Tietojenkäsittelytieteen yhteisvalinta 26.5.2014

Tehtävä 1: Massadata (big data) ja sen käytön haasteet

Lue ensin seuraava taustamateriaali huolellisesti ja vastaa sen jälkeen siihen liittyviin kysymyksiin.

Maailmassa mitataan jatkuvasti monenlaisia asioita, kuten vaikkapa radiosignaalin laatua, ja tästä syntyviä mittausarvoja nimitetään dataksi. Myös sosiaalisen median päivitykset tai kauppaketjujen kassatiedot ovat dataa. Erilaista ja eri lähteistä peräisin olevaa dataa syntyy maailmassa koko ajan enemmän ja sen jakaminen hetkessä ympäri maailmaa on helppoa. Puhutaan massadatasta ("big data"), kun tarkoitetaan niin laajoja ja monimutkaisia datamääriä, että niiden analysointi ja hyödyntäminen on erittäin haastavaa ja vaatii uusien menetelmien kehittämistä. Tästä huolimatta massadatan käytölle on asetettu suuria odotuksia. Ajatellaan, että se tuottaa uusia liiketoimintamahdollisuuksia ja toisaalta antaa mahdollisuuden ratkoa sellaisia ongelmia, joita ennen ei ole pystytty ratkaisemaan.

Kun massadataa luonnehditaan, siihen liitetään usein seuraavia sanoja:

- *Volyymi:* Dataa syntyy jatkuvasti niin valtavia määriä, että sen käsitteleminen ja tallentaminen on hyvin haastavaa. Se myös saattaa sijaita fyysisesti useassa eri paikassa, mikä edelleen lisää haastetta datan käsittelyyn.

- *Vauhti:* Dataa syntyy myös jatkuvasti kiihtyvällä tahdilla. Jo tällä hetkellä on mahdotonta tallettaa kaikkea dataa ja tulevaisuudessa tilanne on aina vain pahempi. Toisaalta kuitenkin olisi tärkeää päästä käsiksi kaikkeen syntyvään dataan, jotta sitä voidaan hyödyntää älykkäästi.

- *Variaatio:* Dataa on monen erityyppistä ja erilaisten datalähteiden määrä kasvaa sekin koko ajan. Syntyvän datan standardoiminen käyttämään tiettyjä tallennusmuotoja ei ole toistaiseksi vielä toiminut tyydyttävällä tavalla. Tämä hankaloittaa datan analyysiä ja hyötykäyttöä. Eroja dataan syntyy myös muun muassa erilaisista käytetyistä mittaustavoista, mittayksiköistä, tavoista kirjata numeerisia tietoja ja jopa mittareiden kalibroinnista. Myös kieli- ja kulttuurierot varsinkin tekstimuotoisen datan tallentamisessa voivat aiheuttaa päänvaivaa. Visuaalisen datan eri esitysmuodot voivat nekin aiheuttaa ongelmia, samoin paikkatieto erilaisten koordinaattijärjestelmien takia.

- *Oikeellisuus:* Syntyvä data ei välttämättä ole kovin täsmällistä. Se voi myös olla epäselvää tai jopa virheellistä. Joskus datan keruu- ja mittausmenetelmät ovat erilaisia tai niitä ei tunneta, ja siksi eri lähteistä

tuotetun datan vertailukelpoisuus ei ole itsestään selvää. Pahimmassa tapauksessa esimerkiksi lämpötiladatasta ei tiedetä onko sen mittayksikkö Celsius vai Fahrenheit.

- *Oleellisuus ja arvo:* Data on usein reaaliaikaista, jolloin sen hyödyllisyyskin on sidottuna ajanhetkeen. Data ei välttämättä ole kovin kiinnostavaa tai oleellista tietyllä ajanhetkellä tai tietyssä asiayhteydessä, mutta esimerkiksi myöhempänä ajanhetkenä tai toisessa asiayhteydessä sama data saattaa olla hyvin kiinnostavaa ja arvokasta. Toisaalta yhdelle toimijalle tai organisaatiolle data saattaa olla arvotonta, kun taas toiselle se on kullanarvoista. Joskus datan arvo, merkitys, tulee nimenomaan siitä että se tarjoaa tietoa datan muutoksista ajan suhteen. Esimerkkinä tästä on ilmastotutkimukseen liittyvä data, jossa esimerkiksi meren pinnan korkeuden muutokset nimenomaan ajan suhteen ovat olennainen ja kiinnostava tieto.

Nämä luonnehdinnat kertovat, että massadata-ilmiössä on kyse muustakin kuin vain valtavan suurista datamääristä. Jotta suuresta määrästä dataa on hyötyä, se pitää jalostaa tiedoksi. Data muuttuu tiedoksi silloin, kun data jäsennetään ja esitetään niin, että sillä on jokin merkitys. Vaikkapa matkapuhelinoperaattorilla voi olla tiedossa kaikkien asiakkaidensa kännyköiden reaaliaikainen sijainti, mutta tämä data muuttuu tiedoksi vasta, kun sitä hyödynnetään sopivasti, esimerkiksi tarjoamalla tietyssä sijainnissa olevien kännyköiden kautta palvelua siten, että kaupungin keskustassa tarjotaan eri palveluja kuin esikaupunkialueella. Vastaavasti sääasemien havaintodatalla ei tee mitään, ennen kuin se on analysoitu sääennusteeksi tai ilmaston muuttumista kuvaavaksi historiatiedoksi. Säähän liittyvän datan määrä kasvaa jatkuvasti ja tänään tietoa voidaan kerätä eri muodossa, eri mittarein ja mittayksiköin kuin vuosia tai vuosikymmeniä sitten. Ilmastodata (lämpötilat, kaasupitoisuudet, meren pinnan korkeudet, jne.), joka ulottuu tästä päivästä taaksepäin aina satojen tuhansien vuosien päähän menneisyyteen, on kerätty eri tekniikoin ja menetelmin, osin laskennallisesti, ja se on erittäin haasteellista analyysin kannalta. Massadatan haasteena esimerkiksi tällaisessa aineistossa on se, miten tuloksista saadaan merkityksellistä ilman että alkuperäistä dataa tulkitaan väärin tai virheellisin oletuksin.

Jotta data voi muuttua tiedoksi, tarvitaan paljon erilaisia apuvälineitä: monimutkaisia ohjelmistoja, laskentamalleja ja ratkaisuja, ja niitä tekemään tarvitaan monenlaista eri osaamista: muun muassa ohjelmointitaitoa, tilastojen ymmärtämistä ja liiketoiminnan tajua. Lisäksi itse datasta pitää olla riittävästi tietoa (miten, missä ja kuka mittasi ja millä tekniikoilla).

Massadatan käyttöön liittyy usein myös se, että monet toimijat avaavat syntyvän datan avoimesti kenen tahansa käytettäväksi. Esimerkiksi Suomes-

sa Maanmittauslaitos on avannut keräämänsä karttadatan avoimesti kaikkien saataville, ja tätä dataa hyödyntävät useat yritykset erilaisissa karttapalveluissa. Samoin on avattu Suomen Ilmatieteen laitoksen säätutkien tuottama data, jota hyödynnetään esimerkiksi säätietoja kertovissa sovelluksissa. Esimerkeistä näkyy kuinka uutta liiketoimintaa voi syntyä, kun data on avointa.

Tärkeänä osa-alueena massadatan käytössä on luottamus: data ei synny tyhjästä ja väärinkäytönkin vaara on olemassa. Paikkakuntien lämpötilatietoja ei ehkä kovin helposti käytetä väärin, mutta jos kännyköiden sijaintitietoihin kytketään käyttäjän henkilökohtaista tietoa, onkin jo väärinkäytön mahdollisuus suurempi. Toisaalta suurista tietomassoista on kuitenkin mahdollisuus jalostaa hyvinvointia, palveluja ja uutta liiketoimintaa. Sen vuoksi jokaisen yrityksen tai organisaation, joka harkitsee datan avaamista julkiseksi, on tehtävä se harkiten ja etukäteen mietittävä, mitä dataa voidaan julkaista ja millaisilla pelisäännöillä.

Kysymykset

Vastaa seuraaviin kysymyksiin edellä olevan taustamateriaalin ja yleistietosi perusteella. Kysymyksien yhteenlaskettu maksimipistemäärä on 25 pistettä.

Kysymys 1. Kerro konkreettinen esimerkki viidestä eri massadataa luonnehtivasta piirteestä perusteluineen. Valitse esimerkit niin että ne poikkeavat toisistaan, eivätkä ole peräisin taustamateriaalista. Anna siis yhteensä 5 esimerkkiä. Anna 1-2 lausetta perusteluja yhtä esimerkkiä kohden. (maksimipistemäärä 5)

Kysymys 2. Viime aikoina on keskusteltu ajoneuvojen reaaliaikaisesta seurannasta autoverotuksen yhteydessä. Jotta ajoneuvon liikkumista voidaan seurata, siihen on käytännössä asennettava GPS-paikannin, joka lähettää jatkuvasti dataa auton sijainnista. Jos oletetaan, että Suomessa otetaan käyttöön tällainen järjestelmä, niin

a) mihin muuhun hyödylliseen voitaisiin sillä kerättyä tietoa käyttää,

b) mitä eettisiä ongelmia tiedon keräämisessä ja hyödyntämisessä voidaan nähdä,

c) mitä teknisiä haasteita järjestelmän luomisessa ja käytössä voidaan kohdata,

d) mitä haasteita datan muuttamisessa tiedoksi voidaan tätä järjestelmää käyttäessä kohdata?

Mainitse kaksi esimerkkiä kustakin kohdasta (a, b, c, d). Perustele. (maksimipistemäärä 8)

Kysymys 3. Usein massadataa hyödynnettäessä olemassa olevaa dataa kerätään eri lähteistä analyysiä ja tiedon tuottamista varten. Eri lähteistä kerättävän datan mittaustapa ja tallennusmuoto voi kuitenkin aiheuttaa ongelmia. Esimerkkinä tästä on vaikkapa päivämäärän tallentaminen eri muotoon: amerikkalainen tapa 04/11/2014 ilmaisee vuoden 2014 huhtikuun 11. päivää, kun taas suomalaiselle kyseessä on marraskuun 4. päivä.

Anna neljä taustamateriaalista ja toisistaan poikkeavaa esimerkkiä massadatan analyysistä, jossa voi syntyä ongelmia siihen liittyen, että tiedot ovat peräisin eri lähteistä ja datan mittaus-, keräys- ja tallennusmuodot vaihtelevat. Pohdi myös kunkin ongelman kohdalla, miten näitä ongelmia voitaisiin ratkoa. (maksimipistemäärä 12)