LAUREA
AMMATTIKORKEAKOULU
Yhdessä enemmän

Tiedonlouhintaprosessien soveltuvuus tietoliikenne-elementtien vikatiedostojen analysoimiseen

Sjöblom, Johanna

2016 Laurea

Laurea-ammattikorkeakoulu

Tiedonlouhintaprosessien soveltuvuus tietoliikenne-elementtien vikatiedostojen analysoimiseen

Sjöblom, Johanna
Tietojenkäsittelyn ko
Opinnäytetyö
Marraskuu, 2016

Laurea-ammattikorkeakoulu **Tiivistelmä**
Laurea Leppävaara
Tietojärjestelmäosaamisen ylempi AMK-tutkinto

Johanna Sjöblom

Tiedonlouhintaprosessien soveltuvuus tietoliikenne-elementtien vikatiedostojen analysoimiseen

Vuosi 2016 Sivumäärä 52

Tässä opinnäytetyössä tutkitaan tiedonlouhintamenetelmien soveltuvuutta laajojen vikatiedostojen analysointiin ja erityisesti työssä on keskitytty elementteihin, joista ei löydy testeissä vikaa. Opinnäytetyö on osa projektia, jossa on tutkittu syitä tietoliikenne-elementtien vikaantumiseen. Näistä yksiköistä aiheutuu turhia huolto- ja kuljetuskustannuksia. Työn tutkimuskysymys on ”miten tiedonlouhintaprosesseja sovelletaan tietoliikenne-elementtien vikatiedostojen analysoimiseen?”

Tiedonlouhinta tarkoittaa suurten tietojoukkojen kuten tietokantojen, dokumenttien tai mittaustulosten analyysia sekä mallien muodostamista. Työssä käsitellään tiedonlouhintaan liittyviä prosesseja ja keskitytään datan esikäsittelyvaiheisiin. Knowledge Discovery in Databases, KDD viitataan koko prosessiin, jossa etsitään hyödyllistä tietoa datasta ja tiedonlouhinta on oma prosessi KDD-prosessissa.

KDD-prosessi voidaan jakaa viiteen päävaiheeseen, jotka ovat tietämyksen vaatimukset, datan valinta, tiedonlouhinta, tulosten analysointi ja tietämyksen sisällyttäminen tietämyskantaan. Tiedonlouhinta voidaan erikseen jakaa myös viiteen osaan, jotka ovat datan pelkistäminen, menetelmän valinta, datan esikäsittely, datan valmistelu ja metodien käyttö.

Esikäsittelyvaiheen tavoite on mahdollistaa analyysimenetelmät, joilla erotetaan oikeat ja asiaankuuluvat tiedot datasta. Työssä käytetään tiedonlouhinnan esikäsittelytekniikoita. Reaalimaailman data vaatii yleensä puhdistamista, koska siinä on häiriöitä ja epätarkkuuksia.

Tutkimuksen aluksi selvitetään ongelman yleisyyttä muussa teollisuudessa ja määritellään keskeisistä käsitteistä. Tavoitteena oli tutkia teollisuudessa käytettyjä tiedonlouhintaprosesseja ja niissä saatuja testaustuloksia. Tietoliikenne-elementtien vika-analyysit ja tietoliikenneympäristö oli tuttua aikaisemmasta projektista. Tutkimuksen viitekehyksenä tutkittiin Hevnerin (2007) tietojärjestelmätutkimuksen kolmea sykliä. Suunnittelututkimus valittiin tutkimusmetodologiaksi, koska se soveltuu ongelman ratkaisemiseen ja sen arvioimiseen.

Tutkimuksen päätavoitteena oli tietämyksen lisääminen kentällä vikaantuneista tietoliikenneelementeistä, joista ei löydetä vikaa huollon testeissä. Tiedonlouhintaprosessit sopivat huollosta saatujen vikatiedostojen analysointiin, koska niiden avulla datasta pystytään selvittämään mitä tietoja on saatavilla ja myös datan puutteet tulevat selville. Tietojen esikäsittelytekniikoilla voidaan parantaa datan laatua, mikä osaltaan parantaa tarkkuutta ja tehokkuutta myöhemmissä louhintaprosesseissa.

Asiasanat: Tiedonlouhinta, Esikäsittely, Tietojärjestelmätutkimus, No Fault Found (NFF)

Laurea University of Applied Sciences **Abstract**
Laurea Leppävaara
Information Systems (Master's degree studies)

Johanna Sjöblom

The feasibility of data mining methods to analyze fault files of the telecommunication network elements

Year 2016 Pages 52

The purpose of this thesis was to determine the feasibility of data mining methods to analyze large fault file. In particular I concentrate on the no fault found elements. The thesis is a part of the project, in which has been studied the causes of failure of the telecommunication network elements. No fault found units cause unnecessary maintenance and transportation costs. The research question of the thesis is "how can data mining processes be applied to analyze the fault file of the telecommunication elements?"

Data mining refers to large data sets such as databases, documents, and measurement analysis, as well as the formation of models. The processes of data mining are covered and the pre-processing techniques are presented in more detail. Knowledge Discovery in Databases, KDD refers to the entire process of searching for useful information from the data. Data mining is own process which belong to the KDD process.

KDD process can be divided into five main phases which are the knowledge requirements, data selection, data mining, result analysis and the knowledge incorporation. Data mining can be separately divided into the five parts which are data reduction, data mining method selection, data pre-processing, data preparation and use of data mining methods. The main target of the pre-processing phase is to ensure that during the analysis stage, the required and relevant information are extracted from data. The data mining pre-processing techniques are used. Real-world data usually requires cleaning, as it is noisy and contains inaccuracies.

The research started with a literature review in order to get an understanding of the frequency of the no fault found problem in other industry areas and to get a knowledge of the essential concepts related to malfunction. The aim was to learn about former data mining processes which were used in industries, and test results that can be obtained from processes. Fault analysis of the telecommunications elements, and the environment of the telecommunication was familiar from the earlier project. Hevner (2007) "A Three Cycle View of Design Science Research" was used as the subtext of the thesis. The research methodology chosen was design science research because it is suitable for solving the problem and evaluating.

The main objective of this thesis was to increase knowledge about the no fault found elements of the telecommunication network. Data mining processes can be applied to analyze the fault file of the maintenance, because the data mining processes help to find out problem related information from the data and also the lack of the information of the data. Data pre-processing techniques can improve data quality, which improve the accuracy and efficiency of the following mining process.

Keywords: Data Mining, Knowledge Discovery in Databases Pre-processing, Design Science Research, No Fault Found (NFF)

Sisällys

1 Johdanto 6
1.1 Tietoliikenneverkot 7
1.2 Tukiaseman elementit 10
1.3 Ympäristön ja ongelman kuvaus 12
1.4 Opinnäytetyön rakenne ja yhteenveto 12
2 Kirjallisuustutkimus 13
3 Suunnittelututkimus 15
3.1 Tiedon keräys 21
3.2 Aineiston analyysi 21
3.3 Tutkimuksen triangulaatio 22
3.4 Menetelmän yhteenveto 23
4 Tiedonlouhinta 23
4.1 KDD-prosessi 25
4.2 Tiedonlouhintaprosessi 27
4.3 Datan esikäsittelyvaiheet 29
4.4 Poikkeavuuksien tunnistaminen (Anomaly Detection) 31
4.5 Klusterointi 32
5 Tutkimustulokset 34
5.1 Elementtien vikaantuminen 34
5.2 Tutkimustulosten tarkastelu 37
5.3 Tietojärjestelmien suunnittelututkimuksen syklit 38
5.4 Tutkimustulosten vertailua 39
6 Keskustelu 40
6.1 Johtopäätökset 40
6.2 Tutkimuksen luotettavuus ja validiteetti 41
6.3 Datan samanlaisuuden ja erilaisuuden mittaaminen 42
6.4 Jatkotutkimusaiheet 44
Lähteet 45
Kuviot 49
Taulukot 50
Liitteet 51

1 Johdanto

Matkapuhelinverkkoon kuuluvista tukiasemista, kerätään jatkuvasti paljon tietoa. Esimerkiksi päivittäin monitoroidaan tukiaseman soluja, jotta havaitaan niissä esiintyvät ongelmat. Laitteista kerätään esimerkiksi lokitiedostoja, hälytyksiä, tilanmuutoksia ja häiriöitä. Tietovarastot voivat sisältää miljoonia tietueita. Näiden hahmotus ei ole mahdollista manuaalisin keinoin, tähän tiedonlouhinta on yksi ratkaisu. Se on monitieteellinen tutkimusalue, jolle läheisiä tieteitä ovat tilastotiede, tietokannat, koneoppiminen, hahmontunnistus, tekoäly sekä visualisointi. Tiedonlouhinnalla tarkoitetaan suurten tietojoukkojen kuten tietokantojen, dokumenttien tai mittaustietojen analyysia sekä mallien muodostamista. (Nurminen 2003.)

Näiden tietojoukkojen analysoinnissa voidaan käyttää eri menetelmiä. Niitä ovat tietämyksen etsintä (Knowledge Discovery in Databases, KDD) ja tiedonlouhinta (Data Mining, DM). Poikkeamien havaitseminen (Anomaly Detection) ja klusterointi (Clustering) ovat tiedonlouhinnan perustehtäviä. Tässä työssä käytetään tietämyksen etsintä -menetelmästä yksinkertaisuuden vuoksi pelkästään lyhennettä KDD. Se voidaan määritellä aikaisemmin tuntemattoman ja mahdollisesti hyödyllistä tietoa sisältävän datan erotteluna. KDD-prosessin tarkoituksena on löytää hyödyllistä tietoa kerätystä datasta. KDD-prosessi koostuu useista eri vaiheista ja tiedonlouhinta on yksi prosessi KDD-prosessin sisällä. (Talonen 2015.)

Tiedonlouhinta jakaantuu myös useisiin eri vaiheisiin. Tiedonlouhinnassa etsitään ongelmiin ratkaisua analysoimalla olemassa olevaa dataa. Tiedonlouhintaprosessien avulla pyritään löytämään malleja ja hyödyllistä informaatiota suurista data-aineistoista. Poikkeamien havaitseminen on yksi tärkeimmistä tiedonlouhinnan tehtävistä ja se on myös tärkeä osa prosessin monitorointia useilla teollisuudenaloilla. Automaattinen poikkeamien ilmaisun sovellus voidaan kuvata välineenä, joka auttaa suodattamaan suurta osaa normaalia käyttäytymistä ja paljastaa poikkeavan käyttäytymisen loppukäyttäjälle tai järjestelmälle. (Hätönen 2009; Talonen 2015.)

Poikkeavuuksien havaitsemisen menetelmät on laajasti käytetty suorituskyvyn, vikojen ja turvallisuuden hallinnassa. Klusteronti on myös tiedonlouhinnan perusmenetelmä. Siinä pyritään jakamaan alkiot ryhmiin siten, että alkiot kussakin ryhmässä ovat keskenään mahdollisimman samanlaisia mutta eri ryhmissä alkiot taas olisivat mahdollisimman erilaisia keskenään. Muodostetut ryhmät ovat klustereita. Poikkeamien havaitseminen on esitetty tarkemmin kappaleessa 4.4 ja klusterointi 4.5. (Witten 2005; Talonen 2015.)

Tämä tutkimus on osa projektia, jossa on tutkittu tukiasemaelementtien vikaantumista. Projektin aluksi tutkittiin radiomoduuleiden vikaantumista, ja nyt tässä tutkimuksessa tarkasteltiin

keskusyksiköiden vikatiedostoja. Tutkimuksen data on kerätty todellisessa tietoliikenneverkossa toimivien tukiasemien vikaantuneista laitteista. Kentällä vikaantuneet laitteet toimitetaan huoltoon. Tutkimusmateriaalina ovat näiden huoltoon toimitettujen elementtien vikaraportit ja lisäksi on saatu elementeistä kerätyt raakahälytiedostot. Ongelmana on, että vikaantuneista elementeistä ei kuitenkaan aina ei löydetä vikaa huollon testeissä. Näistä aiheutuu ylimääräisiä kuljetus- ja huoltokustannuksia.

Työn tutkimuskysymyksenä on selvittää ”miten tiedonlouhintaprosessit soveltuvat tietoliikenne-elementtien vikatiedostojen analysoimiseen”. Ja erityisesti näiden menetelmien soveltuvuudesta ”ei vikaa löydy” -elementtien löytämiseksi. Lisäksi tutkittiin tiedonlouhinnan integroimista Hevnerin (2007, 2) suunnittelutieteen sykleihin. Tiedonlouhintaprosesseja on kuvattu kappaleessa 4.2 ja Hevner (2007, 2) “A Three Cycle View of Design Science Research” on esitetty kappaleessa 3.

Lopuksi työssä verrattiin tuloksia muiden teollisuusalojen vastaaviin ”ei vikaa löydy” -tutkimuksiin. Beniaminy & Joseph (2002) tutkivat lentokoneiden huoltoon palautettuja laitteita, jotka saivat huollossa ”ei vikaa löydy” määrityksen. Block, Tyrberg & Söderholm (2009) tutkivat sotilaslentokoneiden vastaavia laitteita ja Jones & Hayes (2001) tutkivat elektroniikkateollisuuden ”ei vikaa löydy” -laitteita tietokoneista sotilaslaitteistoon (katso esim. " Beniaminy, I., Joseph, D. 2002. "Reducing the “No Fault Found” Problem: Contributions from Expert-System Methods"; " Block, J., Tyrberg, T., Söderholm, P. 2009. "No Fault Found Events During the Operational Life of Military Aircraft Items”; Jones, J., Hayes, J. 2001. "Investigation of the Occurrence of: No-Faults-Found in Electronic Equipment"; James, I., Lumbard, D., Ian, W., Goble, J. 2003 "Investigating No Fault Found in the Aerospace Industry"; Santoro, M. 2008. "New Methodologies for eliminating no trouble found, no fault found and other non repeatable failures in depot settings). Näistä tarkemmin on käsitelty Beniaminy & Joseph (2002), Block, Tyrberg & Söderholm (2009) ja Jones & Hayes (2001).

Tässä tutkimuksessa käsitteellä ”tieto” ymmärretään olevan eri merkityksiä. Lyhyesti kuvattuna datalla tarkoitetaan numeerista arvoa, joka ei yksin pelkkänä numerona kerro mitään erityistä. Se tarvitsee kontekstin, mikä selittää luvun tarkoitusta. Sen jälkeen siitä tulee tietoa ja tietämys tarkoittaa ymmärrystä tiedon merkityksestä. (Zins 2007.)

1.1 Tietoliikenneverkot

Tietoliikenneverkoista on tullut väistämätön osa jokapäiväistä elämää. Matkaviestinverkon palveluilta odotetaan korkeatasoista luotettavuutta ja laatua. Tietoliikenneverkon hallinta ja toi-

minta ovat avaimet verkon luotettavuuteen ja laatuun. Verkon hallinnan tarkoituksena on optimoida tietoliikenneverkon toimintakykyä. Tämä sisältää muun muassa verkkokäyttöjärjestelmien pitämisen huippukunnossa, tiedottamisen operaattorille verkon heikkenemisestä ja välineet, joilla löydetään syitä suorituskyvyn heikkenemiseen. (Kumpulainen 2014.)

Matkapuhelinverkko on yksi osa tietoliikenneverkkoa, ja sen valvottavaan kokonaisuuteen voi kuulua tuhansia tukiasemia. Tietoliikenneverkot tarjoavat monella eri tekniikalla toteutettuja palveluja. Niistä kerätään jatkuvasti valtavia määriä käyttötietoja. Verkonhaltijoiden yksi tärkeimmistä sovelluksista on havaita poikkeavuuksia kerätyistä tiedoista. (Kumpulainen 2014; Hätönen 2009.)

Tietoliikenneverkon tarjoamien palvelujen laatua ja käyttöä valvotaan. Operaattorit käyttävät valvontadataa muun muassa laskutuksessa, ylläpidossa ja suunnittelussa. Valvontajärjestelmän tietokantaan kerätään yhdestä verkkoelementistä jopa satoja aikasarjoja sekä useita tuhansia hälytyksiä päivittäin. Hätönen (2009) toteaa väitöskirjassaan, että täytyy ymmärtää miten tietoliikenneverkot toimivat ja toisaalta myös miten ne on rakennettu, jotta voidaan ymmärtää tiedonlouhinnan vaatimuksia tietoliikenneverkkojen monitoroinnissa.

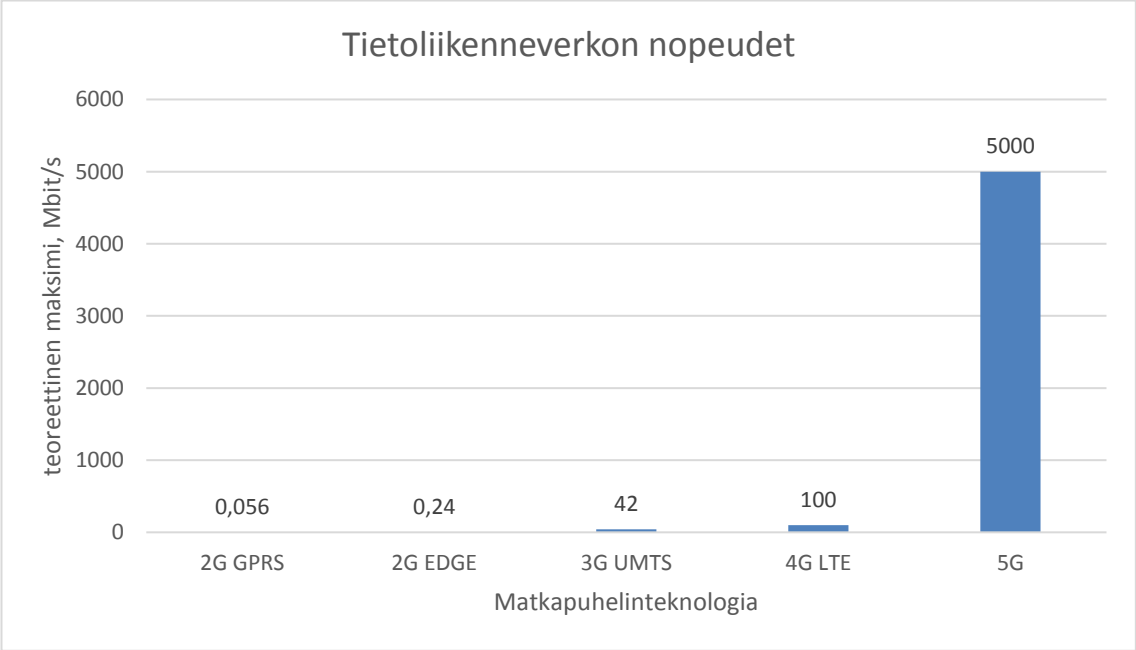
Tietoliikenneverkoissa on vikoja ja poikkeavuuksia perinteisesti etsitty vianhallinnan tai suoritusarvojen optimointijärjestelmän avulla. Havaitut virheet sekä niiden vakavuusaste näkyy verkon ylläpitäjälle, joka voi niiden perusteella päättää tarvittavat korjaustoimenpiteet. Tietoliikenneverkoissa voi olla tuhansia verkkoelementtejä, joista jokainen ilmoittaa verkon valvontaan erilaisia ilmoituksia. Tuhansien hälytysten analysointi on haastavaa verkon ylläpidolle, ja tärkeitä hälytyksiä voi kadota inhimillisten virheiden tai verkon ruuhkautumisen takia. (Jerome 2015.)

Radioyhteys muodostetaan tilaajan ja tukiaseman välille, ja yhteydet muodostetaan tukiasemien ja niiden solujen avulla. Digitaaliset tietoliikenneverkot ovat erittäin monimutkaisia systeemejä ja niiden suunnittelu, hallinta ja optimointi eivät ole triviaaleja tehtäviä. Monimutkaiseen systeemiin kuuluu suuri määrä elementtejä, joita ovat esimerkiksi radioverkko-ohjaimet, lähetys- ja vastaanottoasemat. Monimutkaisen verkon perusyksiköt ovat tukiasemat, joissa olevat antennit osoittavat radiopeiton alueen, niitä kutsutaan soluiksi. (Vehviläinen, Hätönen, Kumpulainen 2003; Jerome, 2015.)

Tukiasemat toimivat tietoliikenneverkoissa, joissa on eri matkapuhelinteknologiaa. Nykyään on käytössä useita mobiiliverkkoteknologioita GSM (Global System for Mobile communication) eli 2G toimii 900 MHz:n tai 1800 MHz:n taajuusalueella. GSM:n täysin digitalisoitu verkko tukee tavallisten puheluiden lisäksi muun muassa datapuheluita ja tekstiviestejä. UMTS (Universal Mobile Telecommunications System) eli 3G toimii 900 MHz:n ja 2100 MHz:n taajuusalueilla.

2G:n ja 3G:n suurin ero on se, että 3G suunniteltiin myös datasiirtoon ja sen tiedonsiirto on huomattavasti nopeampi. LTE (Long-Term Evolution) eli 4G toimii taajuusalueilla 800 MHz:n, 1800 MHz:n ja 2600 MHz:n. Alun perin 4G oli tarkoitettu vain datan siirtoon, nykyään 4G-verkossa on tuki myös puheluille VoLTE (Voice over LTE). (Dahlman, Parkvall & Sköld 2011; Järvinen 2013; Palat & Godin 2009.)

Tietoliikenneverkkojen kehityksessä on ollut monta vaihetta ja niiden kehitys on nopeuttanut jatkuvasti datan siirtonopeuksia tietoliikenneverkoissa. Kuviossa 1 on esitetty tietoliikenneverkon nopeuksien kasvua teknologian kehittyessä.



Kuvio 1: Tietoliikenneverkon nopeuksien kehitys.

Kuviossa 1 tietoliikenneverkon nopeuksien yksikkönä on Mbit/s, ja nopeudet on esitetty verkon teoreettisina maksimeina. Nopeudet ovat kasvaneet suhteellisen tasaisena ennen 5G-teknologiaa. 5G-verkon nopeudet ovat alustavista testeistä, jotka tehtiin Elisan matkapuhelinverkossa. Noin vuonna 2020 aloittavan 5G-verkon kapasiteetin on laskettu olevan nykyistä 4G-verkkoa 100 kertaa nopeampi. (Elisa testasi 5G:ta ensimmäisenä operaattorina Suomessa 2016). Kuviossa 1 on käytetty tietoja nopeuden teorettisista maksimiarvoista Elisan asiakaspalvelusivuilta. (Elisan matkapuhelin- ja mobiililaajakaistaverkon nopeudet.)

1.2 Tukiaseman elementit

Keskusyksikkö, radio, antenni, siirtolaite ja näiden väliset kaapelit muodostavat yksinkertaisimmillaan tietoliikenneverkon tukiaseman. Keskusyksikkö ja radio eli RF-yksikkö (Radio Frequency) ovat tukiaseman aktiivilaitteita, jotka hallitsevat antennipiirin toimintaa ja ohjaavat signaalit oikeisiin portteihin. Tukiaseman ydin on keskusyksikkö. Siihen kytketään RF-yksiköt, siirtolaitteet, sähkönsyöttö ja ulkoiset hälytykset. Passiivisia laitteita tukiasemassa ovat diplekseri ja triplekseri. Diplekseri vastaanottaa kahden eri taajuusalueen signaalin ja yhdistää ne. Triplekseri tekee saman kolmelle taajuudelle. (Määttänen 2015.)

Antennit lähettävät ja vastaanottavat radioaaltoja, ne voidaan jakaa suuntaaviin antenneihin ja ympärisäteileviin antenneihin. Ympärisäteilevissä antenneissa signaalin lähetys- ja vastaanottosuunta muuttuvat jatkuvasti. Suuntaavia antenneja käytetään esimerkiksi linkkiyhteyksissä, joita tarvitaan tukiasemien välillä. Lisäksi tukiaseman toiminta vaatii useita eri kaapeleita, kuten valokuitu-, koaksiaali- ja kuparikaapelia. Tukiaseman, radion ja antennien välissä käytetään valokuitu- ja koaksiaalikaapeleita, ja laitteiden sähköistämiseen ja maadoittamiseen kuparikaapelia. Tukiasema ei toimi jos sitä ei ole yhdistetty runkoverkkoon. Tukiaseman siirtoyhteys eli transmission toteutustapa voidaan toteuttaa monilla tavoilla, kuten radiolinkeillä, toiselta tukiasemalta ketjuttamalla tai kytkimillä. Siirtoyhteyden avulla tukiasema yhdistetään runkoverkkoon. (Määttänen 2015.)

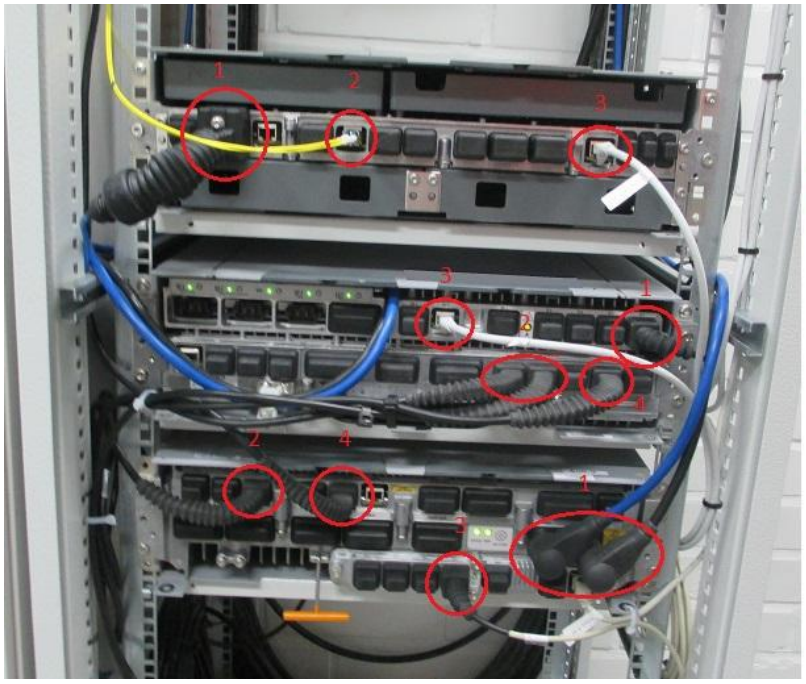
Tukiasemia voidaan asentaa kiinteistöön tai maastoon. Harvaan asutuilla seuduilla käytetään enemmän tukiasemamastoja. Mastoon asennetut laitteet tarvitsee myös maadoittaa. Kiinteistöihin asennetaan tukiasemia, jotta alueen ulkokuuluvuus paranisi tai taataan kuuluvuus kiinteistön sisätiloissa. Kuviossa 2 on kuvattu radioiden mastokiinnitys. (Määttänen 2015.)

Kuvio 2: Putkimastoon kiinnitettyjä radioita. (Määttänen 2015)

Kuviossa 2 on esimerkki mastoon asennetuista tukiaseman radioista ja muista elementeistä. Ne pyritään asentamaan lähekkäin, jotta vältytään tuulikuormalta. Maksimissaan neljä RF-yksikköä voidaan kiinnittää yhteen telineeseen. RF-kaapelit on tärkeää kytkeä antennissa oikean taajuusalueen portteihin, koska jokainen radion tx/rx- ja rx-diversiteettiporteista on kytketty antennin sisäiseen elementtiin, ja se on riippuvainen tietystä taajuusalueesta. Tx/rx-portti tarkoittaa lähettävää ja vastaanottavaa porttia ja rx-diversiteettiportti tukee kahta erillistä tulkintaa vastaanotettavasta signaalista, jolloin häiriöiden vaikutus vähenee. (Määttänen 2015.)

Antenniverkkoon kuuluu yleensä kolme tai neljä antennipiiri. Antennipiirin rakentaminen aloitetaan laitetelineen kalustamisella ja muiden komponenttien asentamisesta. Se päättyy tukiaseman toiminnan testaamiseen. Tukiaseman rakentamisen valmistuttua tehdään tukiaseman komissiointi eli käyttöönotto. Tässä yhteydessä määritellään tukiaseman konfiguraatio ja parametrit. Määritettäviä parametreja on huomattava määrä, ja tukiaseman toiminnan voi estää pienikin virhe. Soluliitokset eli tukiaseman solujen liittäminen oikeisiin radioihin ja niiden portteihin, sekä solun teho ovat tärkeitä parametreja. Rx-diversiteetin käyttö ja tukiaseman käyttämät IP-osoitteet valitaan komissioinnissa. (Määttänen 2015.)

Kuviossa 3 on esimerkki tukiaseman keskusyksikön monimutkaisista kytkennöistä ja eri matkapuhelinteknologioilla toimivista elementeistä. Antennipiirin vaatimat kytkennät vaihtelevat asennettujen laitteiden mukaan. Sähkönsyöttö, siirtoyhteys eli transmissio sekä RF-yksikkö ovat kytkettynä jokaiseen keskusyksikköön.



Kuvio 3: Keskusyksikön kytkennät ylhäältä alas 4G, 3G sekä 2G laitteet. (Määttänen 2015)

Kuviossa 3 ylimpänä on 4G keskusyksikkö, joka on kytketty yhteen RF-yksikköön. Alemmat keskusyksiköt 2G ja 3G jakavat yhden RF-yksikön ja 3G:lle on lisäksi kytketty myös toinen RF-yksikkö, joka palvelee suuremmalla taajuusalueella. Sähkökaapeli ja valokuitu kytketään RF-yksikön vastaavaan porttiin. Kuviossa 3 DC-in portit on merkitty numerolla 1, niihin kytketään sähkönsyöttö. RF-yksikön kytkennät on ympäröity ja osoitetaan numerolla 2, numerolla 3 on merkitty lähetysyksikkö. RF-sharing tarkoittaa saman radion käyttämistä kahdella eri tekniikalla, esimerkiksi 900 MHz:n taajuudella sekä 2G että 3G. Jaettua radiota kutsutaan multiradioksi. (Määttänen 2015.)

1.3 Ympäristön ja ongelman kuvaus

Tukiasemat sijaitsevat usein hyvin hankalissa maastoissa ja ne voivat olla kaukana kaupungeista. Laitteen vikaantuessa niitä yritetään korjata ensin paikanpäällä. Jos laitetta ei saada toimimaan, se lähetetään huoltoon. Huoltopisteet on keskitetty, joten se voivat sijaita kaukana. Laitteen tullessa huoltoon siitä selvitetään muun muassa laitteeseen ladatut ohjelmistot, mahdolliset vikailmoitukset, radioteknologia ja laitteen päälläoloaika. Tiedot ja asiakkaan ilmoittamat viat tallennetaan työkaluun. Näitä tietoja käytetään hyväksi laitteen analyysissä, jossa selvitetään laitteen vikaa ja sen syytä.

Ongelmana on että kaikista käytössä vikaantuneesta laitteesta ei huollon testeissä löydy syytä vikaantumiselle, tällöin laite saa testien jälkeen merkinnän ”vikaa ei löydy”. Tutkimuksen tarkoituksena on lisätä tietämystä näistä ”vikaa ei löydy”-vikamäärityksen saaneista laitteista. Tavoitteena on ”No Fault Found” eli NFF-elementtien löytäminen aikaisemmassa vaiheessa.

Kirjallisuudessa löytyy ilmiölle ”vikaa ei löydy” useita eri nimityksiä, esimerkiksi ”No Fault Found” (NFF), ”No Defect Found” (NDF), ”No Hardware Defect” (NHD), ”No Problem Found” (NPF), ”Retest-OK” (RETOK) tai ”Could Not Duplicate” (CND). Tässä tutkimuksessa käytetään lyhennettä NFF, kun viitataan “vikaa ei löydy”. (Beniaminy & Joseph 2002.)

1.4 Opinnäytetyön rakenne ja yhteenveto

Tutkimusaiheeseen liittyvä kirjallisuustutkimus käydään läpi seuraavassa kappaleessa. Kappaleessa käsitellään opinnäytetyön teoreettisena viitekehyksenä toimivan tietojärjestelmien suunnittelututkimuksen viitekehyksen kehitystä. Lisäksi käydään läpi tutkimustulosten analysointia varten tiedonlouhintaprosessit. Työssä perehdyttiin NFF yksiköiden tutkimiseen, sen

vuoksi NFF-ilmiö on esitelty eri teollisuuden näkökulmista. Kolmannessa kappaleessa on käsitelty suunnittelututkimukseen liittyvä metodologia ja tarkastellaan tietojärjestelmien suunnittelututkimuksen syklejä. Tiedonlouhinta ja siihen liittyvät esikäsittelyprosessit on esitetty kappaleessa neljä. Tutkimustulokset elementtien vikaantumisesta on käsitelty kappaleessa viisi ja johtopäätökset on esitetty lopuksi keskusteluosiossa.

2 Kirjallisuustutkimus

Kirjallisuudesta löytyy hyvin paljon artikkeleita tietojärjestelmätutkimuksen viitekehyksen kehityksestä. Hyvin usein on artikkeleissa viitattu tietojärjestelmätutkimuksen viitekehysten malleihin Nunamaker, Chen ja Purdin (1991), March & Smith (1995), Hevner, March, Park ja Ram (2004). Luonnollisen ja suunnittelutieteen eroja on pohtinut Simon (1996) kirjassaan ”The sciences of the artificial”, se on alun perin ilmestynyt 1969 ja siihen on viitattu useissa myöhemmissä artikkeleissa. Nunamakerin ym. (1991) kaaviossa esitetään, että systeemin kehitys liittyy läheisesti teorian kehittämiseen. March & Smith (1995) jakoivat tutkimuksen luonnontieteelliseen ja suunnittelutieteelliseen. Hevnerin ym. (2004) viitekehysmallissa ympäristö määrittelee ongelma-alueen ja se muodostuu ihmisistä, organisaatioista ja olemassa olevista tai suunnitelluista teknologioista. Myöhemmin Hevner (2007) lisäsi tietojärjestelmätutkimuksen viitekehykseen kolmen syklin näkemyksen. Tutkimusmetologia on käsitelty kappaleessa kolme.

Tiedonlouhinnasta kertovia kirjoja löytyy paljon ja aiheesta on myös kirjoitettu paljon artikkeleita. Alue on kuitenkin hyvin laaja ja erilaisia menetelmiä on useita. Sen vuoksi aihepiirin rajaaminen on tärkeää. Menetelmien avulla havaitaan asioiden välisiä yhteyksiä kuten toimintamalleja tai poikkeamia niistä. Käsittelen tarkemmin kappaleessa 4 tiedonlouhintamenetelmiä, joita käytetään tietoliikenneverkkojen tutkimuksessa. Näitä ovat KDD, tiedonlouhinta, poikkeamien havaitseminen ja klusterointi.

Useissa julkaisuissa on käsitelty poikkeamien havaitsemista tietoliikenneverkosta. Poikkeamien havaitsemista tietoliikenneverkon monitoroinnissa on paljon tutkittu (Gajic, Novaczki, S. & Mwanje 2015; Kumpulainen 2014; Kumpulainen & Hätönen 2008; Vehviläinen ym. 2003). Kumpulainen (2014) käsittelee väitöskirjassaan poikkeamien havaitsemista tietoliikenneverkon toiminnasta. Poikkeamien tunnistaminen auttaa verkon hallinnassa ja kehittämisessä. Viallisista tukiasemista johtuva poikkeuksellisen runsas puheluiden katkeilu tai dataliikenteen hidastuminen voivat näkyä verkon epätavanomaisena toimintana. Hän kehitti menetelmän, jonka avulla voidaan löytää ja esittää oleellinen tieto suuresta määrästä mittaustietoa. Lisäksi verkon hallintaa ja kehittämistä tuetaan, kun poikkeamat voidaan havaita ja analysoida.

Datamäärien keräystä tietoliikenneverkoista kuvaavat Hätönen (2009), Kumpulainen (2014) ja Jerome (2015). Hätönen (2009) käsittelee suurten tietomäärien analysointia tiedonlouhintamenetelmillä. Keskikokoinen tietoliikenneverkko voi tuottaa päivässä useita tuhansia hälytyksiä ja kymmeniä gigatavuja lokia sekä suorituskykytietoja. Tietoliikenneoperaattorien on mahdotonta manuaalisesti analysoida kaikkia tietoja. Hätönen (2009) esittelee väitöskirjassaan kaksi menetelmää, jotka on tarkoitettu suurten lokitietokantojen analysointiin. Jerome (2015) käsittelee diplomityössään tietoliikenneverkossa poikkeamien havaitsemisen esikäsittelytekniikoita. Työssä perehdytään tunnistamaan oikeanlaisen esikäsittelyn askeleet, joita voidaan käyttää reaaliaikaisen tietoliikenneverkon mittaamiseen.

NFF:ään liittyviä tapauksia ei ole tutkittu juuri elektroniikkateollisuudessa vaan enemmän lentokoneteollisuudessa. Beniaminy & Joseph (2002) tutkivat NFF-laitteiden aiheuttamia turhia kustannuksia lentokoneteollisuudessa. Heidän tutkimusten mukaan 4500 NFF-tapausta maksoi Air Transport Association (ATA) ilmailujäsenille 100 miljoonaa dollaria vuosittain ja samalla ne aiheuttivat tuhansien lentojen peruutuksia ja viivästyksiä. Artikkelissa on esitetty erityyppisiä NFF-vikoja. Niitä ovat muuan muassa poistettu yksikkö, joka on oikeasti viallinen, mutta uudelleen testauksessa siitä ei löydetä vikaa. Yksikkö palautetaan varastoon, ja kun osa otetaan uudelleen käyttöön, siinä ilmenee samoja vikaoireita kuin aikaisemmin. Viat esiintyvät epäsäännöllisesti, laitteet voivat olla oikeasti viallisia mutta esimerkiksi laboratorioiden testeissä on vaikea saada lentotilannetta vastaavat olosuhteet. (Beniaminy & Joseph 2002.)

Syitä miksi vikoja ei löydetä, testauksessa ovat: testit eivät ole riittävän kattavia, viat esiintyvät vain esimerkiksi lento-olosuhteissa, ja laboratorioissa on vaikea simuloida nopeasti muuttuvia ja dynaamisia olosuhteita. On myös vaikea määritellä testausmenetelmiä, jotka tarkistavat kaikki testatun laitteen toiminnot, niin että mahdolliset kyseisen laitteen vian voidaan havaita. Yhtä vaikeaa on arvioida testausmenetelmien todellista kattavuutta. Vaikeudet johtuvat tuotteiden lisääntyvästä monimutkaisuudesta, sekoituksesta toisistaan riippuvaisista laitteisto- ja ohjelmistoteknologioista. Näitä ovat laitteiston puolella analogiset, digitaaliset ja radiot ja ohjelmiston puolella reaaliaikaisuus ja viestintä. Lisäksi voi olla puutteellista tiedon siirtoa laitteen ja testauksen suunnittelijoiden välillä. (Beniaminy & Joseph 2002.)

Beniaminyn & Josephin artikkelin (2002) mukaan NFF-ongelman kanssa kamppailua ei voida tehdä tehokkaasti, jos keskitytään vain yhteen näkökulmaan. NFF:ien aiheuttamien kustannuksien alentamiseen voidaan päästä vain useiden parannusten avulla monilla eri alueilla. He käsittelevät NFF-ongelman ratkaisua asiantuntijaohjelmistojen avulla, joilla saavutettaisiin esimerkiksi testikattavuuden ja reaaliaikaisen tiedon analysointi.

Block ym. (2009) tutkivat Ruotsin Ilmavoimien hävittäjien vikaantuneiden laitteiden NFF-ilmiötä. Heillä oli tutkimusmateriaalia vuosilta 1977—2006, ja se sisälsi 330 lentokonetta ja 605

000 lentotuntia. Materiaalissa oli mukana koneiden ylläpitotiedot, jotka sisälsivät esimerkiksi varastointiajat, asennukset, viat sekä korjaukset. Koneissa lennon aikana esiintyneistä vioista suurimmasta osasta ei löydetä huollossa vikaa, jolloin ne saavat NFF-merkinnän. Tutkimuksessa verrattiin Ruotsin Ilmavoimien tutkan neljää eri osaa eri testausvaiheessa. He havaitsivat, että NFF:iä esiintyy enemmän tietyissä tuotteissa ja lisäksi niihin vaikuttaa korjauskäyntien määrä. Yhteenvedossa esitettiin että useimmat NFF-tapaukset voitiin jäljittää operaation aikana huomioituihin vikoihin. Tutkimustulosten analysointiin Block ym. (2009) käyttivät standardeja ohjelmistotyökaluja, kuten MATLAB ja Microsoft Excel.

Jones & Hayes (2001) tutkivat NFF-ongelmaa elektroniikkateollisuudessa. Tutkimusaineistoa kerättiin noin kymmenen vuotta ja se sisälsi hyvin laaja-alaisesti erilaista materiaalia tietokoneista armeijan laitteistoon. Tutkituista vikaantuneista komponenteista noin 40 prosentille ei löydetty vian syytä. NFF:ien aiheutuminen voi johtua piirilevyllä olevista useita monimutkaisia tekijöitä. Tällaisia ovat esimerkiksi mikroprosessorit, suuret muistit ja mikropiirit. On epäilty, että liittimien suuri määrä voisi aiheuttaa NFF esiintymisen. Liittimissä voi olla korroosiotuotteita, jotka aiheuttavat epäsäännöllistä toimintaa laitteelle. Huollossa liittimet kuitenkin yleensä puhdistetaan rutiinin omaisesti, jolloin laite voi testeissä toimia moitteettomasti. Jones & Hayes (2001) esittää myös että inhimilliset virheet tai kentällä saatu virheilmoitusten laatu ovat syitä, joiden vuoksi laitteista ei löydetä huollossa vikaa vaan ne saavat NFF-tilan. Tutkimuksessaan he eivät löytäneet yhteyttä laitteen käytölle tai laitteen monimutkaisuudelle ja NFF:ien esiintymiselle.

Näissä NFF-ongelmaa käsittelevissä artikkeleissa (Beniaminy & Joseph 2002; Block ym. 2009; Jones & Hayes 2001) on esitetty useita syitä, siihen miksi laite vikaantuu kentällä mutta samat viat eivät tule esille korjaamon testeissä. Tutkimuksissa todetaan että NFF-ongelmaa on vaikea ratkaista sen monimutkaisuuden vuoksi. Lisäksi todetaan, että NFF-ongelmaan ei ole yhtä ratkaisua vaan se voidaan saavuttaa useiden eri alueiden parannusten kautta. Parannuksia pitäisi hakea lisäämällä esimerkiksi automaattisia testausjärjestelmiä, joita voidaan integroida testausjärjestelmiin tai lentokoneen järjestelmiin. Näillä saataisiin parempi kuva koko järjestelmästä, jolloin vikakuvaukset olisivat tarkempia.

3 Suunnitteletutkimus

Tässä kappaleessa kuvataan tutkimuksessa käytetty tutkimusmetodologia. Suunnittelutieteellistä tutkimusta voidaan pitää prosessina, jolla ratkaistaan ongelma. Suunnittelutieteen tavoitteena on tuottaa uutta suunnittelutietämystä. Suunnittelutietämyksen voidaan katsoa käsittävän kohteen, toteutuksen ja prosessin suunnittelun. Tämän vuoksi olen valinnut tutkimuksen

menetelmäksi suunnittelututkimuksen. Suunnittelutietämyksen kohteessa artifakti suunnitellaan, toteutuksessa laaditaan suunnitelma artifaktin toteuttamiseksi ja prosessissa ammattilainen laatii oman suunnitelman ongelman ratkaisemiseksi. (Nunamaker ym. 1991; Gregor & Jones 2007; March & Smith, 1995.)

Tutkimuksen tavoitteena on tukiaseman vikaantuneiden laitteiden vikatietojen yhdistämisen ja analysoimisen kautta suunnitella mallia, jonka avulla vioista saataisiin enemmän tietoa jo aikaisemmassa vaiheessa. Nunamaker ym. (1991) mukaan tutkimusaihe voidaan jakaa kolme osaan: alkuperäiset kysymykset eli se mitä halutaan tietää, tutkimuksen perustelu eli miksi se halutaan tietää ja täsmentävät kysymykset eli mitä kysymyksiä pitää tutkia jotta saadaan vastaukset alkuperäisiin kysymyksiin. Suunnittelututkimuksella pyritään saamaan aikaan pysyvä muutos tai parannus systeemissä.

Seuraavaksi esitellään lyhyesti tietojärjestelmätutkimuksen mallin kehitystä. Nunamaker ym. (1991) esittelivät artikkelissaan idean käyttää systeemin kehitysprosessia tietojärjestelmätutkimuksen tutkimusprosessissa. Artikkelissa on esitetty kaavio, jossa on systeemin kehityksen yleinen luonne tutkimuksen elinkaaressa. Kaaviossa systeemin kehitys liittyy läheisesti teorian kehittämiseen, havaintoihin sekä kokeelliseen tutkimukseen. Jotta saadaan täydellinen ymmärrys monimutkaisesta tutkimusalueesta, tutkimuksen monimetodologinen lähestymistapa on tehokkain toimintasuunnitelma. (Nunamaker & Briggs 2011.) Simonin (1996) mukaan tieteen keinotekoinen ilmiö on aina välittömässä vaarassa hävitä ja kadota. Van Aken (2004) mukaan sekä rakentaminen että parantaminen käyttävät samanlaista lähestymistapaa ja tuottavat samanlaisen tuloksen, jota kutsutaan teknologiseksi säännöksi.

Nunamakerista pidemmälle tutkimuksen viitekehystä kehittivät March & Smith (1995). He jakoivat tutkimuksen kahteen osaan luonnontieteelliseen ja suunnittelutieteeseen. Luonnontieteellinen tutkimus sisältää fyysisen, biologisen ja sosiaalisen perinteisen tutkimuksen sekä käyttäytymistieteen. Näiden tieteiden tutkimuksella on tarkoituksena ymmärtää todellisuutta. Kun taas suunnittelutiede yrittää kehittää tuotteita ihmisen tarkoituksia varten ja se on teknologiapainotteinen. Sen tuotteet arvioidaan arvon ja hyödyllisyyden mittareiden kautta. Esimerkiksi toimiiko se? Tai parantaako se toimivuutta?

March & Smith (1995) kehittivät tutkimuksen viitekehyksen luokittelun, joka esitetään taulukossa 1. Siinä on jaoteltu suunnittelutieteen toiminta rakentamiseen ja arvioimiseen. Luonnontieteen toiminta jaotellaan teorian luontiin ja testaukseen. Tutkimussuoritteet ovat molemmissa toiminnoissa käsitteet, mallit, menetelmät ja toteutukset. Nämä tutkimussuoritteet voidaan selittää, siten että tutkimuksen sanasto muodostaa käsitteet, joiden välisiä suhteita kuvataan malleilla.

Tutkimus-suoritteet	Suunnittelutiede		Luonnontiede	
	Rakentaa	Arvioida	Luoda teoriaa	Testata teoriaa
Käsitteet	x	x		
Mallit	x	x		
Menetelmät				
Toteutukset				

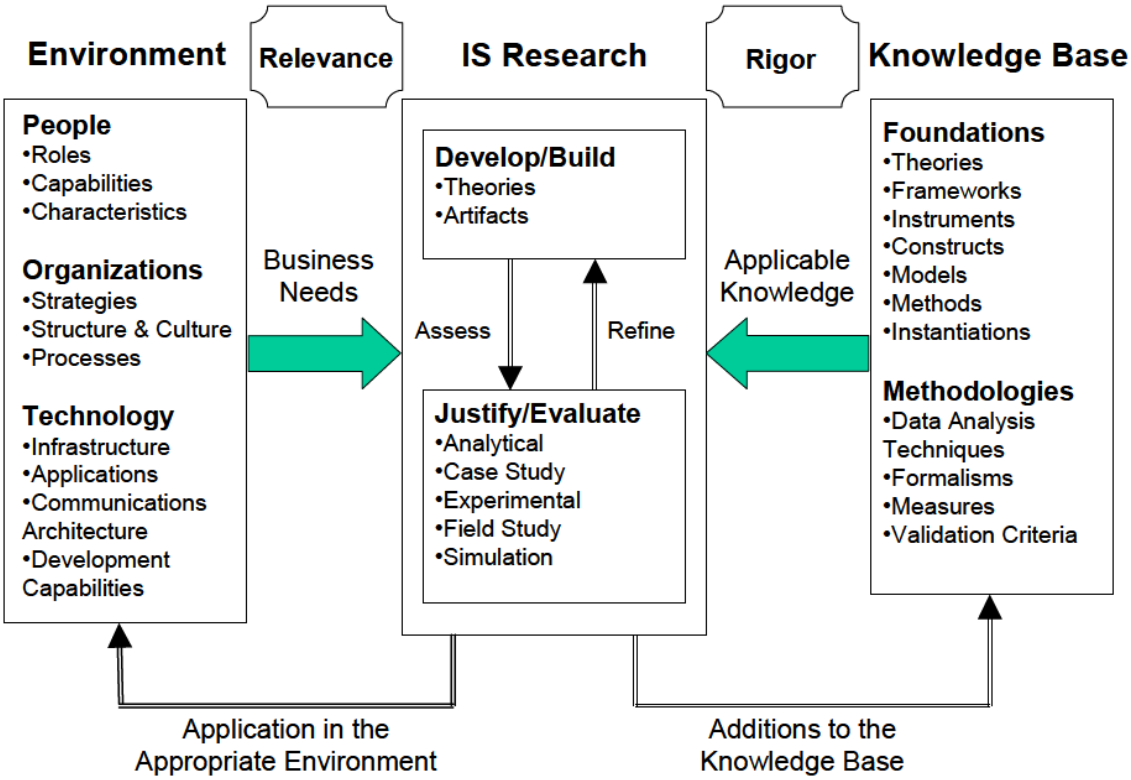
Taulukko1: Tutkimuksen viitekehys (March & Smith, 1995)

Taulukossa 1 ohjeita, joiden avulla suoritetaan itse tehtävä, kutsutaan menetelmäksi ja toteutuksen avulla testataan artifaktia sen omassa toimintaympäristössä. Toteutuksella testataan artifaktin käsitteitä, malleja ja menetelmiä niiden omissa toimintaympäristöissä. Taulukkoon 1 on merkitty tässä tutkimuksessa käytetyt luokittelut. Käsitteen rakentamisessa määritellään tutkimuksen vikatiedostoista löytyviä tutkittavia ominaisuuksia ja niitä myös arvioidaan tutkimuksen edetessä kun tietämys asiasta lisääntyy.

Ominaisuuksien välisiä suhteita voidaan kuvata malleilla, myös mallien arviointi on tärkeä vaihe. Menetelmä- ja toteutusvaiheita, joissa olisi rakennettu ohjeet itse tehtävään tai testattu artifakti, ei ollut mahdollista toteuttaa tämän tutkimuksen yhteydessä. March & Smithin (1995) mukaan suunnittelututkimuksessa on tarkoitus saavuttaa jotakin hyötyä rakentamalla uusi artifakti tai systeemi. Heidän mukaan tarvitaan suunnittelutieteellistä ja luonnontieteellistä toimintaa, jotta voidaan varmistaa että IT-tutkimus on sekä merkityksellistä että tehokasta.

March & Smithin viitekehystä täydensivät Hevner ym. (2004) ja Hevner & Chatterjee (2010): ympäristö määrittelee ongelma-alueen, ja se muodostuu ihmisistä, organisaatioista ja olemassa olevista tai suunnitelluista teknologioista. Liike-elämän tarpeet määritellään ympäristöön liittyvien tavoitteiden, tehtävien, ongelmien ja mahdollisuuksien mukaan siten kuin organisaatioon kuuluvat ihmiset ne ymmärtävät.

Kuviossa 4 esittää käsitteellistä viitekehystä informaatiosysteemin ymmärryksen, toteutuksen ja arvioinnin yhdistämisestä käyttäytymistieteen ja suunnittelutieteen ajatusmalleihin.



Kuvio 4: Tietojärjestelmien tutkimuksen viitekehys. (Hevner ym. 2004, 80)

Kuvion 4 keskellä on kuvattu käyttäytymistieteellinen ja suunnittelutieteellinen tutkimusvaihtoehto. Käyttäytymistieteellisessä vaihtoehdossa luodaan ja testataan teoriaa. Suunnittelutieteen tutkimuksessa rakennetaan ja arvioidaan artifakti. Näitä molempia tarkastellaan liiketoiminnan tarpeiden näkökulmasta. Tietämyskannan muodostavat peruspalikat ja metodologiat. Peruspalikoita tarvitaan teorian luonti- ja artifaktin rakentamisvaiheessa. Metodologiat tarjoavat ohjeita, joita voidaan käyttää testaus- ja arviointivaiheessa. (Hevner ym. 2004.)

IT-teknisten artifaktien suunnittelua, toteuttamista ja arviointia voidaan tarkastella Hevner ym. (2004) esittämien ohjeiden mukaan. Heidän mukaan on tärkeää ymmärtää, että suunnitteluongelman tieto ja ymmärrys ja sen ratkaisu on hankittu artifaktin rakentamisella ja soveltamisella. Seitsemän kohdan ohjeet ovat artifaktin suunnittelu (suuntaviiva 1), jossa suunnittelutieteellisen tutkimuksen tulos (IT-artifakti) pyrkii ratkaisemaan tärkeän ongelman ja on rakennettu sitä varten. Olennaisen ongelman painottaminen suunnittelussa (suuntaviiva 2), jolloin tavoitteena on hankkia tietämystä ja ymmärrystä joiden avulla saadaan ratkaisun rakentaminen ja sen toteutus aiemmin ratkaisemattomiin ongelmiin.

Artifakti täytyy arvioida (suuntaviiva 3), jossa sen hyödyllisyys, laatu ja vaikutus voidaan osoittaa. Tutkimuksella tuotetaan uutta tietoa, uusia menetelmiä tai merkittävä artifakti (suuntaviiva 4). Tieteellisen tarkkuuden painottaminen (suuntaviiva 5), näitä ovat suunnittelutieteessä

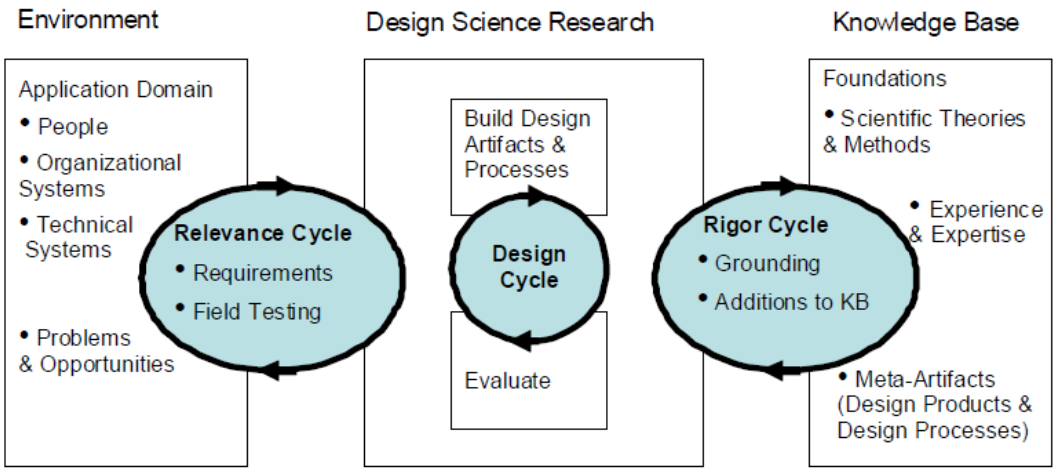
sovellettavat metodit, jotka huomioidaan sekä rakentamisessa että arvioinnissa. Suunnitteluprosessia käytetään ratkaisujen etsintäprosessina (suuntaviiva 6), ja siihen päästään iteraation avulla. Lopuksi on tärkeää uusien tulosten esittäminen sekä tutkija- että soveltajayhteisöille (suuntaviiva 7). (Hevnerin ym. 2004.)

Hevnerin ym. (2004) mukaan suunnittelutieteen tuloksena luodaan IT-artifakti tiettyyn tarkoitukseen ja artifakti käsittelee organisaation ongelmaa. Tavoitteena on hyödyllisyys. Suunnittelututkimuksen tarkoitus on kehittää tietämystä artifaktien suunnitteluun ja toteuttamiseen, esimerkiksi ratkaista rakentamisongelmia, tai käytettäväksi olemassa olevien ominaisuuksien suorituskyvyn parantamiseksi esimerkiksi ratkaista parantamisongelmia.

Tietojärjestelmätutkimusta on harjoitettu tietotekniikassa, ohjelmointituotannossa ja tietojärjestelmissä vuosikymmeniä. Alusta alkaen tietokoneen tutkijat ovat kehittäneet uusia arkkitehtuureja tietokoneita varten, uusia ohjelmointikieliä, uusia käännösohjelmia, uusia algoritmeja, uutta dataa ja tiedostojen rakenteita, uusia datamalleja, uusia tietokannan hallintajärjestelmiä ja niin edelleen. Lyhyesti sanottuna he ovat tehneet tietojärjestelmätutkimusta koko ajan. (Iivari 2007.)

Tietojärjestelmätutkimusprosessin ymmärtämisessä ja viestinnässä on hyvin tärkeää paitsi saada tukea tietojärjestelmäammattilaisten keskuudessa, on myös luoda uskottavuutta muilla tietojärjestelmätutkimuksen aloilla esimerkiksi eri tekniikan, arkkitehtuurin ja taiteen aloilla. (Hevner 2007.)

Kuviossa 5 on Hevnerin (2007) tietojärjestelmätutkimuksen viitekehyksen kolme tutkimussykliä: merkityssykli (Relevance Cycle), suunnittelusykli (Design Cycle) ja tarkkuussykli (Rigor Cycle). Merkityssyklissä yhdistetään tutkimusprojektin yhteydestä riippuva ympäristö suunnittelutieteen toimintoihin, suunnittelusyklissä iteroidaan rakentamisen perustoimintojen ja suunnitteluartifaktien arvioinnin sekä tutkimusprosessin välillä. Tarkkuussyklissä yhdistetään suunnittelutieteen toiminnat yhteen tieteellisen tietopohjan, kokemuksen ja tutkimusprosessiin liittyvien asiantuntijan kanssa. Hevner (2007) esittää että nämä kolme sykliä täytyvät olla läsnä ja selkeästi yksilöitävissä tietojärjestelmätutkimusprojektissa.



Kuvio 5: Tietojärjestelmien suunnittelututkimuksen syklit. (Hevner 2007)

Kuviolla 5 Hevner (2007) esittää, että tietojärjestelmätutkimus on motivoitunut parantamaan ympäristöä esittelemällä uusia ja omaperäisiä artifakteja sekä prosesseja näiden artifaktien rakentamiseen. Sovellusalue (Application Domain) koostuu ihmisistä, organisaatiojärjestelmistä ja teknisistä systeemeistä, jotka tekevät yhteistyötä saavuttaakseen tavoitteen. Hyvä tietojärjestelmätutkimus alkaa yleensä tunnistamalla ja esittämällä mahdollisuudet sekä ongelmat todellisessa sovellusympäristössä.

Merkityssykli aloittaa tietojärjestelmätutkimuksen sovellusalueella. Se ei vain tarjoa vaatimuksia panoksina tutkimukselle, vaan myös määrittelee hyväksymiskriteerit lopulliselle tutkimustulosten arvioinnille. Voiko suunnitteluartifakti parantaa ympäristöä ja kuinka tätä parannusta voidaan mitata? Työn tulos tietojärjestelmätutkimuksesta pitää palauttaa sovellusalueen ympäristölle tutkittavaksi ja arvioitavaksi. Kenttätutkimuksen tulokset määrittelee tarvitaanko uusia merkityssyklin iteraatioita tässä tietojärjestelmätutkimuksessa. Uudella artifaktilla voi olla puutteita toiminnallisuudessa tai ominaispiirteissä esimerkiksi suorituskyvyssä tai käytettävyydessä, jotka voivat rajoittaa sen käytännön hyötyä. Toinen tulos kenttäkokeista voi olla, että vaatimusten lähtötiedot tietojärjestelmätutkimukselle olivat virheellisiä tai puutteellisia, jonka tuloksena artifakti täyttää vaatimukset mutta on silti riittämätön esitettyyn ongelmaan. Toinen merkityssyklin iterointi alkaa kenttäkokeiden palautteesta ja todellisten kokemusten perusteella tehtyihin tutkimusvaatimusten parannuksiin. (Hevner 2007.)

Suunnittelutiede ammentaa laajasta tieteellisten teorioiden tietopohjasta ja suunnittelun menetelmistä, jotka tarjoavat perustan täsmälliselle tietojärjestelmätutkimukselle. Yhtä tärkeää on että tietämyskanta sisältää lisäksi kahden tyyppistä tietämystä. Ne ovat kokemus ja asiantuntemus sekä jo olemassa olevat artifaktit ja prosessit. Tarkkuussykli tarjoaa aikaisempaa tietoa tutkimushankkeelle varmistaakseen sen innovaation. Sykli on riippuvainen tutkijoista, että

he perusteellisesti tutkivat ja viittaavat tietopohjaan, näin voidaan taata, että syntyneet tuotteet ovat tutkimuksen tuloksia. Sisäinen suunnittelusykli on minkä tahansa tietojärjestelmien suunnittelututkimuksen projektin sydän. Tutkimustoiminnan sykli iteroi nopeammin artifaktin rakentamisen ja sen arvioinnin ja saadun palautteen välillä jalostaakseen suunnittelua eteenpäin. (Hevner 2007.)

3.1 Tiedon keräys

Työ liittyy projektiin, jossa aluksi tutkittiin syitä tukiaseman radiomoduulien vikaantumiseen, ja laajennettiin tutkimuksia myöhemmin systeemimoduuleihin. Tutkimusta varten tietoja kerättiin useilla menetelmillä. Projektin alussa tutkin huolellisesti NFF-ilmiöstä löytyvän kirjallisuuden, koska tutkimuksen pääpaino oli vikaantuneiden laitteiden syiden tutkiminen ja erityisesti NFF-määrityksen saaneet laitteet. Tarkemmin tutkin NFF-ilmiöstä löytyviä tutkimus- ja testaustuloksia, jotka olivat lentokone- ja elektroniikkateollisuuden aloilta (n=5). Perehdyin erilaisten tukiasemaelementtien ominaisuuksiin lukemalla sisäisiä dokumentteja (n=4). Tutkimuksen tavoitteena oli analysoida huollosta saatuja vikatiedostoja, ja kartoittaa mitä tietoa tutkimusmateriaalista oli saatavilla. Vikatiedostojen tulkintaa varten, täytyi selvittää tiedostoissa olevien parametrien ja vikakoodien merkityksiä. Vikakoodien merkitykset on hyvin dokumentoitu, laitteiden korjausprosessista ja vikatiedostojen muiden parametrien merkityksistä keskustelin asiantuntijoiden kanssa (n=2).

Kirjallisuusosassa on kuvattu tietojärjestelmätutkimuksen viitekehystä ja suunnittelututkimusta (n=20). Tiedonlouhinnan menetelmiä ja malleja tutkin laajasti, erityisesti niiden käyttöä tietoliikenneverkkojen tutkimuksissa (n=22). Tutkimuksen attribuutit on tarkemmin esitetty liitteessä 1. Tutkimustulosten analysointimenetelmäksi valittiin tiedonlouhintamalli (Vehviläinen, Hätönen & Kumpulainen 2003), jossa tiedonlouhintaprosessia on sovellettu tietoliikenteestä saatuun dataan. Materiaalia kerättiin myös omista palaverien muistiinpanoista sekä projektin aikana kirjoitetuista dokumenteista (n=18). Alustavana tutkimuksena tälle opinnäytetyölle voidaan pitää tapaustutkimusta ”Miten radio- ja systeemimoduulien vikaantumista voidaan ymmärtää?”, jonka tein Työn ja työelämän kehittämisen opintojaksolla.

3.2 Aineiston analyysi

Tietojen analysointi tehtiin käyttäen laadullisia menetelmiä. Denzin ja Lincolnin (1994) mukaan laadullinen tutkimus on monimenetelmäisenä keskipisteenä, johon kuuluu tulkitseva naturalistinen aiheen lähestymistapa. Tämä tarkoittaa, että pätevät tutkijat tutkivat asioita niiden luon-

nollisessa ympäristössä, yrittäen ymmärtää tai tulkita ilmiöitä merkityksien suhteen, joita ihmiset tuovat niille. Laadullinen tutkimus käsittää erilaisia empiirisiä tutkimustapoja, joita ovat muun muassa tapaustutkimus, henkilökohtainen kokemus, haastattelu, havainnoiva, historiallinen ja vuorovaikutuksellinen. (Denzin & Lincoln 1994.)

Tutkimusten tulokset ovat tulkintoja tutkittavasta materiaalista (Miles ym. 2014; Walsham 2006). Tulkitsevasta tutkimuksesta on kirjoittanut Walsham (2006) ja hänen mukaan merkittävää teorian valinnassa on tutkijan omakohtaisuus. Tutkimuksen ja siinä käytetyn menetelmän valintaan vaikuttaa tutkijan oma kokemus, tausta ja kiinnostuksen kohteet (Walsham 2006). Tulkinnallisesta tutkimuksesta voi katsoa tarkemmin (Walsham 2006 ; Orlikowski & Baroudi 1991 tai Walsham 1995).

Tutkija voi osallistua tutkimukseen ulkopuolisena tai sitoutuneena (Walsham 1995). Tätä tutkimusta olen tehnyt ulkopuolisena, sillä tutkimusaineisto saatiin valmiina. Tutkimusaineiston sisältöön pystyi vaikuttamaan, koska vikaraporttiin voidaan kerätä laitteista eri parametrejä. Tutkimustulosten tulkinnassa on ollut apua siitä, että olen ollut pitkään mukana tietoliikenneverkkojen tuotekehityksessä. Muut projektiin kuuluvat henkilöt olivat apuna tutkimustulosten analysoinnissa

3.3 Tutkimuksen triangulaatio

Triangulaatio vahvistaa tutkimusta yhdistämällä eri menetelmiä (Campbell & Fiske 1959). Triangulaation voidaan ajatella tarkoittavan erilaisten aineistojen, tutkijoiden, teorioiden tai menetelmien yhdistämistä tutkimuksessa. Triangulaatiolla voidaan osoittaa ettei jokin tulos ole pelkästään sattumanvarainen, jos sama tulos on saavutettu useilla eri lähestymistavoilla. Triangulaation käytöllä vahvistetaan tutkimuksen luotettavuutta ja validiutta, kun käytetään tutkimuksen erilaisia aineistotyyppejä, teorioita, näkökulmia tai analyysimenetelmiä. Aineisto- ja menetelmätriangulaatiolla tarkoitetaan sitä, että samassa tutkimuksessa voidaan yhdistellä useampia aineistoja ja menetelmiä. (Patton 2002; Gerring 2007; Yin 2014; Miles ym. 2014.)

Tutkimuksen syventämisessä apuna käytettiin triangulaarista aineiston, teorian ja menetelmän lähestymistapaa, jossa tutkimusta katsottiin eri näkökulmista. Yin (2014) mukaan tutkimuksen tietojen keräämiselle tulee vahvuutta siitä, että tietoja kerätään useasta eri paikasta. Tutkimuksessa selvitettiin tiedonlouhintaprosessien soveltuvuutta erilaisiin prosesseihin ja vertailemalla muiden teollisuusalojen laitteista saatujen NFF-vikojen tutkimustuloksia. Aineistotriangulaatiossa käytin apuna artikkeleita, tutkimuspalavereja ja -raportteja sekä teknisiä dokumentteja. Menetelmätriangulaatiossa hyödynnettiin useampia menetelmiä: tutkimalla tiedon-

louhinnanprosessien soveltuvuutta vikatiedostojen analysointiin sekä niiden integroimista tietojärjestelmien suunnittelututkimuksen sykleihin. Muu tutkimuksen apuna käytetty aineisto oli luotettavaa ja suurin osa siitä oli projektista saatua tutkimustietoa, jota voidaan pitää luotettavana ja oikeellisena.

3.4 Menetelmän yhteenveto

Usein tutkimuksissa käytetään monimenetelmällisyyttä, yksi syy on että tutkimusprojektit sisältävät yleensä useita erilaisia tutkimuskysymyksiä, joten yksi tutkimusmetodi ei välttämättä sovi kaikkiin. Toinen syy on että se mahdollistaa triangulaation. Työssä käytettiin induktiivista lähestymistapaa, koska vikaantumistiedostoista yritettiin löytää laitteiden erilaisia ominaisuuksia. (Gray 2004.)

Työssä tutkittiin laajasti tiedonlouhinnan KDD-prosessin malleja Fayyad, Piatetsky-Shapiro & Smyth (1996) ja Vehviläinen ym. (2003). KDD viittaa koko prosessiin, jonka osa tiedonlouhinta on. KDD sisältää tiedon valmistelun ja valinnan sekä tiedon puhdistuksen. Tiedonlouhintaprosessimalli (Vehviläinen ym. 2003) oli soveltuva vikatiedostojen tutkimiseen. Prosessien datan käsittelyvaiheiden avulla voidaan tehdä tiedostojen esikäsittelyä ja parantaa näin tutkittavaa lähtömateriaalia. Tiedonlouhinta on interaktiivinen ja iteroiva prosessi, jonka integroimista Hevnerin (2007) tietojärjestelmien suunnittelututkimuksen sykleihin tutkittiin osana vikatiedostojen analysointia.

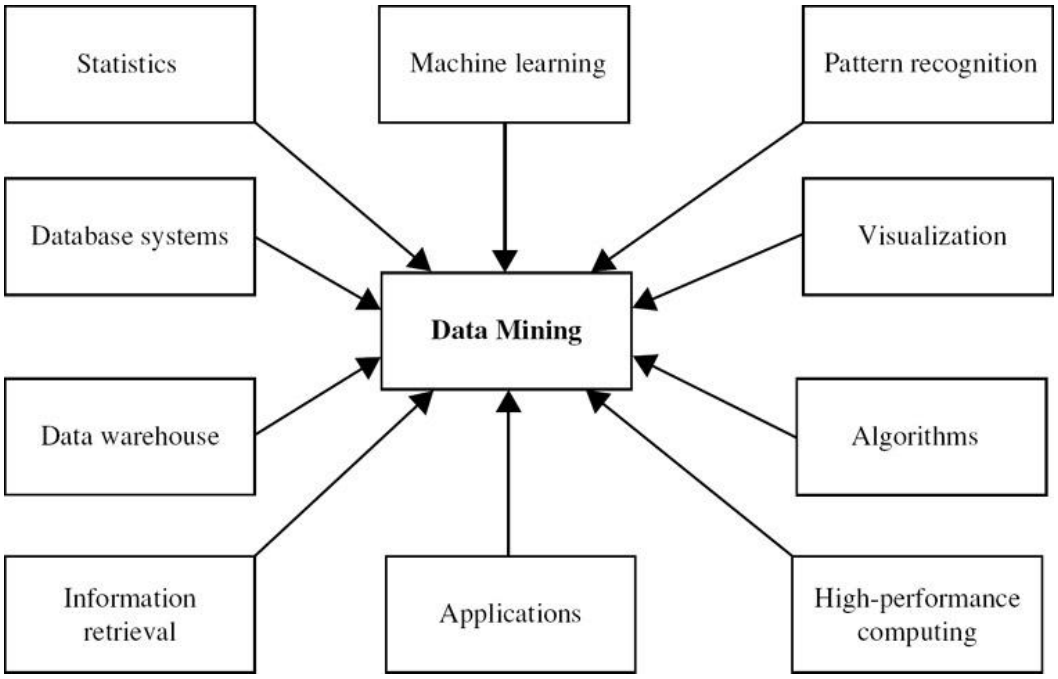
Tutkimusta varten kerättiin materiaalia useista lähteistä kirjallisuudesta, projektipalavereista, projektidokumenteista sekä tapaustutkimuksesta. Jotta saataisiin parempi ymmärrys vikatiedostojen parametrien merkityksestä, selvitettiin niitä dokumenttien sekä projektiin kuuluvien asiantuntijoiden avulla. Tutkimustuloksista kirjoitettiin tämä opinnäytetyö.

4 Tiedonlouhinta

Tiedonlouhinnalla (Data Mining) on monta määritelmää ja se on monitieteellinen tutkimus-alue. Sen avulla pyritään löytämään hyödyllistä informaatiota suurista data-aineistoista. Tiedonlouhinta- ja analysointitehtävät sisältävät yleensä klusterointia, luokittelua, tietojen regressiota, määrittämiseksi parametririippuvuuksien määrittämistä ja erilaisten poikkeamien etsiminen datasta (Laiho, Raivio, Lehtimäki, Hätönen & Simula 2005). Tilastollisilla ja koneoppimismenetelmillä pyritään muuttamaan dataa ymmärrettävään muotoon. Siten on helpompi ha-

vaita siinä esiintyviä poikkeavuuksia ja käyttää tuloksia päätöksenteon tukena. Tiedonlouhinnalle läheisiä tieteitä ovat tilastotiede, tietokannat, koneoppiminen, hahmontunnistus, tekoäly sekä visualisointi. Raaka-analyysivaiheen lisäksi siihen liittyy tietokannan ja datan hallinnointi näkökulmat, tietojen esikäsittely-, malli- ja päättelynäkökohdat, kiinnostavuusmittarit, monimutkaisuuden huomioonottaminen, löydettyjen rakenteiden jälkikäsittely, visualisointi ja online päivitys. Tiedonlouhinta on KDD–prosessin analyysivaihe. (Han, Kamber & Pei 2001; Talonen 2015; Nurminen 2003.)

Kuviossa 6 on kuvattu tiedonlouhinnan monitieteellisyyttä. Sen piiriin kuuluu joukko erilaisia menetelmiä ja algoritmejä. Tiedonlouhinnassa datan esikäsittely ja tulkinta ovat keskeisiä työvaiheita.



Kuvio 6: Tiedonlouhinta omaksuu tekniikoita monilta aloilta. (Han ja muut, 2001)

Witten & Frank (2005) käsittelee kirjassaan laajasti tiedonlouhinnan tekniikoita. Heidän mukaan lähdetieto voi olla käsite, tapaus tai ominaisuus ja tulostiedot taas koostuu tiedon esittämisestä, niitä ovat yleisesti päätöksentekokaavio ja luokittelusäännöt. Myös klusterointi kuuluu tulostiedon puolelle. Heidän mukaan tiedonlouhinta on ongelmien ratkaisemista analysoimalla dataa, joka löytyy valmiiksi tietokannoista. Tiedonlouhinta määritellään prosessina, joka löytää rakenteita tai malleja datasta. Prosessin pitäisi olla automaattinen tai puoliautomaattinen. Löydetyt ominaisuudet täytyy olla merkityksellisiä, jotta niistä olisi taloudellista hyötyä, ja datamäärät ovat poikkeuksetta suuria.

Talosen mukaan (2015) mukaan tiedonlouhinta on prosessi, jossa dataryhmien ominaisuudet saadaan selville. Päätarkoituksena on erotella tietoa laajasta datamäärästä ja muuttaa se ymmärrettävään muotoon. Se järkevöittää dataa erottelemalla mahdollisesti käyttökelpoisen tiedon ymmärrettävään muotoon. Nurminen määrittelee (2003) sen olevan tiettyä tarkoitusta varten kerättyjen suurten tietojoukkojen analyysi. Ja sen tarkoituksena on löytää odottamattomia suhteita ja tiivistää dataa helpommin ymmärrettäväksi ja että käyttökelpoisemmaksi. Tiedonlouhintaprosessi on iteratiivinen ja interaktiivinen, sen keskeisiä vaiheita ovat tiedon esikäsittely ja tulkinta. Tulkinnassa tiedonlouhinnassa saatuja tuloksia arvioidaan ja siinä voidaan arvioida tulosten visualisointia.

Vaikka tiedonlouhinta ja KDD voidaan usein pitää joukkona laskennallisia ja tilastollisia menetelmiä, joilla ratkaistaan tiedonhankintaan liittyviä ongelmia, ovat ne ensisijaisesti vuorovaikutteisia ja iteroivia prosesseja, joihin sisältyy lukuisia vaiheita analyysista tulkintaan ja tulosten hyödyntämiseen. (Äyrämö 2006.)

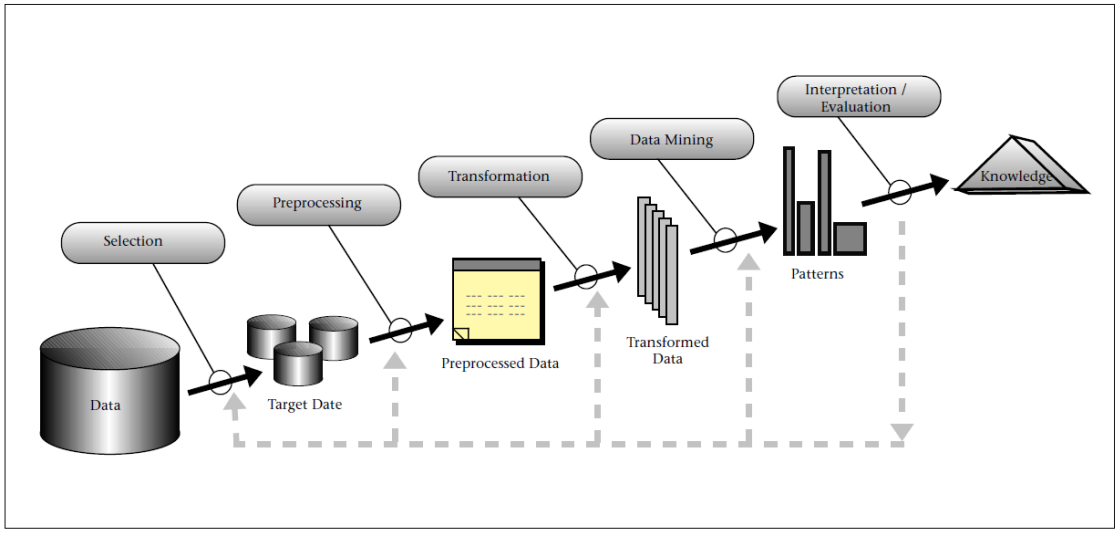
4.1 KDD-prosessi

Tämä tiedonlouhinnassa laajalti käytetty tekniikka on prosessi, joka sisältää tiedon valmistelun ja valinnan sekä tiedon puhdistuksen. KDD viittaa koko prosessiin, jossa etsitään hyödyllistä tietoa datasta ja tiedonlouhinta viittaa tämän prosessin tiettyyn toimenpiteeseen. Tiedonlouhinta käsittelee algoritmien keinoilla ominaisuuksien erottamista ja laskentaa datasta. KDD-prosessi on vuorovaikutteinen ja iteroiva, johon liittyy hyvin monta vaihetta, joissa käyttäjä tekee päätöksiä. (Fayyad, Piatetsky-Shapiro & Smyth 1996.)

Fayyadin ym. (1996) artikkelissa esitellään KDD-prosessin perusvaiheita. Ensiksi pitää ymmärtää sovellusalueet, asianmukainen aiempi tieto ja tunnistaa tavoite. Toiseksi luodaan kohteen datajoukko: valitsemalla datajoukko tai keskittymällä tiettyihin muuttujiin tai datanäytteisiin, joiden perusteella havaintoja datasta tehdään. Kolmas vaihe on tietojen puhdistus ja esikäsittely. Neljäntenä on datan yksinkertaistaminen ja suunnittelu, datan hyödyllisten ominaisuuksien löytäminen ja edustaminen riippuen tehtävän tavoitteesta. Viides on asettaa tavoitteet tiettyyn tiedonlouhintamenetelmään. Kuudentena on tutkimusanalyysi ja malli sekä hypoteesiin valinta: valitsemalla tiedonlouhinnan algoritmi ja valitsemalla menetelmä, joita käytetään datan ominaisuuksien etsimiseen.

Seitsemäs vaihe on tiedonlouhinta: etsitään kiinnostavia ominaisuuksia esityksen muodosta tai esityksien joukosta, joita ovat luokittelusäännöt tai -puut, regressio ja klusterointi. Kahdeksas vaihe tulkitsee louhinnasta saatuja ominaisuuksia ja mahdollisesti palaa johonkin aikaisempaan

vaiheeseen niiden uudelleen iteroimiseksi. Tämä vaihe voi myös sisältää visualisointia erotetuista ominaisuuksista ja malleista. Viimeisenä tutkitaan löydettyjä tiedostoja: käyttämällä tietoja suoraan, tiedon sisällyttämistä toiseen järjestelmään lisätoimia varten, tai yksinkertaisesti dokumentoida sitä ja raportoimalla siitä kiinnostuneille osapuolille. KDD prosessiin voi liittyä merkittävää iterointia ja se voi sisältää silmukoita minkä tahansa kahden vaiheen välillä. Nämä vaiheet on esitetty kuviossa 7. (Fayyad ym. 1996.)



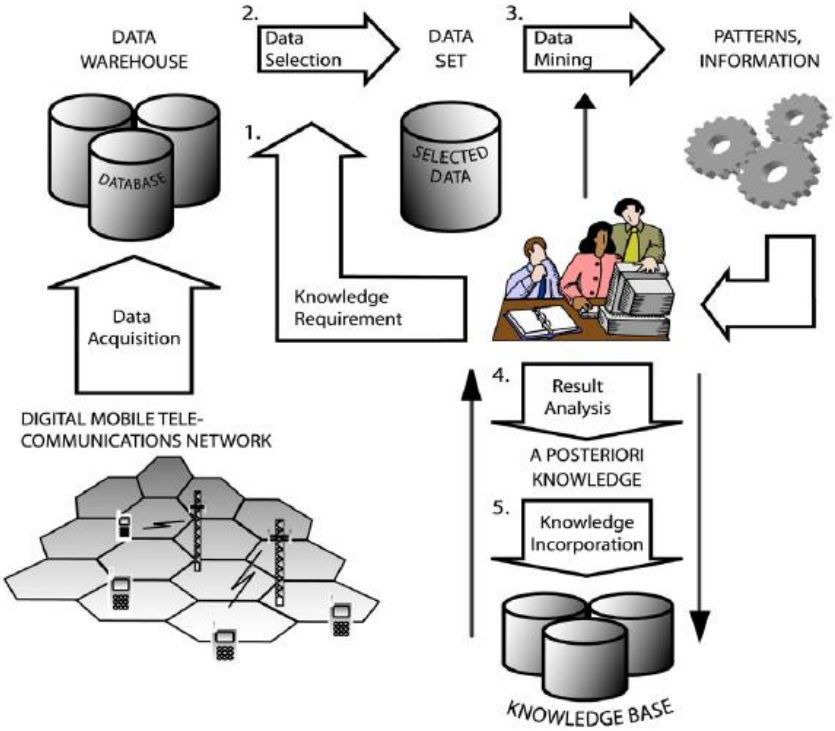
Kuvio 7: Yleiskuva vaiheista, jotka muodostavat KDD-prosessin. (Fayyad ym. 1996)

Kuviossa 7 on tiedonlouhinta esitetty yhtenä vaiheena liittyvää tiedonhankintaprosessia, vaikkakin olennaisena osana, koska se paljastaa arvioinnille piileviä malleja. Kuitenkin useasti termiä tiedonlouhinta käytetään viittaamaan koko tiedonhankintaprosessia. Laaja näkemys tiedonlouhintatoiminnoista on, että tiedonlouhinta on prosessi, joka löytää mielenkiintoisia malleja ja tietämystä suurista määristä dataa. Tietolähteet voivat sisältää tietokantoja, tietovarastoja, Web, muiden tietojen arkistot tai tietoja, joita tulee järjestelmään dynaamisesti. (Fayyad ym. 1996.)

Vehviläinen ym. (2003) käsittelevät myös KDD- ja tiedonlouhintaprosessia. KDD-prosessin tarkoituksena on löytää uutta tietoa sovellusalueelta ja prosessi sisältää monia erillisiä peräkkäisiä tehtäviä, joista tiedonlouhintavaihe tuottaa malleja ja informaatiota analysoitavaksi. KDD käsittelee dataa, joka on tietovarastoissa tai tietokannoissa. Artikkelissa esitetään KDD-prosessille viisi päävaihetta, koska datan esikäsittely ja muutos on sisällytetty tiedonlouhinnan viiteen vaiheeseen. Nämä molempien prosessien vaiheet on esitetty kuvioissa 8 ja 9.

KDD-prosessin ensimmäisessä vaiheessa täsmennetään tietämyksen vaatimukset, eli määritellään mitä analysoija haluaa tietää sovellusalueelta. Jotta pystyy täsmentämään tietämyksen

vaatimuksia analysoijalla pitää olla ennakkotietämys. Toinen vaihe on valita mahdollisesti useista lähteistä, oikeanlaista dataa, joka tukee vaatimuksia. Tässäkin analysoijalta vaaditaan taitoja ja ennakkotietämystä asiasta. Kolmas vaihe on tiedonlouhinta ja se on itsessään prosessi, joka käsitellään tarkemmin myöhemmin. KDD-prosessiin kuuluu myös tulosten analysointi ja tiedon sisällyttäminen tietokantaan. Tässä työssä keskitytään tarkemmin tiedonlouhintaprosessiin. (Vehviläinen ym. 2003.)



Kuvio 8: KDD-prosessi. (Vehviläinen ja muut, 2003)

Kuviossa 8 KDD-prosessin vaiheet muodostuvat: 1. Tietämyksen vaatimuksesta tai informaation tarpeesta, 2. Datan valinta, 3. Tiedonlouhinta, 4. Tulosten analysointi ja 5. Tietämyksen tai informaation sisällyttäminen.

4.2 Tiedonlouhintaprosessi

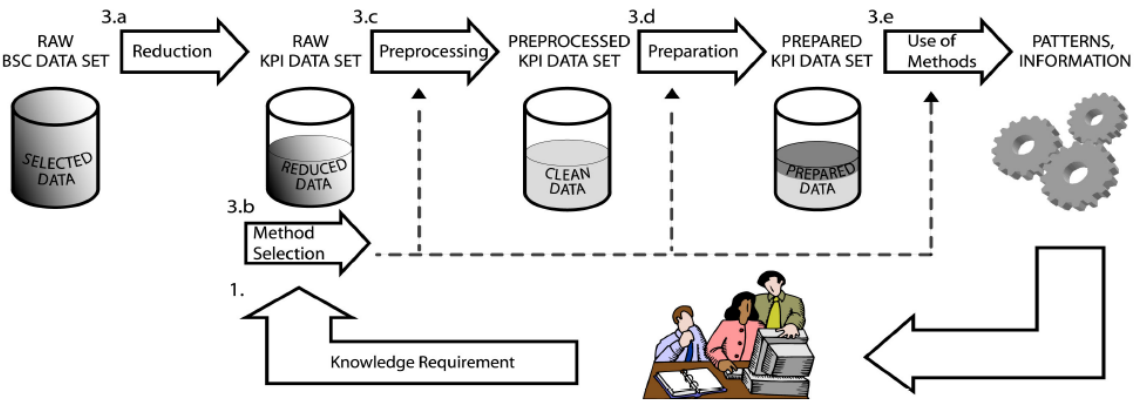
Tiedostojen analysoinnissa olisi houkuttelevaa hypätä suoraan datan louhintaan, mutta ensin täytyy saada data valmiiksi. Tämä merkitsee sitä, että tarkastellaan ominaisuuksia ja dataarvoja. Reaalimaailman data sisältää tyypillisesti häiriöitä, sitä on valtavia määriä esimerkiksi gigatavuja tai enemmän, ja se voi olla peräisin sotkuisista heterogeenisistä lähteistä. Datan

tietämys on ensimmäinen tärkeä vaihe tiedonlouhintaprosessissa ja se on hyödyllistä myös datan esikäsittelyprosessissa.

Han ym. (2001) listaavat seuraavia asioita mitä datasta halutaan tietää: Mitkä ovat eri ominaisuuksia tai kenttiä, jotka muodostavat tiedon? Millaisia arvoja ominaisuudet sisältävät? Mitkä ominaisuudet ovat erillisiä, ja mitkä arvioidaan jatkuviksi arvoiksi? Miltä data näyttää tai miten arvot jaetaan? Onko olemassa keinoja visualisoida dataa, jotta siitä saadaan parempi käsitys? Voidaanko mahdolliset poikkeavuudet havaita? ja Voidaanko joidenkin datan kohteiden samankaltaisuutta mitata suhteessa toisiin?

Näiden tarkastelujen avulla saadaan käsitys datasta ja se on apuna myöhemmissä analyyseissä. Tiedonlouhintasovelluksissa, kuten klusterointi tai poikkeavuuksien analysointi, täytyy arvioida kuinka tyypillisiä tai epätyypillisiä objektit ovat toisiinsa verrattuina. Klusteri on data pisteiden joukko, jonka objektit klusterin sisällä ovat samanlaisia toistensa kanssa ja erilaisia kuin objektit toisissa klustereissa. Poikkeavuuksien analysoinnissa käytetään myös klusteriperusteista tekniikkaa, jolla identifioidaan mahdolliset poikkeavat objektit jotka ovat hyvin erilaisia kuin toiset. Tietoa objektien yhtäläisyyksistä voidaan käyttää lähin naapuri luokituksessa. (Han ym. 2001.)

Tiedonlouhintaprosessi, joka on esitetty tarkemmin kuviossa 9, on osa KDD-prosessia ja ensimmäinen vaihe näissä prosesseissa on määrittää tietämyksen vaatimukset. Se tarkoittaa sitä, että rajataan tarkasti mitä sovellusalueelta halutaan tietää. Jotta voidaan määritellä tietämyksen vaatimukset, tarvitaan ennakkotietoa sovellusalueesta. Seuraava vaihe on valita, mahdollisesti useista eri lähteistä oleva data, joka tukee tietämyksen vaatimuksia. Myös tässä vaiheessa ovat tarpeen taidot ja ennakkotietämys aihealueesta, koska niitä tarvitaan määritettäessä mitä tietoa tietokannasta haetaan. Vasta sen jälkeen alkaa varsinainen tiedonlouhintaprosessi. (Vehviläinen ym. 2003.)



Kuvio 9: Tiedonlouhintaprosessi. (Vehviläinen ja muut, 2003)

Kuviossa 9 esitetyt vaiheet ovat: 3a. datan pelkistäminen tai yksinkertaistaminen, 3b. menetelmän valinta, 3c. datan esikäsittely, 3d. datan valmistelu ja 3e. metodien käyttö. Prosessin aluksi otetaan käyttöön sitä varten valittu data. Seuraavat vaiheet ovat datan vähentäminen ja metodin valinta. Esikäsittelyssä vähennetään häiriöitä, käsitellään ääritapaukset ja puuttuvat arvot, tasapainotetaan muuttujat ja niiden raja-arvot. Esikäsittely on datajakautumien analysoinnin ja muokkauksen iteratiivinen prosessi. Esikäsittelyvaiheen tavoite on mahdollistaa analyysimenetelmät, joilla erotetaan oikeat ja asiaankuuluvat tiedot datasta. (Hätönen 2009; Vehviläinen ym. 2003.)

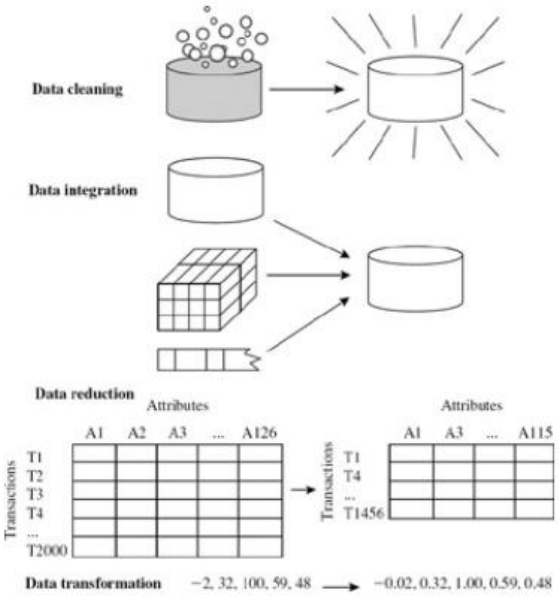
Esikäsittelyä seuraa analyysimenetelmän valinta, valinnassa otetaan huomioon tilastollinen tieto ja analyysivaiheen päämäärät. Kun menetelmä on valittu, halutut mallit voidaan erotella. Vaatimusten louhinta varmistaa että data on esikäsitelty niin että valitut analyysimenetelmät pystyvät tarjoamaan vaaditut tiedot siitä. Analyysin jälkeinen vaihe koostuu informaation ja tiedon tulkinnasta, esittelystä ja validoinnista. (Hätönen, 2009; Vehviläinen ym. 2003.)

4.3 Datan esikäsittelyvaiheet

Tämän päivän reaalimaailman tietokannat ovat erittäin herkkiä häiriöille, puuttuville ja ristiriitaisille tiedoille. Tämä johtuu siitä että datan koko on usein monia gigatavuja ja ne ovat todennäköisesti lähtöisin useista, heterogeenisista lähteistä. Alhaisen laadun data johtaa alhaisen laadun louhintatuloksiin. Datan esikäsittelytekniikoita on useita, datan puhdistusta voidaan käyttää häiriöiden poistamiseen ja korjaamaan ristiriitaisuuksia datasta. Datan yhdistämisessä liitetään yhteen useista lähteistä olevat datat. Datan yksinkertaistaminen voi vähentää datan kokoa, esimerkiksi kokoamalla yhteen, poistamalla tarpeettomat ominaisuudet tai klusteroimalla. (Jerome 2015; Talonen 2015.)

Datan muutosta tai normalisointia käytetään, kun datan skaalausta muutetaan pienemmälle asteikolle. Tämä voi parantaa louhinnan algoritmien tarkkuutta ja tehokkuutta, joihin liittyy etäisyyden mittausta. Nämä tekniikat eivät ole toisiaan poissulkevia, vaan ne voivat toimia yhdessä. Esimerkiksi datan puhdistus voi liittyä muutoksiin korjata väärät tiedot, kuten muuntamalla kaikki päivämääräkentän merkinnät yleiseen muotoon. (Lumme 2012; Han ym. 2001.)

Kuviossa 10 on koottu esikäsittelyn vaiheet. Vaiheiden edellinen luokittelu ei ole toisiaan poissulkevia. Esimerkiksi tarpeettoman datan poistamista voidaan pitää datan puhdistuksena, kuin myös datan yksinkertaistamisena.



Kuvio 10: Datan esikäsittelyprosessit. (Han ym. 2001)

Kuviossa 10 kuvataan esikäsittelyprosesseja. Esimerkiksi datan puhdistuksessa täytetään puuttuvia arvoja, tasoitetaan datan häiriöitä, tunnistetaan tai poistetaan poikkeavuuksia ja epäjohdonmukaisuuksia. Dataa voidaan joutua yhdistämään useista kohteista, jolloin tiedostot ovat eri formaateissa. Datan määrä on tiedonlouhinnassa suuri, datan vähentämisellä parannetaan saatavien tulosten laatua. Viimeisenä kohtana datan muokkaaminen, jolla data muutetaan tiedonlouhintaan sopivaksi. (Han ym. 2001; Jerome 2015.)

Yhteenvetona voidaan sanoa että reaalimaailman data yleensä likaista, epätäydellinen ja epäjohdonmukaista. Tietojen esikäsittelytekniikat voivat parantaa datan laatua, mikä osaltaan parantaa tarkkuutta ja tehokkuutta myöhemmissä louhintaprosesseissa. Tietojen esikäsittely on tärkeä vaihe tiedonhankintaprosessissa, koska laatupäätösten on perustuttava laadulliseen dataan. Havaitsemalla datan poikkeavuudet, korjaamalla ne aikaisessa vaiheessa, ja vähentämällä analysoitavan datan määrää voi johtaa valtaviin tuloksiin päätöksenteossa. (Han ym. 2001.)

Esikäsittely on tärkeää, koska se määrittelee merkittävän vaihteluvälin jokaiselle muuttujalle datajoukossa ja myös sen mitkä datapisteet ovat sopivia analyysiin. Usein kuitenkin käy niin että vaikka esikäsittelyllä on tärkeä merkitys data-analyysiin, siihen liittyvät asiat jätetään käsittelemättä tieteellisessä kirjallisuudessa. (Jerome 2015.)

4.4 Poikkeavuuksien tunnistaminen (Anomaly Detection)

Säännöttömyyden havaitsemisesta käytetään termejä poikkeamat (anomalies) tai ulkopuolinen havainto (outliers). Poikkeamat voidaan määritellä, että ne poikkeavat standardista, normaalista tai odotettavissa olevasta. Ulkopuolinen havainto voidaan määritellä olevan datapiste kuvaajassa tai joukossa tuloksia, jotka ovat paljon isompia tai pienempiä kuin seuraavaksi lähin datapiste. Sana poikkeama voi tarkoittaa kaikkia abstrakteja ilmiöitä, joita ei ole odotettu. Kun taas ulkopuolisen havainnon määritelmä eroaa tästä viittaamalla mitattuun datajoukkoon. Ne voivat olla virheitä tai merkkejä ei-toivotuista suorituksista ja ne tulisi havaita niin pian kuin mahdollista. Datassa olevat poikkeamat voivat olla merkkejä virheistä tai prosessin toimintahäiriöistä, ja niihin sisältyy virheet mittalaitteissa, tiedonsiirrossa tai varastoinnissa. Ne voivat olla myös merkkejä järjestelmän luvattomasta käytöstä, esimerkiksi tunkeutumisen havaitseminen (intrusion detection) tai petoksen paljastaminen (fraud detection). (Jerome 2015; Kumpulainen 2014; Fayyyad ym. 1996.)

Poikkeamien havaitseminen on yksi tärkeimmistä tiedonlouhinnan tehtävistä ja se on myös tärkeä osa prosessin monitorointia useilla teollisuudenaloilla. Koska teollisuuden sovelluksien tuottama datamäärä on yleensä suuri, prosessinhoitajille olisi hankalaa selata kaikki tiedot manuaalisesti. Sen vuoksi on tarvetta automatisoiduille sovelluksille, jotka löytävät kriittisimmän tiedon koko datamäärästä, joilla voidaan tukea operaattoreita heidän päätöksenteossaan. Automaattinen poikkeamien ilmaisun sovellus voidaan kuvata välineenä, joka auttaa suodattamaan suuren osan normaalia käyttäytymistä ja paljastaa poikkeavan käyttäytymisen loppukäyttäjälle tai järjestelmälle. Poikkeavuuksien havaitsemisen menetelmät on laajasti käytetty suorituskyvyn, vikojen ja turvallisuuden hallinnassa. Poikkeamien perusteella voidaan löytää erilaisia huomiota tarvitsevia tilanteita. Ne voivat aiheutua laitteiden toimintahäiriöstä tai ohjelmistokomponenteista. (Höglund, Hätönen. & Sorvari 2000; Kumpulainen 2014.)

Eskin (2000) esittää artikkelissaan kolme olettamusta, jotka tulee täyttää ennen kuin poikkeavuudet pystyy havaitsemaan: normaalia dataa voidaan mallintaa käyttämällä todennäköisyysjakautumaa, poikkeavuudet eroavat riittävästi normaalista datasta ja poikkeavuuksien määrä on pienempi verrattuna normaalihavaintojen määrään. Poikkeavuuksien havaitsemistekniikalla on erittäin suuri rooli tietoliikenneverkon monitoroinnissa ja sen käyttöä tietoliikenneverkkojen monitoroimisessa on käsitelty laajasti (Kumpulainen ja muut, 2008; Kumpulainen, 2014; Talonen, 2015; Jerome, 2015). Poikkeavuuksien havaitseminen on keskeinen työkalu petoksien tunnistamisessa esimerkiksi tietoliikenneverkoissa. Sitä on myös käytetty radiorajapinnassa ja palvelimien lokien monitoroimiseen (Höglund 2006; Kumpulainen 2014).

Yleisimmin käytetyssä poikkeamien havaitsemistekniikassa jako perustuu saatavilla olevan datan ominaispiirteisiin ja ne voidaan jakaa kolmeen osaan. Valvomattomien poikkeamien ilmaisutekniikka (Unsupervised anomaly detection technique) havaitsee poikkeamat merkitsemättömistä testituloksista olettaen, että suurin osa tapauksista tietojoukossa ovat normaaleja, ja näin etsimällä ilmentymiä, jotka sopivat vähiten jäljelle jääneeseen dataan. Käytännössä on yleistä että dataa ei voi merkitä nimilapuilla, tämän vuoksi valvomattomien poikkeamien menetelmä on yleisesti käytössä. Tämä menetelmä vaatii erityisen suuren datamäärän, jotta tunnistetaan prosessin normaalitila. Tunnistamisessa käytetään riittävän suurta määrää havaintoja, jotka on saatu normaalista käyttäytymisestä. (Jerome 2015; Kumpulainen 2014.)

Ohjattu poikkeamien ilmaisutekniikka (Supervised anomaly detection technique) vaatii tietoaineistoa, joka on merkitty "normaali" ja "epänormaali". Puolivalvottujen poikkeamien ilmaisutekniikat (Semi-supervised anomaly detection technique) olettaa, että toinen havainnoista on merkitty joko normaali tai poikkeavuus. Normaalidata on helpompi merkitä, kun taas kaikenlaisten poikkeavuuksien merkintä on lähes mahdotonta. Menetelmä tunnistaa mallin normaalikäyttäytymisen ja mikä tahansa muu, joka eroaa merkittävästi normaalista, pidetään poikkeavana. (Zhang & Zulkernine 2006; Yamanishi & Takeuchi 2001; Kumpulainen 2014.)

4.5 Klusterointi

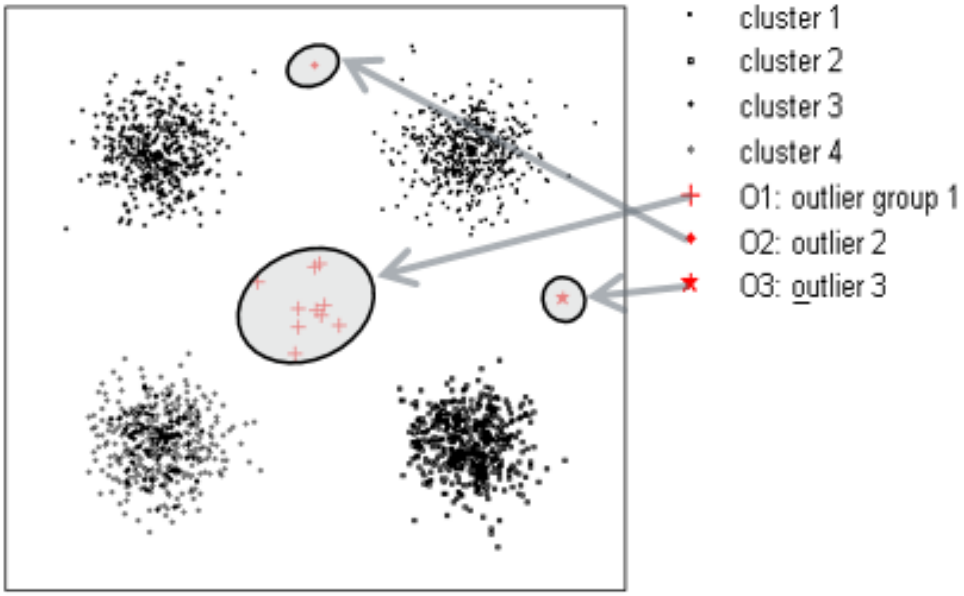
Tiedonlouhinnassa käytetään tietoalkioiden ryhmittelyä eli klusterointia (clustering). Klusterontia voidaan pitää tiedonlouhinnan perusmenetelmänä. Klusteroinnissa pyritään jakamaan alkiot ryhmiin siten, että alkiot kussakin ryhmässä ovat keskenään mahdollisimman samanlaisia mutta eri ryhmissä alkiot taas olisivat mahdollisimman erilaisia keskenään. Näitä muodostettuja ryhmiä kutsutaan klustereiksi. Eräs vaatimus on, että kahdella eri klusterilla ei ole yhteisiä alkioita. Klusterointia varten tarvitaan keino mitata aineiston alkioiden erilaisuutta tai samankaltaisuutta. Voidaan olettaa aineiston olevan joukko tason pisteitä ja aineiston alkioiden erilaisuusmittana käytettävän kolmiulotteista etäisyyttä. (Jain & Dubes 1988; Nurminen 2005.)

Klusterointimenetelmiä on monia ja ne voidaan jakaa hierarkkisiin menetelmiin ja osittaviin menetelmiin (katso esim. Jain & Dubes 1988; Äyrämö 2006; Laiho ym. 2005). Klusterointi kuuluu koneoppimisen kannalta ohjaamattoman oppimisen menetelmiin, kun taas tilastotieteessä sitä pidetään monimuuttujamenetelmiin kuuluvana ryhmittelyanalyysinä. Menetelmän valinnassa on otettava huomioon klusteroitavan aineiston tyyppi sekä klusteroinnin käyttötarkoitus. (Nurminen 2005.)

Klusterianalyysiksi kutsutaan datapisteiden jakoa erotusprosessissa osajoukoiksi, joilla on merkitystä tietyssä ongelman asiayhteydessä. Jokainen näin saatu klusteri edustaa datapisteiden

ryhmää, joilla on samanlainen käyttäytyminen. On olemassa erilaisia klusterointimekanismeja ja -algoritmeja, joita käytetään eri tilanteissa. Kaikki nämä menetelmät yrittävät saavuttaa saman tavoitteen pyrkimällä erilaisten mekanismien kautta sisäiseen yhdenmukaisuuteen ja ulkoiseen erottamiseen. Eräs hyvin yleisesti käytetty osittava klusterointitekniikka on k-means clustering, menetelmää on kuvattu viitteissä (Jerome 2015; Talonen 2015).

Kuviossa 11 on Jeromen (2015) esimerkki visuaalisesti havaittavissa olevista datajoukon klustereista, outliereista sekä outlier ryhmästä. Siinä on kuvattu keinotekoinen kaksi ulotteinen datajoukko. Datassa olevat poikkeamat tai outlierit ovat usein merkkejä virheistä tai häiriöistä.



Kuvio 11: Esimerkki klustereista ja ulkopuolisista havainnoista (outliers). (Jerome 2015)

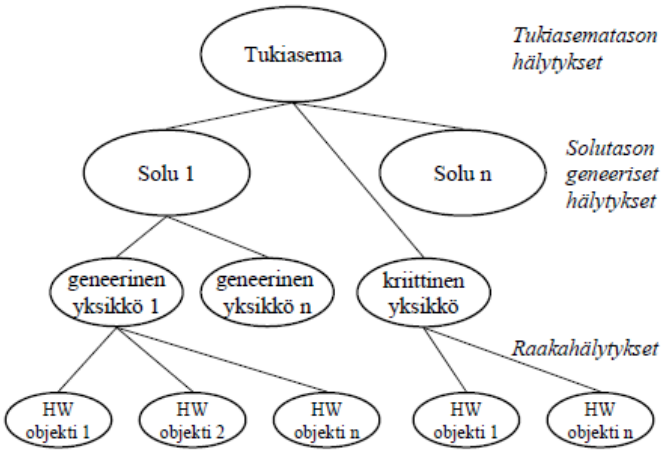
Kuviossa 11 on havainnollistettu neljä eri klusteria ja keskellä kuviota on outlier-ryhmä, joka on liian kaukana kaikista neljästä klusterista. Yksittäiset pisteet, jotka merkitty 02 ja 03, ovat myös outlier-pisteitä, koska sijaitsevat kaukana (Jerome 2015). Klusterointia voidaan käyttää hyödyksi verkon solujen parametrien optimoinnissa, jolloin solut esimerkiksi jaotellaan liikenneprofiilien ja -määrien mukaan, solutyyppien tai radioresurssien hallinnan suorituskyvyn mukaan. Kun yhtenäinen ryhmä löydetään, voidaan sen kaikille soluille asettaa samat parametrit. (Laiho ym. 2005.). Siten myös poikkeamia näissä ominaisuuksissa on helpompi seurata.

5 Tutkimustulokset

Tässä työssä selvitettiin miten tiedonlouhintaprosesseja voidaan käyttää avuksi huollosta saatujen tietoliikenne-elementtien vikatiedostojen analysoimiseen, ja erityisesti näiden menetelmien käyttöä NFF-elementtien löytämiseksi. Kaikista käytössä vikaantuneista laitteista ei huollon testeissä löydy syytä vikaantumiselle, jolloin laite saa testien jälkeen merkinnän NFF. Erityisesti näistä laitteista aiheutuu turhia kuljetus- ja huoltokustannuksia. Tutkimuksen tarkoituksena on lisätä tietämystä NFF vikamäärityksen saaneista laitteista. Samalla tutkin voidaanko tiedonlouhintaa integroida sujuvaksi osaksi Hevnerin (2007) tietojärjestelmän sykleihin. Vertasin myös NFF-laitteiden huollossa saatuja testaustuloksia muiden teollisuusalojen vastaaviin NFF-laitteiden testaustuloksiin. Tutkimusmateriaalina olivat huollon vikaraportit, jotka oli kerätty huoltoon tulleista vikaantuneista tukiaseman elementeistä. Tietoliikenneverkon elementeistä kerätään jatkuvasti paljon tietoa, sen vuoksi tietoa on runsaasti saatavilla myös vikatilanteesta.

5.1 Elementtien vikaantuminen

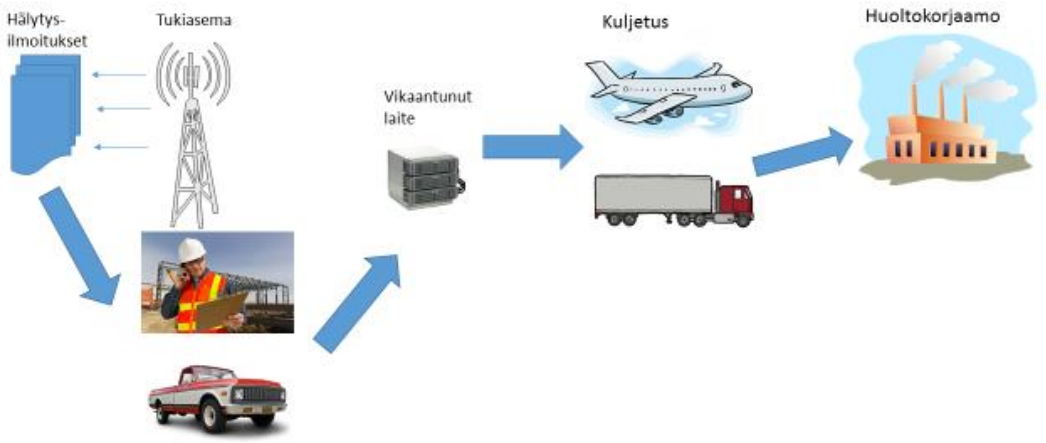
Tietoliikenneverkot ovat monimutkaisia ympäristöjä, joiden laitteiden tulisi olla helppokäyttöisiä ja luotettavia. Sen vuoksi niitä on rakennettu mahdollisimman itse-organisoituviksi. Vikadiagnostiikka toimii siten, että laitteessa ilmenneestä viasta ilmoitetaan käyttäjälle ja sen tarkoituksena on estää vikaa vahingoittamasta muita laitteistoja. Vikadiagnostiikan avulla käyttäjälle tulisi ilmetä virhetilanteessa missä vika on ja millaisia toimintoja pitäisi tehdä, jotta vika saataisiin poistettua. Kuviossa 12 on esitetty vikadiagnostiikan periaate, johon tukiasemien hälytysten luokittelu perustuu.



Kuvio 12: Tukiasemahälytysten luokittelu verrattuna Tervoseen (2002)

Kuten kuviosta 12 voidaan havaita tukiasemat lähettävät kolmen tasoisia hälytyksiä: tukiasema-, solu- ja yksikkötasoisia ns. raakahälytyksiä. Hälytiedostojen tutkiminen on hyvin vaativaa, juuri sen takia että hälytyksillä on monta eri tasoa. Niiden tutkimisessa tarvitaan tukiasemien hälyihin perehtyneitä asiantuntijoita.

Tukiasemien vikaantuneet elementit noudetaan usein hankalista olosuhteista esimerkiksi korkeista mastoista. Riippuen tukiaseman lähettämän hälytyksen tasosta vikaantunut elementti yritetään saada toimintakuntoon ensin paikan päällä esimerkiksi resetoimalla eli ottamalla laitteesta virrat pois. Jos elementtiä ei saada toimimaan se vaihdetaan ja vikaantunut elementti lähetetään huoltokorjaamoon. Huoltokorjaamot on keskitetty, joten ne voivat sijaita hyvinkin kaukana. Kuviossa 13 on havainnollistettu vikaantuneiden elementtien kuljetusprosessia huoltoon.



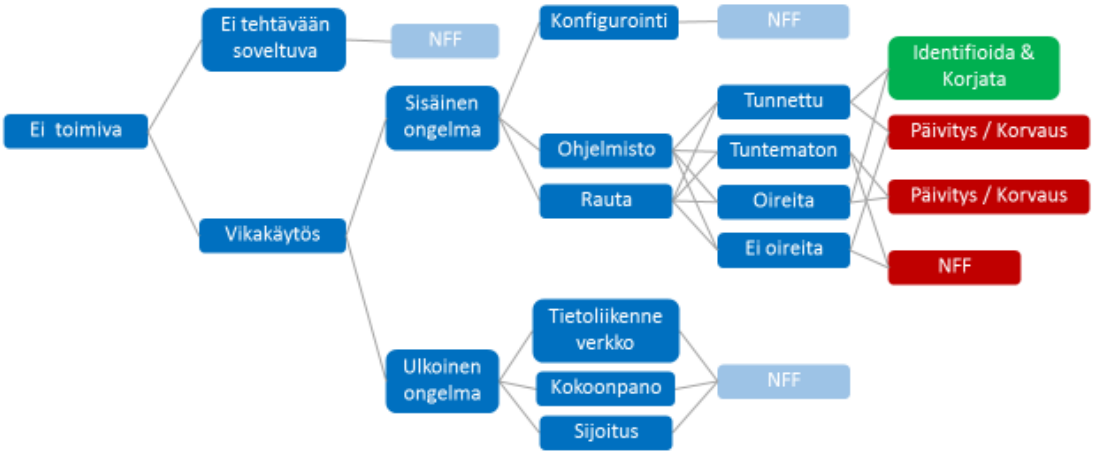
Kuvio 13: Esimerkki vikaantuneiden elementtien kuljetusketjusta.

Kuviossa 13 kuvataan tukiaseman lähettämän hälytysten jälkeen paikan päälle menee huoltohenkilökuntaa. Laite voidaan korjata tai se vaihdetaan ja vikaantunut laite lähetetään huoltokorjaamoon. Elementin tullessa huoltoon sen kunto tarkistetaan ja esitestataan, jonka jälkeen sille luodaan vikaraportti. Siihen kootaan operaattorilta saadut vikakuvaukset sekä huollon testien perusteella määritetyt vikakoodit. Elementtien korjausprosessissa niille tehdään erilaisia testejä. Testitulokset kerätään vikaraporttiin. Viat voivat olla sekä laitteisto- että ohjelmistolähtöisiä, joskus myös luonnon laitteille aiheuttamia vaurioita muuan muassa vesi, hyönteiset, korroosio ja lika.

Tähän tutkimukseen saatiin vikaantuneista elementeistä raakahälytystiedostot sekä huollon vikaraportti. Raakahälytiedostosta saadaan elementtien loki- ja hälytystiedot. Näistä saadaan

laitteen vikaantumisen yhteydessä esiintyneet hälytykset ja lokit. Vikakoodeja päivitetään vikaraporttiin testauksen edetessä ja tehdyt huollon korjaustoimenpiteet kirjataan myös vikaraporttiin.

Kuviossa 14 esitetään erilaisia vika-analyysivaihtoehtoja huoltoon saapuvalle elementille eli laitteelle, joka saa mahdollisesti NFF-määrityksen. Näitä vaihtoehtoja on hyvin paljon. Selkeästi viallinen laite voidaan korjata tai toinen sitten todetaan, ettei korjaus ole mahdollinen. Sen vuoksi nämä laitteet on jätetty pois kuviosta 14.



Kuvio 14: Huoltoon tulleiden laitteiden vika-analyysiä.

Kuviossa 14 on eroteltu eri vaihtoehtoja havaitun vikakäyttäytymisen jälkeen. Vika johtaa analysointiin onko se sisäinen vai ulkoinen ongelma. Sisäiset ongelmat voivat liittyä konfiguraatioon, ohjelmisto- tai laitteistovikaan. Konfiguraatio-ongelmat voivat näkyä laitteissa NFF-vikakoodina, koska laitteesta ei aina voida päätellä minkälaisessa konfiguraatiossa se on ollut mukana. Ohjelmisto- ja laitteistoviat voivat molemmat näkyä tunnettuina tai tuntemattomina vikoina ja näkyvinä tai ei-näkyvinä oireina. Tunnetut viat ja oireet voidaan identifioida ja korjata, näille voidaan tehdä myös päivitys tai korvata uudella osalla laitteella. Tuntemattomat ja ei-oireita olevat laitteet voidaan päivittää tai korvata uudella osalla tai laitteella. Nämä laitteet voivat saada myös NFF-vikakoodin. Ulkoisia ongelmia laitteelle voi syntyä tietoliikenneverkosta, asennuksesta tai sijoittamisesta. Nämä kaikki ulkoisista ongelmista aiheutuneet viat voidaan tulkita huollossa NFF:ksi.

5.2 Tutkimustulosten tarkastelu

Tutkimuksen aluksi on erittäin tärkeää, että ymmärretään mitä tietoja materiaalista on saatavilla. Reaalimaailman data on tyypillisesti häiriöitä sisältävää, sitä on paljon ja se on peräisin useista lähteistä. Sen vuoksi tutkimusmateriaali täytyy analysoida ja tutkia huolellisesti. Yleensä tiedonlouhinta aloitetaan raakadatan prosessoinnilla, jonka avulla saadaan parannettua datan ymmärrettävyyttä ja käytettävyyttä. Raakadataa voidaan prosessoida automaattisesti, puoliautomaattisesti tai manuaalisesti.

On myös erittäin tärkeää tuntea tietoliikenneverkkoa ja sitä miten se on rakennettu, jotta voidaan paremmin ymmärtää tutkimusmateriaalia. Tutkimusmateriaalista tutkitaan datan ominaisuuksia ja mitä arvoja data sisältää. Tämän tutkimuksen aineisto sisälsi vikaantuneiden laitteiden huollon raportit ja laitteista kerätyt hälytys-ja lokitiedot. Hälytys- ja lokitietojen lisäksi laitteista oli saatavilla yksityiskohtaisempia tiedot kuten esimerkiksi käytetty radioteknologia ja parametri, joka kertoo laitteen päälläoloajan. Vikatiedostot olivat eri formaateissa, vikaraportti oli Microsoft Excel-taulukko ja hälytys-ja lokitiedostot olivat tekstitiedostoina.

Tätä tutkimusta varten huollosta saatua vikatiedostoa voidaan pitää jo valmiiksi prosessoituna, koska tiedostoon on kerätty vain vikaantuneiden laitteiden tietoja. Datan vähentämistä on toteutettu myös siinä, että on valittu kolmen laitetyypin vikatiedot puolen vuoden ajalta. Koska tiedostot sisältävät vikaantuneiden laitteiden tiedot, voidaan tiedostoja käsitellä anomaaleina verrattuna koko tietoliikenneverkon tukiasemien laitteiden toimintaan. Näissä laitteissa on poikkeavia ominaisuuksia verrattuna normaalisti toimiviin laitteisiin.

Aloitin vikatiedoston tutkimisen käymällä sitä läpi manuaalisesti. Tutkimuksen kannalta mielenkiintoisin kohde oli NFF-vikakoodin saaneet laitteet. Tarkemmin tutkimusaineistoa analysoitaessa huomattiin, että datassa oli puutteita. Vikaantuneiden laitteiden vertailua haittasi se, ettei laitteista ollut saatavilla samaa määrää tietoa. Kaikista laitteista ei ollut saatavilla tarkempia hälytys- ja lokitiedostoja, vaan useista laitteista oli vain huollon vikaraportit.

Analysoinnin kannalta olisi tärkeää, että laitteista voidaan määrittää samat tutkittavat parametrit, joita voidaan vertailla. Nyt tietojen puutteen vuoksi mahdollisia tutkittavia parametrejä oli hankala määrittää. Tämä johti siihen, että tutkittavien laitteiden määrä väheni huomattavasti. Eräs hankaluus oli asiakkaalta saadut vikakuvaukset, jotka tallennetaan manuaalisesti. Laitteiden vikakuvaukset olivat epätarkkoja ja hyvin samankaltaisia, ja joidenkin laitteiden kohdalla vikakuvaus näytti olevan täsmälleen sama.

Tutkimusmateriaalia analysoitaessa kävi ilmi, että laitteistosta olisi tarvittu myös historiatietoja. Vianselvityksessä olisi tarpeellinen tieto missä tilanteessa vika tuli esiin ja minkälaisessa

konfiguraatiossa laite oli kiinnitettynä. Kuten aikaisemmin todettiin konfiguraatio ja parametrien määritys ovat hyvin merkityksellisiä tukiaseman käyttöönotossa. Parametreja määritellään useita, ja tukiaseman toiminnan voi estää pienikin virhe. Tämän vuoksi lisätiedot laitteiden asennuksista ja konfiguraatioista olisi erittäin tärkeitä tietoja vika-analyysin kannalta.

Tiedonlouhinnan soveltuvuudesta vikatiedostojen analysoimiseen voidaan sanoa, että tiedonlouhinnan menetelmät soveltuvat tämän tutkimuksen datan puhdistukseen ja datan pelkistämiseen. Tiedonlouhintaprosessia voidaan jatkaa, kun saadaan tutkimusmateriaalin puutteet korjatuiksi.

5.3 Tietojärjestelmien suunnittelututkimuksen syklit

Tässä tutkimuksessa selvitettiin myös tiedonlouhintatekniikoiden integroimista Hevnerin (2007) tietojärjestelmien suunnittelututkimuksen sykleihin. Ensimmäinen sykli eli merkitysykli käsittelee ympäristöä. Se yhdistää tutkimusprojektiin liittyvän ympäristön suunnittelutieteen toimintaan. Ympäristöön kuuluu sovellusalueen ihmisiä, erilaisia järjestelmiä ja ongelmia sekä mahdollisuuksia. Merkityssyklin vaatimukset ja kenttätestit ovat vuorovaikutuksessa suunnittelusyklin kanssa. Suunnittelusykli on tietojärjestelmien suunnittelututkimuksen projektin sydän, jossa arvioidaan ja rakennetaan suunnittelun artifakteja ja prosesseja.

Työssä perehdyttiin ensimmäiseen sykliin, koska tutkimusmateriaalin analysointia ei voitu jatkaa esikäsittelyä pidemmälle. Tiedonlouhinnassa lähtökohta on ongelman selvittäminen ja todellinen tutkimusympäristö. Tiedonlouhinnassa syötetieto voi olla käsite, tapaus tai ominaisuus. Ongelmien ratkaisemiseksi analysoidaan tietokannoista löytyvää dataa. Tiedonlouhinta on ensisijaisesti vuorovaikutteinen ja iteroiva prosessi, joihin sisältyy lukuisia vaiheita analyysista tulkintaan ja tulosten hyödyntämiseen.

Tämän tutkimuksen tukiasemien vikaraportit sopivat tietojärjestelmien suunnittelututkimuksen kenttätutkimuksiksi ja vaatimukset määritellään vikaraportista saaduista parametreistä, joiden avulla laitteiden välisiä eroja voidaan vertailla. Ensimmäisen syklin alueeseen soveltuu hyvin KDD-prosessin alkuvaihe, joka aloitetaan rajaamalla tarkasti mitä sovellusalueelta halutaan tutkia. Jotta voidaan määritellä tietämyksen vaatimukset, tarvitaan ennakkotietoa sovellusalueesta. Tiedonlouhintaprosessi on osa KDD-prosessia.

5.4 Tutkimustulosten vertailua

Tällä tutkimuksella oli yhteneväisyyttä aikaisempiin tutkimuksiin lentokone- ja elektroniikkateollisuudessa (Beniaminy & Joseph 2002; Block ym. 2009; Jones & Hayes 2001) siinä, että testausolosuhteissa ei välttämättä pystytä toistamaan kentällä tai käytön aikana esiintyneitä virheitä. Tämä johtuu siitä, että testauksessa on vaikea saada vastaavat olosuhteet. Havaittiin myös että viat esiintyvät epäsäännöllisesti, johtuen esimerkiksi epäpuhtauksista laitteen elektroniikkakomponenteissa. Huollossa komponentit puhdistetaan ennen testausta, jolloin tilanne on erilainen kuin kentällä. Yksi syy laitteiden NFF-luokitukselle voivat olla inhimilliset virheet tai kentältä saatu virheilmoitusten laatu.

Testausmenetelmiä on vaikea toteuttaa niin, että tarkistettaisiin kaikki testatun laitteen toiminnot ja niihin liittyvät vikavaihtoehdot. On myös hankala arvioida testausmenetelmien todellista kattavuutta. Laitteet ovat monimutkaisia ja sisältävät useita sekä laitteisto- että ohjelmistoteknologioita. Yleisenä ongelmana NFF-laitteisiin liittyy se, että nämä huollossa käyneet laitteet palautuvat varastoon ja uudelleen käytön yhteydessä laitteissa ilmenee samoja vikaoireita kuin aikaisemmin. NFF-ongelma on monimutkainen ja sen vuoksi ongelmaan ei ole vain yhtä ratkaisua vaan parannuksia pitää tehdä useilla eri alueilla. Olisi hyvä jos laitteiden vikaantumisesta saataisiin tarkempaa tietoa jo aikaisemmassa vaiheessa. Taulukossa 2 on esitetty yhteenvetoa yllä esitetyistä syistä, joiden tuloksena laite on saanut NFF-määrityksen.

Käytössä havaitut viat	**Laite määritellään NFF testauksessa**
Vikaantuminen käytön aikana	Testausolosuhteet eivät ole samat
Epäpuhtaudet mm. piirilevyjen pinnalla	Laite puhdistetaan ennen testejä, siitä seuraa ettei vikaa löydy testeissä
Laitteiden asennus, inhimilliset virheet tai konfiguraatiovirhe	Kun laite on irroitettu ympäristöstä, ei vika tule esille laitetta testattessa.
Vika laitteessa	Testitilanteessa ei voida testata kaikkia mahdollisia vikatilanteita
Vika laitteessa	Testien kattavuus, laitteistoviat ovat helpommin havaittavissa testauksessa. Ohjelmistoviat voivat jäädä piiloon.

Taulukko 2: Yhteenveto testauksessa NFF-määrityksen saaneista laitteista.

Taulukossa 2 on listattu esimerkkejä laitteiden vikatilanteista, ja syitä miksi näissä tilanteissa laitteista ei ole löytynyt vikaa huollon testeissä.

6 Keskustelu

Tiedonlouhinnassa on tärkeää ensin huolellisesti selvittää mitä tietoa materiaalista on saatavilla ja mitä siitä halutaan selvittää. Tässä tutkimuksessa selvitettiin miten tiedonlouhinnan prosessit sopivat huollosta saatujen tietoliikenne-elementtien vikatiedostojen analysoimiseen. Suuren tietomäärän analysointi tilastollisten menetelmien avulla ei ole yksinkertaista, se vaatii alkuperäisten tiedostojen muokkaamista tiedonlouhintaprosessin mukaisesti. Tutkittavan datan laadulla on suuri merkitys, analysoitaessa huonolaatuista dataa saadaan huonolaatuisia lopputuloksia. Datan puhdistus ja pelkistäminen vaikuttavat millaisia tuloksia analysoinnista saadaan. Isoa tietomäärää voidaan analysoida ja sitä visualisoimalla siitä on helpompi hahmottaa kokonaisuuksia. Analyysiprosessin luotettavuuden parantamiseksi havainnot, jotka ovat ”virheellisiä” pitää poistaa tutkimusaineistosta.

6.1 Johtopäätökset

Koska tutkimusmateriaalia oli paljon, aloitettiin tutkimus analysoimalla NFF-vikakoodin saaneiden laitteiden vikatiedostoja. Tutkimuksessa havaittiin, että vikatiedostojen sisältämää dataa on hyvin vaikea analysoida, koska asiakkaalta kerätyt vikakuvaukset eivät olleet kovin selkeitä. Operaattoreiden manuaalisesti tallennetut vikakuvaukset sisältävät paljon tulkinnanvaraista tietoa. Yksi syy ettei datan laatu ollut riittävän hyvää tiedonlouhinnan jatkotutkimuksiin, oli se että tutkimusdataa kerätään useasta eri paikasta ja osa kirjataan manuaalisesti.

Vikakuvauksia tarkastellessa huomattiin, että osa kuvauksista on täsmälleen samoja. Tällöin vikakuvaus ei kuvaa yksilöllisesti laitteen vikaa, jos se on kopioitu toisen laitteen vikakuvauksesta. Tarkempien vika-analyysien saamiseksi laitteiden historia- ja asennustiedot olisi tärkeä saada huoltoon laitteen mukana. Tällöin olisi mahdollista saada tietoa laitteen tilasta ennen kuin se on irrotettu mastosta tai muusta kiinnitystelineestä. Laajemmat tiedot juuri vikaantumishetkellä tulleista hälyistä ja lokeista auttaisivat vianselvityksessä.

Vikaraporteista voidaan sanoa että laitteiden vikakoodit ovat selkeitä ja ne on dokumentoitu tarkasti. Vikakoodien päivitys korjausten edetessä antaa korjaustoimenpiteistä hyvän kuvan. Tarkastelussa ollut kuuden kuukauden ajanjakso ei ole riittävä kattavaan tiedonlouhintaan vikaantuneista laitteista, vaan ajanjakso pitää olla pidempi. Pidemmän ajanjakson aikana voidaan seurata paremmin laitteiden kiertokulkua huoltoon. Tutkimukset on tehty yrityksen sisäisen tutkimusprojektin tuloksien perusteella, sen takia kaikkia tutkimustuloksia ei voida julkaista.

Tiedonlouhintaprosessit sopivat huollosta saatujen vikatiedostojen analysointiin, koska niiden avulla datasta pystytään selvittämään mitä tietoja on saatavilla ja myös datan puutteet tulevat selville. Vikatiedostojen muokkauksessa voidaan käyttää tiedonlouhintaprosessin menetelmiä, kun datan esikäsittelyssä havaitut puutteet saadaan korjatuiksi. Niiden avulla voidaan etsiä esimerkiksi jos laitteissa esiintyy systemaattisia virheitä. Lisäksi havaittiin että tiedonlouhintamenetelmiä voidaan soveltaa tietojärjestelmien suunnittelututkimuksen sykleihin. Tässä tutkimuksessa perehdyttiin ensimmäiseen sykliin, mutta tiedonlouhintaprosessin edistyessä voitaisiin soveltaa prosessia myös muihin sykleihin.

6.2 Tutkimuksen luotettavuus ja validiteetti

Tutkimusten luotettavuuden arvioinnissa keskeisiä käytettyjä mittareita ovat reliabiliteetti ja validiteetti. Laadullisten tutkimusten luotettavuutta ja perusteltavuutta on laajasti käsitellyt muun muassa Yin (2014), Miles & Huberman (1994) ja Gerring (2007). Validiteetilla ilmaistaan kuinka hyvin tutkimuksessa käytetty mittaus- tai tutkimusmenetelmä mittaa tutkittavana olevan ilmiön ominaisuutta. Validiteetilla voidaan tutkia väitteen pätevyyttä, eli tukeeko käytetty aineisto, tutkimusmenetelmät ja saadut tulokset esitettyjä väitteitä. Käytetyn mittaus- tai tutkimusmenetelmän luotettavuus ja toistettavuus, millä haluttua ilmiötä mitataan, sekä johtuuko tutkimustulos vain sattumasta vai kyetäänkö tulokset riippumattomasti toistamaan, ilmaistaan reliabiliteetilla. (Yin 2014; Gerring 2007.)

Miles ym. (2014) mukaan aineiston analysointi koostuu tietojen tiivistämisestä, valintaprosessista, keskittymistä sekä yksinkertaistamisesta. Tutkimuksen luotettavuutta voidaan lisätä triangulaation avulla, esimerkiksi tutkimalla ilmiötä useista eri näkökulmista, käyttämällä useita erilaisia aineistoja ja tiedonkeruumenetelmiä. (Patton 2002; Gerring 2007; Miles ym. 2014; Yin 2014.) Tähän tutkimukseen kerättiin materiaalia useilta eri tahoilta kentältä, dokumenteista ja muusta empiirisestä materiaalista. Lisäksi tutkimustuloksia sovellettiin tiedonlouhintaprosessiin ja integroitiin tiedonlouhintaprosessia tietojärjestelmien suunnittelututkimukseen. Mielestäni tutkimuksen laatuvaatimukset täytyivät, koska laitteiden vikaantumisprosessia tarkasteltiin useista näkökulmista.

Voidaan sanoa että data on hyvä laatuista, jos se täyttää sen käyttötarkoituksen vaatimukset. On monia tekijöitä, joista muodostuu datan laatu, kuten tarkkuus, täydellisyys, johdonmukaisuus, ajanmukaisuus, uskottavuus ja tulkittavuus. (Han ym. 2001.) Tutkimuksen data täytti suurelta osalta nämä laatukriteerit, koska vikakuvaukset olivat ajanmukaisia, vikaraporttien vikakoodit antoivat tarkan tiedon laitteen tilasta. Tutkimusdataa voidaan pitää johdonmukaisena, koska testausmenetelmät ovat kaikille laitteille samat. Tulkittavuutta datassa on paljon. Vikaanalyysien luotettavuutta parantaisi, jos vikaantuneesta laitteesta olisi saatavilla hälytiedot

vikaantumishetkellä. Viantutkimusmenetelmiä olisi tärkeää parantaa, jotta voitaisiin aikaisemmassa vaiheessa löytää syy vikaantumiseen. Myös vikaantuneen laitteen tarkempi analysointi paikanpäällä vähentäisi turhia huoltokustannuksia.

Ulkoiseen validiteettiin liittyy myös tutkimuksen tarkastelu suhteessa muihin samankaltaisiin tutkimuksiin (Mitchell & Jolley 2001). Yksi vertaishanke tälle työlle on Euroopan merenkulun valvontaviranomaisten yhteistyön lisäämiseksi perustettu H2020 -ehdotus MARISA (Maritime Integrated Surveillance Awareness), joka liittyy SEC-19-BES-2016 hakuun ja teemaan. MARISA –hankkeessa hyödynnettäisiin laajaa tietojen saatavuutta, joka on käytettävissä useilta merenkulkuun liittyviltä viranomaisilta muun muassa PERSEUS ja EU_CISE_2020. MARISA -hankesuunnitelma sisältää moduulitason kuvauksia ulkoisten tietolähteiden keräämiseksi, ja siinä toteutettaisiin käyttöjärjestelmien tuotteita. Lisäksi big data infrastruktuuria käytettäisiin järjestämään ja hyödyntämään kaikkea sisääntulevaa dataa. SEC-19-BES-2016 hakutekstin perusteella tavoitteena on suurten eri lähteistä saatavien tietomäärien yhdistäminen ja työkalujen kehittäminen näiden tietojen fuusiointiin ja hyödyntämiseen päätöksenteossa ja tilannekuvan rakentamisessa.

Tutkimuksen validiteettia paransi myös tutkimuksessa mukana olleet työnohjaajat, joilla on pitkä kokemus tietoliikenteen luotettavuuden tutkimisessa. Heidän väitöskirjan aiheet olivat Salmela: ”Reliability Assessment of Telecommunications Equipment” ja Hätönen: ”Data mining for telecommunications network log analysis” (Salmela 2005; Hätönen 2009).

6.3 Datan samanlaisuuden ja erilaisuuden mittaaminen

Tiedonlouhinnan sovelluksissa, kuten klusterointi ja poikkeavuuksien etsintä, tarvitaan tapoja määrittää kuinka todennäköisiä tai epätodennäköisiä objektit ovat verrattuna toisiinsa. Tässä kappaleessa on esitetty yksi tiedonlouhintamenetelmä laitteiden eroavaisuuksia määrittämiselle binääriattribuuttien avulla.

Datan erilaisuuden mittaamien on esitetty Han ym. (2001) kirjassa. Datajoukot muodostuvat dataobjekteista. Dataobjektit edustavat kokonaisuutta, esimerkiksi myyntitietokannassa se voi olla asiakkaat tai lääketieteellisessä tietokannassa objekti voi olla potilaat ja yliopiston tietokannassa ne voivat olla opiskelijoita, professoreita, ja kursseja. Dataobjektit kuvaavat tyypillisesti ominaisuuksia. Tietokannassa rivit kuvaavat dataobjekteja ja sarakkeet vastaavat ominaisuuksia.

On useita tapoja mitata objektien samankaltaisuutta tai erilaisuutta, jolla viitataan läheisyyden mittaukseksi. Samankaltaisuuden mittauksessa kahdelle objektille palautetaan tyypillisesti

arvo 0 jos objekti on eroava. Yleisesti arvo 1 osoittaa täydellistä samankaltaisuutta. Erilaisuuden mittauksessa arvo toimivat päinvastoin, eli siinä arvo 0 kuvaa objektien olevan samat ja näin ollen kaukana erilaisuudesta. (Han ja muut, 2001.)

Eräs mittaustapa on läheisyysmittaus binääriattribuuteille, jossa saa joko arvon 0 tai 1. Nolla tarkoittaa että attribuutti on poissa ja yksi että se on läsnä. Yksi tapa on merkitä että 1 tarkoittaa samaa kuin Y (yes) ja 0 tarkoittaa N (no). Vikatiedostossa Excelin rivi kuvaa yksittäisen laitteen vikatiedostoa, ja sarakkeissa on esitetty eri ominaisuuksia. Näiden perusteella voidaan taulukossa 3 esittää esimerkin omaisesti laitteiden vertailua. Taulukossa 3 on esitetty mahdollinen luokittelutapa käyttäen vikatietoja binääriattribuutteina. Han ja muut (2001) esittää kaavan, jolla erovaisuutta kahden objektin i ja j välillä voidaan mitata:

$$d(i,j) = \frac{r+s}{q+r+s}$$

missä r = attribuuttien määrä, joka saa arvon 1 objektille i mutta arvon 0 objektille
s = attribuuttien määrä, joka saa arvon 0 objektille i mutta arvon 1 objektille j
q = attribuuttien määrä, joka saa arvon 1 molemmille objekteille i ja j

Laite ID	NFF	1. huoltokerta	hälytykset	lokitiedot
A	Y	Y	N	N
B	Y	N	Y	Y
C	N	N	Y	Y
D	N	Y	N	N
E	Y	Y	N	N
…	…	…	…	…

Taulukko 3: Esimerkki miten laitteiden eroavaisuuksia voidaan määritellä binääriattribuuttien avulla.

Taulukossa 3 on laitteiden verrattaviksi ominaisuuksiksi valittu onko laite saanut NFF määrityksen, onko laitteen 1. huoltokerta, löytyykö siitä hälytiedostot ja lokitiedot. Jos laite on saanut NFF tilan tai sillä on 1. huoltokerta se saa arvon 1 (yes). Myös jos laitteesta on saatavilla hälytys- ja lokitiedostot se saa arvon 1 (yes). Jos näitä ei ominaisuuksia ei esiinny laite saa arvon 0 (no).

Laitteiden eroavaisuutta voidaan mitata:

$$d(A,B) = \frac{1+2}{1+1+2} = 0{,}75$$

$$d(B,C) = \frac{1+0}{2+1+0} = 0{,}33$$

$$d(C,D) = \frac{2+1}{0+2+1} = 1$$

Mittausten mukaan laitteilla C ja D olisi vähiten samankaltaisia ominaisuuksia, koska tulos on yksi. Samoja ominaisuuksia olisi laitteilla B ja C, koska tulos on lähellä nollaa.

6.4 Jatkotutkimusaiheet

Jatkotutkimuksia varten laitteiden vikakuvauksista pitäisi saada tarkempaa tietoa, erityisen tärkeää olisi saada lisätietoa laitteessa esiintyvistä hälytyksistä ja lokeista vikaantumishetkellä. Nämä tukisivat hyvin laitteista saatavaa tietoa, koska laitteiden vikakoodit ja niiden selitykset ovat jo saatavilla tällä hetkellä. Mielenkiintoista olisi verrata vikakoodeja ja -kuvauksia esimerkiksi ohjelmistoviallisten laitteiden, jotka ovat korjattu, ja NFF-laitteiden kesken. Ohjelmistovikojen analysoimiseen voisi keskittyä sen vuoksi, että ohjelmistoviat voivat usein peittää laitteistovikoja. Jatkotutkimuksia varten olisi laitteista saatavilla tarkempaa tietoa, koska laitteista kerätyt hälytys- ja lokitiedostot kehittyvät koko ajan.

Parannettavaa olisi loki- ja hälytystietojen tallennuksessa, koska yleisenä korjaustoimenpiteenä yritetään vikaantunutta laitetta ensin saada toimimaan resetoimalla se paikan päällä. Tämä aiheuttaa sen, että viimeisimmät hälytiedot ovat järjestelmäkäynnistyksiä ja näin ollen voivat peittää vikatilanteesta tulleet hälytykset. Hälytystiedostojen laajempi tarkastelu olisi tarpeellista, jotta saataisiin parempi kuva missä tilanteessa tiettyjä hälytyksiä on esiintynyt.

Tiedonlouhintamenetelmän datan erilaisuuden mittaamisen avulla voisi luokitella laitteiden ominaisuuksia. Näin voitaisiin saada enemmän tietoa laitteissa esiintyvistä eroista. Voi kuitenkin olla, että laitteiden kaikkia tutkittavia ominaisuuksia on hankala luokitella binääriattribuuttien avulla.

Lähteet

Beniaminy, I., Joseph, D. 2002. Reducing the "No Fault Found" Problem: Contribution from Expert-System Methods. Viitattu 31.08.2015. http://ieeexplore.ieee.org/document/1036138/

Block, J., Tyrberg, T., Söderholm, P. 2009. No fault found events during the operational life of military aircraft items. Viitattu 31.08.2015. http://ieeexplore.ieee.org/abstract/document/5269968/

Campbell, D. T. ja Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81—105.

Dahlman, E. Parkvall,S. ja Sköld, J. 2011. 4G LTE/LTE-advanced for mobile broadband.

Denzin, N. ja Lincoln, Y. 1994. Introduction: Entering the Field of Qualitative Research. In Denzin Norman and Lincoln Yvonna (eds.). Handbook of Qualitative Research. Sage Publications, 1—17.

Elisan matkapuhelin- ja mobiililaajakaistaverkon nopeudet. Viitattu 18.11.2016 https://elisa.fi/asiakaspalvelu/aihe/matkapuhelinliittymat/ohje/verkon-nopeudet/

Elisa testasi 5G:ta ensimmäisenä operaattorina Suomessa. Viitattu 18.11.2016 https://palsta.elisa.fi/elisan-tiedotteet-2/elisa-testasi-5gta-ensimmaisena-operaattorina-suomessa-501677

Eskin, E. 2000. Anomaly Detection over Noisy Data using Learned Probability Distributions. Proceedings of the Seventeenth International Conference on Machine Learning.

Fayyad, U., Piatetsky-Shapiro, G. ja Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. Viitattu 25.05.2016. https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Gajic, B., Novaczki, S. ja Mwanje, S. 2015. An Improved Anomaly Detection in Mobile Networks by Using Incremental Time-aware Clustering. IFIP/IEEE IM Workshop: 1st Workshop on Cognitive Network & Service Management.

Gerring, J. 2007. Case study research : Principles and practices. New York: Cambridge University Press.

Gray, D. E. 2004. Doing research in the real world. SAGE Publications.

Gregor, S. ja Jones, D. 2007. Journal of the Association for Information Systems

Han, J., Kamber, M., Pei, J. 2001. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers.

Hevner, A. 2007. A Three Cycle View of Design Science Research," Scandinavian Journal of Information Systems: Vol. 19.

Hevner, A. & Chatterjee, S. 2010. Design Research in Information Systems.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. Design science in information systems research. MIS Quarterly.

Hätönen, K. 2009. Data mining for telecommunications network log analysis. Väitöskirja Helsinki.

Höglund, A. 2006. Advanced mobile network monitoring and automated optimization methods. Väitöskirja. Espoo.

Höglund, A. J., Hätönen, K. ja Sorvari, A. S. 2000. A computer host-based user anomaly detction system using the self-organizing map. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN). 411416.

Iivari, J. 2007. A paradigmatic analysis of information systems as a design science. Scandinavian Journal of Information Systems, 19(2), 39—64.

Jain, A. K ja Dubes, C. R. 1988. Algorithms for Clustering Data, New Jersey: Prentice Hall PTR

James, I., Lumbard, D., Ian, W., Goble, J. 2003 "Investigating No Fault Found in the Aerospace Industry", Proceedings Annual Reliability and Maintainability Symposium.

Jerome, R. B. 2015. Pre-processing techniques for anomaly detection in telecommunication networks. Pro gradu tutkielma Aalto Yliopisto. Viitattu 07.02.2016 https://aaltodoc.aalto.fi/bitstream/handle/123456789/16240/master_Babujee%20Jerome_Robin_2015.pdf?sequence=2

Jones. J, Hayes. J. 2001. Investigation of the Occurrence of: No-Faults-Found in Electronic Equipment. viitattu 31.08.2015. http://ieeexplore.ieee.org/abstract/document/974126/

Järvinen, M. 2013. Mobiiliverkkojen kehitys. Opinnäytetyö. Tampereen ammattikorkeakoulu. Viitattu 29.08.2016. https://publications.theseus.fi/bitstream/handle/10024/68899/Jarvinen_Mika.pdf?sequence=2

Kumpulainen, P. 2014. Anomaly Detection for Communication Network Monitoring Applications. Väitöskirja Tampere. Viitattu 01.03.2016 https://dspace.cc.tut.fi/dpub/handle/123456789/22104

Kumpulainen, P., Hätönen, K. 2008. Anomaly detection algorithm test bench for mobile network management. MathWorks/MATLAB User Conference Nordic. The MathWorks Conference Proceedings. Viitattu 01.03.2016. https://tutcris.tut.fi/portal/fi/publications/anomaly-detection-algorithm-test-bench-for-mobile-network-management%28ba218189-5a6f-4a54-8cc8-099c2c787d5b%29.html

Laiho, J., Raivio, K., Lehtimäki, P., Hätönen, K. ja Simula, O. 2005. Advanced Analysis Methods for 3G Cellular Networks. IEEE Transactions on wireless communications. 930—942.

Lumme, V. 2012 Intelligent Interpretation of Machine Condition Data. Väitöskirja Tampere viitattu 31.08.2015. http://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/21256/lumme.pdf?sequence=3&isAllowed=y

March, S. T., ja Smith, G. F. 1995. Design and natural science research on information technology. Decision Support Systems.

Miles, M.B., Huberman, A. M. ja Saldana, J. 2014. Qualitative Data Analysis, A Methods Sourcebook. 3rd ed.

Miles, M. B. ja Huberman, A. M. 1994. Qualitative data analysis : An expanded sourcebook 2nd ed. Sage.

Mitchell M. L. ja Jolley J. M. 2001. Research Design Explained. New York

Määttänen, T. 2015. Mobiiliverkon tukiaseman rakentaminen https://www.theseus.fi/bitstream/handle/10024/93391/Maattanen_Tommi.pdf?sequence=1

Nunamaker, J. R., Chen, M., and Purdin, T. 1991. Systems development in IS research MIS Quarterly, 7(3), 89—106.

Nunamaker, J. R., ja Briggs, R. O. 2011. Toward a broader vision for information systems. ACM Transactions on Management Information Systems.

Nurminen, M. 2003. Tiedonlouhinta rakenteisista dokumenteista. Seminaarityö Jyväskylä. Viitattu 07.03.2016. http://users.jyu.fi/~minurmin/opiskelu/gradusem/semma.pdf

Nurminen, M. 2005. Tiedonlouhinta rakenteisista dokumenteista, pro gradu -tutkielma, Jyväskylän yliopisto. Viitattu 26.03.2016. https://jyx.jyu.fi/dspace/handle/123456789/12507

Orlikowski, W. J. ja Baroudi, J. J. 1991. Studying Information Technology in Organizations: Research Approaches and Assumptions.The Institute of Management Sciences.

Palat, S. ja Godin, P. 2009. The UMTS Long Term Evolution: From Theory to Practice. Wiley.

Patton, M. 2002. Qualitative evaluation and research methods. 3nd ed. Sage.

Salmela, O. 2005. Reliability Assessment of Telecommunications Equipment. Väitöskirja. Espoo.

Santoro, M. 2008. "New Methodologies for eliminating no trouble found, no fault found and other non repeatable failures in depot settings. IEEE Autotestcon.

Simon, H. 1996. The sciences of the artificial. 3rd ed. USA: MIT Press.

Talonen, J. 2015. Advances in Methods of Anomaly Detection and Visualization of Multivariate Data. Väitöskirja Aalto Yliopisto. Viitattu 07.03.2016 https://aalto-doc.aalto.fi/bitstream/handle/123456789/15255/isbn9789526061122.pdf?sequence=1&isAllowed=y

Tervonen, U. 2002. WCDMA-tukiaseman vikadiagnostiikka ja testaus. Opinnäytetyö. Kajaanin ammattikorkeakoulu. Viitattu 05.11.2016. https://www.theseus.fi/bitstream/handle/10024/10901/KAT9TUlla-RiittaT.pdf?sequence=1

van Aken, J. E. 2004. Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. Journal of Management Studies.

Vehviläinen, P., Hätönen, K., Kumpulainen, P. 2003. Data Mining in Quality Analysis of Digital Mobile Telecommunications Network.In *Proceedings of XVII IMEKO World Congress*, Dubrovnik, Croatia, June 2003. IMEKO.

Walsham, G. 2006. Doing interpretive research. European Journal of Information Systems, (15), 320-330.

Walsham, G. 1995. The emergence of interpretivism in IS research. Institute for Operations Research and the Management Sciences, 376—394.

Witten, I. H. & Frank, E. 2005 Data Mining. Practical Machine Learning Tools and Techniques. USA.

Yamanishi, K. ja Takeuchi, J. 2001. Discovering outlier filtering rules from unlabeled data. Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining.

Yin, R.K. 2014. Case study research: design and methods. 5th ed.

Zhang, J. ja Zulkernine, M. 2006. Anomaly based network intrusion detection with unsupervised outlier detection. IEEE International Conference on Communications.

Zins, C. 2007. Journal of the American Society for Information Science and Technology. Wiley.

Äyrämö, S. 2006. Knowledge Mining using Robust Clustering. Jyväskylän Yliopisto.

Julkaisemattomat lähteet:

Niemi, A. 2014. Black_box_&_maintenance_data_analysis. Sisäinen dokumentti.

Niemi, A. 2014. New Findings from Maintenance Data. Sisäinen dokumentti.

Niemi, A. 2014. NSN Black Box Data. Sisäinen dokumentti.

Sjöblom, J. 2015. Tapaustutkimus: Miten radio- ja systeemimoduulien vikaantumista voidaan ymmärtää? Espoo

Kuviot

Kuvio 1: Tietoliikenneverkon nopeuksien kehitys. 9
Kuvio 2: Putkimastoon kiinnitettyjä radioita. 10
Kuvio 3: Keskusyksikön kytkennät ylhäältä alas 4G, 3G sekä 2G laitteet. 11
Kuvio 4: Tietojärjestelmien tutkimuksen viitekehys. 18
Kuvio 5: Tietojärjestelmien suunnittelututkimuksen syklit. 20
Kuvio 6: Tiedonlouhinta omaksuu tekniikoita monilta aloilta. 24
Kuvio 7: Yleiskuva vaiheista, jotka muodostavat KDD-prosessin. 26
Kuvio 8: KDD-prosessi. 27
Kuvio 9: Tiedonlouhintaprosessi. 28
Kuvio 10: Datan esikäsittelyprosessit. 30
Kuvio 11: Esimerkki klustereista ja ulkopuolisista havainnoista (outliers). 33
Kuvio 12: Tukiasemahälytysten luokittelu verrattuna Tervoseen. 34
Kuvio 13: Esimerkki vikaantuneiden elementtien kuljetusketjusta. 35
Kuvio 14: Huoltoon tulleiden laitteiden vika-analyysiä. 36

Taulukot

Taulukko1: Tutkimuksen viitekehys 17
Taulukko 2: Yhteenveto testauksessa NFF-määrityksen saaneista laitteista. 39
Taulukko 3: Esimerkki miten laitteiden eroavaisuuksia voidaan määritellä binääriattribuuttien avulla..... 43

Liitteet

Liite 1: Research attributes. 52

Liite 1: Research attributes.

Title of study	Tiedonlouhintaprosessien soveltuvuus tietoliikenne-elementtien vikatiedostojen analysoimiseen.
Research questions	Miten tiedonlouhintaprosesseja sovelletaan tietoliikenne-elementtien vikatiedostojen analysoimiseen?
Research agreement	The researcher is permitted to use the collected research data for this thesis.
Unit of analysis	Tiedonlouhintaprosessi (Data Mining Process), vianselvitystieto (the fault data) ja vian kuvaus (description of failure).
Importance of study	Transparent of troubleshooting. Better knowledge about the fault reasons and shorter repair time
Methodological focus	Design science research, analysis of fault data and Information Systems Research Framework.
Form of analysis	Qualitative and quantitative analysis.
Research Approach	Inductive research for fault information management.
Specification of constructs	Design Science Research, Data Mining, No Fault Found, Knowledge Discovery in Databases
Theoretical approaches	"Data Mining Preprocessing" an iterative process ensures that the analysis methods are able to extract the required information from the data.
Theoretical literature	Reference of the Data Mining Eskin 2000; Fayyad et al. 1996; Gajic et al. 2015; Han et al. 2001; Hätönen 2009; Höglund 2006; Höglund et al. 2000; Jain & Dubes 1988; Jerome 2015; Kumpulainen 2014; Kumpulainen & Hätönen 2008; Laiho et al. 2005; Lumme 2012; Nurminen 2003; Nurminen 2005; Salmela 2005; Talonen 2015; Vehviläinen et al. 2003; Witten & Frank 2005; Yamanishi & Takeuchi 2001; Zhang & Zulkernine 2006; Äyrämö 2006.
First research target	Faster troubleshooting helps to decrease the number of NFF units.
Outcome validation comparison	Occurrence of a failure without an identifiable root cause is known as a NFF. Beniaminy et al. 2002, studied Aircraft, Block et al. 2009 studied Military Aircraft Items, Jones et al. 2001 studied electronic equipment.
Research design	Design Science Research.
Logic of evidence	Collected data and analysis.
Methodological and analysis literature	Denzin & Lincoln 1994; Gerring 2007; Gray 2004; Gregor 2007; Hevner 2007; Hevner & Chatterjee 2010; Hevner et al. 2004; Iivari 2007; March & Smith 1995; Miles et al. 2014; Miles & Huberman 1994; Nunamaker et al. 1991; Nunamaker & Briggs 2011; Orlikowski & Baroudi 1991; Patton 2002; Simon 1996; van Aken 2004; Walsham 2006; Walsham 1995; Yin 2014.
Data collection methods	The data collection of this study was cumulative. Research data was collected between October 2015 and April 2016. Research data is field data of equipment, literature and discussions at work. Collected data includes definitions, requirements, fault codes, alarms, theory, customer faults.
Coding	Each document was read carefully and then then similar or identical answers were marked with certain color.
Notes	Researcher used notes from meetings.
Main results	Analyze faults and alarms of the No Fault Found (NFF) units.
Main implication	The suitability of data mining process to analyze fault files of the telecommunication elements.
Role description	Researcher as outsider (objective) and as a participant in project as insiders (subjective).