VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

Aivaras Voveris

VARTOTOJO PROFILIO INFORMACIJOS ANALIZĖ POTENCIALIAI PAVOJINGOS VEIKLOS APTIKIMUI

USER PROFILE ANALYSIS FOR IDENTIFYING POTENTIALLY HARMFUL ACTIVITY

Baigiamasis magistro darbas

Informacijos ir informacinių technologijų saugos studijų programa, valstybinis kodas 621E14007

Informatikos inžinerijos studijų kryptis

Vilnius, 2018

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

Aivaras Voveris

VARTOTOJO PROFILIO INFORMACIJOS ANALIZĖ POTENCIALIAI PAVOJINGOS VEIKLOS APTIKIMUI

USER PROFILE ANALYSIS FOR IDENTIFYING POTENTIALLY HARMFUL ACTIVITY

Baigiamasis magistro darbas

Informacijos ir informacinių technologijų saugos studijų programa, valstybinis kodas 621E14007
Informatikos inžinerijos studijų kryptis

Vadovas doc. dr. Nikolaj Goranin _____ _____
(Pedag. vardas, vardas, pavardė) (Parašas) (Data)

Lietuvių kalbos konsultantas lekt. Regina Žukienė_____ _____ _____
(Pedag. vardas, vardas, pavardė) (Parašas) (Data)

Vilnius, 2018

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

BAIGIAMOJO MAGISTRINIO DARBO (TYRIMO)
UŽDUOTIS

…….....Nr. .....
Vilnius

Studentui (ei) Aivarui Voveriui
(Vardas, pavardė)

Baigiamojo darbo (tyrimo) tema: Vartotojo profilio informacijos analizė potencialiai kenksmingos veiklos aptikimui

patvirtinta 2016 m. spalio 27 d. dekano potvarkiu Nr. 385 fm

Baigiamojo darbo (tyrimo) užbaigimo terminas 201___ m. _____ d.

BAIGIAMOJO DARBO (TYRIMO) UŽDUOTIS:

Atlikti vartotojo profilio informacijos struktūros ir gavimo šaltinių analizę, išanalizuoti egzistuojančius vartotojo profiliavimo būdus bei jų efektyvumą, paruošti metodą, leidžiantį aptikti potencialiai pavojingus vartotojus (teroristai; ir pan.) bei atlikti metodo efektyvumo bandymus

Baigiamojo magistro darbo (tyrimo) konsultantai: _____
(Pedag. vardas, vardas, pavardė)

Vadovas _____ _____
(Parašas) (Pedag. vardas, vardas, pavardė)

Užduotį gavau

_____
(Parašas)

_____
(Vardas, pavardė)

_____
(Data)

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Informacinių sistemų katedra

ISBN ISSN
Egz. sk. ……..
Data ….-….-….

Inžinerinės informatikos studijų programos baigiamasis magistro darbas
Pavadinimas **Vartotojo profilio informacijos analizė potencialiai pavojingai veiklai aptikti**
Autorius **Aivaras Voveris** Vadovas doc. dr. **Nikolaj Goranin**

Anotacija

Atsižvelgiant į kenksmingos veiklos plitimą (internetinės atakos, teroristiniai sprogdinimai) didėja poreikis identifikuoti potencialiai kenksmingus vartotojus ir užkirsti kelią jų veiklai iki nutinkant incidentui. Darbe nagrinėjami metodai rinkti vartotojų informaciją, identifikuoti vartotojus tarp skirtingų įrenginių, sisteminti gautą informaciją ir paruošti metodą potencialiai kenksmingiems vartotojams identifikuoti.

Prasminiai žodžiai:

kenksmingi vartotojai, internetinės atakos, vartotojo demografinė informacija, mašininis

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Information System

ISBN ISSN
Copies No. ……
Date ….-….-….

Master Degree Studies Engineering Informatics study programme master thesis.

Title: **User profile analysis for identifying potentially harmful activity**

Author **Aivaras Voveris** Academic supervisor **Nikolaj Goranin**

Annotation

As harmful activity is spreading (internet attacks, suicide bombings) it is vital to identify potentially harmful users and prevent any harmful activity before it occurs. In this thesis we analyze methods to collect information about user, identify them cross-device and use this information for preparing a method to identify potentially harmful users.

Keywords:

harmful users, internet attacks, user demographic information, machine learning

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Santrumpos

CSS – pakopiniai stilių šablonai (angl. *Cascading Style Sheet*)

DOM – dokumentų objekto modelis (angl. *Document Object Model*)

HTML – hiperteksto žymėjimo kalba (angl. *Hyper Text Markup Language*)

ISP – internetinių paslaugų tiekėjas (angl. *Internet Service Provider*)

JS – JavaScript programavimo kalba (angl. *JavaScript*)

KPK – kenksmingas programinis kodas

WWW – Interneto tinklas (angl. *World Wide Web*)

RTB – realaus laiko aukcionai (angl. *Real time bidding*)

DSP – paklausos pusės platforma (angl. *Demand Side Platform*)

SSP – pasiūlos pusės platforma (angl. *Suply Side Platform*)

ADX – reklamos tinklas (angl. *Ad Exchange*)

ADN – reklamos rodymo tinklas (angl. *Ad Delivery Network*)

DX – duomenų mainai (angl. *Data Exchange*)

DMP – duomenų tvarkymo platforma (angl. *Data Management Platform*)

HTTP – žiniatinklio protokolas (angl. *Hypertext Transfer Protocol*)

IAB – internetinės reklamos biuras (angl. *Internet Advertisement Beaurau*)

URL – universalusis adresas (angl. *Uniform Resource Locator*)

TOR – svogūninis maršrutizatorius (angl. *The Onion Router*)

MGRS – karinė koordinačių žymėjimo sistema (angl. *Military Grid Reference System*)

1. Įvadas

Pasaulinis informacijos tinklas (angl. *World Wide Web*) tapo neatsiejama žmogaus dalimi. Esant galybei įvairios informacijos, kiekvienas vartotojas gali rasti sau tinkamus informacijos šaltinius.

Vartotojai dažnai ieško informacijos juos dominančiomis temomis, lankosi nišiniuose puslapiuose ir pagal tai jie gali būti skirstomi į interesų grupes. Interesų nustatymas yra neatsiejama internetinės reklamos dalis. Vis dažniau media kampanijoms keliami reikalavimai reklamas rodyti specifinei žmonių grupei, taigi kyla poreikis atpažinti vartotojo grupes / interesus. Interesų bei poreikių atpažinimas taip pat naudojamas suasmenintoje paieškoje, straipsnių ir kitų šaltinių pasiūlymuose vartotojui, todėl iš šio poreikio išsivystė modeliai, kaip atpažinti vartotojo interesus, panašius interesus arba tai, kas vartotojui gali būti įdomu – *youtube,* vaizdo ir panašūs puslapiai. Šie modeliai pradėti formuoti dar 1980 m. su pirmosiomis naršyklėmis. Daugelis autorių nagrinėjo, kaip pasiūlyti vartotojui papildomus šaltinius naudojantis naršymo istorija [8], [9]. Taip pat yra ir kitų modelių vartotojo demografinei informacijai atrinkti. Visi šie metodai naudojami norint nustatyti vartotojų interesų grupes, siekiant efektyvinti naršymą ar pasiūlyti papildomą informaciją, tačiau šiuos metodus galima patobulinti ir identifikuoti galimai pavojingą veiklą. Darbe nagrinėjama galimybė identifikuoti vartotoją, kuris potencialiai gali būti suinteresuotas atlikti pavojingą veiklą (KPK kūrimas, interesai šaunamaisiais ginklais, sprogmenimis, apgaule). Šiam tikslui pasirinkta identifikuoti ir rinkti vartotojo interesus, kategorizuojant jo lankomus puslapius ir naudojantis mašininio mokymosi modeliais prognozuoti galimai pavojingos veiklos interesus.

1.1. Tyrimo objektas

Vartotojo profilio informacijos ir naršymo istorijos, gaunamos iš realaus laiko reklamos pirkimo sistemos, panaudojimas mašininio mokymosi klasifikatorių pritaikymui, siekiant identifikuoti potencialiai kenksmingus vartotojus.

1.2. Darbo tikslas ir uždaviniai

Paruošti metodą identifikuoti potencialiai kenksminus vartotojus renkant jų profilius ir naršymo istorijos informaciją.

- Atlikti vartotojo profilio informacijos struktūros ir gavimo šaltinių analizę
- Išanalizuoti egzistuojančius vartotojo profiliavimo būdus bei jų efektyvumą
- Patobulinti vartotojo profiliavimo būdus ir metodus
- Paruošti metodą, leidžiantį aptikti potencialiai pavojingus vartotojus
- Atlikti metodo efektyvumo bandymus

1.3. Temos naujumas

Yra atlikta daug tyrimų, kaip išnaudoti turimą informaciją apie vartotoją, siekiant pagerinti jo informacijos paiešką internete, siūlant susijusias temas ar interesus. Nagrinėjami vartotojo naršymo šablonai, aplankytų puslapių grupės, identifikuojama vartotojo demografinė informacija net vartotojui fiziškai pakeitus vietą ar įrenginį, tačiau ši informacija taip pat gali būti panaudota identifikuoti galimai pavojingą veiklą. Išanalizuoti moksliniai straipsniai bei konferencijų pranešimai koncentruojasi į naršymo efektyvumo gerinimą ar demografinės informacijos gavimą, potencialios veiklos aptikimą naudojantis mašininio mokymosi metodais, tačiau nėra žinomas metodas, kaip identifikuoti potencialiai kenksmingus vartotojus realaus laiko reklamos pirkimo sistemoje. Darbas panaudoja jau egzistuojančias mokslines žinias ir papildo modeliu potencialiai pavojingai vartotojo veiklai identifikuoti.

1.4. Temos aktualumas

Egzistuojant daugybei įvairaus tipo informacijai internete galima rasti šaltinių apie naujienas, kenksmingo programinio kodo (KPK) kūrimą, ginklus, sprogmenis ar komunikacijos kanalus pavojingos

veiklos planams derinti. KPK auga eksponentišku greičiu, 2018 m. įvykdyta 1,35 Tbs DDoS ataka, kurios mastas leistų atkirsti ne vieną šalį nuo interneto. Pastaraisiais metais Europoje nugriaudėjo sprogimai, sutrikdę sveikatą ar mirtinai sužaloję daugybę žmonių viešuose vietose. 2010 m. atakas įvykdęs savižudis sprogdintojas Abdulwahab al-Abdaly yra pavyzdys, kad ir teroristai naudojasi visuomenei prieinamais portalais, skaito ir rašo žinutes diskusijų forumuose, turi socialinių tinklų anketas. Pavyzdžiui, prieš įvykstant dviems teroristinėms atakoms Norvegijoje, „Facebook“ ir „Twitter“ paskyrose buvo parašyta žinutė „2083, Europos nepriklausomybės deklaracija“. Kol kas pagrindinis dėmesys internete skiriamas vartotojo patogumui, tačiau taip pat turi būti plėtojamas ir potencialiai pavojingos veiklos aptikimas. Manoma, kad įrenginių, skirtų naršyti internete, 2020 metais bus apie 30 milijardų, taigi fiziškai peržiūrėti kiekvieno įrenginio duomenis turės būti pakankamai sudėtinga, todėl darbe nagrinėjamas būdas – identifikuoti galimai pavojingus vartotojus automatiniu būdu. Įgyvendintas modelis gali būti panaudojamas platesniame tinkle (ISP).

1.5. Tyrimo metodika

Analitinėje darbo dalyje atliekama profilio informacijos struktūros ir gavimo šaltinių analizė. Išsiaiškinami metodai identifikuoti vartotoją, rinkti lankymosi informaciją, galimybė identifikuoti interesus ir aptikti tą patį vartotoją, jam pakeitus lokaciją bei internetinį naršymo įrenginį. Išanalizuotos metodikos yra papildomos ir toliau taikomos darbe siekiant ištirti vartotojo interesus bei identifikuoti potencialiai pavojingus vartotojus pagal jų naršymo istoriją, įrenginių informaciją, demografinės informacijos ir papildomų interesų atpažinimą.

1.6. Mokslinė darbo vertė

Praplėsti mašininio mokymosi algoritmai potencialiai pavojingai veiklai identifikuoti gali būti panaudoti platesniu mastu. Sistemą galima plėsti pritaikius naujas taisykles vartotojų pavojingumo požymiams žymėti, taip siekiant sukurti našesnį identifikavimo algoritmą.

1.7. Siektini darbo rezultatai

- Papildyti egzistuojančios realaus laiko reklamos pirkimo sistemos architektūrą;
- Paruošti metodą rinkti vartotojų profilius ir stebėti jų pokyčius;

- Išanalizuoti surinktus duomenis;
- Paruošti metodą, kaip identifikuoti potencialiai kenksmingus vartotojus;
- Išbandyti metodą su realiais duomenimis;
- Identifikuoti potencialiai kenksmingus vartotojus.

1.8. Darbo struktūra

Antrajame skyriuje yra nagrinėjama realaus laiko reklamos pirkimo sistema, gaunanti daug informacijos apie vartotojus, jų demografinę informaciją, naršymo istoriją. Nagrinėjami metodai norint atpažinti vartotojo unikalumą, privatumą (ne)saugumą internete, tinklalapio ir vartotojo interesų ryšį bei anonimiškų duomenų (demografinės informacijos) atskleidimą. Apžvelgiami naudojami metodai potencialiai kenksmingiems vartotojams aptikti. Nagrinėjami mašininio mokymosi metodai siekiant sukurti potencialiai kenksmingų vartotojų klasifikatorių.

Trečiajame skyriuje aprašomas siūlomas metodas norint identifikuoti potencialiai kenksmingus vartotojus, aptariama metodo įgyvendinimui reikalinga architektūra, pateikiama informacija apie duomenų rinkimą, tvarkymą, paruošimą mašinio mokymosi klasifikatoriams. Apžvelgiami rezultatai iš kelių klasifikatorių ir atliekama identifikuotų potencialiai kenksmingų vartotojų apžvalga.

2. Susijusios literatūros analizė

Šiame skyriuje nagrinėjama, kaip veikia realaus laiko reklamos pirkimo sistema, kokie yra metodai vartotojui pateikti tinkamą reklamą. Siekiama išsiaiškinti, kaip renkama vartotojo informacija, padedanti pateikti vartotoją sudominančią reklamą. Atliekama vartotojo privatumo ir anonimiškumo analizė, pateikiami metodai vartotojo demografinei informacijai ir interesams nustatyti. Apžvelgiami egzistuojantys metodai potencialiai kenksmingiems vartotojams aptikti.

2.1. Realaus laiko reklamos pirkimo sistema

Darbe parengiamas būdas norint aptikti potencialiai kenksmingus vartotojus naudojantis informacija, gaunama iš realaus laiko reklamos pirkimo sistemos.

Pirmoji reklama internete pasirodė 1994-aisias metais, kuomet buvo apie 30 milijonų interneto vartotojų. Iki 2005-ųjų buvo išbandyti įvairūs reklamos pateikimo modeliai [10] ir galiausiai buvo sukurta realaus laiko reklamos pirkimams (angl. *Real time bidding (RTB)*) skirta platforma. Platforma pradėta naudoti *„ADSDAQ“, „AdECN“, „DoubleClick Advertising Exchange“, „adBrite“* ir *„Rich Media Exchange“*, kuriuos bendrai vadiname reklamos tinklais (angl. *ad exchanges*). Priešingai nei tradiciniai reklamos kanalai, šie tinklai agreguoja keletą reklamos kanalų, balansuoja pasiūlos ir paklausos rinkas ir naudoja aukciono principą siekdami parodyti vartotojui reklamą, kai jis krauna puslapį. Reklamos tinklai ir individualūs leidėjai gali pasipelnyti iš tokio modelio: reklamos leidėjai parduoda parodymus reklamuotojams, kurie domisi susijusiais vartotojais, o reklamuotojai savo ruožtu gali susijungti su daugiau leidėjų geresnei vartotojų atitikčiai pasiekti ir, gaudami varotojų informaciją, gali pirkti parodymus realiu laiku.

Tuo pačiu laiku pasirodė platformos su skirtingomis funkcijomis: paklausos pusės platforma (angl. *deman side platform (DSP)*) skirta aptarnauti reklamdavius ir pasiūlos pusės platforma (angl. *supply side platform (SSP)*), skirta aptarnauti leidėjus.

Realaus laiko reklama bėgant laikui sugeneruoja didelius kiekius duomenų. DSP Fikusu teigia, kad apdoroja 32 milijardus reklamos parodymų per dieną, DSP Turn teigia, kad piko metu apdoroja 2.5 milijono reklamos parodymo užklausų per sekundę. Kad geriau įsivaizduotume srauto dydį, galime jį palyginti su Niujorko vertybinių popierių birža, kurioje prekiaujama po 12 milijardų akcijų kasdien, todėl galime teigti, kad reklamos sandorių apimtis viršija finansų rinkos sandorių apimtį. Galbūt dar svarbiau yra tai, kad

reklamos pramonė kompiuterių mokslininkams ir ekonomistams suteikia unikalią galimybę mokytis ir suprasti internetinį srautą, vartotojų elgesį ir paskatas bei internetinius sandorius. Reklamos pramonėje visas interneto srautas, kaip reklamos pardavimo sandorių forma, yra sujungtas svetainėse ir jungia vartotojus visame pasaulyje. Svarbus techninis reklamos internete tikslas - automatiškai pateikti reikiamus parodymus tinkamiems naudotojams tinkamu laiku ir tinkama kaina, dėl kurios susitarė reklamuotojas ir leidėjai.


```
{
  "id": "c6987c2b-edb4-4b7b-b8cf-157af1d485e3",
  "site": {
    "id": "gumgum_www.answers.com_ed2265d8",
    "domain": "answers.com",
    "page": "http://www.answers.com/article/31029589/insanely-useful-life-hacks-to
      -make-everything-easier?paramt=null&param4=fb-us-de-red&param1=tattoo&param2
      =67660042&s=8"
  },
  "user": {
    "id": "5e29eb00-c30a-416e-9d2a-2e18901f0916",
    "buyeruid": "CAESEHL-9O4oJOAiC1Y0O2EHTcE"
  },
  "device": {
    "language": "en",
    "w": 1920,
    "geo": {
      "country": "US",
      "lon": -80.237,
      "city": "West Palm Beach",
      "lat": 26.638,
      "zip": "33414",
      "region": "FL"
    },
    "os": "Windows",
    "h": 1080,
    "ip": "73.139.39.18",
    "ua": "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:47.0) Gecko/20100101 Firefox/47.0"
  },
  "imp": [
    {
      "bidfloor": 3.213,
      "id": "1",
      "banner": {
        "pos": 1,
        "h": 600,
        "w": 160
      },
      "bidfloorcur": "USD"
    }
  ]
}
```


1 pav. „Realaus laiko reklamos pirkimo sistemos aukciono užklausa“

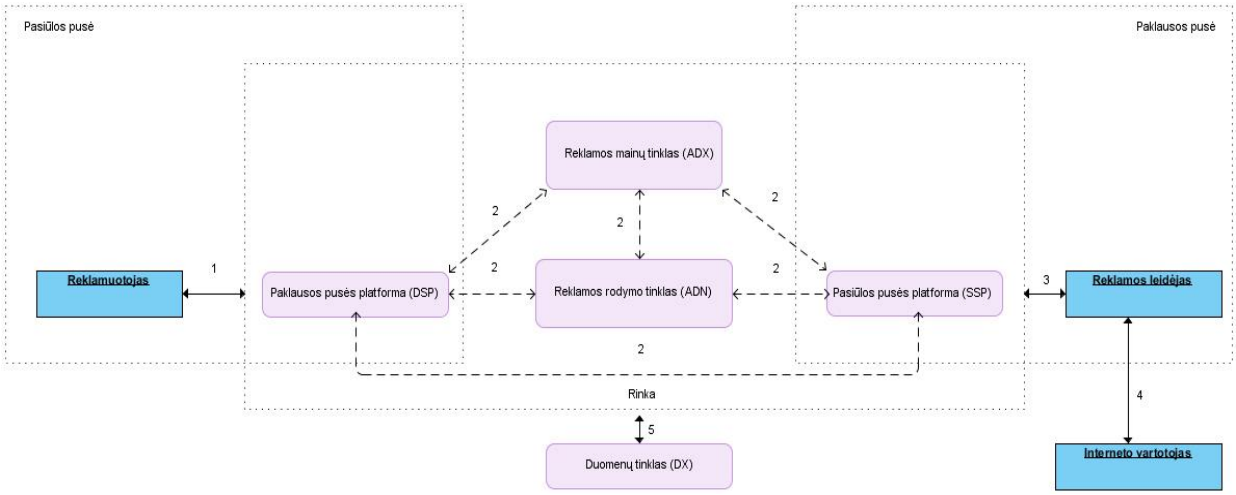
Realaus laiko reklamos pirkimo sistemoje reklamos parodymas sujungtas su vartotojo sekimo slapuku, suteikia ekosistemai galimybę sekti vartotoją, rinkti informaciją apie interesus, modeliuoti elgseną, siekiant sukurti personalizuotą reklamą. Tai leidžia automatiniais algoritmais automatizuoti ir optimizuoti reklamos ir interneto vartotojų atitikimą. Realaus laiko reklamos sistema tapo reikšmingu duomenų mokslo tyrimų mūšio lauku. Tai tapo daugelio tyrimų temų taikymo vieta. Tyrimai vykdomi įvairiomis temomis:

vartotojo reakcijos vertinimas (ar reklama bus paspausta?), elgsenos modeliavimas (vartotojas mėgsta automobilius), žinių išgavimas, sukčiavimo aptikimas bei rekomenduojančių ir personalizuojančių sistemų gamyba.

2.1.1. Realaus laiko reklamos pirkimo platformų ekosistema

Be keturių pagrindinių žaidėjų tipų, realaus laiko reklamos sistema (RTB) sukūrė naujus įrankius ir platformas, kurios yra unikalios ir vertingos.

- Paklausos pusės platformos (DSP): aptarnauja reklamuotojus ar reklamos agentūras automatiškai siūlydamos reklamos kampanijas keliuose tinkluose;
- Tiekimo pusės platformos (SSP): aptarnauja leidėjus, registruojant inventorizaciją (reklamos rodymo pozicijas) keliuose reklamos tinkluose ir automatiškai priima naudingus reklamos parodymus;
- Reklamos tinklai (ADX): sujungia kelis reklamos tinklus. Kai leidėjai prašo atitinkančios kontekstą reklamos, reklamos tinklas susijungia su reklamos rodymo tinklais (ADN) realiu laiku, kad turėtų didesnį pasirinkimą potencialiai reklamai;
- Duomenų tinklai (DX), dar vadinami duomenų valdymo platformomis (DMP), aptarnauja DSP, SSP ir ADX perduodami vartotojo istorinę informaciją (dažniausiai realiu laiku) geresniam reklamos parinkimui.



2 pav. „RTB ekosistema“

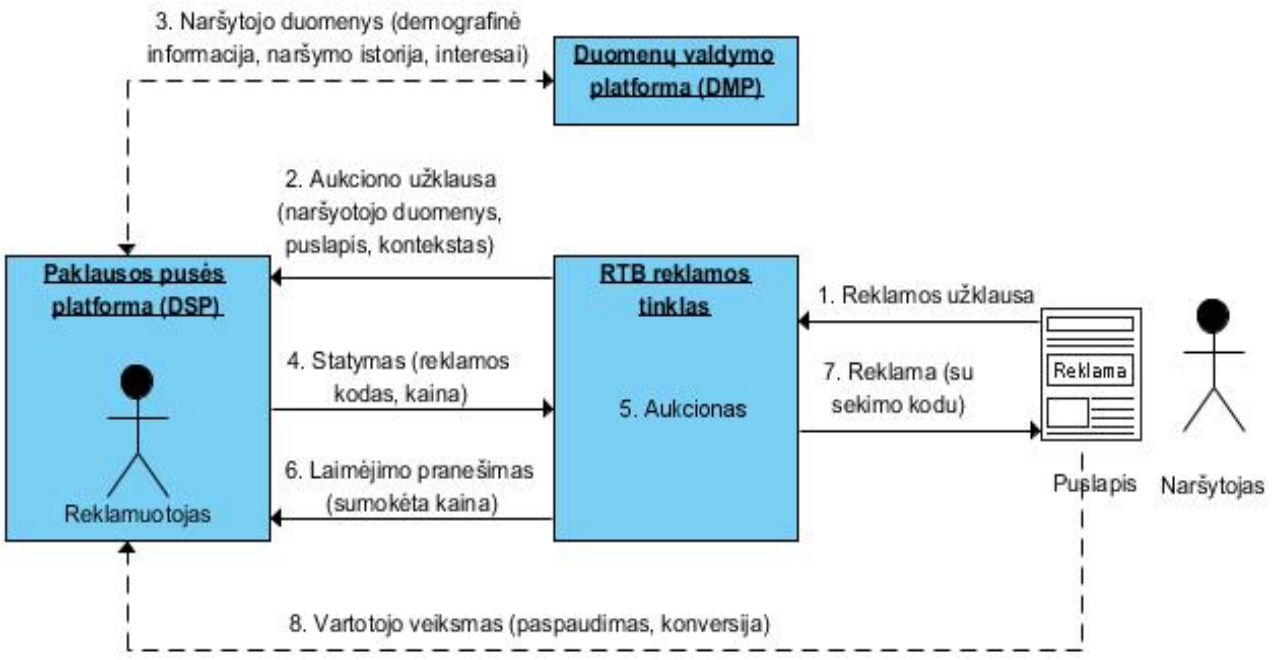
1 pav.: Skirtingi žaidėjai RTB ekosistemoje:

1. Reklamuotojas sukuria reklamos kampanijas rinkoje;
2. Rinka įgyvendina reklamos kampanijas balansuodma impresijas tarp pasiūlos ir paklausos pusių, taip siekdama užtikrinti gerą rezultatą;
3. Leidėjas registruoja parodymus rinkoje;
4. Vartotojas generuoja reklamos pirkimo užklausas;
5. Rinkos dalyviai gali naudotis duomenų platforma (DMP), kad gautų vartotojo profilį realiu laiku.

DSP, SSP, ADX ir DX atsirado dėl tūkstančių internetinės reklamos tinklų, kurie tapo kliūtimi reklamuotojams ir leidėjams patekti į internetinės reklamos verslą. Reklamuotojai turėjo dažnai kurti ir palaikyti reklamos kampanijas, kad padidintų reklamuojamų plotų apimtis. Taip pat turėjo analizuoti kampanijų rezultatus daugelyje platformų. Leidėjai turėjo atidžiai stebėti ir palyginti keletą reklamos tinklų, siekdami optimizuoti pajamas. ADX pasirodė, kaip kelių reklamos tinklų junginys, sprendžiantis paminėtas problemas. Reklamuotojai gali kurti ir analizuoti reklamų kampanijos efektyvumą vienoje platformoje, o leidėjai gali užsiregistruoti ADX ir surinkti optimalų pelną be jokių rankinių trukdžių. ADX galima suskirstyti į dvi kategorijas: SSP ir DSP. DSP veikia kaip reklamdavių agentūra, siūlydama ir stebėdama parinktus reklamos tinklus. SSP veikia kaip leidėjų agentūra – parduoda parodymus parinkdama optimaliausius nustatymus. Pagrindinė abiejų platformų idėja yra ta pati – jos bando sukurti suvienodintą klientų rinką tam, kad sumažintų darbą, reikalingą atlikti žmonėms, ir subalansuotų pasiūlą įvairiose rinkose. Dėl galimybių ir pelno riba tarp šių platformų tampa mažiau apčiuopiama. DX renka vartotojo duomenis ir anonimiškai parduoda DSP, SSP, ADX ir reklamdaviams realaus laiko reklamos pirkimo aukciono metu, kad pagerintų reklamos ir vartotojo sugretinimą, taip didinamas reklamos efektyvumas ir mažinami kaštai. Ši technologija paprastai vadinama elgsenos sekimu ir panaudojimu, taigi jei vartotojo ankstesni duomenys rodo susidomėjimą reklamuotojų produktais ar paslaugomis, reklamuotojai turi didesnę tikimybę gauti pajamų iš reklamos parodymo ir yra linkę mokėti daugiau.

2.1.2. Vartotojo elgsenos sekimas

Naudojantis RTB ekosistema reklamuotojai gali pirkti mažesnes vartotojų dalis ar juos dominančius segmentus, jeigu turi informaciją apie vartotojų elgesį, kuris buvo anksčiau pastebėtas ir išsaugotas DX.



3 pav. „Vartotojo elgsenos sekimas“

Diagramoje parodoma, kaip RTB veikia su DX, kad sektų vartotojo elgsenos ir interesų informaciją nuo momento, kai vartotojas atsiunčia puslapio krovimo užklausą iki per 100ms įvykusio aukciono ir parodytos reklamos su sekimo kodu:

0. Vartotojo naršyklei kraunant puslapį yra kraunamas reklaminis plotas;
1. RTB reklamos tinklui siunčiama reklamos užklausa;
2. Reklamos tinklas siunčia užklausą aukcionų dalyviams (paklausos pusės platformoms) su pirmine vartotojo informacija (naršyklė, IP adresas, informaciją apie puslapį, reklamos pozicija);
3. Paklausos pusės platforma (DSP) gali kontaktuoti su duomenų valdymo platforma (DMP), kad gautų papildomos informacijos apie vartotoją (demografinė informacija, naršymo istorija, interesai);
4. Jeigu reklamuotojas nusprendžia pirkti reklamos parodymą, tuomet sugeneruoja statymo užklausą ir siunčia RTB reklamos tinklui (internetinei drabužių parduotuvei gali būti aktualu, jei vartotojas ieško batų, todėl tikėtina, kad šis vartotojas gali būti linkęs pirkti ir parduotuvė yra pasirengusi mokėti daugiau už reklamą, parodytą tokiam vartotojui);
5. Laimėtojas parenkamas reklamos tinklų ir pasiūlos pusės platformų (dažniausiai pagal antros kainos aukcioną – didžiausiai statęs pirkėjas moka antrą pagal dydį kainą);

6. Laimėjimo pranešimas siunčiamas laimėjusiai paklausos pusės platformai (DSP);
7. Reklama (kuri buvo siunčiama statymo metu), kartu su sekimo kodu yra rodomi vartotojui;
8. Sekimo kodas seka informaciją apie vartotojo veiksmą (ar vartotojas pamatė reklamą, ar paspaudė? Ar pirko batus?).

Toks reklamos pirkimo/pardavimo procesas fundamentaliai skiriasi nuo ankstesnių reklamos prekybos procesų, kuomet informacija apie vartotojus nebuvo žinoma ir reklama būdavo rodoma atsitiktiniams vartotojams, naršantiems internete. RTB sistemoje aukcionų užklausų metu visuomet yra papildomos informacijos apie vartotoją, lankomą puslapį, poziciją lankomame puslapyje. Paprasčiausioje vartotojo elgsenos sekimo formoje, sekimas naudojamas antriniam vartotojo ir reklamos suderinimui (angl. *re-targeting*): elektroninės parduotuvės gali sekti vartotojus, kurie suruošė pirkinių krepšelį, tačiau išėjo iš parduotuvės nebaigę užsakymo. Tokiu atveju, pasinaudojant sekimo kodais ir turima informacija, galima vartotojui rodyti pasiūlymą užbaigti užsakymą, net kai jis naršo kituose puslapiuose.

Kiekvienos impresijos aukciono sprendimas įgalina vartotojo sekimą per puslapius, tačiau išlieka problema kaip identifikuoti vartotoją tarp skirtingų pasiūlos pusės platformų (SSP), kadangi kiekviena platforma gali atsiųsti skirtingą vartotojo identifikatorių (vartotojas ABC „DoubleClick Ad Exchange“ turi identifikatorių CASEhagst459583d, „adBrite“ – fega-saneneja-jnbdsahj-54as).

2.1.3. Vartotojo sekimas

Dažniausiai vartotojas yra atpažįstamas naudojantis HTTP slapukais (angl. *cookie*), kurie sukurti tam, kad puslapiai atsimintų individualių vartotojų statusus: tokius kaip prekių krepšelis ar naršymo istorija, pagal kurią generuojamas kintantis, suasmenintas puslapio turinys. Slapukas yra nedidelis informacijos kiekis, kuris siunčiamas iš puslapio ir saugomas vartotojo naršyklėje. Kievienąkart vartotojui ateinant į puslapį, naršyklė automatiškai prideda slapuko duomenis kartu su HTTP užklausa ir siunčia ją puslapiui. Slapukai yra pririšti prie svetainės adreso, tad jei naršyklė siunčia užklausą į „pavyzdys.lt“, pavyzdys.lt gali „prikabinti“ sausainėlį su vartotojo ID (vartotojas337) ir kitąkart kraunant puslapį pavyzdys.lt į serverį bus siunčiamas vartotojo naršyklėje užsaugotas slapukas ID: vartotojas337. Kita sveitainė, pvz. vgtu.lt negali perskaityti slapuko, skirto pavyzdis.lt (naršyklė tiesiog nesiunčia šių duomenų, kadangi jie priklauso kitam internetiniam adresui). Toks veikimas įsigaliojo, kai buvo pradėta taikyti kryžminio ryšio politika (angl. *Cross-origin policy*). Kryžminio ryšio politika įsigaliojo siekiant saugoti vartotojo privatumą.

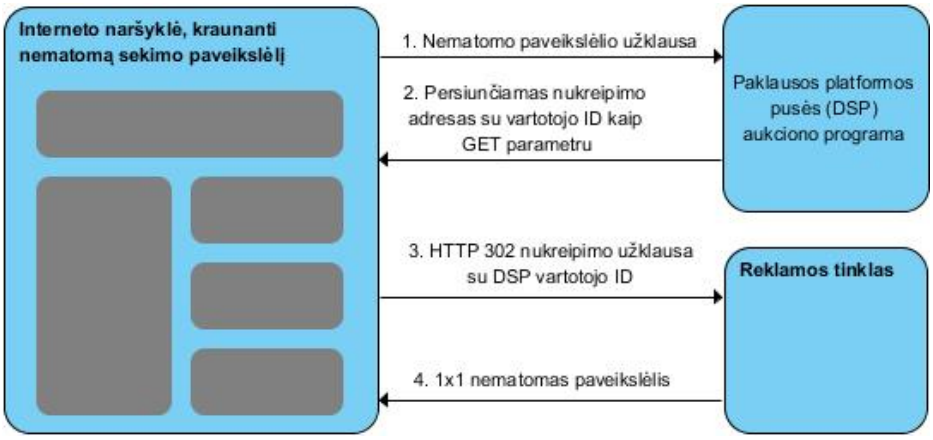
Rodant reklamą internete, kiekvienas paslaugos tiekėjas (paklausos pusės platforma DSP ir duomenų tvarkymo platforma DMP) naudojasi vienu internetiniu adresu tam, kad sukurtų savo vartotojo ID sistemą tarp skirtingų interneto puslapių. Tai įgyvendinama pasinaudojus HTML kodu, kuris integruojamas į puslapius, prie kurių turimas priėjimas. Yra žinoma nemažai trečiųjų šalių, kurios naudoja savo slapukus, pvz. „New York Times“ turi keletą šaltinių, rašančių savo slapukus.

Visos sekimo sistemos turi tik vietinį vartotojo vaizdą. Kadangi egzistuoja labai daug internetinių puslapių, tad turėti vien vietinį vaizdą neužtenka, kad būtų galima efektyviai parinkti reklamas vartotojui. Kai naršyklė siunčia užklausą reklamos parodymui, ji persiunčia tik tą informaciją, kuri yra pririšta prie reklamos tinklo adreso (reklamos-tinklas.lt), vadinasi reklamos tinklas nežino, kokią informaciją turi kiekvienas aukciono dalyvis ar duomenų platformos, kadangi jos veikia kitais interneto adresais (ad.dsp.lt), tad reklamos tinklui siunčiant aukciono užklausas, paklausos pusės platformoms siunčiamas reklamos tinklo vartotojo ID. Paklausos pusės platformos nežino duomenų apie tokį vartotoją, todėl yra sunku nuspręsti kokią reklamą rodyti. Siekiant išspręsti šią problemą, paklausos pusės platforma ir reklamos tinklas turi sinchronizuoti slapukus tam, kad paklausos pusės platforma galėtų sekti vartotojo veiksmus kituose internetiniuose puslapiuose. Procesas, kurio metu yra suderinami slapukai, vadinamas slapukų sinchronizacija (angl. *Cookie Syncing*).

2.1.4. Slapukų sinchronizacija

Slapukų sinchronizacija yra techninis procesas, kurio metu vartotojų sekimo platformos sužino skirtingus to paties vartotojo ID. Dažniausiai procesas įgyvendinamas pasinaudojant HTTP 302 nukreipimo užklausa tam, kad informacija būtų pasiekiama daugiau nei per vieną internetinį adresą. Sinchronizacijos procesas prasideda, kai vartotojas aplanko puslapį (puslapis.lt), kuris turi integravęs HTML kodą su trečiosios šalies kodu (ad.dsp.lt). HTML kodas dažniausiai turi 1x1 nematomą GIF paveikslėlį, pvz.: `<img src="http://ad.dsp.lt/pixel?vartotojoId=3545afi-eawjh-fnjwwa&dest=">`.

Nepaisant to, kad vizualiai paveiksliukas nėra matomas, naršyklė vis vien siunčia užklausą jį užkrauti, taigi puslapis (puslapis.lt) yra kraunamas iš vieno internetinio adreso, o nematomas paveikslėlis (ad.dsp.lt) – iš kito, todėl sekimo sistema (ad.dsp.lt) gali naudoti savo slapuką. Tokiu būdu trečioji šalis gali skaityti ir rašyti slapukus. Slapukų sinchronizacija pradeda veikti, kadangi vietoj iškart rodomo 1x1 nematomo paveikslėlio pirma yra įvykdomas vartotojo nukreipimas į kitą paslaugą, kuri prideda savo vartotojo ID ir tuomet nukreipia į adresą, perduotą GET parametru.



4 pav. „Slapukų sinchronizacija“

Kaip parodyta paveiksle:

1. Naršyklė siunčia užklausą užkrauti paveikslėlį iš ad.dsp.lt (kartu siunčiant ir slapukus priklausančius ad.dsp.lt). Jeigu ad.dsp.lt neatpažįsta vartotojo, t.y. laiko jį nauju, tuomet sugeneruoją naują ID ir įrašo slapuką;
2. Sekimo kodas iš ad.dsp.lt gauna vartotojo ID, bet vietoj rodant 1x1 nematomą paveiksliuką, įvykdo HTTP 302 nukreipimą į reklamos tinklą (įrašydamas vartotojo ID į GET parametrą);
3. Naršyklė siunčia užklausą į reklamos tinklą (kartu su reklamos tinklo slapuku jei jis egzistuoja);
4. Reklamos tinklas rodo 1x1 nematomą paveikslėlį ir įsirašo vartotojo ID savo sistemoje.

Aprašytas slapuko sinchronizavimo procesas taikomas, kai ad.dsp.lt gali įdiegti sekimo kodą į savo puslapius, tačiau galimas ir kitas atvejis – slapukų sinchronizavimas rodant laimėtų reklaminių pozicijų reklamas realaus laiko reklamos pirkimo sistemoje. Tokiu atveju reklamos tinklas nukreipia vartotoją į ad.dsp.lt kartu su reklamos tinklo vartotojo ID. Ad.dsp.lt gauna užklausą, perskaito savo slapuką ir išsaugo vartotojo ID duomenis (reklamosTinkloID: AAAA-BBBB-CCCC-DDDD, dspID: user2564).

Vartotojo ID sinchronizacija nėra patikima tik tokiu atveju, jei vartotojas išvalo naršyklės slapukus ar naudoja slaptą (angl. *incognito*) režimą, kai naršyklė išjungia slapukų naudojimą, todėl atsiranda poreikis kitokioms sekimo technikoms. Naršyklės piršto antspaudas (angl. *fingerprinting*) – technika, kurioje naudojama informacija apie nutolusį kompiuterį ar įrenginį tam, kad atpažintų unikalų vartotoją.

2.2. Vartotojo profilio informacijos struktūros ir gavimo šaltinių analizė

Egzistuoja keli vartotojo informacijos atpažinimo metodai. Informacijos rinkimas ir atpažinimas pradėjo tobulėti gausėjant paieškos varikliams bei informacijos įvairovei internete. Kilo poreikis vartotojui pasiūlyti jam tikslingą informaciją, todėl atsirado rekomendaciniai varikliai. Rekomendaciniai varikliai yra puikus pavyzdys kaip galima rinkti informaciją apie vartotojo interesus. Analizuojami paieškos varikliai ir kiti metodai leidžiantys kaupti duomenis apie vartotojo interesus.

2.2.1. Vartotojo unikalumo atpažinimas

Aplankytų puslapių istorija gali būti laikoma vartotojo piršto antspaudu internete, pagal kurį galima identifikuoti unikalų naršytoją [1]. Naršančio vartotojo naršyklės nustatymai yra tiesiogiai susiję su vartotojo interesais, tad manoma, kad tokie nustatymai yra patys iš saves individualūs tarsi žmogaus biometriniai duomenys – pirštų antspaudai ar akies rainelė. Vartotojo naršymo istorija gali būti priskiriama interesų grupėms, norint sužinoti puslapio kategoriją iš kategorizavimo paslaugos ir priskirti kiekvieną įstorijos įrašą atitinkamai grupei. Tuomet vartotojo profilis sudaromas pagal gautas kategorijas. Ištirta vartotojų naršymo istorija parodo, kad didžiosios daugumos vartotojų naršymo sesijos yra unikalios ir pagal tai galima sudaryti abstraktesnes internetinių puslapių naršymo kategorijas nei analizuojant kiekvieno unikalaus puslapio kategoriją. Taip pat įmanoma identifikuoti vartotojo interesą pagal naršymo istorijos individualumą skaičiuojant kad ir mažą duomenų kiekį turinčią istoriją, pvz., kokiuose populiariuose puslapiuose vartotojas lankėsi – tokie duomenys yra pakankamai nekintantys laiko atžvilgiu. Tam tikrais atvejais užtenka vos vienos naršymo sesijos visiškai naujam vartotojui identifikuoti ir sukategorizuoti interesus. Kadangi kiekvieno vartotojo naršymo sesijos yra skirtingos ir unikalios, egzistuoja privatumo praradimo rizika. Naršymo įgūdžius pakeisti yra pakankamai sudėtinga, tad žinant vartotojo naršymo šablonus galima identifikuoti tą patį vartotoją netgi skirtinguose naršymo įrenginiuose (kompiuteriai, planšetės) ar vartotojui pakeitus savo gyvenamąją vietą.

2.2.2. Vartotojo privatumas internete

Dažnai vartotojo naršymo istorijos privatumą sunku apsaugoti dėl aplinkinių faktorių, pavyzdžiui vartotojo naršyklės pažeidžiamumų. 2010 metais į viešumą pasklido informaciją apie galimybę sužinoti vartotojo naršymo istoriją naudojant CSS :visited selector'ių. Į internetinio puslapio HTML kodą įterpiamas vartotojui nematomas dizaino elementas su dominančiomis nuorodomis į kitus puslapius. CSS pagalba jau

aplankytoms nuorodoms yra priskiriama kita nei standartinė nuorodos spalva ir, panaudojus JavaScript skaitant DOM (angl. *Document Object Model*), galima identifikuoti, ar nuorodos spalva atitinka jau aplankyto puslapio nuorodos spalvą taip identifikuojant puslapį kaip jau aplankytą. Modernios bei dauguma senesnių naršyklių įdiegė apsaugą nuo šio pažeidžiamumo, tačiau dar egzistuoja metodai aplankytiems puslapiams identifikuoti. Tyrimai rodo, kad apie 25% stalinio kompiuterio tipo naršyklių išlieka pažeidžiamos kaip ir dauguma mobiliųjų naršyklių [2].

2.2.3. Tinklalapio ir vartotojo interesų koreliacija

Pastaraisiais metais mokslininkai išreiškia didėjantį interesą populiariais, dažnai lankomais ar neseniai lankytais puslapiais, kadangi jie atspindi vartotojo elgesį ir nišinius interesus, retėjantį naujos informacijos poreikį, pramogavimo būdą ir hobius.

Internetiniai puslapiai gali būti skirstomi į kategorijas, pavyzdžiui www.google.com yra paieškos variklis, o www.delfi.lt yra naujienų portalas, tad žinant puslapių kategorijas jas galima skirstyti į bendresnes kategorijas, kurios gali būti siejamos su naršančiojo interesais.

Profilių įvairovė ir panašumas gali būti apskaičiuojami pagal aplankytų puslapių aibių panašumus naudojantis Žakardo indeksu. Žakardo indeksas parodo, kad aibės yra vienodos jei Žakardo koeficientas yra lygus 1 arba aibės yra panašios jei koeficientas yra 0,7.

Autoriai tyrė vartotojų naršomų puslapių turinį ir pastebėjo panašumus tarp puslapių turinio. Išnagrinėjus vartotojo naršymo sesijos metu aplankytus puslapius P1, P2, P3, .., P7 pastebėta, kad žodis „dirbtuvė" buvo keturiuose paskutiniuose puslapiuose, o puslapis P6 buvo Google paieškos rezultatų puslapis, kuriame paspausta nuoroda vedė į puslapį taip pat turintį žodį "dirbtuvė" savo turinyje. Apmokytas mašininio mokymosi modelis identifikavo, kad jei žodis Ž yra paskutiniuose dvejuose aplankytuose puslapiuose labai tikėtina, kad sekančio puslapio turinyje taip pat bus žodis „dirbtuvė" [3].
Dauguma naršomų puslapių yra pakartotiniai apsilankymai jau aplankytuose šaltiniuose. Dauguma pakartotinių apsilankymų yra siauroje aibėje – naršytojo mėgstamiausias paieškos variklis, elektroninės parduotuvės, socialiniai tinklai ir nesenai lankyti puslapiai [4].

Autoriai atliko tyrimą prašydamį naršančių vartotojų pasižymėti dominantį turinį. Jei vartotojas aukoja laiką pažymėdamas žymes prie turinio ar parašydamas komentarą, tai rodo, kad vartotojui pateiktas turinys yra galimai įdomus [5]. Tradiciškai jei vartotojas spaudžia nuorodas, kurios yra sukategorizuotos, jis neženkliai padidina savo intereso susidomėjimo indeksą. Padidėjimas yra neženklus, kadangi vartotojai naršo puslapius dėl įvairių priežasčių, kad ir iš smalsumo, nebūtinai parodant susidomėjimą konkrečia tema.

Didesnė koreliacija tarp vartotojo elgesio ir intereso būtų, jei vartotojas pasižymėtų konkretų turinį, tad toks veiksmas didina intereso indeksą ženkliai ar labai ženkliai. Yra mažai tikėtina, kad vartotojas aukotų savo laiką pasižymėdamas turinį ar rašydamas komentarą jei nebūtų susidomėjęs naršomo puslapio ar turinio bendrine kategorija.

Letizia vartotojų intereso vertinimas susideda ne vien iš puslapio turinio, bet ir iš jame praleisto laiko. [8] Stebint vartotojo elgesį galima daug ką pasakyti apie jo intereseus. Kiekviena iš rekomendacinių dalių turi savo neapibrėžtumą, tačiau sudedant jas kartu gaunami tikslesni duomenys apie vartotojų interesus. Paspaudimas ant nuorodos gali indikuoti keletą dalykų: pirmiausia vartotojas galimai išreiškia interesą, kai nusprendžia spausti ant nuorodos, tačiau nežinant puslapio turinio spėjimas apie vartotojo interesą yra tik preliminarus. Jei vartotojas staiga grįžta atgal ar pereina į kitą puslapį neišsaugojęs jokios informacijos ar nepaspaudęs ant kitų lankomo puslapio nuorodų, galima teigti, kad vartotojas nesusidomėjo pateikta informacija. Letizia taupo vartotojui laiką mokydamasi kokiomis nuorodomis jis seka. Geras indikatorius apie vertingą resursą yra nuorodos paspaudimas dabartiniame puslapyje. Pastovus apsilankymas ir daug laiko praleisto puslapyje parodo vartotojo interesą.

Suformuluoti paieškos algoritmai, kurie iš anksto apsvarsto vartotojo sąveiką su įvairiomis turinio formomis pagal dabartinę paieškos sesiją [9]. Raidžių junginio “IR” paieška grąžina įvairiapusišką rezultatų sąrašą: kompanijos Ingersoll-Rand akcijų kainas, arabiškus puslapius iš Irano (su .ir domeno galūne), informaciją apie infraraudoną šviesą ir keletą rezultatų apie informacijos paiešką (angl. *Information retrieval*). Autoriai studijuoja kaip personalizuoti paieškos rezultatus pagal ieškantį žmogų tam, kad finansų analitikui pirma rodytų informaciją apie kompanijos Ingersoll-Rand akcijų kainas, o puslapius apie infraraudonąją šviesą – fizikui.

Atliktas tyrimas parodė, kad vartotojai ieško labai skirtingų dalykų netgi naudodami tą pačią paieškos užklausą. Išsiaiškinta, kad paieškos mechanizmai grąžina tinkamus rezultatus įvairiems žmonėms, tačiau prasčiau personalizuoja pagal kiekvieno naršančiojo interesus.

Dideli skirtumai pastebimi internetinio mokymose puslapiuose, todėl vartotojų elgesio kategorizavimas atskirais laiko intervalais yra tikslesnis. Geriausio laiko intervalo atrinkimas priklauso nuo pačio puslapio turinio, naujienų puslapius vartotojai naršo vienu metu, o socialinius tinklus ir asmeninius – kitu.

Kadangi vartotojo interesai bėgant laikui gali kisti, neseniai naršyti puslapiai geriau atspindi vartotojo interesus nei seniau naršytieji.

Priklausomybę turi ir puslapyje patalpintos informacijos tipas – socialiniai tiklai lankomi dažniau nei virtualaus mokymosi portalai, tačiau tai nereiškia, kad mokymasis vartotojui yra mažiau aktualus.

Peržiūrėjus puslapį yra gaunama kontekstinė informacija ir patalpinama faile. Teksto ištraukimas vykdomas dviem žingsniais: HTML kodai „<script>“, „</script>“ ar „<style>“, „</style>“ ir tekstas esantis tarp jų yra pašalinami; bet koks tekstas tarp „<“ ir „>“ yra pašalinamas. Tokių tekstų turinys saugomas dviejuose aplankuose – trumpojo ir ilgojo laikotarpių [11].

Trumpojo laikotarpio tekstinių failų sąrašas yra ribojamas iki penkių vienetų. Užsipildžius šiam katalogui, pirmas į jį įrašytas įrašas yra keliamas į antrąjį katalogą.

Puslapio duomenys gula į ilgojo laikotarpio katalogą ir ten būna žymiai ilgiau todėl, kad reikia aplankyti daug daugiau puslapių, kad ką tik įrašytas įrašas į ilgojo laikotarpio aplanką būtų iš jo pašalintas. Puslapių turinio byla yra pavadinta turinio MD5 kontrolinės sumos reikšme.

Tokiu vadinimo principu galima nustatyti, ar puslapis jau buvo aplankytas ir atnaujinti apsilankymo datą taip pažymint puslapį kaip trumpalaikį susidomėjimą.

Vartotojo profiliui sudaryti pirma imami puslapiai iš trumpalaikio susidomėjimo aplanko. Jų turinys grupuojamas pagal žodžius – taip gaunama unikalių duomenų aibė vadinama žodžių vektoriu.

Kiekvienas vektorius turi papildomą informaciją apie pasikartojimą tekste (kiek kartų žodis pasikartojo / visų žodžių tekste kiekis). Prieš apskaičiuojant puslapių vektorius pašalinami netinkami žodžiai.

Gavus vektorius apskaičiuojamas trumpalaikio aplanko bendrasis vektorius sujungiant atskirų puslapių vektorius, bet nesumuojant pasikartojimo koeficiento.

2.2.4. Anonimiškų duomenų išaiškinimas

Pastaraisiais metais reklama internete auga ypač greitai, todėl demografinis vartotojų atrinkimas yra labai svarbus. Yra pakankamai sudėtinga sužinoti vartotojo demografinę informaciją (tokią kaip amžius ar lytis). Autoriai [6] tyrė vartotojų demografinės informacijos gavybos būdus analizuodami vartotojų aplankytus puslapius pagal jau turimus įrašus apie kitus vartotojus, kurių demografinė informacija yra žinoma. Antras būdas – amžiaus ir lyties spėjimas pagal Bayesian metodą naudojant informaciją apie lankomų puslapių ryšį su amžiumi ir lytimi (gaunant duomenis iš puslapio log įrašų analizės).

Google paieškos sistema personalizuoja vartotojo paieškos rezultatą pagal ankstesnes paieškos sesijas ir paspaudimus iš rezultatų į puslapius.

Kadangi demografinės informacijos gavimas yra sudėtingas, vartotojai pirmiausiai suskirstomi į grupes pagal panašų naršymo šabloną [7]. Darant prielaidą, kad vartotojai elgiasi skirtingai skirtingu laiko metu ar skirtingomis naršymo sesijomis, vartotojai suskirstomi į pogrupius pagal jų naršymo sesijos laiką.

Reklamos pardavimo realiu laiku aukcionuose reklamos vienetas yra parduodamas aukcione per 200 milisekundžių po puslapio užkrovimo užklausos [10]. Per pirmas kelias milisekundes puslapio turinio kalbos, kokybės ir anonimiško vartotojo identifikacinis kodas yra siunčiamas į reklamos pardavimo tinklus. Reklamos pardavimo tinklai, kurie ir yra reklamos vienetų savininkai, išsiunčia aukciono užklausas reklamos tiekėjams. Reklamos tiekėjai turi programinius aukcionų dalyvius, kurie pagal pateiktą informaciją sprendžia maksimalią galimą statymą už reklamos vienetą. Informacija yra pateikiama atgal reklamos pardavėjui, kuris, naudodamas statymo agentus (angl. *bidding agents*), nusprendžia nugalėtoją pagal aukščiausią statomą pinigų sumą už konkretų reklaminį vienetą.

Šis procesas vartotojui yra nematomas, kadangi kol puslapis pilnai užsikrauna aukcionas jau būna baigtas. Kadangi aukcionų vyksta dešimtimis tūkstančių per sekundę, tokie duomenys yra puikus didelių duomenų kiekių šaltinis. Statymo agentai realaus laiko reklamų pardavimuose naudoja informaciją apie vartotojo neseniai atliktus veiksmus (istoriją) ir papildomai gaunamus duomenis (angl. *cookies*), kad vartotojui parinktų tinkamiausią reklamą ir atvaizdavimo tipą. Vartotojai dažnai tai pastebi: aplankoma internetinė parduotuvė, kurioje ieškoma spintelės ir kituose kraunamuose puslapiuose dažnai matomos spintelių reklamos. Tai vadinama pakartotiniu taikymu (angl. *re-targeting*). Žinoma realaus laiko reklamos pardavimai neapsiriboja tokia paprasta informacija iš reklamos tinklų pateikiamos informacijos, kadangi dažniausiai aukciono užklausos metu nėra pateikta pakankamai informacijos vartotojo interesams identifikuoti ir taip pateikti geresnę reklamą. Autoriai [10] siūlo papildomą būdą identifikuoti vartotojo interesus analizuojant vartotojų naršymo istoriją ir teigia, kad vartotojo interesai gali būti gana tiksliai identifikuoti naudojantis vien šiais duomenimis.

2.3. Metodai potencialiai kenksmingiems vartotojams aptikti

Internetinės socialinės platformos susiduria su neapykantos turiniu. Tai turinys, kuris išreiškia neapykantą asmeniui ar žmonių grupei. Toks turinys gali išgąsdinti, įbauginti ar net pastūmėti kitus vartotojus atlikti smurtinį nusikaltimą.

Autoriai [19], atliko tyrimą kaip kaip atpažinti tokio tipo vartotojus socialinėse platformose „Reddit“, „Voat“ ir pokalbių forumuose. Vartotojų grupės skirstomos pagal ryšį tarp kalbos manieros parinkimo ir socialinių taisyklių laikymosi, tokiu principu leidžiant kategorizuoti vartotojus naudojančius panašų kalbos braižą. Neapykantos grupės ir neapykantos kalba yra taikoma tikslinei bendruomenei, todėl

abi pusės negali egzistuoti viena be kitos. Siekdami efektyvaus tyrimo autoriai išanalizavo tris egzistuojančius metodus:

- Rasizmo aptikimas naudojantis „Naive Bayes“ klasifikatoriumi. Šis metodas nėra pakankamai efektyvus, kadangi identifikavus potencialiai rasistinę kalbą ir šiuos rezultatus peržvelgus žmonėms efektyvumas buvo įvertintas 33%;
- Kitas metodas naudojo atraminių vektorių klasifikatorius (angl. *support vector machine*) siekiant identifikuoti antisemitinius komentarus „Yahoo“ naujienų grupėse. Klasifikatoriaus treniravimo duomenys buvo parinkti rankiniu būdu, todėl, kaip ažvelgta šaltinyje, duomenys nėra vienareikšmiškai teisingi ir sunkiai identifikuoja komentarus, kuriuose neapykanta yra išreikšta netiesiogiai;
- Trečiame tyrime buvo naudojamas kalbinių taisyklių principu pagrįstas požiūris į socialinio tinklo „Twitter“ žinutes, kurios renkamos pagal iš anksto apibrėžtus užgaulius raktinius žodžius. Šis metodas taip pat turi reikšmingą neigiamą poveikį rezultatams.

Metodai nesuveikė, nes raktiniai žodžiai buvo parenkami ranka. Autoriai parinko užgauliai kalbai skirtas „Reddit“ bendruomenes ir panaudojo jų turinį apmokyti mašininio mokymosi modeliams. Taip pat, atsitiktiniu keliu buvo parinkti 460 tūkst. vartotojų komentarų, kurie nepriklausė užgaulios kalbos bendruomenėms taip mokydami modelius kuo skiriasi paprasta kalba nuo užgaulios. Mokymui panaudoti dažniausiai naudojami „Naive Bayes“, „Support vector machines“ ir logistinės regresijos algoritmai. Komentarų duomenys buvo apdoroti: tekstas pakeistas paverčiant didžiasias raides į mažasias, pašalinant nuorodas, taškus, kablelius ir kitus skiriamuosius simbolius. Prie kiekvieno komentaro pridėta „Reddit“ bendruomenė. Išmokytus modelius panaudoję bandomajam duomenų rinkiniui, paruoštus mašininio mokymosi klasifikatorius autoriai įvetino kaip geresnius nei nagrinėtus darbe.

Švedų gynybos tyrimų agentūra [20] viena iš pagrindinių problemų analizuojant informaciją internete įvardija tai, kad jis yra didžiulis, todėl neįmanoma rankiniu būdu ieškoti informacijos ir analizuoti visus gaunamus duomenis apie radikalizacijos procesus ir teroristų planus potencialių pavienių teroristų. Be visos medžiagos ir informacijos esančios internete, kurią indeksuoja paieškos sistemos taip pat yra milžiniškas informacijos kiekis vadinamąjame paslėptame internete (angl. *deep web*). Slaptasis internetas nėra indeksuojamas paieškos sistemų, todėl sekti ten sklindančią informaciją yra keblu. Sukurti visiškai automatines kompiuterines terorizmo paieškos sistemas, kolkas yra neįmana tiek dėl didelio duomenų kiekio tiek dėl gilių žinių, kurių reikia suprasti kas rašoma (ar skleidžiama kitu būdu – paveikslėliais, vaizdo

įrašais) ir apie tai daryti išvadas ar turinys yra potencialiai susijęs su kenkėjiška veikla. Tačiau egzistuoja kompiuterinės priemonės, kurios padeda analitikams atrasti skaitmeninius pėdsakus.

Siekiant išaiškinti potencialius pavienius teroristus autoriai aptaria pusiau automatinius įrankius ir technologijas, naudojamas žvalgybos analitikų, stebėti kenkėjiškus puslapius ir pokalbių platformas. Priešingai nei kitose metodikose, gynybos tyrimų agentūra sutelkia dėmesį į pavienių teroristų, o ne kenkėjiškos grupės identifikavimą ir pusiau automatinių įrankių (vietoj visiškai automatinių) panaudojimą. Identifikacijos problema yra padalinta į keletą mažesnių užduočių (ketinimas/motyvas, pajėgumas, tikimybė), vėliau agreguojant rezultatus į bendrą sistemą. Jei visi identifikatoriai atpažįsta rezultatą (yra motyvas, pajėgumas ir galimybė) nagrinėjamą atvejį pripažista kaip potencialiai kenksmingą. Toks užduoties išskaidymas dalimis yra naudingas tuo, kad nereikia turėti bendros visumos sprendimui atlikti. Jeigu žvalgybos darbuotojas tikina, kad reikalingas motyvas ir pajėgumas vartotojui įvertinti kaip potencialiam teroristui, viena posistemė gali ieškoti galimų motyvų (radikalūs pasisakymai), kita – pajėgumų (dažnas lankymasis įvairiuose puslapiuose diskutuojančiuose kaip gaminti sprogmenis). Rezultatai yra apjungiami ir vertinami kiek tikėtina, kad nagrinėjamas vartotojas yra ar linksta tapti teroristu. Agreguoti duomenys tampa sąrašu pavojingų internetinių vartotojų, kuriems taikomas papildomas stebėjimas. Kiekviena hipotezė yra temos indikatorius. Temos indikatoriai gali būti sudaryti iš kelių pavienių indikatorių: tiesioginis indikatorius – kai asmuo parašo radikalią žinutę pokalbių svetainėse, netiesioginis – atvejis kai asmuo gali turėti ryšių su pokalbių svetainės administratoriais. Surinkta informacija gali būti panaudota autoriams, naudojantiems skirtingus slapyvardžius apjungti (dažnai potencialiai kenksmingi vartotojai keičia savo slapyvardžius siekdami panaikinti savo skaitmeninius pėdsakus).

Kiekvienai indikacinei temai, žvalgyba siūlo ieškoti susijusių puslapių ar diskusijų platformų ta tema. Kadangi turinio kiekis internete yra labai didelis, nebūtų labai logiška ieškoti potencialiai pavojingų vartotojų be gairių, todėl temų paieška yra mažinama. Nepaisant to, kad yra daug internetinių puslapių apie ekstremistus, kurių neindeksuoja paieškos sistemos, nemažai jų yra prieinami plačiajai auditorijai, kadangi idėjos dalis yra iškomunikuoti ideologiją ir kitas žinutes dideliam kiekiui žmonių. Dažnai ekstrimistų puslapiai turi nuorodas į kitus puslapius ta pačia tematika, todėl tokie puslapiai yra naudojami kaip indikatoriai žymėti internetiniams vartotojams, galimai susidomėjusiems potencialiai kenkminga veikla. Siekiant rasti tokių teminių puslapių tinklą naudojama metodika lankyti puslapiuose esančias nuorodas ir rinkti informaciją apie jas (angl. *crawling*). Nuorodų lankymo procesas kartojamas tol, kol nebėra

neaplankytų nuorodų ar pasiektas nustatytas kiekis aplankytų šaltinių. Tokiu būdu sudaromas temos šaltinių tinklas. Būdas nėra vienareikšmikai teisingas, kadangi sekant nuorodas tikėtina patekti į su teroristinėmis idėjomis nesusijusius tinklapius. Siekiant atskirti teroristinius šaltinius nuo neįdomaus (teroristui) turinio atliekamas natūralios kalbos apdorojimas (angl. *natural language processing*). Švedų žvalgyba siūlo sudaryti raktažodžius susijusius su potencialiai kenksminga veikla ir sekti kiek ir kiek dažnai raktažodžių yra naudojami šaltinyje, taip nustatant ar šaltinis turi pakankamą kiekį raktažodžių priskirti jį kaip susijusį su teroristine veikla. Tokia pati metodika taikoma ir nesusijusiam turiniui žymėti – parenkami raktažodžiai, kurie indikuoja apie puslapio turinio nekenkėjiškumą. Skirtingų vartotojų vardų surišimas su vienu internetiniu vartotoju įgyvendinamas sekant informaciją apie bendrą elgesį: jeigu keli skirtingi vartotojų vardai rašo tuose pačiuose diskusijų platformose, komentuoja vienodose temose, reguliariai rašo tokio pačio tipo komentarus, tikėtina, kad tai yra ta pats vartotojas.

Įgyvendinus sistemą, žvalgyba teigia, kad rezultatus visuomet turi patikrinti žmogus, kad įvertintų kokios priežastys lėmė vartotojo priskyrimą prie potencialiai pavojingų ir nutartų ar reikalinga išaiškinti kas konkrečiai yra vartotojas ir pradėti papildomus jo sekimo veiksmus.

Tiriant galimybę išaiškinti kenkėjiškus vartotojus su minimalia turima informacija [21] identifikuotos penkios pagrindinės kenkėjiškų vartotojų charakteristikos:

1. Kenkėjiški vartotojai yra įvairūs ir kompleksiški. Vartotojai dažnai kuria kompleksišką ir ivairų turinį. Kompleksiškumas vertinamas nagrinėjant vartotojų vardus pasinaudojant Kolmogorov‘o kompleksiškumo metodu. Kolmogorovo vartotojo vardo sudėtingumas apibrėžiamas kaip trumpiausios programos ilgis, galintis atkurti naudotojo vardą universaliame kompiuteryje, pvz., Turingo mašinoje;
2. Kenkėjiški vartotojai yra demografiškai panašūs. Kriminologijos literatūroje yra žinoma, kad nusikaltimai koreliuoja su demografine informacija, todėl išaiškinant vartotojų demografiją (lytis, kalba, raštingumas) galima didinti metodų efektyvumą;
3. Kenkėjiški vartotojai yra anonimai. Kenkėjiška veikla reikalauja tam tikro anonimiškumo. Tyrime buvo aiškinamasi koks vartotojo vardo raidžių pasiskirstymas abecelėje;
4. Kenkėjiški vartotojai yra panašūs. Vartotojai reklamuojantys nelegalų produktą „pavojinga tabletė“ dažnai vartotojo varde naudoja žodžių junginį „pavojinga tabletė“;

5. Kenkėjiški vartotojai yra efektyvūs. Kartu su kompleksiškumu kenkėjiški vartotojai taip yra efektyvūs – jų tikslas yra atlikti dažną, greitą, didelio masto kenkėjišką veiklą.

Siekiant užkirsti apgavikiškus pinigų pervedimus, autoriai [22] analizavo mašininio mokymosi metodus identifikuoti tokio tipo pinigų pervedimus. Neuroninių tinklų pranašumą įvardija kaip gerus, šabloną atitinkančius klasifikatorius, gebančius priimti apibendrintus sprendimus naudojant nepilną ar nevisiškai tikslią informaciją. Priešingai nei tradiciniai statistiniai metodai kaip regresija, neuroniniai tinklai gali dirbti su palyginti mažai treniravimosi duomenų, kuriant spėjimo modelius, todėl yra puikiai panaudojami nusikaltėlių aptikimo situacijose, kadangi pakankamai mažas procentas pinigų pervedimuose yra apgavikiškas.

Apžvelgti kenkėjiškų vartotojų aptikimo metodai pasižymi ta pačia savybe – taikomas mašininis mokymasis siekiant iš dalies automatizuoti tokių vartotojų atpažinimą. Dažnai, po mašinio mokymosi atlikto klasifikavimo rezultatai yra perduodami žmogui galutiniam patikrinimui atlikti ir nuspręsti ar surinkti požymiai byloja apie potencialiai kenksmingą veiklą.

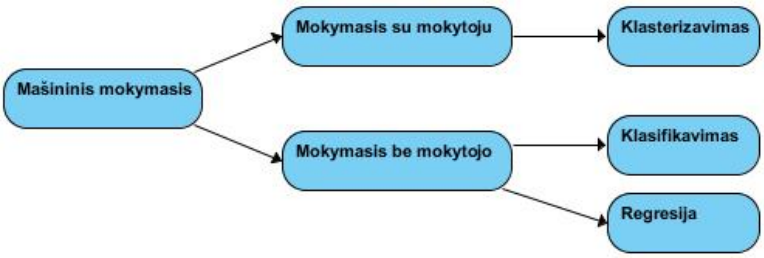
2.4. Mašininio mokymosi metodai ir pritaikymas

2.4.1. Mokymasis su mokytoju

Mašininio mokymosi algoritmai taikomi daugelyje sričių, ir kiekvienoje srityje jie tobulinami. Vienoje srityje patobulintas mašininio mokymosi algoritmas gali būti sėkmingai panaudotas kitoje srityje. Mašininio mokymosi modelio treniravimas – procesas, kurio meto mokymosi algoritmui yra perduodami mokymosi duomenys. Mokymosi duomenys turi turėti „teisingą atsakymą“ kuris yra žinomas kaip tikslas. Mokymosi algoritmas ieško šablonų mokymosi duomenyse, kurie siejasi su tikslu (kurį modelis bando prognozuoti) ir perduoda duomenis mašininio mokymosi modeliui kai randa šablonus. Mašininio mokymosi modelį galima panaudoti naujiems duomenims, kuriuose nėra žinoma koreliacija tarp šablono ir tikslo taip siekant gauti atsakymą į rūpimą klausimą (ar nagrinėjamas laiškas yra brukalas?). Brukalo paieškos atveju mašininio mokymosi modeliui turėtų būti perduoti paprasti (sukurti žmogaus su tikslu informuoti) laiškai ir brukalai, juos pažymint kaip tokius. Modeliui apdorojus duomenis galima duoti naujus laiškus ir prognozuoti ar šie laiškai yra brukalai.

2.4.2. Mokymasis be mokytojo

Mokymasis be mokytojo yra pats lanksčiausias metodas, kadangi nenaudojami klasifikuoti duomenys, taip pat nėra skirtumo tarp mokymosi ir testavimo duomenų. Neprižiūrimas mokymas anomalijų paieškos atveju ieško anomalijų tikrindamas atstumus tarp taškų taip sprendžiant kas yra normalu ir kas yra anomalija.



5 pav. „Mašininio mokymosi metodai“

Paveikslėlyje pavaizduoti mašinio mokymosi metodai.

2.4.3. Mašininio mokymosi modeliai

Pagrindiniai mašininio modelio tipai yra binarinio klasifikavimo, daugiasluoksnio klasifikavimo ir regresijos modeliai. Kiekvienas iš modelių stengiasi prognozuoti skirtingus atsakymų tipus.

Binarinio klasifikavimo modelis

Mašininio mokymosi binarinio klasifikavimo modeliai numato vieną iš dvejų galimų atsakymų – „taip“ arba „ne“. Modelio treniravimui naudojamas logistinės regresijos algoritmas. Galimi modelio klausimai:

- Ar šis elektroninis laiškas yra brukalas?
- Ar vartotojas pirks šį produktą?
- Ar šis produktas yra mašina ar šampūnas?
- Ar atsiliepimas parašytas žmogaus ar roboto?

Daugiasluoksnio klasifikavimo modelis

Daugiasluoksnio klasifikavimo modelis numato vieną iš daugiau nei dvejų galimų atsakymų. Treniravimui naudojamas daugialypės logistinės regresijos algoritmas. Galimi modelio klausimai:

- Šis produktas yra knyga, spinta ar papuošalas?
- Ar šis filmas yra komedija, drama, kriminalas?
- Kuri programavimo kalba labiausiai patinka programuotojui?

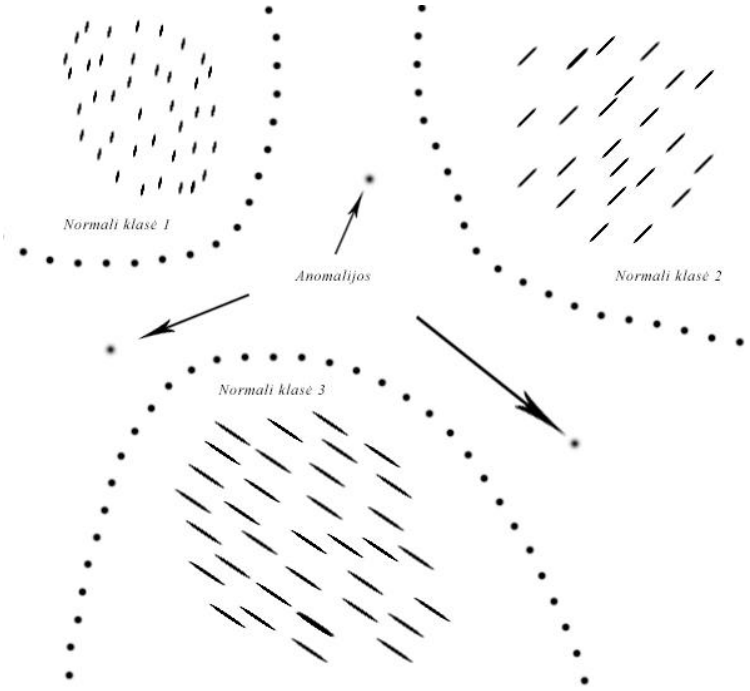
Regresinis klasifikavimo modelis

Regresinis modelis numato skaitinę reikšmę. Treniravimui naudojamas tiesinės regresijos algoritmas. Galimi modelio klausimai:

- Kokia rytoj bus temperatūra Vilniuje?
- Kiek bus parduota šios knygos kopijų?
- Už kiek bus parduotas šis butas?

Klasifikavimo metodai

Klasifikavimas yra naudojamas apmokyti modelį iš turimų sužymėtų duomenų taškų, kurie yra priskirti grupėms. Tuomet apmokytas modelis gavęs naujus duomenis gali juos suklasifikuoti į atskiras grupes. Atsižvelgiant į turimus duomenis testavimo fazėje, klasifikavimu paremta anomalijų paieška gali būti sugrūpuota į vienos ar kelių klasių paieškos kategorijas. Kelių klasių klasifikavimu paremtos technikos daro prielaidą, kad treniravimosi duomenys priklauso kelioms normalioms klasėms. Tokios technikos moko klasifikatorių atskirti duomenis tarp kiekvienos normalios ir likusių klasių.



6 pav. „Kelių klasių klasifikavimas“

Taškas pažymimas kaip anomalija jeigu jis nepriklauso nei vienai iš klasių.

Vienos klasės klasifikavimu paremti anomalijų paieškos algoritmai daro prielaidą, kad visi treniravimo duomenys turi tik vieną. Bet kuris taškas, kuris nėra priskiriamas klasei yra laikomas anomalija.

Neuroniniais tinklais paremti algoritmai taikomi tiek vienos tiek kelių klasių klasifikavime. Paprastai kelių klasių anomalijų paieška naudojantis neuroniniais tinklais veikia dviem žingsniais. Pirma neuroninis tinklas yra apmokomas su normaliais duomenimis tam, kad atskirtų keletą normalių klasių. Tuomet algoritmui yra duodami testuojamieji duomenys ir jeigu neuroninis tinklas priima duomenis, jie laikomi normaliais, jeigu atmeta – anomalijomis.

Toliau trumpai aptariami klasifikavimo metodai:

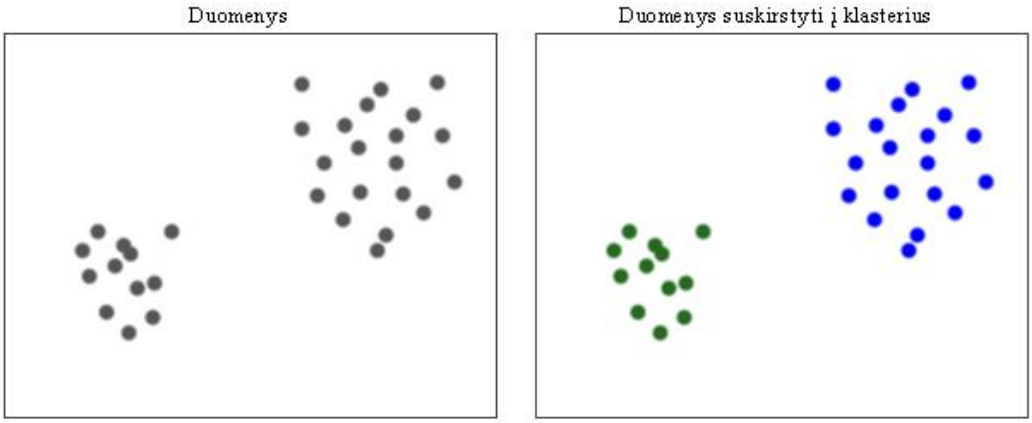
- Bajeso klasifikatorius (angl. *Naive Bayes classification*);
- k-artimiausių kaimynų (angl. *k-nearest neighbours*);
- sprendimų medis (angl. *decision tree*);
- daugiasluoksnis neuroninis tinklas (angl. *multilayer perceptron*);
- atraminių vektorių klasifikatorius (angl. *support vector machine*).

Naive Bajeso klasifikatorius remiasi Bajeso taisykle. Laikoma, kad visi duomenų požymiai yra nepriklausomi ir kiekvienas iš požymių daro įtaką klasifikavimo rezultatui. Klasifikatorius skaičiuoja tikimybes kiekvienai klasei. Objektas priskiriamas tai klasei, kuri įgyja didžiausią tikimybę. Sprendimų medžio algoritmo rezultatą galima pavaizduoti struktūra, panašia į medį, kurio kiekvienas išsišakojimas reiškia vienos ar kitos sąlygos tenkinimą [14]. Taip sudaromos taisyklės, kurios leidžia nagrinėjamą duomenų aibę suklasifikuoti atsižvelgiant į požymių savybes. K-artimiausių kaimynų metodo idėja yra naujo objekto palyginimas su mokymo aibės objektais, kurie yra panašūs į jį. Norint naują objektą priskirti kuriai nors klasei, skaičiuojami atstumai nuo to objekto iki visų mokymo aibės objektų. Dažniausiai naudojamas Euklido atstumas. Naujas objektas priskiriamas tai klasei, kuriai priklauso dauguma iš artimiausių k-jo kaimynų. Dirbtinio neuroninio tinklo struktūra primena biologinius neuroninius tinklus. Daugiasluoksnis neuroninis tinklas sudarytas iš kelių sluoksnių: įvesties, išvesties ir vieno ar daugiau paslėptų neuronų. Be kitų uždavinių, neuroniniai tinklai naudojami ir klasifikavimo uždaviniui spręsti. Tuomet įvesties sluoksnyje pateikiama požymius aprašanti informacija, o išvesties sluoksnyje gaunamas rezultatas – priklausymas klasėms. Atraminių vektorių klasifikatorius – algoritmas, kuris transformuoja pradinius duomenis į didesnę dimensiją, kur randama hiperplokštuma, skirianti dvi klases kiek galima

didesniu atstumu tarp klasifikuojamų duomenų. Radus šią hipreplokštumą, duomenis galima suskirstyti į dvi atskiras klases.

2.4.4. Klasterizavimas

Mašininiame mokymesi klasterizavimas yra svarbus žingsnis siekiant suskirstyti turimus duomenis į grupes – klasterius. Klasteris – panašių objektų grupė. Klasterinė analizė yra matematinių metodų visuma, kurios pagalba galima objektų arba reikšmių aibes pagal jų panašumus suskirstyti į tam tikras grupes (klasterius). Tikslas yra suskirstyti objektus taip, kad tos pačios grupės objektai būtų „artimi“ vienas kitam, kitaip tariant skirtumai būtų kuo mažesni, o objektai iš skirtingų grupių - „tolimi“, kitaip tariant siekiama suskirstyti duomenis tarp skirtingų klasterių taip, kad jie butų kuo labiau nutolę vieni nuo kitų. Skirstydami objektus į klasterius dažniausiai nežinome, kiek klasterių tiriamoje populiacijoje realiai egzistuoja. Kitaip tariant klasterinė analizė yra tam tikrų egzistuojančių struktūrų paieška. Šių algoritmų pagrindinis privalumas – gebėjimas atpažinti grupavimo struktūrą be jokios išankstinės informacijos. Duomenų suskirstymas yra naudingas siekiant skirstyti vartotojus į grupes pagal jų pomėgius ar interesus.



7 pav. „Duomenų klasterizavimas“

2.4.4.1. Klasterizavimo metodai

K-vidurkių metodas yra efektyvus, plačiai naudojamas vienas iš nehierarchinių klasterinės analizės metodų. Nehierarchiniai metodai paprastai taikomi tada, kai iš anksto žinomas (pasirenkamas) klasterių skaičius ir norima klasterizuoti tiriamus objektus. Prieš naudojant algoritmą yra apsibrėžiama atsumo tarp duomenų taškų skaičiavimo funkcija, pavyzdžiui Euklido atstumo funkcija, ir nustatatomas klasterių kiekis (k). Algoritmas parenka „k“ taškų kurie yra laikomi klasterių centrais. Pasirinkus atstumo funkciją ir klasterių centrus algoritmas atlieka šiuos žingsnius:

1. Priskyrimas: kiekvienas iš „m“ taškų esančių duomenų erdvėje yra priskiriamas tam klasteriui, prie kurio centro taškas yra arčiausiai;
2. Atnaujinimas: kiekvienam klasteriui yra parenkamas naujas centras apskaičiuojant vidutinį atstumą tarp visų taškų priskirtų šiam klasteriui.

Po kiekvienos algoritmo iteracijos klasterių centrai juda arčiau klasterių centrų. Algoritmo žingsniai iteruojami tol kol įvyksta konvergencija – klasterių centriniai taškai nepajuda. Finaliniai klasifikavimo rezultatai smarkiai priklauso nuo pradinių klasterio centrų parinkimo, todėl yra svarbu skirti pakankamai dėmesio šiai problemai. Vienas iš sprendimo variantų yra algoritmą leisti keletą kartų taip kiekvienąkart parenkant skirtingus klasterių centrinius taškus. Tuomet galima pasirinkti geriausią rezultatą apskaičiuojant duomenų taškų atstumų sumą iki klasterio centrinių taškų. Galima alternatyva yra parinkti pradinius taškus kurie yra nutolę. Galima tikėtis geresnių rezultatų, tačiau renkantis nutolusius taškus yra tikimybė parinkti taškus kurie yra pernelyg nutolę, tad galimai nėra tinkami klasterio centriniams taškams.

Lūkesčių maksimizavimo (angl. *Expectation Maximization*) klasterizavimo metodas.

„K-vidurkių“ algoritme kiekvienas taškas yra priskiriamas tik vienam klasteriui. Tai nėra labai lankstus algoritmas, kadangi negalime teisingai priskirti taškui klasterio, jei klasteriai persidengia arba jei klasteriai nėra apskritimo formos. Lūkesčių maksimizavimo (angl. *expectation maximization,* toliau EM) algoritmas yra pažangesnis, kadangi yra parenkamas klasterio centrinis taškas, nustatoma kovariacija (suteikianti galimybę klasteriui tapti elipsės formos) ir suteikiamas svoris įtakojantis klasterio dydį. Tikimybė, kad taškas priklauso klasteriui priskiriama naudojantis Gauso tikimybių paskirstymo metodu. Algoritmo procedūra yra panaši į „k-vidurkių“ – taškams paskaičiuojama tikimybė priklausyti klasteriui. Procedūra vadinama „E“ (lūkesčio) žingsniu. Jei kažkuris klasteris turi didelę tikimybę taško priskyrimui, dažniausiai taškas yra netoli klasterio centro, tačiau keli klasteriai gali turėti tokią pačią tikimybę taško priskyrimui. Ši algoritmo savybė, kad taškai gali priklausyti keliems klasteriams vadinama „lanksčiu“ klasterizavimu. Žingsnis „M“ perskaičiuoja kiekvieno klasterio parametrus naudodamas taškų pasiskirstymą iš ankstesnio klasterių rinkinio. Norint apskaičiuoti naują vidurkį, kovariaciją ir klasterio svorį, kiekvieno taško duomenys yra apskaičiuojami pagal jo tikimybę priklausyti klasteriui, taip pat kaip skaičiuota praeitame etape. Kaip ir „k-vidurkių“ algoritme – naudojant algoritmą tiems patiems duomenis keletą kartų galima pagerinti taškų priskyrimą į klasterius. Norint parinkti tik vieną klasterį taškas

priskiriamas klasteriui pagal didžiausią tikimybę jam priklausyti. Turint tikimybių modelį galima sukurti papildomų taškų kurie bus panašūs į jau atrinktus taškus.

Hierarchinių klasterizavimo metodų rezultatai nusako klasterių tarpusavio hierarchiją, t. y. visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos – dar mažesni ir t. t. Taikant šiuos metodus, nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik paskui sprendžiama, koks klasterių skaičius optimalus. Hierarchinis jungimo metodas smulkius klasterius jungia į stambesnius, kol galų gale lieka vienas.

2.5. Antro skyriaus apibendrinimas ir pagrindiniai rezultatai

Susijusios literatūros analizėje apžvelgta realaus laiko reklamos pirkimo sistema, vartotojo elgsenos sekimas, slapukų sinchronizacija. Vartotojas, naršantis internete gali ir yra sekamas įvairių duomenų rinkimo platformų. Ši informacija dažnai yra naudojama siekiant didinti reklamos kampanijų efektyvumą, tačiau informaciją galima panaudoti ir tokiems tikslams kaip potencialiai kenksmingo vartotojo atpažinimas.

Vartotojo profilio informacijos struktūros analizėje išsiaiškinta, kad egzistuoja įvairūs skirtingi metodai identifikuoti vartotojo informacijai, ją apdoroti iš kontekstinės informacijos (lankomų puslapių istorija, jų kategorijos, temos). Anonimiškumo atskleidimo metodai leidžia susieti iš pažiūros skirtingus vartotojus į vieną vartotoją, kadangi vartotojo atlikti veiksmai palieka skaitmeninį pėdsaką.

Išnagrinėti metodai potencialiai kenksmigų vartotojų aptikimui pasižymi bendra savybe – mašininio mokymosi taikymo siekiant sumažinti rankinį darbą ir didinti potencialiai kenksmingų vartotojų aptikimą internete.

Išnagrinėti mašinio mokymosi metodai siekiant išsiaiškinti koks mašininio tipo mokymasis yra tinkamas iškelto uždavinio sprendimui.

2.6. Antro skyriaus išvados

Realaus laiko reklamo pirkimo sistemos gali surinkti pakankamai informacijos, kad galėtų identifikuoti vartotoją skirtinguose reklamos tinkluose, mobiliuosiuose įrenginiuose. Išnaudojus slapukų

sinchronizavimo galimybes galima sekti vartotojo veiksmus internete net ir nerodant jam reklamos. Šis metodas yra tinkamas informacijos rinkimo būdas siekiant identifikuoti potencialiai kenksmingą vartotoją.

Išnagrinėjus vartotojo profilio struktūros gavimą paaiškėjo, kad galima identifikuoti vartotojo interesus, demografinę informaciją ir identifikuoti tą patį vartotoją skirtinguose įrenginiuose, todėl vienareikšmiškai išlikti anonimišku yra pakankamai sudėtinga.

Išnagrinėjus metodus potencialiai kenksmingų vartotojų aptikimui paaiškėjo, kad mašinino mokymosi metodai yra pakankamai efektyvūs identifikuoti tokius vartotojus ir sumažina rankinį darbo poreikį.

Išnagrinėjus mašinio mokymosi metodus nustatyta, kad potencialiai kenksmingų vartotojų aptikimui labiausiai tinkamas mašininio mokymosi tipas yra mokymasis su mokytoju.

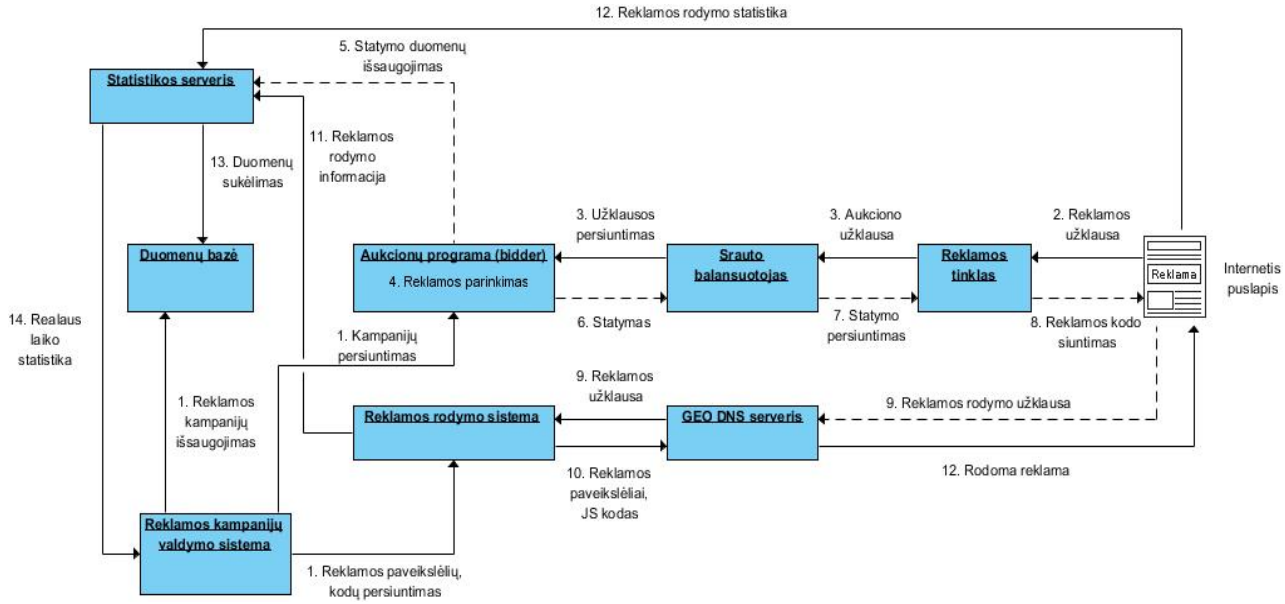
3. Siūlomo būdo arba metodo aprašymas

Darbe nagrinėjama galimybė identifikuoti potencialiai kenksmingus vartotojus realaus laiko reklamos pirkimo sistemoje. Išanalizavus metodus potencialiai kenksmingų vartotojų aptikimui nuspręsta naudoti mašinio mokymosi metodus siekiant identifkuoti tokius vartotojus realaus laiko reklamos pirkimo sistemoje. Šiame skyriuje nagrinėjama egzistuojanti tokios sistemos architektūra ir jos praplėtimai tikslui pasiekti.

Siekiant identifikuoti vartotojus įvairiuose reklamos tinkluose ir puslapiuose papildyta egzistuojanti sistema, pradėta rinkti vartotojų profilio informacija, kuri naudojama identifikuoti kitiems potencialiai kenksmingiems vartotojams.

3.1. Egzistuojanti architektūra

Šiame skyriuje apžvelgiama egzistuojanti realaus laiko reklamos pirkimo aukciono sistemos architektūra.



8 pav. „Egzistuojanti sistemos architektūra“

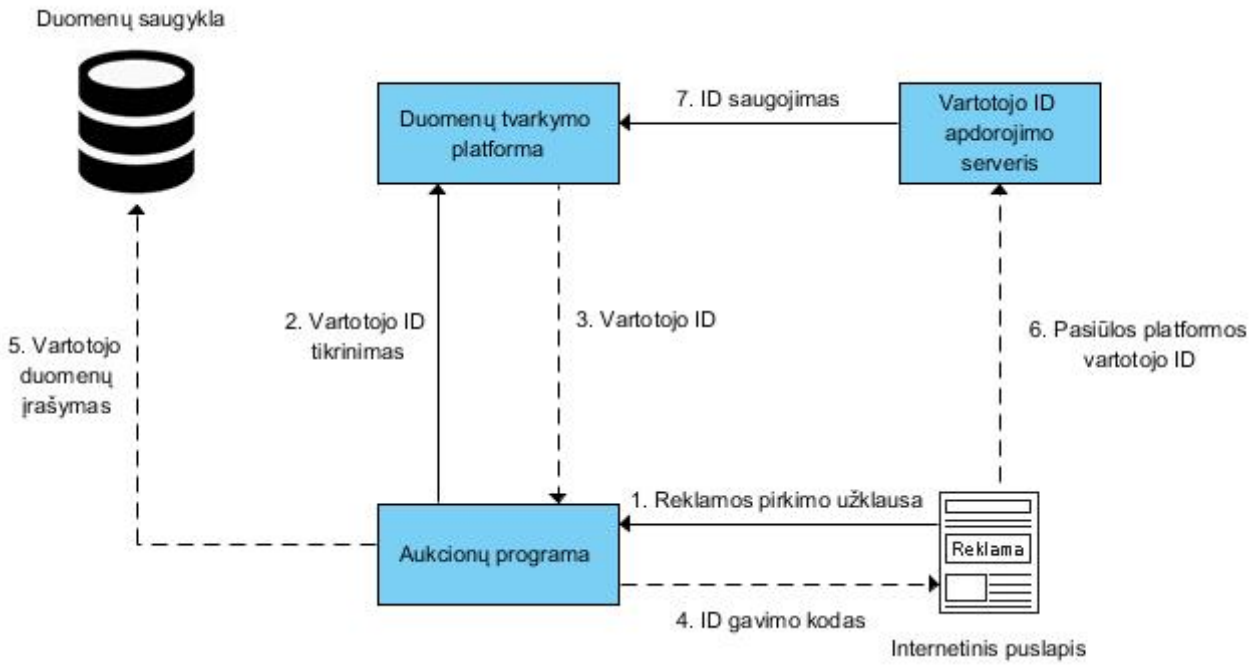
Diagramoje pavaizduota sistemos architektūra ir realaus laiko reklamos pirkimo sistemos sujungimas su ja:

1. Paklausos pusės platformoje (DSP) kuriamos reklamos su norimu vartotojų filtru (demografinė informacija, interesai, vartotojų segmentai), taiomų puslapių parinkimu. Įkeliama reklamos media – paveikslėliai, video failai, javascript failai. Sistema išsaugo duomenis duomenų bazėje, persiunčia informaciją aukcionų programai (bidder). Reklamos failai persiunčiami reklamos rodymo sistemai, kuri turi keletą serverių skirtinguose pasaulio regionuose;
2. Vartotojas naršantis internete užsuka į sveiainę, kuri turi reklaminio ploto ir kodą, sujungtą su pasiūlos pusės platforma (SSP). Vartotojo naršyklei nusiuntus užklausą krauti internetinį puslapį, kartu ir reklamos kodą, pasiūlos pusės platforma sukuria aukcioną ir siunčia informaciją apie puslapį bei vartotoją paklausos pusės platformoms;
3. Reklamos aukciono užklausa patenka į srauto balansuotoją, kuris persiunčia užklausą vienai iš aukcionų programų (bidder);
4. Pagal sukonfigūruotas kampanijas reklamos aukciono programa atlieka tikrinimą ar vartotojas ir puslapis atitinka kampaniją. Atitikimo atveju sistema nusprendžia kurią iš atitikusių kampanijų rodyti ir paruošia statymo užklausą;
5. Reklamos statymo sistema sugeneruoja statymo statistiką (įvyko statymas kampanijai Z puslapyje K vartotojui E) ir ją persiunčia statistikos serveriui;
6. Paruošta statymo užklausa siunčima srauto balansuotojui;
7. Srauto balansuotojas persiunčia reklamos statymo užklausą reklamos tinklui;
8. Po įvykusio aukciono, jeigu aukciono laimėtojas yra sistemos siųsta kampanija, reklamos tinklas siunčia HTML kodą į internetiniame puslapyje esantį reklamos plotą;
9. Reklamos plotas pradeda krauti gautą HTML kodą. Naršyklė siunčia reklamos paveikslėlių/video/javascript krovimo užklausą į GEO DNS serverį. GEO DNS serveris parenka arčiausiai esantį reklamos serverį ir persiunčia užklausą;
10. Reklamos rodymo sistema randa prašomas bylas (paveikslėlius/video/javascript) ir siunčia jas GEO DNS serveriui;
11. Reklamos rodymo sistema įrašo statistiką apie rodomą reklama statistikos serveryje (aukciono laimėjimas, sumokėta kaina už reklamą);
12. Reklama užkraunama vartotojo naršyklės ir yra matoma puslapyje. Reklama turi kodą, kuris persiunčia informaciją statistikos serveriui apie reklamos parodymo, bei jos matomumo faktą.

Egzistuojanti sistema yra veikianti paklausos pusės platforma, tačiau norint rinkti informaciją apie varotojo naršymo istoriją, nėra pakankamai išpildyta. Siekiant rinkti vartotojo naršymo istoriją reikalinga galimybė ją kaupti. Istorijos kaupimas paruošiant kampaniją, kuri pirktų visą interneto vartotojų generuojamą srautą būtų labai brangus, todėl nuspręsta pasinaudoti vartotojo elgsenos sekimo metodais ir paruošti duomenų tvarkymo platformą, kuri galėtų rinkti varotojų naršymo istoriją be tiesioginių reklamos parodymų.

3.2. Vartotojo elgsenos sekimo modelis

Siekiant sekti vartotojo elgseną be tiesioginių reklamos pirkimų reikalinga galimybė saugoti duomenis apie vartotojus ir jų identikiatorius pasiūlos pusės platformoje. Kuriama duomenų tvarkymo platforma saugoti duomenims apie vartotoją ir jo identifikatorius skirtinguose pasiūlos pusės platformose.



9 pav. „Vartotojo elgsenos sekimo modelis“

Diagramoje parodytas vartotojo elgsenos sekimo modelis:

1. Naršydamas internete vartotojas sugeneruoja reklamos užklausą, kuri patenka į pasiūlos pusės platformą ir iš jos – į aukcionų programą;
2. Aukcionų programa aukciono užklausoje gauna reklamos kontekstinę informaciją, kartu ir pasiūlos platformos vartotojo ID. Aukcionų programa kreipiasi į duomenų tvarkymo

platformą siekdama išsiaiškinti ar pasiūlos pusės platformos (SSP) siunčiamas vartotojo ID yra žinomas duomenų tvarkymo platformoje (DMP);

3. Duomenų tvarkymo platforma ieško pasiūlos pusės platformos atsiųsto vartotojo ID atitikties su duomenų tvarkymo platformos vartotojo identifikatoriumi (ID). Jei ID neegzistuoja – sistema sugeneruoja atsitiktinį ID ir jį persiunčia aukcionų programai. Jeigu ID egzistuoja, duomenų tvarkymo platforma tikrina kuriuose pasiūlos pusės platformose žinomi vartotojo ID. Jei randama nežinomų pasiūlos pusės platformų identikatorių, jų adresai siunčiami aukcionų programai;
4. Aukcionų programa, gavusį ID susiejimo adresus juos prideda į reklaminį kodą ir siunčia pasiūlos puės platformai, kuri, aukciono laimėjimo atveju, persiunčia reklaminį kodą kartu su ID susiejimo nematomais paveikslėliais į vartotojo naršyklę;
5. Aukciono užklausos metu gautus duomenis aukciono programa siunčia į duomenų saugyklą vėliasniam apdorojimui;
6. Reklamos užkrovimo metu, vyksta HTTP nukreipimas į pasiūlos pusės platformą, kuri prideda savo vartotojo ID ir daro HTTP užklausą nurodytu adresu (į vartotojo ID apdorojimo serverį);
7. Vartotojo ID apdorojimo serveris užsaugo šią informaciją duomenų tvarkymo platformoje.

Toks vartotojo elgsenos sekimo modelis sprendžia kelias problemas: nebūtina pirkti reklamos, jei duomenų tvarkymo platformoje žinomas vartotojo ID, todėl aukciono užklausos metu kontekstinę informaciją galima išsaugoti į duomenų saugyklą, kuri vėliau agreguos duomenis apie vartotoją ir taip papildys informaciją apie vartotoją. Toks informacijos pildymas yra tarsi vartotojo istorija laike, kadangi iš išsaugotos informacijos laiko galima stebėti kaip keičiasi vartotojo profilis bėgant laikui. Kita svarbi sprendžiama problema yra vartotojo ID pildymas skirtingose pasiūlos pusės platformosse. Prie aukcionų sistemos prijungus daugiau paklausos pusės platformų, duomenų tvarkymo platforma neranda naujai prijungtų pasiūlos pusės platformos vartotojų ID, todėl pirkdama reklamas prisega nematomų paveikslėlių adresus ir taip sužinomi nauji vartotojo ID kitose sistemose. Šios problemos sprendimas padeda gerinti duomenų kokybę ir susieja duomenis su konkrečiu duomenų tvarkymo platformos vartotoju.

Vartotojo ID ir pasiūlos pusės platformų ID saugojimo lentelių struktūra:

Pasiūlos pusės platformos ID	Pasiūlos pusės platformos vartotojo ID	Duomenų tvarkymo platformos vartotojo ID
1	CAESEIcS1pC2TBvb-4SLDjMqsY9	7d588315-b75c-46a3-9de9-dd52b14a6d8d
2	5e29eb00-c30a-416e-9d2a-2e18901f0916	e75c3d68-9b88-4962-81d7-0175a57b090d
N	....	....
N+1	3485E719-C68E-495B-945E-C89D3DF4287D	7d588315-b75c-46a3-9de9-dd52b14a6d8d

1 lentelė „Vartotojo ID ir pasiūlos pusės platformų ID saugojimo lentelė“

Lentelėje pavaizduota vartotojo ID ir pasiūlos pusės plarformų ID saugojimo struktūra: kiekviena prie aukcionų programos pajungta pasiūlos pusės platforma turi unikalų identifikatorių. Duomenų tvarkymo platformos vartotojo ID yra sugeneruojamas aukcionų pirkimo sistemos naudojantis universaliu unikaliu identifikatoriumi (UUID), todėl yra unikalus kiekvienam vartotojui (kadangi pririštas prie generavimo laiko [17]). Pasiūlos pusės platformos vartotojo ID yra gaunamas reklamos aukciono metu. Toks duomenų talpinimo principas leidžia lengvai rasti vartotojo ID pagal pasiūlos pusės platformos siunčiamą vartotojo ID, tačiau kelia problemų norint žinoti kokiuose pasiūlos pusės platformose vartotojas yra identifikuotas.

Siekiant užtikrinti greitą duomenų patikrinimą ir perdavimą aukcionų programai reikalinga papildoma lentelė, kurioje saugoma duomenų tvarkymo platformos ir atpažintų pasiūlos pusės platformų ID informacija.

Duomenų tvarkymo platformos vartotojo ID	Pasiūlos pusės platformų žinomi ID
7d588315-b75c-46a3-9de9-dd52b14a6d8d	2
....	....
e75c3d68-9b88-4962-81d7-0175a57b090d	78

2 lentelė „Duomenų tvarkymo platformos ir pasiūlos pusės identifikuotų ID saugojimo struktūra“

Lentelėje saugomas duomenų tvarkymo platformos vartotojo ID ir identifikuotų pasiūlos pusės platformų ID. Identifikuotų pasiūlos pusės platformų ID saugomi sveikojo skaičiaus formatu. Sveikasis

skaičius gaunamas pasiūlos pusės platformos ID pakeliant kvadratu ir pridedant kitus pasiūlos pusės platformos ID pakeltus kvadratu, pvz.:

- Vartotojo ID identifikuotas pasiūlos platformose 1, 2, 4, 7. Žinomų ID sveikasis skaičius apskaičiuojamas: $2^1 + 2^2 + 2^4 + 2^7 = 2 + 4 + 16 + 128 = 150$.

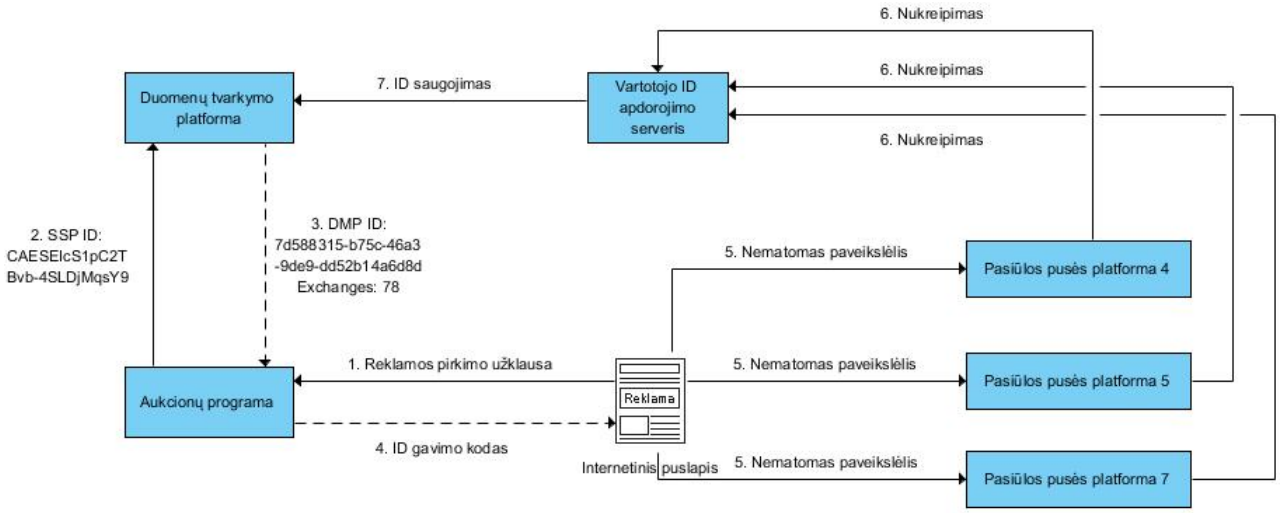
Toks metodas tinkamas todėl, kad yra efektyvus užimamos atminties atžvilgiu ir aukcionų programa, naudojantis dvejetaine aritmetika ir matematika gali greitai patikrinti kurių pasiūlos pusės platformų vartotojų ID nėra žinomi, pvz.: gautas pasiūlos pusės platformų ID sveikasis skaičius 78 (dvejetainėje sistemoje – 110010), aukcionų pirkimo sistema tikrina ar vartotojo ID atpažintas pasiūlos pusės platformose 1 – 7 skaičiavimas atliekamas naudojantis dvejetainiu skaičiavimu „ir“:

Tikrinamas pasiūlos pusės vartotojo ID egzistavimas skaičiuje	Tikrinimo sąlyga	Skaičiavimas	Skaičiavimo rezultatas	Vartotojo ID atitikmuo pasiūlos pusės platformoje yra žinomas
1	$78 \,\&\, 2^1 = 2$	$\& \frac{1001110}{0010}$	0010 = 0010	Taip
2	$78 \,\&\, 2^2 = 4$	$\& \frac{1001110}{0100}$	0100 = 0100	Taip
3	$78 \,\&\, 2^3 = 8$	$\& \frac{1001110}{1000}$	1000 = 1000	Taip
4	$78 \,\&\, 2^4 = 16$	$\& \frac{1001110}{0010000}$	0000 0000 ≠ 0010 000	Ne
5	$78 \,\&\, 2^5 = 32$	$\& \frac{1001110}{0100000}$	0000 0000 ≠ 0100 000	Ne
6	$78 \,\&\, 2^6 = 64$	$\& \frac{1001110}{1000000}$	1000 0000 = 1000 000	Taip
7	$78 \,\&\, 2^7 = 128$	$\& \frac{1001110}{100000000}$	0 0000 0000 ≠ 1 0000 0000	Ne

3 lentelė „Vartotojo ID atitikmens pasiūlos pusės platformoje tikrinimas“

Pateiktoje lentelėje patikrinta ar vartotojo pasiūlos pusės platformos ID yra žinomi duomenų tvarkymo platformoje. Šiuo atveju aukcionų sistema siųstu nematomus paveikslėlius atpažinti vartotojo ID 4, 5 ir 7 pasiūlos pusės platformose.

Vartotojo ID ir pasiūlos pusės platformų ID susiejimas detaliau parodytas diagramoje:



10 pav. „Vartotojo ID ir pasiūlos pusės platformų ID susiejimas“

Vartotojo ID pasiūlos pusės platformoje susiejimas su duomenų tvarkymo platformos žingsniai:

1. Gaunama aukciono užklausa iš pasiūlos pusės platformos (kuri gauna užklausą iš naršančio vartotojo);
2. Aukcionų programa kreipiasi į duomenų tvarkymo platformą su pasiūlos pusės vartotojo ID (CAESEIcS1pC2TBvb-4SLDjMqsY9);
3. Duomenų tvarkymo platforma randa pasiūlos pusės platformos atitikmenį ir atpažintų pasiūlos pusės platformų skaičių (DMP ID: 7d588315-b75c-46a3-9de9-dd52b14a6d8d Exchanges: 78);
4. Aukcionų programa tikrina kuriuose pasiūlos pusės platformose vartotojo ID nėra žinomas (pagal lentelėje 3 pateiktus duomenis nežinomį paslaugos pusės platformų vartotojo ID yra platformose 4, 5 ir 7) ir sugeneruoja nematomų paveikslėlių kodus:

```
<img src="//ssp-
4.lt/pixel?dest=https%3A%2F%2Fdmp.lt%2Fpixel%3FdmpId%3D7d588315-b75c-
46a3-9de9-dd52b14a6d8d%26sspId%3D4%26userId%3D" style="display:none">


<img src="//ssp-
5.lt/pixel?dest=https%3A%2F%2Fdmp.lt%2Fpixel%3FdmpId%3D7d588315-b75c-
46a3-9de9-dd52b14a6d8d%26sspId%3D5%26userId%3D" style="display:none">


<img src="//ssp-
7.lt/pixel?dest=https%3A%2F%2Fdmp.lt%2Fpixel%3FdmpId%3D7d588315-b75c-
46a3-9de9-dd52b14a6d8d%26sspId%3D7%26userId%3D" style="display:none">
```

5. Pasiūlos pusės platforma, gavusi paveikslėlio užklausą prideda savo vartotojo ID prie HTTP GET „dest: parametro ir, naudojantis HTTP antraštėmis, atlieka nukreipimą į reikšmę, perduotą „dest“ parametre:

   SSP 4:

   ```
   HTTP/1.1 302 Found
   Location: https://dmp.lt/pixel?dmpId=7d588315-b75c-46a3-9de9-
   dd52b14a6d8d&sspId=4&userId=CAESEO561pC4EVa-uyu76UJH08
   ```

   SSP 5:

   ```
   HTTP/1.1 302 Found
   Location: https://dmp.lt/pixel?dmpId=7d588315-b75c-46a3-9de9-
   dd52b14a6d8d&sspId=5&userId= 765bc858-4fe9-4470-852f-e2f4e1259fb5
   ```

   SSP 7:

   ```
   HTTP/1.1 302 Found
   Location: https://dmp.lt/pixel?dmpId=7d588315-b75c-46a3-9de9-
   dd52b14a6d8d&sspId=7&userId= 5eee3313-8b8e-4eed-a631-fcefa5c77eeb
   ```

6. Vartotojų ID apdorojimo serveris (https://dmp.lt/pixel) gauna HTTP užklausas iš pasiūlos pusės platformų ir atsako su permatomu GIF paveikslėliu;
7. Vartotojų ID apdorojimo serveris išsaugo gautas reikšmes duomenų varkymo platformoje ir atnaujina atpažintų vartotojo ID skaitinę reikšmę (Exchanges = Exchanges | 2^4 | 2^5 | 2^7).

3.3. Vartotojo profilio sudarymas

Reklamos aukciono siunčiama užklausa turi kontekstinę informaciją apie lankomą puslapį ir vartotoją. Pasinaudojus siunčiama informacija galima rinkti duomenis apie vartotoją iš skirtingų pasiūlos pusės platformų. Vienos platformos siunčia daugiau informacijos, kitos – mažiau, todėl reikalinga galimybė kaupti visus gaunamus duomenis apie vartotoją, juos kaupti, apjungti ir taip sudaryti naujausią vartotojo profilį. Kadangi ne visa siunčiama informacija yra aktuali, bus naudojama informacija iš „OpenRTB BidRequest“ labiausiai susijusi su vartotoju.

Pagal OpenRTB API specifikaciją, aukciono užklausa (angl. *bid request*) yra siunčiama JSON formatu. Siekiant atrinkti tik aktualius saugojimui duomenis apžvelgiama OpenRTB API speficikacija.

3.3.1. Aukciono užklausos objektas

Lentelėje aprašomas aukciono užklausos objektas ir jo sudedamosios dalys.

Atributas	**Tipas**	**Aprašymas**
id	tekstas; privalomas	Unikalus aukciono užklausos ID
imp	objektų masyvas; privalomas	„Imp“ objektų masyvas reprezentuojantis galimus reklamos parodymus
site	objektas; rekomenduojamas	„Site“ objekto informacija apie puslapį kur bus rodoma reklama. Siunčiama jei reklama rodoma pulapyje
app	objektas; rekomenduojamas	„App“ objekto informacija apie mobiliąją programėlę. Siunčiama jei reklama rodoma mobiliojoje programėlėje
device	objektas; rekomenduojamas	„Device“ objektas. Informacija apie naršantį įrenginį
user	objektas; rekomenduojamas	„User“ objektas. Informacija apie naršantį vartotoją

test	skaičius; 0 pagal nutylėjimą	Indikuoja testinį režimą
at	skaičius; 2 pagal nutylėjimą	Aukciono tipas. 1 = Pirmos kainos aukcionas, 2 = antros kainos aukcionas
tmax	skaičius	Maksimalus laikas milisekundėmis per kurį turi būti pristatytas reklamos pirkimo statymas
wseat	teksto masyvas	Agentūrų ir reklamuotojų leidžiamas sąrašas. Gali būti siunčiamas tik vienas iš „wseat“ ar „bseat“. Jei neatsiųstas nei vienas, laikoma, kad nėra reklamos pirkimo apribojimų
bseat	teksto masyvas	Agentūrų ir reklamuotojų draudžiamas sąrašas. Gali būti siunčiamas tik vienas iš „wseat“ ar „bseat“. Jei neatsiųstas nei vienas, laikoma, kad nėra reklamos pirkimo apribojimų
allimps	skaičius; 0 pagal nutylėjimą	Indikuoja ar visos siunčiamos reklaminės pozicijos yra rodomos
cur	teksto masyvas	ISO-4217 valiutos kodas
wlang	teksto masyvas	Reklamos turinio kalbos ISO-639-1-alpha-2 kodas
bcat	teksto masyvas	Uždraustos „IAB“ kategorijos
badv	teksto masyvas	Uždrausti reklamuotis internetiniai adresai (pvz.: „ford.com“)
bapp	teksto masyvas	Uždraustos reklamuoti mobiliosios programėlės (pvz.: „com.foo.mygame“)
source	objektas	„Source“ objekto informacija. Reklamos šaltinio informacija
regs	objektas	„Regs“ objektas. Informacija apie reklamos reguliavimą
ext	objektas	Vieta „OpenRTB“ praplėtimui

4 lentelė „OpenRTB API specifikacija“

Iš pateiktos lentelėje informacijos naudosime tik tą informaciją, kuri yra susijusi su konkrečiu, naršančiu vartotoju, todėl detalizuojami tik šie objektai: „Site“, „App“, „Device“, „Geo“, „User“ ir „IAB“ turinio kategorijos.

3.3.2. „Site“ objektas

Lentelėje aprošomas „site“ objektas, kuris yra siunčiamas jei reklama bus rodoma internetiniame puslapyje.

Atributas	**Tipas**	**Aprašymas**
id	tekstų masyvas	Pasiūlos pusės platformos puslapio ID
name	tekstas	Puslapio pavadinimas
domain	tekstas	Puslapio adresas (pavyzdys.lt)
cat	tekstų masyvas	„IAB“ puslapio kategorijų masyvas
sectioncat	tekstų masyvas	„IAB“ kategorijų masyvas aprašantis konkrečios puslapio vietos kategorijas
pagecat	tekstų masyvas	„IAB“ kategorijų masyvas aprašantis konkretaus puslapio kategorijas
page	tekstas	Naršomo puslapio internetinis adresas (URL)
ref	tekstas	Praeito puslapio internetinis adresas (URL)
search	tekstas	Paieškos frazė, per kurią vartotojas pateko į puslapį
mobile	skaičius	Indikuoja ar puslapis yra paruoštas prieinamumui per mobilųjį įrenginį
privacypolicy	skaičius	Indikuoja ar puslapis turi privatumo politiką
publisher	objektas	Informacija apie reklamuotoją
content	objektas	Informacija apie puslapio turinį
keywords	tekstas	Puslapiui prisikirtas, kableliais atskirtas, „IAB“ kategorijų sąrašas
ext	objektas	Vieta „OpenRTB“ praplėtimui

5 lentelė „Site objektas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus naudojami laukai: „domain“, „cat“, „ref“, „search“, „keywords“.

3.3.3. „App“ objektas

Lentelėje aprašoma objekto „App“ struktūra ir turinys.

Atributas	**Tipas**	**Aprašymas**
id	tekstas	Pasiūlos pusės platformos mobiliosios programėlės ID
name	tekstas	Pasiūlos pusės platformos mobiliosios programėlės pavadinimas
bundle	tekstas	Pasiūlos pusės platformos mobiliosios programėlės programinis pavadinimas (com.foo.mygam)
domain	tekstas	Mobiliosios programėlės internetinis adresas (mygame.foo.com)
storeurl	tekstas	Mobiliųjų programėlių parduotuvės (App store) adresas
cat	teksto masyvas	„IAB“ mobiliosios programėlės kategorijų masyvas
sectioncat	teksto masyvas	„IAB“ mobiliosios programėlės sekcijos kategorijų masyvas
pagecat	teksto masyvas	„IAB“ mobiliosios programėlės vaizdo kategorijų masyvas
ver	tekstas	Mobiliosios programėlės versija
privacypolicy	skaičius	Indikuoja ar mobilioji programėlė turi privatumo politiką
paid	skaičius	Indikuoja ar mobilioji programėlė yra mokama
publisher	objektas	Informacija apie reklamuotoją
content	objektas	Informacija apie programėlės turinį
keywords	tekstas	Mobiliajai programėlei prisikirtas, kableliais atskirtas, „IAB“ kategorijų sąrašas
ext	objektas	Vieta „OpenRTB“ praplėtimui

6 lentelė „App objekto aprašymas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus naudojami laukai: „bundle“, „cat“, „keywords“.

3.3.4. „Device“ objektas

Lentelėje aprašoma objekto „Device“ struktūra ir turinys.

Atributas	**Tipas**	**Aprašymas**
ua	tekstas	Naršyklės pavadinimas (angl. *user agent*)
geo	objektas	„Geo“ objektas
dnt	skaičius	Indikacija apie vartotojo sekimą (1 – sekimas draudžiamas)
lmt	skaičius	Indikacija apie reklaminio kodo sekimo draudimus (1 – sekimas draudžiamas)
ip	tekstas	IP adresas esantis arčiausiai įrenginio
ipv6	tekstas	IPv6 adresas esantis arčiausiai įrenginio
devicetype	skaičius	Įrenginio tipas
make	tekstas	Įrenginio gamintojas (pvz.: „Apple“)
model	tekstas	Įrenginio modelis (pvz.: „iPhone“)
os	tekstas	Įrenginio operacinė sistema (pvz.: „iOS“)
osv	tekstas	Įrenginio operacinės sistemos versija (pvz.: „.3.1.2“)
hwv	tekstas	Įrenginio geležies versija (pvz.: „5S“, jei įrenginys yra „iPhone 5S“)
h	skaičius	Ekrano aukštis pikseliais
w	skaičius	Ekrano plotis pikseliais
ppi	skaičius	Ekrano dydis išreikštas pikseliais per colį
pxratio	skaičius	Fizinių pikselių ir įrenginio pikselių santykis
js	skaičius	Indikuoja JavaScript palaikymą

geofetch	skaičius	Indikuoja ar geolokacijos API bus prieinamas iš JavaScript pusės reklamoje
flashver	tekstas	Naršyklės palaikoma „Flash“ versija
language	tekstas	Naršyklės kalbos kodas pagal ISO-639-1-alpha-2 kodą
carrier	tekstas	Paslaugų tiekėjas (pvz.: „Tele2“)
mccmc	tekstas	Mobiliojo tiekėjo „MCC-MNC“ kodas
connectiontype	skaičius	Tinklo tipas
ifa	tekstas	Įrenginio ID
didsha1	tekstas	Įrenginio IMEI kodas užšifruotas SHA1 algoritmu
didmd5	tekstas	Įrenginio IMEI kodas užšifruotas MD5 algoritmu
dpidsha1	tekstas	Įrenginio platformos ID (pvz.: Android ID) užšifruotas SHA1 algoritmu
dpidmd5	tekstas	Įrenginio platformos ID (pvz.: Android ID) užšifruotas MD5 algoritmu
macsha1	tekstas	Įrenginio „MAC“ adresas užšifruotas SHA1 algoritmu
macmd5	tekstas	Įrenginio „MAC“ adresas užšifruotas MD5 algoritmu
ext	objektas	Vieta „OpenRTB“ praplėtimui

7 lentelė „Device objekto aprašymas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus naudojami laukai: „geo“, „ifa“, „make“, „model“, „ua“.

3.3.5. „Geo“ objektas

Lentelėje aprašoma objekto „Geo“ struktūra ir turinys.

Atributas	**Tipas**	**Aprašymas**
lat	skaičius	Platuma nuo -90.0 iki +90.0, kur neigiama – pietūs
lon	skaičius	Ilguma nuo -180.0 iki +180.0, kur neigiama – vakarai

type	skaičius	Geografinių duomenų šaltinis
accurancy	skaičius	Apytikslis koordinatės tikslumas metrais
lastfix	skaičius	Sekundžių kiekis nuo koordinatės gavimo
ipservice	skaičius	Pslaugos tiekėjas naudotas gauti GEO informaciją iš IP adreso
country	tekstas	Šalies kodas pagal ISO-2166-1-alpha-3 kodą
region	tekstas	Regiono kodas pagal ISO-3166-2 standartą; dviraidis valstijos kodas jeigu įrenginys yra JAV
regionfips104	tekstas	Šalies regionas pagal FIPS 10-4 žymėjimą
metro	tekstas	„Google“ metro kodas
city	tekstas	Miestas
zip	tekstas	Pašto indeksas
utcoffset	skaičius	Laiko pokytis nuo UTC laiko juostos
ext	objektas	Vieta „OpenRTB“ praplėtimui

8 lentelė „Geo objekto aprašymas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus naudojami laukai: „lat“, „lon“, „country“, „city“.

3.3.6. „User“ objektas

Lentelėje aprašoma objekto „User“ struktūra ir turinys.

Atributas	**Tipas**	**Aprašymas**
id	tekstas	Pasiūlos pusės vartotojo ID
buyerid	tekstas	Reklamuotojo vartotojo ID
yob	skaičius	Gimimo metai (4 skaitmenų)
gender	tekstas	Lytis, kur „M“ – vyras, „F“ – moteris, „O“ – kita

keywords	tekstas	Kableliais atskirtas vartotojo interesų sąrašas
customdata	tekstas	Galimybė pasiūlos pusei pridėti papildomą informaciją apie vartotojo slapukus
geo	objektas	„Geo“ objektas
data	objektų masyvas	„Data“ objektų masyvas
ext	objektas	Vieta „OpenRTB“ praplėtimui

9 lentelė „User objekto aprašymas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus naudojami laukai: „id“, „yob“, „gender“, „keywords“, „geo“.

3.3.7. „IAB“ turinio kategorijos

Lentelėje aprašomos „IAB“ turinio kategorijų grupės.

Reikšmė	**Aprašymas**
IAB1	Menas ir pramogos
IAB2	Automobiliai
IAB3	Verslas
IAB4	Karjera
IAB5	Edukacija
IAB6	Šeima ir tėvystė
IAB7	Sportas ir sveikata
IAB8	Maistas ir gėrimai
IAB9	Hobiai ir pomėgiai
IAB10	Namai ir sodas
IAB11	Teisė, įstatymai ir politika

IAB12	Naujienos
IAB13	Asmeniniai finansai
IAB14	Visuomenė
IAB15	Mokslas
IAB16	Naminiai gyvūnai
IAB17	Sportas
IAB18	Mada ir stilius
IAB19	Kompiuteriai ir technologijos
IAB20	Kelionės
IAB21	Nekilnojamas turtas
IAB22	Prekinimasis
IAB23	Religija
IAB24	Nekategorizuota
IAB25	Nestandartinis turinys
IAB26	Nelegalus turinys
IAB26-1	Nelegalus turinys
IAB26-2	Vogta programinė įranga (angl. *„Warez“*)
IAB26-3	Kenksmingo programinio kodo programos (angl. *„Spyware/Malware“*)
IAB26-4	Autorių teisės pažeidimai

10 lentelė „IAB kategorijų sąrašas“

Pagal lentelėje pateiktus duomenis vartotojo profilio informacijai rinkti bus renkamos ir kategorijos.

Iš lentelių duomenų vartotojo profilis bus formuojamas iš šių laukų: site.domain, site.cat, site.ref, site.search, site.keywords, app.bundle, app.cat, app.keywords, device.geo, device.ifa, device.make,

device.model, device.ua, geo.lat, geo.lon, geo.country, geo.city, user.id (bus naudojamas gauti duomenų tvarkymo platformos ID), user.yob, user.gender, user.keywords, user.geo.

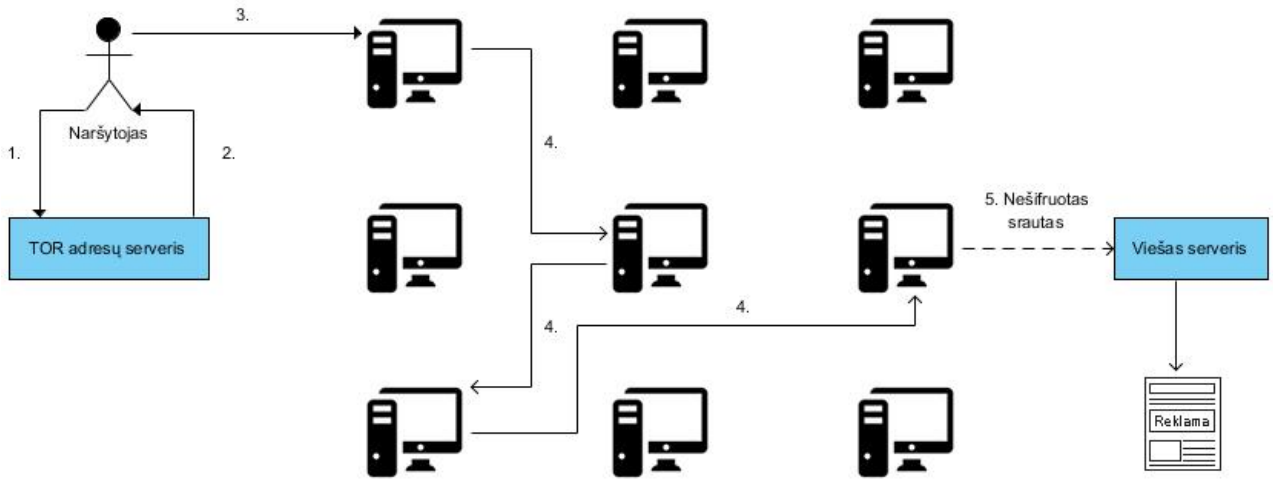
Vienas profilio gali turėti tik vieną iš „site“ arba „app“ laukų, profilio pavyzdys pateiktas paveikslėlyje:


```
{
  "site": {
    "domain": "pavyzdys.lt",
    "cat": [
      "IAB1",
      "IAB1-1"
    ],
    "ref": "https://www.google.com/search?q=pavyzdys",
    "keywords": [
      "IAB4",
      "IAB4-1"
    ]
  },
  "device": {
    "ifa": "ffeea-mmmfa-vuylac-llrda",
    "geo": {
      "lat": 54.69,
      "lon": 25.26,
      "country": "LTU",
      "city": "Vilnius"
    },
    "user": {
      "id": "CASEasgud-321bhaiwqmx",
      "yob": 1988,
      "gender": "M",
      "keywords": [
        "IAB-19",
        "IAB-16"
      ]
    }
  }
}
```


11 pav. „Vartotojo profilio duomenų pavyzdys“

3.3.8. TOR IP aptikimas

TOR yra savanorių administruojama serverių grupė skirta naudotis internetu anonimiškai. Kiekviena vartotojo užklausa yra šifruojama ir siunčiama atsitiktiniu keliu tarp keletos serverių, kol galiausiai pasiekia internetą išėjimo taške:



12 pav. „TOR puslapio užkrovimo užklausa“

Paveikslėlyje parodyta kaip vartotojo užklausa keliauja TOR tinklu:

1. Vartotojas įsijungia TOR naršyklę, kuri siunčia užklausą į TOR adresų serverį;
2. Serveris atsako su prisijungimo nuorodų sąrašu;
3. Vartotojo naršyklė parenka atsitiktinę nuorodą ir siunčia užklausą į ją;
4. Šifruota užklausa keliauja atsitiktiniu keliu tarp keletos TOR serverių;
5. Viename iš tor serverių užklausa pasiekia internetą ir naršyklė užkrauna internetinės svetainės turinį.

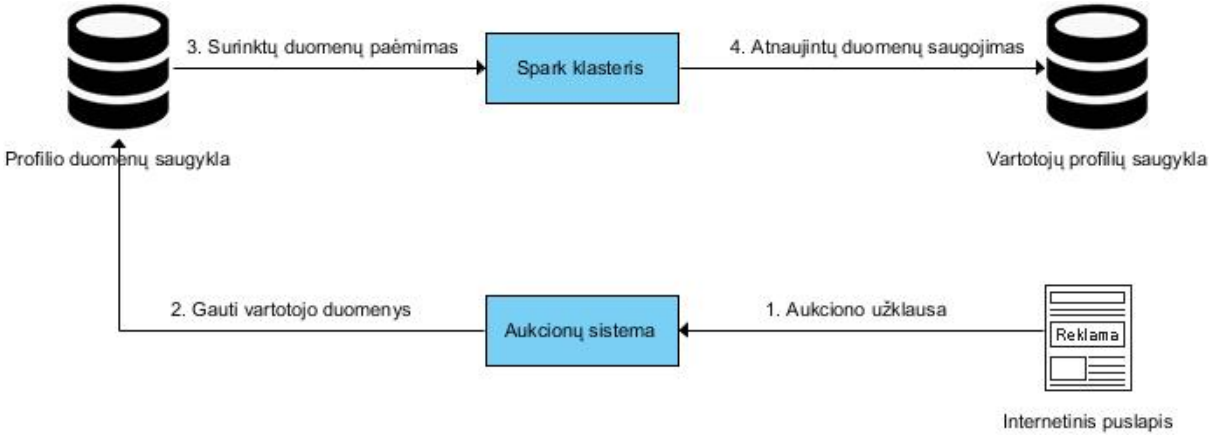
Kadangi užklausa iš TOR tinklo patenka į internetą atsiranda galimybė identifikuoti TOR serverio IP adresą. Tokie adresų sąrašai yra nuolat atnaujinami ir talpinami internete. Vartotojo profilio informacijos sudarymo metu yra patikrinama ar vartotojo IP adresas yra iš žinomų TOR serverių adresų ir šis požymis yra įtraukiamas į profilio duomenis.

3.3.9. Kalėjimo ribose esančių koordinačių aptikimas

Aukcionų užklausų metu gaunama ir GEO vartotojo/įrenginio informacija. Ši informacija panaudojama kaip papildomas profilio informacijos vienetas, jeigu vartotojo koordinatės sutampa su pasaulyje esančių kalėjimų koordinatėmis. Tokiu atveju profilyje atskirai pažymimas požymis „kalėjimo koordinatė“.

3.4. Vartotojo profilio atnaujinimas

Kiekvienos aukciono užklausos metu gali būti atsiunčiami duomenys apie vartotoją. Dažnai šie duomenys yra nepilni, kartais esami duomenys kinta, todėl reikalinga sistema apdoroti vartotojų duomenis ir turėti pilną visų gautų duomenų vaizdą. Procesas detaliau atvaizduotas paveikslėlyje:



13 pav. „Vartotojo profilio duomenų pavyzdys“

Siekiant įgyvendinti vartotojo profilio duomenų atnaujinimą ruošiama sistema sugebanti apdoroti didelius duomenų kiekius ir kasnakt atnaujinti vartotojo profilį:

1. Kartu su aukciono užklausa gaunami vartotojo duomenys;
2. Aukcionų sistema gauna duomenų tvarkymo platformos vartotojo ID, paruošia duomenis pagal 3.3 skyriuje aprašytą formatą ir išsaugo duomenų saugykloje;
3. Po vidurnakčio kuriamas „spark“ klasteris, kuris apdoroja dienos duomenis;
4. Apdorotus (sugrupuotus pagal vartotoją duomenis) sujungia su iš anksčiau žinomais vartotojo duomenimis taip atnaujindamas profilį ir patalpina į vartotojų profilių saugyklą.

Per dieną vartotojai galimai sugeneruoja daugiau negul po vieną reklamos aukciono užklausą, todėl duomenis reikia apjungti. Apjungimo algoritmas įgyvendintas „spark“ klasterio pagalba: skaitomi visi dienos duomenys, grupuojami pagal duomenų platformos vartotojo ID ir taip gaunamas vieno vartotojo sugeneruoti profilio duomenų masyvas. Masyvo elementai apdorojami naudojantis duomenų rinkinio mažinimo algoritmu (angl. *reduce*). Algoritmo veikimo principas:

1. Gaunami vartotojo profilio duomenys A ir B;
2. Jeigu A ir B turi duomenis, naudojami duomenys A:

A: {„user“: {„gender“: „M“}};

B: {„user“: {„gender“: „M“}}.

3. Jeigu A neturi duomens, naudojamas duomuo B:

   A: {„user“: {}};

   B: {„user“: {„gender“: „M“}}.

4. Jeigu B neturi duomens, naudojamas domuo A:

   A: {„user“: {„gender“: „M“}};

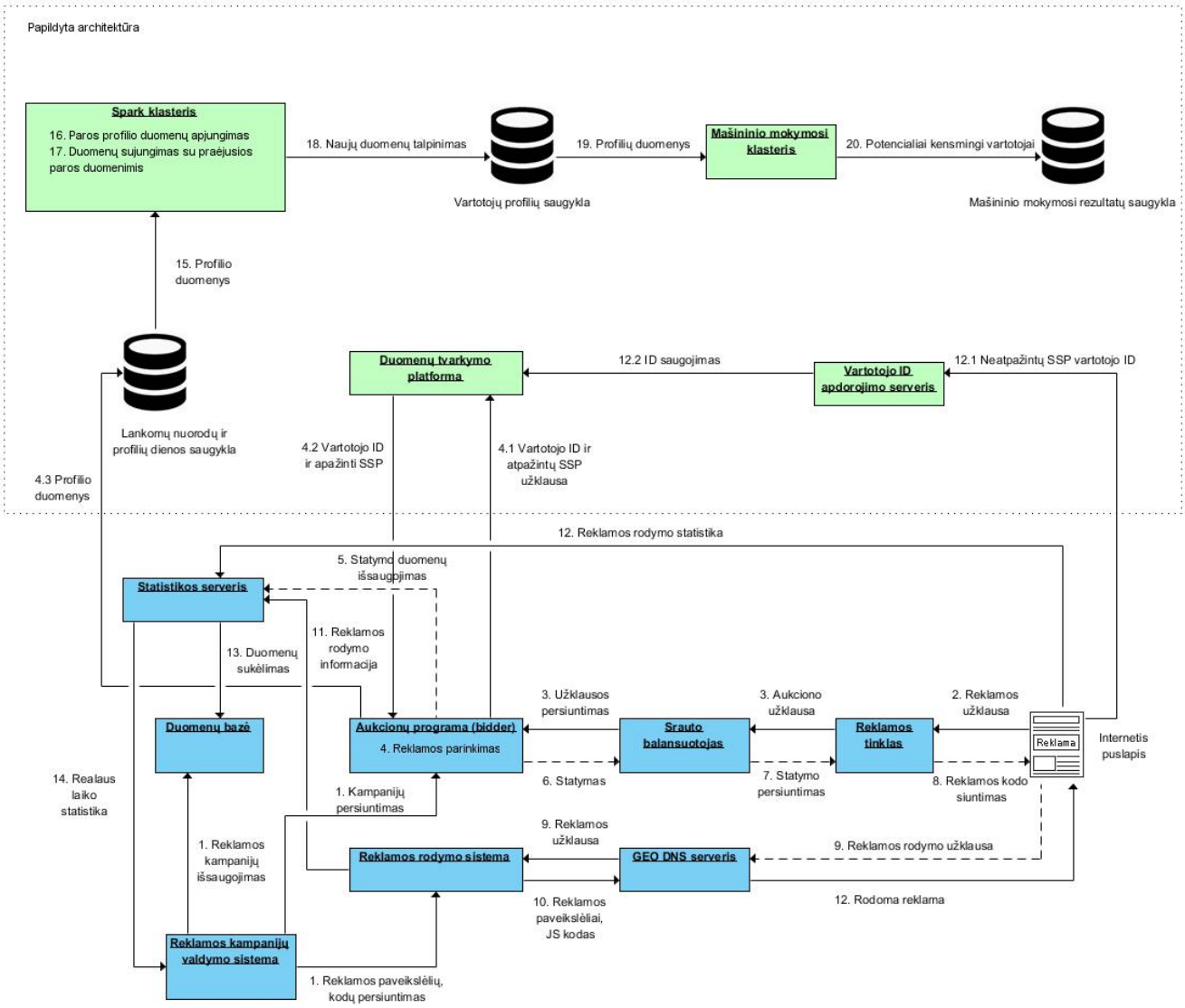
   B: {„user“: {}}.

Taikant tokį algoritmą, galutinis vartotojo profilis turės pilniausius duomenis iš to ką sistema gavo apdorodama reklamos aukciono užklausas.

Gautas rezultatas yra duomenų platformos vartotojo duomenų atvaizdas paros laikotarpyje. Jeigu tai yra pirmoji duomenų atnaujinimo iteracija, duomenys yra rašomi į vartotojų profilių saugyklą. Jeigu vartotojų profilių saugykloje yra praeitos paros duomenų, vykdomas dienos duomenų apjungimas su praeitos paros duomenimis naudojantis duomenų rinkinio mažinimo algoritmu ir galutinis rezultatas įrašomas kaip šios paros rezultatas. Praeitos paros rezultatai yra nebereikalingi ir yra ištrinami.

3.5. Papildyta architektūra

Sistemos veikimui reikalinga papildoma įranga, todėl buvo papildyta egzistuojančios sistemos architektūra. Architektūra buvo pildoma, kadangi egzistuojančioje realaus laiko reklamos pirkimo sistemoje labai didelis dėmesys yra skiriamas efektyviam sistemos veikimui. Įdiegti sistemą į jau egzistuojančius serverius būtų neefektyvus sprendimas. Išnagrinėjus „spark“ veikimo principus, duomenų saugyklas ir duomenų keliavimą iš „spark“ į saugyklas galutinai nuspręsta atskirti egzistuojančia sistemą nuo naujosios ir taip papildyti architektūrą.



14 pav. „Papildyta architektūra“

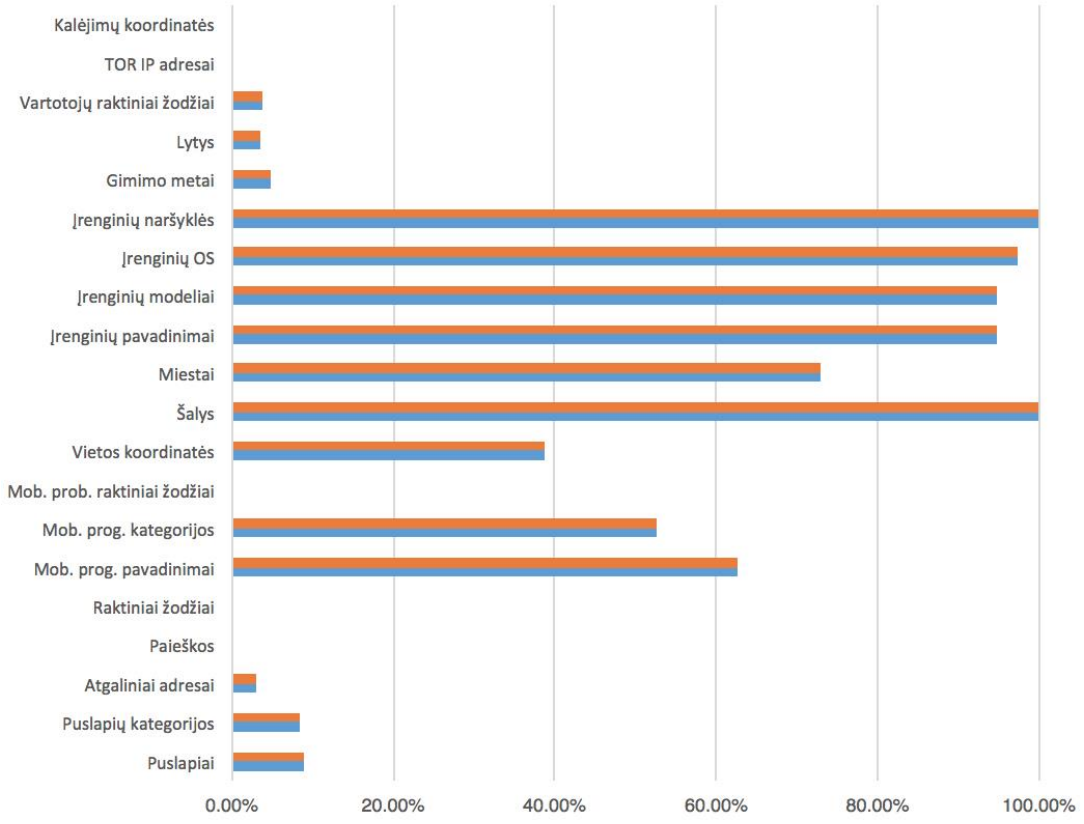
Egzistuojanti sistema papildyta duomenų tvarkymo platformos funkcionalumu. Paruošti metodai aukcionų programai gauti informaciją iš duomenų tvarkymo platformos, ją panaudoti siekiant identifikuoti vartotojų ID dar neįdentikuotuose pasiūlos pusės tinkluose ir ją kaupti. Architektūros papildymai:

- Vartotojo ID apdorojimo serveris atsakingas už užklausų atėjusių iš pasiūlos pusės platformų apdorojimui, jas paverčiant į duomenų struktūrą ir talpinant duomenų tvarkymo platformoje;
- Duomenų tvarkymo platforma laiko žinomus vartotojų identifikatorius ir realiu laiku perduoda informaciją apie juos aukcionų programai;
- Aukcionų programa siunčia informaciją apie vartotoją į lankomų nuorodų ir profilio duomenų saugyklą;

- Kiekvieną naktį atliekamas profilio duomenų apjungimas su iš anksčiau žinomais duomenimis taip vis didinant žinomą informacijos kiekį apie vartotoją;
- Apjungtus profilių duomenis naudoja mašininio mokymosi serveriai siekiant identifikuoti potencialiai kenksmingus vartotojus ir saugoti informaciją apie juos tokių duomenų saugykloje.

3.6. Surinktų duomenų paruošimas mašininiam mokymuisi

Mašininio mokymosi algoritmui buvo renkami vartotojų duomenys. Tryimui surinkta ir panaudota 432.8 GB duomenų, kuriuose yra apie dešimt milijonų (10524080) unikalių vartotojų profilių. Iš visų renkamų duomenų nerasta tik „site.search“ tipo duomenų. Visi vartotojai turi duomenis apie įrenginių naršykles ir vartotojo buvimo šalį aukciono užklausos metu. Pakankamai mažai duomenų yra apie kalėjimų koordinates, TOR tinklo IP adresus, lytis, gimimo metus, mobiliųjų programėlių raktinius žodžius, puslapių atgalinius adresus (angl. *referrer*), puslapių kategorijas ir konkrečias naršytas nuorodas. Iš turimų duomenų galima teigti, kad didžioji dalis vartotojų srauto buvo stebėta mobiliosiose programėlėse.



15 pav. „Stebėtų vartotojų duomenų pilnumas“

Turimų duomenų aibė apskaičiuota naudojantis tikimybiniu duomenų unikalumo skaičiavimo algoritmu HyperLogLog. Surinkti unikalūs duomenys aprašyti lentelėje.

Duomenų tipas	**Unikalios reikšmės**
Vartotojų ID	10471397
Naršomų svetainių internetiniai adresai (vgtu.lt)	13281
Naršomų svetainių kategorijos	354
Naršomų svetainių atgaliniai adresai (angl. *referrer*)	549495
Paieškos frazės per kurias vartotojas pateko į puslapį	0
Naršomų svetainių raktažodžiai	16088
Mobiliųjų programėlių programinis pavadinimas (bar.foo.com)	161590
Mobiliųjų programėlių kategorijos	98
Mobiliųjų programėlių raktiniai žodžiai	19
Geografinės koordinatės	8048592
Šalys	48
Miestai	4738
Įrenginių gamintojai	3448
Įrenginių modeliai	23593
Įrenginių operacinės sistemos	37
Įrenginių naršyklės	939726
Vartotojų gimimo metai	211
Vartotojų lytys	3
Vartotojų raktiniai žodžiai/interesai	721797
TOR IP adresai	2615

Naršymai iš kalėjimo teritorijoje esančių koordinačių	3585

11 lentelė „Vartotojų duomenų unikalios reikšmės“

Iš lentelėje pateiktų duomenų matoma duomenų anomalija, kadangi egzistuoja bent 3 lytys ir 195 vartotojų unikalūs metai (jeigu vartotojas gimęs prieš 3 metus, seniausias vartotojas turėtų būti bent 198 metų amžiaus), todėl vartotojai su tikėtina neteisingais duomenimis yra įtraukiami į potencialiai kenksmingų kandidatų sąrašą.

3.7. Duomenų vektoriaus sudarymas

Siekiant duomenis perduoti mašiniam mokymui pirma juos reikia apdoroti. Kasnakt vykdomas profilio duomenų atnaujinimas leidžia turėti naujausią informaciją apie vartotojus. Surinkti vartotojų duomenys buvo agreguojami kasdien ir taip gaunami vartotojų interesų, naršytų puslapių, raktinių žodžių ir kitų duomenų dažnumo rodikliai. Mašininio mokymosi algoritmuose duomenys yra naudojami skaičių pavidalu, dažniausiai išreikštų vektoriais, todėl surinktus duomenis reikia normalizuoti (paversti skaitinėmis reikšmėmis). Internete egzistuoja beglinė aibė mobiliųjų programėlių, puslapių adresų, atgalinių adresų pavadinimų, todėl duomenis paversti vektoriais yra sudėtinga. Tam, kad tekstinius duomenis ir jų dažnumo požymius pavertimu vektoriais reikia išreikšti tekstines reikšmes skaičiais. Dėl begalinės aibės tokių tekstinių reikšmių pasirinkta naudoti duomenų hešavimo metodą, plačiai naudojamą mašininio mokymosi duomenų normalizavimui. Algoritmas ima tekstinę reikšmę, apskaičiuoja jos hash reikšmę, gautą rezultatą dalina iš potencialaus unikalių rezultatų kiekio ir naudoja liekaną kaip vektoriaus indekso reikšmę. Požymio dažnumas naudojamas tiesiog kaip skaičius:

```
val data = Seq(
  Visits(Map("vgtu.lt"  -> 4, "google.com" -> 19)),
  Visits(Map("spark.apache.org"  -> 9)),
  Visits(Map.empty[String, Int]),
  Visits(Map("vgtu.lt"  -> 21, "github.com" -> 301))
).toDS().map(v => Features(v.domains, freqToVec(v.domains)))
data.show()
```

```
+--------------------------------+--------------------------------------+
|domains                         |features                              |
+--------------------------------+--------------------------------------+
|[vgtu.lt -> 4, google.com -> 19]|(262144,[153771,210147],[19.0,4.0])   |
|[spark.apache.org -> 9]         |(262144,[97640],[9.0])                |
|[]                              |(262144,[],[])                        |
|[vgtu.lt -> 21, github.com -> 301]|(262144,[210147,241859],[21.0,301.0])|
+--------------------------------+--------------------------------------+
```

16 pav. „Apsilankymo dažnumo požymis paverstas vektoriumi“

Tokiu algoritmu paruošiami visi turimi duomenys mašininiam mokymuisi.

Vartotojų geografinių koordinačių aibė taip pat yra labai didelė, todėl kiekviena koordinatė buvo paversta į MGRS koordinačių sistemos reikšmę.

Iš turimų duomenų kuriamas vartotojo duomenų vektorius ir kaip vartotojo pavojingumo požymiai laikomi šie kriterijai:

1. Naršymas naudojantis TOR IP;
2. Naršymas iš kalėjimo teritorijose esančių koordinačių;
3. Turinio naršymas kai turinys pažymėtas potencialiai kenksmingomis kategorijomis (nestandartinis turinys: neapykantos turinys, nemoderuojamas turinys, nelegalus turinys: kenksmingas programinis kodas, „*warez*“);
4. Turinio naršymas potencialiai kenksmingą informaciją platinančiuose tinklalapiuose (raymond.cc).

Tokius vartotojų duomenų vektorius perduodame mašinio mokymosi algoritmams, kurie aptinka potencialiai kenksmingus vartotojus.

3.8. Mašininio mokymosi modelio rezultatų apžvalga

Mašininio mokymo klasifikacijos uždavinys spręstas naudojantis trimis klasifikatoriais: logistinės regresijos, atraminių vektorių ir Naive Bayes. Buvo atliekami 3 tyrimai, kad išsiaiškinti kaip skirtingas duomenų kiekis įtakoja klasifikatoriaus rezultatus, rezultatai pateikiami chronologiška tvarka:

1. 117.9 GB, 5071350 vartotojų profilių algoritmų klasifikacijos rezultatai:

Vartotojas pažymėtas kaip potencialiai kenksmingas	**Vartotojas klasifikuotas kaip potencialiai kenksmingas**	**logistinė regresija**	**atraminiai vektoriai**	**Naive Bayes**
Taip	Taip	9068 (95%)	7805 (82%)	12 (0.1%)
Taip	Ne	479 (5%)	1742 (18%)	9249 (99.9%)
Ne	Taip	128705	89672	0

Ne	Ne	4933098	4972131	5061803

12 lentelė „Pirmoji klasifikacija. Klasifikatorių rezultatai identifikuojant potencialiai kenksmingus vartotojus“

2. 268.8 GB, 7845175 vartotojų profilių algoritmų klasifikacijos rezultatai:

Vartotojas pažymėtas kaip potencialiai kenksmingas	**Vartotojas klasifikuotas kaip potencialiai kenksmingas**	**logistinė regresija**	**atraminiai vektoriai**	**Naive Bayes**
Taip	Taip	15114 (93%)	12404 (77%)	31 (0.2%)
Taip	Ne	1011 (7%)	3721 (23%)	16094 (99.8%)
Ne	Taip	189646	139264	4
Ne	Ne	7639404	7689786	7829046

13 lentelė „Antroji klasifikacija. Klasifikatorių rezultatai identifikuojant potencialiai kenksmingus vartotojus“

3. 432.8 GB, 10524080 vartotojų profilių algoritmų klasifikacijos rezultatai:

Vartotojas pažymėtas kaip potencialiai kenksmingas	**Vartotojas klasifikuotas kaip potencialiai kenksmingas**	**logistinė regresija**	**atraminiai vektoriai**	**Naive Bayes**
Taip	Taip	23639 (91%)	19963 (77%)	85 (1%)
Taip	Ne	2270 (9%)	5946 (23%)	25824 (99%)
Ne	Taip	215561	188049	11
Ne	Ne	10282610	10310122	10498160

14 lentelė „Trečioji klasifikacija. Klasifikatorių rezultatai identifikuojant potencialiai kenksmingus vartotojus“

Kadangi Naive Bayse klasifikatoriaus rezultatai yra abejotini, šiuo metodu klasifikuotų vartotojų tyrime nenaudota.

Atmetus pagal taškų skaičiavimo taisykles identifikuotus vartotojus, logistinės regresijos ir atraminių vektorių klasifikatoriai identifikavo 176230 potencialiai kenksmingų vartotojų pirmuoju bandymu, 277049 – antruoju ir 344957 – trečiuoju bandymu. Identifikuoti vartotojai sudaro 3% tirtų duomenų. Potencialiai kenksmingų vartotojų unikalių duomenų pasiskirstymas:

Duomenų tipas	**Unikalios reikšmės**
Vartotojų ID	344957
Naršomų svetainių internetiniai adresai (vgtu.lt)	2277
Naršomų svetainių kategorijos	215
Naršomų svetainių atgaliniai adresai (angl. *referrer*)	14681
Paieškos frazės per kurias vartotojas pateko į puslapį	0
Naršomų svetainių raktažodžiai	988
Mobiliųjų programėlių programinis pavadinimas (bar.foo.com)	46522
Mobiliųjų programėlių kategorijos	84
Mobiliųjų programėlių raktiniai žodžiai	19
Geografinės koordinatės	2938047
Šalys	35
Miestai	2975
Įrenginių gamintojai	1268
Įrenginių modeliai	10488
Įrenginių operacinės sistemos	18
Įrenginių naršyklės	88292
Vartotojų gimimo metai	76
Vartotojų lytys	2
Vartotojų raktiniai žodžiai/interesai	403538

TOR IP adresai	0
Naršymai iš kalėjimo teritorijoje esančių koordinačių	0

15 lentelė „Potencialiai kenksmingų vartotojų unikalių duomenų pasiskirstymas“

Klasifikatoriai taip pat rado vartotojų, neatitinkančių realybės: vartotojas A matytas su 2108 skirtingomis naršylėmis, 2 lytimis, 10 skirtingų gimimo metų reikšmių (skirtumas tarp didžiausios ir mažiausios – 34 metai), 565 skirtingais įrenginių modeliais.

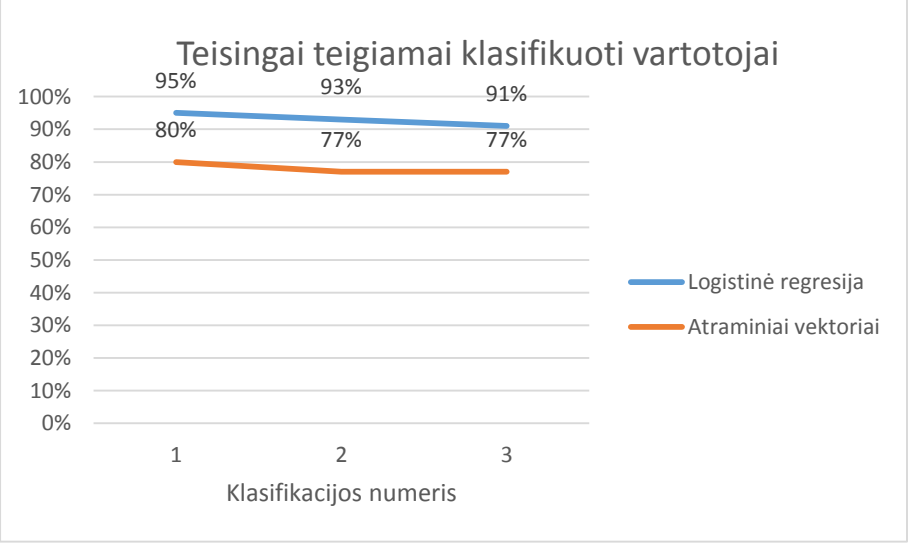
Klasifikavimas atliktas naudojantis „spark“ klasteriais: 8 serveriai, kiekvienas turintis 16 branduolių ir 32 GB operatyviosios atminties.

Sparčiausiai veikia logistinės regresijos ir „Naive Bayse“ klasifikatoriai:

	logistinė regresija	**atraminiai vektoriai**	**Naive Bayse**
117.9 GB	6 min.	48 min.	5 min.
268.8 GB	8 min.	1 val. 3min.	6 min.
432.8 GB	10 min.	1 val. 34min.	9 min.

16 lentelė „Klasifikavimo laikų palyginimas“

Surinktų vartotojų klasifikavimo atveju greičiausiai veikia logistinės regresijos ir „Naive Bayse“ klasifikatoriai. Atraminių vektorių veikimo greitis skiriasi 9 kartus.



17 pav. „Klasifikatorių teisingai teigiamai nustatytas pavojingumo požymis“

Iš bandymų matyti, kad logistinės regresijos klaidingai teigiamų vartotojų kiekis auga kartu su duomenų kiekiu, tačiau atraminių vektorių atveju – išlieka stabilus.

Iš trijų bandymų su skirtingais duomenų kiekiais pastebėta, kad didėja klaidingai neigiamų rezultatų kiekis, kuris stabilizuojasi naudojantis atraminių vektorių klasifikatoriumi. Paruoštas sprendimas identifikuoja 3% potencialiai kenksmingų vartotojų, kurie gali būti perduoti žmogui papildomam įvertinimui atlikti.

4. Išvados

Viena iš pagrindinių problemų aptinkant potencialiai kenksmingus vartotojus yra ta, kad nėra konkretaus tipinio tokio vartotojo profilio. Surinkti duomenys parodė didelį unikalių reikšmių pasiskirstymą tarp potencialiai kenksmingų vartotojų. Dėl tokio duomenų pasiskirstymo egzistuoja galimybė neteisingai identifikuoti paprastą vartotoją kaip potencialiai kenksmingą. Tokių vartotojų identifikavimas paruoštu metodu yra tik vienas iš šios sudėtingos problemos sprendimų būdų.

Atlikus darbą padarytos šios išvados:

1. Atlikus literatūros analizę paaiškėjo, kad egzistuoja daug skirtingų metodų vartotojo unikalumui internete atpažinti, išaiškinti demografinius duomenis, interesus. Mašininio mokymosi algoritmai yra tinkamas sprendimas klasifikuoti turimus duomenis taip išskiriant vartotojus potencialiai linkusius atlikti kenksmingą veiklą;
2. Išanalizavus egzistuojančius vartotojo profiliavimo būdus bei jų efektyvumą paaiškėjo, kad dažniausiai pavojingų vartotojų identifikavimui yra sisteminami duomenys apie juos ir naudojami mašininio mokymosi algoritmai siekiant sumažinti rankinį darbą;
3. Vartotojo profiliavimo būdai patobulinti renkant vartotojų informaciją iš susistemintų duomenų per skirtingas platformas, todėl profiliai yra platesni negu nagrinėtuose sprendimuose;
4. Paruoštas vartotojo duomenų rinkimo metodas, leidžiantis aptikti potencialiai pavojingus vartotojus identifikavo apie 3% tokių vartotojų iš surinktų duomenų, todėl atrinktus vartotojus galima perduoti kitoms sistemos tolimesniems tyrimams;
5. Atlikus metodo efektyvumo bandymus paaiškėjo, kad randami nauji potencialiai pavojingi vartotojai, tačiau negalima vienareikšmiškai teigti, kad tokie vartotojai yra linkę įgyvendinti potencialiai kenksmingą veiklą, kadangi jų duomenų pasiskirstymas yra platus.

Literatūra

1. Lukasz Ilejnik, Claude Castelluccia, Artur Janc. Why Johhny Can‘t Browse in Peace: On the Uniqueness of Web Browsing History Patterns, 2012.
2. I Still Know What You Visited Last Summer: Leaking browsing history via user interaction and side channel attacks, Zack Weinberg, Eric Chen, Pavithra Ramesh Jayaraman, and Collin Jackson, 2011.
3. T. Zhu, R. Greiner, and G. Häubl. An effective complete-web recommender system, 2003.
4. R. Kawase, E. Herder, Classification of user interest patterns using a virtual folksonomy, 2011.
5. Carmagnola F., Cena F., Console L., Cortassa O., Gena C., Goy A., Torre I., Toso A., Vernero F.: Tag-based user modeling for social multi-device adaptive guides. User Model. User-Adap. Inter, 2008.
6. Jian Hu , Hua-Jun Zeng , Hua Li , Cheng Niu , Zheng Chen, Demographic prediction based on user's browsing behavior, 2007.
7. F. H. Wang, H. M. Shao, Effective personalized recommendation based on time-framed navigation clustering and association mining, 2004.
8. Henry Lieberman, Letizia: an agent that assists web browsing, 1995.
9. Jaime Teevan , Susan T. Dumais , Eric Horvitz, Personalizing search via automated analysis of interests and activities, 2005.
10. Stange, Martin and Funk, Burkhardt, Predicting online user behavior based on real-time advertising data, 2016.
11. M. Grcar, D. Mladenic, M. Grobelnik, User profiling for interest-focused browsing history, 2005.
12. S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, 2007.
13. Kęstučio Bytauto Magistro baigiamasis darbas, LRS Seimo narių grupavimas pagal balsavimą ir balsavimo kitimo aptikimas, 2012.
14. Jolitos Bernatavičienės Daktaro disertacija, Vizualios žinių gavybos metodologija ir jos tyrimas, 2008.

15. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütz, An introduction to information retrieval, 2004.
16. Anomaly detection: A survey. [Interaktyvus]. [Žiūrėta 2018-01-09]. Prieiga per internetą: http://doi.acm.org/10.1145/1541880.154188
17. A Universally Unique IDentifier (UUID) URN Namespace. [Interaktyvus]. [Žiūrėta 2018-02-15]. Prieiga per iternetą: https://tools.ietf.org/html/rfc4122#section-4.1.4
18. Real Time Bidding (RTB) Project OpenRTB API Specification Version 2.5. [Interaktyvus]. [Žiūrėta 2018-05-19]. Prieiga per internetą https://www.iab.com/wp-content/uploads/2016/03/OpenRTB-API-Specification-Version-2-5-FINAL.pdf
19. Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, Derek Ruths, Web of Hate: Tackling Hateful Speech in Online Social Space, 2010.
20. Joel Brynielsson, Andreas Horndahl, Fredrik Johansson, Lisa Kaati, Christian M˚artenson, Pontus Svenson, Analysis of Weak Signals for Detecting Lone Wolf Terrorists, 2012.
21. Reza Zafarani, 10 Bits of Surprise: Detecting Malicious Users with Minimum Information, 2015.
22. Jesus Mena, Investigative Data Mining for Security and Criminal Detection, 2003.