srivastava (1997)	5	15
Clarke & Willett (1997)	3	30
Gordon & Pathak (1999)	7	33

Relevansbedömningar gjordes dels av författarna själva och dels av användare. Följande poängbedömningar, kriterier och kategorier användes:

- Tregradig skala – 1 för relevant, 0,5 för något relevant och 0 för icke relevant.
- Sexgradig skala anpassad för varje enskild fråga.
- Tregradig skala - 1 för relevant, 0,5 för delvis relevant och 0 för icke relevant utifrån följande kriterier: 1 för sida som matchade frågan mycket väl 0,5 för sida med användbara hyperlänkar, 0 för dubletter, spegelsidor, "file not found" och "There was no response".
- Fyrgradig skala – högst relevant, till viss del relevant, till viss del irrelevant och högst irrelevant.
- Fem relevanskategorier –dubletter och inaktiva länkar ,tekniskt sett relevanta dokument, potentiellt användbara dokument och högst sannolikt användbara dokument.

Val av mått:

- *Precision* och responstid.
- *Precision, salience* och *relevance concentration*.
- *Precision*.
- *First twenty precision*.
- *Precision*, relativ *recall* och *coverage*.
- *Precision* och relativ *recall*.

Övrigt :

- Undersökte de 10 eller 20 första träffarna.
- Undersökte överlappningar.
- Ställde frågorna inom så kort tidsintervall som möjligt.
- Undersökte de verkliga dokumenten.
- Undersökte endast referenserna i söktjänsternas resultatlistor.
- Signifikantstade resultaten.
- Signifikantstade inte resultaten.

## 4 Metod

### 4.1 Ämnesval

I inledningen av detta arbete (se kap. 1.3) berördes ytterst kortfattat, att det ligger i mitt intresse att undersöka söktjänsters återvinningseffektivitet med frågor utifrån det akademiska ämnesområdet Antikens kultur och samhällsliv. Detta ämnesområde är alltså centralt för uppsatsen som helhet, och därmed krävs det också enligt min uppfattning en generell beskrivning av ämnet för att skapa en förståelse för informationsbehov och sökformuleringar. Den huvudsakliga anledningen till att just detta ämne valts, presenterades också i inledningen. Detta tillsammans med att jag läst 80 poäng i ämnet ger mig även den nödvändiga kompetens som krävs för att göra de relevansbedömningar som krävs.

Antikens kultur och samhällsliv hette från början Klassisk fornkunskap och antikens historia och blev en vetenskaplig disciplin vid svenska universitet vid början av 1900-talet. 1969 ändrades namnet till dess nuvarande form Antikens kultur och samhällsliv. Detta för att betona att det var förståelse av helhetsbilder av det antika samhällena som studiet syftar till, ett studium där man förutom arkeologi och historia även studerar religion, filosofi, konst och litteratur. Antikens kultur och samhällsliv behandlar främst de samhällen som växte fram i de områden som numera utgörs av Grekland och Italien, men intresseområdet täcker in hela medelhavsområdet. Den tidsperiod man studerar sträcker sig från stenålder fram till medeltid. Källorna går till vis del in i varandra, men man kan generellt sett säga utgöras av arkeologiskt material, arkitektur, konstföremål och litterära källor av olika slag.

### 4.2 Val av söktjänster

Som jag nämner i inledningen av arbetet är jag alltså intresserad av att mäta och jämföra återvinningseffektiviteten hos två frågebaserade söktjänster – dels en generell och dels en ämnesspecifik.

När det gäller en generell söktjänst spelar det egentligen inte, i detta sammanhang, så stor roll vilken man väljer, så länge man väljer bland de största och mest utnyttjade. De fungerar alla på liknande sätt, dock med vissa undantag och Google är ett exempel på ett sådant, som skiljer sig något från de övriga. Man kan alltså se valet här som en representant för stora frågebaserade söktjänster och i denna studie har jag valt Google som denna representant. Detta av följande anledningar:

- Google är, när detta skrivs, den söktjänst med flest indexerade sidor(Search Engine Watch 2001b).
- Google indexerar PDF-filer samt word- och exceldokument..
- Personliga erfarenheter av att använda Google.

Beträffande den ämnesspecifika söktjänsten var den första målsättningen att hitta en söktjänst inriktad på just det ämnesområde från vilka mina informationsbehov skulle hämtas. Som jag nämnde i inledningen så har det förts fram att generella söktjänster inte lämpar sig för akademiskt bruk. I den aktuella artikeln som jag refererar till nämns även Argos som ett exempel på en ämnesspecifik söktjänst mer lämpad att användas i en akademisk kontext (the Chronicle of Higher Education 1996). Argos förekommer även i listorna över söktjänster hos Archaeology (Archaeology 2001) – officiell publikation för Archaeological Institute of America (Archaeological Institute of America 2001), vilket även belyser söktjänstens anseende till viss del.

#### 4.2.1 Google

Enligt Landoni & Bell (2000) bör man för att kunna genomföra en utvärderingsstudie, samla och beskriva de funktioner och egenskaper som utmärker de aktuella söktjänsterna. Följande uppgifter, när inget annat nämns, är hämtade från Googles webbplats(Google 2001b).

Google grundades 1998 av Larry Page och Sergey Brin, doktorander vid Stanford University, vilka utvecklat en metod att ranka information som återvunnits av en söktjänst. Tekniken kallas *PageRank* och är enligt Google baserad på webbens ”demokratiska natur”(jfr kap. 2.7.2). Google tolkar en länk från sida A till sida B som en röst (*vote*) från sida A för sida B. Man tar även hänsyn till hur viktig en sida är. Länkarna från en sida som anses viktig rankas högre än länkar från en mindre viktig sida. För återvinningen används sedan ett textåtervinningsprogram. Från Googles håll hävdar man även att deras index uppdateras ungefär en gång per månad.

*PageRank* simulerar enligt Baeza-yates & Ribeiro-Neto (1999, s. 381) en användares navigering slumpvis på webben som hoppar till en slumpvis vald sida med sannolikheten  $q$  eller följer en slumpvis vald hyperlänk med sannolikheten  $1 - q$ . Om man låter  $C(a)$  vara antalet utgående länkar från sida  $a$  och antar att sida  $a$  pekar till sidorna  $p_1$  till  $p_n$ , så kan man enligt författarna definiera *PageRank* av  $a$ ,  $PR(a)$ , på följande vis, där  $q$  bestäms av systemet:

$$PR(a) = q + (1 - q) \sum_{i=1}^n PR(p_i)/C(p_i)$$

Föra att beräkna *PageRank* för en sida A måste man känna till *PageRank* för alla sidor som länkar till sida A, samtidigt som dessa sidors *PageRank* påverkas av om sida A eller någon annan sida länkar till dem. *PageRank* för sida A påverkas alltså av en länk från sida B,

samtidigt som fler länkar från sida B till övriga sidor minskar *PageRank* för sida A i detta fall. Det är detta som kallas för röst (*vote*) – om en sida B endast länkar till sida A får sida A ett högre *PageRank*-värde än om sida B även skulle länka till ett flertal andra sidor (Ridings 2001, s. 4). Den grundläggande idén här kan jämföras med Vektormodellen (se kap. 2.2.2) där man tar hänsyn till termförekomst och där termer som förekommer i många dokument i en samling får en lägre vikt än söktermer som enbart förekommer i vissa dokument i samlingen.

**Tabell 2a. Sökfunktioner hos Google.**

Nivåer av sökning	Implied Boolean <sup>15</sup>	Booleska operatörer	Närhets-sökning	Fras-sökning	Fält-sökning	Trunk-ering	Känner igen versaler	Läser meta-taggar	Tillgång till hjälp
enkel, avancerad, begränsad	+/- framför termer som måste vara med alt. uteslutas	automatisk AND, stöder OR, men inte NOT	automatiskt NEAR	frasen inom ""	domän, språk, titel, länk, URL, PDF, bild	nej	nej	nej	Help

I tabell 2a finns de olika sökfunktionerna hos Google presenterade. Vid en sökning presenteras hos Google antalet träffar och man kan välja mellan 10, 30 och 100 visade resultat. För varje träff finns följande information: rubrik på återvunna sidan, sammanfattning, adress, storlek och *cached link*<sup>16</sup>. Indraget resultat innebär att det hittats på samma webbplats som resultat ovan i listan.

#### 4.2.2 Argos

Följande information är hämtad från Argos (Argos 2001a). Argos utvecklades vid University of Evansville 1996 och är en frågebaserad söktjänst som indexerar ett förgranskat avgränsat material på webben. Söktjänsten benämns som LASE (*Limited Area Search Engine*). Målsättningen med Argos är bl.a. följande:

*“With Argos, we have aimed to create an academically viable resource for students, teachers and scholars of the ancient and medieval worlds.”*(Argos 2001b)

Söktjänsten är utformad så att man använder sig av ett tvåskiktprotokoll som bestämmer vad som ska genomsökas och inte genomsökas. Protokollet fyller två funktioner, dels begränsas området som genomsöks och dels kontrolleras den övergripande kvalitén på indexet. Detta sker genom att Argos söker ett mindre antal förgranskade medverkande webbplatser (*Associate sites*) och de sidor som dessa är länkade till, med undantag för sidor som Argos uteslutit, så som personliga hemsidor. Man förlitar sig alltså på det granskningsarbete som utförs vid de respektive medverkande webbplatserna. Man kan utifrån detta till viss del betrakta Argos som en metasöktjänst. Argos söker först och främst efter söktermerna i titelfältet, de dokument som anses relevanta sorteras sedan efter förekomst av söktermerna i dokumentet som helhet. Från Argos sida hävdar man att deras index uppdateras en gång per vecka.

<sup>15</sup> Fungerar ungefär som booleska operatörer fast man använder + och – tecken istället.

<sup>16</sup> Hyperlänk till sidan så som den såg ut när den indexerades av Google.

**Tabell 2b. Sökfunktioner hos Argos**

Nivåer av sökning	Implied Boolean <sup>17</sup>	Booleska operatörer	Närhets-sökning	Fras-sökning	Fält-sökning	Trunk-ering	Känner igen versaler	Läser meta-taggar	Tillgång till hjälp
enkel	nej	automatisk AND, stöder inte OR eller NOT	Nej	nej	nej	* före och/eller efter en term	nej	framgår ej	Help

Argos stöder ett ytterst begränsat antal funktioner avsedda för att kunna formulera en effektiv frågesyntax. Man ger inte heller på sin webbplats någon utförlig information om vad som stöds och vad som inte stöds. I tabell 2b kan man utläsa den information som ges på hjälpsidorna hos Argos. Vidare skrivs Ä skrivs som ae, ö som oe, ü som ue, æ som ae och tyskans ß som ss. Vid sökning med flera termer söker Argos alla dokument som innehåller den första av termerna och sedan efter dokument som även innehåller de övriga termerna. Detta betyder att om man t.ex. söker efter "Great Wall" så återvinns dokument som innehåller "Great" och "Wall", dock nödvändigtvis inte bredvid varandra. Även dokument som innehåller "Alexander the Great" och "He built a Wall", skulle i detta fall återvinnas.

Argos visar antalet återvunna dokument för varje fråga och presenterar 25 av dessa åt gången. Vidare ges information om dokumentens titel, en kort beskrivning av dem, deras adresser och storlek.

### 4.3 Val av informationsbehov och utformning av sökformuleringar

Vilket redan påpekats så menar Leighton & Srivastava (1999, s. 872) att utvecklandet av en testbädd utgörs av två delar: (1) Hitta och välja ut vilka informationsbehov som man skall söka efter, och (2) bestämma exakt vilka frågor och vilken syntax som skall användas.

#### 4.3.1 Informationsbehov

Gordon & Pathak (1999) skriver:

*"Experimenters who personally think up searches for an experiment may introduce biases into an experiment (say by composing searches which favor a particular search engine)..."* (Gordon & Pathak 1999, s. 146)

Även Landoni & Bell (2000, s. 125) poängterar det fördelaktiga med att utgå från verkliga informationsbehov. Med anledning av detta, har jag här 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, med ekonomiskt stöd från Forskningsrådsnämnden (Vetenskapsrådet 2001). Informationsbehoven är uppdelade i följande, för ämnet relevanta, kategorier:

- Antiken och nutiden.
- Fältarkeologi.
- Antika citat.
- Politik och styrelseformer.

<sup>17</sup> Fungerar ungefär som booleska operatörer fast man använder + och – tecken istället.

- Kvinnors villkor.
- Dagligt liv.
- Filosofi, idé- och lärdoms historia och religion.
- Teknik och vetenskap.
- Etruskerna.
- Det antika Athen.
- Det antika Rom och Romarriket.

Det sammanlagda antalet informationsbehov i denna frågespalt är 78 och antalet under varje kategori varierar, under vissa finns endast en fråga medan andra har fler.

### 4.3.2 Utformning av sökformuleringar

Vid en närmare betraktelse så visade det sig att alla informationsbehov, av olika anledningar, inte var lämpliga att ligga till grund för sökformuleringar. Mot denna bakgrund har jag valt att ställa upp ett antal kriterier för att ett informationsbehov skulle komma med i urvalet till att ligga som grund för en sökformulering.<sup>18</sup> Kriterierna var följande:

- Sökningen skulle kunna formuleras på engelska. T.ex. "Hur har antiken fått sitt namn och varför har den fått just namnet antiken?", rör det svenska namnet och kunde därför inte tas med.
- Informationsbehov som ansågs för generella valdes bort. T.ex. "Hur såg politiken ut i praktiken?" Jag måste i detta fall veta var någonstans.
- Informationsbehov som ansågs för komplexa valdes bort. T.ex. "Varför utvecklade Platon och Aristoteles ett förakt mot arbete och arbetande. Varför finns det inga inslag av detta i naturfilosofernas tankar?"
- Informationsbehov rörande felaktig fakta eller tidsperiod valdes bort.

Dessa urvalskriterier har självfallet också inneburit att jag i stor utsträckning personligen också styrt vilka sökformuleringar som kommit med i undersökningen, vilket därmed också innebär en begränsning av arbetets objektivitet. Jag är väl medveten om detta, men anser, detta till trots, att det viktigaste i sammanhanget är att jag utgått från verkliga informationsbehov.

Av de 78 informationsbehoven återstod det efter att de kontrollerats mot ovan nämnda kriterier 37. Av dessa 37 sällades sju slumpmässigt bort, så att 30 informationsbehov återstod. Dessa har sedan legat till grund för de 30 sökformuleringar som använts för respektive söktjänst. Jag valde antalet 30 till min stickprovsundersökning eftersom det är praxis inom statistisk metod att urvalet inte skall vara mindre än just 30 (Körner & Wahlgren 1998, s. 96). Eftersom söktjänsterna stöder olika funktioner när det gäller frågesyntax - Google erbjuder betydligt fler funktioner för att precisera sökformuleringarna än vad Argos gör, ligger det en fara i att jag i och med detta kommer att jämföra olika saker. Å andra sidan anser jag det inte som rättvisande om jag endast skulle använda de enklaste funktionerna hos Google, av den anledningen att det är dessa funktioner som stöds av Argos. Liksom i ett antal tidigare studier (se kap. 3.1) - Leighton (1995), Ding & Marchionini (1996), Tomaiuolo & Packer (1996) och Gordon & Pathak (1999), där man strävat efter att optimera frågesyntaxen hos sökformuleringarna för varje söktjänst, ligger det även i mitt intresse att optimera

---

<sup>18</sup> Jag använder fortsättningsvis benämningen sökformulering för informationsbehoven omvandlade till frågor till de respektive söktjänsterna.

sökformuleringarna hos varje enskild söktjänst, för att på så vis återvinna så många relevanta dokument som möjligt.

Jag har strävat efter att använda i grunden samma termer för respektive söktjänst( se tabell 3.). Hos Google har jag utfört mina sökningar i formuläret för avancerad sökning (Google 2001c), vilket bl.a. inneburit att söktermer bestående av olika böjningsformer av samma ord har placerats i fältet ”med något av dessa ord”.<sup>19</sup> Detta med anledning av att Google då placerar ett automatiskt OR mellan dem. För att återvinna dokument med olika former av samma ord hos Argos<sup>20</sup> har jag, när detta varit aktuellt, valt att trunkera dessa söktermer. I detta sammanhang har det även uppstått ett problem, då OR funktionen hos Google inneburit att även synonymer etc har kunnat tas med. Hos Argos, med automatisk AND, skulle detta innebära en katastrof, varför t.ex. synonymer inte kunnat användas. Jag menar dock fortfarande att Google inte bör bestraffas för detta faktum, och har därför i vissa fall försett Google med fler söktermer än Argos. Eftersom Google inte är ämnesspecifik som Argos har jag av uppenbara anledningar även valt att ta med *ancient* i vissa sökformuleringar hos Google. Samma sak gäller för informationsbehov nr. 9, där jag valt att utesluta Pythagoras sats, för att minimera risken att ett stort antal matematiska dokument skulle återvinnas.

**Tabell 3. Informationsbehov och sökformuleringar.**

Informationsbehov	Söktermer Google	Söktermer Argos
1. Skulle ni kunna översätta detta citat från latin till svenska? <i>Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit.</i> <sup>21</sup>	<i>“Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit”</i>	<i>Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit</i>
2. Hur styrdes Sparta?	[sparta spartan] govern society	sparta* govern* society
3. Hur var kvinnors ställning i Sparta?	women [sparta spartan]	women sparta*
4. Vad är hellenism?	hellenism	hellenism
5. Vilken var den hellenistiska världen?	”hellenistic world”	hellenistic world
6. Vad är en stadsstat?	ancient ”city state” polis	city state polis
7. Hur fungerade Athens demokrati?	[athens athenian] ancient democracy	athen*, democracy
8. Demokratins framväxt och fakta rörande demokratin i det antika Grekland.	ancient greece democracy	greece democra*
9. Vet man med säkerhet att Pythagoras personligen existerat?	pythagoras life existence -theorem	pythagoras life existence
10. Kan man säga att matematik var det ämne som lockade flest (även om de var få) kvinnliga vetenskapspionjärer under antiken?	[mathematic mathematics mathematician mathematicians] ancient women	mathematic* women
11. Hade antikens kvinnor någon makt?	ancient women [authority power]	women authority power
12. Var det någon skillnad på kvinnans ställning i Rom jämfört med i Athen?	ancient women [rome roman athen athenian]	women rom* athen*
13. Du vill veta mer om antiken mode och nöje; vad man gjorde. <sup>22</sup>	“ancient entertainment” [entertainments amusement amusements greece greek rome] roman]	pleasure* entertainment* amusement*

<sup>19</sup> I tabell 3 är dessa termer markerade inom hakparentes.

<sup>20</sup> Hos Argos finns som tabell 2b visar endast ett formulär för enkel sökning (Argos 2001b).

<sup>21</sup> Här engelsk översättning.

<sup>22</sup> Har här utgått från kommentarerna till det ursprungliga informationsbehovet, då detta inte fanns med. Därav utformningen.

**Tabell 3. forts.**

14. Jag undrar hur en teater såg ut och hur de fungerade?(de antika grekiska) <sup>23</sup>	ancient [Greece greek theatre theatres]	greek theatres
15. Vad spelade man upp där? (de antika grekiska)	ancient [greece greek theatre theatres play plays drama dramas]	greek theatre* play* drama*
16. Var kan vi få fakta om de olympiska spelen?	ancient "olympic games"	olympic games
17. Hur var vardagslivet under antiken?	ancient "daily life"	daily life
18. Vart kan jag finna prisuppgifter på allt från mat till vapen under kejsardömet? (helst under Nero)	[price prices rome] "roman empire"	price* rom* empire
19. text om Pompeji...bilder...hur grävde man upp all aska?	pompeii [texts pictures paintings excavation]	pompei* text* pictures excavation
20. På ungefär 400-talet e. Kr., vilken världsbild hade man då? Trodde man att jorden var platt eller rund? Cirkulerade solen runt jorden eller tvärt om?	"5th century AD" ["world picture" conception]	fifth century AD world picture conception*
21. Hur mycket vet man om de grekiska gudarna och pågår det forskning om dem?	"greek gods" religion research	greek gods religion research
22. Jag undrar också om historierna från den grekiska antiken som handlar om de olika stjärnbilderna finns samlade i någon bok?	ancient greek [constellations astronomy stars]	greek constellation astronomy star*
23. Vad är det för likheter och skillnader mellan Platons filosofi och Plotinos?	plato plotinus philosophy differences similarities	plato plotinus philosophy difference* similar*
24. Jag skulle vilja veta lite om Thales, var kan jag hitta information om honom?	thales philosopher	thales philosopher
25. Finns det i grekisk mytologi ett folk som associeras till myror?	greek mythology ants people	greek mythology ants people
26. Jag skulle vilja veta om det finns något organ för den senaste forskningen inom etruskologi?	"etruscan studies" [research agency institution authority]	etruscan studies research
27. När upphörde Akropolis i Athen att vara en försvarsborg?	acropolis [athens athenian fortification defence] ceased	acropolis athen* fortification defence ceased
28. Var kan jag få tag på information om de länder som lydde under Nero? Bara lite om deras geografi, invånare, uppror osv.	["roman empire" rome] Nero reign	roman empire rome Nero reign
29. Var kan jag få tag på kartor över Rom under Nero samt en bra karta över hela Romarriket?	["roman empire" rome map maps] nero	rom* empire map* nero
30. ...det står inte det jag är ute efter. Det är något så enkelt som en kronologisk översikt på Roms kejsare...	"roman emperors" [chronology chronological]	roman emperors chronology*

<sup>23</sup> Institutionens anmärkning inom parentes.

## 4.4 Utvärderingskriterier

Följande avsnitt är en ytterst central del av min metod och därmed också helt avgörande för uppsatsens resultat. Jag har valt att kalla det utvärderingskriterier vilket i sin tur, i stort, innefattar hur jag förhållit mig till relevans och vilket effektivitetsmått som använts. Det bör redan här nämnas att denna del av min metod är baserad på den metod som Leighton & Srivastava (1997) utvecklat, och som jag i följande avsnitt motiverar varför jag valt.

### 4.4.1 Relevanskategorier

I arbetet med att utforma en gångbar metod är det troligen i frågor kring hur man skall förhålla sig till relevans man stöter på flest problem. Många tidigare utvärderingsstudier, t.ex. Chu & Rosenthal (1996) och Clarke & Willett (1997), har utgått från flergradiga skalor i sina relevansbedömningar. Problemet med detta är att det finns en överhängande risk för att man inte är konsekvent i sina relevansbedömningar, att ett visst mått av subjektivitet gör sig gällande.

Som jag skrev i avsnittet om relevansbegreppet (se kap. 2.5) kommer jag att ta avstamp i föreställningen 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 detta att relevans betraktas som ett mått på till vilken nivå ett dokument tillfredsställer ett informationsbehov utifrån ett antal på förhand definierade kriterier. För att genomföra en undersökning utifrån dessa premisser har jag här alltså valt att utgå från Leighton & Srivastavas (1997) metod där man utformat ett antal generella kriterier för att kategorisera relevansen hos de återvunna dokumenten. Detta för att göra det möjligt att testa söktjänsternas effektivitet vid olika definitioner av relevans. Här föreligger givetvis en liknande risk för subjektivitet, som i fallet med relevansskalor, men jag anser att man också till viss del synliggör problematiken genom att resultaten i hög grad illustrerar vilken skillnad olika definitioner av relevans kan innebära. Följande kategorier med respektive kriterier användes vid relevansbedömningen:

- **Dubbletter** - dokument med i grunden samma URL<sup>24</sup> som ett dokument vilket förekom tidigare i resultatlistan hamnade i denna kategori, oavsett om det var relevant eller inte. Spegelsidor betraktades inte som dubletter.
- **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.<sup>25</sup>
- **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 allt. 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.

---

<sup>24</sup>URL (*Uniform Resource Locator*) – adress till en webbplats.

<sup>25</sup> Nummerbetäckningarna 404 och 606 är de felkoder som används av systemet och som visas på skärmen vid ovan nämnda fel.

- **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.

#### 4.4.2 *First twenty precision* som effektivitetsmått

Även valen av effektivitetsmått har varierat i de tidigare utvärderingsstudierna (se kap. 3.1). Jag har här valt att använda det mått som Leighton & Srivastava (1997) benämner *first twenty precision*, ett mått som tar hänsyn till hur bra söktjänsterna är på att återvinna relevanta dokument bland de 20 första träffarna. Det är alltså detta mått som jag använt för att kunna besvara de två första frågorna i min frågeställning. Anledningen till att valet föll på just detta mått är att jag betraktar det som lämpligt och användbart i en studie av det slag som jag tänkt genomföra. Jag har samtidigt, liksom Leighton & Srivastava, valt bort *recall* som mått, dels med anledning av att det är omöjligt att beräkna, och dels med anledning av att det i detta fall inte är så viktigt att veta hur många relevanta dokument det finns totalt på webben för en specifik fråga.

*First twenty precision* mäter även *precision* med vikter för rankingseffektivitet. Måttet tar alltså hänsyn till var, i 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.

För att kunna analysera och jämföra *first twenty precision* hos de två söktjänsterna har jag liksom Leighton & Srivastava använt en formel som ger resultatet för varje fråga ett ental mellan 0 och 1. Inledningsvis har jag alltså konverterat de ovan nämnda kategorierna till binära värden av 0 eller 1. Vid t.ex. ett test där man definierat relevansen efter dokument som tekniskt sett tillfredställt informationsbehovet (se kap. 4.4.3), tilldelades 1 till dokument som hamnade i kategorierna ett, två och tre, medan övriga kategorier tilldelades 0. Vidare har jag för att ge dokument högre upp i resultatlistorna ett högre värde, valt att dela in de första 20 träffarna i tre grupper – de första tre, näst följande sju och de sista tio. Varje dokument i de tre grupperna tilldelades sedan tre olika vikter – de första tre 20, näst följande sju 17 och de sista tio en vikt av 10. Anledningen till denna uppdelning är att det var 20 träffar som skulle undersökas, att man vanligen får upp 10 träffar på första sidan av resultatlistan samt att man vanligen får de tre första träffarna synliga på skärmen när resultatlistan först presenteras.

För att beräkna ett slutligt mått användes följande formel där  $R(i-j)$  är lika med relevanta träffar i positionerna  $i$  till och med  $j$  i resultatlistan och  $T$  är antalet återvunna dokument:

$$\frac{R(1-3) \times 20 + R(4-10) \times 17 + R(11-20) \times 10}{279 - ((20 - T) \times 10)}$$

Om en söktjänst för en fråga t.ex. återvann fem relevanta dokument bland de fem första skulle man få följande värde:  $(3 \times 20) + (2 \times 17) = 94$ . Skulle det istället vara de fem sista dokumenten som bedömdes som relevanta skulle man få följande värde:  $(5 \times 10) = 50$ .

Nämnumaren beräknades som summan av samtliga vikter upp till 20 där en söktjänst för en fråga återvunnit 20 dokument eller fler, alltså:  $(3 \times 20) + (7 \times 17) + (10 \times 10) = 279$ . Om en söktjänst för en fråga återvann färre än 20 dokument, så beräknades nämnumaren i detta fall genom att från 279 subtrahera 10 för varje dokument som resultatlistan understeg 20 med. Om en söktjänst t.ex. endast återvann 15 dokument skulle man få följande nämnumare:  $279 - (5 \times 10) = 229$ . I bilaga 1 finns ett antal verkliga exempel på beräkningar.

#### 4.4.3 Tester vid olika definitioner av relevans

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, baserade på Leighton & Srivastavas metod, med fem olika definitioner av relevans. Med testerna har jag även velat testa i vilken grad *precision* påverkas beroende på hur man förhåller sig till dubletter och inaktiva länkar. Det är alltså dessa tester som avses i den andra frågan i min frågeställning. För att förtydliga nedanstående resonemang har jag valt att i en bilaga sammanställa ett antal verkliga exempel på hur jag gått till väga för att få fram de olika resultaten för de olika testen (se bilaga 1).

De första tre testerna visar söktjänsternas prestanda vid olika trösklar för *precision*. Här bestraffades även söktjänsterna för eventuella dubletter och inaktiva länkar genom att dessa sänkte summan av täljaren men inte nämnumaren. 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 dubletter 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 dubletter bland de 20 första träffarna, togs dubletterna bort och nämnumaren beräknades som om resultatlistan hade bestått av 17 träffar. På detta vis bestraffades inte söktjänsterna för eventuella dubletter och inaktiva länkar i resultatlistorna eftersom både täljarens och nämnumarens 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.

## 4.5 Signifikanstest

Det jag med ett signifikanstest vill undersöka är om skillnaden mellan resultaten mellan de båda söktjänsterna beror på att det finns en verklig skillnad eller om det enbart kan förklaras med slumpen. För att kunna avgöra om det föreligger någon signifikant skillnad mellan de olika resultaten för *first twenty precision* som Google och Argos presenterade, har jag valt att med Wilcoxons teckenrangtest<sup>26</sup> signifikantstesta resultaten. Detta görs för att undersöka om resultaten utifrån stickprovets storlek (mina 30 sökformuleringar) kan sägas gälla generellt och med ett signifikanstest beräknas sannolikheten för att så är fallet.

Vid ett signifikanstest formulerar man en nollhypotes ( $H_0$ ) som säger att den effekt man letar efter är noll. Man ställer nollhypotesen mot en alternativ hypotes ( $H_1$ ) som innebär att den effekt man letar efter inte är noll. Man räknar sedan ut sannolikheten för de observerade mätvärdena om nollhypotesen vore sann. Utifrån denna sannolikhet bestämmer man sedan om nollhypotesen skall förkastas eller om den skall behållas. Att förkasta nollhypotesen innebär att anse att en alternativ hypotes är mest trolig. Följande hypoteser användes i denna studie.

$H_0$ : Söktjänsterna är lika effektiva med avseende på *first twenty precision*.

$H_1$ : Söktjänsterna är inte lika effektiva med avseende på *first twenty precision*.

De uträknade sannolikhetsvärdena jämförs sedan med s.k. signifikansnivåer. I detta fall innebär ett sannolikhetsvärde som överstiger signifikansnivån att det är osannolikt att man skulle få de observerade mätvärdena om den effekt man letar efter vore noll, att detta talar för att nollhypotesen kan förkastas och den alternativa hypotesen kan då anses vara mest trolig. Det bör även sägas att signifikansnivån är en fast gräns<sup>27</sup> som bestäms i förväg innan signifikanstestet genomförs. Oftast arbetar man med nivåerna 0,05, 0,01 och 0,001, vilka då betecknas som signifikant\*, signifikant\*\* resp. signifikant\*\*\*. Med Signifikant \*\* avser man alltså att det är 1% sannolikhet, men inte mindre, att den alternativa hypotesen kan förkastas. (Blom & Holmquist 1998, s. 123)

## 4.6 Praktiska aspekter

Sökningarna genomfördes mellan 2001-11-06 – 2001-11-20. Sökningarna utfördes så att de två sökformuleringarna för ett informationsbehov ställdes till de båda söktjänsterna, följt av nästa två sökformuleringar o.s.v. Samma informationsbehov behandlades alltid samma dag inom ett så kort tidsintervall som möjligt. De inaktiva länkar som kommit med i resultatlistorna kontrollerades ytterligare en gång en vecka efter första sökningen.

<sup>26</sup> Ekvationerna för dessa beräkningar är hämtade från Blom & Holmquist (1998, s. 264)

<sup>27</sup> För signifikansnivån i detta arbete har jag använt tabell 2 i Blom & Holmquist (1998, s. 373)

## 5 Resultat och diskussion

### 5.1 Resultat

#### 5.1.1 Fördelning av resultat – relevanta kategorier

Även om måttet *first twenty precision* tar hänsyn till var i resultatlistan ett dokument hamnar, vilken kategori det hamnar i o.s.v., har jag här även valt att i tabellform presentera själva fördelningen av träffarna. Detta dels för att läsaren skall kunna härleda resultaten för *first twenty precision*, och dels för att på ett mer explicit sätt illustrera fördelningen. Som man kan se i tabell 4a fördelade sig resultaten bland de relevanta kategorierna, högst naturlig, på det viset att ju striktare definition som användes desto färre träffar blev resultatet. Google presenterade sammanlagt i kategori ett 258 träffar, 185 i kategori två och 75 träffar i kategori tre. Motsvarande siffror för Argos var 228 träffar i kategori ett, 113 i kategori två och 15 i kategori tre. Tabell 4a illustrerar alltså hur träffarna fördelade sig i resultatlistorna. För att underlätta läsningen av tabellen kan vi använda den första sökformuleringen som exempel. Denna sökformulering är baserad på infobehov 1 i tabell 4a. Nästa del av tabellen hur många dokument som hamnade bland träffarna 1-3, 3-10 och 10-20 i de respektive kategorierna ett, två och tre. För den första sökformuleringen blev resultatet för Google att fyra dokument bedömdes som tillhörande kategori ett, samt att dessa dokument återfanns bland träffarna 4-10 i resultatlistan. För denna sökformulering bedömdes även tre dokument som tillhörande kategori tre, varav två av dessa återfanns bland träffarna 1-3 samt att ett av dem återfanns bland träffarna 4-10 i resultatlistan.

**Tabell 4a. Fördelning av resultat – relevanta kategorier.**

Info.behov	träff	Google			Argos		
		ett	två	tre	ett	två	tre
1	1-3			2	1		
	4-10	4		1			
	11-20						
2	1-3		2	1	1		2
	4-10	1	1	1	4	1	1
	11-20	3	3	2	6		
3	1-3		1	1	1	2	
	4-10	2	2	3	6	1	
	11-20	4	3	3	7	2	
4	1-3	2				2	
	4-10	4	1	1	3	3	
	11-20	8		1	7	2	
5	1-3		2	1	2		1
	4-10	1	2	4	2	2	
	11-20	6	2		3	4	
6	1-3			3	1	1	
	4-10	2	4	1	2	2	
	11-20	5	4		6	1	
7	1-3	1	2			1	1
	4-10	4	2	1	5	1	1
	11-20	5	3	1	4	2	

**Tabell 4a.forts.**

8	1-3	3			1	1	
	4-10	2	3		5		
	11-20	4	5	1	5	2	
9	1-3	3				1	2
	4-10	7			6		
	11-20	6	2	1	8		
10	1-3		1	1	1	1	
	4-10	6	1		1	2	
	11-20	6	2	1	8	1	
11	1-3		3		1		
	4-10	6			3	2	1
	11-20	6	2	1	6	1	
12	1-3	2			1		
	4-10	3	3		3	3	
	11-20	6	4		2	6	
13	1-3		2				
	4-10	2	3		2	1	
	11-20	7	2				
14	1-3		2		1		
	4-10	3	2	1	4		
	11-20	5	2	3	9		
15	1-3			2	2	1	
	4-10	2	2	3	1		
	11-20	4	4	1	5	1	
16	1-3			1		1	1
	4-10	2	1	2	1	1	1
	11-20	1	6	3	5	2	
17	1-3		2		1	1	1
	4-10	2	2	1	5	1	
	11-20	10			9		
18	1-3	1		2	1		
	4-10	2	4		4	2	
	11-20	8	1		6	1	
19	1-3	1	1			3	
	4-10	2	5			3	
	11-20	5	5				
20	1-3	3			1	1	
	4-10	6	1		4	1	
	11-20	7	3		4	5	
21	1-3		3		1	1	
	4-10	2	4	1	1	1	
	11-20	3	7		5	1	
22	1-3		1	2	1	2	
	4-10		6	1		3	
	11-20	3	7		3		

**Tabell 4a. forts.**

23	1-3	2			2	1	
	4-10	4	2		1	4	
	11-20	8	2				
24	1-3			3		1	1
	4-10		3	3		3	1
	11-20	1	3	5	2	5	1
25	1-3	1	1		3		
	4-10	2	3				
	11-20	9	1				
26	1-3	3			1		
	4-10	3	4		3	1	
	11-20	5	2		5	4	
27	1-3	2	1			2	
	4-10	6	1		2	1	
	11-20	4	1				
28	1-3		3		1		
	4-10	2	2	1	1	4	
	11-20	4	2		2	3	
29	1-3			3	3		
	4-10	2	4	1	4	1	
	11-20	1	7	1			
30	1-3		1	1	1		
	4-10	2	2	2	3	1	
	11-20	4	2		6	2	
summa	1-3	24	28	23	29	23	9
	4-10	86	97	28	76	45	5
	11-20	148	60	24	123	45	1
totalt	1-20	258	185	75	228	113	15

**5.1.2 Fördelning av resultat – icke-relevanta kategorier**

De icke-relevanta kategorierna var inaktiva länkar, dubbletter och kategori noll. I tabell 4b kan man se att Google sammanlagt presenterade 18 inaktiva länkar, 38 dubbletter och 20 dokument i kategori noll, medan Argos presenterade 40 inaktiva länkar, 90 dubbletter och 14 dokument i kategori noll. I tabellen redovisas inte var i resultatlistorna träffarna fördelade sig, då man med utgångspunkt från tabell 4a kan härleda detta.

**Tabell 4b. Fördelning av resultat - icke-relevanta kategorier.**

Info.behov	Google			Argos		
	inaktiva länkar	dubletter	kategori noll	inaktiva Länkar	dubletter	kategori noll
1	0	1	0	0	0	0
2	1	1	3	1	3	1
3	1	0	0	0	1	0
4	0	2	1	3	0	0
5	0	1	1	1	6	0
6	0	0	1	3	4	0
7	0	1	0	0	5	0
8	2	0	1	3	3	0
9	0	0	1	0	2	1
10	0	1	1	2	3	0
11	1	2	0	2	4	0
12	0	2	0	2	3	0
13	0	4	0	6	0	0
14	0	2	0	0	6	0
15	1	1	0	0	10	0
16	2	2	0	1	7	0
17	0	3	0	2	0	0
18	0	1	1	0	5	0
19	1	1	0	1	2	0
20	2	0	0	1	3	0
21	0	0	0	3	4	3
22	2	0	0	2	1	0
23	1	2	0	1	2	0
24	0	1	1	1	0	4
25	1	1	1	0	0	0
26	1	1	2	1	4	1
27	1	1	2	1	1	0
28	1	3	1	2	6	0
29	0	1	0	1	0	2
30	0	3	3	0	5	2
summa	18	38	20	40	90	14

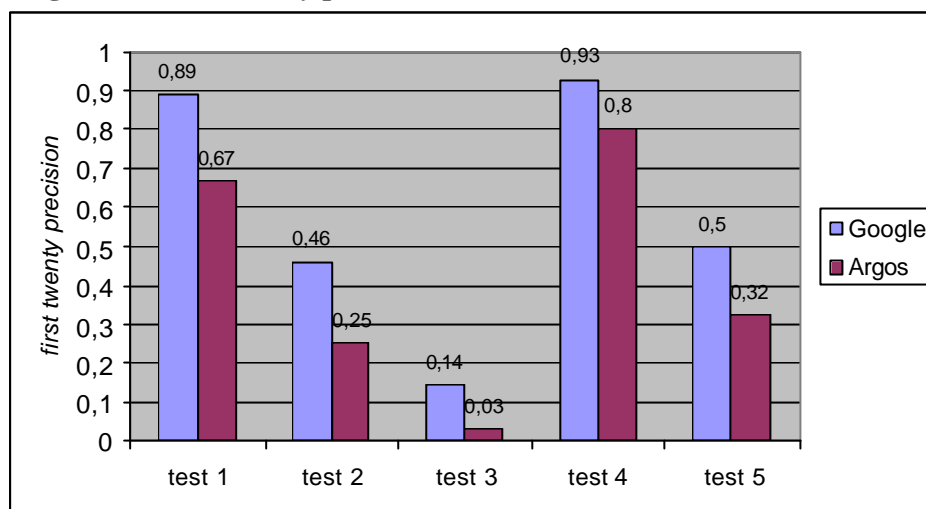
### 5.1.3 First twenty precision vid olika definitioner av relevans

I tabell 5 kan man utläsa de slutliga medelvärdena för resultaten för *first twenty precision*. I det första testet, med en låg tröskel för relevans, uppnådde Google ett resultat av 0,89 och Argos 0,67. Det andra testet och en moderat tröskelnivå resulterade i 0,46 för Google och 0,25 för Argos. Den tredje nivån för relevans och därmed också den högsta tröskelnivån för relevans, innebar för Google ett resultat på 0,14 medan Argos presterade 0,03. Det fjärde testet byggde på samma definition av relevans som det första, med den skillnaden att dubletter och inaktiva länkar här eliminerats från resultatlistorna. Detta innebar ett resultat på 0,93 för Google och 0,8 för Argos. Det femte testet motsvarande det andra, samt i likhet med det fjärde, att dubletter och inaktiva länkar tagits bort och gav resultaten 0,5 för Google och 0,32 för Argos. För att ytterligare illustrera de uppnådda resultaten har jag även valt att presentera dem i diagramform (se diagram 1).

**Tabell 5. First twenty precision medelvärden**

Söktjänst	test 1	test 2	test 3	test 4	test 5
Google	0,89	0,46	0,14	0,93	0,5
Argos	0,67	0,25	0,03	0,8	0,32

**Diagram 1. First twenty precision**



## 5.2 Resultat för signifikanstest

Utifrån Wilcoxons teckenrangtest presenteras i tabell 6 de aktuella sannolikhetsvärdena för varje test. Vid en signifikansnivå (se kap. 4.5) på 1% och därmed ett gränsvärde på 2,33 så överstiger samtliga sannolikhetsvärden för de fem testerna denna nivå. Man kan därmed konstatera att det föreligger ett signifikant\*\* samband mellan samtliga tester, samt att nollhypotesen -  $H_0$  kan förkastas och att det finns ett starkt stöd för mothypotesen  $-H_1$ . Detta innebär i praktiken att sannolikheten för att  $H_1$  kan förkastas endast är 1%.

**Tabell 6**

Test	sannolikhetsvärde
1	4,56
2	3,51
3	3,7
4	4,25
5	2,81

## 5.3 Diskussion

Vilket resultatet med tydlighet visar så presterade alltså Google bättre, med avseende på *first twenty precision*, i samtliga fem tester. I följande avsnitt diskuteras dessa skillnader i förhoppningen om att även finna förklaringar till dem.

Eftersom Argos är en specifik söktjänst för det aktuella ämnet, misslyckades söktjänsten sällan helt, d.v.s. att dokument i kategori noll kom med i resultatlistorna. De återvunna dokumenten berörde nästan alltid någon aspekt av ämnet för sökformuleringen. Trots en såpass låg tröskel för relevans som i det första testet, och trots att Argos sällan återvann dokument i kategori noll, presterade alltså Google även i detta första test ett signifikant bättre

resultat än Argos. Problemet för Argos, vilket även är en viktig del i förklaringen till det lägre resultatet i jämförelse med Google, och då främst i det första testet, står att finna i det stora antalet dubletter och inaktiva länkar som Argos presenterade i sina resultatlistor. Argos presenterade t.ex. för sökformulering 15 hela 10 dubletter, vilket givetvis också fick stora konsekvenser för *first twenty precision*. Detta förhållande illustreras tydligt av det fjärde testet, där dubletter och inaktiva länkar eliminerats från resultatlistorna, vilket också tydligt påverkade Argos i positiv riktning i högre grad än vad Google påverkades.

Ett ständigt återkommande exempel på dubletter med förödande konsekvenser för Argos var en webbplats där avhandlingar och uppsatser relevanta för ämnet finns indexerade. Med anledning av den höga termförekomsten rankades dessa dokument ofta högt och upptog inte sällan de fyra första träffarna av resultatlistorna. Indexet innehöll visserligen söktermerna men utan möjlighet att länka sig vidare till själva texterna. Indexet innehöll inte heller något mer än namnen på avhandlingarna och uppsatserna samt författare.

I ovan nämnda sammanhang blir även konsekvenserna av Argos andra stora begränsning uppenbar, nämligen de små möjligheterna att utifrån de sökfunktioner som stöds (se kap. 4.2.3) förfina sökformuleringarna. Även om Argos är ämnesspecifik så blir söktjänsten till ett relativt trubbigt instrument i informationsåtervinning av ett mer komplext slag - som man kan utgå från att det ofta handlar om i ett vetenskapligt sammanhang. Denna begränsning blev särskilt tydligt i testerna 2 och 3, där framförallt den sistnämnda resultatsiffran var häpnadsväckande låg med endast ett värde på 0,03 för *first twenty precision*. Att Argos endast tar hänsyn till termer i titel och den totala termförekomsten, men inte var i dokumenten termerna förekommer, och i vilket sammanhang, får till följd att Argos inte heller lyckas särskilt bra när det gäller att återvinna dokument med en högre tröskelnivå för relevans än i det första testet.

Inte heller Google presenterade särskilt många dokument i kategori noll – 20 mot 14 hos Argos. Tar man även i beaktande att Google sammanlagt återvann fler dokument än Argos blir denna siffra ytterst marginell. Google stöder en långt mer förfinad frågesyntax och rankingalgoritm än Argos, vilket också är en del i förklaringen till de få dokumenten i kategori noll. Google tycktes dock i vissa situationer, när det gällde att återvinna dokument i kategori tre, falla offer för sin egen rankingalgoritm *PageRank*. Exempel på detta är de många länksamlingar med länkar till högre relevanta dokument som återvanns och som hamnade i kategori två. Jag utgår från att även dessa högre relevanta dokument som det ofta länkades till, fanns med i Googles index, men att länksamlingarna av popularitet rankades högre. Ytterligare ett exempel på denna popularitetsrankning är sökformulering 27 där Google tog med ett dokument från *Lonely Planet* som handlade om Albanien och där endast två av termerna för sökformuleringen fanns med. Förmodligen kom detta dokument med just p.g.a. *Lonely Planets* popularitet. Den högsta tröskeln för relevans som det tredje testet utgjorde innebar även för Google ett lågt resultat för *first twenty precision*, om man t.ex. jämför detta resultat med resultatet som uppnåddes i test 1.

Hos Google ser man tydligt, med anledning av indraget i resultatlistan, när det rör sig om en dublett. Google presenterade alltså även ett avsevärt mindre antal dubletter än Argos (se tab. 4b) När det gäller inaktiva länkar och Google har jag som bekant utnyttjat funktionen *cached* (se not 16), då det verkliga dokumentet inte kunde nås. Detta har även fått till följd att Google returnerat ett ytterst begränsat antal inaktiva länkar som resultat. Då så har varit fallet har detta berott på problem med texten i *cached*. De få dubletterna och inaktiva länkarna hos

Google har fått till följd att siffrorna för resultaten i test 4 och 5 inte skiljer sig så mycket från resultaten testerna 1 och 2 hos Google som hos Argos.

Ytterligare en skillnad mellan de båda söktjänsterna var att Argos vid ett flertal tillfällen inte lyckades återvinna 20 dokument. Det finns förmodligen flera förklaringar till detta. Även om Argos inte ger någon siffra över det totala antalet indexerade sidor, är det högst troligt att Google indexerat fler, och att detta också kan ses som en förklaring till förhållandet. En annan förklaring ligger i redan nämnda problematik kring de begränsade sökfunktionerna, att mer komplexa sökformuleringar tenderade att resultera i mindre antal träffar, men då inte nödvändigtvis mer relevanta sådana.

Vad som inte undersökts i denna studie, men som dock kan vara av yttersta intresse, är källkritiska aspekter av de återvunna dokumenten. Om någon av söktjänsterna återvinner källkritiskt mer relevanta dokument än den andra? Med anledning av Argos *Associate Sites*, kan man t.ex. i denna studie endast spekulera i om detta även innebär källkritiskt mer relevant information. Detta skulle indirekt kunna undersökas genom att kontrollera överlappningarna av relevanta träffar söktjänsterna sinsemellan för att på så vis kunna dra slutsatser om det ofta är samma dokument som återvinns. Mer lämpligt vore kanske att utifrån, på förhand uppställda kriterier för källkritik, granska de återvunna dokumenten. I vilket fall som helst är ovan nämnda problematik värd att ha i åtanke vid fortsatt forskning.

Liksom med alla typer av undersökningar av detta slag finns det en inbyggd begränsning i och med den subjektivitet som uppstår i samband med relevansbedömningar. Detta gäller såväl Cranfield och TREC som med studier av återvinningseffektivitet på webben. Även om jag i denna undersökning utgått från verkliga informationsbehov samt ställt upp ett antal kriterier för relevansbedömning, kommer man inte undan denna subjektivitet, och gränsdragningarna mellan relevant och icke-relevant har därför också i många situationer varit mycket svåra, samt att det trots allt är jag som konstruerat sökformuleringarna. Detta faktum tillsammans med att man inte exakt kan uttala sig om hur söktjänsterna återvinner och rankar dokumenten, får till följd att de matematiska och statistiska beräkningarna som studien är uppbyggd kring inte är lika väl underbyggda som i ett sammanhang där denna subjektivitet och ovisshet kan undvikas. Det innebär givetvis också en begränsning att inte *recall* använts i undersökningen, att man inte kunnat uttala sig om andelen relevanta dokument som återvunnits. Vilket poängterats vid flertal tillfällen, finns det med webbens karaktär inte heller någon möjlighet att använda detta mått.

### 5.3.1 Slutsatser

Man kan således, med avseende på *first twenty precision* 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 precision*. 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.

## 6 Sammanfattning

Syftet med denna studie har varit att mäta och jämföra återvinningseffektiviteten hos två frågebaserade söktjänster, Google och Argos, med frågor från det akademiska ämnesområdet Antikens kultur och samhällsliv som utgångspunkt. Det har även legat i mitt intresse att undersöka i vilken utsträckning dessa typer av söktjänster klarar av att tillgodose ett informationsbehov grundat i en utbildningsmässig kontext. För att undersöka detta ställdes följande frågor.

- 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?

För att förankra mitt arbete i en teoretisk grund samt ge en bakgrundsbild av forskningsområdet inleds arbetet med en kortfattad genomgång av IR, IR-system och IR-modeller. I avsnittet om utvärdering av IR-system presenteras de vanligaste måtten *recall* och *precision*. I nästföljande avsnitt tas tre betydande utvärderingsstudier av återvinningseffektivitet upp och av vilka särskilt cranfieldmodellen kommit att utgöra ett paradigm. Nästa avsnitt behandlar relevansbegreppet. Här belyses problematiken kring relevans och dess subjektivitet och det konstateras bl.a. att det existerar olika typer av relevans samt att relevans kan ses som ett mått på effektiviteten av kontakten mellan en källa och en destination i en kommunikationsprocess. Nästa avsnitt behandlar IR och webben och inleds med att skilja mellan utvärderingar av söktjänster gjorda av t.ex. populärtidskrifter och utvärderingar gjorda som typiska IR-utvärderingar. Det konstateras att *recall*, p.g.a. webbens karaktär, är omöjligt att beräkna varför också utvärderingsstudier på webben skiljer sig något från traditionella utvärderingar baserade på cranfieldmodellen. För att till viss del handskas med denna problematik presenteras utifrån Oppenheim et al. (2000) fyra metoder. I detta kapitel presenteras även en kortfattad historik kring söktjänster och webben följt av en genomgång av olika typer av söktjänster. Utifrån denna genomgång definieras söktjänsterna för undersökningen i detta arbete som frågebaserade söktjänster.

Kapitlet om tidigare forskning består av en genomgång av sju tidigare utvärderingsstudier av söktjänster på webben. Valet av utvärderingsstudier för denna genomgång är baserat på två av de metoder som Oppenheim et al. presenterar – de som använt sig av relativ *recall* och de som undvikit *recall* som mått helt och hållet. Syftet med detta är att ge en metodologisk bakgrund, att lyfta fram hur man i tidigare studier förhållit sig till vilka söktjänster som skall undersökas, vilka och hur många frågor som skall användas, hur man skall förhålla sig till relevans samt vilka mått som använts.

Metodkapitlet inleds med en presentation av ämnet Antikens kultur och samhällsliv och fortsätter med motivering av valet samt beskrivning av söktjänsterna Google och Argos. Detta följs av en genomgång av informationsbehov och sökformuleringar. Jag utgår i arbetet från verkliga informationsbehov hämtade från Antiken på Internet. Utifrån dessa konstruerades 30 sökformuleringar till respektive söktjänst. De utvärderingskriterier som använts är baserade på Leighton & Srivastavas (1997) metod där man valt att utforma ett antal generella kriterier för att kategorisera relevansen hos de återvunna dokumenten. Effektivitetsmålet *first twenty precision* som använts är också hämtat från Leighton & Srivastavas metod, ett mått som tar hänsyn till hur bra söktjänsterna är på att återvinna relevanta dokument bland de 20 första

träffarna. Med anledning av att man kan definiera vad som är relevant på en mängd olika sätt har jag i studien valt att mäta och jämföra återvinningseffektiviteten i fem olika tester med olika definitioner av relevans. Metodkapitlet består slutligen av hypoteser för att utifrån Wilcoxons tecken-rangtest signifikantesta resultaten för *first twenty precision*.

I det avslutande kapitlet, resultat och diskussion, presenteras slutligen fördelningen av träffarna och resultaten för *first twenty precision*. Det konstateras att Google uppnår ett signifikant bättre resultat än Argos i samtliga fem tester. Som möjlig förklaring till detta förhållande framhålls Argos begränsade möjligheter att stöda förfinade sökformuleringar och dess rankingalgoritm. Ytterligare en förklaring till Argos sämre resultat stod att finna i det stora antalet dubletter som presenterades.

# Källförteckning

Antiken på Internet (2001) *Frågelåda*.

<http://www.hum.gu.se/~akswww/Questionbox/first.html> [2001-10-22]

Archaeological Institute of America (2001) <http://www.archaeological.org/> [2001-11-19]

Archaeology (2001) *The world of archaeology – general archaeology*.

<http://www.archaeology.org/cgi-bin/site.pl?page=wwwarky/general> [2001-11-19]

Argos (2001a) *About Argos*. <http://argos.evansville.edu/about.htm> [2001-10-19]

Argos (2001b) *Argos*. <http://argos.evansville.edu/>

Back, J. (2000) An Evaluation of Relevancy Ranking Techniques used by Internet Search Engines. *Library & Information Research News* 24(77), s. 30-34.

Bell, D. (1973) *The Coming of Post-Industrial Society – A Venture in Social Forecasting*. New York: Basic Books.

Baeza-Yates, R. & Ribeiro-Neto, (1999) *Modern Information retrieval*. New York: ACM.

Blom, G. & Holmquist, B. (1998) *Statistikteori med tillämpningar*. 3. uppl. Lund: Studentlitteratur.

Chowdhury, G.G. (1999a) The Internet and Information Retrieval Research a brief Review. *Journal of Documentation* 55(2), s. 211-225.

Chowdhury, G.G. (1999b) *Introduction to Modern Information Retrieval*. London: Library Association Publishing

*The Chronicle of Higher Education* (1996) Scholar Creates Search Engine for “Peer-Reviewed” Material oct. 18 s. 23.

Chu, H. & Rosenthal, M. (1996) Search engines for the World Wide Web: a comparative study and evaluation methodology. *ASIS 1996 Annual Conference Proceedings, Baltimore, MD, October 19-24, 1996*, s. 127-135.

Clarke, S.J. & Willet, P. (1997) Estimating the recall performance of web search engines. *Aslib Proceedings* 49(7), s. 184-189.

Clarke, S.J. (2000) Search engines for the world wide web: an evaluation of recent developments. *Journal of internet cataloging* 3/4 2000, s. 81-93.

Ding, W. & Marchionini, G. (1996) A comparative study of web search service performance. *ASIS 1996 Annual Conference Proceedings, Baltimore, MD, October 19-24, 1996*, s. 136-142.

- Dong, X. & Su, L.T. (1997) Search Engines on the World Wide Web and Information Retrieval from the Internet: a Review and Evaluation. *Online & CDROM Review* 21(2), s. 67-81.
- Google (2001a) *Google*. <http://www.google.com> [2001-10-05]
- Google (2001b) *All About Google*. <http://www.google.com/about.html> [2001-10-19]
- Google (2001c) *Google Advanced Search*. [http://www.google.com/advanced\\_search](http://www.google.com/advanced_search)
- Gordon, M. & Pathak, P. (1999) Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management* 35, s. 141-180.
- Gudivada, V.N. et al. (1997) Information Retrieval on the World Wide Web, *IEEE Internet Computing*, sep-oct, s. 58-68.
- Harter, S.P. & Hert, C.A. (1997) Evaluation of information retrieval systems: approaches, issues, and methods. *Annual Review of Information Science and Technology*, 32, 1997, 3-79.
- Korfhage, R.R. (1997) *Information Storage and Retrieval*. New York: Wiley.
- Körner, S. & Wahlgren, L. (1998) *Statistiska metoder*. Lund: Studentlitteratur.
- Landoni, M. & Bell, S. (2000) Information retrieval techniques for evaluating search engines: a critical overview. *Aslib Proceedings* 52(3), 124-129.
- Lawrence, S. & Giles C.L. (1999) Accessibility of information on the web. *Nature* vol 400, 8 july, s. 107-109.
- Lawrence, S. (2000) Context in Web Search *IEEE Data Engineering Bulletin* 23(3), s. 25-32.
- Lebedev, A. (1997) *Best Search engines for finding scientific information in the Web*. <http://www.chem.msu.su/eng/comparison.html>
- Leighton, H.V. (1995) Performance of Four World Wide Web (WWW) Index Services: Infoseek, Lycos, Webcrawler and WWWorm. <http://www.winona.msus.edu/library/webind.htm>.
- Leighton, H.V. & Srivastava, J. (1997) *Precision among World Wide Web search services (search engines): Alta Vista, Excite, HotBot, Infoseek, Lycos*. <http://www.winona.msus.edu/library/webind2/webind2.htm>
- Leighton, H.V. & Srivastava, J. (1999) First 20 Precision among World Wide Search Services (Search Engines) *Journal of American Society for Information Science* 50(10) s. 870-881.
- Lyman, P. & Varian, H.R. (2000) *How much Information?* <http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html#intro> [2001-10-15]

- Mizzarro, S. (1997) Relevance: The whole history. *Journal of the American Society for Information Science*. 48(9), 1997, s. 810-832.
- Nicholson, S. (2000) Raising Reliability of Web Search Tool Research through Replication and Chaos Theory. *Journal of the American Society for Information Science* 51(8) s. 724-729.
- Oppenheim, A., Morris, A. & McKnight, C. (2000) Progress in documentation the evaluation of www search engines. *Journal of Documentation* 56(2) 2000, s. 191-211.
- Ridings, C. (2001) *Page Rank Explained or Everything you've always wanted to know about Page Rank* <http://search.engine-submission.co.uk/> [2002-02-19]
- Saracevic, T. (1997) Relevance: a review of and a framework for the thinking of the notion in information science, i Sparck Jones, K. & Willett, P.(eds.) *Readings in Information Retrieval*. San Fransisco, s. 143-165.
- Savoy, J. & Picard, J. (2001) Retrieval effectiveness on the web. *Information Processing and Management* 37, s. 543-569.
- Schwarz, C. (1998) Web Search Engines. *Journal of the American Society for Information Science* 49(11), s. 973-982.
- Schatz, B.R. (1997) Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science* vol 275, 17 jan, s. 327-334.
- Schimmrich, S.H. (1997) Searching the World Wide Web for Geoscience Resources. *Computers & Geoscience* 23(5), s. 559-562.
- Search Engine Watch (2001a) *Search Engine Watch*. <http://searchenginewatch.com/>
- Search Engine Watch (2001b) *Search Engine Sizes*. <http://searchenginewatch.com/reports/sizes.html> [2001-10-17]
- Search Engine Watch (2001c) *Speciality Search Engines*. [http://searchenginewatch.com/links/Specialty\\_Search\\_Engines/](http://searchenginewatch.com/links/Specialty_Search_Engines/) [2001-11-08]
- Smeaton, A.F. & Harman, D. (1997) The TREC experiments and their impact on Europe. *Journal of Information Science* 23(2), s. 169-174.
- Sparck Jones, K. & Willett, P.(eds.) (1997) *Readings in Information Retrieval*. San Fransisco.
- Sullivan, D. (1998) Counting Clicks and Looking at Links. *Search engine watch* <http://www.searchenginewatch.com/sereport/98/08-clicks.html> [2001-10-15]
- Sullivan, D. (2001) Google Does PDF & Other Changes. *Search engine watch* <http://www.searchenginewatch.com/sereport/01/02-google.html> [2002-01-10]
- Svenska datatermgruppen (2001) *Term- och språkmateriel version19, 24 januari 2001*. <http://www.nada.kth.se/dataterm/>, [2001-09-16]

Tague-Sutcliffe, J. (1997) The Pragmatics of Information Retrieval Experimentation, Revisited. I Sparck Jones, K. & Willett, P.(eds.) *Readings in Information Retrieval*. San Fransisco, s. 205-216.

Telecordia Technologies (2001) *Evaluating the Size of the Internet*. <http://www.netsizer.com/> [2001-10-15]

Tomaiuolo, N.G. & Packer, J.G. (1996) An analysis of internet search engines: assessment of over 200 search queries. *Computers in Libraries* 16(6), s. 58.

Vetenskapsrådet (2001) *Vetskap*. <http://www.vetskap.frn.se/> [2001-10-22]

Zeizig, M. & Lattermann, A. (1996) *Stora guiden om World Wide Web*. Stockholm: IDG.

# Bilaga 1

Här nedan följer ett verkligt exempel på hur jag räknat ut resultaten för de olika fem testerna och de fem olika definitionerna av relevans. Exemplet gäller informationsbehov tre hos Google. Från tabell 4a och 4b kan man hämta följande information:

## Utdrag från tabell 4a.

		Google		
Info.behov	träff	ett	två	tre
3	1-3		1	1
	4-10	2	2	3
	11-20	4	3	3

## Utdrag från tabell 4b.

		Google		
Info.behov		inaktiva länkar	dubletter	kategori noll
	3	1	0	0

Med informationen från de båda tabellerna kan man sedan göra beräkningarna för de fem testerna enligt följande:

### Test 1

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. För detta informationsbehov gjordes därför följande beräkning:

$$\frac{(1+1) \times 20 + (2+2+3) \times 17 + (4+3+3) \times 10}{279} = 0,93$$

Bland träffarna 1-3 har vi alltså här två relevanta träffar (1+1) och en inaktiv länk, bland träffarna 4-10 har vi sju relevanta träffar (2+2+3) och bland träffarna 10-20 har vi tio relevanta träffar (4+3+3). Detta divideras sedan med 279 och vi får resultatet 0,93 för informationsbehov tre och den lägsta tröskeln för relevans.

### Test 2

Det andra testet utgick från en moderat tröskelnivå genom att tilldela 1 till samtliga dokument som återfanns i kategorierna två och tre och följande beräkning gjordes:

$$\frac{(1+1) \times 20 + (2+3) \times 17 + (3+3) \times 10}{279} = 0,66$$

Bland träffarna 1-3 har vi här alltså två relevanta, bland träffarna 4-10 har vi fem relevanta och bland träffarna 10-20 har vi 6 relevanta. Detta divideras med 279 och vi får resultatet 0,66 för detta test. Skillnaden här jämfört mot test 1 är att kategori ett dokument tagits bort.

### Test 3

Det tredje testet utgick från en hög tröskelnivå genom att endast tilldela 1 till de dokument som återfanns i kategori tre och följande beräkning genomfördes:

$$\frac{20 + (3) \times 17 + (3) \times 10}{279} = 0,36$$

Här återfinns inga relevanta träffar bland träffarna 1-3, tre bland träffarna 4-10 och tre bland träffarna 10-20. Liksom i de två tidigare testerna dividerades detta med 279 och resultatet för det tredje testet blev 0,36. skillnaden här jämfört med test två är alltså att även de dokument som klassades som kategori två har tagits bort från beräkningen.

### Test 4

I det fjärde och femte testet eliminerades dubletter och inaktiva länkar från resultatlistorna samtidigt som jag behandlade återstoden som en resultatlista bestående av färre än 20 återvunna dokument.

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. Följande beräkning gjordes:

$$\frac{(1+1) \times 20 + (2+2+3) \times 17 + (4+3+3) \times 10}{279 - (1 \times 10)} = 0,96$$

Detta test är alltså identiskt med test 1, med det undantaget att söktjänsten här inte bestraffas för den inaktiva länk som presenterats i resultatlistan. Detta beräknas genom att 10 subtraherats från nämnaren. Även om det här endast rör sig om en inaktiv länk blir resultatet något högre än i test – 0,96 istället för 0,93.

### Test 5

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. Följande beräkningar genomfördes med detta som utgångspunkt:

$$\frac{(1+1) \times 20 + (2+3) \times 17 + (3+3) \times 10}{279 - (1 \times 10)} = 0,69$$

På samma sätt som test 4 bygger på test 1 så bygger test 5 på test 2. Även här rör det sig alltså om en inaktiv länk som sänker summan för nämnaren och vi får resultatet 0,69 till skillnad från 0,63 som i test 2.