TARKVARA

ANDMEKAEVANDAMINE

SISSEJUHATUS

Andmete maht, mida oleks vaja töödelda, kasvab pidevalt. Kümmekond aastat tagasi peeti utoopiliseks analüüsitavaks andmemahuks gigabaite. Tänaseks haldavad paljud rahvusvahelised korporatsioonid terabaitides andmeid, suurematel neist tuleb juba opereerida petabaitidega. Ilmselt ei ole siin midagi imestada – andmete maht peabki kogu aeg kasvama, sest salvestatakse pidevalt ajalugu (mis peab jäädvustuma, isegi kui hetkel ei osata seda kõige otstarbekamalt kasutada) ning tänast päeva, mille kohta iga hetk aina rohkem informatsiooni talletatakse. Lihtsaimaks näiteks on ostukeskused – iga triipkoodiga kassast läbi läinud kaup salvestatakse unikaalselt andmebaasi, nõnda kõikide kassade ning ketis olevate poodidega. Mõistagi muutub taolise andmehulga analüüs pidevalt keerulisemaks. Seetõttu oleks andmehulkade paremaks töötlemiseks hädasti tarvis uusi suundi ning ideid, s.t mitte ainult optimeeritumaid algoritme, vaid pigem erisuguseid lähenemisviise. Suurenenud informatsiooni-hulka oleks tarvis pöörata koondatud teadmiseks ning ellu rakendada. Valdkonda, mis just selle probleemiga tegeleb, nimetatakse andmekaevandamiseks.

ANDMEKAEVANDAMISE OLEMUS

"Mine sinna – ei tea kuhu, too seda – ei tea mida!" (Vene muinasjutt)

Andmekaevandamine (*data mining*) ei erine põhimõtete poolest traditsioonilisest kaevandamisest – teatud oskusteabe ning vahendite abil püütakse pinnasest kätte saada väärtuslikke maavarasid. Andmete puhul täpselt samamoodi: kindla oskusteabe ning vahendite abil proovitakse suurest andmehulgast kätte saada väärtuslikku ning isegi ootamatut informatsiooni. Eesmärgiks viimasest saadud teadmist edukalt ka mingis kindlas valdkonnas rakendada.

Julgeksin üsna kindlalt ka väita, et niisama loomulik kui täna on traditsioonilises kaevandamises erinevad load, keelud ja piirangud, ei ole väga kaugel ka ajaloo kordumine andmekaevanduse kontekstis. Privaatsuspoliitika ning paranoiline (paraku tihti mitte ka alusetu) hoiak erinevate eraeluliste andmete kogumisele on juba käesoleval hetkel tekitanud avalikke diskussioone erinevate andmete analüüsimise eetikas, seda ka Eestis. Tõepoolest, iga turundusdirektori unistus oleks ekspluateerida klientide alateadvust ning panna neid seeläbi rohkem ostma. Lähemalt selle üle arutledes võib siiski jõuda tõdemuseni, et tegelikult ongi alati püütud seda teha – lihtsalt vahendid on olnud teised.

Loodetavasti kirjeldabki „andmekaevandamine" kõige paremini juba termini enda läbi ning traditsioonilise kaevandamise analoogia abil tegevuse põhiolemust.

Kirjanduses on välja pakutud mitmeid erinevaid laiemalt levima jäänud definitsioone:

- etapp teadmushõives, mille üldine eesmärk on leida andmetest paikapidavaid, uudseid, potentsiaalselt kasulikke ning lõppkokkuvõttes mõistetavaid mustreid [1]
- mahukate andmete analüüs leidmaks uusi seaduspärasusi ja ootamatuid seoseid ning summeerida andmed sellisel uudsel viisil, et need oleksid omanikule samaaegselt arusaadavad kui ka kasulikud [2]
- mustrite avastamise protsess, mis peab olema automaatne või (sagedamini) poolautomaatne. Leitud mustrite sisu peab olema selline, et nad suudaksid juhatada teed mõne teatud eeliseni, tüüpiliselt ärilise konkurentsieeliseni [3]

Autori arvates on eelnenud definitsioonid oma tehnilisusega ning terminoloogia tõttu valdkonnaga alles tutvuda soovijatele keerukad. Lisaks ei tohiks olla üldine definitsioon liigselt mõne kindla tehnika keskne ega seotud konkreetse tegevusalaga. Pigem olgu andmekaevanduse definitsioon üldisem (nagu pakutud – analoogia traditsioonilise kaevandamisega ning väärtusliku informatsiooni leidmine), hilisemalt annab alati minna vastavalt täpsemale eesmärgile spetsiifilisemaks.

Kõige üldisemas mõttes on andmekaevandamisel kaks eesmärki: [1]

- **kirjeldamine** – keskendub andmete selgitamisele, mis võimaldaks analüütikul nende sisse näha ning neid interpreteerida
- **prognoosimine** – võimaldab olemasolevate tunnuste põhjal luua ennustusmudeleid tundmatute või tulevikuväärtuste leidmiseks.

Tuuakse välja ka kolmas – **juhtimata ning järelevalveta avastamine** (nt [2],[4]), ent tegelikult võiks selle pigem liigitada ikkagi **kirjeldamise** alla, sest tegemist ei ole niivõrd eesmärgiga kui pigem protseduuri tüübiga, jaotades tehnikad selle järgi veel eraldi:

- **juhitud tegevused** – küsimus, millele vastust otsitakse, on juba olemas – sihikindlalt liigutakse selle vastuse leidmise suunas
- **juhtimata, järelevalveta ning juhuslik tegevus** – ettevalmistatud andmehulgale rakendatakse erinevaid tehnikaid, lootes leida midagi huvitavat;

Siiski ei tohiks eelnevat mõista selliselt, et arvutid ning algoritmid kaevandavad – seda teevad ikkagi inimesed ning kaevandamisvahendite roll on abistav – võimaldada hakkama saada tohutute andmehulkadega ning neid interpreteerida.

Loomulikult on võimalik ka eelnevaid protsesse automatiseerida, ent sellisel juhul pigem juba leitud mudelite taasrakendamise kujul. Loovust ning sellele vastavat käitumist on arvutitelt veel vara loota.

ANDMEKAEVANDAMISE JA STATISTIKA ERINEVUS

Statistika, masinõpe, andmebaasid ja andmeaidandus, mustrite leidmine, tehisintellekt, andmete visualiseerimine – andmekaevandus oma interdistsiplinaarse olemusega on seotud kõigi nendega ning tihti ka kirjeldatud läbi nimetatute omavahelist lõikumist. Autori arvates toob statistikaga kõrvutamine ja võrdlemine andmekaevandamise olemuse kõige paremini esile.

Seda enam, et ühised jooned ei tundu liigselt juhuslikud ka ülejäänud maailma jaoks – statistikute ringkondades on juba alustatud [5] diskussioone andmekaevandamise temaatikal eesmärgiga:

- proovida kasu lõigata väga lähedase valdkonna populaarsuse kasvust
- otsustada, kas hakata artiklites avaldama survet, et andmekaevandamine kuulutatakse lihtsalt statistika alamdistsipliiniks.

Põhiline erinevus [6] andmekaevandamise ja traditsioonilise statistika vahel, lähtudes toimimise loogikast, on see, et formaalne järeldav statistika on juhitud oletustest – formaliseeritakse hüpotees ning kontrollitakse seda teatud etteantud olulisuse nivool. Andmekaevandus on aga, vastupidi, juhitud avastustest – mustrid ja hüpoteesid genereeritakse andmetest automaatselt. Teisisõnu, andmekaevandust juhivad pigem andmed ning statististilist analüüsi inimesed.

Proovides eelnevat väidet laiendada, võiks öelda, et statistika puhul räägime andmete esmasest analüüsist – me teame juba ette, mida me soovime kontrollida ning kogume vastavalt ka andmeid. Andmekaevandamisel aga vastupidi – üldjuhul kasutatakse selliseid andmeid, mida mingil muul põhjusel on juba varem kogutud ning nüüd viiakse läbi andmete sekundaarne (või järjekorras veelgi hilisem) analüüs. Seetõttu on ka andmekaevandamist vahetevahel defineeritud kui „suurtes andmehulkades läbiviidud sekundaarne andmeanalüüs eesmärgiga leida ootamatuid ning uudseid tulemusi." [7]

Suurimateks erinevusteks statistika ja andmekaevandamise vahel võib pidada [7]:

- **andmetabelite suurust** – statistikud peavad andmete hulka suureks juba mõnesaja tunnusega, igal juhul on tuhanded tunnused analüüsimisele juba tõsine katsumus. Kindel on aga see, et maailma ühe juhtiva telekommunikatsioonifirma AT&T ligi 500 000 000 000-objektise andmetabeli [8] analüüsimisega jääksid traditsioonilised vahendid hätta
- **puuduvad ja vigased andmed** (sh ülekaetus ja kordumised) – probleemi olemus on tegelikult tihedalt seotud eelmise erinevusega. Näiteks 0,1% puuduvaid või vigaseid andmeid avaldaks tavapärastes statistilistes analüüsides väga vähe mõju, suurte andmemahtude puhul tähendaks see aga näiteks miljardist miljonit kirjet, mida ei saa enam analüüsi läbi viies ignoreerida. Lisaks, et andmekaevandamisi viivad suurelt jaolt juba läbi ka mittestatistiku taustaga analüütikud, siis puuduvate ja vigaste andmete maht võib olla veelgi suurem. Nad aktsepteerivad seda, sest nad ei soovigi kätte saada lõplikku kindlust ja kinnitust, vaid vihjet hüpoteesile, mille paikapidamist tuleks veel kontrollida
- **mittestatsionaarsus** – tihti ei ole enam aega koguda andmeid ning hakata neid analüüsima, vaid andmebaas suureneb pidevalt. Äärmuslikes olukordades tuleb isegi olla võimeline analüüsima reaalajas – loomulikult sellised tegevused on rangete piiridega ning seeläbi automatiseeritavad. Mittestatsionaarsuse probleemi toob kõige selgemalt välja informatsiooni vajamise kiirus – eelmise kuu müügi- või mõõtmiste tulemuste analüüsi täna kätte saada võib olla juba liiga hilja. Loomulikult on need kaks probleemi külge vasturääkivad – soov saada andmeid ääretu kiirusega ning tohutud andmehulgad, paraku andmekaevandamisprotsess toimubki pidevalt kompromissina nende piirangute vahel
- **mittearvulised väärtused**. Klassikaline statistika tegeleb puhtalt numbrilise analüüsiga ja kuidas analüütik oma praktilised vajadused nendeks kodeerib, on iga kord spetsiifiline. Andmekaevandamisel tegeldakse aga ka eriomaste andmetega – näiteks pildid, tekst ning geograafilised andmed. Üldine eesmärk – leida huvitavaid mustreid ning avada andmete sisemist struktuuri – kohaldub täiesti edukalt ka neile.

Vaadates nüüd eelmainitud punkte, võime näha, et enamasti on statistika poolelt tegemist teatud piiride seadmisega ning üleastumiste mittetolereerimisega. Selge on see, et andmekaevandamist ei ole võimalik vaadata ilma statistikata, ent eelneva põhjal võiks teadlaste täiesti loogiline käik olla andmekaevandamise lahterdamine statistika alla. Viimast pigem isegi mitte klassikaliseks ja uudseks lähenemiseks, vaid moodustamaks terviklikumat süsteemi, mis võimaldaks ühest küljest genereerida hüpoteese poolautomaatselt tohututest andmehulkadest ja seejärel analüütiku valikul nende korrektsust ning usaldusväärsust kontrollida. Poolautomaatsuse aitaks ära hoida aktsepteeritavate intelligentsete valikute reaalne toimimine.

TEADMUSHÕIVE ANDMEBAASIDEST NING SELLE PROTSESS

Teadmushõive andmebaasidest (*knowledge discovery in databases*) on mittetriviaalne protsess, mille käigus leitakse andmetest paikapidavaid, uudseid, potentsiaalselt kasulikke ning lõppkokkuvõttes mõistetavaid mustreid. [1]

Meenutades nüüd selle definitsiooni kõrval andmekaevandamise definitsiooni, võime näha hämmastavat, isegi segadusse ajavat sarnasust. Mis on siis ikkagi nende erinevus või nad on sünonüümid? **Tegemist ei ole sünonüümidega** – teadmushõive andmebaasidest kujutab endast tervet protseduuride jada, mida on tarvis teha selleks, et andmetest peidetud teadmisi kätte saada. Seega **andmekaevandamine on üks etapp teadmushõives andmebaasidest**.

Teisest küljest ei saa andmekaevandamist läbi viia ilma eelnevate protseduurideta ega pole temast kasu ilma hilisema rakendamiseta, seetõttu lihtsustamise mõttes loetakse neid tihti teadlikult sünonüümideks. Vaatamata sellele, et seni kõige autoriteetsema definitsiooni autorid Fayyad et al. [1] sõnastasid üsnagi selgesti sellise definitsiooni hoopis kogu teadmushõive kohta ning kirjeldasid andmekaevandamist kui ühte etappi (rakenduslikku vahendit) terves pikas protsessis, avaldatakse jätkuvalt ja korduvalt ülalesitatud definitsiooni andmekaevandamise definitsioonina, kusjuures autoriks viidataksegi Fayyad et al.

Alguse tegijad [9] ning edasiarendajad [1] näevad teadmushõivet andmebaasidest järgmisse iteratiivse ning interaktiivse protsessina:

1. **Valdkonnaga tutvumine** ning piisavad eelnevad teadmised, võimaldamaks protsessi eesmärke näha tellija (kliendi) vaatepunktist.

2. **Andmete valik**, millest omakorda selekteeritakse sobivad atribuudid ning vajadusel ka alamhulk kirjeid.

3. **Andmete puhastamine ning eeltöötlus** – võimaluse korral eemaldatakse müra, pannakse paika strateegia vigaste ja puuduvate andmetega ringikäimiseks, silutakse episoodilisi andmeid.

4. **Andmete lihtsustamine ning neile õige kuju andmine**.

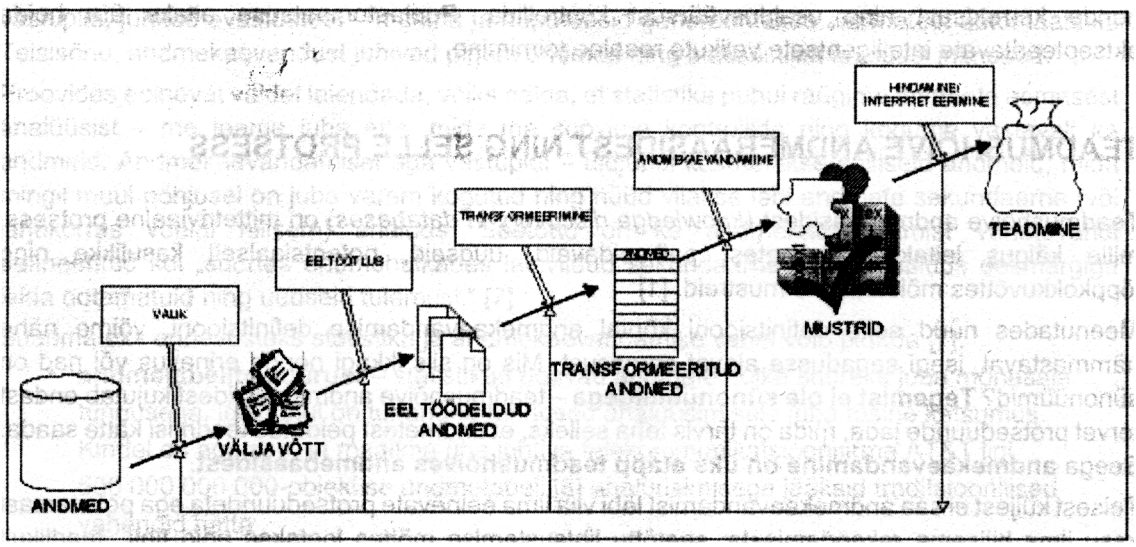
5. **Esimese etapi eesmärgid seotakse kindla andmekaevandamise tehnikaga** (nt summeerimine, klassifitseerimine, regressioonianalüüs, klasterdamine).

6. **Avastav analüüsimine, andmekaevandamise algoritmide ja meetodi valik** mustrite leidmiseks.

7. **Andmekaevandamine** – valitud meetodi ning konkreetse algoritmi rakendamine.

8. **Leitud mustrite ja vihjete interpreteerimine**, võimalik tagasipöördumine kõigi esimese seitsme etapi juurde – selle etapi lõpuks võidakse proovida tulemust ka visualiseerida või välja pakkuda konkreetne mudel.

9. **Leitud teadmistele vastav käitumine** – mudelite integreerimine asutuse süsteemidesse automatiseeritult, lihtne dokumenteerimine ja aruandlus või mudeli rakendamine turunduses või asutuse strateegia kujundamisel.

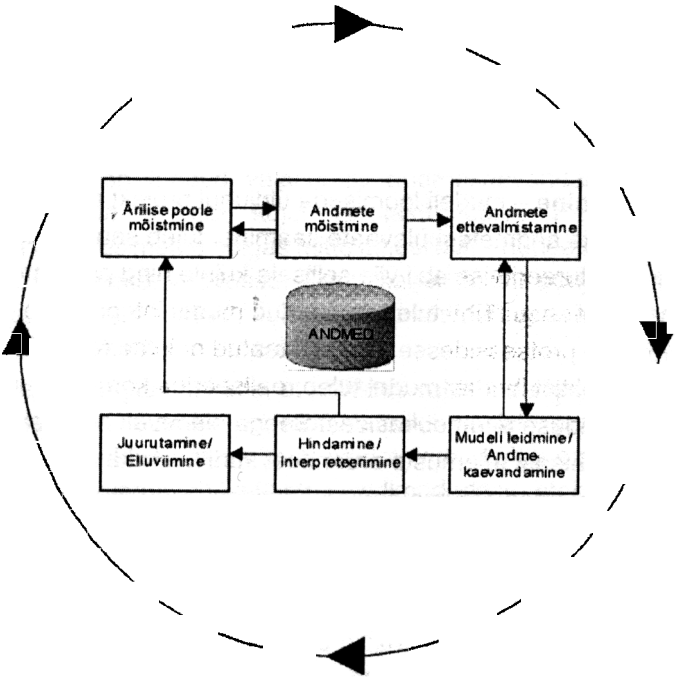


Joonis 1. Teadmushõive etapid [1].

On veelgi loomulikum, et uuel ning üha populaarsemaks saaval tegevusel võib areneda välja mitu konkureerivat protsessimudelit. Õnneks on üldises teadmushõives ja andmekaevanduses suudetud pigem teineteist täiendada ning koostöös ressursirikaste ettevõtetega panna paika ka kokkulepitud protsessikirjeldus.

Tohutuks edasiviivaks jõuks kujunes kolme ettevõtte initsiatiiv panna kokku ühtne protsessimudel [10], mille autoriteks on Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) ning Rüdiger Wirth (DaimlerChrysler). Statistikute juures kõrgelthinnatud SPSS omandas andmekaevandamise oskusteabe teise ettevõtte – ICL ostmisega, mille tulemusena on ka SPSS koosseisus uus tarkvara SPSS Clementine.

Töörühm pani tulemusele nimeks *Tegevusalast sõltumatu standardiseeritud protsess andmekaevandamiseks (CRoss Industry Standard Process for Data Mining* – CRISP-DM). Erinevus, võrreldes eelnevalt väljapakutud protsessidega, oli selgelt äriliste huvide kaitsmine – s.t iga taoline projekt peab algama ärilisest vajadusest ning lõppema tulemuste rakendamisega konkurentsieeliste saavutamise eesmärgil.



Joonis 2. CRISP-DM etapid.

Nende esitatud protsessimudel võiks välja näha selline (detailselt [10]):

- **ärilise poole mõistmine** – esimene etapp keskendub projekti ärilistele eesmärkidele ja nõuetele, püüab formuleerida selle teadmise andmekaevanduse probleemipüstitusena ning pakkuda välja esialgse plaani eesmärkide täitmiseks
- **andmete mõistmine** – etapp algab andmete kogumisega, sisaldab tegevusi nende struktuuri ja sisuga tutvumiseks ning kvaliteediprobleemide tuvastamiseks. Lisaks saadakse selles etapis juba esimesi vihjeid andmete kohta ning moodustatakse väljavõtte, millest võiks hüpoteese genereeruda kõige edukamalt
- **andmete ettevalmistus** – kolmas etapp sisaldab endas kõiki vajaminevaid tegevusi algsetest allikatest lõpliku andmetabeli moodustamiseks. Lõplikuks kutsutakse andmetabelit siis, kui seda on sobiv ette sööta kaevandamisvahendile. Andmete ettevalmistuseks vajaminevaid tegevusi sooritatakse suure tõenäosusega korduvalt ning ilma kindla järjekorrata. Selliste tegevuste hulgas on näiteks tabelite, kirjete ning atribuutide valikud, samuti kõikvõimalikud puhastamised ja üldkuju transformeerimised
- **andmekaevandamine** – valitakse sobivaid tehnikaid ning rakendatakse neid andmetele. Tüüpiliselt on olemas mitu erinevat lähenemist samale probleemile, lisaks eeldavad mõned tehnikad andmetelt teatud kuju, mistõttu andmete ettevalmistamise juurde tagasipöördumine ei ole siin etapis harv juhus

- **hindamine/interpreteerimine** – selleks hetkeks olete juba välja töötanud mudeli (või ka mitmeid mudeleid), mis näivad olevat väärtuslikud andmeanalüüsi seisukohalt. Enne elluviimist on tähtis,et mudel köidaks korralikult ka äriliste eesmärkide ning nõudmiste mõttes taas läbi kontrollimaks, kas mõnda eeldust või nõuet ei ole unustatud. Peale hindamise ning kogu senise protsessi ülevaatamise pannakse täpselt paika järgmised sammud
- **juurutamine/elluviimine** – mudeli loomisega üldjuhul projekt ei lõpe. Isegi kui mudeli eesmärgiks on näiteks andmetest ülevaate saamine, tuleb saadud teadmine korrastamise ja struktureerimise abil viia sellisele kujule ning presenteerida taoliselt, et klientidel oleks sellest kasu. Tihti tuleb ka saadud mudel integreerida olemasolevatesse otsuste vastuvõtmise protsessidesse. Näiteks teatud objekte (ettevõtte kliente, tooteid) mõnest kindlast aspektist hindav mudel tuleb realiseerida korduvate ja regulaarsete arvutustöödena turunduse andmebaasides. Seega olenevalt nõudmistest võib kogu projekti väljund olla lihtsast tulemuste aruandest kuni keeruka korduva andmekaevandamisprotseduuri implementeerimiseni kogu ettevõttes. Tihti on antud etapi läbiviivaks pooleks juba töö tellinud klient, mitte täitev andmeanalüütik. Isegi kui andmeanalüütik ise ei tegele juurutamisega, peab ta siiski kliendile juba ette täpselt määratlema kõik vajalikud sammud mudelite elluviimiseks.

Esmalt tasub kohe märkida, et kui paljud artiklid püüdsid kummutada mingil ajahetkel tekkinud valearusaama, justkui teadmushõive andmebaasidest oleks sama mis andmekaevandamine, siis [10] nullis suures osas nende tehtud töö. Vaatamata sellele, et ka [10] kirjelduses on andmekaevandamine vaid üks etapp, sisaldab kogu protsessi nimetus ikkagi teadmushõive asemel andmekaevandamist. Loomulikult võib süvenemisel tõlgendada seda ka õigesti, et ülejäänud etapid on lihtsalt kohustuslikud eelnevad ja järgnevad tegevused, kuid keskne tegevus on ikkagi andmekaevandamine. Arvestades aga, et isegi artiklis [1] suudeti pealiskaudsel tõlgendamisel terminoloogias palju segadust tekitada, siis artiklist [10] on seda loota veel rohkemgi – seda enam, et sihtgrupiks ning esimeseks filtriks ei ole enam teadusasutustes töötavad inimesed, vaid erineva taustaga äriettevõtete töötajad.

Vaatamata terminoloogiale andis [10] siiski tohutu panuse andmekaevandamise rakendamise, eelkõige formaalsema raamistiku loomisega ning tegevuse tugevama sidumisega eesmärkidega ja hilisema rakendamisega. Lisandus ka etapp, millele [1] veel tähelepanu pööranud polnud – *arendamine, jälgimine ning hooldamine*. Eelkõige kirjeldas [10] küll tehnilist hooldamist eesmärgiga süsteemide muutumisel vigadele kiiresti jälile saada, kuid tegelikult võis (eriti rõhutuse tõttu, et kogu protsess on iteratiivne) välja lugeda ka ühe lisanüansi, millele märksa rohkem pani rõhku [6]: ka äriline keskkond (loomulikult teisedki valdkonnad, kus andmekaevandamist kasutatakse) muutub pidevalt, konkurendid võivad välja tulla uute toodetega, elukvaliteet muutub – kõik see võib muuta klientide käitumist ning seetõttu ei pruugi varasema käitumise põhjal kokkupandud mudel igavesti töötada – teda tuleb pidevalt korrigeerida. Viimane teeb muidugi investeeringute tasuvuse osas ettevõtetele kogu taolise projekti üsnagi riskantseks, sest pideva rahastuseta võib kogu tegevus mõttetuks osutuda. Probleem nii tõsine siiski pole, kuna hinnanguliselt 80% kogu töömahust on valdavalt eelnev andmete korrastamine, õigele kujule viimine ning tehnikate valimine. Seda enam, et tänapäeval on andmebaasidesse n-ö regulaarsed transformeerimisteenused juba sisse integreeritud, võimaldades pärast andmevoogude esimest transformeerimise kirjeldamist hilisemad muutused juba automaatseks viia.

Olles üldjoontes korrastanud protsessi korrektse läbiviimise, on üldine trend taas välja töötada uusi ning täiendada olemasolevaid tehnikaid ja algoritme.

PRAKTILISED RAKENDUSED

Järgnevates alapunktides on toodud valdkonniti andmekaevandamise võimalikud praktilised rakendused, nimekiri pole sellisel kujul kindlasti lõplik. Pangandus, kindlustus ning telekommunikatsioon ei ole esimeste hulgas mitte juhuslikult – nendes valdkondades on seni kõige rohkem investeeritud andmekaevandusse. Seda kahel lihtsal põhjusel: neil on kapitali, mida sellesse investeerida, ning nad teavad, kuidas see investeering ennast üsnagi kiiresti ära tasub ning väga võimsalt neile kasumit genereerib.

Loomulikult ei taha keegi jäägitult uskuda, et kõikides valdkondades peab kõike mõõtma rahas – heaks näiteks on meditsiin. Paraku on motivaatoriks taoliste projektide puhul siiski kulude kokkuhoid (meditsiini alapunktist saab lugeda näidet, kuhu kulub ameeriklastel miljardeid), mitte üllas soov paremini ravida. Sarnaselt ka teistes valdkondades – efektiivsus tähendab raha.

Seega on järgmised alapunktid pühendatud eelkõige andmekaevandamisprojekti võimaldajate motivatsioonile – ehk siis kuidas teenida andmekaevandamise abil omanikele rohkem raha.

TURUNDUS JA MÜÜK

Esimese rakendusena toome välja turunduse ja müügi üldiselt, sest sellega katame ettevõtete ja organisatsioonide tüüpilise ühisosa: kõik soovivad kellelegi midagi müüa. Vastus, miks kasutada turunduses ja müügis andmekaevandamist, on üsna lihtne – selleks et mõista paremini klientide huve ja käitumist.

Eestis ei ole rakendamine võrreldav suuremate riikidega, sest piisavalt suuri ettevõtteid on vähe ning maailma mastaabis suuri polegi. Kui aga tinglikult kuhugi alampiir tõmmata, siis autori hinnangul võiks andmekaevandamisest kasu saada järgmise suurusega ettevõtted:

- käive >50 miljoni krooni aastas;
- kliente kokku >1000 ja/või müügiarvete ridu aastas keskmiselt >50 000.

Eelnev ei ole kindlasti lõplikult määrav, ent filtreerib enam-vähem õiglaselt välja ettevõtted, kus ei ole mõtet andmekaevandamise peale mõelda (investeeringuteks raha raisata). Kindlasti võiks see ka väiksematele firmadele anda väärtuslikku informatsiooni, ent investeering ei tasuks ennast majanduslikult ära. Seda enam, et eelnevalt toodud piir on üsnagi leebe iseloomuga ning kaasab ka palju kohalikke keskmise suurusega ettevõtteid.

Järgnev tabel peaks andma ülevaate, millistele küsimustele üldse andmekaevandamise abil vastust leida võib:

Tabel 1

Tehnoloogiate võrdlus

Tehnoloogia	Äriline küsimus	Olemus
Andmete salvestamine (alates 60ndad)	Kui palju on raha sisse toonud klient X?	Informatsioon
Andmeaidad ja mitme-dimensioonilised andmebaasid (alates 90ndatest)	Kuidas on ettevõttel läinud toodete ning maakondade kaupa, võrreldes eelmise aastaga? Ida-Virumaal on toimunud tõus ... mis on mõjutanud sealset müüki?	Analüüs
Andmekaevandamine (massidesse jõudmas alles tänasel päeval)	Kui tõstaksime toote X hinda 5%, kui palju kliente me kaotaksime? Miks?	Kirjeldused Vihjed Ennustamine

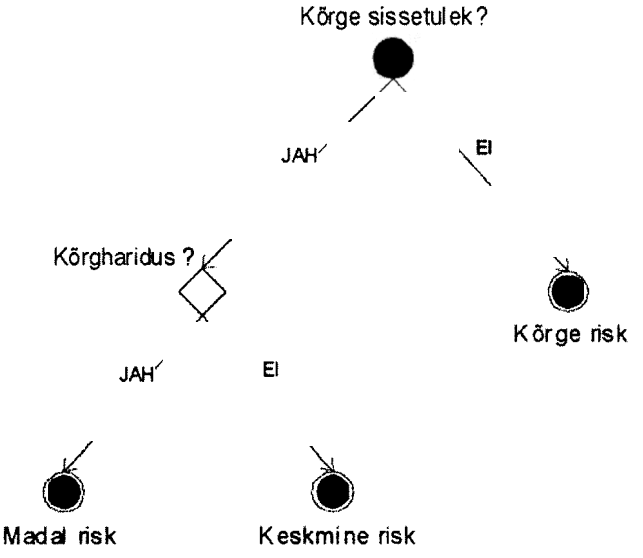
Põhilised turundusalased rakendused on järgmised [11]:

- **kliendiprofiilide leidmine ning segmenteerimine** (*profiling and segmentation*) – otseturunduses ehk individualiseeritud turunduses ei ole kliendiprofiilide leidmine ning segmenteerimine uus – mida täpsemalt ja võimalikult väikese kuluga potentsiaalsed ostjad ära tabada, seda võimsam kasum. Andmekaevandamine pakub võimaluse töödelda suuremat kliendibaasi rohkema ja kaudsema informatsiooniga, kus tunnusteks on peale traditsioonilise demograafilise bloki ka ostukäitumised ja -harjumused. Lisaks pakub tugevamat ennustusaparatuuri
- **ristmüük** (*cross-selling and up-selling*) – juba rahulolev klient ostab keskmisest suurema tõenäosusega samalt pakkujalt ka teise toote (näiteks lisaks sõiduki kindlustusele ka elukindlustuse) või annab olemasoleva toote asemele müüa suurema/parema (näiteks elamu kindlustamine suurema summa peale). Tundes toodetevahelisi seoseid, on võimalik ühe toote kampaania abil võimendada teise müüki
- **kliendikaotuse vältimine** (*customer retention, customer attrition, churn management*) – kõrge konkurentsiga tegevusalades tähendab uue kliendi saamine teisele firmale kliendi kaotust – sellest võimalikult varakult teadasaamine võib aidata seda vältida. Kliendi kadumine on igal juhul ettevõttele suurem kahju kui uue kliendi saamisest tulenev kasu (millest tuleb maha arvestada ka saamisele kulunud ressurss). Teisest küljest kasutatakse kliendikaotuse mudeleid ka hinnakujundamisel – näiteks hinnatõusu puhul arvestades teadlikult loobumistega. Viimast loomulikult eeldusel, et hinnatõusust saadav raha on suurem loobujatelt saadavast
- **eluaegse väärtuse hindamine** (LTV – *lifetime value*) – kogu kliendiksolemise aja jooksul sissetuleva raha määratlemine, lojaalne klient on lisaks ka vähem hinnatundlik. Kindlaksmääratud tõenäosusega potentsiaalse teenitava summa teadmisega on võimalik toetada tulevasi investeeringuid.

PANGANDUS

Lisaks eelmises punktis väljatoodud toodete ristmüügile ning klientide paremale tundmisele, profileerimisele ja säilitamisele, on panganduses ka mitmeid spetsiifilisi rakendusi, näiteks:

- investeeringute optimaalne juhtimine ning riskide hindamine
- krediidiriskide hindamine
- krediidi kulukuse määra võimalikult täpne hindamine.



Joonis 3. Laenuriski lihtsustatud otsustuspuu.

KINDLUSTUS

Kindlustussektoris hakati andmekaevandamist rakendama esimeste seas. Täpsemalt, olemasolevaid statistilisi ning adaptiivseid mudeleid prooviti korrigeerida ning rakendada üha suuremate andmebaaside puhul – lihtsalt algselt ei kutsutud seda andmekaevandamiseks. Mahtude suurenemisel tuli paratamatult seda tööd tegema hakata andmekaevandamisvahendite abil, sest spetsiaalne (sisuliselt olemasolevate vahendite) tarkvaraarendus ei oleks majanduslikult otstarbekas. Põhilisteks andmekaevandamise kasutusvõimalusteks kindlustusvaldkonnas on:

- riski ennustamine ja hindamine
- hinnakujundus
- kahjunõuete töötlemine ja analüüs
- kindlustuspettuste ning keerukamate petuskeemide avastamine

Kindlustuspettustest saab rääkida ka riiklikes organisatsioonides, näiteks haigekassa hüvitiste väljapetmine.

TELEKOMMUNIKATSIOON

Lisaks turunduslikule aspektile (toodete paremale kujundamisele, positsioneerimisele ning ristmüügile) ning pettuste ja krediidiriskide hindamisele on telekommunikatsiooni-spetsiifiliselt põhjalikumalt uuritud andmekaevandamise rakendamist ka ulatusliku võrgu monitoorimisel (nt [12], [13], [14], [15]).

Koostöös telekommunikatsioonifirmadega on välja töötatud tarkvara nimega TASA (*Telecommunication Alarm Sequence Analyzer*), mille prototüübid on juba reaalses kasutuses. Tänapäevased keerulised võrguseadmed genereerivad terve süsteemi peale päevas tohutul hulgal alarme – enamik neist pole tähtsad ning osade puhul hindavad seadmed lokaalselt viga ebatäpselt. Kasutusele võeti sagedaste episoodide analüüs (*sequential patterns*), mis olemuselt on assotsiatsioonireeglite leidmine koos lisandunud ajalise mõõtmega.

Analüüsi tulemusel suudetakse paremini:

- tuvastada korduvaid ja ülemääraseid alarme
- senise kogemuse põhjal ennustada sõltumatute lokaalsete alarmide põhjal ülesüsteemilist viga – kindel järjestus lokaalseid alarme üldjuhul viitab mõnele suuremale üldisele veale;
- tõrked võivad olla omavahel ka seotud, mistõttu teatud tõrgete järel osatakse juba ennustada, kus järgmisi alarme oodata on, ning ennustada ka üldiseid süsteemi vigu.

MAKSUAMET

Enim levinud rakenduseks on rahapesu skeemide tuvastamine. Põhjalikumalt saab protsessi ning väljatöötatud mudelitega tutvuda [16]. Hakates inimjõul tõestama, et mõni ettevõte tegeleb pettusega, tasub tohutuid andmemasse klasterdades näiteks igaks juhuks tähelepanelikuma pilguga üle käia kõik arvuti poolt samasse lahtrisse asetatud ettevõtted. Väljapakutud ettevõtted ei pruugi tegelda veel kelmustega, ent taoliselt võib tuvastada ka mõne üldisema petmisskeemi mudeli, mida analüütikud seni hoomanud pole.

Ka Eestis on maksuametis ümbrikupalkade maksmise tuvastamiseks rakendatud andmebaaside abi. Viimasel juhul tehti siiski väljavõte mustrist, mida genereeris oma ala spetsialist, mitte automaatselt mõni algoritm: kõik suurte käivetega ettevõtted, kelle tööjõukulud on väiksemad selle piirkonna keskmisest palgatasemest. Selle põhjal võiks arvata, et korralik algus on tehtud ning varem või hiljem rakendatakse ka vähemate kahtlustunnustega maksupetturite leidmiseks andmekaevandamist.

Autori hinnangul on pankade ning kindlustusasutuste kõrval just riiklikud organisatsioonid need, kes võiksid igapäevasest andmekaevandamise rakendamisest kõige rohkem võita.

KURITEGEVUSEGA VÕITLEMINE

Eelnevates punktides sai mitu korda käsitletud kõikvõimalikke kelmusi ning üldise süsteemi nõrkade kohtade ärakasutamist. Kelmuste avastamine (*fraud detection*) on seni olnud kaevandamise praktilise rakendamise lipulaev, sest kurjategijad proovivadki ekspluateerida inimeste võimetust suurte hulkade puhul avastada seda, et mängitakse väikestele kõrvalekalletele. Kuna viimaste avastamine ongi andmekaevanduse üks põhilisi tugevaid külgi,

siis on tänaseks enamik suure klientide arvuga ettevõtteid suutnud edukalt arendada endale taolisi süsteeme.

Tasapisi on võimalik rääkida andmekaevandamise kasutamisest ka riiklikul tasandil kuritegevusega võitlemiseks, seda seni eelkõige suhtlemisvõrgustike (mustrite!) avastamiseks ning analüüsimiseks. Suuremad ettevõtted on proovinud oma töötajate suhtlemist e-posti ning telefoni teel kaardistada – sellisel juhul oleksid tulemuseks omavahel suhtlevad osapooled, mida on isegi kaalutud graafina võimalik üles joonistada. Taolise graafi analüüsimine tooks välja omavahel kõige tihedamalt suhtlevad osapooled, mis võimaldaks analüüsida, kas meeskonnad on kõige optimaalsemalt planeeritud ning kas töötajad ei suhtle põhiajast ettevõttesiseselt selliste inimestega, kelle peale ei tohiks nad tegelikult aega kulutada.

Riiklikul tasandil on suhtlusmustrite avastamist proovitud rakendada [17] kuritegelike grupeeringute piiritlemiseks, struktuuri ning võimuhierarhia tuvastamiseks.

Lisaks on juba mitmeid näiteid, kuidas püütakse lahendada traditsioonilise tööga lahendamata jäänud kuritegusid. SPSS suutis Suurbritannias [18] juurutada politseitöösse andmete analüüsi, mis võimaldaks tabatud kurjategijate käitumismustrite järgi siduda neid vanade lahendamata kuritegudega.

TOOTMINE

Üldiste näidetena võiks välja tuua:

- kvaliteedikontrolli mudelite parandamine
- protsessimudelite korrigeerimine
- garantiide juhtumikäsitlus (tugev analoogia kindlustusega) – garantiipettuste avastamine, teatud mudelitel tüüpiliste ehitusvigade tuvastamine, varuosade vajaduse prognoosimine
- automaatse diagnostika ekspertsüsteemid.

TEKSTIANALÜÜS, DOKUMENDIHALDUS

Teksti kaevandamises on tõstatatud küsimus: kas on võimalik näha ka teksti sisse samamoodi, nagu me püüame avada traditsiooniliste andmetabelite sisu. Strateegiaid ning lähenemisi on mitmeid [19]:

- **statistiline** – töödelda dokumente nagu suurt hulka sõltumatuid tunnuseid (analüüsida võib sõnade või n-grammide kaudu, viimaseid on tarvis selleks, et paremini suuta analüüsida mürarikkaid tekste ning dokumente, kus on korraga esindatud mitu keelt)
- **lingvistiline** – analüüsida dokumendi süntaksit ning semantikat
- **graafiline** – käsitleda dokumente visualiseeritavate objektidena; niisugune lähenemine nõuab siiski analüütikutelt tugevat võimet vastavaid mustreid avastada.

Tüüpilised väljakutsed teksti kaevandamisele võiksid olla:

- kas need dokumendid on kirjutatud sama inimese poolt?
- kas need dokumendid puudutavad samu küsimusi ning temaatikat?

John Madison, John Jay ning Alexander Hamilton kirjutasid aastal 1787 konstitutsiooni kiiremaks läbisurumiseks terve seeria esseesid, mis avaldati nime all „The Federalist Papers" [21]. 11 autorit 88-st on teada vaid oletuslikult.

Kjelli ja Friederi [22] hüpoteesiks oli, et n-grammide abil on võimalik leida tekstides mustreid ning seeläbi tuvastada autor. Valiti välja võimalikult unikaalsed n-grammid ning püüti neid segmenteerida. Leiti, et John Madison kirjutas tegelikult kõik üksteist tundmatu autori esseed.

Sarnaseks mahukamate tekstide analüüsinäiteks või tuua veel Dr. Charles Nicholas' [19] piibliteksti sügavamad uurimused (seda enam, et piibel on kirjutatud heebrea, kreeka ja aramea keeles ning kaks esimest on *on-line*-versioonidena Internetis levinud), kui ta proovis vastata teoloogidele kaua arutlusainet pakkunud küsimustele:

- kas prohvet Jesaja puhul oli tegemist ainult ühe inimesega?
- kes on Deuteronoomiumi ehk Viienda Moosese raamatu autor?
- kes on esimese ja teise Ajaraamatu autor?
- kas apostel Paulus kirjutas kõik Epistlid?

Kõige rohkem saavad tekstianalüüsiga seotud uuringutest mõjutusi tulevased dokumendihaldussüsteemid, mis peaksid võimaldama tekste automaatselt lahterdada autori ning teema kaupa.

MEDITSIIN

Andmekaevandamise kasutusvõimaluste uurimisel meditsiinis on suurimaid investeeringuid teinud ilmselgelt ravimifirmad, sest uute ravimite projekteerimine (*drug design*) on juba oma olemuselt ääretult üldiselt võetuna kombinatsioon ainetest, millele organism üht või teist moodi reageerib (uurimiseesmärgiks seega eri kombinatsioonide lahterdamine). Erinevad vahendid tohutute katsetulemuste ning kõrvalnähtude andmebaasist teadmiste kaevandamiseks pakuvad juba täna uute ravimite väljatöötamises kulude kokkuhoidu.

Heade näidetena on järgnevalt ära toodud veel kaks projekti, mille puhul on erasektori asemel tegemist riiklikul tasandil algatatud uuringutega.

Singapuri elanikest umbes iga kümnes põeb suhkruhaigust, millel on mitmeid kõrvalnähte – suurem risk silmahaiguste, neeruhaiguste ning muude tüsistustega. Varajane haiguse avastamine ning korralik ravi võimaldavad neid vältida. Haiguste vastu võitlemiseks alustas Singapur aastal 1992 haigete regulaarset jälgimist – patsientide informatsioon, kliinilised sümptomid, silmahaiguste diagnoosid ning raviinfo salvestati andmebaasi. Tänaseks on suutnud nad antud süsteemi üsnagi hästi tööle rakendada, põhjalikuma ülevaate saamiseks ning tehnoloogiaga tutvumiseks võib lugeda [23]. Uuringus kasutati seaduspärasuste otsimiseks andmekaevandamise assotsiatsioonireeglite leidmise tehnikat.

Teine, ilmselt veelgi võimsama toetusega praktiline rakendus on käsil neeru dialüüsi patsientidega Ameerikas [24]. Umbes 370 000 ameeriklast on neeruvaeguste puhul sellises staadiumis, kus dialüüs või neeru transplantatsioon on eluliselt vajalik. Aastane kulu neeruhaigete ravile on 12 miljardit dollarit. Hemodialüüsis patsientide jälgimisel jääb maha tohutu suur hulk meditsiinilist infot, mistõttu arstidel on mustrite nägemine üle pikema aja üsnagi problemaatiline. Tehnika ei paku uusi lahendusi, vaid laiendab analüüsimisel ajalist akent, mille sisse mahtuvat spetsialist suurte andmemahtude tõttu enam haarata ei suudaks.

Seniste juhtumianalüüside põhjal ning patsientide ajalooliste raviandmete analüüsimine võimaldab vastavalt andmekaevanduse olemusele kirjeldada hetkeolukorda täpsemalt ning ehitada prognoosimiseks paremaid mudeleid.

JAEKAUBANDUS

Andmekaevandamise populariseerimine ning eriti assotsiatsioonireeglite leidmise probleem (*association rules, affinity analysis*) võlgneb suuresti tänu ka jaekaubandusega tegelevate ettevõtete investeeringutele. Algselt tuntigi assotsiatsioonireeglite temaatikat rohkem ostukorvi analüüsina (*market basket analysis*).

Analüüsi sisu on tegelikult lihtsalt mõistetav: isegi korvi isikuga sidumata on võimalik kõikide kassas registreeritud ostukorvide sisu analüüsides leida omavahel tugevalt seotud kaubad, täpsemalt, milliseid kaupu ostetakse koos. Analüüsi eesmärk on leida huvitavaid seoseid, mis ei oleks liiga triviaalsed (näiteks sai ja leib), kuid mis näitaksid piisava kindlusega, et kahte (või enamat kaupa) ostetakse tihti koos.

Taolise uuringu tulemust saab jaekaubanduses tõhusalt rakendada mitmel moel:

- paigutada koosostetavad kaubad teineteise lähedale, suurendades nõnda nende mõlema müüki (soovitud toote mitteleidmisel võib klient ka loobuda)
- asetada koosostetavad kaubad teineteisest võimalikult kaugele, suurendades nii võimalust, et teel teise kauba juurde ostetakse emotsioonide ajendil ka muid kaupu
- toodete paigutus riiulitel, riiulite paigutus, kliendi liikumise optimeerimine
- sooduskampaaniate ning kupongide abil võimendada ühte kaupa reklaamides teise müüki. Sooduskampaania varjus klient tegelikult ei anna endale aru, et teine toode on samavõrra (või isegi rohkem) kallim
- odavamat kaupa on alati lihtsam müüa – seega võib olla kasulik odava kauba (millega koos tegelikult alati ostetakse ka seotud kallim kaup) reklaami rohkem investeerida;

Andmete maht, mida jaekaubanduse andmebaasides töödeldakse, esitab juba väga tõsise väljakutse riist- ja tarkvara tootjatele, sest mitmed eksperdid on andnud hinnangu, et Wal-Marti andmebaasid (eriti tulevase RFID tehnoloogia mõjul) võivad juba lähiaastail ületada 1 petabaidi (= 1 000 terabaiti = 1 000 000 gigabaiti). Lisaks sügavale analüüsimisele peab nende süsteem Retail Link hakkama saama [20] ka enam kui 7500 tarnijale täpse jooksva müügiinfo serveerimisega, võimaldades neil oma tootmist ning ladusid paremini planeerida.

TULEVIK

Esmalt tuleks kindlasti välja tuua üha valjenev poliitiline trend, mille arengut kärpivaid mõjusid võib tunda ilmselt juba lähiaastatel – uuritavate range privaatsuse tagamine. Olgugi, et andmekaitse on alati olnud tundliku informatsiooniga tegelejatele kohustuslik nõue, on andmete kogumise ning laiatarbe analüüsitarkvara levikuga tekkimas olukord, kus informatsiooni lekkimise risk on kõrge. Teiseks küljeks on privaatsuse eetiline aspekt – kas on õige tunda inimeste käitumist ning seda ekspluateerida?

Üldiseks trendiks süsteemides ja rakendustes on suund muutuda automaatsemaks, kuhu professionaali oskusteave oleks juba üha rohkem integreeritud. Teisest küljest kaotaks see põhilise konkurentsieelise, mida spetsialisti loovus võimaldaks. Seetõttu usub autor, et antud valdkond peaks jääma alati teatud mõttes poolautomaatseks ning täisautomaatsena oleks süsteemil mõtet ainult ettevõttesisesena (mille kitsad piirid seadistab eelnevalt siiski spetsialist). Laiatarbetarkvara, mis sisseehitatud oskusteabe abil lubab konkurentsieelist, on nonsenss. Konkurentsieelise tekitavad siiski inimesed, olgugi et andmekaevandamise abil suurema võimendusega.

KIRJANDUS

[1] Fayyad,U., Piatetsky-Shapiro, G., Smyth P. "The KDD process for extracting useful knowledge from volumes of data" //Communications of the ACM, 39 (11): November 1996, pp. 27 –34.

[2] Hand, D., Mannila, H., Smyth, P. "Principles of Data Mining". Cambridge: MIT Press, August 2001, 425 p.

[3] Witten, I. H., Frank, E. "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations." Morgan Kaufmann Publishers, 2000, San Francisco, CA., 416 p.

[4] Berry, M. J. A., Linoff, G. S. "Mastering Data Mining." New York: Wiley, 2000, 512 p.

[5] Friedman, J. H. Data mining and statistics: what's the connection? // Proc. of the 29th Symposium on the Interface: Computing Science and Statistics, May 1997, Houston, Texas, pp.5–10.

[6] Zhang, C., Zhang, S. "Association Rule Mining: Models and Algorithms." Berlin, Springer, 2002, 238 ps.

[7] Hand, D. J. "Data mining: Statistics and More?" // The American Statistican, May 1998 Vol. 52, No. 2, pp.112–118.

[8] Winter Corporation: "Top Ten Data Warehouses" [WWW] http://www.wintercorp.com/VLDB/2003_TopTen_Survey/TopTenWinners.asp (05.05.2005).

[9] Brachman, R. J., Anand, T. "The Process of Knowledge Discovery in Databases: A First Sketch" // KDD Workshop 1994, Seattle, Washington, USA, pp.1–12.

[10] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. "CRISP-DM 1.0." 2000. [WWW] http://www.crisp-dm.org/ (05.05.2005).

[11] Rud, O. P. "Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management." New York: Wiley, 367 p.

[12] Hätönen, K., Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. Knowledge Discovery from Telecommunication Network Alarm Databases // Proceedings of the 12th International Conference on Data Engineering (ICDE'96), New Orleans, Louisiana, IEEE Computer Society Press, February 1996, pp. 115–122.

[13] Hätönen, K., Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. TASA: "Telecommunications Alarm Sequence Analyzer, or "How to enjoy faults in your network" // In IEEE/IFIP 1996 Network Operations and Management Symposium (NOMS'96), Kyoto, Japan, IEEE Computer Society Press, April 1996, pp. 520–529.

[14] Klemettinen, M., Mannila, H., Toivonen, H. Exploration of interesting findings in TASA // Information and Software Technology 41, 9 (1999), pp. 557–567.

[15] Klemettinen, M., Mannila, H., Toivonen, H. Rule discovery in telecommunication alarm data // Journal of Network and Systems Management 7, 4 (December 1999), pp. 395–423.

[16] Zhang, Z., Salerno, J.J., Yu, P.S. "Applying data mining in investigating money laundering crimes" // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., pp. 747–752.

[17] 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" // Proceedings of the National Conference for Digital Government Research (dg.o 2003), May 18-21, 2003, Boston, Massachusetts, pp. 45–48.

[18] Crime detection – A case study [WWW] http://www.spss.com/uk/westmidlands.pdf (05.05.2005).

[19] Charles K. Nicholas' homepage [WWW] http://www.cs.umbc.edu/~nicholas/ (05.05.2005).

[20] About WalMart.com [WWW] http://www.walmart.com/cservice/aw_index.gsp (05.05.2005).

[21] An Outline of American History / H. Cincotta, D. Brown, S. Burant, M. Green, J. Holden, R. Marshall. United States Information Agency, 1994, 407 p.

[22] Kjell, B., Frieder, O. "Visualization of literary style" // IEEE International Conference on Systems, Man and Cybernetics, IEEE, 18–21, October 1992, pp.656–661.

[23] Hsu, W., Lee, M.L., Liu, B., Ling, T.W. "Exploration mining in diabetic patients databases: findings and conclusions" // Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), New York: ACM Press, 2000, pp.430–436.

[24] Shah, S., Kusiak, A., Dixon, B. "Data Mining in Predicting Survival of Kidney Dialysis Patients". // Proceedings of Photonics West–Bios 2003, Bass, L.S. et al. (Eds), Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems XIII, Vol. 4949, SPIE, Belingham, WA, January 2003, pp. 1–8.

Innar Liiv
TTÜ informaatikainstituut