

Statistik från webbplatser

– problem och möjligheter

Ulf Kronman

Föredragets huvuddelar

- Frågorna och motfrågorna
 - Vilka frågor ställer chefen, BIBSAM och ISO?
 - Varför ställer webmastern krångliga motfrågor?
- Tekniken bakom webben
 - Webbens funktion och webbsidornas uppbyggnad
 - Nätets komponenter och komplikationer
- Definition av kvantitetsmått
 - Besökare, besök och sida
- Vad kan mätas?
- Hur ska man mäta?

Frågorna från ledningen och BIBSAM

Chefen

- Hur många besökare hade vår webbplats under förra året?
- Varifrån kom besökarna?
- Hur många sidor tittade de på vid varje besök?
- Vilka sidor efterfrågades mest?

BIBSAM

- Rapportera "Virtuella besök per år, i tusental"

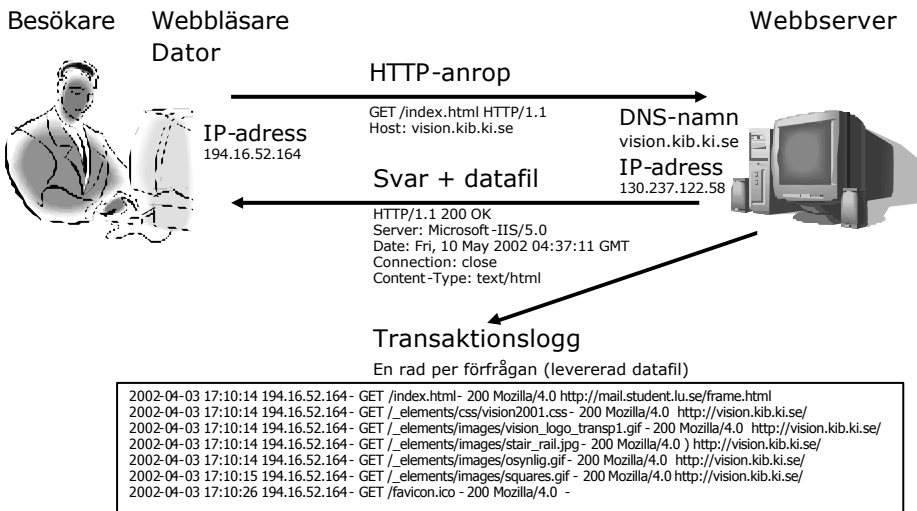
ISO (ISO 2789:2001)

- Collect, if possible, the number of virtual visits

Problemen – webbmasterns motfrågor

- Vad är en besökare ?
 - Webbservern ser inga människor – bara datorer
- Vad menas med varifrån ?
 - Webbservern ser inga platser – bara adresser på nätet
- Vad är ett besök ?
 - Webbservern ser inga besök – bara efterfrågade datafiler
- Vad är en sida ?
 - Webbservern ser inga sidor – bara levererade datafiler
- Vad menas med vår webbplats ?
 - Vilka servrar och tjänster ska räknas?

Tekniken bakom World Wide Web



13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

HTTP – hypertext transport protocol

- Protokollet HTTP är tillståndslöst (stateless)
 - Inga uppkopplingar – sessioner – bevaras mellan överföringarna av filerna
 - Webbläsare och webbserver kopplar upp och ned en förbindelse för varje levererad datafil
- Bristen på sessioner ger problem vid insamling av besökarstatistik
 - Tekniskt sett finns det inga "besök"
 - Sessioner (besök) kan konstrueras med hjälp av så kallade cookies

13 januari 2003

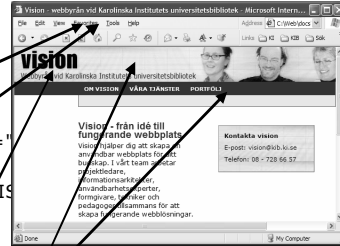
Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Webbsidans uppbyggnad

En webbsida är vanligen uppbyggd av flera olika datafiler

```
<html>
<head>
  <title>Vision </title>
  <link rel="stylesheet"
    href="_elements/css/vision_2002.css" type="text/css" />
  <script type="text/javascript"
    src="_elements/scripts/display.js" charset="ISO-8859-1" />
</head>
<body>
<table background="_elements/images/background.jpg">
  <tr>
    <td><a href="index_se.html">
      
    </a></td>
    <td></td>
  </tr>
</table>
...

```

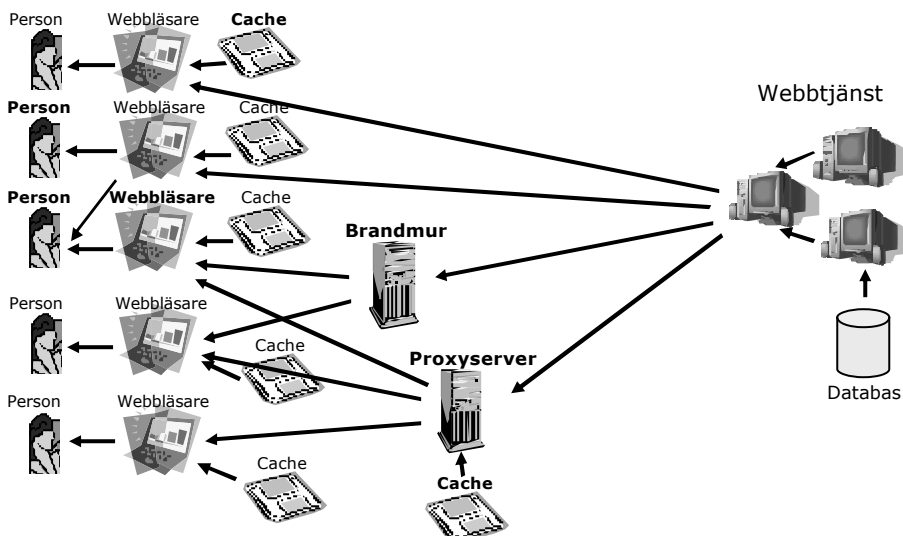


Varje leverans av en datafil genererar en rad i transaktionsloggen

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Cache och proxyserverar komplicerar



13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Cookies identifierar webbläsarna

- En cookie är en liten textfil som lagras av webbläsaren
 - Webbläsaren skickar med cookien till servern vid varje förfrågan
- Permanenta cookies kan identifiera en webbläsare
 - Även bakom brandmurar och proxyserverar
- Temporära cookies kan skapa sessioner (besök)
- Användning av cookies är komplicerad
 - Hantering av cookies kräver någon form av programmering
 - Registrering av cookies kräver utvidgad transaktionslogg
 - Analys av cookies kräver avancerade och kommersiella program

Kan vi nu besvara webmasterns frågor?

- Vad är en besökare?
- Var befinner besökaren sig?
- Vad menas med ett besök?
- Vad menas med en sida?

Problem med begreppet besökare

- Vi vill identifiera personer men har bara IP-adresser
 - Mellan personen och IP-adressen finns webbläsare och dator
 - Ibland även brandmur och proxyserver
 - Proxyserverar gör att många besökare får samma IP-adresser
- Med en cookie kan man identifiera en webbläsare
 - Relationen personer och webbläsare är dock inte entydig
 - En publik webbläsare kan användas av många personer
 - En person kan använda flera olika webbläsare
 - Hemma, på arbetet, på publika datorer
- Inloggning är det säkraste sättet att identifiera personer
 - Inte realistiskt att kräva inloggning på en publik webbplats

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Definition av webbplatsbesökare

- IP-besökare (datorbesökare)
 - En identifierad IP-adress – finns i webbserverns transaktionslogg
 - Många personer får samma IP-adress bakom en proxyserver
- Webbläsarbesökare
 - Använd cookies för att identifiera webbläsarprogram
 - Kräver cookiehantering, utvidgad transaktionslogg och avancerat analysprogram
- Söktjänsternas insamlingsprogram är inte besökare
 - Över 300 olika insamlingsprogram kan avläsas med uppgiften user agent i en utvidgad transaktionslogg

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Var befinner besökaren sig?

- För att bli meningsfulla måste IP-adresser översättas till domännamn av en DNS-server
 - Exempel: 130.237.122.58 -> vision.kib.ki.se
 - Domännamnet visar i vilken organisation en dator finns – ibland även land
- Runt 20% av IP-adresserna går inte att koppla till ett domännamn
 - Domännamn måste registreras manuellt
- Besökare med adresser i domäner som .com, .org, .net och .nu kan befinna sig varsomhelst i världen
- En brandmur/proxyserver mellan besökaren och servern gör att webbservern bara ser proxyserverns IP-adress och domännamn

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Vad är ett besök?

- En svit av leveranser av datafiler till samma besökare
 - Ett uppehåll mer än 30 minuter avslutar besöket
- IP-besök (datorbesök)
 - Svit av datafiler levererade till samma IP-adress
 - Problem med att olika personers besök blandas i proxyserverar
- Webbläsarbesök
 - Svit av datafiler levererade till samma webbläsare
 - Kräver cookiehantering, utvidgad transaktionslogg och avancerat analysprogram
- Leveranser till söktjänsternas insamlingsprogram ska inte räknas som besök
 - Kräver utvidgad transaktionslogg med user agent

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

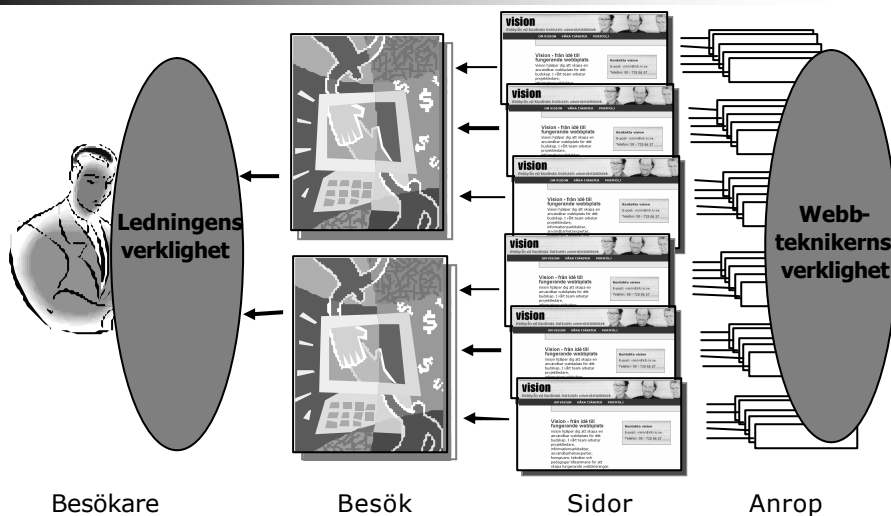
Vad är en sida?

- Vår uppfattning av sidor och webbserverns data skiljer sig åt
 - I webbserverns transaktionslogg representeras varje sida av en rad leveranser av datafiler (anrop)
 - Bara en av datafilerna representerar själva innehållet
- Sortera bort datafiler som inte är innehåll
 - Bilder, stilmallar, JavaScript, ramar (framesets)
- Räkna inte sidor levererade till söktjänsternas insamlingsprogram
 - Kräver utvidgad transaktionslogg med user agent

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Relation besökare – besök – sidor - anrop



13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Summering: Förslag till definitioner

- Webbläsarbesökare / IP-besökare
 - En unik webbläsare / IP-adress, använd av en fysisk person, som hämtat filer från webbplatsen inom mätperioden
- Webbläsarbesök / IP-besök
 - Ett antal sidor levererade med mindre uppehåll än 30 minuter till samma webbläsare / IP-adress använd av en fysisk person
- Sidor (pageviews, impressions)
 - Innehållsfiler levererade till en fysisk person
 - Dekorationselement och navigationsdelar oräknade
- Anrop (hits, requests)
 - Antal datafiler som efterfrågats och levererats
 - Främst ett mått på hur mycket webbservern arbetar

Vilken statistik är mest rättvisande?

- IP-besök – relativt enkelt, men missvisande
 - Långt ifrån bilden av en besökande person
- Webbläsarbesök – rättvisande, men komplicerat
 - Mycket närmare bilden av en besökande person
 - Kräver cookiehantering och avancerad analys
- Sidleveranser – enkelt och relativt rättvisande
 - Ingen cookiehantering och enkel analys
 - Se webbtjänsterna som system för dokumentleveranser

Möjligheterna – vad kan vi mäta?

- Det chefen och BIBSAM frågar efter finns inte i loggfilen
 - Vi kan inte säkert se mänskliga besökare, deras besök eller var de finns
 - Vi kan räkna datorer eller webbläsare och konstruera besök
 - Vi får stora felmarginaler på grund av proxyservrar och cache
- Möjliga kvantitetsmått
 - Antal webbläsarbesökare och webbläsarbesök
 - Antal IP-besökare och IP-besök (felmarginal för proxyservrar)
 - Antal levererade sidor (felmarginal för cache)
 - Antal levererade sidor per IP-besök eller webbläsarbesök
- Besökarnas organisationshemvist
 - Räkna med 20-30% felmarginal
 - Vi kan ofta inte se besökarnas fysiska placering

Mer möjligheter – vad bör vi studera?

- Kvalitetsmått
 - Vilka sidor och delar av webbplatsen som används mest och minst
 - Från vilka webbplatser besökarna kommit (refererande webbplatser)
 - Vilka sökord besökarna använt för att hitta webbplatsen
 - Hur besökarna navigerat genom webbplatsen (path analysis)
 - Hur besökarna söker inom den egna webbplatsen
- Teknisk information
 - Efterfrågade sidor som inte finns (trasiga länkar)
 - Svarstider och mängd levererad data
 - Besökarnas webbläsare och operativsystem

Hur ska man mäta?

- Ställ realistiska frågor
 - Fråga inte efter personer, platser eller "virtuella besök"
 - Fråga efter levererade sidor och relativa kvalitetsmått
- Planera för statistik – helst när webbplatsen skapas
 - Webbplats och webbserver måste anpassas innan man börjar efterfråga statistik
 - Ställ in webbserverna för utvidgad transaktionsloggning
 - Uteslut insamlingsprogram och registrera refererande webbplats
- Skaffa ett analysredskap
 - Analog – gratis och kraftfullt. Mäter ej besök.
 - Webalizer – gratis. Mäter IP-besök.
 - WebTrends – bra och billigt för en webbserver (runt 6000 kr)
 - 123 Log Analyzer – sägs vara bra (\$150)

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek

Frågor & diskussion

Dagens föredrag

<http://vision.kib.ki.se/portfolio/statistics/libraries/>

13 januari 2003

Ulf Kronman - Karolinska Institutets Universitetsbibliotek