Gemensam intagning till datavetenskap 26.5.2014

Uppgift 1: Big data och deras användningsutmaningar

Läs först noggrant bakgrundsmaterialet nedan och besvara sedan de påföljande frågorna.

Ute i världen mäts kontinuerligt en stor mängd olika saker, till exempel radiosignalers kvalitet, och dessa mätvärden utgör data. Också uppdateringarna i sociala medier och butikskedjornas kassauppgifter utgör data. Olika format data från olika källor skapas hela tiden mera och mera och det är enkelt att distribuera data över hela världen. Vi talar om ”big data” då det gäller så stora och komplicerade mängder data att deras analys och utnyttjande är utmanande och kräver utveckling av nya metoder. Trots detta har man ställt stora förväntningar på användningen av big data. Man väntar sig att de ger nya businessmöjligheter och möjligheter att lösa problem som tidigare inte kunnat lösas.

Då man karaktäriserar big data, relaterar man ofta följande ord till dem:

- *Volym:* Data produceras kontinuerligt i så stora mängder att deras behandling och lagring är utmanande. De kan också vara fysiskt på olika ställen, vilket ställer tilläggsutmaningar på databehandlingen.
- *Fart:* Farten med vilken data produceras stiger hela tiden. Redan nu är det omöjligt att spara alla data och i framtiden är situationen än värre. Å andra sidan skulle det vara viktigt att ha åtkomst till alla data, så att de kunde användas förnuftigt.
- *Variation:* Data finns av olika form och också mängden olika datakällor växer hela tiden. Man har inte ännu lyckats standardisera datas lagringsformat på ett tillfredsställande sätt. Detta gör det svårare att analysera data och utnyttja dem. Skillnader i data uppstår också på grund av olika mätningsförfarande, mätningsenheter, noteringssätt för numeriskt data och till och med mätarnas kalibrering. Också språk- och kulturskillnader skapar huvudbry, i synnerhet för lagring av data i textformat. Olika format för visuella data kan förorsaka problem, likaså olika koordinatsystem i fråga om geodata (geografiska data).
- *Korrekthet:* Data som produceras är inte nödvändigtvis speciellt exakta. De kan också vara oklara eller direkt felaktiga. Ibland är datas mätnings- och uppsamlingsmetoder olika, eller de är okända, och därför är det inte uppenbart att data från olika källor är jämförbara. I värsta fall vet man inte om temperaturen är given i enheten Celsius eller Fahrenheit.

- *Relevans och värde:* Data är ofta realtidsdata, i vilket fall deras nytta är bunden till tidpunkten. Data kan vara av smärre intresse eller betydelse vid en viss tidpunkt eller i ett visst sammanhang, för att senare eller i ett annat sammanhang vara mycket intressanta och värdefulla. Å andra sidan kan data vara värdelösa för en viss aktör eller organisation, medan de kan vara guld värda för en annan. Ibland ligger datas värde, deras betydelse, i att de ger information om hur data förändras med tiden. Som exempel kan nämnas data som hänför sig till klimatförändringen, där de intressanta och betydelsefulla data är hur till exempel havsytans höjd förändras med tiden.

Dessa karaktäriseringar ger vid handen att det ligger mera bakom fenomenet big data än bara väldigt stora datamängder. För att de stora mängderna data ska vara till nytta bör de raffineras till information. Data blir till information när de analyseras och presenteras så att de har någon betydelse. Exempelvis kan en mobiloperatör känna till alla dess kunders mobiltelefoners realtida plats, men det blir information av data först när de används på ett lämpligt vis, till exempel genom att erbjuda en tjänst till mobiler enligt plats, så att tjänsten är annorlunda i stadens centrum än i förorterna. På motsvarande sätt gör man inget med observationsdata från en väderstation, innan de har analyserats till en väderrapport eller till historiska data som beskriver klimatförändringen. Mängden väderdata växer hela tiden och data kan idag samlas i olika form, med olika mätare och mätenheter än tidigare. Klimatdata (temperatur, gashalter, havsytans höjd osv.) som sträcker sig från våra dagar tillbaka hundratusentals år har samlats med olika tekniker och metoder, delvis genom beräkningar, och dessa data är utmanande att analysera. En utmaning för big data i material som dessa är hur man får ut det väsentliga ur data utan misstolkningar och utan felaktiga antaganden.

För att data skall kunna bli information behövs en massa olika hjälpmedel: komplicerad programvara, beräkningsmodeller och lösningar, och för att framställa dem olika slags kompetens: bland annat programmeringskunskap, förståelse för statistik och businessmedvetande. Dessutom bör det finnas tillräcklig information om själva data (hur, när och vem mätte med vilka tekniker).

Till användningen av big data hör också att många aktörer öppnar sina data för användning av vem som helst. Som exempel har Lantmäteriverket i Finland öppnat kartdata som de samlat, så att det är tillgängligt för alla, och dessa data används av många företag i olika tjänster baserade på kartor. På motsvarande sätt har Meteorologiska institutet i Finland öppnat väderradardata och de används till exempel i applikationer för väderinformation.

Exemplen visar hur ny business kan uppstå när data är öppna.

En viktig aspekt för användningen av big data är förtroende: data uppstår inte från ingenting och det finns risk också för missbruk. Data om temperaturen på olika platser används kanske inte så lätt fel, men om geodata för mobiltelefonerna kombineras med personlig information är redan möjligheten för missbruk större. Å andra sidan ger de stora datamassorna möjligheter att producera välstånd, tjänster och ny business. Således bör alla företag och organisationer som överväger att lägga ut data öppet göra det med besinning och avgöra på förhand vilka data som kan öppnas och med vilka spelregler.

Frågorna

Svara på följande frågor på basen av bakgrundsmaterialet ovan och dina allmänkunskaper. Frågornas sammanlagda maximala antal poäng är 25.

Fråga 1. Ge konkreta exempel med motiveringar på fem olika drag som karaktäriserar big data. Välj exemplen så att de avviker från varandra och inte är tagna ur bakgrundsmaterialet ovan. Ge alltså sammanlagt fem exempel. Ge 1-2 meningar med motiveringar per exempel. (maximalt antal poäng 5)

Fråga 2. Under senare tid har man diskuterat spårning i realtid av fordon i samband med fordonsskatten. För att fordonets färd skall kunna följas måste i praktiken en GPS-mottagare installeras i fordonet. GPS-mottagaren sänder kontinuerligt data om bilens plats. Om vi antar att ett sådant system tas i användning i Finland

a) till vilka andra nyttiga ändamål kunde data som samlas med detta system användas?

b) vilka etiska problem kan finnas i samlandet och utnyttjandet av dessa data?

c) vilka tekniska utmaningar kan finnas för att sätta upp och använda dessa system?

d) vilka utmaningar för att data skall bli information kan man möta då man anväder detta system?

Nämn två exempel per punkt (a, b, c, d). Motivera. (maximalt antal poäng 8)

Fråga 3. Ofta när man utnyttjar big data har man samlat data från olika källor för analys och produktion av information. Mätningssättet och lagringsformatet kan emellertid förorsaka problem då data har olika källor. Som exempel på detta kan nämnas lagring av data i olika format: det amerikanska sättet 04/11/2014 anger 11 april 2014, men för finländare är det 4 november.

Ge fyra exempel, som inte har getts i bakgrundsmaterialet och som avviker från varandra, på analys av big data, där problem kan uppstå på grund av att data har olika källor och mätningen, insamlingen och lagringsformatet av data varierar. Begrunda för varje problem också hur det kunde lösas. (maximalt antal poäng 12)