

MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP  
VID INSTITUTIONEN BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN  
2007:21  
ISSN 1654-0247

# Ask.com, Web Wombat och Yahoo

En studie av två globala sökmotorer och en lokal sökmotor

JONAS EKSTEIN  
CHRISTIAN RUNESSON



© Ekstein/Runesson  
Mångfaldigande och spridande av innehållet i denna uppsats  
– helt eller delvis – är förbjudet utan medgivande.

Svensk titel: Ask.com, Web Wombat och Yahoo: En studie av två globala och en lokal sökmotor.

Engelsk titel Ask.com, Web Wombat and Yahoo: A study of two global and one local search engines.

Författare: Jonas Ekstein & Christian Runesson

Kollegium: 2

Färdigställd: 2006

Handledare: Anders Stenström

Abstract: This thesis is focusing on how global and local search engines retrieve information from the local domain. The three search engines tested are the global search engines Yahoo, Ask.com and the local search engine Web Wombat. The questions we examined were: which search engine has the best retrieval effectiveness? Could there be other reasons than retrieval effectiveness, to choosing a local search engine? For our test we constructed 20 questions related to Australia. We chose to divide the questions into topics like nature, sports and culture. For all questions we evaluated the relevance of the first 20 hits. We used the following measures in our test: Jaccard's index, precision and average precision. We also looked at factors such as duplicates and error pages, because we consider this to be an important aspect to consider, when looking at the relevance of the first 20 hits. The results of our study showed that Yahoo had the best performance for precision. Web Wombat had faulty precision but results from Jaccard's index revealed that Web Wombat had many unique documents. Web Wombat had the best average precision on one of our questions. In spite of Web Wombats faulty precision, we think that Web Wombat serve a purpose as an alternative to global search engines.

Nyckelord: Yahoo, Ask.com, Web Wombat, sökmotorer, Australien, återvinningseffektivitet, precision, Jaccard's index

1. INLEDNING .....	4
1.1 PROBLEMFÖRMULERING.....	5
1.2 SYFTE OCH FRÅGESTÄLLNINGAR .....	5
1.3 AVGRÄNSNINGAR.....	5
2. SÖKMOTORER .....	6
2.1 GLOBALA SÖKMOTORER.....	6
2.2 LOKALA SÖKMOTORER.....	6
2.3 METASÖKMOTORER .....	6
2.4 HUR FUNGERAR EN SÖKMOTOR? .....	7
2.5 UTVECKLINGEN AV SÖKMOTORER.....	10
2.6 INFORMATION RETRIEVAL TEORI.....	12
2.6.1 Olika sätt att mäta återvinningseffektivitet .....	12
2.6.2 Rankning.....	13
2.6.3 Olika söktekniker .....	15
3. TIDIGARE FORSKNING.....	16
3.1 COMPARING RANKINGS OF SEARCH RESULTS ON THE WEB.....	17
3.2 COVERAGE, RELEVANCE, AND RANKING: .....	18
THE IMPACT OF OPERATORS ON WEB SEARCH ENGINE RESULTS. ....	18
3.3 THE EFFECTIVENESS OF WEB SEARCH ENGINES .....	19
FOR RETRIEVING RELEVANT ECOMMERCE LINKS. ....	19
3.4 MEASURING SEARCH ENGINE QUALITY .....	21
3.5 THINK LOCAL, SEARCH GLOBAL? COMPARING SEARCH ENGINES FOR SEARCHING GEOGRAPHICALLY SPECIFIC INFORMATION .....	22
3.6 NEW MEASUREMENTS FOR SEARCH ENGINE EVALUATION PROPOSED AND TESTED. ....	24
3.7 HOW DO SEARCH ENGINES RESPOND TO SOME NON-ENGLISH QUERIES?.....	26
3.8 GLOBALA OCH LOKALA SÖKMOTORER: EN UTVÄRDERING AV GOOGLE, MSN SEARCH OCH SVESÖK .....	27
3.9 SAMMANFATTNING .....	28
4. METOD .....	30
4.1 INFORMATIONSBEHOVET .....	30
4.2 SÖKFRÅGORNA.....	31
4.3 RELEVANSBEDÖMNING AV ÅTERVUNNA DOKUMENT.....	32
4.4 ÅTERVINGSEFFEKTIVITETEN .....	32
4.5 BESKRIVNING AV SÖKMOTORERNA .....	34
4.5.1 Ask.com.....	34
4.5.2 Web Wombat.....	35
4.5.3 Yahoo.....	35
5. RESULTAT .....	36
5.1 ASK.COM.....	36
5.1.1 Ask.com precision vid DCV 20.....	36
5.1.2 Ask.com genomsnittsprecision.....	37
5.2 WEB WOMBAT .....	41
5.2.1 Web Wombat precision vid DCV 20 .....	41
5.2.2 Web Wombat genomsnittsprecision .....	42
.....	44
5.3 YAHOO .....	45
5.3.1 Yahoo precision vid DCV 20 .....	45
5.3.2 Yahoo genomsnittsprecision .....	46
5.4 ALLA TRE SÖKMOTORER.....	50
5.4.1 Jaccards index.....	50
5.4.2 ANTAL FELMEDDELANDEN .....	53
5.4.3 Dubbletter.....	53

6. DISKUSSION .....	55
6.1 DISKUSSION AV RESULTAT .....	55
6.2 SÖKFRÅGOR OCH SÖKTERMER .....	56
6.3 DISKUSSION RUNT TIDIGARE FORSKNING .....	57
6.4 ALLMÄNT OM SÖKMOTORER .....	58
7 SLUTSATSER .....	59
8. SAMMANFATTNING .....	61
KÄLLFÖRTECKNING .....	63

# 1. Inledning

I en globaliserad värld erbjuds vi som användare en hel uppsjö av olika sökverktyg som kan användas för att lösa våra informationsbehov. Ett av dessa sökverktyg är en sökmotor. Andra exempel på sökverktyg är ämneskataloger och bibliotekskataloger. Vi tror att många användare bara känner till ett litet antal sökmotorer och några känner bara till de mest populära och omtalade sökmotorerna. Exempel på omtalade och populära sökmotorer är Google, Yahoo, Allthewebb, Ask.com etc.

Alla dessa sökmotorer är globala, vilket innebär att användaren kan söka efter dokument som täcker in stora delar av webben. De flesta och mest kända sökmotorerna kommer från USA och indexerar främst dokument på engelska<sup>1</sup>. Ett exempel på detta är att enligt forskningsundersökningen *How do search engines respond to some non-English queries?* så är bara cirka 3 % av de franska dokument som är indexerade<sup>2</sup>.

Uttrycket att ”googla” tror vi har lett till att användare kanske bara använder sig av en sökmotor, Exempelvis användare använder sig av Google. En anledning kan vara att sökmotorn har ett stort index och verkar återvinna många dokument.

Det många människor kanske inte är medvetna om eller känner till är att många länder har egna sökmotorer. Dessa sökmotorer kallas för lokala sökmotorer. I Sverige har vi t. ex. den lokala sökmotorn Svesök. Den indexerar dokument som handlar om Sverige. Motsvarigheten för Australien heter Web Wombat och den indexerar dokument från Australien samt Nya Zeeland.

Vi ville från början testa två lokala australiensiska sökmotorer, Web Wombat och Howzat mot Ask.com. Vi märker dock ganska snart att Howzat inte var lämplig att använda i undersökningen, eftersom den hade så få dokument indexerade. Ett exempel en sökning på Sidney gav bara 599 träffar. Ett fåtal dokument skulle leda till en missvisande precision i relation till de andra sökmotorerna. Så istället för Howzat valde vi den globala sökmotorn Yahoo.

Vi vill i uppsatsen utreda om det kan finnas några generella slutsatser rörande globala och lokala sökmotorer. Vi kommer att utveckla detta närmare under kapitlet 1.2 syfte och frågeställning.

---

<sup>1</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. s 374

<sup>2</sup> Bar –Ilan Judith och Gutman Tatyana(2005) *How do search engines respond to some non-English queries?*

## **1.1 Problemformulering**

De lokala sökmotorerna står ofta i skuggan av de globala sökmotorerna. Vi tror också att webbsidor som utvärderar olika sökmotorer förstärker denna bild genom att de bara verkar ta hänsyn till storleken hos sökmotorn. Alltså hur stort index sökmotorn har. Några exempel på webbsidor som utvärderar sökmotorer är Search Engine Showdown och Search Engine Watch.com. Det vi har utrett i vår uppsats är, om en lokal sökmotor kan prestera bättre än global sökmotorer vid sökning på specifik geografisk information. I vårt fall Australien. Om så inte är fallet, finns det då andra anledningar till att använda en lokal sökmotor?

Inom Biblioteks- och informationsvetenskap föreslås det ofta att användaren vänder sig till flera olika informationskällor med sitt informationsbehov. Enligt detta synsätt bör användaren även överväga om denne skall använda sig av flera olika sökmotorer för att lösa sina informationsbehov. Detta är nog dock en ideal bild för hur bibliotekarier och informationsvetare vill att användarna skall lösa sina informationsbehov. Vi tror att de flesta användare vill lösa sitt behov så snabbt som möjligt med så lite tidsåtgång som möjligt. Detta kan vara en anledning till att användare väljer att använda sig av endast en sökmotor för att lösa sitt informationsbehov.

## **1.2 Syfte och frågeställningar**

Huvudsyftet med denna uppsats är att undersöka huruvida den lokala australiensiska sökmotorn Web Wombat presterar bättre, sämre eller lika bra som de två globala sökmotorerna Yahoo och Ask.com. Med presterar bättre, sämre eller lika bra syftar vi på den lokala sökmotorns återvinningseffektivitet för frågor som är kopplade till Australien. Bra återvinningseffektivitet är en viktig anledning till varför användare väljer en sökmotor framför en annan. Med bra återvinningseffektivitet menas att sökmotor återvinner relevanta dokument som svarar mot användarens query. Ett annat syfte vi har med uppsatsen är att se om det kan finnas andra anledningar än återvinningseffektivitet vid val av lokal sökmotor. Detta antagande bygger på vår föreställning att en lokal sökmotor hypotetiskt skulle prestera sämre återvinningseffektivitet, än en global sökmotor. Vi kommer i vår uppsats att jämföra återvinningseffektiviteten mellan de tre valda sökmotorerna, för att se om så är fallet.

- Vilken sökmotor presterar bäst utifrån återvinningseffektivitet?

## **1.3 Avgränsningar**

I vår uppsats funderade vi ett bra tag på att mäta hur sökmotorerna behandlar kontroversiellt material. Dock kom vi fram till att tolkningen av vad som är kontroversiellt material skulle bli en högst subjektiv bedömning. Det kan finnas en anledning till att en åsikt får större utrymme bland de indexerade dokumenten. Vi spekulerar i om en ståndpunkt får mycket utrymme av en specifik anledning. Denna anledning kan vara att personer bakom webbplatserna som får stort utrymme är bättre på att förstå hur sökmotorerna fungerar i form av hur de indexerar och rankar dokument. Vi kommer att gå in på ranking och rankningsfunktioner i kapitel 2.6.2

Vi hade också tänkt se hur täckningen av Australiensiskt material ser ut för respektive sökmotor. Vi kände dock att uppsatsen skulle bli för ofokuserad och spretig, så vi koncentrerade vi oss på återvinningseffektiviteten istället.

När de gäller val av mått som vi har använt i vår undersökning så har vi valt bort recall. Anledningen till detta är att det inte går att mäta korrekt, eftersom vi inte befinner oss i en labbmiljö. För att mäta recall måste den som utför testet känna till storleken på samlingen enligt IR litteraturen MIR (Modern Information Retrieval). I vårt metodkapitel, kapitel 4 förklarar vi vilka mått vi använt och motiverar varför.

## 2. Sökmotorer

För att generalisera kan sökmotorer delas upp i de två huvudkategorierna: sökmotorer och metasökmotorer. Den indelning som främst används är dock den där sökmotorer delas in i efter hur de indexerar. Exempel på olika former av indexering är fulltext och ämnesspecifik indexering.

Danny Sullivan som är skribent, journalist och skapare av webbplatsen Search Engine Watch.com. har delat upp sökmotorerna i åtta olika kategorier. Detta är de åtta olika sökmotorerna enligt Sullivan. Vi har i vår uppsats valt att beskriva tre av dessa som vi fann relevanta för uppsatsen, de var globala sökmotorer, lokala sökmotorer och metasökmotorer.

- **Globala sökmotorer**
- **Lokala sökmotorer**
- **Metasökmotorer**
- *Special sökmotorer*
- *Nyhetssökmotorer*
- *Barnsökmotorer*
- *Multimedia sökmotorer*
- *Search utilities*<sup>3</sup>

### 2.1 Globala sökmotorer

De globala sökmotorerna söker igenom och indexerar dokument ifrån hela världen och ”hela webben”. Hela webben är en sanning med modifikation eftersom den inte finner dokument ifrån den dolda webben. Förutom huvudsökmotorn erbjuds oftast lokala editioner på en mängd olika språk. Exempel på globala sökmotorer är Google, Yahoo och Ask.com.

### 2.2 Lokala sökmotorer

En lokal sökmotor har uppgiften att indexera sidor från ett specifikt land eller region. Detta innebär att om en sökmotor specialiserar sig på Nya Zeeland letar den upp och indexerar webbsidor med domännamnet nz. Istället för att försöka täcka in ”hela Internet” har de fokuserat på en mindre del av den. Exempel på lokala sökmotorer är Web Wombat, SearchNZ och Funnel.

### 2.3 Metasökmotorer

En metasökmotor har inget eget index. Den hämtar istället resultatet för varje sökfråga hos andra sökmotorer. De allra enklaste metasökmotorerna visar bara vilka resultat för varje

---

<sup>3</sup> Chowdhury C G, Chowdhury Sudatta (2001) *Information sources and searching on the World Wide Web* s.19

sökmotor. De mer avancerade placerar in dokumenten i olika kategorier som t. ex. dokument med samma domän.<sup>4</sup>

## 2.4 Hur fungerar en sökmotor?

Innan vi går närmare in på hur en sökmotor fungerar vill vi bidra med en definition av begreppet sökmotor: ”Search engines are databases containing fulltext indexes of web pages.”<sup>5</sup> För att övergå till den tekniska biten, så fungerar sökmotorer på en mängd olika sätt. Trots detta har vi konstaterande att de flesta sökmotorer har tre gemensamma beståndsdelar, dessa beståndsdelar beskriv här i denna punktlista.

Detta är de tre beståndsdelar som är gemensamma för alla sökmotorer:

- Sökmotorer söker av hela Internet eller delar av Internet
- Sökmotorer upprättar index över ord/fraser och har information om dem
- Sökmotorer tillåter användare att söka på ord/fraser i deras index<sup>6</sup>

Sökmotor utför fyra huvuduppgifter:

- Den är en dokumentprocessor
- Den behandlar sökfrågorna
- Den söker och matchar olika funktioner
- Den rankar och rangordnar sökträffarna.<sup>7</sup>

Detta är några exempel på hur dessa uppgifter genomförs:

Dokumentet indexeras och formas till ett format, som passar sökmotor. Sökmotorer har tidigare inte kunnat återvinna alla format, så som pdf och doc. Men utvecklingen av sökmotorerna och konkurrensen mellan dem har lett till att sökmotorer som Google och Yahoo nu återvinner dessa format. Sökmotorer skall även skapa en hanterbar mängd av dokument som kan återvinnas

- Hitta metataggar i dokumenten

Identifiera lämpliga indexeringstermer, vilket innebär termer och ord som beskriver dokumentet och vad det handlar om.

- Ta bort stoppord som och, eller jag m.m.
- Termstämning, ta bort suffix i termer i dokumenten som indexeras
- Vikta olika termer i dokumenten.

Enligt Chowdhury består en sökmotor av tre komponenter. Dessa komponenter är enligt Chowdhury Index, Spindeln samt mjukvaran och gränssnittet. Chowdhury räknar mjukvara och gränssnitt som samma komponent.<sup>8</sup>

---

<sup>4</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. Lund Studentlitteratur. s. 28-29

<sup>5</sup> Sherman Chris och Price Gary (2002) *The Invisible web* uncovering information sources search engines can't see,

<sup>6</sup> Chowdhury C G, Chowdhury Sudatta (2001) *Information sources and searching on the World Wide Web* s.16.

<sup>7</sup> Liddy, Elizabeth (2001) *How a search engine works* Searcher Vol. 9, no 5

<sup>8</sup> Chowdhury C G, Chowdhury Sudatta (2001) *Information sources and searching on the World Wide Web* s.16.

## Spindeln

Spindeln (The spider) är en robot som används av sökmotorer<sup>9</sup>. Namnet spindel hänger ihop med att den kryper över nätet. Automatiskt samlar och hanterar den miljontals webbsidor åt sökmotorn. Spindeln söker upp och identifierar element som titlar och undertitlar som har betydelse för att kunna återvinna information. Eftersom spindlarna väljer olika element att fokusera på leder det till att sökresultaten varierar mellan olika sökmotorer.<sup>10</sup> Genom att skicka sitt domännamn till sökmotorn kan användaren be den att lista sidan.<sup>11</sup>

Roboten har även andra namn som worms, web crawlers och web ants. Worms och web ants är även namn på program som används av hackers. Vi har dock funnit att det är vanligast att roboten kallas för antingen spider eller crawlers, i den litteratur vi funnit om sökmotorer<sup>12</sup>. Vi kommer därför att använda oss av spindel i vår uppsats. Spindeln för de olika sökmotorerna har olika strategier för hur de samlar in och indexerar dokument. Generellt så söker spindeln från en lista med URL:adresser. Det kan t. ex. vara att spindeln söker igenom de nyaste eller populäraste sidorna. Spindeln söker efter vissa nyckelelement när den granskar en webbsida. Dessa nyckelelement bygger på strukturer om hur en webbsida bör vara utformad. En webbsida bör innehålla ett huvud (head) med viktiga taggar eller märken som det heter på svenska. Ett exempel på ett sådant märken är title, som talar om titeln på dokumentet. I huvudet finns viktiga uppgifter som metadata, dessa uppgifter söker spindeln igenom. Följer användaren anvisningar om hur ett webbdokument bör vara uppbyggt och förstår hur spindeln är uppbyggd och vad den letar efter, så är det troligare att ens dokument hamnar högt upp på träfflistan. Skillnaden på de olika sökmotorerna brukar ligga i spindeln, eller rättare sagt hur spindeln samlar in och indexerar dokument efter en given sökfråga.

## Indexet

Det som spindeln hittar kommer att lagras i indexet. När ny information från spindeln återvinns kommer indexet att automatiskt uppdateras<sup>13</sup>. Det som är vanligast för sökmotorer att indexera kallas "high value" fields. Dessa fält innefattar webbadressen (URL) och titeln på dokumentet som indexerats. För att underlätta sökningar är det vanligt att stoppord exkluderas från indexet. På grund av att de förekommer så frekvent och ofta i dokument tillför de inget vid sökningar.<sup>14</sup>

Ett index blir alltså beroende av en bra spindel. Om inte spindeln kan visa på nya förändringar av webbsidor kan sökmotorn inte inkludera informationen och indexet blir inte uppdaterat.

## Mjukvaran och gränssnitt

Mjukvaran (software) har till uppgift stödja återvinningen av relevanta dokument. Detta genom att gränssnittet skall visa de träffar som spindeln sökt igenom i indexet på jakt efter matchande träffar. Matchande träffar innebär hur väl sökmotorn kan besvara användarens query genom att återvinna relevanta dokument för queryn. De relevanta dokumenten eller träffarna rankas sedan i en lista. Listan som visas sorteras efter olika relevans kriterier.

---

<sup>9</sup> Ince ,Darrel (2001) "robot"

<http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t12.e2772> [2006-04-25]

<sup>10</sup> Chowdhury C G, Chowdhury Sudatta (2001)

*Information sources and searching on the World Wide Web*

<sup>11</sup> Thorpe Peter *search engine secrets* <https://www.cia.com.au/thorpe/secrets/howtheywork.html> [2006-05-05]

<sup>12</sup> Chowdhury C G, Chowdhury Sudatta (2001) *Information sources and searching on the World Wide Web*

Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*.

<sup>13</sup> Chowdhury C G, Chowdhury Sudatta (2001)

*Information sources and searching on the World Wide Web*

<sup>14</sup> Sherman Chris och Price Gary (2002) *The Invisible web: uncovering information sources search engines can't see*.

Relevans kriterierna kan variera beroende på respektive sökmotors policy. Olika sökmotorer arbetar på olika sätt för att ange viktning för termer i indexet<sup>15</sup>. Eftersom algoritmerna för olika sökmotorer varierar gör det att resultaten för sökningarna på samma fråga kan skilja sig markant<sup>16</sup>.

Det webbaserade gränssnittet (interface) är det som användaren ser på sin datorskärm. I de flesta fall består gränssnittet av en sökruta där användaren kan söka med hjälp av enkel sökning. Sedan brukar det även ges möjlighet att använda sig av avancerad sökning. Vissa sökmotorer erbjuder även användaren att göra sina personliga inställningar till hur sökmotorn skall fungera och se ut. Det ökade utbudet av alla olika sorters dokument, från film till bilder har gjort att vissa sökmotorer har ändrat sitt utseende. De erbjuder möjlighet att söka under t.ex. sådana kategorier som bilder, film, ljud, olika filformat etc.

Trenden inom informationssökning och sökmotorer är att gränssnittet blir mer användarvänligt. Gränssnitten blir mer och mer interaktiva, vilket nog är kravet för att dagens sökmotorer skall överleva i framtiden. En anledning till att sökmotorerna skall ha ett bra gränssnitt är att det skall ge en bra feedback, visa träfflistan tydligt och lätt överskådligt för användaren. Detta skall leda till en bättre "Information Retrieval (IR) återvinning" för användaren.<sup>17</sup>

Ett gränssnitt i sökmotorerna har två sidor. Ena sidan kallas *query interface*, där användaren ställer sin sökfråga till sökmotorn. Den andra kallas *answer interface*, där användaren får svaret på sin sökfråga. Oftast återvinns svaret som en träfflista där träffarna har rankats<sup>18</sup>.

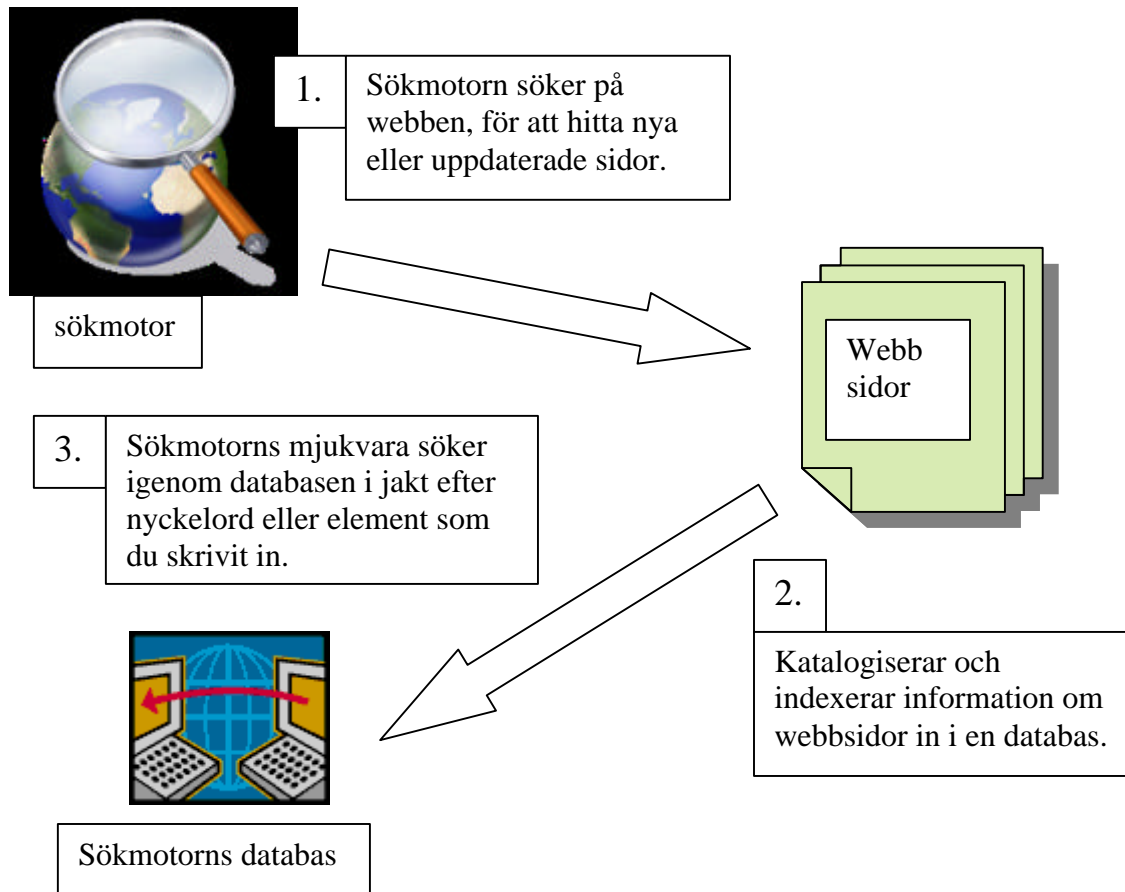
---

<sup>15</sup> Chowdhury C G, Chowdhury Sudatta (2001) *Information sources and searching on the World Wide Web*  
*Information sources and searching on the World Wide Web*

<sup>16</sup>Thorpe Peter *search engine secrets* <https://www.cia.com.au/thorpe/secrets/howtheywork.html> [2006-05-05]

<sup>17</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. s. 96

<sup>18</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*.s. 377



Figur 1: Bilden illustrerar hur en sökmotor söker efter, återvinner och lagrar dokument. Vår egen modell.

### Missuppfattningar runt sökmotorer

Fast vi idag är ett välutvecklat informationssamhälle existerar det myter och felaktiga påståenden runt sökmotorer och hur de fungerar. Detta är några exempel på missuppfattningar som råder gällande sökmotorer.

- Hela webben söks igenom av sökmotorn: Fel, sökmotorn indexerar bara delar av webben och dokumenten som finns där
- Alla sökmotorer söker igenom samma antal webbsidor: Fel, sökmotorerna söker ofta igenom olika samlingar av dokument
- Alla sökmotorer fungerar på samma sätt: Fel, som vi kommer att ta upp senare i uppsatsen kan sökmotorerna skilja sig åt på flera sätt
- Att det alltid är bäst att använda sökmotorns avancerade sökning
- Många användare tror att sökmotorer och webbkataloger ofta är samma sak<sup>19</sup>.

## 2.5 Utvecklingen av sökmotorer

I detta kapitel presenterar vi en kort historisk överblick kring sökmotorer och hur dessa har utvecklats. I kapitlet presenteras tidigare aktörer och sökmotorer som har haft inflyttande på marknaden och som skapat standarder och sökverktyg, som i sin tur drivit branschen framåt. Detta har i slutändan lett till bättre informationsöknings verktyg för användaren. 1993 fanns

<sup>19</sup> Feldman, Susan, Liddy, Elizabeth (2001) *The Searching quagmire*. Searcher no. 5. Vol. 9,

experimentella sökmotorer och dessa indexerade dock inte hela dokument utan oftast bara titeln på dokumentet eller delar av dokumentet. 1994 kom några av de första sökmotorerna. En av dessa var Webcrawler. Webcrawler skapade standarden för hur en sökmotor skulle fungera och så fungerar de flesta sökmotorer även idag. Den standarden innebar att sökmotorn sökte igenom hela dokument, och inte som förut delar av den. År 1995 kom ytterligare en ny aktör, Altavista in på marknaden. Altavista kom att ha stort inflyttande, för vilka sökverktyg en sökmotor bör innehålla. 1998 kom ytterligare en aktör, Google. Google representerade en ny sorts sökmotor med ett unikt rankingsystem. Rankingsystemet används som ett starkt konkurrensmedel mot andra konkurrenter. Detta rankingsystem kallas PageRanking. Det nya med Google var dess enkla och avskalade gränssnitt, med en enkel sökruta. Där kan användaren skriva in sin sökfråga. Detta utseende har flera andra sökmotorer idag<sup>20</sup>.

Googles roll och inflyttande över marknaden har vuxit stadigt. Google är idag den största sökmotorn och verkar också vara den sökmotor som används mest. Konkurrenten på marknaden för sökmotorer är hård och flera andra sökmotorer har tagit upp kampen med Google, om att vara störst och populärast. Några av dessa konkurrenter är Allthewebb, Yahoo, Teoma, Ask jeeves etc.

Sökmotorbranschen är en omvälvande bransch och utvecklingen har lett till att många sökmotorer försvunnit eller inte finns kvar i sitt ursprungliga skick och en del har köpts upp av konkurrenter. Aktörer som t. ex. Northern Light(1997-2003) och Excite(1995-2001) som tidigare var ledande på marknaden och förekommit i tidigare forskningsstudier har idag försvunnit<sup>21</sup>. IT-bubblan under 2000-talet inverkar negativt på utvecklingen av nya sökmotorer men under denna kris i branschen utvecklades ytterligare en aktör som likt Google konkurrerade med en egen unik relevansranking. Denna sökmotor var Teoma som senare blev uppköpta av Ask.com, som tidigare hette Ask jeeves. När Ask.com köpte upp Teoma(köpet skedde i 27 februari 2006) gav de Ask.com en möjlighet att utveckla sin sökmotor och idag använder sig Ask.com bara av Teomas relevansrankingsystem. Vårt test har inte påverkats av uppköpet eftersom vårt test skedde efter köpet. Det kan vara riskabelt att göra ett test av sökmotorer under en längre tid just därför att de kan bli uppköpta och ändra karaktär och utseende. Under historiens gång har flera sökmotorer satsa på tjänster utöver informationsökning. Dessa tjänster har bland annat varit haft chatt, shopping, e-post etc. Sökmotorerna som gått i konkurs har även gjort det misstaget att enbart satsa på annonsintäkter som den enda inkomstkällan<sup>22</sup>. Aktörerna idag verkar dock inte lärt sig så mycket av de misstag som tidigare aktörer gjort, då flera sökmotorer idag just har flera olika tjänster att erbjuda. Det verkar också som att själva söktjänsten inte är sökmotorns högsta prioritet. De vill ofta att det skall verka som om deras främsta uppgift är att hjälpa användaren att söka och hitta relevant information. Efter att vi krisiskt studerat sökmotorernas webbplats och läst deras material, så fann vi att många sökmotorer fortfarande ser annonsintäkter som sin viktigaste inkomstkälla. Detta genom att sökmotorernas högsta prioritet verkar vara att göra annonsörerna glada och lyckliga. Användaren av sökmotorn verkar komma i andra hand.

---

<sup>20</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. Lund Studentlitteratur. s. 16-18

<sup>21</sup> Se kap 3.4 Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathleen (2001) *Measuring search engine quality*

<sup>22</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. Lund Studentlitteratur. s. 16-18

Trenden de senaste åren är att antalet sökmotorer verkar ha minskat<sup>23</sup>. Kanske beror detta på att ett fåtal aktörer har en så stor del av marknaden. Det kan också bero på att de stora aktörerna köper upp sina konkurrenter.

## 2.6 Information Retrieval Teori

### 2.6.1 Olika sätt att mäta återvinningseffektivitet

Information Retrieval (IR) för en sökmotor kan bedömas efter flera olika effektivitetsmått. De vanligaste måtten som används inom IR är recall och precision. Kritik som riktas mot dessa två mått, är att relevansbedömningen som sker av de återvunna relevanta dokumenten många gånger är subjektiv. Vilket kommer att bli fallet i vår undersökning, där bedömningen av precision kommer att bli högst subjektiv från vår sida. Då vi bedömer vilka dokument som verkar vara relevanta för oss, till skillnad mot vad sökmotorn bedömer som relevanta dokument. Ur användbarhetssynpunkt kanske användaren inte heller vill ha tag i alla dokument inom ett ämne, alltså är inte användandet av måttet recall, så relevant ur vår synpunkt. Det är också lämpligare att använda precision än recall, eftersom det är svårt att bedöma recall för en sökmotor. Detta eftersom det är svårt att veta hur stor deras samling av dokument är. Definitionen av precision kommer att tas upp under kap 4.4 återvinningseffektivitet. Definitionen av recall följer här nedanför.

#### Definitionen av recall

##### Antalet av relevanta dokument som återvunnits

Antalet relevanta dokument i kollektionen

Recall är andelen relevanta dokument som effektivt har återvunnits i en specifik samling.<sup>24</sup>

Kritiken mot måtten precision och recall har lett till att forskare inom IR har utvecklat andra mått för att mäta återvinningseffektiviteten hos ett IR-system. Dessa mått tar mer hänsyn till om de återvunna dokumenten är relevanta för användaren. Inom IR-forskningen har forskarna även använt sig av mått som skall utreda om återvinningen av dokument är objektiv eller subjektiv, i vilken grad sökmotorerna återvinner subjektivt material. Detta mått används inte i så stor utsträckning inom IR, eftersom det är svårt att bedöma vad som är objektivt eller subjektivt material. Forskare inom IR har också börjat se annorlunda på användare. Detta genom att de inte ses som en homogen grupp, utan att de tar hänsyn till att olika användare inte bedömer återvinningen på en sökfråga på samma sätt.<sup>25</sup> Forskare inom IR har tagit fram ett mått som kombinerar recall och precision som heter "The Harmonic Mean", måttet lägger lika stor vikt vid att nå högt värde på både recall och precision. Detta mått är annorlunda mot när man annars mäter recall och precision eftersom om du får ett högt värde på recall så brukar du få ett lågt värde på precision och tvärtom, men detta mått eftersträvar höga värden genom att viktningen för båda är lika stor. Ett annat mått som kombinerat och använt både recall och precision är "The E Measure". Meningen med det här måttet är att användaren själv

<sup>23</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. Lund Studentlitteratur. S. 18

<sup>24</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. s.75

<sup>25</sup> ibid s. 82-83

skall kunna vikta om denne är mer intresserad av hög recall och lite mindre intresserad av att få hög precision.

## 2.6.2 Rankning

Hur bra en sökmotor presterar mäts oftast efter hur bra dess ranking fungerar. Rankingen innebär att spindeln identifierar var de relevanta träffarna hamnar i träfflistan. Användare av sökmotorer brukar för det mesta inte orka titta igenom flera träfflistor, utan ser om de första träffarna är relevanta och nöjer sig med dessa som svar. Forskarna i undersökningen iProspect från 2002 har bl. a kommit fram till att de flesta inte tittar på fler än tre träfflistor<sup>26</sup>. För de flesta sökmotorer brukar träfflistan innehålla tio träffar, vilket innebär att användaren i de flesta fall bara studerar de första 30 träffarna. Undersökning kom också fram till att ovana användaren bara tittar på första träfflistan alltså de första tio träffarna. En annan undersökning från Pennsylvania State University kom fram till att 54 % tittar på första sidan i träfflistan, 19 % av användarna såg på andra sidan och endast 10 % tittade på tredje sidan<sup>27</sup>.

Hur sökmotorer rangordnar och rankar dokument i träfflistan styrs av matematiska algoritmer. Dessa matematiska algoritmer och hur sökmotorn rankar dokument är en stor hemlighet för användarna. Detta är några exempel på hur sökmotorn rankar dokument.

- Ordens placering i dokumentet eller i sökfrågan
- Termfrekvens och termviktning
- Närhet mellan ord
- Vektorrymdsmodellen<sup>28</sup>

Sökmotorerna har även andra rankingsfunktioner som just är unika för deras sökmotor och som sagt är välbevarade företagshemligheter. Vilka dokument rankas då högst av en sökmotor? Några generella slutsatser om hur dokument rankas är, att sidan innehåller väsentlig metainformation. T. ex. bör det finnas information i titelmärket, vad det är för typ av dokument, vad det handlar om m.m. Det existerar myter och missuppfattningar om hur användare kan gå tillväga för att kunna få sökmotorerna att ranka ens hemsidor högre. Många användare tror att det är metataggen, metaname som innehåller keywords alltså nyckelord i dokumentet kan påverka hur högt ens dokument rankas. Därför försöker många att manipulera och skriver in populära söktermer i denna metatagg. T. ex. så är det populärt att söka på kändisar och artister, så de skriver in ett namn som är helt irrelevant för dokumentets innehåll. Lyckligtvis för användaren som söker relevanta dokument så ignorerar sökrobotarna ofta att helt enkelt studera denna metatagg. Denna myt eller missuppfattning är så spridd att även självaste New York Times trodde på den<sup>29</sup>.

För att ett dokument skall rankas högt bör viktiga termer förekomma ofta i dokument. Som i vårt fall där vi letar efter relevanta dokument som handlar om Australien. Förekommer termen Australia ofta i dokumenten, så rankar sökmotorerna dessa som relevanta för oss utifrån den query vi skapade. Vanliga förekommande ord brukar ofta ignoreras av sökroboten och har alltså ingen betydelse för rankingen av dokumenten. Dessa ord kallas stoppord. Exempel på stoppord kan vara och, jag, du m.m. Termerna bör

---

<sup>26</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. s.90

<sup>27</sup> ibid. s. 90

<sup>28</sup> ibid s. 91

<sup>29</sup> ibid. s. 92

också förekomma så tidigt som möjligt i dokument. Dokument kan vara väldigt relevanta men har de sina viktiga termer sist i dokumentet, vilket kan leda till att det rankas lägre. Ett annat relevant kriterium är närheten mellan termerna i dokumenten. Därför använder sig en del sökmotorer av närhetsoperatorer som NEAR, NOT, etc. Detta för att användaren skall kunna återvinna relevanta dokument där närheten mellan termerna styr.

De flesta sökmotorer använder sig av någon av följande IR modeller Booleska modellen, Vektormodellen eller Probabilistiska modellen. Den Booleska modellen bygger på mängdlära. Kritiken mot den booleska modellen är att återvinningen sker binärt, alltså relevant eller irrelevant utan att ranka i vilken grad det återvunna dokumentet skulle vara relevant. En annan nackdel är att partiell matchning mellan queries och dokument ej sker. Fördelen med den är att den är enkel och formaliserad. Queries som använder sig av boolsk sökformulering är exakt semantiskt formulerade och använder sig av AND eller OR. Vektormodellen är icke binär och bygger på algebra genom termviktning. Icke binära vikter tilldelas ett värde för indexerade termer i queries och i dokumenten. Viktningen sker genom att mäta likheten mellan dokumenten och användarens queries. Graden av likhet mäts genom användning av ett cosinusmått, som mäter likheten mellan vektorerna. Fördelen gentemot den booleska modellen är att vektormodellen tillåter partiell matchning. Kritiken mot denna modell är att den inte tar hänsyn till om det finns någon relation mellan termerna i dokumentet.

Probabilistiska modellen bygger på sannolikhetslära. Modellen bedömer sannolikheten för att dokumentet är relevant för queryn. Fördelarna med denna metod är att relevansrankingen av dokument sker fallande och att användaren inte behöver skapa boolska sökformuleringar. Nackdelarna är att användaren måste bedöma sannolikheten för att en term skall vara relevant för dokumentet och att modellen heller inte tar hänsyn till termfrekvens<sup>30</sup>.

De flesta av sökmotorerna använder sig av olika variationer av den booleska modellen eller vektormodellen för sin ranking. Andra former av ranking är t. ex. PageRanking som Google använder sig av eller Expertrank, är en annan form av ranking som Ask.com tillämpar efter å ha köpt upp Teoma. Vi kommer att beskriva ExpertRank i Kapitel 4.5.1.

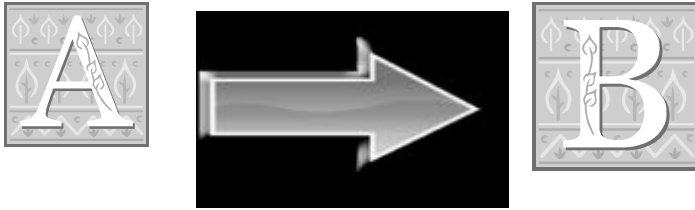
PageRank är en metod där länkpolariteten mäts mellan olika webbsidor för att sedan rankas. Rankingen sker genom att PageRank bedömer hur högt värde sidan skall få beroende på antal relevanta inlänkar och utlänkar som webbsidorna innehåller. Metoden använder sig också av en dämpande faktor för att inte de skall bli en obalans mellan inlänkar och utlänkar. Andra metoder innan PageRank har haft denna obalans då nya sidor med få inlänkar rankas lågt jämfört med en annan webbplats som har många inlänkar men som inte uppdateras så ofta och alltså inte länkas till ifrån andra sidor<sup>31</sup>.

---

<sup>30</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. s. 25-33

<sup>31</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. s. 96-97

Här är vår tolkning av PageRank.



*Figur 2: En länk från sidan A till sidan B betraktas som en röst på sidan B. Beroende på hur viktig sidan A är värdet för dess länk till sidan B Att förändras. Standardvärde för en länk är 0.85.*

### 2.6.3 Olika söktekniker

Sökmotorerna stöder olika söktekniker genom t. ex. sökoperatorer. Dessa operatorer är t. ex. booleska operatorer som AND och OR. De två operatorerna AND eller OR eller båda två kan vara inkluderade i sökfrågan. Detta betyder att användaren inte behöver skriva in dem. Webbens storlek, och i sin tur att antal dokument som indexeras av sökmotorerna har växt har i sin tur avgjort vilken förinställning/defaultinställning de olika sökmotorerna har valt. Ett exempel på detta är att tidigare hade många sökmotorer OR som förinställning mellan termerna.

Nu har de flesta sökmotorer förinställt operatören AND för att begränsa och precisera sökningarna. Som användare gäller det att veta vilken av operatorerna som är förinställda av sökmotorn. Detta eftersom en informationsökning ger helt olika resultat beroende på vilken operator som är förinställd. Så en sökning med termer som *bilar*, *Sverige* kan ge helt olika återvinning. I sökmotorn med OR som defaultinställning skulle den bedöma sökfrågan som att användaren vill ha dokument som handlar om antingen *bilar* eller antingen *Sverige*. I den andra sökmotorn däremot som använder AND, skulle resultatet bli att dokumenten måste innehålla båda termerna för att vara giltigt. En annan sökteknik är frassökning, vilket innebär att användaren söker på en fras t.ex. *de giftigaste ormarna i Australien*

### 3. Tidigare forskning

#### Inledning

Forskning av sökmotorer i ett jämförande syfte har pågått länge. Det som följer är en genomgång av vad några undersökningar som är gjorda under 90-talet.

Chu och Rosenthal (1996) utvärderade sökmotorerna Alta Vista, Lycos och Excites precision med hjälp av den tregradiga skalan "relevant, somewhat relevant and irrelevant". Det som utvärderades efter en tregradiga skalan för respektive sökmotor var topp tio länkar. För varje sökmotor använde de tio sökfrågor.

Ding och Marchionni (1996) studerade tre sökmotorer utifrån fem ämnen. Förutom att ta med relevanta länkar använde de sig av en sexpoängs skala för att bedöma tre sorters relevans. Tomaiuolo och Packer (1996) undersökte Alta vista, Infoseek och Lycos. De tittade på de 200 första träffarna samt de 10 första länkarna. Till sin hjälp hade de evalueringsverktygen Magellan och Point. Gordon och Pathak (1999) evaluerade åtta sökmotorer med en fyra-gradig relevansskala. Relevansgraderna som användes var "highly relevant, somewhat relevant, somewhat irrelevant och highly irrelevant" Till sin hjälp hade de 33 sökfrågor och topp 200 länkar.<sup>32</sup>

När vi har letat efter tidigare forskning i databaser har vi funnit att mycket av forskningen handlade om de största globala sökmotorerna. När vi sökte i INSPEC om tidigare forskning och sökmotorer, fann vi mest information om "Google", dess utveckling, tjänster, problem med sökmotorn etc.

Den forskning vi närmare valt att undersöka, med början på sidan 16 är alla publicerade på 2000-talet. Skälet till att konsekvent välja ny forskning är att utvecklingen går snabbt och undersökningar snabbt blir irrelevanta och inaktuella. Forskningen koncentrerad kring jämförelse mellan globala och lokala sökmotorer upplever vi som ganska begränsad. Konsekvensen av detta blir att vi även tar med forskning mellan olika globala sökmotorer.

---

<sup>32</sup> Molina R Paulo och Jansen J. Bernard (2006) *The effectiveness of web search engines for retrieving relevant ecommerce links*

### 3.1 Comparing rankings of search results on the Web.

#### Syfte

Det är enligt undersökningen ett känt faktum att sökmotorer har låg överlappning vad gäller täckningen. I studien har de som syfte att jämföra och mäta hur lika sökmotorers ranking är gällande överlappande resultat. Även en jämförelse av rankingen av resultatet för identiska sökfrågor återvunna från flertalet sökmotorer gjordes.

#### Metod

För undersökningen valde de noggrant ut 15 sökfrågor inom området Information Retrieval. Sökmotorerna som användes var Google, Alta Vista, Alltheweb och HotBot. Sökfrågorna som ställdes var som tidigare nämnts identiska och alla resultat sparades. Bara sökmotorer som kunde uppvisa en komplett rankinglista togs med i undersökningen.

För undersökningen användes ett mått som kallas Spermans rho. Måttet används för att finna relation mellan två variabler. Måttet tar bara hänsyn till webbadresser (URL) som förekommer på båda träfflistorna. Utifrån Spermans rho jämförs samband mellan ranking för två sökmotorer utefter identiska sökfrågor. Spermans rho har en skala som sträcker sig från -1 till 1. På skalan står 0 för inget samband mellan två rankingar, 1 för total överensstämmelse och -1 för brist på överensstämmelse. Måttet är en variant av Pearson's r.

När det är fler än två listor som jämförs används Kendalls W med 0 för ingen överensstämmelse och 1 för total överensstämmelse. Spermans rho beräknades för varje par av sökmotorer och för varje sökfråga.

#### Resultat

De största sambanden fanns mellan Google och Alta Vista. Det som kan konstateras är att sambanden var högt för de flesta sökfrågorna. Undantaget var att en av frågorna gav ett extremt lågt samband. Google och Alltheweb hade svagast samband emellan sig men båda hade ett starkt samband till Alta Vista. Det fanns vissa oenigheter gällande definitionerna av samband. J. Cohen menade att under 0.3 var lågt, 0,3-0,49 medium och över 0,5 högt medan D. Rowtree menade att den var hög bara över 0,7 och att medium var mellan 0,4 -0,7. Beroende på hur man väljer att definiera det kan man se att sambandet mellan Google och Alta Vista och Google och HotBot är medium enligt Rowtree och högt enligt Cohen. Sambanden mellan Google och AlltheWeb samt AlltheWeb och HotBot är väldigt låga<sup>33</sup>.

Ranking	Samband mellan sökmotorer	Genomsnitt Spermans
1	Google-Alta Vista	0,563
2	Google-HotBot	0,504
5	Google-AlltheWeb	0,138
6	AlltheWeb-HotBot	0,098

*Tabell 1: Det som kan utläsas av tabellen är att sambanden mellan sökmotorerna ligger mittemellan inget samband och total överensstämmelse. Enligt Cohen ligger de på högt och enligt Rowtree medium i samband. De två nedersta är låga oberoende av om man går efter Cohen eller Rowtree*

<sup>33</sup> Bar –Ilan Judith (2005) *Comparing rankings of search results on the web*

## **3.2 Coverage, Relevance, and Ranking:**

### ***The Impact of Operators on Web Search Engine Results.***

#### **Syfte**

Forskning har visat att endast 10 % använder sig av avancerade operatörer, medan 90 % använder mycket enkla söksträngar. Det tas ofta för givet att hjälpmedel som boolesk logik och frassökning ska ge en bättre återvinningseffektivitet. Undersökningen tittar närmare på effekterna av att ha operatörer i sökfrågor i jämförelse med sökfrågor utan operatörer.

#### **Metod**

De tre sökmotorerna som granskades var MSN, AOL (AmericaOnline) och Google. Anledningen till att de valde dessa var att de vid tidpunkten för undersökningen var de mest populära. Den centrala frågan som ställdes var om sökfrågor med operatörer ger ett större antal träffar än utan operatörer. Samtidigt undersökte de hur de olika sökmotorerna hanterar operatörer. De utgicks ifrån områdena täckning, ranking och relevans. Sökfrågorna valdes från Excite den 1 maj 2001.

Utifrån vad statistiska undersökningar visat valde användarna att bara undersöka de första tio resultaten. Studierna visade att ungefär 80 % av användarna aldrig studerade mer än de första tio resultaten. Om dubletter fanns bland de första tio resultaten räknades bara en av dem. I beräkningen av täckningen kunde de inte utesluta att det fanns dubletter. I ett första skede ställde de 100 sökfrågor med operatörer till sökmotorerna. Operatörer som användes var MUST APPEAR, AND, OR och PHRASE och varje operator ingick i 25 frågor var. I nästa steg tog de bort operatörerna. Hela undersökningen tog fem minuter så inga förändringar av sidor var trolig. Tiden för undersökningen var mellan den 12 och 22 mars, 2002. Antalet träffar antecknades och top tio URL: s sparades för evaluering. Fyra utomstående granskare undersökte sedan resultaten individuellt.

#### **Resultat**

Totalt återvanns över 560 miljoner dokument. Räckvidden för täckningen sträckte sig mellan 0 och 61 540 009 dokument. I undersökningen ställde de totalt 600 sökfrågor varav 570 av dessa med mer än 10 träffar. 13 sökfrågor fick inga träffar. Utifrån de 600 sökfrågorna återvann de 5748 dokument. Dessa 5784 dokument fördelade sig sedan som 3328 relevanta och 2420 irrelevanta dokument.

Resultatet visade att en frassökning minskade täckningen hos alla sökmotorerna. Operatör ”MUST APPEAR” gav ökad täckning hos MSN men minskad hos AOL och Google. Operatör ”AND” gav ökad täckning för MSN men oförändrat hos AOL och Google.

Vid mätningen av den relativa precisionen kunde de se att Google utmärkte sig i relation till de andra sökmotorerna. Vid nivån precision DCV 10 (P@10) hade Google ett medel på 73 % utan operatörer och 68 % med operatörer. För MSN var motsvarande värden 49 % respektive 52 %.<sup>34</sup>

---

<sup>34</sup> Eastman M. Caroline och Jansen J. Bernard 2003 *Coverage relevance and ranking: The impact of query operators on web search engine results.*

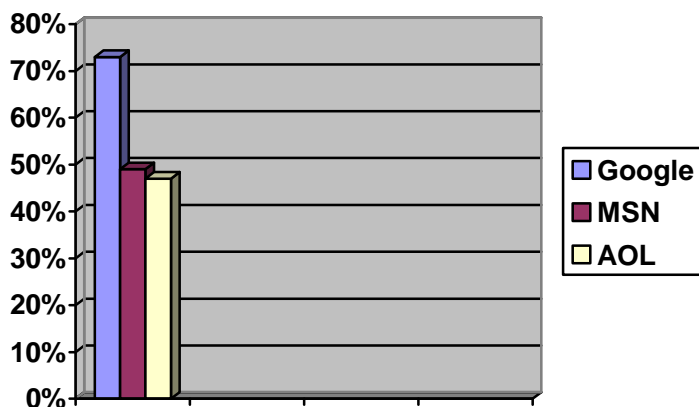


Diagram 1: Genomsnitt för P@10 utan operatorer. P@10= precision för DCV 10

### 3.3 The effectiveness of Web search engines

#### for retrieving relevant ecommerce links.

##### Syfte

Undersökningen fokuserar på att ställa ekonomiska frågor för att kunna återvinna bra ekonomiska länkar. Forskarna vill utvärdera hur bra fem olika typer av sökmotorer presterar i fråga om ekonomiska sökfrågor.

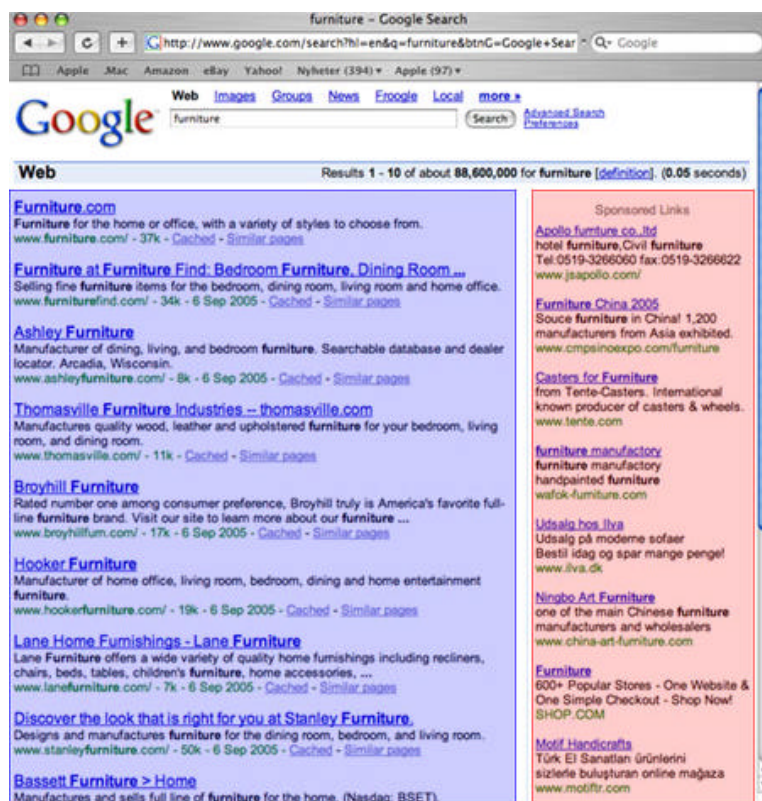
##### Metod

Tre forskningsfrågor ställdes i undersökningen.

1. Vilken sökmotor var bäst utifrån aspekten kommersiell sökning?
2. Är de högst rankade länkarna de mest relevanta?
3. Är sponsrade länkar bättre än organiska?

Samankopplade med dessa frågor fanns tre stycken hypoteser.

- Den första hypotesen utgick från att det kommer att finnas en skillnad i antalet relevanta länkar beroende på om man använder kommersiella eller icke kommersiella sökmotorer
- Den andra hypotesen är att relevansen på länkarna kommer att sjunka när rankingen av länkarna ökar (De mest relevanta länkarna har den lägsta rankingen) En ökad ranking innebär att länkarna enligt sökmotorn är mindre relevanta.
- Den tredje hypotesen var att sponsrade länkar skulle vara mer relevanta än de organiska



Figur 2: Bilden föreställer en vanlig träfflista samt en lista för sponsrade länkar. Ibland definieras posterna i träfflistan som organiska länkar. Till vänster är de organiska länkarna och till höger de sponsrade.<sup>35</sup>

Sökmotorerna som användes var utvalda från fem olika kategorier.

- Metasökmotor: **Excite**
- Kommersiell sökmotor: **Froggle**
- Sökmotor för generella ändamål: **Google**
- Pay per click: **Overture**
- Förtecknings sökmotor: **Yahoo directories**

Efter att vi presenterat de olika kategorierna är det dags att ta en närmare titt på dem.

**Metasökmotorn** söker av flera olika sökmotorer och rankar dess länkar efter en speciell algoritm.

Den **kommersiella sökmotorn** är formad så att den erbjuder produktinformation, kostnader och recensioner till användaren.

Det som i undersökningen benämns som **sökmotor för generella ändamål** vänder sig till den breda massan och har dokument för alla typer av informationsbehov.

**Pay per click** rankar sidor efter den som betalar mest istället för vad sidan har för innehåll. Idén är att man bjuder för att få topp- positionerna på sidan.

<sup>35</sup> Google bildsökning: "Sponsrade länkar" [http://www.jajja.com/images/google\\_search.jpg](http://www.jajja.com/images/google_search.jpg) [2006-05-07]

**Förtecknings sökmotorn** har en förteckning över webbsidor i en mängd olika kategorier. Vanligt för denna typ av sökmotorer är att de bedöms av människor.

Sökmotorerna är valda efter popularitet utifrån en ranking av Alexa research<sup>36</sup>.

I urvalet av kommersiella sökfrågor låg Excites transaction lågt. Sökfrågorna hade ett vitt spektrum med allt från frågor som "houses for sale in Lewis county" till "region free DVD software" Med hjälp av mjukvaran Webposition Gold skickade de ut alla frågorna och återvann länkarna. De fick totalt 3540 träffar men kunde få ner det till 3221 när forskarna tagit bort alla dubletter.

Fyra collage studenter i åldrarna 20-23 fick i uppdrag att utvärdera de mer än 3200 länkarna som återvunnit. De fick ranka länkarna från *1 relevant till 3 inte relevant* och ange en anledning till varför studenterna rankat en länk på ett visst sätt. För att underlätta arbetet sattes ingen deadline för när utvärderingen av länkarna skulle vara klar. Efter att de bedömt dokumenten fick de till uppgift att omevaluera 15 helt slumpvist utvalda frågor som fått samma värde. Av dessa 15 frågor var varje värde från 1-3 representerat med fem frågor. När studenterna utfrågades om varför de rankat en viss länk som de gjort så fick de välja bland följande skäl: Sid innehåll, laddningstid, sidans proffsighet, sammanfattningen, titeln, trovärdighet och URL.

### Resultat

Det visade sig att i 85 % av fallen angavs titeln som skäl för att en länk var relevant. Så många som 90 % av länkarna med ranking 1 hade titel som basis. Detta är att jämföra med 85 % på ranking 2 och 80 % på ranking 3. Utvärderarna märkte att komponenterna sammanfattning och trovärdighet rankades väldigt lågt. Den andra forskningsfrågan angående topp- rankade länkar var mest relevanta fick avfärdas då forskarna inte hittade några belägg. Fråga tre handlade om att sponsrade länkar skulle vara bättre än organiska. Det fanns totalt 3079 organiska länkar med en medelvärdes relevans på 2.17 och 461 sponsrade länkar med en medelvärdesrelevans på 2.07. Hypotesen blev falsifierad då det visade sig att sponsrade länkar inte var bättre utan snarare tvärtom.<sup>37</sup>

## 3.4 Measuring Search Engine Quality

### Syfte

Syftet med undersökningen är att utvärdera sökmotorernas två huvudkomponenter Crawlern<sup>38</sup> och återvinningssystemet. Crawlern har som jobb att hämta, välja och identifiera dokument. Återvinningssystemet har till uppgift att arbeta med den information, som Crawlern tar fram. En sökmotor är beroende av båda komponenterna för att prestera bra. Av just den anledningen så kommer det att ske en evaluering av båda komponenterna i kombination.

### Metod

Tjugo sökmotorer från hela världen valdes ut för undersökningen. Bland sökmotorerna finns bl.a. globala sökmotorer som AltaVista, Google, Infoseek och lokala som Web Wombat och ANZwers. För undersökningen valdes det ut sex stycken australiensiska "domare" som alla hade en universitetsexamen. Deras uppgift blev att avgöra om dokumenten var relevanta för en specifik sökfråga. Varje dokument bedömdes oberoende av varandra och

---

<sup>36</sup> [www.alexa.com](http://www.alexa.com)

<sup>37</sup> Paulo R Molina och Jansen J Bernard (2005) *The effectiveness of web search engines for retrieving ecommerce links*

<sup>38</sup> Kan även benämnas som spindel (spider)

dokument utan direkt koppling till sökfågan bedömdes som relevanta. Dokumenten benämns som "on topic". Ett exempel skulle kunna vara frågan om vem som är kung i Sverige? För användaren ett dokument om Sveriges statsskick eller drottningen är det "on topic". För bedömning av de första 20 resultaten användes olika former av precisionsmått. Mått som användes i undersökningen var precision vid mindre än 20 dokument återvunna (P@n), Inbördes ranking av första återvunna dokumentet (MRR1), TREC genomsnittsprecision (TSAP)<sup>39</sup> och precision vid nivåerna 1 till 5 (P@1-5)

## Resultat

Northern Light var topprankad på vad som angavs som P@1-5. Detta innebär medelvärde för varje nivå från dokument 1 till 5. Trots Northern Lights toposposition var det små statistiska skillnader mellan den och de följande nio sökmotorerna. Även för relevansrankingen av den första sidan (MMR1) var Northern Light högst upp. I undersökningen fick sidor utan relevanta träffar bland de första sju träffarna värdet noll. För ovanstående undersökningar placerar sig Web Wombat trea från slutet. För TREC genomsnittsprecision Var Inquirus i topp och Northern Light denna gång trea. Web Wombat placerade sig än en gång tredje sist. För samtliga tre mått finns ett mönster där de tre sista sökmotorerna var lokala.<sup>40</sup>

## 3.5 Think local, search global? Comparing search engines for searching geographically specific information

### Syfte

Studien syftar till att undersöka hur information om Nya Zeeland återvinns av lokala, globala och metasökmotorer. En underliggande fråga var att undersöka om globala sökmotorer indexerar lika många nya zeeländska domäner som de lokala gör?

### Metod

De fyra globala sökmotorerna som valdes ut var Google, HotBot, Altavista och AlltheWeb. De tre lokala sökmotorerna från Nya Zeeland var SearchNZ, SearchNow och NZExplorer. De tre metasökmotorerna var Excite, Vivi'simo och SurfWax.

De globala och metasökmotorerna är utvalda efter deras utbredning och att de har presterat bra i diverse tester. De lokala var de enda tillgängliga vid undersökningens genomförande.

I undersökningen valde forskarna ut 10 frågor och de 20 första träffarna bedömdes. Sökfrågorna som valdes ut härstammar från sökfrågor ställda till den australiensiska söktjänsten OzSearch.

En annan aspekt var att undersöka om det möjligtvis är så att de globala sökmotorerna har samma antal nz domäner som de lokala. Sökfrågorna anpassades sedan för att passa Nya Zeeland. Det fanns vissa problem med att definiera vilka sidor som var relevanta. De bedömde att bara de indexerade sidorna var relevanta. Detta innebar att sidor som länkade till relevanta sidor inte var relevanta. I relevansbedömningen exkluderade forskarna spegelsidor. Spegelsidor är webbsidor som har samma innehåll men olika webbadresser<sup>41</sup>.

---

<sup>39</sup> TSAP= TREC-style avrage precision som innebär att ett ämne grundar sig bara på precisionspunkter där ett relevant dokument har återvunnits.

<sup>40</sup> Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathleen 2001 *Measuring search engine quality*

<sup>41</sup> Ding och Marchionini (1996) *A Comparative Study of Web Search Service Performance*.

För att undersöka utbredningen av nz-domäner genomförde de 40 sökningar på frågor med bara ett ord. Hälften var generella och den andra hälften nz-specifika. Exempel på ett generellt ord är Hubble medan ett nz-specifikt ord är Waitangi<sup>42</sup>. På grund av att metasökmotorer saknar egna databaser exkluderades dessa från den aktuella övningen.

## Resultat

Hela tre av fyra globala sökmotorer presterade högre övergripande recall än den bästa lokala sökmotorn SearchNZ. Övergripande recall räknas ut genom ta antalet träffar för var och en av de tio sökmotorerna och sökfrågorna delat med antalet totalt relevanta sidor som återvunnits. Eftersom de två andra Nya Zeeländska sökmotorerna presterade relativt dåligt drog de slutsatsen att de lokala sökmotorerna från Nya Zeeland inte har några större fördelar mot de globala sökmotorerna. På det stora hela kan man säga att de globala sökmotorerna har liknande eller bättre täckning av den lokala nz-domänen. Något förvånande var att ingen sökmotor nådde mer än 42 % recall. Så mycket som 36 % var relevant information men utanför den lokala domänen.<sup>43</sup>

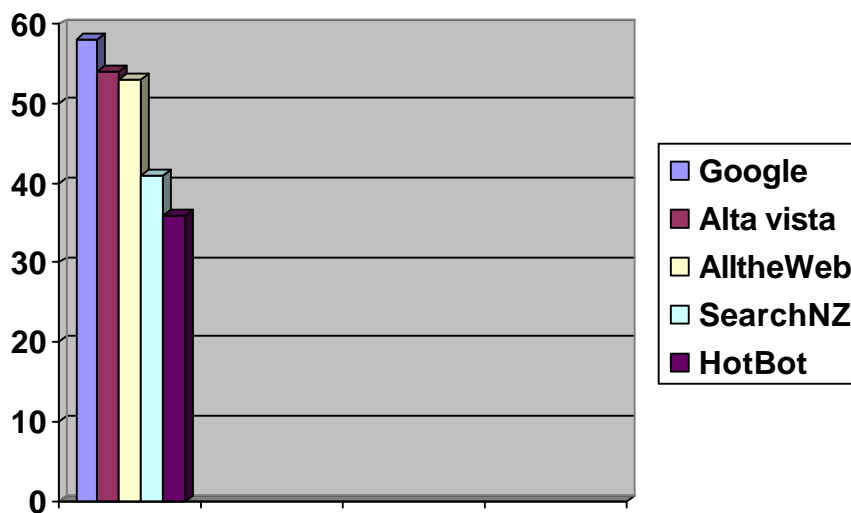


Diagram 2: Över alla frågor återvanns totalt 139 olika relevanta dokument. Diagrammet visar antalet relevanta träffar för varje sökmotor räknat över alla sökfrågor.

<sup>42</sup> Plats för förteckande av ett avtal mellan maorier och briter 1840.

<sup>43</sup> . Smith Alistair G (2002) *Think local, search global? Comparing search engines for searching geographically specific information*

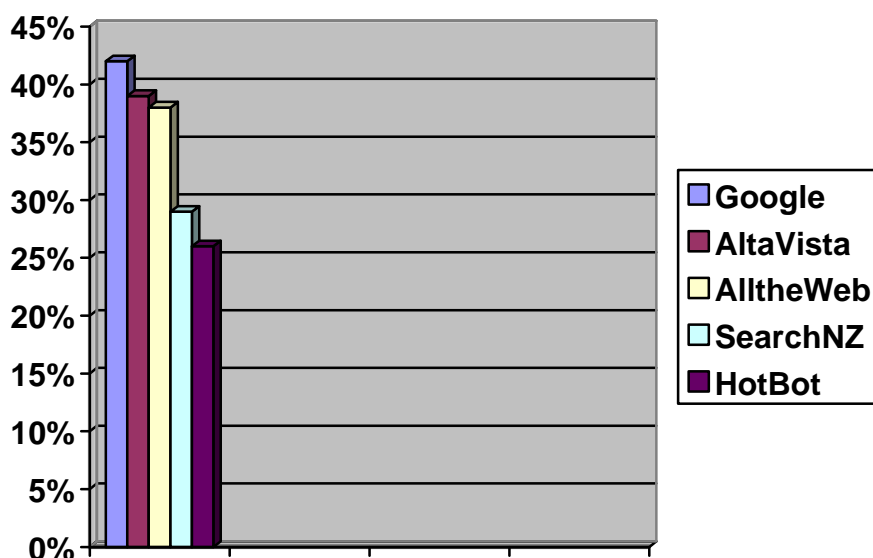


Diagram 3: Relativ recall för den nya zeeländska domänen. Som nämnts tidigare är antalet dokument 139. Recallen blir alltså antalet träffar för en sökmotor delat med det totala antalet träffar. För Google blir det  $58/139=0,42$ .

### 3.6 New measurements for search engine evaluation proposed and tested.

#### Syfte

Att testa nya mått genom att använda dem i en jämförande studie för sökmotorerna Google, Alta Vista och Teoma.

#### Metod

Google valdes för att den är störst, Alta Vista för att den är en av de äldsta och Teoma för att den är relativt ny och har en unik rankingsalgoritm. Fyra stycken sökfrågor konstruerades och de valde att blanda frassökning, boolesk logik och enkelordsökningar. För att se hur sökmotorn presterar över tid använder forskarna tre mått. Dessa är 1. stabiliteten för antalet sidor som återvinns, 2. antalet sidor top 20 som är samma i två på raken följande test sett över en tidsperiod och 3. måttet är antalet sidor top 20 som förblir samma och har samma rangordning i två på raken följande test.<sup>44</sup>

Undersökningen genomfördes i två faser. I första fasen gjorde de en sökmotorranking och i andra fasen en "human ranking". Resultatet av båda faserna jämfördes slutligen. Man får fram kvalitén av resultat rankingen genom jämföra sökmotorsranking med "Human Ranking". Sperman ranking ordnings samband korrelation räknades ut och används här som motvikts mått till precision. Desto högre koefficient desto större samband mellan sökmotor och "Human Ranking". "Human ranking" genomfördes av ett antal studenter från Ontario, Kanada. Alla gjorde först en individuell bedömning för att sedan diskutera och kunna diskutera och omvärdera sina resultat. Undersökningen pågick från den 1 maj till 3 juli, 2002 och en gång i veckan gjorde de nya sökningar i de olika sökmotorerna.

<sup>44</sup> Tidsperioden mellan testen är en vecka för både mått två och tre.

## Resultat

Google var den sökmotor som fick bäst på resultat rankingen. Sökmotorn hade ett genomsnitt på 0,72 över de fyra sökfrågorna. Detta kan betecknas som högt på spermans rho-skala. Alta Vista klarade sig bra och hade högt samband på tre av fyra frågor, vilket gav ett medelvärde på 0,49. Teoma däremot presterade dåligt och utan något tydlig samband med "Human ranking" över de fyra frågorna. Det som var en bidragande orsak var att de på en fråga fick -0,42. Slutvärdet blev låga 0,19 i likhet.

<b>Likhet mellan human och sökmotorsranking</b>	<b>OGS<sup>45</sup></b>	<b>Genomsnitt</b>
Google	0.66	0,72
Alta Vista	0.07	0,49
Teoma	-0,42	0,19

*Tabell 2: Spermans rho*

Vid mätningarna under tioveckorsperioden hade Alta Vista under vecka 7 nästan fördubblat antalet återvunna i relation till hur det var när studien startade. De misstänkte en större ökning av databasens storlek. Google låg på en stabil nivå under hela tioveckorsperioden. Teoma ligger även stabilt men mellan Google och Alta Vista. Om vi ser på top 20 som är samma för en tvåveckors intervall imponerar Google med 98,5 % i genomsnitt, vilket innebär 19.7/20 sidor. Teoma intar andraplatsen och Alta Vista den tredje.

För sidor top 20 som är samma och är i samma rankingordning i två på varandra följande veckor återfinns Google också i topp följt av Teoma och Alta Vista.<sup>46</sup>

<b>Sökmotorer</b>	<b>Medelvärde, antalet sidor med samma rankings ordning två veckor i rad</b>
Google	19.7
Teoma	18.4
Alta Vista	17.9

*Tabell 3: medelvärde över en tioveckorsperiod.*

<sup>45</sup> OGS står för Ontario Graduate School.

<sup>46</sup> Vaughan Liwen (2004) *New measurements for search engine evaluation proposed and tested*

### **3.7 How do search engines respond to some non-English queries?**

#### **Syfte**

Att jämföra hur lokala sökmotorer anpassade för språken ungerska, hebreiska, ryska och franska stod sig i konkurrens med stora globala sökmotorer. Språken som forskarna valde ut har en bredd som innefattar flera språkgrupper. Dessa språk innehåller mängder av möjliga hinder som prefix och apostrofer, vilket gör det till en utmaning för de globala sökmotorerna.

#### **Metod**

Sökfrågor valdes ut för att testas på både globala och lokala sökmotorer för tidigare nämnda språk. De globala sökmotorerna som valdes var AlltheWeb, Google och Altavista. För varje sökmotor valdes lokala domäner för respektive språk. För de lokala sökmotorerna valde de tre morfologiskt anpassade sökmotorer. För hebreiska var det två sökmotorer. Det söktes på både ord och fraser för det lokala språket. I undersökningen försökte de dessutom välja ord som kunde karaktärisera svårigheterna med det specifika språket. För varje fråga granskades de tio första sidorna. Det fanns inget intresse att relevansbedöma resultaten, utan forskarna ville se om de kunde hitta dokument som innehöll den exakta fras man sökte.

#### **Resultat**

För de första tio resultaten för ryska kunde de se att de lokala sökmotorerna genomförde morfologisk analys på sökfrågorna medan de globala sökmotorerna inte tog hänsyn till de egenheter som språket hade. Det som kunde utläsas var endast svaga samband mellan sökfrågor och orden i träfflistan

Franska blev ytterligare en besvikelse för de globala sökmotorerna och gav liknande problem, som för det ryska språket. Både karakteristiska element som apostrofer, accenter och singular/plural ignorerades. De flesta språkspecifika verktygen ignorerades dessutom.

För ungerska hade de globala sökmotorerna inte stöd för olika ordformer. Detta är ett problem när man ska hantera ungerskan. AlltheWeb gjorde inte någon skillnad för vokaler med eller utan apostrof.

För hebreiska sökte de globala sökmotorerna bara på exakt form. För ordet bait som betyder hus eller hem var många av träffarna fristående men trots det innebar sökmotorn morfiks återvinning av olika morfologiska former att träffarna ökade med 60 %.

Den övergripande slutsatsen för undersökningen var att man inte kan sammankoppla www med Web of english. Av det totala antalet Internetanvändare har uppskattningsvis över 50 % av dem inte engelska som sitt modersmål. Ungefär 1/3 av alla webbsidor är på andra språk än engelska.

Google är väldigt populär bland icke engelskspråkiga användare, men tyvärr har de tydliga brister för vissa språk. Användare är omedvetna om dessa brister och förlorar därför information på grund av det.

Andra problem för Google är att de har en gräns på tio ord för en sökfråga och att OR- operatören ibland inte fungerar.

Engelska är ett morfologiskt lätt språk, medan många andra språk har en betydligt svårare morfologisk uppbyggnad. På grund av ekonomiska faktorer är det troligt att sökmotorerna inte producerar tillräckligt bra sökverktyg för de mindre språken. Men om vi fortfarande anser att webben ska vara en plats för alla, måste vi arbeta för att ge alla språk bra sökverktyg anser Bar –Ilan Judith och Gutman Tatyana.<sup>47</sup>

Sökfråga	Sökmotor	Träffar	grammatik
électricité	AOL France	230 000	Alla former, apostrof ignorerad.
électricité	Google.fr	149 000	Exakt form bara
électricité	AlltheWeb.fr	791 590	Alla former

Tabell 4: Illustrerar hur en lokal fransk sökmotor och två internationella sökmotorer behandlar det franska språkets morfologi.

### 3.8 Globala och lokala sökmotorer: en utvärdering av Google, MSN Search och Svesök

#### Syfte

Syftet för denna magisteruppsats är att jämföra återvinningseffektiviteten mellan de tre sökmotorer, Google, MSM Search och Svesök. Petterson ville studera vilken sökmotor som var det lämpligaste valet för en svensk användare. Är det så att Svesök är bäst på den svenska domänen? Om svaret blir nej vilken är då bäst? Han granskade också precisionen i länkbeskrivningarna och sökmotorernas täckning.

#### Metod

För att genomföra undersökningen hämtade han 30 olika sökfrågor från Eniro. se. Frågorna ställdes sedan till de svenska versionerna av de globala sökmotorerna samt Svesök. Sökningarna skedde utan avancerade operatörer och bara sidor på svenska bedömdes. Det som studerades var det första 20 träffarna, alltså DCV20. Vid bedömningen av länkarna användes den binära skalan med bedömningen relevant eller inte relevant. För att avgöra täckningen beräknas det totala antalet sidor som varje sökmotor har fått för samtliga frågor.

#### Resultat

Resultatet blev att Google fick den högsta precisionen av sökmotorerna. Det var också den sökmotor som hade minst dokument med illrelevanta länkar och också den sökmotor som täckte mest av det material som svenska användare kan tänka sig vilja ha enligt Petterson. Petterson konstaterar även att Svesök hade många felmeddelanden, illrelevanta dokument och illrelevanta länkar. Resultatet för precisionen för länkbeskrivningarna blev att Svesök var bäst med sina 91 %. Täckningen visade att Svesök hade 58 000 träffar mot MSNs 602 000 och Googles 1,175 milj träffar. Intressant att notera var också att Svesök inte hade flest antal träffar på någon fråga och att felmeddelanden var så högt som 220. Antalet dubletter var 22., Slutsatsen blev att globala sökmotorer var bättre än lokala på att hitta lokal information.<sup>48</sup>

<sup>47</sup> Bar –Ilan Judith och Gutman Tatyana(2005) *How do search engines respond to some non-English queries?*

<sup>48</sup> Petterson Marcus (2005) *Globala och lokala sökmotorer: En utvärdering av Google MSN search och Svesök*

### **3.9 Sammanfattning**

Av de åtta undersökningarna som presenterats är fyra av dem jämförelser mellan globala och lokala sökmotorer. De övriga fyra behandlar jämförelser mellan globala sökmotorer.

Bland de lokala sökmotorerna återfinns både australiensiska och nya zeeländska sökmotorer. I respektive undersökningar återfinns de i botten, vilket betyder att de inte är bättre än globala sökmotorer.

Två undersökningar använder sig av likhetsmättet spermans rho. Måttet användes för att i den ena undersökningen jämföra likheter mellan sökmotorer och i den andra sökmotorns likhet med olika sökfrågor. Google var den sökmotor som hade bäst medelvärde med sina 0,72 i likhet. Teoma var sämst med medelvärdet 0,19 i likhet.

Fyra av undersökningarna litade till någon form av utomstående experthjälp för att bedöma dokument. I två av fallen använde forskarna sig av studenter som i det ena fallet gick på högskola och i det andra fallet på universitet. För en tredje undersökning var det personer med en universitetsexamen. I en fjärde undersökning nämndes personerna som utomstående och oberoende granskare. I den näst sista undersökningen jämförde lokala och globala sökmotorer men skiljde sig markant från övriga undersökningar. Forskarna brydde sig inte om relevansen utan koncentrerade sig på morfologiska aspekter. Undersökningen kom fram till att morfologiska aspekter ger lokala sökmotorer en fördel gentemot globala sökmotorer. Den sista undersökningen är en magisteruppsats och jämför globala och en lokal sökmotor. Resultatet av undersökningen var att de globala var bättre än den lokala.

Studie	Forskare	antal sökfågor	Mått	DCV	Val av relevansbedömning
3.1	Bar –Ilan Judith	15	Spermans rho, Kendalls W	Ej angivit	Flergradig
3.2	Eastman M. Caroline Och Jansen J. Bernard	100	relativ recall relativ precision	10	Binär
3.3	Jansen J Bernard och Paulo R Molina	100	Modifierad recall,, Spermans rho	10	flerskalig
3.4	Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathlee	54	Genomsnitts Precision 1-5 , precision vid DCV 20 mm.	20	Flerskalig
3.5	Alistair G. Smith	100	relativ recall	20	Binär
3.6	Vaughan Liwen	4	Stabilitetsmått över tid	10	Binär
3.7	Bar –Ilan Judith och Gutman Tatyana	9 ryska 5 franska 8 ungerska 8 hebreiska = 30	inga mått. Morfologisk analys.	ingen	ingen relevansbedömning
3.8	Petterson, Marcus	30	Makro precision, mikro precision m.m	20	Binär

Tabell 5: En sammanfattning av tidigare forskning

## 4 Metod

I vår undersökning jämförde vi tre sökmotorers återvinningseffektivitet. Informationsbehovet och sökfrågorna var fokuserade på Australien och typiska australiensiska företeelser för att se om den lokala australiensiska sökmotorn, Web Wombat var bättre på att återvinna relevanta dokument om Australien än de två globala sökmotorerna. Frågorna som vi ställde till sökmotorerna var övervägande specifika frågor, men de fanns även några generella frågor. Vissa frågor var både generella och specifika. En generell fråga vi ställde var t. ex. fråga 2. Vilka är de populäraste sporterna att utöva i Australien?

En specifik fråga var fråga 4. Vilka är de giftigaste ormarna i Australien?

Exempel för generell och specifik fråga var fråga 13. Hur ser betygssystemet i Australien ut? Med generella frågor menar vi frågor som kan eller har mer än ett rätt svar. Med specifika frågor menar vi frågor som har ett rätt svar, som bekräftas av en faktauppgift t.ex. vem som är arkitekt för Sidneyoperan. Vårt test av återvinningseffektivitet har skett under flera tillfällen under perioden slutet av mars till slutet av april 2006.

### 4.1 Informationsbehovet

Vi konstruerade sökfrågorna utifrån sådant som vi ansåg vara typiskt för Australien. Frågorna delades upp i kategorierna: sport, natur, film, musik, politik, kultur, turism, historia och utbildning. Vi diskuterade fram och tillbaka om hur många sökfrågor som var tillräckligt för att kunna mäta återvinningseffektiviteten. I den tidigare forskningen så använde sig forskarna av 54 st sökfrågor<sup>49</sup>, i ett annat exempel så använde de 15 st<sup>50</sup> och Petterson använde i sin magisteruppsats av 30 st<sup>51</sup>. Vi gjorde en övervägning över vilket antal som verkade rimligt, för att testet skulle vara tillfredsställande. Vi kom fram till att 20 st verkade vara ett rimligt antal frågor.

Frågorna har vi konstruerat själva. Frågorna är alltså inte hämtade ifrån något färdigt frågeformulär som är utformat för IR bedömning av sökmotorer. I de flesta fallen vi studerat har sökfrågorna varit hämtade från ett sådant färdigt frågeformulär. Petterson har hämtat sina frågor från Eniro. Vi ser varken några fördelar eller nackdelar med något av tillvägagångssätten. Det som vi däremot tror spelar roll är valet av söktermer.

Det utformades sedan frågor som har en koppling till Australien. Med koppling menas att frågorna handlar om just företeelser och personer med stark anknytning till Australien. Ett exempel på en sådan fråga är, vilka de giftigaste ormarna i Australien är? Motivet till att vi bara sökte på frågor om Australien var att se om den lokala sökmotorn kunde ge en mer relevant bild på sökfråga, som handlade om en typisk företeelse i landet eller om någon av de globala gav en likvärdig bild eller till och med bättre. Vi har valt att utforma 20 sökfrågor som inkluderar följande ämnesområden:

---

<sup>49</sup> Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathleen (2001) *Measuring search engine quality*

<sup>50</sup> Bar –Ilan Judith (2005) *Comparing rankings of search results on the web*

<sup>51</sup> se tabell 5 s.27

## Ämneskategorier

Sport

Natur

Turism

Politik

Utbildning

Historia

Film

Musik

Kultur

## Sökfrågor

1. Vilka är reglerna för australiensisk fotboll?
2. Vilka är de populäraste sporterna att utöva i Australien?
3. Hur stora fötter har simmaren Ian Thorpe?
4. Vilka är de giftigaste ormarna i Australien?
5. Hur hög är Ayers rock?
6. Vad äter den Tasmanska djävulen?
7. Hur många regioner består Australien av och vad heter de?
8. Vem byggde Sydneyoperan?
9. Vad heter Australiens premiärmister?
10. Vilka partier ingår i regeringen?
11. Hur ser betygssystemet i Australien ut?
12. Vad är medellönen för en lärare i Australien?
13. Var ligger vingården Barossa valley?
14. Vilken religion har aboriginerna ?
15. Hur spelar man på en didgeridoo?
16. Vem skrev australiensiska nationalsången?
17. Vad finns det för fiskar i stora barriärrevet?
18. Vilken är Australiens näst största stad?
19. Hur tillverkar man en bumerang?
20. Hur ser aboriginernas flagga ut?

## 4.2 Sökfrågorna

Att välja rätt termer för de olika frågorna var ett stort problem. Det gällde att välja så betydelsebärande query som möjligt för att få så relevanta dokument som möjligt. Vi försökte även att ta hänsyn till att vissa termer kan betyda flera olika saker. Under kapitlet 6 diskussion kommer vi närmare att gå in på hur valet av termer kan ha spelat in på hur resultat blev för en del frågor.

Vi har valt att konstruera sökfrågor som var anpassade för enkel sökning. Detta eftersom tidigare forskning om användare av sökmotorer visat att de flesta användare söker endast med enkelsökning. Eastman M. Caroline och Jansen J. Bernard kommer i sin studie bl.a. fram till

att 90 % söker med enkelsökning<sup>52</sup>. Vi tror att detta beror på att, många användare inte förstår sökmotorns gränssnitt och kan då inte använda de olika sökverktygen som sökmotorn erbjuder. Vilket i sin tur innebär att de inte förstår hur de skall använda avancerad sökning.

## Querys

1. aussie fotball rules
2. popular sports australia
3. shoe size ian thorpe
4. snakes venom australia
5. Ayers rock high
6. food tasmanian devil
7. States territories australia
8. Sydney opera house
9. Prime minister australia
10. Politics parties australia
11. Grading scale australia
12. average wage teachers australia
13. Wine Barossa valley
14. aboriginal religion
15. Playing didgeridoo
16. national anthem australia
17. Great barrier reef fish
18. Second largest city
19. Making boomerang
20. aboriginal flag

### **4.3 Relevansbedömning av återvunna dokument**

Relevansbedömning av dokumenten som återvanns skedde genom att vi utgick ifrån vad vi ansåg vara relevant för en användare. Detta gjorde att dokumenten som återvanns skulle uppfylla två kriterier. 1. Dokumentet skall svara på frågan. De skall alltså inte länka vidare utan svaret skall synas på sidan. Därför bedöms även sidor som säljer böcker om frågorna som illrelevanta. 2. Dubbletter och sidor som ej fungerade bedöms som illrelevanta av oss.

### **4.4 Återvingseffektiviteten**

Ett mått vi kommer att använda är precision, genom att vi kommer att mäta precision vid DCV 20. De andra måtten är genomsnittlig precision och jaccard's index. Vi kommer även att se vilken sökmotor som ger mest felmeddelanden och flest dubbletter.

---

<sup>52</sup> Eastman M. Caroline och Jansen J. Bernard (2003) *Coverage relevance and ranking: The impact of query operators on web search engine results.*

## Definitionen av precision

$$\frac{\text{Antalet relevanta dokument som återvunnits}}{\text{Totala antalet dokument som återvunnits}}$$

Precision är andelen återvunna dokument som är relevanta<sup>53</sup>  
Exempelvis ger 15/20 en precision på 75 %.

## Genomsnittlig precision

Tanken är att en enskild summa som grundas på placeringarna av de relevanta dokumenten i rankingen. Exempelvis vid DCV 20 så finns de relevanta dokumenten vid plats 5,10 och 18. Uträkningen blir  $0.2+0.2+0.17/3=0.19$ .<sup>54</sup>

## Jaccards index

Detta är ett enkelt sätt att mäta olikheten mellan två olika samlingar. En samling kan t.ex. vara en träfflista för en sökmotor. Det här är formen för hur man räknar ut Jaccards index.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Anta att Yahoo har återvunnit 12 dokument och Ask.com har återvunnit 14 dokument. Om bara 3 av dokumenten är gemensamma innebär det att 9 dokument är unika för Yahoo och 11 dokument är unika för Ask. Snittet för Yahoo och Ask blir då 3 dokument och unionen blir 20 dokument (9+11) vilket ger  $J=3/20$ .<sup>55</sup>

En svaghet hos Jaccards index är att den är känslig för storleksförhållanden mellan olika samlingar. Detta är inte lika påtagligt hos det liknande måttet Mountfords Index av likhet. Måttet använder sig av de tre variablerna a, b och j. Här står a för antalet träffar i samling ett, b för antal träffar för samling två och j för antalet gemensamma träffar för båda samlingarna.<sup>56</sup>

## Felmeddelanden

Felmeddelanden innebär att användaren får meddelanden som t.ex. page not found och error 404. Vi beräknade antal felmeddelanden för respektive sökmotor. Vi tycker att det är relevant att studera hur många felmeddelanden som en sökmotor erhåller totalt, för att se hur användarvänlig en sökmotor är. Som nämnts i tidigare forskning, så studerar de flesta användare just de första träffarna och om dessa träffar är felmeddelanden, så ger de färre relevanta dokument till användaren.

## Dubbletter

Sidor med adresser som förekommer två eller flera gånger är att betrakta som dubletter. Då det handlar om relevanta dokument registreras inte de irrelevanta dubletterna. Liknande eller samma sida men med olika källor är lite problematiskt men bedömningen blir ändå relevant.

<sup>53</sup> Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval* s.75

<sup>54</sup> Ibid s.80

<sup>55</sup> Egghe, L. and Rousseau, R. (1990). *Introduction to informetrics*. s 288

<sup>56</sup> Meyer R John (1998) *So what's the big deal*

<http://www.cals.ncsu.edu/course/ent591k/gcextend.html#diversity> [2005-05-03]

Förekomsten av dubletter är också ett sätt att se hur användarvänlig sökmotorn är. Likt antalet felmeddelande betyder ett stort antal med dubletter färre relevanta dokument och är negativt för användaren. Vi vill vara tydliga med att definitionerna av både dubletter och felmeddelanden helt och hållet vår egen.

## **4.5 Beskrivning av sökmotorerna**

Innan undersökningen tar vid följer en liten presentation av sökmotorerna. Förutom en kort historisk introduktion kommer vi att presentera olika komponenter som t. ex. index, gränssnitt och mjukvara samt spindeln (roboten).

Anledningarna till att vi valde följande sökmotorer grundar sig främst på deras storlek och algoritmer. Ask.com valde vi för att den har sin unika algoritm ExpertRank, Yahoo för att den är välkänd och har ett stort index. Web Wombat är vald för att den är den största och bästa sökmotorn för Australien samt att vi fann en forskningundersökning där den förekom

### **4.5.1 Ask.com**

Teoma som på gaeliska betyder expert har sedan 2001 varit "Ask search technology: s" flaggskepp, men 2006 sammanlänkandes Teoma med Ask.com. Ask.com som tidigare var känt som Ask Jeeves övertog Teoma algoritmen Subject-specific popularity som nu mera är känd som ExpertRank. Denna algoritm gör att Ask.com profilerar sig som en av världens mest kraftfulla och unika sökmotorer, enligt sig själva. ExpertRank rankar en sida efter hur många sidor inom samma ämne som länkar till den. Exempelvis så rankas en sida om fiske högre om andra sidor om fiske länkas till den<sup>57</sup>

Ask.com erbjuder en hel rad sökverktyg, i form av att användaren kan söka i olika kategorier såsom bilder, kartor och väder. I Ask.com kan användaren söka på webben genom enkel sökning och avancerad sökning, samt även en teusarus. Ask vänder sig också till flera olika målgrupper. T. ex. genom "ask for kids", som erbjuder ett förenklat gränssnitt för barn och ungdomar.

Gränssnittet och mjukvara skall underlätta för användaren och bidra till att återvinningen av relevanta dokument ökar. Det tycker vi verkar det som att ask.com, gör. Exempel på detta är att Ask.com stödjer sökningar med naturligt språk t. ex. "vem skrev romanen"? Ask.com ger även användaren möjlighet att begränsa sin sökfråga eller utöka den. Användaren kan även söka på termer relaterade till sökfrågan. Det är ett exempel på hur gränssnittet hjälper användaren med begränsning av sökfrågan, expandera sökfrågan eller hitta relaterade termer till sökfrågan. Söker användaren på t. ex. Australien, kan användaren begränsa sin sökfråga genom att klicka på några av termerna som föreslås under "narrow search" t. ex. "facts about australia". Användaren kan även expandera sin sökning genom "expand" och då t. ex. klicka på termen "kangaroo".

---

<sup>57</sup> [www.ask.com](http://www.ask.com)

## 4.5.2 Web Wombat

Web Wombat är Australiens äldsta sökmotor och har varit aktiv på webben sedan 1995. Sökmotorn med huvudkontor i Melbourne har nischat sig genom att exklusivt vara en lokal sökmotor. Sökmotorn har inga ambitioner på att indexera hela Internet, utan lägger alla resurser på att försöka vara den bästa sökmotorn för Australien och Nya Zeeland. På sin sida omnämner Web Wombat sig som "The original australian search engine".

2005 blev Web Wombat belönade med patent för sin forskning kring sökmotorsspindlar. En av grundarna Phill Bertolus säger:

"By using computers dotted around the world we're able to spider Australian web sites from offshore using cheap remote processing power to compress and send back only the salient data need by our search engines."<sup>58</sup>

Anledningen till att de benämner det som "search engines" grundar sig i att Web Wombat har en mängd sökmotorer kallade WombatView. WombatView är speciella sökmotorer som arbetar med att förmedla hjälp till företag och myndigheter. De arbetar med säkerhet och att söka igenom databaser och intranät.

Enligt Web Wombat själv har de mer än 20 000 000 australiensiska sidor indexerade. Det man inte får reda på är hur många sidor från Nya Zeeland som indexeras.<sup>59</sup>

## 4.5.3 Yahoo

Sökmotorn grundades 1994 av Stanfordstudenterna Jerry Yang and David Filo. Namnet är ett akronym för "Yet Another Hierarchical Officious Oracle", men grundarna hävdar att namnet är valt med bakgrund av att ordet Yahoo betyder osofistikerad och oartig. Som många andra stora sökmotorer har de sitt huvudkontor i Kalifornien.<sup>60</sup>

Sedan starten har sökmotorn växt till att idag vara en av världens största sökmotorer. 2004 introducerade de sin egen sökmotors databas. Innan hade Yahoo använt sig av andras databaser. Runt om i världen finns en mängd lokala editioner spridda över flera världsdelar och i mars 2005 var 36 språk sökbara för den avancerade sökningen. Yahoo rankar sina dokument utifrån relevans baserat på analys av bl. a webbsidornas titel, t. ex., källa och länkar. Sökmotorn har flera tjänster där bl. a väderinformation, mailservice lokal sökning för adresser och kartor samt Yahoo!igans som är deras barnsökmotor ingår. Sökmotorn verkar ha ett fullgott stöd för booleska operatorerna AND, OR och NOT.<sup>61</sup>

Det användaren dock bör tänka på som användare är att en hel del av kategorierna och sökverktygen kan innefatta endast information från och om USA och Kanada. T.ex. sökning på "local" i Yahoo, så har de endast kartor över USA och Kanada.<sup>62</sup>

---

<sup>58</sup> [www.WebWombat.com](http://www.WebWombat.com) [2006-10-18]

<sup>59</sup> *Ibid* [2006-10-18]

<sup>60</sup> <http://yhoo.client.shareholder.com/press/overview.cfm> (2006)

<sup>61</sup> Notess R. Greg *Review of Yahoo! Search*

<http://www.searchengineshowdown.com/features/yahoo/review.html> senast uppdaterad [2005]

<sup>62</sup> [www.yahoo.com](http://www.yahoo.com)

## 5. Resultat

### 5.1 Ask.com

#### 5.1.1 Ask.com precision vid DCV 20

Sökfrågor	Antal relevanta träffar	Precision
1	1	5%
2	2	10%
3	15	75%
4	2	10%
5	8	40%
6	14	70%
7	10	50%
8	8	40%
9	8	40%
10	7	35%
11	2	10%
12	4	20%
13	6	30%
14	7	35%
15	6	30%
16	5	25%
17	6	30%
18	9	45%
19	13	65%
20	8	40%

Tabell 6: Tabellen visar precisionen vid DCV 20 för Ask.com.

Ask.com Fick högst precision på fråga 3 (75%) och lägst på fråga 1 (5%). Precisionen ligger På 14 av frågorna mellan 20-50 %.

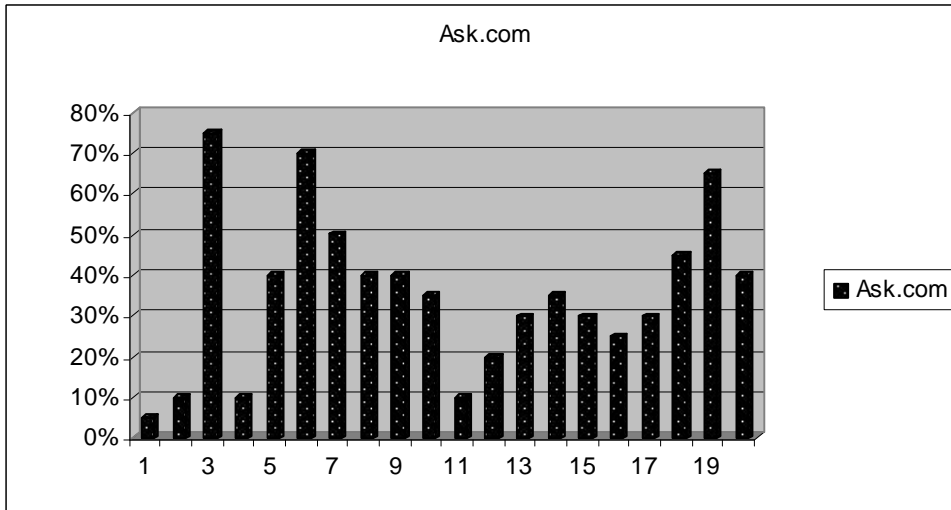


Diagram 4: Precision vid DCV 20 för Ask.com.

### 5.1.2 Ask.com genomsnittsprecision

Fråga	# relevant	Pos	Calc.	Avg.Prec
1.	1	3	0,33	<b>0,33</b>
	<b>Summa:</b>		<b>0,33</b>	
2.	1	4	0,25	<b>0,29</b>
	2	6	0,33	
	<b>Summa:</b>		<b>0,58</b>	
3.	1	1	1	<b>0,76</b>
	2	2	1	
	3	4	0,75	
	4	6	0,66	
	5	8	0,62	
	6	9	0,66	
	7	10	0,7	
	8	11	0,72	
	9	13	0,69	
	10	14	0,71	
	11	15	0,73	
	12	16	0,75	
	13	17	0,76	
	14	18	0,78	
	15	19	0,83	
	<b>Summa:</b>		<b>11,36</b>	
4.	1	10	0,1	<b>0,1</b>
	2	19	0,1	
	<b>Summa:</b>		<b>0,2</b>	
5.	1	1	1	<b>0,54</b>
	2	4	0,5	
	3	5	0,6	
	4	7	0,57	
	5	11	0,45	
	6	17	0,35	
	7	18	0,39	

	8	19	0,42	
	<b>Summa:</b>		<b>4,28</b>	
6.	1	1	1	<b>0,93</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	6	1	
	7	7	1	
	8	8	1	
	9	9	1	
	10	11	0,9	
	11	14	0,78	
	12	15	0,8	
	13	17	0,76	
	14	19	0,74	
	<b>Summa:</b>		<b>12,98</b>	
7.	1	1	1	<b>0,76</b>
	2	2	2	
	3	4	0,75	
	4	9	0,44	
	5	10	0,5	
	6	12	0,5	
	7	13	0,54	
	8	14	0,57	
	9	15	0,6	
	10	16	0,62	
	<b>Summa:</b>		<b>7,52</b>	
8.	1	2	0,5	<b>0,37</b>
	2	7	0,28	
	3	8	0,37	
	4	12	0,33	
	5	16	0,31	
	6	17	0,35	
	7	18	0,39	
	8	20	0,4	
	<b>Summa:</b>		<b>2,93</b>	
9.	1	1	1	<b>0,75</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	9	0,55	
	6	10	0,6	
	7	14	0,5	
	8	15	0,53	
	9	17	0,53	
	<b>Summa:</b>		<b>6,71</b>	
10.	1	1	1	<b>0,68</b>
	2	2	1	
	3	3	1	
	4	7	0,57	
	5	11	0,45	
	6	16	0,37	

		7	18	0,39	
		<b>Summa:</b>		<b>4,78</b>	
11.		1	7	0,14	<b>0,17</b>
		2	10	0,2	
		<b>Summa:</b>		<b>0,34</b>	
12.		1	1	1	<b>0,51</b>
		2	6	0,33	
		3	7	0,43	
		4	14	0,28	
		<b>Summa:</b>		<b>2,04</b>	
13.		1	1	1	<b>0,58</b>
		2	3	0,66	
		3	4	0,75	
		4	7	0,57	
		5	11	0,45	
		6	19	0,31	
		7	20	<b>0,35</b>	
		<b>Summa:</b>		<b>4,09</b>	
14.		1	1	1	<b>0,82</b>
		2	2	1	
		3	3	1	
		4	4	1	
		5	7	0,71	
		6	11	0,54	
		7	14	0,5	
		<b>Summa:</b>		<b>5,75</b>	
15.		1	1	1	<b>0,73</b>
		2	2	2	
		3	7	0,43	
		4	12	0,33	
		5	18	0,28	
		6	19	0,31	
		<b>Summa:</b>		<b>4,35</b>	
16.		1	2	0,5	<b>0,45</b>
		2	4	0,5	
		3	6	0,5	
		4	9	0,44	
		5	16	0,31	
		<b>Summa:</b>		<b>2,25</b>	
17.		1	1	1	<b>0,74</b>
		2	2	1	
		3	3	1	
		4	6	0,66	
		5	11	0,45	
		6	19	0,31	
		<b>Summa:</b>		<b>4,42</b>	
18.		1	2	0,5	<b>0,51</b>
		2	5	0,4	
		3	6	0,5	
		4	8	0,5	
		5	10	0,5	
		6	12	0,5	
		7	13	0,54	

	8	14	0,57	
	9	16	0,56	
	<b>Summa:</b>		<b>4,57</b>	
19.	1	1	1	<b>0,87</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	7	0,86	
	7	9	0,78	
	8	10	0,8	
	9	11	0,81	
	10	13	0,77	
	11	14	0,78	
	12	15	0,8	
	13	20	0,65	
	<b>Summa:</b>		<b>11,25</b>	
20.	1	1	1	<b>0,55</b>
	2	4	0,5	
	3	6	0,5	
	4	7	0,57	
	5	11	0,45	
	6	12	0,5	
	7	16	0,44	
	8	18	0,44	
	<b>Summa:</b>		<b>4,4</b>	

Tabell 7: Genomsnittsprecisionen för Ask.com.

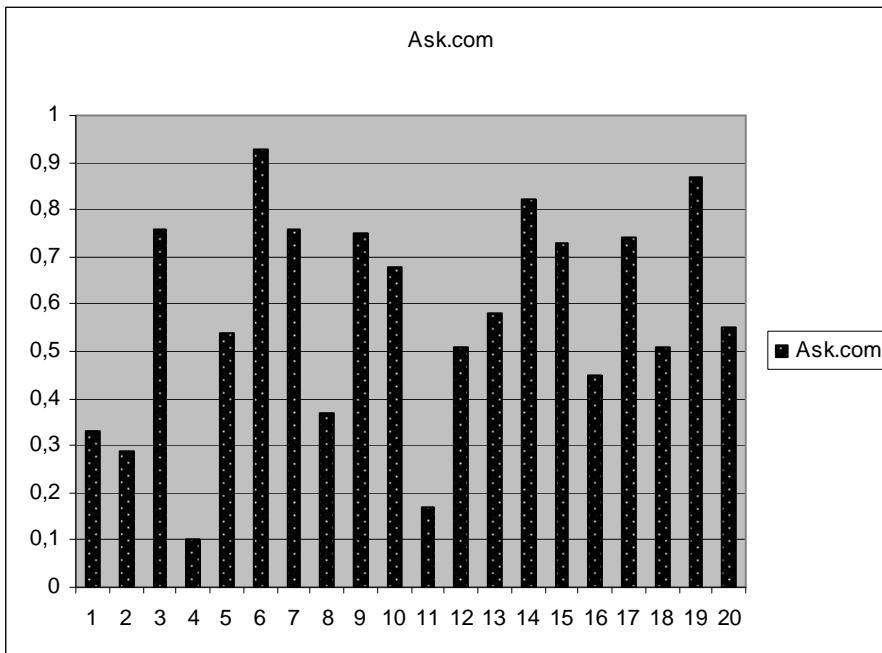


Diagram 5: Genomsnittsprecision för Ask.com fråga för fråga.

Den Högsta genomsnittsprecisionen var 0,93 för fråga sex och lägst var 0,1 för fråga fyra. Sett över alla 20 frågor blev snittprecisionen 0,57.

## 5.2 Web Wombat

### 5.2.1 Web Wombat precision vid DCV 20

Sökfrågor	Antal relevanta träffar	Precision
1	2	10%
2	0	0%
3	2	10%
4	5	25%
5	4	20%
6	4	20%
7	9	45%
8	8	40%
9	5	25%
10	8	40%
11	2	10%
12	7	35%
13	9	45%
14	3	15%
15	5	25%
16	5	25%
17	5	33%
18	8	40%
19	3	15%
20	10	50%

Tabell 8: Tabellen visar precisionen vid DCV 20 för Web Wombat.

Web Wombat har högst precision för fråga 20 (50%) och lägst för fråga 2 (0%).  
15 av frågorna låg mellan 15-45 %.

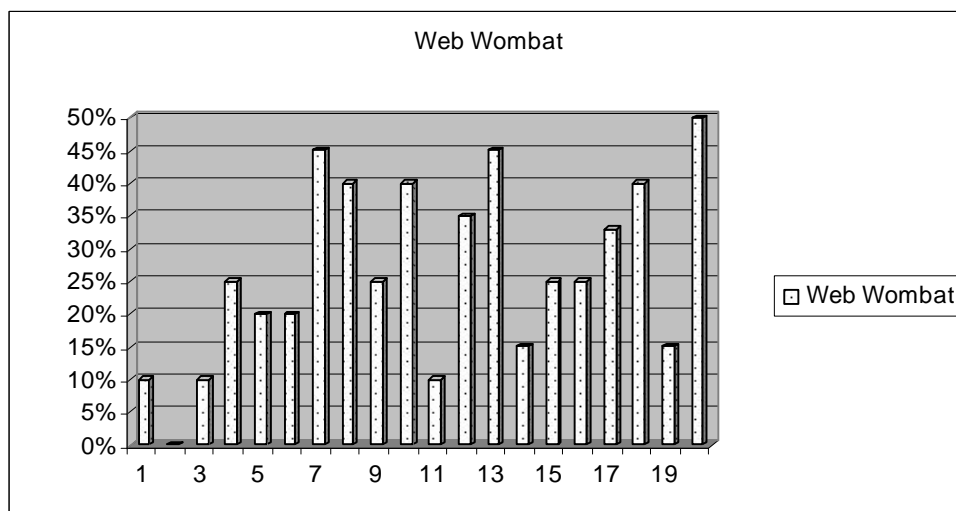


Diagram 6: Precision vid DCV 20 för Web Wombat

## 5.2.2 Web Wombat genomsnittsprecision

Fråga	relevant	Position	Uträkning	Avg.Pres.
1.	1	8	0,12	<b>0,07</b>
	2	19	0,015	
	<b>Summa:</b>		<b>0,135</b>	
2.	0	0	0	<b>0</b>
3.	1	10	0,1	<b>0,14</b>
	2	11	0,18	
	<b>Summa:</b>		<b>0,28</b>	
4.	1	3	0,33	<b>0,27</b>
	2	7	0,28	
	3	15	0,2	
	4	16	0,25	
	5	17	0,29	
	<b>Summa:</b>		<b>1,35</b>	
5.	1	3	0,33	
	2	4	0,5	
	3	7	0,43	
	4	8	0,5	
<b>Summa:</b>		<b>1,76</b>		
6.	1	1	1	<b>0,70</b>
	2	2	1	
	3	8	0,37	
	4	9	0,44	
<b>Summa:</b>		<b>2,81</b>		
7.	1	1	1	<b>0,64</b>
	2	3	0,66	
	3	4	0,75	
	4	6	0,66	
	5	7	0,71	
	6	13	0,46	
	7	14	0,5	
	8	15	0,53	
	9	18	0,5	
<b>Summa:</b>		<b>5,77</b>		
8.	1	2	0,5	<b>0,43</b>
	2	6	0,33	
	3	8	0,37	
	4	9	0,44	
	5	12	0,42	
	6	14	0,43	
	7	16	0,44	
	8	17	0,47	
<b>Summa:</b>		<b>3,4</b>		
9.	1	3	0,33	<b>0,37</b>
	2	5	0,4	
	3	7	0,43	
	4	9	0,44	
	5	19	0,26	
<b>Summa:</b>		<b>1,86</b>		
10.	1	1	1	<b>0,81</b>

	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	12	0,5	
	7	13	0,54	
	8	20	0,4	
	<b>Summa:</b>		<b>6,44</b>	
11.	1	2	0,5	<b>0,31</b>
	2	17	0,12	
	<b>Summa:</b>		<b>0,62</b>	
12.	1	1	1	<b>0,48</b>
	2	5	0,4	
	3	9	0,33	
	4	10	0,4	
	5	12	0,42	
	6	13	0,46	
	7	20	0,35	
	<b>Summa:</b>		<b>3,36</b>	
13.	1	1	1	<b>0,54</b>
	2	2	1	
	3	9	0,33	
	4	12	0,33	
	5	13	0,38	
	6	15	0,4	
	7	16	0,44	
	8	17	0,47	
	9	18	0,5	
	<b>Summa:</b>		<b>4,85</b>	
14.	1	3	0,33	<b>0,2</b>
	2	18	0,11	
	3	19	0,16	
	<b>Summa:</b>		<b>0,6</b>	
15.	1	1	1	<b>0,41</b>
	2	8	0,25	
	3	11	0,27	
	4	14	0,29	
	5	19	0,26	
	<b>Summa:</b>		<b>2,07</b>	
16.	1	1	1	<b>0,55</b>
	2	3	0,66	
	3	5	0,6	
	4	18	0,22	
	5	19	0,26	
	<b>Summa:</b>		<b>2,74</b>	
17.	1	3	0,33	<b>0,33</b>
	2	4	0,5	
	3	12	0,25	
	4	15	0,27	
	5	16	0,31	
	<b>Summa:</b>		<b>1,66</b>	
18.	1	2	0,5	<b>0,54</b>
	2	3	0,66	
	3	4	0,75	

	4	5	0,8	
	5	13	0,38	
	6	14	0,43	
	7	17	0,41	
	8	20	0,4	
	<b>Summa:</b>		<b>4,33</b>	
19.	1	1	1	<b>1</b>
	2	2	1	
	3	3	1	
	<b>Summa:</b>		<b>3</b>	
20.	1	1	1	<b>0,58</b>
	2	3	0,66	
	3	5	0,6	
	4	6	0,66	
	5	9	0,55	
	6	13	0,46	
	7	16	0,44	
	8	17	0,47	
	9	19	0,47	
	10	20	0,5	
	<b>Summa:</b>		<b>5,81</b>	

Tabell 9: Genomsnittsprecisionen för Web Wombat.

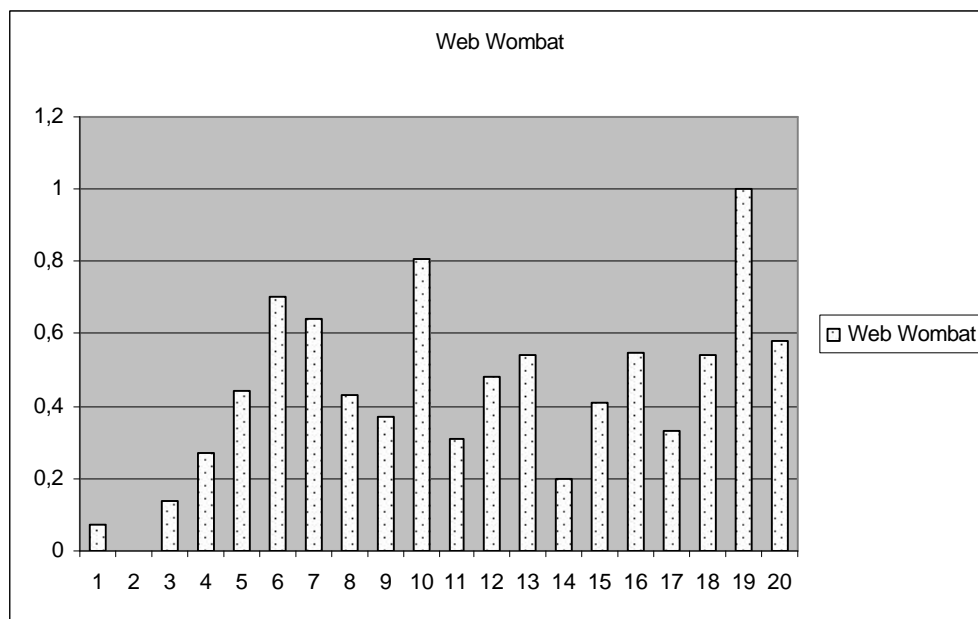


Diagram 7: Genomsnittsprecision för Web Wombat fråga för fråga.

Den högsta genomsnittsprecisionen var 1 på fråga 19 och lägst var 0 för fråga två. Web Wombat var ensam om att ha en genomsnittsprecision på 1 för en fråga. Sett över alla 20 frågorna blev snittprecisionen 0,44.

## 5.3 Yahoo

### 5.3.1 Yahoo precision vid DCV 20

Sökfrågor	Antal relevanta träffar	Precision
1	4	20%
2	5	25%
3	11	55%
4	9	45%
5	10	50%
6	14	70%
7	14	70%
8	10	50%
9	8	40%
10	14	70%
11	3	15%
12	8	40%
13	9	45%
14	8	40%
15	7	35%
16	7	35%
17	9	45%
18	8	40%
19	9	45%
20	11	55%

Tabell 10: Denna tabell visar precision för Yahoo vid DCV 20.

Yahoo hade högst precision för fråga 7 (70%) och lägst för fråga 11 (15%).  
15 av frågorna låg mellan 20-55 %.

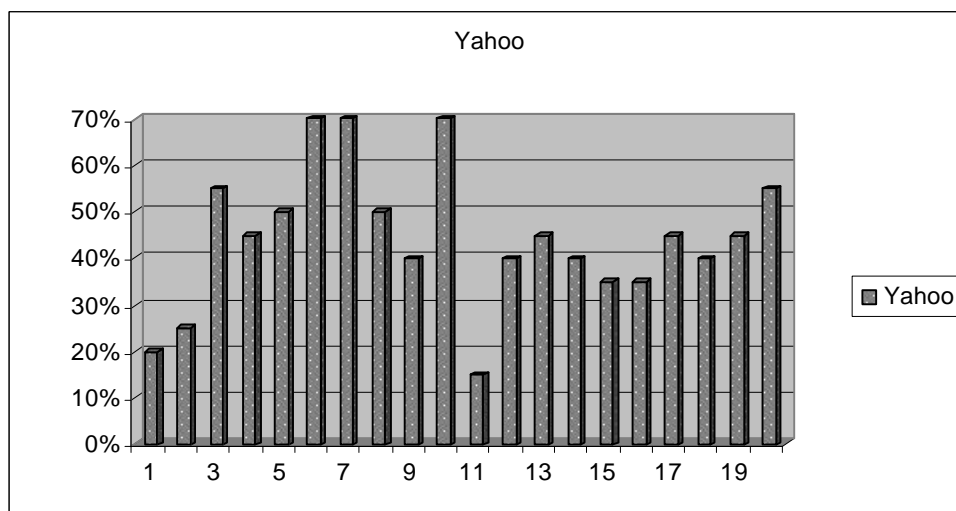


Diagram 8: Precision vid DCV 20 för Yahoo.

### 5.3.2 Yahoo genomsnittsprecision

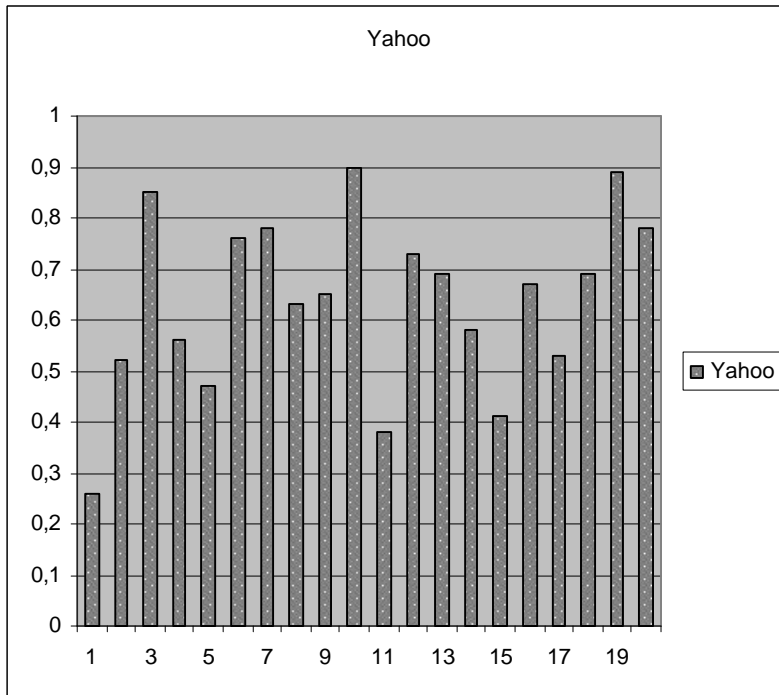
Fråga	relevant	Position	Uträkning	Avg.Pres.
1.	1	5	0,2	<b>0,26</b>
	2	7	0,29	
	3	12	0,25	
	4	13	0,31	
	<b>Summa:</b>		<b>1,05</b>	
2.	1	1	1	<b>0,52</b>
	2	4	0,5	
	3	7	0,43	
	4	10	0,4	
	5	19	0,26	
	<b>Summa:</b>		<b>2,59</b>	
3.	1	1	1	<b>0,85</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	6	0,83	
	6	7	0,86	
	7	8	0,88	
	8	11	0,72	
	9	13	0,69	
	10	14	0,71	
	11	18	0,61	
	<b>Summa:</b>		<b>9,3</b>	
4.	1	2	0,5	<b>0,56</b>
	2	3	0,66	
	3	4	0,75	
	4	8	0,5	
	5	11	0,45	
	6	12	0,5	
	7	13	0,54	
	8	14	0,57	
	9	16	0,56	
	<b>Summa:</b>		<b>5,03</b>	
5.	1	3	0,33	<b>0,47</b>
	2	6	0,33	
	3	7	0,43	
	4	9	0,44	
	5	10	0,5	
	6	12	0,5	
	7	13	0,54	
	8	17	0,47	
	9	18	0,5	
	10	19	0,53	
	11	20	0,55	
	<b>Summa:</b>		<b>5,12</b>	
6.	1	1	1	<b>0,76</b>
	2	3	0,66	
	3	4	0,75	
	4	5	0,8	
	5	6	0,83	
	6	8	0,75	

	7	9	0,78	
	8	10	0,8	
	9	13	0,69	
	10	14	0,71	
	11	15	0,73	
	12	16	0,75	
	13	18	0,72	
	14	20	0,7	
	<b>Summa:</b>		<b>10,67</b>	
7.	1	1	1	<b>0,78</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	8	0,62	
	6	9	0,66	
	7	10	0,7	
	8	12	0,66	
	9	13	0,69	
	10	14	0,71	
	11	16	0,69	
	12	17	0,7	
	13	18	0,72	
	14	19	0,74	
	<b>Summa:</b>		<b>10,89</b>	
8.	1	2	0,5	<b>0,63</b>
	2	3	0,66	
	3	4	0,75	
	4	5	0,8	
	5	7	0,71	
	6	9	0,66	
	7	10	0,7	
	8	14	0,57	
	9	19	0,47	
	10	20	0,5	
	<b>Summa:</b>		<b>6,32</b>	
9.	1	1	1	<b>0,65</b>
	2	2	1	
	3	5	0,6	
	4	7	0,57	
	5	8	0,62	
	6	12	0,5	
	7	14	0,5	
	8	19	0,42	
	<b>Summa:</b>		<b>5,21</b>	
10.	1	1	1	<b>0,90</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	6	1	
	7	7	1	
	8	8	1	
	9	11	0,81	
	10	12	0,83	

	11	14	0,79	
	12	15	0,8	
	13	18	0,72	
	14	20	0,7	
	<b>Summa:</b>		<b>12,65</b>	
11.	1	2	0,5	<b>0,38</b>
	2	5	0,4	
	3	12	0,25	
	<b>Summa:</b>		<b>1,15</b>	
12.	1	1	1	<b>0,73</b>
	2	2	1	
	3	3	1	
	4	5	0,8	
	5	7	0,71	
	6	12	0,5	
	7	18	0,39	
	8	20	0,4	
	<b>Summa:</b>		<b>5,8</b>	
13.	1	1	1	<b>0,69</b>
	2	2	1	
	3	5	0,6	
	4	6	0,66	
	5	8	0,62	
	6	10	0,6	
	7	11	0,63	
	8	13	0,62	
	9	18	0,5	
	<b>Summa:</b>		<b>6,23</b>	
14.	1	1	1	<b>0,58</b>
	2	3	0,66	
	3	6	0,5	
	4	8	0,5	
	5	9	0,55	
	6	12	0,5	
	7	14	0,5	
	8	19	0,42	
	<b>Summa:</b>		<b>4,63</b>	
15.	1	1	1	<b>0,41</b>
	2	10	0,2	
	3	11	0,27	
	4	12	0,33	
	5	15	0,33	
	6	16	0,38	
	7	20	0,35	
	<b>Summa:</b>		<b>2,86</b>	
16.	1	1	1	<b>0,67</b>
	2	2	1	
	3	5	0,6	
	4	7	0,57	
	5	8	0,63	
	6	14	0,43	
	7	15	0,47	
	<b>Summa:</b>		<b>4,7</b>	
17.	1	1	1	<b>0,53</b>

	2	2	1	
	3	8	0,38	
	4	12	0,33	
	5	13	0,38	
	6	16	0,39	
	7	18	0,39	
	8	19	0,42	
	9	20	0,45	
	<b>Summa:</b>		<b>4,74</b>	
18.	1	1	1	<b>0,69</b>
	2	3	0,66	
	3	5	0,6	
	4	6	0,66	
	5	7	0,71	
	6	8	0,75	
	7	10	0,7	
	8	11	0,72	
	9	13	0,69	
	10	17	0,59	
	11	19	0,58	
	12	20	0,6	
	<b>Summa:</b>		<b>8,26</b>	
19.	1	1	1	<b>0,89</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	6	1	
	7	7	1	
	8	15	0,53	
	9	20	0,45	
	<b>Summa:</b>		<b>7,98</b>	
20.	1	1	1	<b>0,78</b>
	2	2	1	
	3	3	1	
	4	4	1	
	5	5	1	
	6	9	0,66	
	7	11	0,63	
	8	12	0,66	
	9	16	0,56	
	10	18	0,55	
	11	20	0,55	
			<b>8,61</b>	

Tabell 11: Genomsnittsprecisionen för Yahoo.



Tabell 9: Genomsnitt precisionen för Yahoo fråga för fråga.

Den högsta genomsnittsprecisionen var 0,90 för fråga 10 och lägst var 0,26 för fråga 1. Sett över alla 20 frågor blev snittprecisionen 0,64.

## 5.4 Alla tre sökmotorerna

### 5.4.1 Jaccards index

Jaccard's index: Yahoo&Ask.com	Gemensamma	Yahoo	Ask.com	överlappning
1.	2	3	1	100%
2.	1	3	1	25%
3.	4	5	8	29%
4.	1	7	2	11%
5.	1	8	5	8%
6.	10	5	5	100%
7.	2	9	6	13%
8.	2	9	3	17%
9.	2	6	3	22%
10.	4	8	2	40%
11.	0	3	1	0%
12.	0	5	5	0%
13.	1	8	8	6%
14.	2	5	2	29%
15.	3	2	3	60%
16.	2	4	5	22%
17.	4	5	2	57%
18.	1	6	0	17%
19.	5	3	7	50%
20.	3	7	5	25%

<b>Samanlagt</b>		<b>46</b>	<b>106</b>	<b>78</b>	<b>25%</b>
<b>Jaccard's index:</b>					
<b>Ask.com-Web Wombat</b>	<b>Gemensamma</b>	<b>Ask.com</b>	<b>Web Wombat</b>	<b>överlappning</b>	
1.	0	3	2	0 %	
2.	0	1	0	0 %	
3.	0	8	3	0 %	
4.	0	2	6	0 %	
5.	0	2	4	0 %	
6.	0	5	3	0 %	
7.	2	6	5	18%	
8.	1	3	7	10%	
9.	1	3	3	17%	
10.	0	2	4	0 %	
11.	0	1	1	0 %	
12.	1	5	4	11%	
13.	2	7	5	17%	
14.	1	2	3	20%	
15.	1	3	1	25%	
16.	3	5	2	43%	
17.	2	4	3	29%	
18.	0	9	8	0%	
19.	1	7	2	11%	
20.	2	5	5	20%	
<b>Samanlagt</b>	<b>15</b>	<b>86</b>	<b>67</b>	<b>9%</b>	
<b>Jaccard's index</b>					
<b>Yahoo-Web Wombat</b>	<b>Gemensamma</b>	<b>Yahoo</b>	<b>Web Wombat</b>	<b>Överlappning</b>	
1.	0	4	2	0 %	
2.	0	3	0	0 %	
3.	0	9	3	0 %	
4.	0	7	6	0 %	
5.	0	9	4	0 %	
6.	1	13	3	18%	
7.	2	9	5	10%	
8.	2	11	7	17%	
9.	1	8	3	0 %	
10.	1	11	4	0 %	
11.	0	3	1	11%	
12.	3	5	4	33%	
13.	0	3	0	20%	
14.	1	7	3	25%	
15.	1	5	1	43%	
16.	3	6	2	37%	
17.	2	7	3	20%	
18.	2	7	2	11%	
19.	1	8	2	20%	
20.	2	10	5	13%	
<b>Samanlagt</b>	<b>20</b>	<b>144</b>	<b>59</b>	<b>9%</b>	

Tabell 12: Visar uträkningen för Jaccard's index som är antalet gemensamma dokument dividerat på antalet unika dokument multiplicerat med 100.

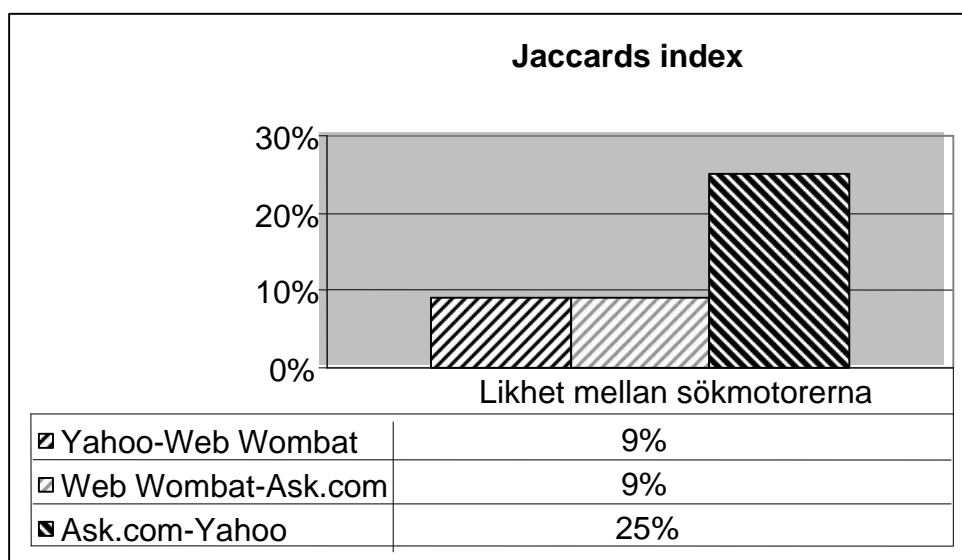


Diagram 10: Överlappningen mellan sökmotorerna.

Ask.com och Yahoo har flest gemensamma dokument. Detta beror på att deras index är större än den lokala sökmotorns. Yahoo hade 144 unika dokument mot endast 59 för Web Wombat. Ask.com hade 78 unika mot 106 för Yahoo. Den högsta överlappningen mellan Ask.com och Web Wombat låg på 43 %. Mellan Yahoo och Ask.com var den 100%.

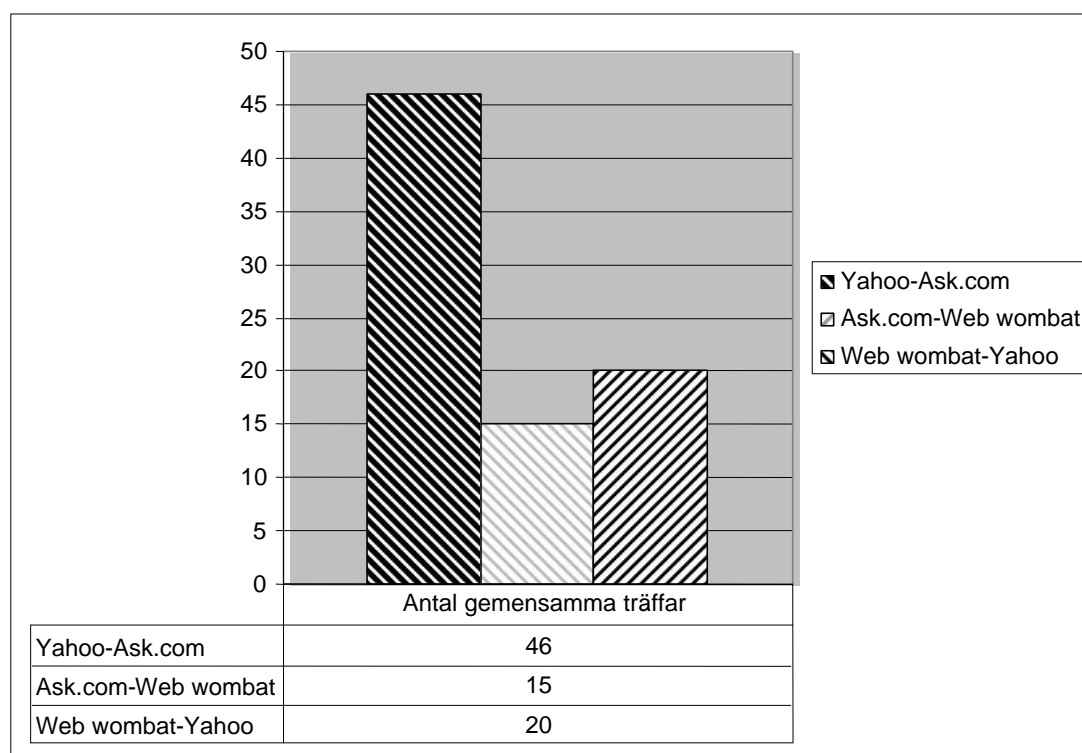


Diagram 11: Även detta digram visar överlappningen genom att se antalet gemensamma träffar mellan de tre sökmotorerna.

Ask.com och Yahoo har flest gemensamma dokument. Detta beror på att deras index är större än den lokala sökmotorns. Flest antal gemensamma var tre för Yahoo och Web Wombat och tio mellan Yahoo och Ask.com.

### 5.4.2 Antal felmeddelanden

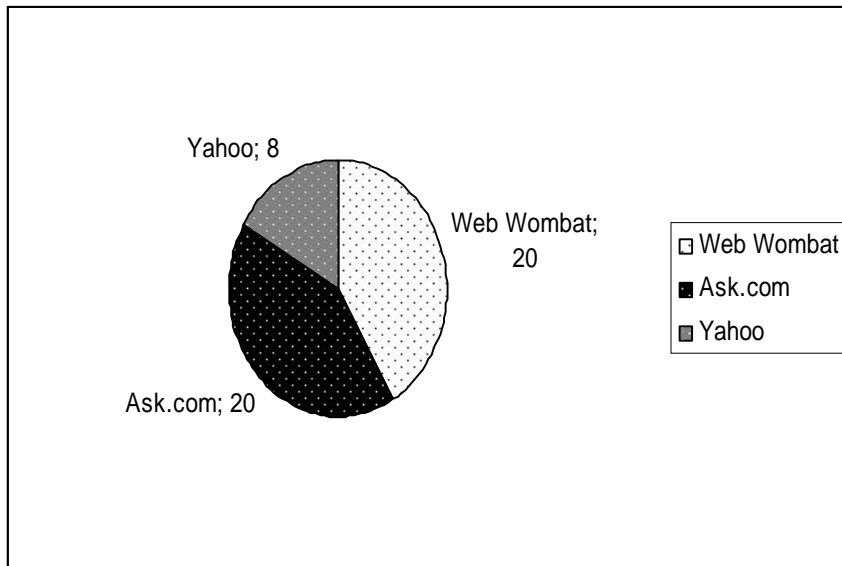


Diagram 12: Diagrammet visar hur många felmeddelanden respektive sökmotor hade sett över alla 20 sökfrågor.

Web Wombat och Ask.com har flest felmeddelanden.

### 5.4.3 Dubbletter

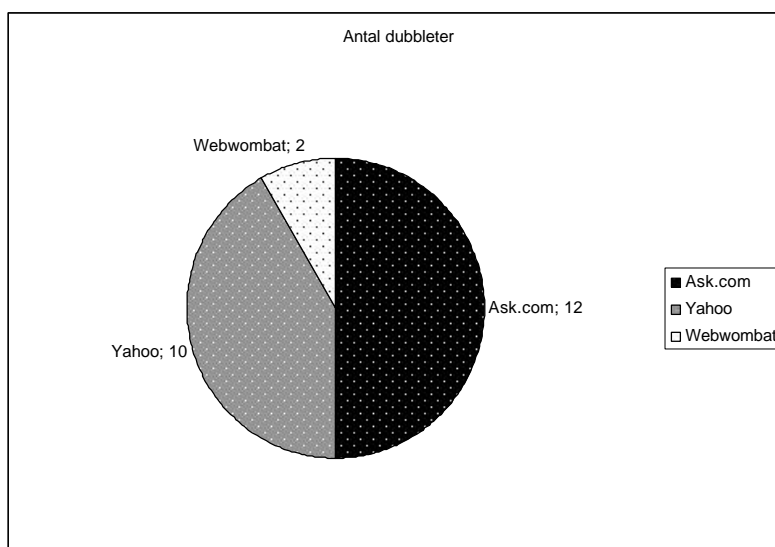
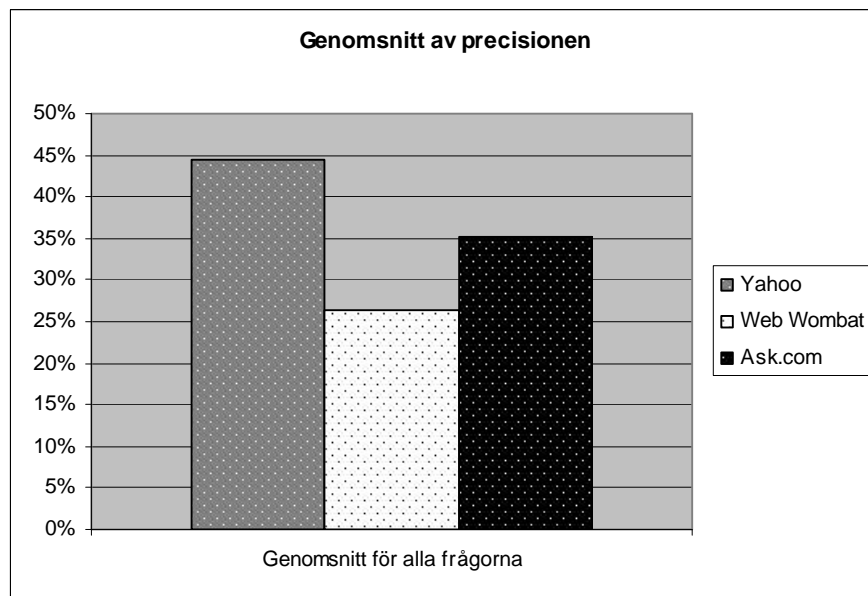


Diagram 13: Diagrammet visar antalet dubbletter per sökmotor sett över alla 20 sökfrågorna.

Web Wombat hade minst antal dubletter. Många dubletter återfanns istället bland dom irrelevanta dokumenten

#### 5.4.4. *Genomsnitt av precisionen vid DCV 20*



*Diagram 14: Diagrammet visar genomsnittet för precision vid DCV 20*

Yahoo har högst genomsnitt med sinna 45 %. Därefter kommer Ask.com med 35 %, och WebWombat med sinna 25 %.

## 6. Diskussion

### 6.1 Diskussion av resultat

I det flesta fall hade Yahoo bäst resultat, för precision och genomsnittsprecision. Detta tror vi beror på en effektiv spindel, ett index som uppdateras ofta och ett effektivt rankningssystem. Att Yahoo lyckades bäst på de flesta av frågorna beror alltså inte enligt oss på att det är den största sökmotorn. Ett exempel på att storlek inte alltid är avgörande för bra återvinning är att Web Wombat hade på ett fåtal frågor bättre återvinning än Ask.com. Yahoo fick säkert också ett bättre resultat än de övriga två, på grund av att den var bättre på att ranka de relevanta träffarna högt i träfflistan. Web Wombat t. ex. visade med procent i sin träfflista hur lika sökmotorn tolkade träffarna mot sökfrågan. Vi upptäckte snart att deras ranking inte alltid är så bra. Ofta fann vi svaret på vår sökfråga bland de träffarna som rankades lägre av Web Wombat.

När det gäller antal felmeddelanden för de tre sökmotorerna, så hade Web Wombat och Ask.com lika många felmeddelanden. Ger sökmotorn ett stort antal felmeddelanden tyder det på att indexet inte uppdateras tillräckligt ofta. Därför blev vi lite konfunderade när Ask.com fick så många felmeddelanden. Vi trodde att den uppdaterades oftare, och skulle då ge mindre felmeddelanden. Det hade varit bra om sökmotorerna tydligare angett hur ofta de uppdaterar sitt index och hur detta går till. T. ex. vilken del av indexet som uppdateras. Sökmotorernas index kan bli lidande av att det blir för dyrt att upprätthålla ett väl fungerande index. Vi tycker att det verkar som att storleken på indexet är viktigare än kvaliteten på det.

Ask.com var även den sökmotor som hade flest dubletter. Web Wombat hade minst antal dubletter. Detta brukar innebära att en sökmotor har en mer unik samling av dokument än en sökmotor med många antal dubletter. I vårt fall fann vi dock, att bara för att Web Wombat hade ett litet antal med dubletter innebar detta inte automatiskt att den hade bra precision. För bland de irrelevanta dokumenten fanns många dubletter.

När det gäller den genomsnittliga precisionen, fann vi följande. Studerarna man varje fråga så var Web Wombat ensam om att ha 0 % på en fråga. Trots låg genomsnittlig precision hade Web Wombat ganska många träffar långt upp i träfflistan. Med långt upp i träfflistan menar vi de fem första dokumenten. Vi anser att Web Wombats ranking medför sökmotorns dåliga genomsnittliga precision. Web Wombat hade många irrelevanta träffar och sidor med reklam på de första sidorna. I ett exempel vad gäller frågan Australiens populäraste sporter så hade Web Wombat:s hela första sida en träfflista med massor av irrelevanta träffar som handlade om lotto och spel.

Jaccard's index mäter överlappningen mellan olika sökmotorer, och i vilken utsträckning olika sökmotorer återvinner samma dokument. Detta mått tycker vi är viktigt att studera och mäta, för att jämföra likheten och olikheten mellan sökmotorerna. Efter att ha genomfört testet kom vi fram till att sökmotorerna inte var särskilt lika och återvann inte samma resultat på sökfrågorna. De globala sökmotorerna var mest lika varandra med ett genomsnitt på 25 %. Ett resultat som visar att sökmotorerna inte återvinner samma träfflista i så hög grad. Vi trodde innan testet, att likheten mellan de globala sökmotorerna skulle vara högre. Den lokala sökmotorn var som vi trodde innan testet olik de andra två sökmotorerna. Web Wombat och Ask.com hade en genomsnittlig likhet på 9%. Web Wombat hade lika mycket gemensamt med Yahoo, där den blev 9%. Som man kan se så skiljer resultatet mellan de två globala och

den lokala ganska mycket. I litteraturen runt sökmotorer bl. a Våge så skrivs det att sökmotorerna bör eftersträvas så hög överlappning, likhet mellan sökmotorerna<sup>63</sup>. Vi är kritiska till detta tänkande att dålig överlappning mellan sökmotorer alltid behöver betyda något negativt. Dålig överlappning betyder ju att sökmotorerna fungerar på olika sätt och inte indexerar samma antal dokument på en sökfråga. Användaren erbjuds alltså flera unika dokument och flera svar på sin sökfråga. Hade överlappningen fungerat bra och alla sökmotorer hade återvunnit från ungefär samma samling av dokument, så kunde användaren lika gärna söka med hjälp av en metasökmotor eller en av de stora sökmotorerna som Google, Yahoo, Ask.com, Alltheweb etc. Sen är det så att Jaccards index är känsligt för stora storleksförhållanden. Detta innebär att de globala sökmotorerna med sinna stora index har mer gemensamt än den lokala med ett mindre index.

Ask.com:s Expertrank trodde vi skulle göra att Ask skulle ha bra återvinning på de flesta av sökfrågorna. Detta eftersom Sökmotorn Google också använder sig av PageRanking och den brukar återvinna många relevanta dokument på olika sökfrågor. Ask.com hade som sagt inte så bra återvinning som vi trodde att den skulle ha. Den presterade som oftast sämre än Yahoo men hade bättre återvinning än Yahoo på några sökfrågor. Det kan ha berott på deras ranking.

Även fast Web Wombat hade sämre återvinning och lyckades sämst i vårt test, så ser vi ändå några anledningar till att använda den. Det visa sig t. ex. att den hade många unika och bra dokument. I vårt test lyckades inte Web Wombat så bra men vi tycker än då att det är positivt att den kan prestera bättre än Ask.com på vissa frågor. Det visar att de lokala sökmotorerna faktiskt kan konkurrera med även de största sökmotorjättarna.

Att en användare bör använda flera olika sökmotorer för att lösa sitt informationsproblem kan bara vår undersökning bekräfta. Detta kan mätas genom att som i vårt fall studera likheten av Jaccard's index för de olika sökmotorerna. Där fann vi också att de var väldigt olika. Sedan har även testet visat att de olika sökmotorerna har presterat olika bra på de varierade frågorna.

När vi jämförde vårt resultat med vad andra kommit fram till i tidigare forskning, så hade de ofta högre precision än oss. Detta kan bero på en hel rad faktorer. Den främsta anledningen tror vi är att vi hade en mycket hög relevansbedömning på vissa frågor. I exempelvis. frågan 16, så nöjde vi oss inte enbart med att texten på nationalsången förekom utan att upphovsmannen skulle förekomma i dokumentet.

## **6.2 Sökfrågor och söktermer**

Termerna som vi använde i de olika sökfrågorna kanske inte alltid var de mest lämpliga. Bland annat så märkte vi att termerna snakes och venom inte var helt lyckade för att få reda på de giftigaste ormarna i Australien då vi fick många träffar som handlade om ormgift och serum och inte just vilka som var de farligaste ormarna i Australien. Precisionen kunde ha höjts genom att vi sökt på frasen most venomest snakes in australia, men vi hade formulerat testet som att vi sökte på enstaka ord eller termer och inte på fraser. Andra termer kunde ha använts för orm bl.a. ordet serpent, viper m.m.. Dock verkar ordet snake vara det mest använda ordet i de flesta av dokumenten. Venom kunde också ha ersatts av poison. Så som vi påpekat kunde vi ha formulerat en sökfråga med alla dessa termer.

---

<sup>63</sup> Våge, Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på internet*. s. 46-47

Ett annat exempel där vi inte tänkte till riktigt var frågan om Australiens premiärminister. Vi ville veta vem som var Australiens premiärminister nu. Vi använde termerna australia och prime minister. Då gav sökfrågan ett resultat, där de bland träffarna förekom andra länders premiärministrar som t. ex. Fiji, Papua nya guinea etc., samt premiärministrar som har varit innan den nuvarande. Detta visar att en del sökfrågor inte gav så bra precision på grund av att de inte var tillräckligt specifika.

En term som rules kan vara knepig då den kan betyda både regera och regler. Det gör att användaren får en del träffar som beskriver hur bra Australien är på australiensisk fotboll istället för reglerna till sporten.

Efter att vi genomfört vårt test kom vi att tänka på vanliga förekomster i Australien som vi glömt att göra sökfrågor av, vi glömde bland annat att fråga om dingo, koala, känguru och wombat. Alla dessa typiska företeelser för Australien. Vi märkte i vårt arbete hur lång tid det tar att utforma sökfrågor och att det är problematiskt att konstruera dem så att de blir bra.

Hur många sökfrågor är då lämpligt att skapa för ett test och ställa till sökmotorn? Vi diskuterade detta fram och tillbaka. Vi anser att 20 stycken är lagom. Problemet är att man inte vill ha för många eller för få sökfrågor, så att resultatet av testet blir missvisande. Om vi t. ex. hade använt för få sökfrågor så hade kanske en sökmotor fått antingen ett väldigt bra resultat eller dåligt, beroende på en extrem. Extremen kunde påverka så genomsnittet blivit missvisande. Extremen skulle i detta fall vara att en fråga skulle ge ett väldigt bra resultat i relation till de övriga nio frågorna.

### **6.3 Diskussion runt tidigare forskning**

Vår undersökning skiljer sig från den som Pettersson gjorde, eftersom den lokala sökmotorn som vi använder har mycket större index än Svesök. Pettersson kom precis som oss fram till att den lokala sökmotorn presterade sämst. Skillnaden ligger i att Web Wombat i jämförelse med Svesök presterade bättre i konkurrens med de globala sökmotorerna. En stor skillnad mellan vår och Petterssons undersökning är antalet dubletter och felmeddelanden. Web Wombat hade endast 2 dubletter mot Svesöks 22 och antalet felmeddelanden för Web Wombat är 20 mot Svesöks 220. Skillnaden i resultat kan bero på att Pettersson hade tio frågor till och att svesök har ett betydligt mindre index.

Likt annan forskning runt sökmotorer, så lyckades de globala sökmotorerna bäst. Yahoo som tillhör en av sökmotorsjättarna likt Google brukar visa god återvinningseffektivitet i tidigare test som gemonförts.

I undersökningen *Measuring Search Engine quality*<sup>64</sup> så presterade Web Wombat sämre än de konkurrerande globala sökmotorerna. De kom fram till ungefär den bilden vi har fått efter vår undersökning. Det som skiljer vår undersökning från deras är att deras är mer omfattande. De har bl. a 54 sökfrågor och jämför tjugo sökmotorer. Dessutom förekommer mer än en lokal sökmotor. I deras undersökning förekommer inte Jaccards index, vilket det gjorde i vårans.

---

<sup>64</sup> se kapitel 3.4 Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathleen (2001) *Measuring search engine quality*

Undersökningen *Think local, search global? Comparing search engines for searching geographically specific information*<sup>65</sup> visar även den att de lokala sökmotorerna presterade sämre. Det som är värt att notera är att undersökningen bara har hade tio frågor. Färre antal frågor ökar risken för extremer och därmed ett missvisande resultat. Det vore intressant att se om samma undersökning med 20 frågor skulle ge samma eller liknande resultat.

*How do search engines respond to some non-English queries?* Säger att när det gäller morfologi så har den lokala sökmotorn en fördel, därför att det är ekonomiskt olönsamt för globala sökmotorer att erbjuda sina tjänster på vissa språk. Sökmotorerna tar inte alltid hänsyn till språkens unika utformning som t. ex. apostrofer och plural/singular. För de som använder Google och inte söker på engelska kan det vara svårt att se att sökmotorns morfologiska brister gör att mycket information går förlorad<sup>66</sup>. Det språkliga utgör för vår del ingen fördel eftersom vårans var engelskspråkig. Men faktum är att de språkliga bristerna framhäver behoven av lokala sökmotorer.

#### **6.4 Allmänt om sökmotorer**

För framtida forskning tycker vi det behövs mer forskning om lokala sökmotorer. Anledningen är att det idag finns väldigt få undersökningar med lokala sökmotorer, vilket minskar deras inflyttande på Internet.

Sökmotorerna måste även ta tag i problemet att de inte alltid ses som en seriös informationskälla. De måste också se till att deras index alltid är väl uppdaterad. Många sökmotorer dras med problemet att det material de indexerar är inaktuellt, vilket innebär att de har många felmeddelanden och dubletter.

De flesta sökmotorer kommer från USA, vilket har gjort att engelska språket blivit det dominerande. En anledning till att engelska är det dominerande språket är att det är enklare att kommunicera och sprida information om det användes ett standardspråk på Internet.

Dessutom är engelska ett språk, som är lätt rent morfologiskt.

Detta kan dock komma att ändras när länder som Kina och Indien kommer ikapp den tekniska utvecklingen. I framtiden kanske dokument på kinesiska kommer att dominera webben? I Indien talar de både engelska och hindi, och med en teknisk utveckling kanske hindi får en mer framträdande roll på Internet.

Det användaren bör tänka på, är att en hel del av kategorierna och sökverktygen kan innefatta endast information från och om USA och Kanada. Vid t. ex. sökning på "local" i Yahoo, så har de endast kartor över USA och Kanada. Eftersom sökmotorn utgår från USA blir därför USA och även Kanada lokalt. Sökmotorerna tycker vi behöver bli tydligare på att förklara för användare vad local innebär och hur andra sökverktyg fungerar och särskilt vilken typ av information användaren kan förvänta sig att få.

---

<sup>65</sup> se kapitel 3.5 Alistair G. Smith (2003) *Think local, search global? Comparing search engines for searching geographically specific information*

<sup>66</sup> se kap 3.7 Bar –Ilan Judith och Gutman Tatyana(2005) *How do search engines respond to some non-English queries?*

## 7 Slutsatser

Våra frågeställningar i undersökningen löd:

- Vilken sökmotor presterar bäst utifrån återvinningseffektivitet?

Det är inte lätt att ge ett enkelt svar på frågan om vilken sökmotor som hade den bästa återvinningseffektiviteten, då sökmotorerna presterade olika bra beroende på vilken av återvinningsmetoderna som användes. Att en sökmotor är bättre eller sämre på att återvinna material för en viss query, kan bero på att sökmotor har mer relevanta dokument indexerade inom vissa ämnen och tvärtom.

### **Precision vid DCV 20 över 20 frågor.**

Den sökmotor som hade högst precision var Ask.com med 75 % på en fråga. Lägst precision hade Web Wombat med 0 % på en fråga. Yahoo hade en jämn nivå med bra återvinning på sin precision med 70 % på tre frågor. Sett till alla frågor hade Yahoo högsta genomsnittet över alla frågor med 45 %. Att Yahoo hade högst genomsnitt berodde på att de inte någon fråga understeg 15 % till skillnad från t. ex. Ask.com som hade 5 % som lägsta värde på en fråga. Så ser vi till bäst precision över enskilda frågor har Ask.com bäst återvinning av precision. Ser vi däremot över alla tjugo frågorna har Yahoo bäst återvinning.

### **Genomsnitt precision för sökmotorerna**

Den sökmotor som hade högst genomsnittlig precision var Web Wombat med 100% Samtidigt var den ensam om att ha 0 % för en fråga. Den bästa genomsnittsprecisionen på en fråga var 93 % för Ask.com och 90 % för Yahoo. Vid användande av detta mått tjänar sökmotorerna på att deras träffar hamnar så högt som möjligt alltså vid DCV nivå ett till fem. De skall även återvinna få dokument. Återvinnns dokument lägre ner på DCV nivån t. ex. att sökmotorn har tre höga träffar vid positionerna 1-3 sjunker ändå genomsnittliga precision om träffar återvinnns vid lägre DCV nivåer som t.ex. vid positionerna 16-20. Högsta genomsnittet av den genomsnittliga precision sett alla frågor var Yahoo, med 64 %.

### **Mätning av överlappning med Jaccards index**

Överlappningen alltså likheten mellan sökmotorerna var högst mellan de Globala sökmotorerna med 25%. De två Globala hade samma överlappning, 9 % med Web Wombat. Den slutsats vi drar är att de globala sökmotorerna har mer gemensamma dokument på grund av storleken på deras index. Att ha en dålig överlappning ser vi inte som en nackdel eftersom det innebär att sökmotorn har många unika dokument.

### **Dubbleter**

Web Wombat hade minst antal dubletter, med två stycken. Detta kan bero på att Yahoo och Ask.com generellt kanske återvinner fler antal dokument per fråga och att detta i sin tur leder till att de riskerar att återvinna fler antal dubletter.

### **Fel meddelanden**

Web Wombat och Ask.com hade lika många fel meddelande, 20 st. Yahoo hade åtta stycken fel meddelanden. Ett högt antal fel meddelanden brukar indikera att sökmotorn inte uppdateras så ofta, så att spindeln söker och återvinner inaktuella webbplatser.

## Anledningar till att välja en lokal sökmotor

Finns det då någon anledning att välja den lokala sökmotorn Web Wombat, Ja, det tycker vi eftersom enligt vårt test så har inte de tre sökmotorerna så många gemensamma träffar. Web Wombat erbjöd många unika och relevanta dokument på våra sökfrågor. De relevanta träffarna som återvanns var ofta väldigt bra och gav en tillfredställande bild av sökfrågan. Dock är det negativa i detta sammanhang att samma sak gäller det motsatta, alltså dåliga träffar var väldigt irrelevant. Web Wombat hade också bäst genomsnittliga precision på en fråga vid DCV 20, samt högst genomsnittlig precision för en enskild fråga.

En annan anledning att välja en lokal sökmotor framför en global kan vara om användaren söker efter dokument på andra språk än engelska. Därför bör forskarna i framtiden genomföra fler test om sökmotorer som använder andra språk än engelska

Vi anser att det finns tre viktiga faktorer till varför en sökmotor har bra återvinningseffektivitet.

- Indexets storlek.
- Uppdateringsfrekvensen på indexet.
- Sökmotorns rankningsfunktioner.

Som användare föredrar vi ofta att använda vissa sökmotorer framför andra. Oftast kanske vi som användare bara använder oss av en sökmotor som vi tycker täcker vårt informationsbehov. Att användare bara söker med hjälp av en sökmotor innebär att de kan missa mycket relevanta information, som inte kan täckas in av bara en sökmotor. Samma sak gäller även för användare av en ”metasökmotor”. Att använda bara en sökmotor gör inte bara att användaren finner information utan att denne också exkluderar information på grund av de olika syntaxerna hos sökmotorerna. Sökverktygen hos de olika sökmotorerna skiljer sig också åt, vilket leder till att t.ex. viktningen av olika termer i en query kan vara olika så att återvinningen av samma query kan se helt annorlunda ut. Även sättet att rankningen skiljer sig mellan sökmotorerna t.ex. har vissa sökmotorer unika rankningssystem så som Ask.com och deras ExpertRank.

Det är svårt att få en balansgång för hur användarna skall kunna få en bra återvinningseffektivitet och att sökmotorerna inte avslöjar sina företagshemligheter så som rankningsfunktioner. Vi tycker att Sökmotorerna bör informera användarna bättre om hur deras rankning fungerar så att användarna skall kunna skapa så bra queries som möjligt. Det finns som sagt en risk med detta om företagen avslöjar för mycket om hur rankningen sker kan illvilliga personer påverka rankningen och att deras dokument hamnar högst upp på träfflistan fast det är irrelevant för den query som användaren har utformat

## 8. Sammanfattning

Denna uppsats tar upp jämförelsen mellan två globala sökmotorer och en lokal sökmotor. Det som jämförs är återvinningseffektiviteten. Huvudsyftet med denna uppsats är att undersöka huruvida den lokala Australiensiska sökmotorn Web Wombat presterar bättre, sämre eller lika bra som de två globala sökmotorerna Yahoo och Ask.com. Med presterar bättre, sämre eller lika bra syftar vi på den lokala sökmotorns återvinningseffektivitet är för frågor som är kopplade till Australien. Dessa frågor handlar bl a om Australiens natur, sport och turism.

Frågeställningarna i uppsatsen är följande:

- Vilken sökmotor presterar bäst utifrån återvinningseffektivitet?

I kapitel två beskrivs Information Retrieval(IR) teori bl.a. diskuterar vi kritiskt runt användningen av de vanligaste måtten precision och recall. Detta kapitel tar även upp hur en sökmotor fungerar och historiken kring den. En sökmotor består av en spindel ett index och ett gränssnitt och mjukvara.

I kapitel tre studerar vi tidigare forskning runt globala och lokala sökmotorer. Åtta olika undersökningar presenteras. Olika mått som används är Spermans rho, relativ recall och precision. I studie 3.5 förekommer Web Wombat. Resultatet för sökmotorn blir att den är sämre än dom globala på. På precision för position 1-5 hamnar Web Wombat tredjen sist. Detta kapitel avslutas med en sammanfattning av tidigare forskning av samtliga studier.

I kapitel fyra, metod beskrivs vårt tillvägagångssätt för hur testet har genomförts, vi har mätt återvinningseffektiviteten hos de tre sökmotorerna genom att använda måtten vi använt är Jaccards index, Genomsnittsprecision och precision vid DCV 20. Här följer definitioner för respektive mått. Precision är Antalet relevanta dokument som återvunnit/totala antalet dokument som återvunnits.

Genomsnittsprecision är att det exempelvis vid DCV 20 finns dom relevanta dokumenten vid plats 5,10 och 18. Uträkningen blir  $0.2+0.2+0.17/3=0.19$ .

Jaccards index definieras som: anta att Yahoo har återvunnit 12 dokument och Ask.com har återvunnit 14 dokument. Om bara 3 av dokumenten är gemensamma innebär det att 9 dokument är unika för Yahoo och 11 dokument är unika för Ask. Snittet för Yahoo och Ask blir då 3 dokument och unionen blir 20 dokument  $(9+11)$  vilket ger  $J=3/20$ .

Vi går även in på val av sökmotorer, hur informationsbehov, sökfrågor har formuleras och varför. För vår undersökning har vi använt oss av en kvantitativ metod.

I kapitel fem presenterar vi resultatet av vårt test genom tabeller och diagram. Tabellerna visar bl a att överlappningen mellan sökmotorerna är högre mellan dom globala (25%), en mellan dom globala och den lokala (9%).Antalet felmeddelanden var hela 20 för Ask.com och Web Wombat men bara åtta för Yahoo. Ask.com och Yahoo hade hela 46 gemensamma träffar medans dom gemensamma träffarna för Yahoo och Web Wombat var 20 och mellan Ask.com och Web Wombat 15.

I kapitel sex, diskussion analyseras orsaker till varför resultatet blev som det blev. Vi jämför även vårt test med tidigare forskning om lokala och globala sökmotorer och ser om vi kommit till samma resultat och slutsatser eller om vårt avviker. Vi analyserar dessutom varför de avviker eller ej. Exempelvis presterade den lokala sökmotor bättre i vårt test än den gjorde i

den andra undersökningen som genomfördes i en magisteruppsats skriven av Petterson. Vi tror att resultatet beror på att vår lokala sökmotor, Web Wombat hade större index som uppdaterades oftare. Sedan har nog den australienska lokala sökmotor mer kapitalresurser till att sköta underhåll, uppdatering och konstruerande av en effektiv spindel.

Vi anser att det finns tre viktiga faktorer till varför en sökmotor har bra återvinningseffektivitet.

- Indexets storlek.
- Uppdateringsfrekvensen på indexet.
- Sökmotorns rankningsfunktioner.

I kapitel sex, diskussion analyseras orsaker till varför resultatet blev som det blev. Vi jämför även vårt test med tidigare forskning om lokala och globala sökmotorer och ser om vi kommit till samma resultat och slutsatser eller om vårt avviker. Vi analyserar dessutom varför de avviker eller ej. Exempelvis kunde vi konstatera att det höga antalet felmeddelanden för Web Wombat och Ask.com troligen berodde på dåligt uppdaterade index. Dessutom diskuterade vi val av söktermer. En term som rules kan exempelvis betyda både regera och regler.

I kapitel sju besvarar vi vår frågeställning. Det finns inget enkelt svar på vår frågeställning, om vilken sökmotor som hade bäst återvinningseffektivitet gällande frågor som handlade om Australien. Det beror på vilket mått som används. Det beror också på om användaren eftersträvar hög återvinning på en fråga eller ett högt genomsnittligt värde. Som exempel så hade Ask.com den bästa precisionen på en enskild fråga, 75 %. Ser vi till vårt test över 20 frågor så hade dock Yahoo den bästa genomsnittliga precisionen med ett genomsnitt på 45 %. Ser vi till hur högt de relevanta träffarna återvanns, så hade Web Wombat bäst genomsnitts precision med 100 % på en fråga. Ser vi till genomsnittet sett över 20 frågor får Yahoo återigen bäst återvinning med ett värde på 45 %. En slutsats vi kommit fram till är, globala sökmotorer tenderar att ha en jämn nivå på återvinningseffektiviteten medan lokala sökmotorer i vissa fall presterar en hög återvinning på enskilda frågor.

## Källförteckning

Ackermann, Ernest C (2000)*The information specialist's guide to searching & researching on the Internet & the World Wide Web* / Ernest Ackermann, Karen Hartman  
Wilsonville, Or. : Franklin, Beedle & Associates.

Alexa.com  
www.Alexa.com (2007-01-24)

Ask.com  
www.ask.com (2006-05-08)

Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*.  
Harlow: Addison-Wesley

Bailey Peter, Craswell Nick, Hawking David och Griffiths Kathleen (2001) *Measuring search engine quality* Information retrieval no 4 s.33-59

Bar –Ilan Judith (2005) *Comparing rankings of search results on the web* Information processing and management no 41

Bar –Ilan Judith och Gutman Tatyana (2005) *How do search engines respond to some non-English queries?* Journal of information science no 31 s.13-28

Chowdhury G.G.(1999) *Introduction to modern information retrieval* /  
London : Library Association Publishing.

Chowdhury G. G.(2001) *Sudatta Chowdhury Information sources and searching on the World Wide Web*, London : Library Association

Eastman M. Caroline och Jansen J. Bernard (2003) *Coverage relevance and ranking: The impact of query operators on web search engine results*.ACM transaction on information systems vol 21 no 4

Egghe, L och Rousseau, R. (1990). *Introduction to informetrics*. Amsterdam: Elsevier.

Feldman, Susan, Liddy, Elizabeth (2001) *The Searching quagmire*. Vol. 9, no 5.

Ince, Darrel (2001) “robot” *Dictionary of the Internet*. University Press  
*Oxford Reference Online*. Oxford University Press. <http://www.oxfordreference.com>  
[2006-01-31]

Liddy, Elizabeth (2001) *How a search engine works*, *Searcher* Vol. 9, no 5

Meyer R John. (1998) *So what's the big deal*  
Department of Entomology *NC State University*  
<http://www.cals.ncsu.edu/course/ent591k/gcextend.html#diversity> [2005-05-03]

Molina R Paulo och Jansen J. Bernard (2006) *The effectiveness of web search engines for retrieving relevant ecommerce links*. Information processing and management no 42 s.1075-1098

Notess R.Greg (2005) *Review of Yahoo! Search*  
<http://www.searchengineshowdown.com/features/yahoo/review.html> [2005]

Pettersson Marcus (2005) *Globala och lokala sökmotorer: En utvärdering av Google MSN search och Svesök*. Borås : Högsk. i Borås, Bibliotekshögskolan/Biblioteks- och informationsvetenskap, 2005

Sherman, Chris & Price, Gary (2001) *The invisible web : uncovering information sources search engines can't see*. Medford, N.J. : Information Today

Smith Alistair G (2003) *Think local, search global? Comparing search engines for searching geographically specific information* Online information review vol 27 no 2 s.102-109

Thorpe Peter *search engine secrets* <https://www.cia.com.au/thorpe/secrets/howtheywork.html> [2006-05-05]

Vaughan Liwen (2004) *New measurements for search engine evaluation proposed and tested*. Information processing and management no 40

Våge,Lars, Dalians, Hercules & Iselid, Lars (2003). *Informationssökning på Internet*. Lund Studentlitteratur.

Web Wombat  
[www.WebWombat.com/au](http://www.WebWombat.com/au) [2006-05-08]

Yahoo  
[www.yahoo.com](http://www.yahoo.com) [2006-05-08]

Yahoo  
<http://yhoo.client.shareholder.com/press/overview.cfm> (2006)