DRI1002 - IKT og informasjonssøking
3. Seminar uke 6: Søkeverktøy og søkestrategier

Hovedpunkter for forelesningen

- Databaser og fritekstsystemer – et overblikk
- Hypertekst og WWW
- Søkemotorer – søkestrategier
- Om arkiver og postjournaler som kilde og verktøy for å søke informasjon
- Diskusjon
- Øvelser i informasjonssøking

Et kort historisk overblikk

Datamaskinen – en regnemaskin (compute (r))

- *Digitaliseringen* : Alt representeres ved 0 og 1 (tall, tekst, lyd, bilder, film,..)
- *Formalisering*: Handlingsregler uttrykkes på presis form (matematiske/logiske algoritmer)
- *Strukturering* : Organisering av data i bestemte, (veldefinerte) mønstre

Eksempler på tidligere programmeringsspråk:

Fortran : *Formula* translator,
Algol : *Algoritmic* language

Som begge er utformet for å kunne uttrykke og utføre matematiske og logiske operasjoner

Behandling av strukturerte data – **databaser**

Det vokste raskt fram et behov for å beskrive data på en strukturert form

De første eksempler på EDB-baserte databaser på 50-tallet :

- Folketellingsdata
- Skatt- og ligningsdata
- Bankenes og forsikringsselskapers kundekonti
- Medlemsarkiver
- Regnskaps- og lønnssystemer
- ……

Hva er en database?

Samling med data som er organisert for å tjene et bruksområde. Organiseringen av data er gjort i henhold til en tenkt struktur som beskriver dataenes karakteristikker og sammenhengen mellom dem:

Et databasehåndteringssystem (DBMS – data base management system) er et programsystem som laget for *opprette* og *vedlikeholde* databaser

Den mest vanlige formen for databaser er tabellformen – eller *relasjonsdatabaser*

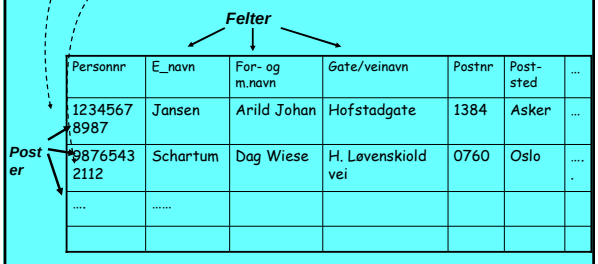
Manuelle databaser –eksempler

- Kirkebøker
- Leksikon, ordbøker
- Kataloger
- Kartoteker,
- Offentlige og private arkiver
- Medlemsregistre
- ....

Eksempel på relasjonsdatabase

Arild Johan Jansen, Hofstadgate , 1384 Asker

Dag Wiese Schartum, Harald Løvenskiolds v , 0760 Oslo



Personnr	E_navn	For- og m.navn	Gate/veinavn	Postnr	Post-sted	...
12345678987	Jansen	Arild Johan	Hofstadgate	1384	Asker	...
98765432112	Schartum	Dag Wiese	H. Løvenskiold vei	0760	Oslo	....
....	.....					

Noen sentrale begreper knyttet til databaser

- *Poster (record) :* En 'linje' i tabellen som inneholder *verdier* i de enkelte feltene
- *Primærnøkkel* : et felt som gir entydig identifikasjon for alle poster (f eks. personnr.)

Vanlige operasjoner

- *Opprette* en database : definere strukturen
- *Vedlikeholde* en database : legge inn/rette data
- *Spørring og rapporter*: Hente ut data fra en database, enkeltvis ellers oversikter

Noen eksempler på databaser

Samordna opptak :

- http://www.samordnaopptak.no/sokerinformasjon_2005/laresteder_studier.html

Visveg

- http://www.visveg.no/visveg/default.jsp

Studentveb

- http://www.uio.no/studier/tilbud/

Statistisk sentralbyrå, f eks.

http://www.ssb.no/navn/

Datamaskinen ble også en tekstbehandler

- Fritekstsystemer :
  - Med *fritekst* mener vi en vanlig prosatekst inndelt i kapitler, avsnitt og setninger - i utgangspunktet uten spesielle skilletegn og markører. Fritekstsystemer har i Norge blitt brukt til databaser over arkeologisk gjenstandsmateriale, utdrag fra middelalderdiplomer og tingbøker innenfor historiefaget.

    Rettslig materiale er kanskje det felt hvor tekstsøking har blitt mest anvendt i Norge, jf de juridiske databasene hos stiftelsen *Lovdata*.

    (hentet fra Heimen 2/1996 Av: *Gunnar Thorvaldsen, se* *http://www.rhd.uit.no/art/heim96-2.html*

Informasjonssøking

- Computer-aided information search and retrieval
  - historie om lag like gammel som datamaskinene
  - første skikkelege gjennombrot på 50-talet i samband med søk og erstatt av uttrykk i lovtekst
  - IR = Information Retrieval
- Før WWW har informasjonssøk særleg vært knyttet til databaser og slik sett databasesøk, men også enkle fritekstsøkesystemer
- Internett/WWW har endra dette ved søk i store, ustrukturerte datamengder

Internett-søk i et historisk lynglimt

- I begynnelsen var.... Archie
  - utvikla i 1990 av Alan Emtage, pre-web søkemotor (ftp)
- The World Wide Web Wanderer (Wandex) – den første søkeroboten på web'en
- Galaxy (1994), den første internett-katalogen
- Excite (1993)
- WebCrawler (1994) – første fulltekstindeksering av web
- Yahoo! (1994)
- 10 år med internett-søk har vist at det skjer raske endringer og mange søketjenester har relativt kort levetid. Yahoo! er en av få tjenester som har vært med heile tida
- AltaVista var ei viktig tjeneste fram til slutten av 90-talet. På berre ca. et halvt år forsvant den nesten helt då Google tok over.

Ulike typer søketjenester

- Katalog
  - menneskeskapt hierarkisk database over nettressursar (Yahoo, Open Directory, LookSmart, Kvasir)
- Søkemotor
  - robot, database, brukargrensesnitt mot database (Google, AltaVista, Teoma, Kvasir...)
  - samme søkemotor kan være motor i ulike tjenester (Google blir brukt i Yahoo, AOL, Kvasir...) – outsourcing av søk!
- Metasøkemotor
  - søkemotor som bruker andre søkemotorer som kilde, parallellsøk i mange underliggende baser
    HotBot, *Queryster*, *DogPile, Excite, MetaCrawler, mamma*
- I praksis er i dag de fleste søketjenester en kombinasjon av kataloger og søkemotorer

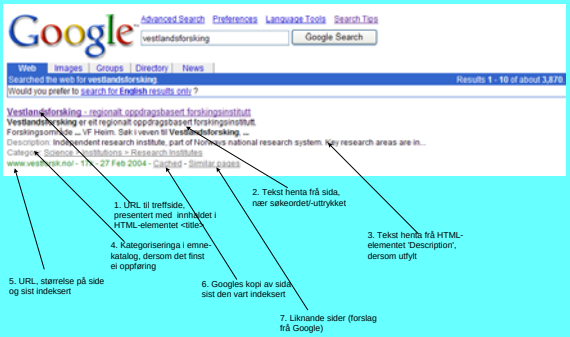
Hva er en søkemotor ?

- I Søkerobot (*crawler, bot, spider, vevkjerring*)
  - program som følger lenker på veven og kopierer informasjon (tekst) inn i den sentrale databasen
- II Database
  - informasjonen samla av roboten blir lagra i en data-base med en del tilleggsinfo
  - indekseringa i etterkant av informasjonsinnhenting inneber m.a. statistikk over ord, plassering av ord i teksten, analyse av lenker m.m.
- III Søkegrensesnitt
  - brukeren sin interaksjon med søkemotoren
  - enkelt søkefelt eller grensesnitt for avansert søk

Søkemotor: Søkerobot

- Søkerobot
  - ikke en, men mange roboter (program) som traverserer nettet og henter inn informasjon
  - en tjeneste som Google vil vanligvis indeksere en vevtjeneste en gang i månaden
  - søkeroboten leser vevsider som en "primitiv" tekstbasert nettleser

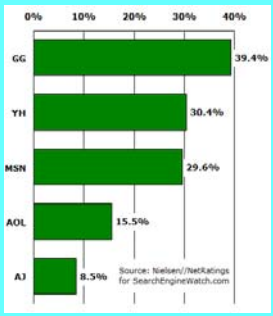
Analyse av resultat-treff, Google



Søkemotor: Søkegrensesnitt

- Søkegrensesnitt
  - Søkeboks for enkle søk
  - Avansert søk med hjelp til avgrensing
  - Problem:
    - Ingen standard for søk i søkemotorar
      - korleis fungerer søket "epler pærer" i Google?
        (finsk undersøking viste at > 60% av brukerene tok feil)
    - Variabel støtte for Boolsk logikk (AND, OR, NOT)
  - For mer informasjon om oppbygging av en søkemotor, les "The Anatomy of a Large-Scale Hypertextual Web Search Engine" av Larry Page og Sergey Brink (grunnleggjarane av Google)

Mest brukte søkemotorar (januar 2004)



GG = Google
YH = Yahoo
MSN = Microsoft
AOL = America Online
AJ = Ask Jeeves

Panel på meir enn 60 000 brukarar i USA

Finst det andre søkemotorar enn Google?

- Fort å se seg blind på en dominerende aktør som Google
  - bør stadig prøve ut andre søkemotorer
  - nisjesøk
    - dersom du søker etter offentlig informasjon i Norge, bør norge.no vera en betre søkemotor enn Google (men ikke heilt sikker på at den er det...)
    - Kvasir er avgrensa til .no-domenet og bør slik sett kunna gi betre resultat enn ei meir omfattande tjeneste (i realiteten er det Google som leverer søkeresultat til Kvasir; skilnaden blir emnekatalogen Kvasir har bygd opp)
    - Startsiden.no tilbyr søk i Google, Kvasir, AltaVista, Yahoo! og Alltheweb (som nå er en del av Yahoo!)
- Google har i dag en for dominerande rolle siden den også blir brukt som motor for mange andre konkurrenter
- Meir informasjon:
  - Search Engines and controversy: http://www.firstmonday.dk/issues/issue9_1/gerhart/

Hvilken søkemotor skal jeg bruke

- Same søkemotoren kan være brukt på mange søketjenester:
  - Google er også søkemotor for tjenester som
    - Yahoo! (heilt fram til årsskiftet)
    - AOL (America Online)
    - Kvasir (Scandinavia Online – SOL)
  - Yahoo! har gjennom oppkjøp følgjande søkemotorar:
    - Inktomi
    - AltaVista
    - AlltheWeb (FAST)
    - gjennom Inktomi gir dei søkeresultat for MSN (Microsoft)

Synlege vevtenester

- For sluttbruker er søkegrensesnittet den synlige delen av søkemotoren
- For tjenestetilbydar er søkeroboten den viktigste delen
  - søkerobotar les vevsider som "primitive" nettlesarar
  - http://www.delorie.com/web/lynxview.html for å sjå korleis søkemotoren les sidene
- Den usynlige (skjulte) vev
  - http://www.hib.no/biblioteket/Kildekritikk.asp#aapen

  eksempel på usynleg side: www.kjornes.no/start.htm

Øvelser i søking 7.2

1. Går inn i Lovdata. Søk etter lover mm som omtaler hvilke rettigheter og plikter dere har som studenter
   - Hva fant dere?
   - Hva slag type søking tror dere ble utført (databasesøk, fritekst,…) ?
2. Bruk minst to åpne søkemotorer og gjør tilsvarende søk, både enkle og mer avanserte
   - Hva fant dere ? Sammenlign ulike resultater fra ulike søkemotorene
   - Forklar hvorfor ble dette forskjellig for ulike søkemotorene. Drøft resultatet.