VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

Didelių duomenų srautų analizė, anomalijų aptikimas

Big data analysis, detection of anomalies

Kursinis darbas

Atliko: 3 kurso 3 grupės studentas
Jokūbas Rusakevičius (parašas)

Darbo vadovas: dr. Vytautas Valaitis (parašas)

Vilnius – 2018

TURINYS

Įvadas

Šis darbas yra Programų Sistemų studijų trečio kurso privalomas kursinis darbas apie anomalijų aptikimą didžiuosiuose duomenyse (angl. *Big Data*) bei anomalijų aptikimą naudojant atviro kodo analitinį įrankį „MacroBase“.

Problematika

Remiantis 2001 metais „Gartner“ pateiktu apibrėžimu (iki šiol laikomu pagrindiniu (angl. *go-to*)), didieji duomenys - tai duomenys, kurie turi didelę įvairovę (angl. *variety*), yra renkami nuolat didėjančiais kiekiais (angl. *volumes*) ir generuojami vis didėjančiais greičiais (angl. *velocity*), šis apibrėžimas dar vadinamas tryjų „V“ [Lan01]. Paprastai, didžiuosius duomenis galima apibrėžti kaip duomenų rinkinius tokius didelius, kad tradiciniai programiniai duomenų apdorojimo įrankiai nesugeba jų sugauti, tvarkyti, organizuoti, apdoroti ar su jais dirbti priimtiname laiko intervale [SMR12], jie yra paprasčiausiai per dideli ir per daug sudėtingi. Tačiau šie duomenys turi milžinišką potencialą ir gali būti panaudojami sprendžiant verslo ir kitas problemos, kurių sprendimas iki šiol buvo neįmanomas.

Generuojant ir saugant milžiniškus kiekius duomenų, natūraliai, užfiksuojami tokie duomenys, kurie išsiskiria ir yra nebūdingi duomenų rinkiniui. Duomenų vienetai kurie yra nukrypę nuo kitų duomenų rinkinyje yra vadinami anomalijomis. Anomalijų aptikimas yra procesas, kurio metu yra aptinkama ir identifikuojama anomalija arba išskirtis (angl. *outlier*) duomenų rinkinyje [Tec18]. Anomalijos yra retas reiškinys, tačiau jų egzistavimas gali reikšti didelį pavojų taikomajai sistemai ar jos naudotojams. Šio darbo metu bus tiriamas anomalijų aptikimas remiantis iš „Backblaze“ 2018 metų (pirmą ketvirtį) duomenų centruose esančių diskų surinkta informacija [Bac18].

Dėl įvairių priežasčių renkamų, saugomų ir operuojamų duomenų kiekiai nuolatos didėja ir netgi gerokai lenkia žmogaus sugebėjimą juos apdoroti ar analizuoti. Didžiosios socialinių tinklų kompanijos Twitter, Facebook ir LinkedIn 2015-2016 metais pranešė kiekviena atskirai fiksuojanti iki 12 milijonų įvykių per sekundę [Ast16; PFT+15; Woo15]. Taip pat negalima pamiršti vis labiau plintančių ir didelius kiekius duomenų generuojančių automatizuotų duomenų šaltinių („Dalykų Interneto“ (angl. *Internet of Things* arba *IoT*)). Be to, tokie palankūs veiksniai kaip kylantis automatizuotų duomenų šaltinių populiarumas, pinganti techninė įranga, išvystyti komunikaciniai tinklai bei mažėjančios duomenų saugojimo kainos paskatino dešimčių milijardų dolerių komercines investicijas šių technologijų vystimui [MCB+15]. Dėl šių ir kitų veiksnių numatoma, kad kiekvienais metais bendras duomenų kiekis išaugs po 40% [EMC14], o iki 2020 metų pranašaujama, kad bendras pasaulinis duomenų kiekis peržengs 40 zetabaitų (4×10^{22} baitų) ribą [HA17].

Didžiųjų duomenų laikomos informacijos paslėpta nauda ir svarba yra visuotinai pripažįstama, tačiau ši informacija nėra lengvai išgaunama. Duomenų peržiūra ir analizė sudaro labai didelį krūvį tiek analitikui, tiek analitiniams įrankiams. Fizinė duomenų peržiūra yra paprasčiausiai neįmanoma, o nuolatos didėjantys duomenų kiekiai vis labiau atskiria dėmesio reikalaujančius

duomenis ir ribotą dėmesį turintį analitiką. Net ir aukščiausios kvalifikacijos analitikai praneša panaudojantys vos iki 6% jų surenkamų duomenų [BGR+17]. Todėl atsiranda iššūkis prioritizuoti žmogaus dėmesį. Nors žmogui yra neįmanoma peržiūrėti visų šių duomenų, tačiau kompiuteriai ir/ar mašinos gali. Informacinės sistemos labiau nei bet kada turi filtruoti, akcentuoti, jungti, grupuoti pateiktus duomenis, jiems suteikti kontekstą ir rodyti naudotojui tik ribotą, svarbią bei apibendrintą informaciją. Visa rodoma, bet nereikalinga informacija reikalauja ir eikvoja žmogaus dėmesį [Sim71].

Standfordo universitetas kartu su Masačusetso technologijos institutu 2017 metais paskelbė kuriantys naują atviro kodo, ne tik didžiųjų, bet ir greitųjų duomenų (angl. *Fast Data*) analitinį paieškos įrankį. „MacroBase“ pagrindinis uždavinys yra žmogaus dėmesio prioritizavimas. Vienas iš „MacroBase“ šio uždavinį sprendimų yra sugeneruoti didžiausio dėmesio reikalaujančią supaprastintą išvestį, kurios neįprastus duomenų vienetus „MacroBase“ padeda aiškinti pagal duomenų atributus [BGM+17; BGR+17]. „MacroBase“ yra naujas ir modernus analitinis įrankis, pateikiantis inovatoriškų sprendimų vis didėjančių ir greitėjančių duomenų srautų analizei atlikti bei galintis grąžinti tikslius rezultatus dirbdamas 2 milijonų įvykių per sekundę greičiu per užklausą per branduolį, dėl to, šiame darbe bus plačiau nagrinėjamas bei eksperimentai atliekami naudojant būtent šį įrankį.

Darbo tikslas ir uždaviniai

Šio darbo **tikslas** - palyginti iki šiol naudotas didelių duomenų analizavimo technologijas ir pasinaudoti „MacroBase“ duomenų analizavimo ir anomalijų aptikimo įrankį anomalijų aptikimui.

Darbui iškelti **uždavyniai**:

1. Paaiškinti kas yra „Didieji Duomenis“.
2. Surasti ir palyginti dabar naudojamus anomalijų dideliuose duomenyse aptikimo įrankius.
3. Išanalizuoti „MacroBase“.
4. Paruošti eksperimentui reikalingą įrašų rinkinį iš „Backblaze“ kiekvieną ketvirtį skelbiamų duomenų.
5. Įsidiegti ir paruošti darbui „MacroBase“ analitinį įrankį.
6. Atlikti eksperimentus ir aptikti anomalijas paruoštuose duomenyse.
7. Pateikti galutines eksperimento išvadas.

Rezultatai ir išvados

Rezultatų ir išvadų dalyje turi būti aiškiai išdėstomi pagrindiniai darbo rezultatai (kažkas išanalizuota, kažkas sukurta, kažkas įdiegta) ir pateikiamos išvados (daromi nagrinėtų problemų sprendimo metodų palyginimai, teikiamos rekomendacijos, akcentuojamos naujovės).

Literatūra

[Ast16] Anthony Asta. Observability at twitter: technical overview, part i. 2016. URL: https://blog.twitter.com/engineering/en_us/a/2016/observability-at-twitter-technical-overview-part-i.html.

[Bac18] Backblaze. Hard drive data and stats. 2018. URL: https://www.backblaze.com/b2/hard-drive-test-data.html.

[BGM+17] Peter Bailis, Edward Gan, Samuel Madden, Deepak Narayanan, Kexin Rong ir Sahaana Suri. Macrobase: prioritizing attention in fast data. *SIGMOD'17- Proceedings of the 2017 ACM International Conference on Management of Data*, p. 541–556, Chicago, Illinois, USA. Stanford Infolab ir Massachusetts Institute of Technology, ACM New York, 2017. ISBN: 978-1-4503-4197-4.

[BGR+17] Peter Bailis, Edward Gan, Kexin Rong ir Sahaana Suri. Prioritizing attention in fast data: principles and promise. *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research*. Stanford Infolab, 2017.

[EMC14] Dell EMC. The digital universe of opportunities: rich data and the increasing value of the internet of things. 2014. URL: http://www.emc.com/leadership/digital-universe/.

[HA17] Makrufa Sh. Hajirahimova ir Aybeniz S. Aliyeva. About big data measurement methodologies and indicators. *International Journal of Modern Education and Computer Science*, 9:1–9, 2017.

[Lan01] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Tech. atask., META Group, 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[MCB+15] James Manyika, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs ir Jacques Bughin Dan Aharon. *McKinsey Global Institute: The internet of things: mapping the value beyond the hype*. 2015.

[PFT+15] Tuomas Pelkonen, Scott Franklin, Justin Teller, Paul Cavallaro, Qi Huang, Justin Meza ir Kaushik Veeraraghavan. Gorilla: a fast, scalable, in-memory time series database. *VLDB Endowment - Proceedings of the 41st International Conference on Very Large Data Bases*, p. 1816–1827, Kohala Coast, Hawaii. Facebook Inc., 2015.

[Sim71] Herbert A. Simon. *Designing organizations for an information-rich world*. *In Computers, communications, and the public interest*. Martin Greenberger, redaktorius. 1971, p. 37–72.

[SMR12] Chris Snijders, Uwe Matzat ir Ulf-Dietrich Reips. „big data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, 7:1–5, 2012. URL: http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.

[Tec18] Techopedia. Anomaly detection. 2018. URL: https://www.techopedia.com/definition/30297/anomaly-detection.

[Woo15] Alex Woodie. Kafka tops 1 trillion messages per day at linkedin. 2015. URL: https://www.datanami.com/2015/09/02/kafka-tops-1-trillion-messages-per-day-at-linkedin/.

Santrumpos

Sąvokų apibrėžimai ir santrumpų sąrašas sudaromas tada, kai darbo tekste vartojami specialūs paaiškinimo reikalaujantys terminai ir rečiau sutinkamos santrumpos.