Aarhus Universitet
Institut for Datalogi

Den 3. oktober 2011

dPersp11 – ugeseddel for uge 40
Internetalgoritmer

Ugens program

Mandag 3/10	10.15-11.00	Internetsøgemaskiner, PageRank, inverterede filer *Gerth Stølting Brodal (Store Aud)*
	11.15-12.00	R-træer, geografisk internetsøgning *Christian S. Jensen (Store Aud)*
Tirsdag 4/10	9.15-11.00	Øvelser omkring mandagens forelæsning (Open Learning Center)
	11.00-11.45	Historisk perspektiv *Erik Meineche Schmidt (Store Aud)*
	12.15-12.45	MapReduce *Gerth Stølting Brodal (Store Aud)*
	13.15-15.00	Øvelser – MapReduce (Open Learning Center)
Onsdag 5/10	14.15-16.00	Peer-to-Peer Networking: Lokalisering og deling af data, *Niels Olof Bouvin (Aud F)*

Øvelser

9.15-10.00	Øvelser omkring PageRank og inverterede filer
10.15-11.00	Øvelser omkring geografiske søgninger
13.15-15.00	Øvelser og afleveringsopgave omkring MapReduce

Om ugens forelæsninger

Denne uges forelæsninger omhandler teknikker og algoritmer der i høj grad har påvirket vores brug af internettet de seneste år.

Internetsøgemaskiner
Med introduktionen af Google i 1998 blev det muligt at søge effektivt blandt de milliarder af websider der findes på internettet. Nogle af de egenskaber der var fundamentet for Google's succes var hastigheden hvormed forespørgsler udføres, at de fleste sider på internettet var indeholdt i Google's indeks, og ikke mindst at de returnerede svar indeholdt de meste relevante sider om det man søger efter. I forelæsningen beskrives de basale idéer der er grundlaget for Google's virkemåde.

R-træer, geografisk internetsøgning
Internettet bruges i stigende grad fra mobile enheder, f.eks. smartphones. Det er i stigende grad muligt at positionere disse enheder geografisk. Desuden har studier vist at hen imod 20% af alle desktop søgninger og 50% af alle søgninger fra mobile enheder har "local intent", hvilket betyder at brugeren ønsker at finde websider, der vedrører objekter der ligger tæt på, f.eks. en webside for et sted, der sælger pizza. I forelæsningen beskrives en udbredt teknik til hurtigt at finde objekter i et geografisk område, og det beskrives hvordan denne teknik kan udvides til også at gøre det muligt at finde websider som matcher en Google-forespørgsel og samtidig repræsenterer et objekt tæt på brugeren.

MapReduce
Et af fundamenterne i Google's succes er at have mange maskiner (størrelsesorden en million) forbundet i et stort netværk, og ved at effektivt udnytte den massive parallelisme dette giver. Traditionelle software metoder til programmering af parallelle algoritmer medfører dog i praksis et meget stort overhead til håndtering af tilbagevendende problemer af parallel natur. Google introducerede derfor i 2004 en elegant software omgivelse kaldet MapReduce, der måske ikke kan løse alle problemer effektivt parallelt, men tillader en simpel måde at designe parallelle algoritmer, der automatisk håndterer alle aspekter der vedrører udnyttelsen af parallelisme. Der findes en open source variant af MapReduce, Hadoop, som bl.a. bruges intenstivt af Yahoo*!*

P2P netværk
To udfordringer indenfor P2P netværk vil blive præsenteret: (1) Hvordan lokaliserer man data? Et P2P netværk er, per definition, decentralt, så det er ikke muligt at referere til en central autoritet (såsom en søgemaskine), når man ønsker at finde ting. Vi vil se på to forskellige metoder, som finder anvendelse i ustrukturerede og strukturede netværk. (2) Hvordan henter man data? Når man først har identificeret en ressource, er næste udfordring at hente den. I dag foregår det ofte ved hjælp af BitTorrent. Vi vil se på, hvordan BitTorrent fungerer og hvordan det kan angribes.