

Tallinna Ülikool
Informaatika Instituut

NÕUDED AUTOMATISEERITUD TEKSTIANALÜÜSILE KURITEOMUSTRITE TUVASTAMISEKS

Magistritöö

Autor: Kai Jääger

Juhendaja: PhD Katrin Niglas

MA Marilis Sepp

Autor: „2010

Juhendaja: „2010

Instituudi direktor: „2010

Autorideklaratsioon

Deklareerin, et käesolev magistritöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

(kuupäev)

(autor)

Sisukord

SISSEJUHATUS	4
Töö eesmärk	5
Magistritöö ülesehitus	5
1. KRIMINAALANALÜÜSI JA TEKSTIKAEVE LÄBIVIIMISE VÕIMALUSED.....	8
1.1. Kriminaalanalüüsi teooriad kuriteomustrite leidmiseks	9
1.2. Kriminaalanalüüsi läbiviimise meetodid	12
1.3. Tekstiliste andmete töötlemine	14
1.4. Ülevaade sarnastest uuringutest	20
1.5. Tekstianalüüsi läbiviimise metoodika ja analüüsimudeli kujundamine	21
2. ANALÜÜSIMUDELI RAKENDAMINE KAHTLUSTATAVA ÜTLUSTELE	29
2.1. Valdkonna mõistmine.....	29
2.2. Ülevaade andmetest, mida koguma hakatakse.....	30
2.3. Andmete ettevalmistamine	31
2.4. Mudeli loomine	32
2.5. Tulemuste hindamine ja rakendamine.....	42
3. NÕUDED AUTOMATISEERITUD TEKSTIANALÜÜSILE KURITEOMUSTRITE LEIDMISEKS	44
3.1. Automatiseeritud tekstianalüüsi eesmärk	44
3.2. Üldised nõuded automatiseeritud tekstianalüüsile	44
3.3. Jätkuarendused	47
4. KOKKUVÕTE	49
5. RESÜMEE.....	50
KASUTATUD KIRJANDUS	52
LISA 1 Kodeerimistabel.....	55
LISA 2 Karistusseadustiku paragrahv 200	56
LISA 3 Sisend valdkondlikku sõnastikku	57
LISA 4 Seosed graafiliselt.....	58

SISSEJUHATUS

Viimaste aastate tendentsid kuritegevusega seotud infomahtude suurenemise ja keerukamaks muutumise suunas, mida politseil on vaja töödelda, on suurendanud vajadust leida traditsiooniliste kuritegude lahendamise kõrvale täiendavaid nõ alternatiivsed vahendeid tõhustamiseks kuritegude lahendamist. Politseis registreeritud kuritegude andmed sisaldavad väärtuslikku informatsiooni, mille *modus operandi* (toimepanemise viis) alusel on võimalik siduda näiteks kurjategija poolt toimepandud erinevaid kuritegusid, siduda gruppide poolt toime pandud kuritegusid jne (Grover jt, 2006).

Täiendavaid vahendeid on vaja selleks, et tagada aina suuremate infomahtude juures analüüside efektiivsus (Chen jt, 2004). Õigeaegsus ja õiged andmed on politseis menetluste läbiviimisel ja tegevuste planeerimisel võtmetähtsusega. Kriminaalanalüüsi efektiivsus sõltub järjest enam sellest, kuidas suudetakse suurest info üleküllusest leida uuritava juhtumiga seotud informatsioon nii juba olemasolevast infost (sidumine toimunud kuritegudega) kui ka uuest tekkivast informatsioonist kasutades seostamist – loogilise struktuuri loomist. Näiteks kurjategijate ülekuulamisprotokollidest saadakse palju kasulikku informatsiooni – meetodid, kuritegude detailid, kuriteoliigid ajas ja kohas (Clarke jt, 2006).

Politseis kogutavad andmed jagunevad andmekandja järgi üldiselt kaheks: paberkandjal ja digitaalseteks andmeteks, millest viimase osakaal järjest kasvab. Digitaalsel kujul olevad andmed jagunevad omakorda andmebaasis struktureeritult hoitavateks ja struktureerimata (tekstilised) andmeteks (nt: politseile antavad ütlused). Väidetavalt koguni 80% avaliku sektori ja korrakaitseorganisatsiooni informatsioonist on tekstikujul (Kollepara jt, 2002). Hoolimata tekstikujul olevate andmete suurest osakaalust on siiani arvutitega läbiviidaval analüüsimisel keskendunud peamiselt struktureeritud andmetele, sest suhteliselt hiljutise ajani oli tekstianalüüsi võimalik teha vaid mahuka „käsitööna“ teksti kodeerimisel, mis on aga liiga aeganõudev ja vigaderohke.

Tänapäeval kasutatakse vahendeid, mis võimaldavad nii struktureeritud kui struktureerimata teksti analüüsida ja visualiseerida. Näiteks kasutatakse laialdaselt andmete töötlemiseks andmekaevet (*Data Mining*) ning teksti osas selle alavalikut tekstikaevet (*Text Mining*). Nende vahenditega on võimalik tõsta analüüsi kvaliteeti ja produktiivsust – andmete kaevandamise eesmärk on võimalikult vähese kuluga saada andmeallikatest vajalik informatsioon – eraldada oluline informatsioon ebaolulisest (Liiv, 2005). Tekstikaeve on selle võrra keerukam, et sellisel kujul olevate andmete puhul tuleb ettevalmistavaid tegevusi

(kirjavigade, trükivigade jms eemaldamine, teksti analüüsimine, lühendid, släng) kaevandamise etapi alustamiseks läbi viia rohkem.

Tekstilisel kujul olevate andmete säilitamise asemel võiks andmeid hoida struktureeritult – oleks lihtsam, kuna siis ei oleks vaja tegeleda eelpool toodud ettevalmistavate tegevustega. Kõiki andmed ei saa kunagi struktureeritud kujule viia või ei ole nende sellisel kujul kogumine otstarbekas. Andmete struktureeritud sisestamine aeglustab andmete talletamist (vaja teha palju valikuid etteantud kitsastest klassifikaatoritest). Lisaks ei suuda sellisel kujul kogutud andmed edasi anda objektide omavahelist täpsemat seotust, vaid talletada saab vaid üldised faktid (Schroeder jt, 2007). Struktureerimata kujul andmete hoidmisel aga ei lähe andmete seos kaduma ning lugeja suudab oma oskuste ja kogemuste põhjal hinnata, kas näiteks kaks sündmust läbivat isikut on omavahel tugevalt seotud või mitte.

Töö eesmärk

Kuritegevuse analüüs tugineb eeldusel, et kuriteod ei ole toime pandud juhuslikult, eraldiseisvate ja unikaalsete juhtumitena, vaid neid võib ühiste ja selgelt eristatavate tunnusoonte järgi koondada rühmadesse (nt: koha, aja, sihtmärgi või ohvritüübi järgi, mida pannakse toime teatud meetodeid (*modus operandi*) kasutades) (Ekblom, 1988) ning selle põhjal kirjeldada kuriteomustrid.

Käesolevas töös vaadeldakse kuriteomustrite kujunemist tekstilistel andmetel lähtuvalt kriminaalanalüüsi teooriatest ja meetoditest, tekstianalüüsi läbiviimisest ning andmekaeve põhimõtetest. Teooriate toel ja läbi praktilise rakendamise luuakse töö tulemusena üldised nõuded automatiseeritud tekstianalüüsile kuriteomustrite tuvastamiseks, mis oleksid eelduseks järgmiste tarkvaraarenduse etappide läbiviimiseks ja rakenduse loomiseks.

Näiteks USAs ja Ühendkuningriigis on juba kasutusel politseis kogutavate tekstiliste andmete töötlemine (vaata „Ülevaade sarnastest uuringutest“). Käesoleva töö uudsus seisneb selles, et eestikeelsete tekstide kohta ei ole Eesti Politseis varem automaatset tekstianalüüsi tehtud ning nii ei ole võimalik toetuda varasematele töödele selles valdkonnas.

Magistritöö ülesehitus

Uuringutüübina kasutatakse arendusuuringut (rakendust loov uurimus), mis koosneb viiest sammust, mis jagunevad magistritöö peatükkide vahel järgmiselt:

Arendusuuringu samm	Vastav magistritöö osa
Eeldatava probleemi/vajaduste tunnetamine	Töö eesmärk
Probleemi analüüs	Kriminaalanalüüsi ja tekstikaeve läbiviimise võimalused
Arendusprotsess	Tekstianalüüsi läbiviimise meetodika
Hindamine	Analüüsimudeli rakendamine kahtlustatava ütlustele
Järeldused ja üldistused	Nõuded automatiseeritud tekstianalüüsile kuriteomustrite leidmiseks

Eeldatava probleemi/vajaduste tunnetamine – vajadus automatiseeritud tekstianalüüsi järgi, kuna suurenenud andmemahtude juures ei rahulda käsitsi analüüsimine enam püstitatud vajadusi, sest palju informatsiooni on elektroonselt ja struktureerimata kujul. On vaja leida võimalus, kuidas need vabateksti kujul andmed viia struktureeritud kujule, millele oleks võimalik rakendada juba traditsioonilisi andmetöötluste ja -analüüsi meetodeid.

Probleemi analüüs – kirjeldab kriminaalanalüüsi läbiviimise lähtekohaks olevad teooriad ja meetodid. Lisaks tuuakse ära tekstiliste andmete töötlemise võimalused, mis aitavad probleemi lahendada – andmekaeve põhimõtete ja tehnikate kasutamine, kontentanalüüs ja automatiseeritud tekstianalüüsi keeletugi.

Arendusprotsess – käesoleva töö arenduseks on analüüsimudeli loomine eelnevalt toodud võimalusi ja piiranguid silmas pidades. Arendusprotsessi tulemuseks ei ole valmis rakendus, kuna käesoleva töö skoobiks ei ole rakenduse loomine, vaid probleemi sõnastamine ja eelanalüüs.

Hindamine – kirjeldatud analüüsimudeli rakendamine valimi põhjal, et anda esmane hinnang kirjeldatud arendusprotsessi eesmärgipärasusele ja täpsustada nõuded automatiseeritud tekstianalüüsile. Lõplik mudeli hindamine toimub käesoleva töö väliste protsessidena.

Järeldused ja üldistused – eelnevates sammudes läbiviidud tegevuste ja saadud tulemuste kohta järelduste kokkuvõtmine nõuetena, millele automatiseeritud tekstianalüüsi läbiviimine politseis peab hakkama vastama ning vajadusel muude soovitude sõnastamine järgmiste tarkvaraarenduse etappide läbiviimiseks.

Järgmistes peatükkides liigutakse teooriatest analüüsimudeli kirjeldamise juurde, peale mida toimub analüüsimudeli rakendamine. Viimases peatükis võetakse nõuetena kokku teooria ja rakendamise tulemused ning kirjeldatakse need kui automaatse tekstianalüüsi üldised nõuded kuriteomustrite tuvastamiseks. Antud töö pinnalt peab tekkima eeldus minna edasi automaatse süsteemi loomiseks, et muuta tekstianalüüsi läbiviimine poliseis automatiseerituks.

1. KRIMINAALANALÜÜSI JA TEKSTIKAEVE LÄBIVIIMISE VÕIMALUSED

Peatüki eesmärk on anda ülevaade kriminaalanalüüsi teooriate ja analüüsimeetodite rakendamisest lähtuvalt käesoleva töö eesmärgist leida kuriteomustreid tekstilistest andmetest. Lisaks antakse ülevaade tekstianalüüsi tehnikatest ja keeletehnoloogiast ning tuuakse näited mujal maailmas sarnase probleemi lahendamiseks. Peatüki lõpetab eelneval põhinev analüüsitudeli kirjeldus.

Kuritegevuse analüüsi keskne fookus on kuritegevus ja korrariikumised, probleemid ja informatsioon, mis on seotud juhtumite olemuse, toimepanijate, ohvrite või kuriteo objektidega. Kuigi distsipliini nimetatakse kuritegevuse analüüsiks, hõlmab see tegelikkuses palju enam kui lihtsalt kuriteosündmuste uurimist (Boba, 2005). Enim uuritakse sotsiaaldemograafiliste (nt: sugu, vanus), ruumiliste (nt: seosed sündmuste ja objektide vahel) ja ajaliste (nt: aasta, kuu, nädal) faktorite infot.

Läbiviidavate tegevuste eesmärk on luua efektiivne tegevusplaan, kasutades selleks, õiget informatsiooni ja meetodeid (Vellani ja Nahoun, 2001). Kriminaalanalüüs on suunatud kuritegevuse mustrite ja trendide seoste kohta õigeaegse ja asjakohase informatsiooni andmisele, et oleks võimalik planeerida ressursse, ennetada ja ohjeldada kuritegevust, püstitada uurimisprotsessi eesmärk ja paremini mõista ning avastada kuritegusid.

Näiteks võib andmete kriminaalanalüüsi abil visualiseerimine võimaldada näha seoseid kuritegude jm sündmuste või objektide vahel (nt: kuidas saab ohver ja kurjategija aegruumis kokku) (Boba, 2005). Samuti võib näiteks kriminaalanalüüsi eesmärk olla ammendavate vastuste saamine järgmistele püstitatud küsimustele (Kollepara jt, 2002):

- Kas kuriteoliigi ja sündmuse toimumise asukoha vahel on seos?
- Kas kuriteoliigi, kasutatud relva ja sündmuse asukoha vahel on seos?
- Kuidas oleks võimalik koondada ja kiirelt saada kätte kindlatele parameetritele vastavad kuriteod? Näiteks teismeliste poolt noa ähvardusel toime pandud röövimised

Traditsiooniliselt viiakse kriminaalanalüüsi läbi ainult struktureeritud andmetele tuginedes või viies läbi mahukat „käsitsi“ analüüsi.

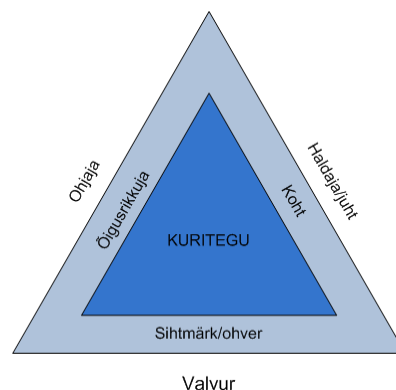
1.1. Kriminaalanalüüsi teooriad kuriteomustrite leidmiseks

Käesoleva alapeatüki eesmärk on anda ülevaade kriminaalanalüüsi peamistest teooriatest, mis toetavad püstitatud eesmärki – kuriteomustrite tuvastamist. Kriminaalanalüüsi läbiviimisel kuritegude ennetamiseks ja lahendamiseks saab toetuda vastavatele teooriatele, kombineerides neid sobivate analüüsimeetoditega.

Järgnevalt kirjeldatakse kolme kriminaalanalüüsi teooriat, millest esimene kirjeldab kuriteo kolme osapoolt, teine on selle teooria edasiarendus lähtudes keskkonnast, mis mõjutab kuriteos osalevaid isikuid ning kolmas lähtub sellest, et kuritegu on protsess, mille käigus viiakse läbi teineteisele järgnevad sammud.

1.1.1. Kuriteokolmnurk

Probleemi analüüsikolmnurga (ehk kuriteokolmnurga) eesmärk on kirjeldada kuriteo kolme olulist elementi: õigusrikkuja, koht ja ohver (Joonis 1). Lawrence Coheni ja Marcus Felsoni teooria järgi pannakse kuritegu toime siis, kui õigusrikkuja ja sobilik sihtmärk saavad kokku ajas ning ruumis, kus puudub kompetentne valve (Clarke jt, 2006). Seega on iga kuriteo puhul võimalik eristada need osapooled.

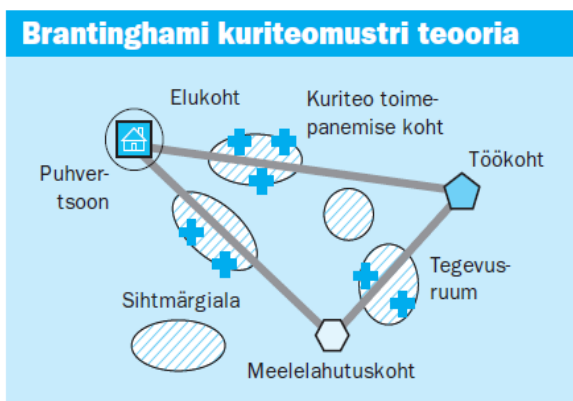


Kuriteokolmnurk kirjeldab küll kuriteo kolme osapoolt, kuid ei arvesta sellega, kuidas õigusrikkuja leiab sobiva ohvri ning kuriteopaiga. Nende oluliste aspektide arvestamiseks arendati teooriat edasi ning kirjeldati kuriteomustri teooria.

1.1.2. Kuriteomustri teooria

Mustrid (sageli küll raskesti tuvastatavad) on hea võimalus kirjeldamiseks, kuidas inimesed suhtlevad oma keskkonnas (Adderley, 2007). Kuriteomustri teooria on keskkonnakriminoloogide P. ja P. Brantinghami edasiarendus kuriteomudeli teooriast (Joonis 2), mis lähtub isiksuse tegevusruumist ning kirjeldab isikute käitumise keskkonnas läbi kolme mõiste: sõlmpunktid, teelõigud ja äärealad (Brantingham ja Brantingham, 1984). Sõlmpunktid on isiku jaoks olulised kohad nagu näiteks kodu, kool, töökoht jms, mille vahel liigutakse regulaarselt. Kõrgendatud oht kuritegevuse ohvriks langeda võib tuleneda sõlmpunkti ümbritsevast keskkonnast. Teelõigud on sõlmpunkte ühendavad ühendused, mida inimesed

igapäevaselt (harjumuse jõu tõttu läheb inimene ühest punktist teise tavapäraselt ühte ja sama teed pidi) kasutavad liikumaks ühest sõlmpunktist teise. Teelõigud on sageli kohad, kus langetakse kuriteo ohvriks ja pannakse toime kuritegusid. Äärealad on sõlmpunkte (nt: kodu) ümbritsevad piirkonnad, kus pannakse toime teatud kuritegusid (nt: röövid ja kauplusevargused), sest kohtuvad inimesed ei tunne teineteist.



Allikas: Kim Rossmo. Geographic Profiling. Boca Raton, FL: CRC Press, 2000.

Kim Rossmo koostas diagrammi, et selgitada Brantinghamide teooriat. Diagramm selgitab õigusrikkuja tegutsemisala (elu-, töö-, meelelahutuskoht ja liikumisteed nende vahel), puhvertsooni, mis asub kodu lähedal, kus õigusrikkuja tavaliselt kuritegu toime ei pane, ning viit potentsiaalset sihtmärgiala (nt auto-parkla). Siniste ristidega on märgitud kohad, kus kuritegu toime pannakse, sest seal puutuvad õigusrikkuja tegutsemisala ja sihtmärgiala kokku. Tähelepanuväärne on, et selles näites ei panda kuritegusid toime õigusrikkuja töökoha lähedal, sest seal pole sobivaid sihtmärke. Lisaks on veel kaks sihtmärgiala, kus puudub kuritegevus, sest õigusrikkuja ei tea neid.

Joonis 2 Brantinghami kuriteomustri teooria Kim Rossmo visuaalne selgitus (Clarke jt, 2006)

Kui on teada kuriteo osapooled ja isiku tegevusruum, siis selleks, et selgitada, kuidas kuritegu on toime pandud, kasutatakse kuriteo stsenaariume.

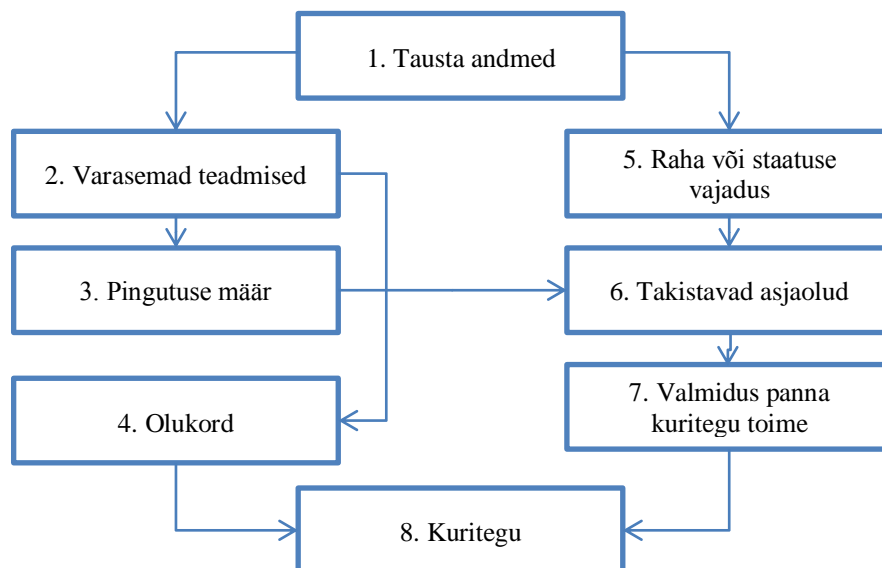
1.1.3. Kuriteo stsenaarium

Derek Cornishi väljatöötatud kontseptsiooni kohaselt toimib iga kuriteo puhul teatud tegevuste kindlas järjekorras läbimine, mis seega on kokkuvõttes pigem protsess kui sündmus. Olenemata kuriteo keerukusest toimub kurjategija poolt pidev otsuste langetamine ning nende otsuste hilisem analüüsimine politsei poolt annab infot kuriteo toimepanemise kohta (Clarke jt, 2006). Kindlat stsenaariumi järgides on võimalik nii keerukad kui ka lihtsamad kuriteod jagada omavahel võrreldavateks etappideks ning selle põhjal kirjeldada kuriteo toimepanemise muster. Omavahel saab nii võrrelda sama tüüpi kuritegusid, mille toimepanemise viis on sarnane. Kurjategija otsusprotsessi elemendid (Adderley, 2007):

1. Tausta andmed – isiksuse omadused vastavat tüüpi kuriteo toimepanemiseks, perekond jne
2. Varasemad kogemused – kurjategija taust ja oskused

3. Pingutuse määr – kuriteoga saadav tulu kaalub üles muud kurjategija meetodid eesmärgi saavutamiseks
4. Olukord – narkootikumid/alkohol või surve teiste poolt mis mõjutavad kuriteo toimepanemist
5. Raha või staatuse vajadus – võib olla motiiv kuriteo toimepanemiseks
6. Takistavad asjaolud – kas on võimalik sama asja saada muul moel
7. Valmidus panna toime kuritegu – kurjategija peab olema õiges meelelises seisundis või endale õigustama, et kuritegu on õige tegu (nt: endale sisendamine, et ohver ei vaja neid esemeid).

Sammude omavahelist seotust illustreerib allolev joonis (Joonis 3).



Joonis 3 Ratsionaalne otsustusmudel (Adderley, 2007)

Eelpool kirjeldatud teooriad seovad kuriteo tervikuks – kuriteos on alati osapooled (isikud), mingi tegevusruum, kus kuritegu toime pannakse ja kus langetakse ohvriks ning kuriteo toimepanemise protsess, mille jooksul tehakse üksteisele järgnevaid otsuseid. Nende teooriate kombineeritud rakendamine aitab andmeid analüüsida ja lähtuvalt eesmärgist leida kuriteomustreid. Järgmises punktis kirjeldatakse teoreetilise poole rakendamiseks sobivad meetodid.

1.2. Kriminaalanalüüsi läbiviimise meetodid

Selle alapeatüki eesmärk on kirjeldada kolme meetodit, mis võimaldavad kriminaalanalüüsi teooriatele toetudes kirjeldada kuriteomustreid. Allpool kirjeldatakse kolme meetodit, mille eesmärk on analüüsi abil luua andmetele analüüsi eesmärgist lähtuv struktuur ning mille tulemuste tõlgendamiseks ja põhjendamiseks saab kasutada eelpool toodud teooriaid.

Esimese meetodi kohaselt tükeldatakse uuritav andmestik lähtuvalt kuuest küsimusest plokkideks, muutes nii andmestikud omavahel võrreldavateks. Teine meetod kõrvutab toimunud sündmused ajas, et leida sündmuse toimepanekuajapõhiseid mustreid. Kolmanda meetodiga eraldatakse tekstist olulised objektid (nt: isikud, ettevõtted jms), eesmärgiga kujundada sündmustevahelised seosed läbi neid läbivate objektide. Kõigi kolme meetodi rakendamine või kombineeritud rakendamine loob eeldused, mille põhjal kujundada kuriteomustreid.

1.2.1. Kolme M-küsimuse ja kolme K-küsimuse meetod

Barry Poyneri meetodi (*5W+1H*) kohaselt peab analüüsi tulemusena oskama kuriteo kohta vastata kuuale küsimusele: mis, kus, millal, kes, miks ja kuidas, mille tulemusena muutub mahukas andmestik üksikuteks koostisosadeks peegeldades terviklikult probleemi (Poyneri, 1986; Clarke jt, 2006).

Mis juhtus? – kirjeldatakse tegevusi, millest toimunud sündmused koosnesid.

Kus juhtus? – kuriteo toimepanemise koht (nt: elukoht, nõ meelelahutuskohat ja selle ümbrus).

Millal juhtus? – millal (nt: kuupäev, kellaeg, ööpäeva osa) kuritegu toime pandi.

Kes osalesid? – vähemalt üks kurjategija. Kannatanuid ja tunnistajaid võib olla rohkem kui üks.

Miks nad nii tegid? – mis vajadust (nt: raha, staatuse saavutamine) selle kuriteoga sooviti rahuldada. Milline oli valmidus panna kuritegu toime.

Kuidas kurjategija oma teo toime pani? – üksikasjalik kirjeldus, kuidas kuritegu toime pandi, milline oli olukord (nt: narko- või alkoholijoove, teiste isikute surve).

Sündmuste tükeldamine nende küsimuste järgi loob esimese aluse sündmuste omavaheliseks võrdlemiseks. Antud meetodi juures kirjeldatakse ka sündmuse toimumise aeg („Millal juhtus?“), mis on oluliseks osaks kuriteomustrite identifitseerimise ja kirjeldamise juures (Clarke jt, 2006) ning millele keskendubki järgmine meetod.

1.2.2. Aeganalüüs

Aeganalüüsi tulemusena on võimalik näha sündmuste toimumise seaduspärasust ajas (Clarke jt, 2006). Mustrit on kergem kujundada sageli toimuvate sündmuste põhjal (nt: vargused) ning keerulisem harvem toimuvate sündmuste puhul (nt: mõrvad). Aegruumi saab vastavalt vajadusele või olemasolevatele andmetele kirjeldada näiteks ööpäeva või tundide lõikes. Võimalik on kasutada ka veel täpsemaid või üldisemaid skaalasid.

Sündmuste märgatav kuhjumine mingisse ajavahemikku aitab leida mustreid. Jerry Ratcliffe kirjeldab kolme tüüpi ajalisi kogumeid (Ratcliffe, 2002; Clarke jt, 2006):

- Hajutatud muster – ööpäeva lõikes on juhtumid võrdlemisi ühtlaselt jaotunud.
- Koonduv muster – esineb kindlate selgepiirilistes ajavahemikes (nt: tipptundidel toimuvad sündmused).
- Teravad ehk akuutsed mustrid – sündmused on kuhjunud lühikestesse ajavahemikesse.

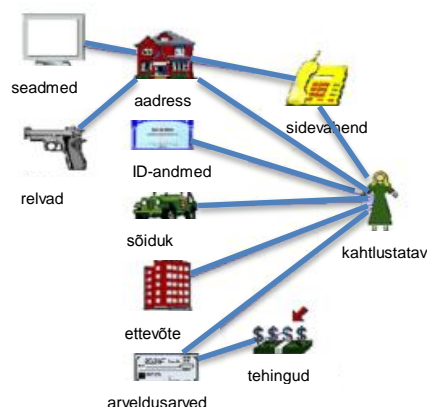
Lisaks toimepandud sündmuse kirjeldamisele ning määratlemisele aegruumis on oluline ka nõ taustsüsteem sündmuses osalenud isikute ja objektide näol ja millised on nende omavahelised seosed, millele keskendub järgmine kirjeldatav meetod.

1.2.3. Sündmuse taustsüsteemi kirjeldamine

Inimene tegutseb igapäevaselt mingis taustsüsteemis, kus toimuvad sündmused, osalevad objektid (nt: sidevahendid mida kasutatakse, aadressid kus elatakse jne) (Caputi jt, 2008). Selle taustsüsteemi mõistmine ja kirjeldamine aitab omavahel siduda toimunud kuritegusid.

Isiku taustsüsteemi võib võtta kokku akronüümi POLE taha, millega mõeldakse isikupõhist analüüsi eesmärgiga eraldada tekstist järgmised andmed: inimesed (*People*), objektid (*Objects*), asukohad (*Localities*), sündmused (*Events*) (NPIA; 2010).

Sündmuste omavahelise võrdlemise üheks aluseks on objektide vahelised ja sündmuste vahelised seosed. Nii näiteks on võimalik kuritegusid siduda läbi ühise isiku või tuvastada isikute vaheliste seoste alusel kuritegude vahelisi seoseid.



Lähtuvalt käesoleva töö eesmärgist viia analüüs [Joonis 4 POLE-mudel](#)

läbi tekstilistel andmetel on oluline andmete ettevalmistamine (sobivale kujule viimine) enne sobivate meetodite rakendamist. Järgmises peatükis keskendutakse tekstianalüüsi põhimõtete kirjeldamisele ning automatiseeritud tekstianalüüsi võimalustele ja piirangutele.

1.3. Tekstiliste andmete töötlemine

Käesoleva alapeatüki eesmärk on anda ülevaade tekstianalüüsist ning automatiseeritud tekstiliste andmete töötlemise võimalustest ja piirangutest.

Tekst on kirjutatud või trükitud sõnade järjend ning keelelise suhtlemise suurim üksus, koosnedes kindlate reeglite järgi ühendatud ja sidusast tervikust (Eesti keele seletav sõnaraamat). Tekst on struktureerimata, amorfne ja raskesti käideldav, kuid samas on ta levinuim viis informatsiooni vahetamiseks (Sharp, 2001). Väidetakse, et koguni 80% andmetest on tegelikult struktureerimata kujul, mida masinal on üldjuhul keeruline tõlgendada.

Andmebaasides hoitakse tavaliselt andmeid struktureeritud kujul (nt: isiku ees- ja perekonnanimi on erinevatel väljadel) nende parema töötlemise eesmärgil (nt: võimalus otsida nii ees- või perekonnanime järgi). Tekstilisi andmeid talletatakse tavaliselt andmebaasis tekstilist infot hoidvatel väljadel, kus ühel väljal on koos mahukas ja andmebaasi mõttes ilma struktuurita tekst. Sellisel kujul andmeid on masinal keeruline töödelda. Sellistelt väljadelt info otsimine eeldab indekseerimist, et päringud oskaksid leida ja leiaksid tekstis olevaid osi. Keerukusest olenemata on neid andmeid siiski vaja töödelda ning tulenevalt suurtest andmemahtudest on selle läbiviimiseks vaja arvutite kaasabi.

Tekstianalüüsi suurimaks probleemiks on keel ise, sest andmetel ei ole ühest tähendust. Probleem on eriti hästi tajutav morfoloogiliselt rikaste keelte juures, kuhu kuulub ka eesti keel. Eesti keeles on umbes 40-50% tekstis esinevatest sõnavormidest mitmeti mõistetavad – homonüümsed (Muischnek jt, 2003).

Järgmises punktis kirjeldatakse tekstianalüüsimiseks kasutatavaid võimalikke tehnikaid.

1.3.1. Tekstianalüüsi tehnikad

Selles punkti eesmärk on kirjeldada tekstide analüüsimiseks sobilikke tehnikaid. Suurte andmemahtude paremaks töötlemiseks kasutatakse andmekaevandamist koos erinevate tehnikate (nt: neuronvõrgud, statistika, reeglid jms) rakendamisega eesmärgiga leida neist kasulikku informatsiooni (Liiv, 2005). Lisaks nõ konkreetsele küsimusele vastuse saamiseks on tekstikaeve üks võimalus ka koguda andmeid, et oskaks esitada küsimusi andmete kohta. Tekstikaeve on üks andmekaeve alaliik ning tema eesmärk ja rakendamise tehnikad ei erine andmekaeve omast muu poolest, kui et neid tehnikaid rakendatakse tekstile.

Järgnevalt kirjeldatakse olulisemad tehnikad, mida saab rakendada andmetele siis, kui vabatekst on viidud struktureeritud kujule. Rakendatavaid tehnikaid nimetatakse andmekaeve tehnikateks. Sobivate tehnikate valik tuleneb probleemist ja eesmärgist, mida lahendada hakatakse ning saadaolevatest andmetest. (Wu jt, 2008; Adderley, 2007). Näiteks saavutatava eesmärgi järgi on Liiv (2005) jaganud meetodid kaheks:

- Kirjeldavateks (avastavateks) meetoditeks – näiteks: klasterdamine, seoseanalüüs, jadamuustrid (*sequence patterns*), mida kasutatakse eelnevalt mitteteadaolevate reeglipärade leidmiseks andmete põhjal muustrite kujundamise teel.
- Ennustavateks meetoditeks – näiteks: klassifitseerimine, regressioon, mida kasutatakse mingite näitajate ennustamiseks olemasolevate andmete põhjal.

Käesolevas töös ei viida läbi automatiseeritud tekstianalüüsi ja -kaevet ning seetõttu on siin punktis loetletud meetodite nimetamisel toetunud erinevate autorite praktikale (Chen jt 2004; Adderley, 2007), mida kasutasid nemad politseis kogutavate andmete analüüsis:

- Klassifitseerimine – kasutakse sageli ennustavas analüüsis, eesmärgiga olemasolevate andmete põhjal ennustada uute andmete eeldefineeritud klassidesse kuulumine. Näiteks politsei jaoks on klassifitseerimisest kasu kuritegevuse trendide loomisel.

- Seostepõhine analüüs – eesmärk on tuvastada sageli koos esinevate objektide paare. Kriminaalanalüüsi puhul näiteks sageli koos esinevate kuriteotunnuste eraldamine (nt: röövimisega kaasneb vägivald). Suhtlusvõrgustike analüüs (*Social Network Analysis*) – kirjeldab objektide vahelised seosed ja rollid, mida kriminaalanalüüsis saab kasutada näiteks isikute vaheliste seoste kaardistamiseks.
- Klasterdamine – eesmärk on jagada objektid nendele omaste teatud tunnuste alusel gruppidesse ehk klastritesse. Grupeerimise tulemusena jaotakse objektid klastritesse nii, et ühte klastrisse kuuluvad objektid on eelnevalt kindlaks määratud valikukriteeriumi põhjal sarnasemad võrreldes teistesse klastritesse kuuluvate objektidega. Kriminaalanalüüsis on näiteks võimalik kuritegusid grupeerida nende liigi järgi ühte klastrisse.
- Reeglite rakendamine – eesmärk on eeldefineeritud reeglite järgi andmete eraldamine tekstist. Kasutada võib ka statistilisi reegleid, mille rakendamise tulemusena leitakse mingi tunnuse statistiline esinemine. Reegleid kasutatakse ka nimeliste objektide (*Named Entity*) eraldamiseks – teksti võrdlemine kirjeldatud reeglitega, et tekstist leida nimelisi objekte. Näiteks suurtähega nimi + suurtähega perenimi >> isik.
- Üldistest andmetest kõrvalekallete tuvastamine (*Deviation detection, outlier detection*) – tuvastatakse andmed, mis olulisel määral erinevad üldistest andmetest. Kriminaalanalüüsi jaoks näiteks kasutajate tavapärasest käitumisest kõrvalekallete leidmine. Eeldab eeldefineeritud nõ „normaalse käitumise“ kirjeldamist.

Paremate tulemuste saavutamiseks ei ole otstarbekas tugineda ainult ühele meetodile vaid neid kombineerida (Chau jt, 2002) ning rakendada näiteks täiendavalt:

- Võrdlemist sõnastikega – olulistest nimedest (isikute nimed, ettevõtete nimed jne) käsitsi koostatud sõnastikud. Andmete võrdlemine meetodis käib sõnastiku alusel.
- Masinõpe – selle meetodi puhul tuginetakse pigem masinõppivatele algoritmidele, kui käsitsi kirjeldatud reeglitele. Masinõpe põhineb näiteks: neuronvõrkudel ja otsustuspuudel jne.

Konkreetsed kaevandamise tehnikad koos eesmärgipärasusega on võimalik esitada tulevikus läbiviidava detailanalüüsi etapis.

Antud töö rakendusliku osa läbiviimisel toetutakse tugevalt järgmises punktis kirjeldatud teksti sisu analüüsimise meetodil, mis on kasutusel juba 20-nda sajandi algusest alates.

1.3.2. Teksti sisu analüüsimine kontentanalüüsiga

Kontentanalüüs on tekstianalüüsi meetod, mille üks definitsioon on kirjeldatud Klaus Krippendorfi poolt järgmiselt: “kontentanalüüs on uurimismeetod – vahend, mille kasutamise käigus tehakse andmete ja nende konteksti osas korratavaid ja tõeseid järeldusi” (Krippendorf, 1980).

Kontentanalüüsi hakati laialdasemalt kasutama 20-nda sajandi alguses, kui algas ajakirjanduse võidukäik USAs ja teadusmaailmas hakati mõtlema ajakirjanduse analüüsimisele (*quantitative newspaper analysis*) (Krippendorf, 1980). Massimeedia on siiani jäänud suurimaks valdkonnaks, kus tekstide analüüsimist kasutatakse. Lisaks kasutavad psühholoogid oma suusõnaliste vastuvõttude materjali analüüsi patsiendi motivatsiooni ning mentaalsete ja isikuliste omaduste selgitamiseks (Krippendorf, 1980). Analoogselt saab ka politsei kasutada tekstianalüüsi tekstiliste andmete (nt: kahtlustatava, tunnistaja või kannatanu antud ütlused) analüüsimiseks.

Kontentanalüüsi jaotatakse kaheks: kvantitatiivseks ja kvalitatiivseks tekstianalüüsiks. Kvantitatiivse meetodi eesmärk on vastata küsimustele „Kui palju?“ ja „Miks?“. Kvantitatiivne kontentanalüüs võimaldab tekste täpsete numbriliste väärtustega mõõta ning näiteks näidata valitud ühiku omaduste esinemise sagedust. Puuduseks on see, et kvantitatiivse meetodiga ei saa me täpseid tulemusi teksti sisu osas. Teksti sisulise ja varjatud (keeleliselt mitte otse välja öeldud) tähenduste väljatoomiseks kasutatakse natukene paindlikumat kvalitatiivset tekstianalüüsi, millega tahetakse saada vastust küsimustele „Kas?“ ja „Kuidas?“ ja mis võimaldab arvestada tekstielementide vaheliste seoste ja suhetega (Kalmus, 2000).

Kvalitatiivse kontentanalüüsi puhul puudub selgelt mõõdetav kodeerimise juhend, kasutatakse nõ avatud kodeerimist. Tekstil „lastakse kõneleda“ ja ei toimu teksti tükeldamist osadeks nende loendamise eesmärgil. Puudusteks on see, et kvalitatiivne kontentanalüüs jätab lahti võimaluse valikulise tõendusmaterjali kogumiseks (Kalmus, 2000).

Kontentanalüüsi jaoks ei ole takistuseks suured andmemahud ning eeliseks on mitmekülgsus ja võimalus teha järeldusi teksti osade esinemissageduste ja omavaheliste seoste kohta, luues

dokumendist nõ loogilise vaate, kus sõnad on koondatud kokku hulga väiksemaks arvuks märksõnadeks/kategooriateks. Meetoditena kasutatakse näiteks sünonüümide ühendamist ja sõnade grupeerimist nende konteksti või sõnade mõttelise seose (tervis ja jõud) järgi (Weber, 1990).

Weber (1990) leiab, et kontentanalüüsi suurimaks probleemiks on kodeerimise usaldusväärsus. Krippendorf (1980) eristab kolme tüüpi usaldusväärsust: stabiilsus (*stability*) – sama kodeerija kodeerib teksti alati samamoodi, korratavus (*reproducibility*) – mitu kodeerijat kodeerib sama teksti samamoodi ja täpsus (*accuracy*) – kodeerimise tulemus vastab standardile või kirjeldatud normile.

Arvuti poolt teostatud kodeerimisega on usaldusväärsust ja valiidsust võimalik märgatavalt suurendada, kuna erinevalt inimesest kodeerib masin infot etteantud algoritmi järgi alati ühte moodi. Weberi nimetatud mure stabiilsuse üle on täiesti mõistetav, kuna käsitsi kodeerimisel on stabiilsust keeruline saavutada, sest isegi professionaalsed kodeerijad ei suuda olla kodeerimisel järjepidevad.

Selle punkti eesmärk oli anda ülevaade kontentanalüüsist ja selle meetodi positiivsetest külgedest ning tuua välja probleemid, millega peab arvestama tekstide kodeerimisel. Automatiseeritud tekstianalüüs eeldab aga erinevate keeletehnoloogiate kasutamist ja neid tutvustatakse lähemalt järgmises punktis.

1.3.3. Keeletehnoloogiad tekstianalüüsi läbiviimisel

Käesoleva punkti eesmärk on lühidalt kirjeldada automatiseeritud tekstianalüüsi arengut ning eesti keele tehnilist tuge. Kirjelduses keskendutakse põhilistele käesoleva töö eesmärki saavutada aitavatele keeletehnoloogiatele.

Juba 1950-ndate aastate lõpus hakati mõtlema sellele, kuidas tõhustada tekstianalüüsi läbiviimist arvutiga. Arvuti laialdasemat kasutusele võtmist mõjutasid arengud muudes valdkondades (lingvistid arendasid mitmeid süntaksi ja semantika interpreteerijaid jne) ning andmete digitaliseerimine (nt: alustati suurte tekstikogude digitaliseerimisega) (Krippendorf, 1980; Muischnek jt, 2003). Keeletehnoloogid tegelevad pidevalt sellega, et viia teksti arvutite jaoks mõistetavamaks ja kergemini töödeldavamaks. Praeguseks hetkeks on maailma mastaabis arvutis loetavaid sõnastikke palju ning praegune kriitiline ülesanne on vajadus neid standardiseerida (Muischnek jt, 2003). Eesti keele kohta on samuti olemas elektroonilisi

sõnastikke, mis on toetanud oluliselt mitmete keeletöötluslike ülesannete lahendamist (nt: süntaktiline ja sõnatähenduse analüüs jpm) arvutite kaasabil (Muischnek jt, 2003).

Kirjaliku teksti töötlemiseks arvutiga on vaja kasutada keeletehnoloogiaid, mille valik sõltub püstitatud eesmärgist. Peamised nõuded kirjalikule tekstianalüüsile (Muischnek jt, 2003) antud töö kontekstis võiksid olla järgmised:

- Tekstide tükeldamine etteantud ühikuteks (nt: sõnadeks, lauseteks, lõikudeks), mis eeldab arvutilt oskust tunda, kus algab ja kus lõpeb valitud ühik.
- Morfoloogiline analüüs:
 - Teksti keele tuvastamine – vajalik selleks, et edaspidi rakendatavad tehnoloogilised võtted annaksid korrektseid tulemusi.
 - Sõnavormide analüüs – tulemuseks on tekstis esinevate sõnade tuletatud algvormid. Eesti keeles on see realiseeritud sõnastike ja reeglite kombineerimise teel. Analüüsi tulemust võib mõjutada kirjavigade esinemine tekstis, kui seda ei arvestata (nt: tüüpilisemad trükkimise vead jne) analüüsi käigus.
 - Morfoloogiline ühestaja – eesmärk on mitmetähenduslikele sõnadele antud kontekstis õigeima tähenduse leidmine. Võrdlemine toimub vahetult nende ees ja järel olevate sõnadega, et arvestada sõna kontekstiga ning vähendada mitmetähenduslike sõnade probleemi. Morfoloogiline ühestaja annab umbes 95% juhtudest positiivse tulemuse.

Tulemuseks peaks tekkima kasutatud sõnade loetelu koos sõnade algvormide ning nende sõnatüüpidega. Peale seda on võimalik neid andmeid töödelda eesmärgist tulenevalt.

Näiteks kui eesmärgiks on leida teksti kohta käivad olulisemad märksõnad, siis kõigepealt eraldatakse tekstist sidesõnad, mis muidu oleksid sageduselt domineerivad. Oluliste märksõnade leidmiseks viiakse tekst väiksema arvu märksõnade alla (nt: sünonüümide ühestamine). Teksti viimine lihtsustatud kujule ehk märksõnade eraldamine tekstist on keeruline ülesanne. Probleemiks on see, et ei ole nii kindlalt määratud märksõnu, mis igal ajahetkel ja olenemata isikust kehtiksid. Ainuke kriteerium on kasutaja ja tema vajadus antud ajahetkel (Hulth, 2004).

Ülesande lihtsustamiseks peab kasutama ka näiteks valdkondlike sõnastikke, et eristada probleemist tulenevaid olulist seost omavaid märksõnu, mille esinemissagedus tekstis muidu oleks madal.

Iga tekst sisaldab lisaks morfoloogiaga tuvastavatele sõnadele nimelisi objekte (*Named Entity*) näiteks isikud, kohanimed, ettevõtted jms. Nende tuvastamine tekstist, eeldab vastavate keelereeglite olemasolu või sõnastike olemasolu, kus analüüs viiakse läbi võrdluse teel sõnastikuga. Selline teksti anoteerimine (*Text Annotation*) sõnastikeks toimub inimeste poolt ja arvuti võrdleb enda kogutud andmeid nende vastu. Parimate tulemuste saavutamiseks ei ole otstarbekas toetuda ainult sõnastikele vaid kombineerida neid morfoloogilise oletaja ja ühestajaga, kasutada mustreid või statistilisi näitajad ning masinõpet (neuronvõrgud, otsustuspuud, masinõppivad algoritmid jne) (Chau, 2002).

Lisaks objektide eraldamisele on vajalik nende koondamine võimalikult vähesteks ja mittekattuvateks klassideks. Eelduseks on oskus omavahel siduda ühesuguseid objekte. Väljakutseks on, kuidas viia kokku näiteks isiku pärisnimi ja tekstis tema kohta kasutatav hüüdnimi.

Peatüki eesmärk oli anda ülevaade praeguseks keeletehnoloogide poolt loodud lahendustest teksti analüüsimiseks arvuti kaasabil. Antud töö püstitatud eesmärgi saavutamiseks on esmajoonel oluline morfoloogiline analüüs ja nimede eraldamise oskus tekstist.

Kuna tegemist ei ole probleemiga, millega varasemalt ei oldaks muude riikide politseis kokku puutunud, siis järgmises peatükis tuuakse mõned näited, kuidas teiste riikide politseid on enda jaoks lahendanud tekstianalüüsi kriminaalanalüüsi toetajana.

1.4. Ülevaade sarnastest uuringutest

Peatüki eesmärk on anda mõne näite põhjal ülevaade sarnaste probleemide lahendamisest mujal maailmas. Kahjuks ei ole võimalik siin viidata varasematele töödele Eesti Politseis, kuna neid ei ole haakuvas kontekstis seni läbiviidud. Maailmas aga ei ole antud probleemi lahendamine politsei jaoks uus. Vajaduse kasv on suurenenud viimaste aastatega ning koostöös ülikoolidega on loodud ka lahendusi tekstide paremaks analüüsiks.

Näiteks võib tuua projekti COPLINK (USA) (<http://ai.bpa.arizona.edu/research/coplink/>). Antud töö raames rakendati politseis kogutavatele andmetele andmekaeve põhimõtteid.

Nende tegevused jagunesid nelja suuremasse gruppi: objektide eraldamine, seoste eraldamine, ennustamine ja mustrite visualiseerimine (Chen jt, 2004).

Täpsemalt viidi läbi sammud, millega kõigepealt töödeldi tekste lähtuvalt keelereeglitest ja eraldati nimisõnafraasid, seejärel arvutati igale fraasile tulenevalt mustrist tema kaal ning lõpuks leiti neuronvõrgu tehnoloogiat kasutades kõige tõenäolisem objekti klass (isik, asutus, aadress vms).

Valeidentiteetide (mida kurjategijad võivad kasutada) vältimiseks kasutasid nad isiku andmete võrdlemist andmebaasis olevate isikute andmete vastu. Kolmanda eesmärgina lahendasid nad kuritegelike võrgustike ja nende võtmeisikute visualiseerimise, mille tulemusena toodi dokumendist välja isikutevahelised seosed, mille omavaheline korrelatsioon arvutati välja selle põhjal, kui tihti neid dokumendis koos mainiti. Oma ülesande nad vajalikus ulatuses täitsid.

Samuti on arendatud sarnaseid süsteeme Ühendkuningriigis, kus R. W. Adderley (Adderley, 2007) oma doktoritöös näitab, kuidas andmekaeve tehnikaid rakendades saab analüüsida kuriteotrende ning profileerida kurjategijat. Oma töös nendib ta, et politsei andmebaasid erinevad tavapäraest andmekaevandamise subjektiks olevatest andmebaasidest, kuna seal hoitakse koos ajutisi, ruumilisi ja geograafilisi andmeid koos tekstiliste andmetega ning tõdeb, et sellest tulenevalt on neid andmeid seni vaid keskmises ulatuses kasutatud. Töö tulemustes nendib ta, et valitud algoritme õnnestus edukalt kasutada nii struktureeritud kui struktureerimata andmete peal ning tõhusus suurenes, kui andmete kodeerimisel lähtuti kriminaalanalüüsi teooriatest. Kokkuvõttes võideti ajas ning analüütilise protsessi täpsuses. Lisaks mõjutati erineva tasandi strateegiaid.

Järgmise alapeatüki eesmärk on võtta kokku esimeses peatükis esitatud teooriad ja meetodid ning nende põhjal kirjeldada politsei jaoks sobiv analüüsimudel.

1.5. Tekstianalüüsi läbiviimise metoodika ja analüüsimudeli kujundamine

See alapeatükk keskendub eelnevalt toodud teoreetilise poole põhjal rakendatava tekstianalüüsi metoodika väljatöötamisele ja kirjeldamisele kuriteomustrite leidmiseks. Tulemuseks on sisend töö teisele ehk praktilisele osale, et loodud mudelit vähendatud mahus rakendada eesmärgiga hinnata üldiselt mudeli vastavust ning kirjeldada üldiseid nõudeid automatiseeritud tekstianalüüsile.

Peatükk jaguneb kaheks osaks, millest esimene kirjeldab loodava analüüsimudeli põhimõtted ja eesmärgi, võttes kokku kuriteomustrite analüüsi teooria ja töö eesmärgi ning teine osa kirjeldab analüüsimudelit lähtuvalt andmekaeve standardi CRISP-DM sammudest.

1.5.1. Analüüsimudeli põhimõtted ja eesmärk

Eesmärk on kujundada analüütiline lahendus, mis koosneb sobivate meetodite kogumist, mida hakatakse rakendama andmetele eesmärgiga võrrelda andmeid ning nende ühisosa ja sarnasuste põhjal kirjeldada kuriteomustrid, mida on võimalik „peegeldada“ seni lahendamata kuritegudele ning nii aidata kaasa nende lahendamisele. Analüütilist lahendust peab uurijatel olema võimalik rutiinselt kasutada selleks, et leida kuriteomustritele vastavaid kuritegusid ja seoseid erinevate sündmuste, asukohtade, aegade jms vahel ning vajadusel luua uusi kuriteomustreid.

Tõhus kriminaalanalüüs kasutab konkreetse kuriteo lahendamiseks kõiki vajalikke meetodeid ja selle läbiviimine on võimalik jagada osadeks nii, et tekivad teiste toimunud kuritegudega võrreldavad mustrid, mis võivad kaasa aidata seni avastamata kuritegude lahendamisele.

Antud analüüsimudeli jaoks defineeritakse kuriteomuster läbi järgmiste tunnuste operatsionaliseerimise: kuriteosündmus; *modus operandi* – kuriteo toimepanemise viis; kuriteo toimepannud isik; sündmuse või isikuga seotud objektid.

Kuriteomustrite kujundamiseks võib kasutada erinevaid politsei poolt kogutud andmeid toimepandud ja lahendatud kuritegude kohta (nt: järgmises peatükis proovitakse kuriteomustreid kujundada kahtlustatava ütluste järgi). Mustrite kujundamiseks kasutatakse eelpool kirjeldatud kuriteomustrite teooriaid läbi analüüsi meetodite rakendamise valimile.

1.5.2. Analüüsimudeli komponendid

Käesoleva punkti eesmärk on kirjeldada eelpool toodud teooriast, meetoditest ning andmekaevandamise põhimõtetest lähtuvalt koostatud analüüsimudel kohandatuna eesmärgiga kuriteo mustrite tuvastamiseks.

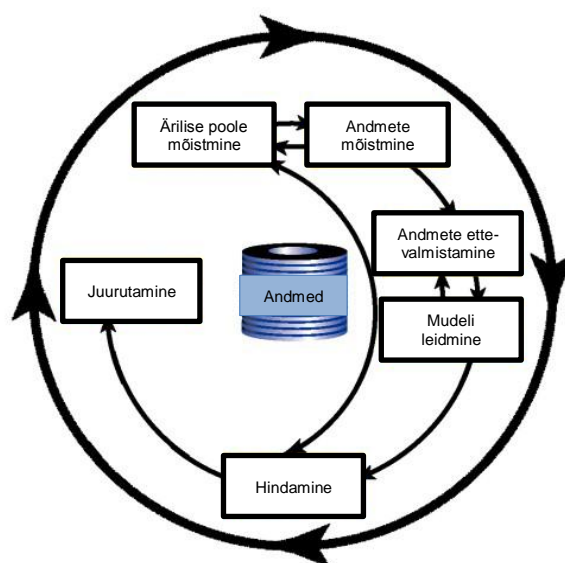
Analüüsimudel luuakse lähtuvalt andmekaevandamise protsessi valdkonnas standardiks kujunenud *The Cross Industry Standard Process for Data Mining* (CRISP-DM) mudelist, mis on loodud küll eelkõige äriettevõtete vajadusi silmas pidades. Kuid Adderley (2007) viitab oma töös, et seda on soovitatud kasutada ka kuritegude andmete kaevandamisele ning

KDNuggets 2002. aasta uuringule, kus 51% vastanutest on öelnud, et kasutab oma andmekaevanduse läbiviimisel CRISP-DM standardit.

CRISP-DM järgi kirjeldatakse andmete kaevandamine protsessina, mis koosneb kuuest omavahel seotud etapist, mida võidakse läbida korduvalt ning paremate tulemuste saamise nimel liigutakse sammude vahel edasi ja tagasi.

Lisatud joonisel (Joonis 5) on näha etapid, mis tuleb läbida ning nooltega on kujutatud enim levinud suunad nende etappide vahel liikumiseks.

Lühidalt koosneb analüüs järgmistest etappidest: protsessi esimeseks etapiks on valdkonna mõistmine, millega täpsustakse küsimused, millele vastust otsima hakatakse. Järgneb andmete kogumise ja nendest ülevaate saamise etapp, millele järgneb kõige ajamahukam ja olulisim etapp, kus toimub andmete puhastamine ja ettevalmistamine (analüüsi jaoks sobivale kujule viimine). Sobival kujul andmetele saab rakendada modelleerimist sobivate meetodite valimisega ja nende rakendamisega. Saadud tulemusi tuleb hinnata seisukohast, kas püstitatud eesmärk täideti, peale mida toimub tulemuste esitamine. Järgmiseks kirjeldatakse kõiki etappe põhjalikumalt, kirjeldades eesmärgid ning saavutatavad tulemused (CRISP-DM 1.0, 2000).



Joonis 5. CRISP-DM modelleerimise protsess

1.5.2.1. Valdonna mõistmine

Andmekaeve protsess algab uuritava valdkonna probleemi ja andmestiku põhjalikust selgitamisest, mis on eelduseks eduka mudeli loomisel. Kokku kogutakse olemasolev teave valdkonna kohta:

- Lahendada hakatava probleemi sõnastamine.
- Andmete asukoht (nt: andmebaasid, võimalusel tabelid), kus ülesande lahendamise jaoks vajalikke andmeid hoitakse.

- Mis kujul andmeid hoitakse (nt: elektrooniliselt, paber kandjal). Esmane ülevaade annab võimaluse näha, milliseid (täiendavaid) samme (nt: vaja andmeid digitaliseerida jne) on vaja astuda selleks, et eesmärgipärase tulemuseni jõuda.
- Andmete kitsendamise kriteeriumid (nt: kuriteoliigi järgi). Kitsendamine võimaldab muuta andmed paremini võrreldavaks. Näiteks kuritegudel on küll palju sarnaseid tunnuseid, kuid siiski on igal kuriteoliigil spetsiifilised tunnused (nt: röövimisega kaasneb vägivald, mis eristab seda vargusest, kus vägivalda isikute suhtes toime ei panda).

Tulemuseks on ülevaade olustikust, mis probleemi tõstatas ning on kirjeldatud probleem, mida lahendada hakatakse.

Ülesande püstitamise juures tuleks kohe kirjeldada ka oodatav tulemus läbi hindamiskriteeriumite sõnastamise selleks, et oleks võimalik saadud tulemusele anda hinnang. Peale nende sõnastamist on võimalik asuda andmete kogumise juurde.

1.5.2.2. Andmete kogumine

Kui eelmises etapis on muuhulgas saanud selgeks see, millised andmeid vajatakse, siis selle etapi eesmärk on need andmed kokku koguda, et neid järgmises etapis (andmete ettevalmistamine) sobivale kujule viima hakata. Andmeid võidakse hoida erineval kujul: paberil (vaja enne digitaliseerida), struktureeritud andmebaasides struktureeritud ja struktureerimata kujul.

Probleemi püstitamisest selgub see, mis andmeid on konkreetselt selle ülesande lahendamiseks vaja ning lähtuvalt sellest koostatakse enne andmete kogumise alustamist detailne andmete kirjeldus:

- Detailne loetelu andmeallikatest. Politseis kogutavaid andmeid võib leida igal eelpool nimetatud kujul – paber kandjal näiteks isikute avaldused toimunud kuriteo kohta, politsei andmebaasides on digitaalsed andmed nii struktureeritud kujul (nt: registreeritud sündmused), kui ka struktureerimata (vabatekstilisel) kujul, milleks on näiteks ülekuulamiste protokollid.
- Detailne ülevaade andmetest. Kirjeldatud on andmete struktuur (struktureeritud või struktureerimata kujul olevad andmed), andmetevahelised seosed ja andmete kvaliteet.

Andmetest ülevaate loomise tulemusena on näha mis tegevusi on vaja viia läbi järgmise andmete ettevalmistamise etapi käigus. Näiteks andmete kirjelduste põhjal saab planeerida andmete saamise meetodid, mille abil kavatsetakse analüüsi jaoks vajalikud andmed kätte saada (nt: andmete ühekordsed väljavõtted, päringud, teenused jms).

Peale andmete kogumist on võimalik kirjeldada selleks hetkeks selgunud probleemide loetelu (nt: andmeid ei ole võimalik saada algselt plaanitud kujul või mahus).

Antud sammu tulemusena tekib loetelu andmeallikatest ja andmete kirjeldusest (nt: andmete struktuur, seosed andmete vahel, andmete kvaliteet). Andmetest ülevaate loomise tulemusena on näha, kui mahukaks kujuneb järgmine etapp, mille eesmärk on valmistada andmed ette analüüsimiseks.

1.5.2.3. Andmete ettevalmistamine

Andmete ettevalmistamine on üks olulisemaid tegevusi, kuna selle etapi tulemuse kvaliteedist sõltub see, milline on edasiste etappide tulemuste kvaliteet. Andmete ettevalmistamise eesmärk on loetleda kasutatavad andmed ning viia andmed maksimaalselt kasulikule kujule – kõrvaldada puuduvad väärtused, dubleerivad andmed jne.

Eelmise etapi tulemusena kokku kogutud andmete ettevalmistamiseks analüüsi jaoks viiakse läbi järgmised tegevused:

- Andmete normaliseerimine – andmete eelnev töötlemine (nt: puuduvate väärtuste asendamine, dubleerivate andmete kõrvaldamine jne). Ettevalmistuse käigus võib toimuda ka andmete ühendamine (mitmest väärtusest kokku), lisaks toimub ka andmete viimine ühele formaadile (sünktahtiline), millega ei muudeta sisulist tähendust. Lisaks on vaja arvestada, et ka järgmises sammus rakendatav andmekäve tehnika võib esitada nõudeid andmetele (nt: puuduvad andmed ei ole lubatud).
- Struktureerimata andmete töötlemine. Kirjeldada tuleb, kuidas toimub struktureeritud andmetest märgatavalt vigasemate („mürasemate“ – kirjavead, trükkimisel tekkivad vead, lühendid, släng jms (Chau jt, 2002)) tekstiliste andmete korrastamine.
- Millise struktuuriga andmeid hoitakse, näiteks kuidas luuakse tekstidele struktuur (tükeldamine loogiliseks osadeks), et oleks võimalik järgmises etapis analüüsi

tehnikatega neid andmeid analüüsida. Sealjuures kasutatakse järgmisi keeletehnoloogiaid:

- Keele tuvastamine – masin saab aru, mis keeles tekstiga on tegu (üks tekst võib sisaldada mitme erineva keele tekste ning keele mittetuvastamise tulemusena ei ole tulemused korrektsed), et rakendada järgnevates sammudes vastava keele loogikaid
- Teksti tükeldamine loogilisteks osadeks – lõikude, lausete, sõnade ja nende piiride tuvastamine.
- Morfoloogiline tekstianalüüs – sõnade algvormide tuvastamine ja morfoloogiline ühestamine – mitmetähenduslikele sõnadele kontekstist tulenevalt õige tähenduse valimine.
- Nimede eraldamine tekstist – isikute, asukohtade, ettevõtete nimede eraldamine tekstist näiteks morfoloogilise oletamise või sõnastike abil (nt: ettevõtete nimed, riikide nimed jms), kus tekstis esinevale sõnale tähenduse andmine käib läbi võrdlemise sõnastikuga.

Kui eelnevalt on kirjeldatud kõik vajalikud tegevused andmete ettevalmistamiseks, siis tuleb need läbi viia, et tekiks korrastatud valim, mis on aluseks modelleerimisele ja analüüsitegevuste läbiviimisele. Korrastatud andmed võimaldavad hakata nende põhjal looma järgmises punktis kirjeldatud mudelit.

1.5.2.4.Mudeli loomine

Selle etapi eelduseks on korrastatud andmed. Tulemuseks on läbi erinevate modelleerimistehnikate sobiva mudeli leidmine. Sobiva valiku jaoks rakendatakse tehnikaid erinevate sisendparameetritega ning leitakse parimat tulemust andev sisendandmete komplekt.

Mudeli koostamise aluseks on oodatava funktsionaalsuse kirjeldus. Konkreetse tehnika valik sõltub andmetest ja eesmärgist mida tahetakse saavutada. Näiteks politsei vajadustest tulenevalt võib olla vajalik:

- Andmete koondamine klastritesse – oluliste märksõnade ja terminite leidmine tekstilisest kirjeldusest (Kollepara jt, 2002). Võimalik töödelda andmeid

eeldefineerimata või eeldefineeritud moel, kus keskendutakse ainult eeldefineeritud teemadele (nt: relvadega, narkootikumidega, koolidega seotud sündmuste leidmine).

- Seoste (korrelatsioonide) leidmine – eesmärk leida kuritegevuse karakteristikute (nt: sündmuse liik ja asukoht) vahelisi seoseid ning need seosed visualiseerida. Graafiliste seoste kuvamine koos korrelatsiooniga võimaldab koheselt teha olulisi järeldusi (nt: joone tüüp ja paksus määrab seose liigi ja selle olulisuse)
- Kuritegude kirjeldamine toimumise aja järgi.
- Tulemuste visualiseerimine.

Konkreetsed andmekaeve tehnikad täpsustuvad käesoleva töö välistes etappides, sest parima mudeli leidmine võib toimuda erinevate tehnikate katsetamise ja modelleeritud mudeli rakendamisel erinevate parameetritega.

Enne mudeli rakendamist tuleb valmis mõelda, kuidas testida loodava mudeli kvaliteeti ja valiidsust. Lõppkokkuvõttes tuleb anda hinnang, milline nendest mudelitest annab parima tulemuse lähtuvalt püstitatud eesmärgist.

Tulemuseks on kirjeldus, mis sisaldab ülevaadet kasutatavatest modelleerimise tehnikatest, plaanist, kuidas hinnata mudeli tulemusi, loodud mudelite kirjeldusi ning hinnangut kirjeldatud mudelitele. Protsess ei lõpe mudeli loomisega, vaid tuleb hinnata loodud mudelit ja kuidas sellega püstitatud eesmärk on täidetud, mida kirjeldatakse tulemuste hindamise etapis.

1.5.2.5. Tulemuste hindamine

Hindamise etapp on oluline etapp, mida tuleks rakendada mitte ainult modelleerimise lõpus, vaid mis on oluline tegevus peale igat etappi, et võimalikult varakult eemaldada tekkivad vead. Lõplik tulemuste hindamine peab andma valdkonna ülevaates püstitatud probleemile vastuse nii, et tulemus vastab toodud hindamiskriteeriumile.

Tulemuseks on kokkuvõtlik ülevaade püstitatud eesmärgist, teostatud tegevustest (nt: tegevused mis jäid sooritamata, tegevused mida peaks kordama jms) ning saadud tulemustest ning lõplik heaks kiidetud mudel. Mudel luuakse üldjuhul korduvate tegevuste läbiviimiseks

ning protsess ei ole lõppenud enne kui mudel rakendatakse näiteks toetatavatesse infosüsteemidesse.

1.5.2.6.Mudeli rakendamine

Eduka mudeli eesmärk on, et see rakendatakse püstitatud probleemi lahendamiseks. Rakendamine eeldab juurutamise planeerimise tegevusi, mudeli töötamise monitoorimist ja hooldamist.

Tulemuseks on lõplik püstitatud probleemi ja selle lahendamise ülevaade, millele on lisatud mudeli rakendamiseks vajalike tegevuste ning monitoorimiseks ja haldamiseks vajalikke tegevuste loetelu.

Kokkuvõtvalt on analüüsimudeli komponendid tuginedes andmekavandamise protsessile CRISP-DM järgmised:

- Andmete eeltöötlemine eesmärgiga viia andmed analüüsi jaoks sobivale kujule.
- Tekstianalüüsi abil oluliste andmete ja terminite välja toomine.
- Identifitseeritud andmete mustrite ja kokkusattumiste analüüsimine.
- Automatiseeritud kuriteomustrite analüüsilahenduse kujundamine.

Käesoleva töö esimese osa eesmärk oli kirjeldada teoreetilise pool – teooriad, meetodid (ka tekstianalüüs), andmete kavandamise põhimõtted ja tehnikad, mis kokku peaksid andma piisava aluse, et viia läbi arendusuuringu teine samm eesmärgiga kirjeldada tekstianalüüsi läbiviimise meetodika kuriteomustrite tuvastamiseks.

2. ANALÜÜSIMUDELI RAKENDAMINE KAHTLUSTATAVA ÜTLUSTELE

Peatüki eesmärk on eelmises peatükis kirjeldatud analüüsimudeli rakendamine. Selleks, et hinnata kirjeldatud protsessi otstarbekust, viiakse läbi kirjeldatud analüüsimudeli protsess vähendatud mahus, et saada üldine ülevaade kirjeldatud mudeli otstarbekusest ning täpsustada nõudeid automatiseeritud tekstianalüüsile. Detailne hindamine toimub järgmises käesoleva töö välises etapis, kui on juba võimalik viia läbi modelleerimine ning tekstikaeve tehnikate rakendamine selleks sobiva rakenduse kaasabil.

Selles peatükis kirjeldatud tulemusteni jõuti ilma andmekaevandamise vahendeid kasutamata. Eelpool kirjeldatud analüüsimeetoditest ei rakendatud selles faasis POLE-mudeli järgi objektide eraldamist, kuna nende andmete eraldamiseks tekstist kasutatakse automaatse tekstianalüüsi puhul keelereegleid või võrdlemist sõnastikega ning käsitsi nende eraldamine siin töös mingit lisaväärtust andnud ei oleks.

Analüüsimudelit rakendati röövimistes koostatud kahtlustatavana ülekuulamise protokollide andmestikule. Järgnevalt kirjeldatakse läbiviidud samme ja tulemusi.

2.1. Valdkonna mõistmine

Käesolevas töös loodava mudeliga soovitakse tekstilisel kujul olevatest kahtlustatava ütlustest saada kätte toimepandud kuriteo kohta andmeid eesmärgiga, et nende andmete järgi oleks võimalik kujundada kuriteomustreid. Töös piiritleti valimisse sattumist kuriteoliigi järgi. Selline piiritletus tulenes sellest, et kuriteo toimepanemise viisid erinevad kuriteoliigiti ning ei ole otstarbekas ühte valimisse võtta erinevat liiki kuritegusid. Kuriteoliigiti on ühised tunnusjooned, millest moodustuvad kuriteomustrid, erinevad.

Tervikpildis huvitavad politseid vastused küsimustele: kes, kuidas, miks, mis, millal ja kus. Nende vastuste põhjal on võimalik toimunud sündmuseid omavahel võrrelda ja näiteks nii aidata lahendada seni lahenduseta kuritegusid ning leida preventiivmeetmeid, mis võivad arvestada just neid ühiseid tunnusjooni (Ekblom, 1988). Käesoleva näite lahenduses püstitati kitsam eesmärk – keskenduti sellele, kuidas kuritegu toime pandi ning sõnastati sellest lähtuvalt eesmärk:

- milline on tüüpiline kuriteo (röövimise) *modus operandi* (edaspidi MO)

Püstitatud probleemi eesmärk on täidetud, kui tekib ülevaade röövimist enim iseloomustavate tunnuste kohta.

2.2. Ülevaade andmetest, mida koguma hakatakse

Sammu eesmärk on saada täpsem ülevaade sellest, kas olemas on kõik vajalikud andmed ja mis on vajalikud sammud enne andmete analüüsi alustamist. Andmetele rakendatud kitsendused:

- Kuriteoliik – röövimine (KarS § 200).
- Kriminaalaja staadium – kriminaalasi on saadetud prokuratuuri kohtueelse menetluse kokkuvõttega. Selline piirang suurendab analüüsivate andmete usaldusväärsust, kuna valimist jäävad suurema tõenäosusega automaatselt välja ekslikud kahtlustused.
- Andmete olek – ütlused peavad olema elektroonilisel kujul, kuna vaatluse all on automatiseeritud analüüsivõimalused, mis eeldavad andmete olemasolu elektroonselt.
- Ajaline piirang – ütlused ei ole vanemad kui 01.01.2009, piirang tuleneb elektroonilisel kujul olevate andmete kättesaadavusest.
- Andmete allikas – kriminaalajas koostatud toiming „Kahtlustatavana ülekuulamise protokoll“ (tekstiline väli „Ütlused“), sai valitud, kuna toimepaneku viisi (MO) kohta saab kõige rohkem informatsiooni kahtlustatava ütlustest.

Juba probleemi tõstatamise juures oli teada, et tegemist ei ole struktureeritud andmetega, millest tulenevalt on vaja arvestada tekstilisel kujul olevate andmete täiendavate erisustega (struktuuri puudumine, vigased andmed jms).

Käsitsi analüüsi läbiviimise juures jõuti järeldusele, et otstarbekas on teha täiendav samm enne MO'd kirjeldava andmestiku kodeerimist ja lugeda kõigepealt läbi kõik dokumendid, et saada kätte üldine pilt ning leida sobivaim suund edasiminekaks. Etapi tulemusena selgus:

- On vaja korrigeerida valimit (mittesobivate andmete eemaldamine valimist), sest selgus, et eesmärgist tuleneva andmete kvaliteedi tagamiseks (kuriteotüübile omaste ühiste tunnusoonte – mustrite leidmiseks) tuleb eemaldada osa andmeid järgmistel põhjustel:

- Esitatud kvalifikatsioon ülekuulamises ei vastanud valimi kitsenduseks olnud kvalifikatsioonile või kui isikut kuulati üle mitme erineva kuriteo osas, siis jäeti alles ainult röövimist kirjeldav sündmus.
- Isik keeldus ütlustest või järgmise ütlusega tunnistas, et eelmine antud ütlus oli valeütlus.
- On vaja luua nõ politseiline sõnastik, et automatiseeritud tekstianalüüsi läbiviimisel ei läheks kaduma (eelkõige) kuidas kuritegu pandi toime, kuna kirjeldustes kasutatakse palju kujundlike väljendeid, slängi jms. Näited võimalikust sisendist politseilisesse sõnastikku on lisatud tööle (LISA 3 Sisend valdkondlikku sõnastikku).

Kokkuvõttes tekkis kindlus valimi ja esitatud probleemi osas ning järgmise sammuna toimus andmete täpsem ettevalmistamine.

2.3. Andmete ettevalmistamine

Antud sammu eesmärk on eraldada vajalikud andmed ning viia need analüüsi läbiviimiseks vajalikule kujule.

Peale eelmises sammus toimunud valimi kitsendamist toimus andmete ettevalmistamise käigus valimisse jäänud toimingute „Kahtlustatavana ülekuulamise protokoll“ tekstiliselt väljalt „Ütlused“ antud vastuste eraldamine. Käsitsi läbiviidava analüüsi mahu piiramiseks ja võrdlevate tulemuste saamiseks tehti otsus, et kodeeritakse teksti seda osa, mis kirjeldab kuriteo toimepanemist (MO) ja kõrvale jäeti kuriteoeelne ja -järgne tegevus, mis oli ütlustes erineval määral kaetud. Stsenaariumipõhisel kuriteomustri analüüsil tuleb aga pidada silmas, et kuriteoeelne ja -järgne tegevus on samuti olulised.

Andmete töötlemisel ei kasutatud automaatseid lingvistilisi vahendeid, millega oleks eelnevalt toimunud sõnade algvormide leidmine ja nende ühestamine jm tegevused ning nende alusel teksti kodeerimine. Samuti tegeleti käsitsi tekstides esineva „müraga“ ehk jooksvalt korrastati andmetes olevaid vigu, mõisteti kasutatud lühendeid ja slängi jms.

Etapi tulemuseks on korrastatud valimi põhjal tehtud andmete väljavõte, mis on aluseks modelleerimisele ja analüüsitegevuste läbiviimisele. Korrastatud andmed võimaldavad hakata nende põhjal looma järgmises punktis kirjeldatud mudelit.

2.4. Mudeli loomine

Antud sammu oodatavaks tulemuseks on valitud meetoditega püstitatud probleemile vastuse leidmine. Eesmärk oli olemasolevate andmete pealt näha, milline võiks olla röövimiste *modus operandi* ning kirjeldada andmete põhjal lihtsustatult kujunevad mustrid. Tekstidest oluliste mõtete ja terminite eraldamiseks viidi läbi järgmised tegevused:

- Andmete kodeerimine lähtudes kontentanalüüsi põhimõtetest ja eelpool toodud kriminaalanalüüsi meetoditest.
- Toimunud kuriteosündmuse aja eraldamine tekstist aeganalüüsi läbiviimiseks.
- Seoste visuaalne esitamine.

Lõplikusse valimisse jäänud tekstide teistkordse lugemise jooksul toimus ka tekstide kodeerimine. Kodeerimisjuhendi koostamisel lähtuti avatud kodeerimise põhimõttest ning esialgne kodeerimisjuhend koostati 3K+3M meetodi järgi kirjeldatud küsimustest: mis, kuidas, miks, kus, millal ja kes. Koodid täpsustusid kodeerimise käigus ning lõplik kodeerimisjuhend on tööle lisatud (LISA 1 Kodeerimistabel). Kujunenud kodeerimisjuhendi ja tulemuste suurema valiidsuse tagamiseks viidi läbi valimi (20% ulatuses) samade tekstide ja sama juhendi alusel kodeerimine veel kahe isiku poolt.

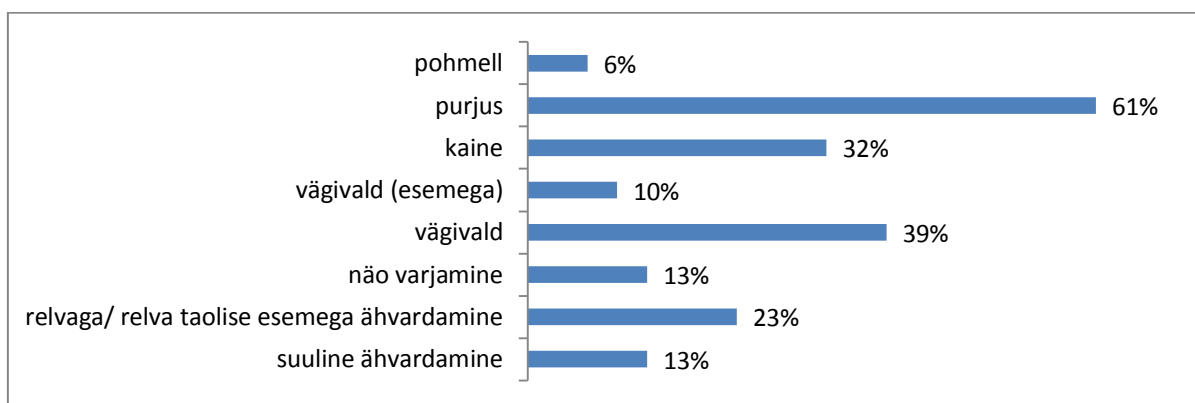
Kodeerimise käigus kujunes arusaam, et küsimusele „Mis toimus?“ on kahtlustatavana ütluste põhjal keeruline vastust anda, kuna ütlustest ei selgu selgelt näiteks isiku plaan, et „otsustasime minna röövima“, sest pigem kirjeldatakse ütlustes enda poolseid tegevusi, kuidas kuritegu toime pandi. Sellest tulenevalt otsustati, et vastused küsimusele „Mis?“ kodeeritakse politsei poolt kahtlustatavale esitatud kahtlustuse järgi, kuna see kõige selgemini väljendas seda, mis kuritegu pandi toime. Nende andmete saamiseks tuli teha täiendav andmete väljavõte.

Kodeerimise käigus selgus, et kodeerimisest on mõtet jätta välja küsimus „Kes?“, sest praeguse mudeli rakendamise eesmärk ei olnud eraldada kõiki kuriteos figureerivaid isikuid. Põhjalikuma analüüsi puhul on kindlasti oluline, et eraldatakse kõik isikud koos nende seoste ja muude kirjeldustega.

Neid kitsendusi arvestades visualiseeriti kodeeritud andmestiku (kuidas?, miks?, kus? ja millal?) tulemused, et näha tunnuste omavahelisi seoseid tööle on lisatud üldine seostepuu

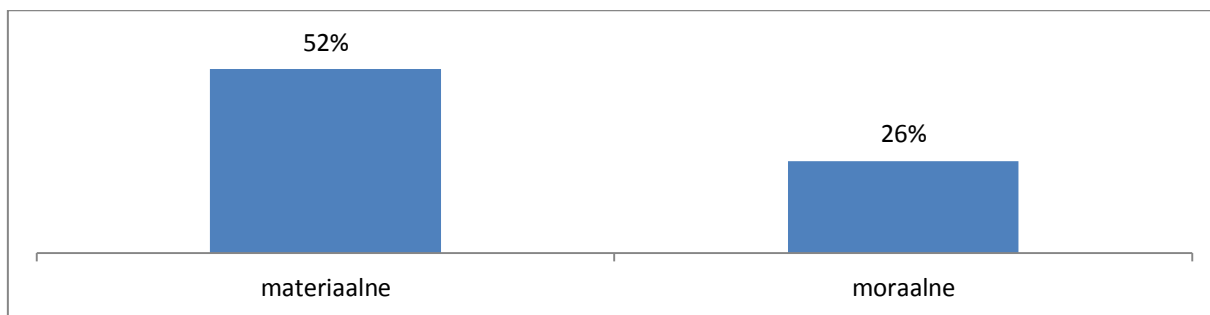
(LISA 4 Seosed graafiliselt). Visualiseerimise tulemusena oli võimalik teha esimesed kokkuvõtted ning näha oli, et valitud kuriteoliigiga enim seotud väärtused oli „purjus“, „asukoht“ (sündmuskoht on olnud mingis ettevõttes (nt: kauplus, bensiinijaam jms)) ja „õhtu“. Järgnevalt kõigepealt üldine ülevaade tulemustest.

Küsimuse „Kuidas?“ (Joonis 6) alla koguti kodeerimisel parameetrid, mis iseloomustavad kuidas kuritegu toimepandi. Kuna röövimisega kaasneb vägivald või sellega ähvardamine, siis kodeerimise aluseks oli eraldada nii suulise kui relva või relvataolise esemega tehtud ähvardamised. Toime pandud vägivalda juures eristati füüsilisest vägivaldast vägivald, mis pandi toime esemega. Kuna ütluste esmasel lugemisel selgus, et kuritegusid sooritades on isikud sageli alkoholijoobes, siis sai ühe mõõtmena kirjeldatud ka kahtlustatava olek (kaine, purjus, pohmell) kuriteo toimepanemise ajal.



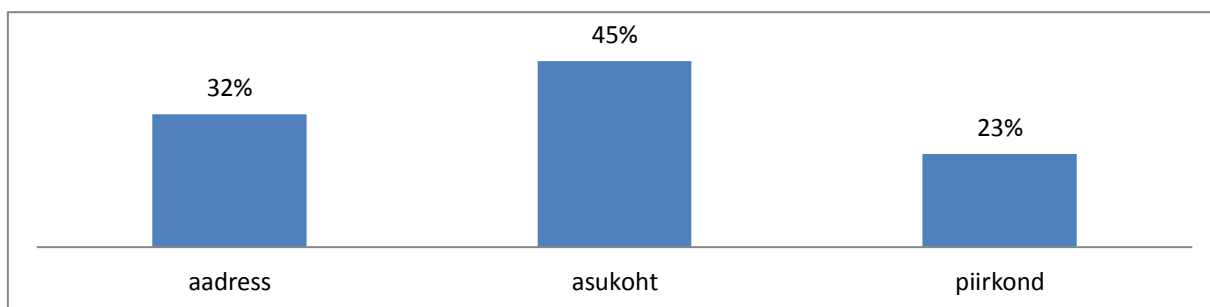
Joonis 6 Küsimuse "Kuidas?" kodeerimise tulemused

Küsimuse „Miks?“ (Joonis 7) alla koguti kodeerimisel väärtused, mis näitavad põhjust/ajendit, miks kuritegu hakati toime panema. Röövimise ajendiks on omakasu saamine ning sellest tulevalt sai kirjeldatud kategooriad (nt: raha, väärtuslik ese jms). Peamiseks ajendiks oli vajadus saada raha, kuid olid ka alkoholi ja söögi saamise vajadus (koondatud väärtusesse „materiaalne“). Esmasel tekstide lugemisel selgus, et sageli ei ole ajendiks varalise kasu saamine, vaid algselt pannakse toime muu kuritegu, mis lõpuks muutub röövimiseks. Sellest tulenevalt oli otstarbekas tuua mõõtmena juurde ka kahtlustatava nõ moraalne ajend (nt: vihastamine ja solvumine).



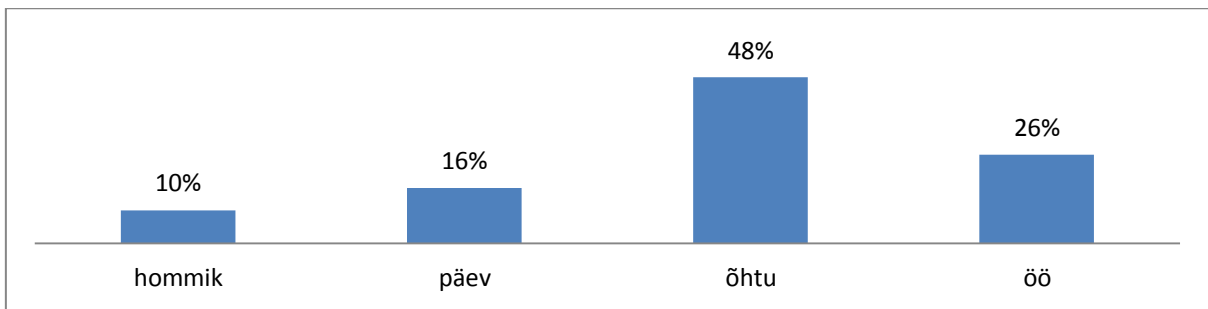
Joonis 7 Küsimuse "Miks?" kodeerimise tulemused

Küsimuse „Kus?“ (Joonis 8) alla koguti kodeerimisel kokku sündmuse toimumise tegevusruum, kus kuritegu toime pandi. Kodeerimisel jagati toimumiskohad kolme suurde gruppi, et eristada eluruumides toimunud kuritegusid (kategooria „aadress“), ettevõtetes (kauplus, bensiinijaam, ööklubi jms) toimunud kuritegusid (kategooria „asukoht“) ning tänaval või laiemalt määratletud alal toimunud kuriteod (kategooria „piirkond“).



Joonis 8 Küsimuse "Kus?" kodeerimise tulemused

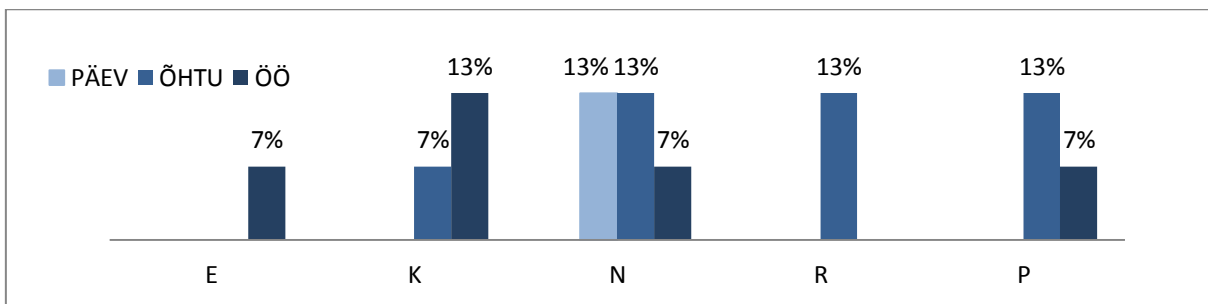
Küsimuse „Millal?“ (Joonis 9) alla kodeerimisel koondatud väärtused on oluliseks kuriteomustrite identifitseerimise ja kirjeldamise osaks (Clarke jt, 2006). Kõigepealt vaadeldi, kuidas toime pandud kuriteod ööpäeva osade vahel jaotuvad. Kodeerimisel jaotati ööpäev neljaks osaks ning eelkõige lähtuti sellest, mida ütlustes öeldi, st ei tehtud määramist kellaajapõhiselt. Kui muud moodi ei olnud võimalik ööpäeva osa määrata, siis võeti aluseks nimetatud kellaeg ning kodeerimisel lähtuti: hommik 07.00-10.59; päev 11.00-16.59; õhtu 17.00-22.59; öö 23.00-06.59. Ööpäeva osa võis jääda ka määramata, kui seda ühegi eelpool nimetatud meetodiga ei olnud võimalik üheselt mõistetavalt määrata.



Joonis 9 Küsimuse "Millal?" kodeerimise tulemused

Kuna ütlustes on kellaag sageli antud ligikaudselt, siis täpse kellaaja järgi ajalist analüüsi praeguses faasis tegema ei hakatud. See on siiski oluline automatiseeritud süsteemi puhul suure hulga kuritegude andmete võrdlemisel ja juhul kui analüüs koostatakse lisaks ka muude kuritegu kirjeldatavate andmete põhjal.

Lähtuvalt andmetest leidis autor, et täiendavalt on mõistlik uurida nädalapäeva ja ööpäeva osa vahelist seost (Joonis 10). Antud valimi põhjal (ühe teo puhul valimist ei olnud võimalik määrata täpsemat aega kui aasta) selgus, et teisipäeval ja laupäeval ühtegi kuritegu toime ei pandud ning et suurim osa (33%) pandi toime neljapäeva päeval. Nädalapäevast olulisemat kaalu röövimiste toimepanemisel omab kellaag. Pigem pannakse kuritegusid toime õhtul ja öisel ajal (53%).

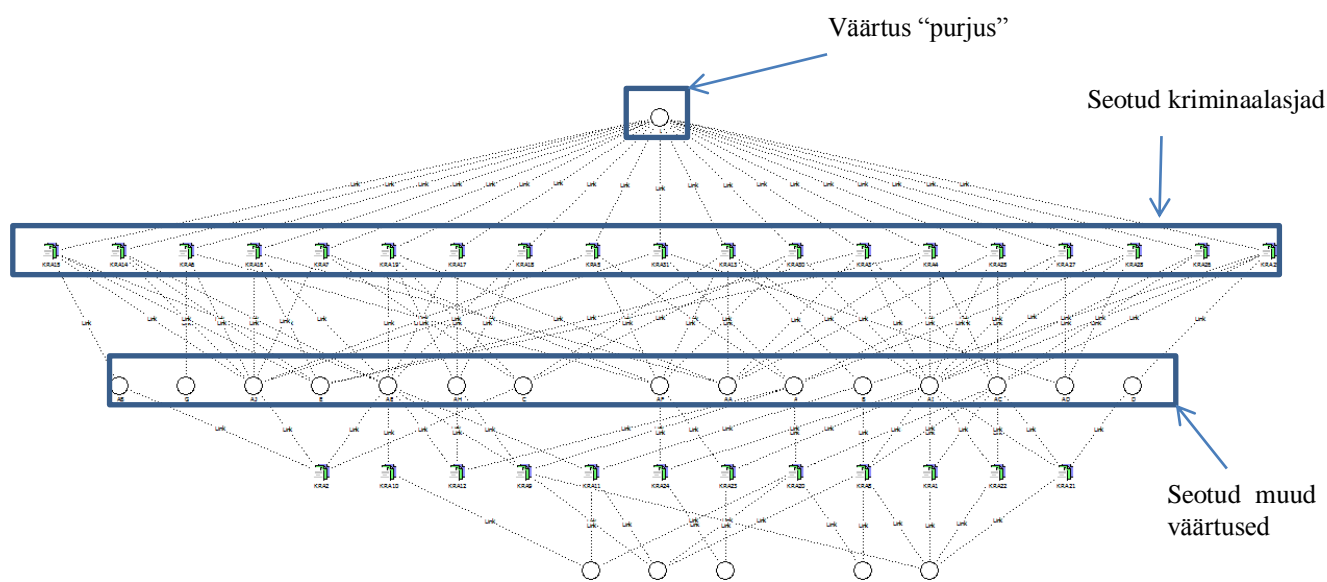


Joonis 10 Toimepandud kuriteod nädalapäevade ja ööpäeva osade lõikes

Kokkuvõtlikult andis küsimuste kaupa tehtud esmane analüüsitulemuste visualiseerimine ülevaate, kuidas valimi andmed nende väärtuste vahel jagunevad. Järgmiseks vaadeldi eelmainitud visualiseerimise tulemusena eristunud väärtuseid „purjus“, „asukoht“ ja „õhtu“ täpsemalt, et nende põhjal kujundada võimalikud kuriteomustrid.

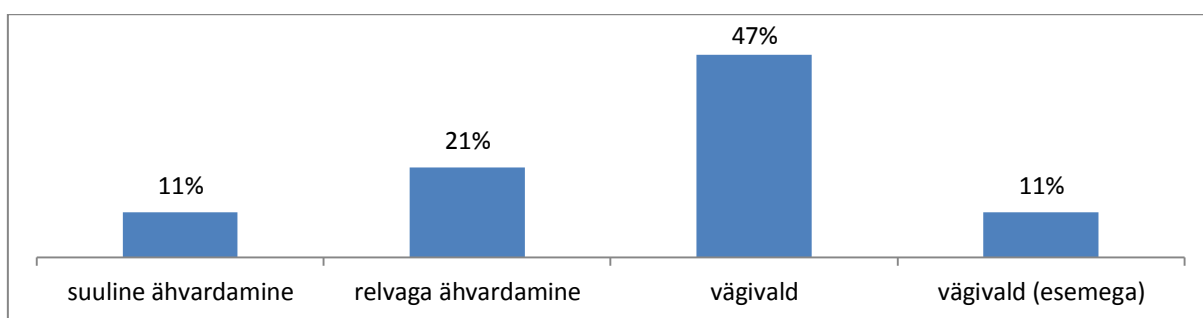
2.4.1. Väärtus „purjus“

Nagu eelpool öeldud ja tulemustest selgub (Joonis 6), on suureks kuritegude toime panemist mõjutavaks teguriks alkoholihoove. Lähtuvalt sellest vaadeldi lähemalt alkoholihooves toime pandud kuritegusid iseloomustavaid fakte. Väärtuse „purjus“ alusel kriminaalasjade seoste visualiseerimisel oli näha, et see oli üks enamesinenud seoseid (Joonis 11). Selle joonise ja edaspidiste sarnaste jooniste visualiseerimiseks kasutati tootja i2 tarkvara Analyst Notebook.



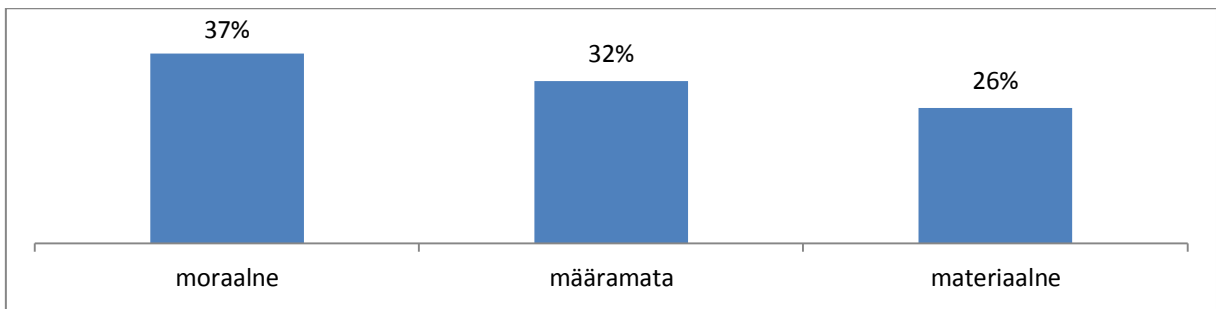
Joonis 11 Visualiseeritud kriminaalasjad, mille toimepanija oli purjus

Väärtusega „purjus“ koos esines vaadeldud andmetes enim (47%) füüsilist vägivalda kajastatavat väärtust (Joonis 12).



Joonis 12 Vägivaldaga toime pandud kuriteod joores olekus

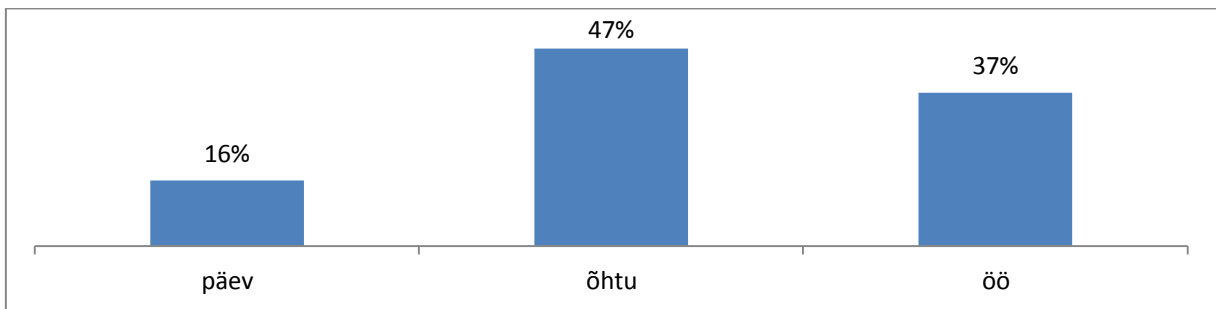
Andmete võrdlemisel lähtuvalt kuriteo toimepanemise ajendist – „miks?“ kuritegu toime pandi – näitab (Joonis 13), et alkoholihooves isikute poolt toime pandud kuritegude enim esinenud ajendiks on nõ moraalne (kannatanu oma käitumisega vihastas, kannatanu solvas kahtlustatavat) faktor.



Joonis 13 Joobes isikute poolt toime pandud kuritegude ajend

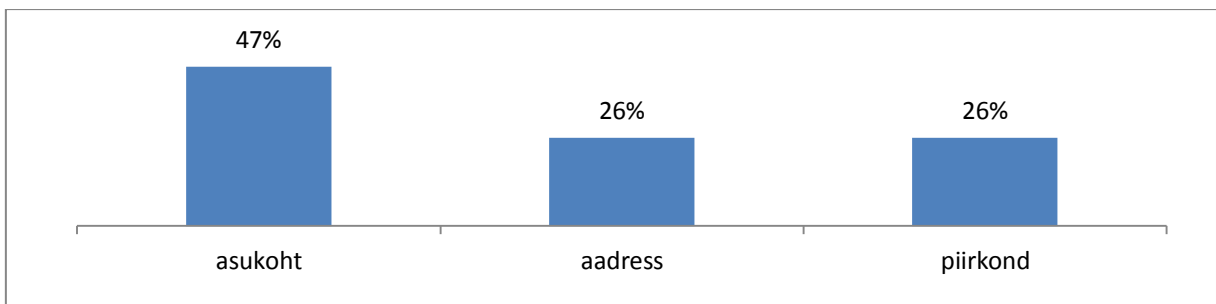
Kuna toimepandud kuritegusid iseloomustas vägivalla kasutamine, siis sellest on näha, et esialgu võis toimuda ähvardamine (§ 120) või kehaline väärkohtlemine (§ 121), mis on kuriteo toime panemise ajal muutunud röövimiseks, kuna kannatanult on lisaks vägivalla kasutamisele võetud ära ka esemeid.

Alkoholijoobes isikute poolt toime pandud kuriteod langevad suures osas õhtusesse ja öisesse (84%) aega (Joonis 14).



Joonis 14 Joobes isikute poolt toime pandud kuritegude jaotus ööpäevas

Need kuriteod on pandud toime peamiselt (47%) ettevõtetes (tinglikult nimetatud „asukoht“), mille alla on kokku võetud erinevad paigad (nt: kauplus, bensiinijaam, lõbustusasutused jms) (Joonis 15). Väärtuse „aadress“ alla on koondatud eluruumides toime pandud kuriteod ja väärtuse „piirkond“ alla on koondatud nõ tänaval toimunud kuriteod.



Joonis 15 Joobes isikute poolt toime pandud kuritegude toimepanemise kohad

Lähtudes eeltoodust saab kirjeldada näiteks järgmised potentsiaalsed joobes isikute poolt toime pandud röövimiste mustrid:

- moraalne (miks?), vägivald (kuidas?), „asukoht“ (kus?), õhtul (millal?)
- materiaalne (miks?), relvaga ähvardamine (kuidas?), „aadress“ (kus?), öö (millal?)

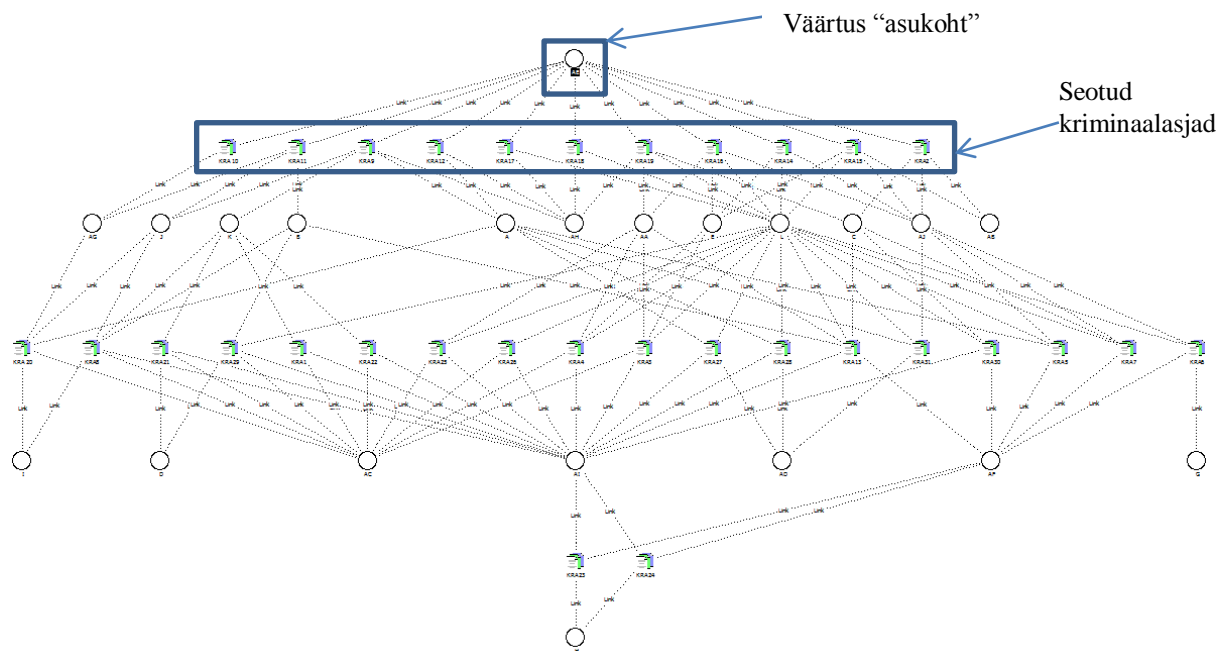
Näiteks võiks kuriteomustrile vastav kuriteokirjeldus olla järgmine:

„Kahtlustatav, olles alkoholijoobes, tundis, et teda solvati kannatanu poolt ning pani sellest tulenevalt õhtusel ajal vägivalda kasutades toime röövimise bussijaamas.“

Järgmisena tehti ülevaade seoste kohta lähtudes väärtusest, kus kuritegu toime pandi.

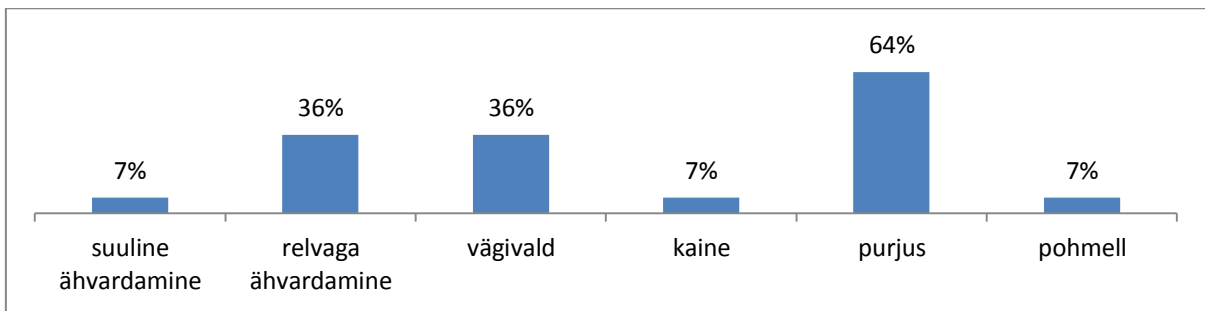
2.4.2. Väärtus „asukoht“

Väärtus „asukoht“ visualiseeriti (Joonis 16), et näha selle esinemissagedust. Valimisse sattunud kuriteod pandi 45% korral toime tinglikult nimetatud asukohas ehk mingis ettevõttes (nt: kauplus, bensiinijaam jms).



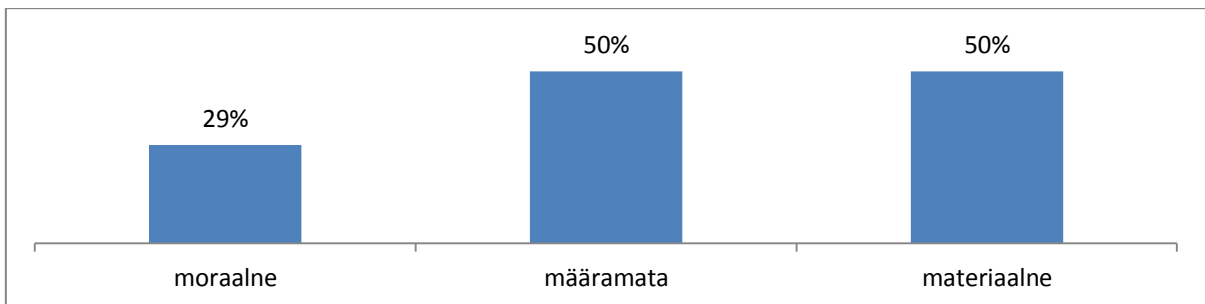
Joonis 16 Visualiseeritud kriminaalasjad, mille toimumiskoha väärtus oli "asukoht"

„Asukohas“ toimepandud röövimisi valimis iseloomustab eelkõige (Joonis 17) nende toime panemine vägivald või relvaga (või relvataolise esemega) ähvardades (kumbki 36%) ning joobes olek (64%).



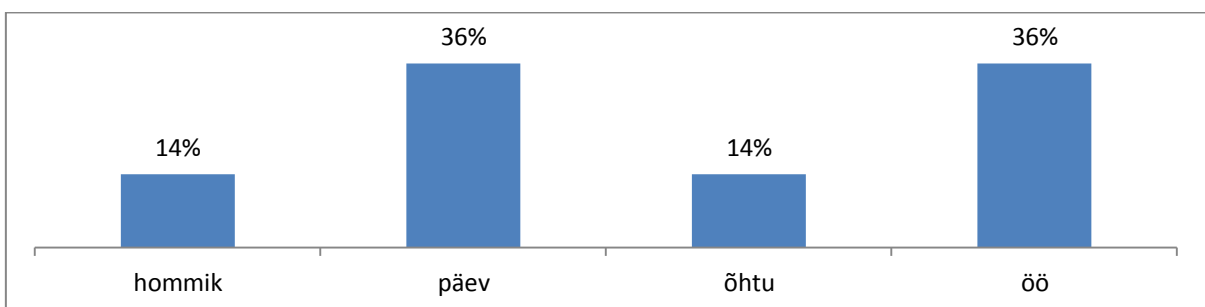
Joonis 17 Väärtuse „Asukoht” järgi röövimiste jaotumine toimepanemisviisi järgi

Andmete vaatamisel lähtuvalt kuriteo toime panemise ajendist – „miks?” kuritegu toime pandi – näitab (Joonis 18), et „asukohas” toime pandud röövimiste ajendiks on materiaalse kasu saamine (50%), mis on ootuspärane, kuna selle väärtuse alla lähevad ettevõtted – kauplused, bensiinijaamad jne, kuhu üldjuhul minnakse kuritegu sooritama konkreetse omakasu saamise kavatsusega.



Joonis 18 Väärtuse "Asukoht" järgi röövimiste jaotumine ajendi järgi

Ettevõttes toime pandud kuriteod langevad suures osas päevasesse (36%) ja öisesse (36%) aega (Joonis 19).



Joonis 19 Väärtuse "Asukoht" järgi röövimiste jagunemine ööpäevas

Lähtudes eeltoodust saab näiteks kirjeldada järgmised potentsiaalsed kuriteo toime panemise kohast tulenevad röövimiste mustrid:

- materiaalne (miks?), vägivald ja purjus (kuidas?), päev (millal?)

- määramata (miks?), relvaga ähvardamine ja purjus (kuidas?), öö (millal?)

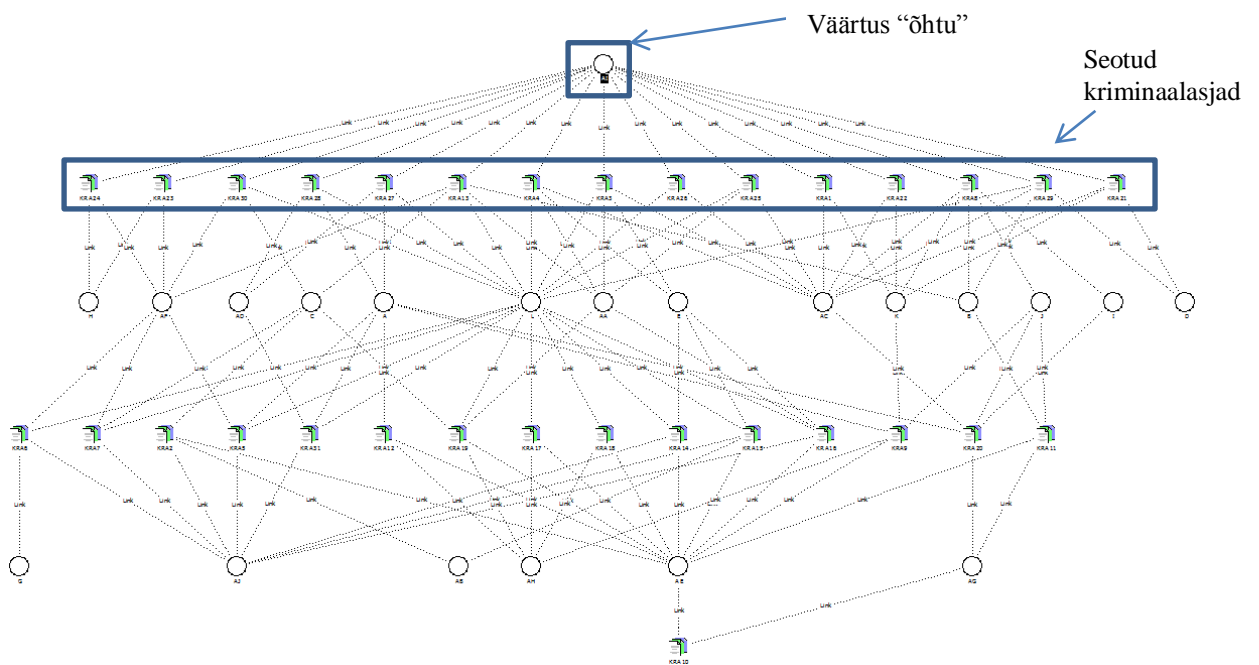
Näiteks võiks kuriteomustrile vastav kuriteokirjeldus olla järgmine:

„Eesmärgiga saada alkoholi on alkoholihoobes isik pannud päevasel ajal vägivalda kasutades toime röövimise“

Järgmiseks ülevaade seoste kohta lähtuvalt levinuimast ajalisest väärtusest.

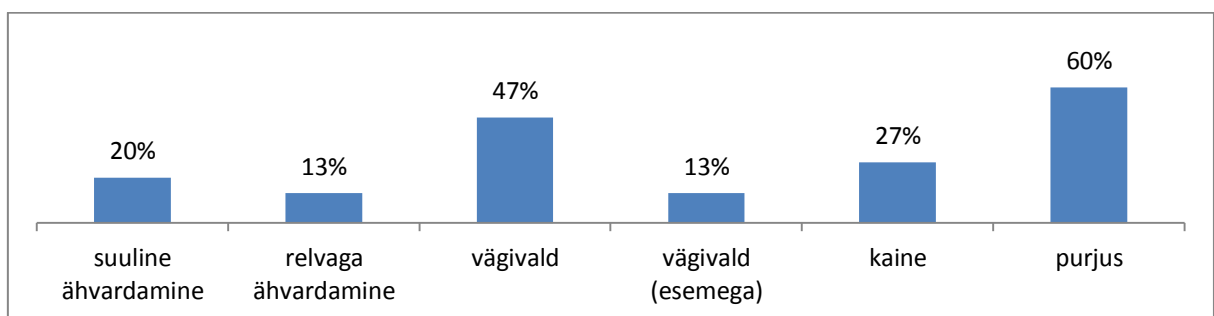
2.4.3. Väärtus „õhtu“

Ajaliste väärtuste võrdlemise tulemusena (Joonis 20) selgus, et enim pannakse käesoleva valimi põhjal kuritegusid toime õhtuti (84%).



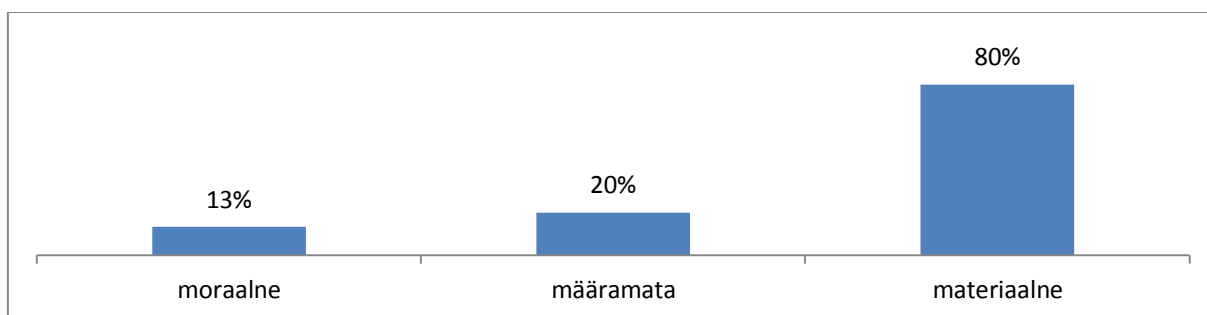
Joonis 20 Visualiseeritud röövimiste seosed lähtuvalt ajalisest väärtusest "õhtu"

Õhtusel ajal pandi enim kuritegusid toime vägivalda kasutades (47%) ja purjus olekus (60%) (Joonis 21).



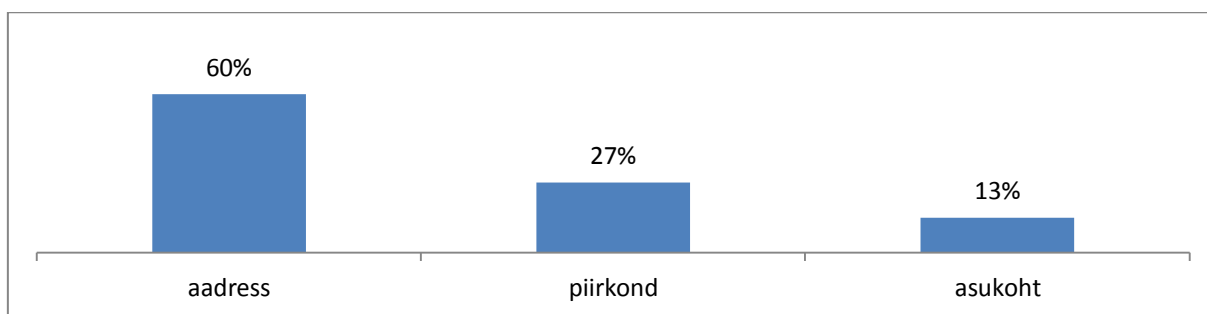
Joonis 21 Väärtuse "õhtu" järgi röövimiste jaotumine toimepanemisviisi järgi

Õhtusel ajal toime pandud kuritegude ajendiks (Joonis 22) on ülekaalukalt materiaalse kasu saamine.



Joonis 22 Väärtuse "õhtu" järgi röövimiste jaotumine ajendi järgi

Toimepanemise koha järgi (Joonis 23) pannakse õhtusel ajal röövimised toime pigem eluruumides (väärtus „aadress“).



Joonis 23 Väärtuse "õhtu" järgi röövimiste jaotumine toimumiskoha järgi

Lähtudes eeltoodust võib näiteks kirjeldada järgmised röövimiste toimumise aja järgi kirjeldatud mustrid:

- materiaalne (miks?), vägivald ja purjus (kuidas?), „aadress“ (kus?)
- materiaalne (miks?), suuline ähvardamine ja purjus (kuidas?), „piirkond“ (kus?)

Näiteks võiks kuriteomustrile vastav kuriteokirjeldus olla järgmine:

„Eesmärgiga saada raha on alkoholihoobes isik on pannud õhtusel ajal vägivalda kasutades toime röövimise eluruumides.“

Läbilugemise ja kodeerimise käigus märgistati ära ka sõnu ja väljendeid valdkondliku sõnastiku loomiseks, mille näited on toodud töö lisas (LISA 3 Sisend valdkondlikku sõnastikku).

Kokkuvõtvalt võrreldi andmeid kolme levinuima väärtuse lõikes. Tehtud üldistavad järeldused on võetud kokku järgmises punktis.

2.5. Tulemuste hindamine ja rakendamine

Eelmises punktis kirjeldatud sammude tulemusena tekkis ülevaade röövimist iseloomustavate väärtuste kohta ning defineeriti esmased lihtsustatud mustrid. Konkreetse mudeli loomine kahtlustatavana ülekuulamiste protokollide põhjal ja selle rakendamine toimus küll kitsendatult, kuid oli vajalik samm, sest selle tulemusena oli võimalik testida analüüsimudeli kriminaalanalüüsi teooriate rakendamist ning täiendada nõudeid automatiseeritud tekstianalüüsile.

Analüüsimudeli lihtsustatud rakendamisel kahtlustatavana ülekuulamise protokollide tulemusena selgus, et:

- Ootuspäraselt esines tekstides „müra“ – vigu, lühendeid jms, mida käsitsi kodeerimisel oli võimalik eirata, kuid millega on vaja kindlasti arvestada automatiseeritud tekstianalüüsi juures, et analüüsi tulemus ei kannataks.
- Vajadus on valdkondliku sõnastiku järele. Käsitsi kodeerimisel oli näha laia sõnade kasutust sündmuse kirjeldamisel ning palju kujundlike väljendite kasutamist (nt: toimus mingi mässamine), mida ei ole võimalik sünonüümide ühendamise kaudu automatiseeritult lahendada. Tekstides kasutati ka palju nõ valdkondlikku kõnekasutust (nt: vanglas midagi kokku keerama). Lisaks sellele selgus, et alati ootuspärane kuriteo toimepanemise ajend ei pruugi vastata sellele, millena kuritegu lõpuks kvalifitseeriti, (nt: röövimiste puhul oli ajendiks vihastamine). Automatiseeritud tekstianalüüsi juures tuleb selliste nüanssidega arvestada, et see info ei läheks tekstidest kaduma. Lahenduseks oleks valdkondliku sõnastiku loomine.
- Seoste visualiseerimine – 3K+3M küsimuste vastuste visualiseerimine annab lisaks tavapärasele objektidevaheliste seoste visualiseerimisele häid tulemusi ülevaatliku pildi saamiseks ja erinevate kuritegude vaheliste seoste illustreerimiseks (joonised: Joonis 6, Joonis 7, Joonis 8, Joonis 9). Analüüsimudeli rakendamise juures käesolevas

töös ei kasutatud kõikide isikute ja objektide eraldamist, kuid selle läbiviimine on automaatse tekstianalüüsi juures oluline, et tagada paremad võimalused kuritegude omavaheliseks sidumiseks.

- Sündmuse toimumise aja eraldamine – töös toimus algandmetes puuduva päevaosa määratlemine etteantud ligikaudsete ajavahemike järgi. Automatiseeritud tekstianalüüsi puhul on oluline, et sündmuse toimumise aeg ei jääks eraldamata ja paremate tulemuste saavutamiseks on ilmselt otstarbekas kasutada ka muid menetluse andmeid (ütlustes ei pruugi konkreetne aeg olla öeldud).
- Töö tulemuseks olid lihtsustatud kuriteomustrid, mis täitsid oma eesmärgi ja andsid lihtsustatult ülevaate erinevate tunnuste väärtuste omavaheliste seoste kohta. Automatiseeritud tekstianalüüsi läbiviimise juures on vaja need viia juba detailsemale tasemele.

Koostatud esialgne mudel on eelduseks taolise analüüsi rakendamisele järgmistes detailanalüüsi etappides juba detailsema mudeli loomiseks.

Mudeli rakendamise ja varasemate teooriate ning meetodite järelduste ja kokkuvõtete põhjal on järgmisesse peatükki koondatud nõuded automatiseeritud tekstianalüüsile.

3. NÕUDED AUTOMATISEERITUD TEKSTIANALÜÜSILE KURITEOMUSTRITE LEIDMISEKS

Peatüki eesmärk on koondada kokku töö eelmistes peatükkides esitatud nõuded automatiseeritud tekstianalüüsile kuriteomustrite leidmiseks ning olla sisendiks jätkuarenduste detailanalüüsile.

3.1. Automatiseeritud tekstianalüüsi eesmärk

Tulenevalt suurenenud ja keerukamaks muutunud andmemahtudest ning aegkriitilisusest on vaja tagada jätkusuutlik ning tulemuslik kriminaalanalüüs, et leida kuritegevuse mustreid ja oleks võimalikult kiirelt tagatud otsuste tegemiseks vajalik informatsioon. Selle protsessi nõuetekohaseks läbiviimiseks on vaja automatiseeritud tuge arvutite näol.

Käesolevas töös keskenduti tekstiliste andmete analüüsile, et tõsta politseis talletatud struktureerimata andmete kasutamise osakaalu kriminaalanalüüsi läbiviimisel eesmärgiga tuvastada nende põhjal kuriteomustreid ja nii kuritegusid, kuritegelikke gruppe ja kurjategijaid omavahel siduda. Loodud kuriteomustreid on võimalik „peegeldada“ uutele andmetele eesmärgiga näiteks suurendada kuritegude lahendamise määra või leida preventiivsed mehhanismid tulevikujuhtumite ennetamiseks.

Loodav automatiseeritud tekstianalüüsi lahendus peab võimaldama uurijatel kasutada seda rutiinselt kuriteomustrite tuvastamiseks ning erinevate sündmuste, asukohtade, isikute, aegade jms omavaheliseks sidumiseks.

Järgmises alapeatükis esitatakse üldised nõuded, mis on eelduseks järgmiste tarkvaraarenduse etappide läbiviimiseks ja rakenduse loomiseks ja mida on vaja täpsustada järgmistes detailanalüüsi etappides.

3.2. Üldised nõuded automatiseeritud tekstianalüüsile

Käesolevasse peatükki on koondatud nõuded lähtuvalt kriminaalanalüüsi teooriatest ja meetoditest, tekstianalüüsi läbiviimisest ning andmekaeve põhimõtetest ning käesoleva töö raames läbiviidud analüüsimudeli rakendamisel saadud kogemusest. Käesoleva töö teoreetilise ja praktilise osa tulemuste põhjal on süsteemi üldised nõuded järgmised:

- Tekstianalüüs ja -kaevandamine peaks politseis toimuma lähtuvalt valdkonna standardist CRISP, mis tagaks läbiviidavate tegevuste süsteemsuse: valdkonna kirjeldamine, andmete kogumine ja ettevalmistamine, mudeli loomine ja selle hindamine ning testimine. Iga

samm koosneb erinevatest järjestikustest sammudest, mille järjekord ja nõuded tuleb kirjeldada lähtuvalt politseis esitatud nõuetest ning reeglitest. Tulemuseks oleks politseis läbiviidavate andmekaevandamiste tekstiliste andmete osa reguleeriv kord.

- Käesoleva töö empiirilises osas piirduti vaid kahtlustatava ütluste peal teksti analüüsi läbiviimisega. Politsei käsutuses on hulgaliselt kuritegelikku käitumist kajastavat informatsiooni vabateksti kujul. Edasine detailanalüüs peab selgitama süsteemi seotuse politsei teiste süsteemidega, millisena on otstarbekas realiseerida automatiseeritud tekstianalüüs, st kas tegemist on iseseisva süsteemiga, kuhu andmeid laetakse ühekordselt või luuakse andmete ülekandmiseks liidesed, et veelgi automatiseerida protsessi.
- Luua sobilike vahendite kogum, millega oleks võimalik tekste enne analüüsimist korrastada või struktureerida, et neile oleks võimalik rakendada juba traditsioonilisi andmetöötluse ja -analüüsi meetodeid, näiteks: tekstide puhastamine (kirjavead, trükivad, lühendid, släng), kirjete ja atribuutide valikud jne ja müra eemaldamine tekstist (sidesõnad jms).
- Keeletehnoloogiline tugi – hädavajalik selleks, et tekste oleks võimalik automatiseeritult analüüsida. Kuna peamine tekstiliste andmete keel on eesti keel, siis on vajalik, et loodav lahendus toetaks eesti keele põhist morfoloogilist analüüsi.
- Töö empiirilises osas läbiviidud tekstide kodeerimine kinnitas, et kuritegude võrdlemiseks ja nende põhjal detailsemate kuriteomustrite kujundamiseks on väga tõhus vaja kuritegude andmed tükeldada lähtudes 3K+3M loogikast, millega eristatakse kuriteo toimepanemise ajend, kuidas kuritegu viidi läbi, kus ja millal see toimus ning kes osalesid kuriteo toimepanemise juures.
- Töö empiirilises osas läbiviidud tekstide kodeerimine ja näidissõnastiku loomine kinnitas, et süsteem peab toetama valdkondlike sõnastike kasutamist, millega tagatakse toimepandud kuritegusid kirjeldava andmestiku „mitte kadumine“ (st automatiseeritud süsteem ei oska neid eralda). Sõnastikke on vaja pidevalt täiendada kuriteoliiki iseloomustava sõnavaraga ning lisaks ootuspärasele kuriteo toimepanemise ajendile tuleb jälgida, et tähelepanuta ei jääks mittereeglipärased ajendid (nt: toimepandud röövimise esialgne ajend oli kannatanu moraalne riive)

- Oskus eraldada tekstist nimelisi objekte, et oleks võimalik kuritegusid omavahel läbi nende siduda. Eraldamine võib toimuda keelereeglite rakendamise või näiteks sõnastike kaasabil. Täpne lahendus peab selguma detailanalüüsi käigus.
- Sündmuste omavaheline ajalises ruumis võrdlemine, et kirjeldada kuriteomustreid toimepandud kuritegude ajalise koondumise alusel. Käsitsi analüüsi juures selgus, et valitud allikast rahuldaval kujul andmeid alati ei saanud. Automatiseeritud tekstianalüüsi läbiviimisel tuleb sellega arvestada, et oleks võimalik andmeid laadida mitmest allikast või realiseerida väärtuse arvutamine mingite tunnuste põhjal.
- Kaevandamisel lähtutakse kriminaalanalüüsi teooriatest (mida tekstist otsitakse: 3M+3K küsimust, POLE-mudel jne). Süsteem peab võimaldama tekstile rakendada kaevandamistehnikaid, millest sobivaimad on võimalik läbi testimise valida detailanalüüsi etapis. Käesolevas töös on toodud loetelu (punktis 1.3 „Tekstiliste andmete töötlemine“) võimalikest tehnikatest teiste riikide kogemusele toetudes.
- Tekstianalüüsi läbiviimise juures on oluline tulemuste visualiseerimine kasutajale arusaadaval kujul. Suurte andmemahtude juures on oluline, et tulemusi oleks võimalik visualiseerida näiteks objektide ja nende seoste lõikes. Tulemuste visualiseerimiseks võib kaaluda ka olemasolevate muude tarkvaraliste vahendite (nt: tootja i2 analüüsivahendid andmete visualiseerimiseks) integreerimist. Kuidas tehniliselt visualiseerimine lõpuks realiseeritakse, peab selguma detailanalüüsi käigus, kus on võimalik täpsustada, milliseid lisavõimalusi visualiseerimiselt oodatakse.
- Korduv mudeli hindamine läbi mudeli rakendamise etteantud parameetritega.

Kirjeldatud oodatud tulemuste realiseerimine automatiseeritult sõltub eelkõige keeletehnoloogilisest toest ning sobivate andmekaevandamise meetodite valikust.

Eelnevalt toodud loetelu nõuetest peab olema sisendiks tulevikus valmivale detailanalüüsile, kus esitatakse täpsem funktsionaalsuse loetelu.

Töö koostamise jooksul kujunenud võimalikud jätkuarendused on koondatud kokku järgmisesse peatükki.

3.3. Jätkuarendused

Struktureerimata andmete töötlemise eesmärk ei pruugi olla ainult kuriteomustrite tuvastamine, mis on vaid üks võimalus, mida on võimalik saavutada, vaid ka näiteks toetada päringute tegemist üle tekstiliste andmete, salvestada andmed struktureeritud andmebaasi juba struktureeritud kujul jms. Seega rakendusvaldkond on lai ja perspektiivikas. Tuleviku edasiarendusteks:

- Kuritegude automaatne menetlejate vaheline sidumine, kui potentsiaalselt huvipakkuvatena (Chen jt, 2002). Eesti väiksuse tõttu võib sageli olla oluline saada võimalikult varakult teada, kui sarnaste tunnustega kuritegu pannakse ka mujal.
- Efektivsem päringute tegemine üle tekstiliste andmete:
 - otsingusõnade väljapakkumine – (nt: *relevance-feedback*. Idee seisneb selles, et kasutaja annab pidevalt tagasisidet, millised saadud vastustest on olulised ja millised mitte, millest tulenevalt korrigeeritakse päringut, võttes positiivsest tagasisidest uusi märksõnu juurde. Päringu täpsuse osas tõstab see tulemuse 40%-lt 60%-le) (Chen jt, 1998).
 - Otsingute juures morfoloogiaga arvestamine – tulemuseks antakse sisu mõistes tihedalt seotud dokumendid ja jäetakse kõrvale vähem olulised dokumendid.
- Võimekus automaatselt vastata küsimusele „Kes tegi mida?“ – tekstianalüüsi tulemusena on võimalik saada teada isikud, kes panid toime kuriteo, millise kuriteo, mis oli nende ajend ja millal ning kus see toimus.
- Sündmuste toimumise ennustamine: sündmuse A ja B toimumise järel toimub sündmus C (Pherson, 2008).
- Kuriteostsenaariumite rakendamine võimaldab leida igal etapil (kuriteo ettevalmistamine, läbiviimine ja järeltegevused) ennetusmeetmeid, kus kuritegu jagatakse etappideks – ettevalmistamine, elluviimine ja järeltegevused.

Viimase peatüki eesmärk oli võtta kokku eelnevas töös esitatud teooriad ja meetodid ning kirjeldada nende põhjal automatiseeritud tekstianalüüsi jaoks üldistatud nõuded, et nende

põhjal oleks võimalik asuda looma detailset kirjeldust. Välja toodi ka töö koostamise käigus sõnastatud ja esile kerkinud võimalikud jätkuarendused ja haakuvad teemaatikad. Valdkonna aktuaalsus ja perspektiivikus leidsid kinnitust.

4. KOKKUVÕTE

Käesoleva töö sissejuhatuses tõdeti, et politsei jaoks on suurenenud andmemahud tõstatanud vajaduse uute ja tõhusamate analüüsivahendite leidmiseks, et jätkuvalt viia läbi ajakohast ja eesmärgipärast kriminaalanalüüsi politseis talletatavate andmete le.

Sellest tulenevalt püstitati töö eesmärgiks kuriteomustrite kujundamine tekstilistel andmetel lähtuvalt kriminaalanalüüsi teooriatest ja meetoditest, tekstianalüüsi läbiviimisest ning andmekaeve põhimõtetest, mille põhjal kirjeldati teoreetilise osa lõpus analüüsimudel.

Töö eesmärgi saavutamiseks viidi läbi analüüsimudeli rakendamine käsitsi ja vähendatud mahus, mille tulemusena tekkinud üldised nõuded automatiseeritud tekstianalüüsile sõnastati töö viimases osas.

Käesoleva tööga taheti saada ülevaade probleemi keerukusest ja üldistest nõuetest automatiseeritud tekstianalüüsile, mida ka õnnestus saada. Kokkuvõtlikult võib öelda, et automatiseeritud tekstianalüüsi funktsionaalsuse loomine annaks väga suurt kasu politsei jaoks, kuna andmeid oleks võimalik suurtest andmemahtudest olenemata kiiremini töödelda kui seda võimaldaks käsitsi analüüsi läbiviimine. Käesoleva töö põhjal võib öelda, et loodava süsteemi suurimaks nõudeks on eesti keele morfoloogilise toe olemasolu, et tekstianalüüs annaks oma oodatavat kasu. Lisaks teksti „mõistmisele“ on oluline leida ja kasutada sobilikke andmekaevandamise tehnikaid.

Järgnevad detailanalüüsi ja rakendamise etapid peaksid käesolevas töös püstitatud üldiseid nõudeid viima täpsemale tasemele, et nende põhjal oleks võimalik luua toimiv rakendus.

5. RESÜMEE

Requirements of Automated Text Analysis for Identification of Crime Patterns

Kai Jääger

Summary

In the foreword of the current thesis a statement is made that the ever increasing volume of data has forced the police to look for better and more efficient analysis tools to continue to perform up to date and result-oriented criminal analysis of the data kept in the police information systems.

The crime data registered in the police contains valuable information, *modus operandi* of which could be used for example to describe crime patterns and connect the different crimes committed by a criminal. Often this information is as far as the database is concerned kept in a non-structured form or free text.

Based on this fact the goal of this thesis was set to design crime patterns from textual data based on criminal analysis theories and methods, text analysis and data mining principles which were described in the theoretical part of this thesis.

Based on the analysis theories and CRISP-DM standard a preliminary analysis model suitable for police needs was designed. To test the analysis model and specify the general requirements for the automated text analysis a crime pattern description exercise based on the suspect interrogation protocols was performed in the second part of the thesis. The exercise was performed by manually coding text following the principles of content analysis. The coded data was visualized using the vendor i2's software Analyst Notebook and based on the coding results preliminary simplified crime patterns were presented.

In the third part of the thesis the general requirements for the automated text analysis were described based on the theoretical and practical part of the thesis, which will in turn act as input for the next phases of software development and application design.

The conclusion follows that the police would very much profit from the availability of automated text analysis functionality because it would enable to process data quicker than manual analysis would allow in spite of the enormous data quantities present. This thesis confirms that to produce the most benefit the most important requirement for this application

is the availability of Estonian language morphology support. In addition to “understanding” the text it is important to find and use suitable data mining techniques.

The phases of detailed analysis and implementation that follow should take the general requirements presented in this thesis to the next level leading to a working application.

KASUTATUD KIRJANDUS

1. Adderley, R. W., „The Use of Data Mining Techniques in Crime Trend Analysis and Offender Profiling“, Doctoral Dissertation, University of Wolverhampton, veebruar 2007
2. Boba, R., “Crime Analysis and Crime Mapping: An Introduction“. Thousand Oaks, CA: Sage Publications, 2005
3. Berson, A., Smith, S., Thearling, K. „An Overview of Data Mining Techniques“, excerpted from the book „Building Data Mining Application for CRM“, <http://www.thearling.com/text/dmtechniques/dmtechniques.htm> (14.12.2010)
4. Brantingham, P., Brantingham. P. ”Patterns in Crime”, Macmillan Inc, 1984
5. Chau, M., Xu, J.J., Chen, H. „Extracting Meaningful Entities from Police Narrative Reports“, Proceedings of the 2002 annual national conference on Digital government research, 2002
6. Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J.J., Wang, G., Zheng, R., Atabakhsh H. „Crime Data Mining: An Overview and Case Studies" II Proceedings of the National Conference for Digital. Government Research (dg.o 2003), May 18-21, 2003, Boston, Massachusetts, pp. 45-48.
7. Chen, H., Chung, W., Qin, Y., Chau, M „Crime Data Mining: A General Framework and Some Examples“, IEEE Computer Society, 2004
8. Chen. H., Shankaranarayanan, G., She, L. „A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing“ Journal of the American society for information science, 49(8):693–705, 1998
9. Clarke, R., Eck. J., „Probleemikeskseks kriminaalanalüütikuks: 55 lihtsa sammuga“, Tallinn 2006
10. CRSIP-DM 1.0, 2000, www.crisp-dm.org/Process/index.htm (14.12.2010)
11. Eesti keele seletav sõnaraamat, <http://www.eki.ee/dict/ekss/> (21.12.2010)

12. Ekblom, P. "Getting the Best out of Crime Analysis". Crime Prevention Unit: Paper 10. London Home Office", 1988
13. Grover, V., Adderley, R., Bramer, M. „Review of Current Crime Prediction Techniques“, University of Portsmouth, UK, 2006
14. Hulth, A. „Coming Machine Learning and Natural Language Processing for Automatic Keyword Extraction“ Doctoral Dissertation, Stockholm University, aprill 2004
15. Kalmus, V., „Tekstianalüüsi meetodid“, loengukonspekt, 2000
16. Kollepara, V., Ananyan, S. "Crime Pattern Analysis", Megaputer Case Study in Text Mining, www.megaputer.com, 2002
17. Krippendorf, Klaus "Content Analysis. An Introduction to its Methodology", Thousand Oaks: Sage, 1980
18. Liiv, I. "Andmekaevandamine", A&A, vol 5, 2005, pp. 28-43, ISSN 1406-345X
19. Muischnek, K., Orav, H., Kaalep, H.-J., Õim, H. „Eesti keele tehnoloogilised ressursid ja vahendid: arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara“, Eesti Keele Sihtasutus, Tallinn 2003
20. National Policing Improvement Agency (NPIA) „Guidance on the Management of Police Information“, 2010
21. Pherson Associates, "Handbook of Analytic Tools & Techniques", Pherson Associates, LLC, 2008
22. Poyner, B., „A Model for Action', in Heal and Laycock (eds.), Situational Crime Prevention: From Theory into Practice“, Her Majesty's Stationery Office, London, 1986
23. Ratcliffe, J. „Aoristic Signatures and the Spatio-Temporal Analysis of High Volume Crime Patterns“, Journal of Quantitative Criminology, no. 18 (1), pp. 23-43, 2002
24. Schroeder, J., Xu, J., Chen, H., Chau, M. "Automated Criminal Link Analysis Based on Domain Knowledge" Journal of the American Society for Information Science and Technology, 58(6):842-855, 2007

25. Sharp, M., "Text Mining", Seminar in Information Studies
http://comminfo.rutgers.edu/~msharp/text_mining.htm, 2001, (21.12.2010)
26. Weber, R.P., „Basic Content Analysis“ Newbury Park: Sage, 1990
27. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., jne "Top 10 algorithms in data mining",
Known Inf Syst (2008) 14:1-37

LISA 1 Kodeerimistabel

1. Mis toimus – kuriteo kvalifikatsioon:

1. lg2p7	4. lg2p7p8p10	7. lg2p7p9
2. lg2p4p7	5. lg2p8	8. lg2p7p8
3. lg2p4p7p8	6. lgp8p10	9. lg2p10

2. Kuidas kuritegu pandi toime:

Ähvardamine:

1. ähvardamine puudus	2. vägivaldaga ähvardamine – suuline	3. vägivaldaga ähvardamine - relvaga/ relva taolise esemega
-----------------------	--------------------------------------	---

Löömine/peksmine

1. vägivald puudus	3. löömine millegagi	6. pipragaasi laskma
2. löömine (peksmine jms) - nõ kätega	4. kätega-grupp	
	5. esemega-grupp	

Alkoholi joove:

1. kaine	2. purjus	3. pohmell
----------	-----------	------------

3. Miks kuritegu pandi toime:

Varalisel eesmärgil:

1. raha	2. süüa/juua saada	3. muu
---------	--------------------	--------

Moraalne:

1. vihastamine	2. solvamine
----------------	--------------

5. Kus kuritegu pandi toime:

1. aadress (eluruum)	2. asukoht (kauplus, ööklubi, bensiinijaam jms)	3. piirkond (tänav, linnaosa, jms)
----------------------	---	------------------------------------

6. Millal kuritegu pandi toime:

Kellaeg

1. Määramata	2. Täpne kellaeg	3. Ligikaudne kellaeg
--------------	------------------	-----------------------

Kuupäev:

1. Määramata	2. Täpne kuupäev	3. Ligikaudne kuupäev
--------------	------------------	-----------------------

Kuu:

1. Määramata	2. Täpne kuu	3. Ligikaudne kuu
--------------	--------------	-------------------

Aasta:

1. Määramata	2. Täpne aasta	3. Ligikaudne aasta
--------------	----------------	---------------------

Ööpäeva osa:

1. määramata	3. päev	5. öö
2. hommik	4. õhtu	

LISA 2 Karistusseadustiku paragrahv 200

(RT I 2001, 61, 364):

§ 200. Röövimine

(1) Võõra vallasasja äravõtmise eest selle ebaseadusliku omastamise eesmärgil, kui see on toime pandud vägivallaga, – karistatakse kahe- kuni kümneaastase vangistusega.

(2) Sama teo eest, kui:

- 1) teo objektiks on tulirelv, laskemoon, lõhkeaine või kiirusallikas;
- 2) teo objektiks on narkootiline või psühhotroopne aine või nende lähteaine;
- 3) teo objektiks on suure teadusliku, kultuuri- või ajalooväärtusega ese;
- 4) see on toime pandud isiku poolt, kes on varem toime pannud röövimise või tapmise seoses röövimisega või muul omakasu motiivil või väljapressimise;
- 5) see on toime pandud raske tervisekahjustuse tekitamisega;
- 6) see on toime pandud suures ulatuses;
- 7) see on toime pandud grupi või kuritegeliku ühenduse poolt;
- 8) see on toime pandud relva või relvana kasutatava muu esemega või sellega ähvardades;
- 9) see on toime pandud sissetungimisega;
- 10) see on toime pandud näo varjamisega näokatte või maskiga või muul viisil, mis takistas isiku tuvastamist, – karistatakse kolme- kuni viieteistaastase vangistusega.

LISA 3 Sisend valdkondlikku sõnastikku

Valimisse sattunud tekstide lugemise ja kodeerimise käigus kujunes ülevaade kasutatava sõnavara kohta, mille põhjal tuleks kaaluda valdkondliku sõnastiku koostamist.

Allpool on esitatud mõned näited kasutatud sõnadest ja väljenditest.

KUIDAS?:

Sukad või maskid peas, müts tõmmatud näo ette (eesmärk teadlikult varjata nägu), „hunt kodu juurest ei murra“, püstolit või relva suunama, noaga vehkima, tungisid peale, patsutas tema taskutes (läbiotsimiseks), tuulas mööda tuba ringi, salaja sees käima (vargil käima), kähmlus, rüselus, füüsiline madin, mässamine (ehk kaklus), lendama maha (ma kukkuma), käed/jalad kinni teipima/siduma, põrutada saanud (peksa saanud, peksmise tagajärg), sõnelema (vaidlema), lõugama, „teeme ta paljaks“ (röövima), veits raha olema, hakkas midagi koitma (meelde tulema), salgasin kõike (tehtu eitamine) jne.

Toimepanemise vahend: õhupüstol/relv, püstol, relv, mängupüstol/püstol ei olnud ehtne jne.

MIKS? rahaline seis vilets, laenu on peal, raha võlgu küsima, küsima raha, (raha) ei ole kusagilt võtta, pohmakas, ei tea mis pähe tuli, viha, otsustasin sõimamise eest karistada, vajun täiesti ära (peale alkoholi tarbimist), süüdlasena tundma, väljakutsuvalt käituma, ei taha endale probleeme, ei ole neid jamasid vaja, patust eemale hoida, vanglas midagi kokku keeranud, vana põhi oli all (st ennem alkoholi tarbinud), peaparandust (jooki) jne.

Seotud isikud: tundmatud isikud, noormees, organiseerija, mendid, politsei, sõber, tuttav, võõras jne.

LISA 4 Seosed graafiliselt

Kogu kodeeritud andmestiku kujutamine graafilise seostepuuna:

