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

Renkamų, saugomų ir operuojamų duomenų kiekiai nuolatos didėja ir gerokai lenkia žmogaus sugebėjimą juos apdoroti, peržiūrėti ar analizuoti. Didžiosios socialinių tinklų kompanijos Twitter, Facebook ir LinkedIn praneša kiekviena atskirai fiksuojanti iki 12 milijonų įvykių per sekundę [Ast16; PFT+15; Woo15]. Taip pat, vis daugiau duomenų yra surenkama iš automatizuotų duomenų šaltinių (pvz.: „Dalykų Interneto“ (angl. „Internet of Things“)). 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]. Numatoma, kad kiekvienais metais bendras duomenų kiekis išaugs po 40% [EMC14]. Tačiau žmogaus skiriamo dėmesio kiekis yra fiksuotas ir nekinta. Dėl šios priežasties, yra vis labiau neįmanoma remtis tik fizine šių „Didžiųjų Duomenų“ (angl „Big Data“) peržiūra. Taip pat, dėl didžiulio duomenų kiekio ir limituoto žmogaus dėmesio kiekio, dabartiniai aukščiausios klasės taikomųjų programų operatoriai praneša panaudojantys vos 6% jų surenkamų duomenų [BGR+17].

Dabartinė „Didžiųjų Duomenų“ era su savimi atsinešė milžiniškus kiekius duomenų ir padėjo pagrindus naujai „Greitųjų Duomenų“ (angl. „Fast Data“) erai. „Greitųjų Duomenų“ era bus pažymėta didžiuliu duomenų pertekliumi ir resursų jų apdorojimui ir interpretavimui trūkumu [BGR+17]. Todėl atsiranda iššūkis, prioritizuoti dėmesį. Nors žmogui yra neįmanoma peržiūrėti visų šių duomenų, tačiau kompiuteriai ir mašinos gali. Informacinės sistemos labiau nei bet kada turi filtruoti, akcentuoti, jungti, grupuoti pateiktus duomenis ir rodyti naudotojui tik ribotą bei apibendrintą informaciją. Visa rodoma, bet nereikalinga informacija reikalauja ir eikvoja žmogaus dėmesį [Sim71]. Tačiau, ypač dideli duomenų kiekiai gali būti per dideli ne tik žmogui, bet ir mašinai arba jos sugebėjimui ekonomiškai apdoroti duomenis dėl atliekamų skaičiavimų. Viso to pasekmė yra dažnas svarbaus funkcionalumo nepastebėjimas, dėl to krentantis efektyvumas bei atsakymų praradimas.

Stadfordo Universiteto ekspertai bei doktorantai, norėdami sugretinti žemo lygio duomenų srautų apdorojimo variklius su efektyviais ir tiksliais analitiniais varikliais gebančiais prioritizuoti dėmesį greituosiuose duomenyse, pradėjo kurti MacroBase [BGM+17], atviro kodo greitųjų duomenų analitinį variklį. MacroBase pagrindinis veikimo principas yra žmogaus dėmesio prioritizavimas. Šito uždavinio sprendimui naudojami analitiniai operatoriai, kurie identifikuoja reikšmingus taškus duomenų sraute, ir kurie sugretina, sulygina ir sugrupuoja panašumus tarp jų. Šie ir kiti operatoriai užtikrina, kad gražinamas rezultatas ir jo požymiai yra tikrai reikšmingi. Mažas pagrindinių greitųjų duomenų operatorių kiekis leidžia šiai sistemai likti lanksčiai ir lengvai pritaikomai įvairioms bei skirtingoms sistemoms.

Atviro kodo analitinis paieškos variklis MacroBase yra nuolatos kuriamas ir naujinamas. Šiuo darbu bus siekiama, įsidiegus šį variklį, ištirti MacroBase „Pašalinių“ (angl. „Outlier“) duomenyse aptikimo tikslumą naudojant skirtingus MacroBase algoritmus. Taip pat, priklausomai nuo gautų rezultatų pabandyti pagerinti tikslumą. Šiuo darbu tikimasi....

1. Greitųjų duomenų dėmesio prioritizavimas

Greitųjų duomenų analizavimo sistemos privalo prioritizuoti programuotojų, kūrėjų, naudotojų ir mašinų dėmesį, akcentuodamos tik tikrai reikšmingus duomenis. Šiame skyriuje bus pristatyti trys kūrimo principai nepakankamo dėmesio prioritizavimui [BGR+17], kuriais remiantis yra kuriamas MacroBase. Pirma, rodomi rezultato duomenys gali būti apibendrinti, sujungti ir kontekstualizuoti, taip sufokusuojant srities eksperto dėmesį 1.1. Antra, iteracinio kūrimo palengvinimui, inžineriniai resursai turėtų būti protingai skirstomi į sąsajas (angl. „Interfaces“) 1.2. Trečia, arčiausiai duomenų šaltinio esantys skaičiavimo resursai gali būti skiriami tik didžiausią įtaką rezultatui turintiems duomenims 1.3.

1.1. Rezultato prioritizavimas

Žmogaus dėmesys yra fiksuotas, todėl vos keli per sekundę neapdorotų duomenų (angl. „Raw Data“) rezultatai parodyti sistemos naudotojui gali būti neįveikiamai daug. Priklausomai nuo sistemos masto, net sekundė skirta kiekvienam neapdorotų duomenų rezultatui stabdo darbo procesą. Be to, duomenų srautai fiksuojantys šimtus tūkstančių ar net daugiau įvykių per sekundę nebėra retas atvejis. Todėl su tokiais duomenų srautais dirbančios greitųjų duomenų sistemos negali sau leisti grąžinti ir rodyti sistemos naudotojui rezultatą su rodomais neapdorotais duomenimis.

Rezultato prioritizavimas yra pasiekiamas grąžinant daugiau informacijos su mažesne išeiga. Tai reiškia, kad rodoma naudotojui ne daug konkrečių rezultatų, o keli apibendrinti ir sugrupuoti. Be to, greituosius duomenis prioritizuodamos bei atlikdamos sprendimą, kokius ir/ar kokio tipo duomenis rodyti naudotojui, sistemos turi tai atlikti protingai. Kaip rašoma knygoje „Dieter Rams“, greitųjų duomenų sistema turėtų grąžinti mažiau, bet geresnius rezultatus [LKI11].

Pasiekti dėmesio prioritizavimo prioritizuojant rezultatą galima keliais būdais. Vienas iš būdų greitųjų duomenų sistemai tai padaryti yra apibendrinant ir apjungiant duomenis, taip suteikti jiems kontekstą ir akcentus pagrindiniams duomenų požymiams ar elgsenai duomenų junginio viduje ar duomenų visumoje. Pavyzdžiui, vietoje visų 200 tūkstančių probleminių įrašų gautų iš įrenginio 1225 grąžinimo ir atvaizdavimo, sistema galėtų grąžinti tik įrenginio identifikacinį kodą (šiuo atveju 1225) ir įtartinų įrašų skaičių (200 tūkst.). Dar vienas būda tolesniam dėmesio prioritizavimui gali būti rodymas duomenų pagal jų svarbumą ir reikšmingumą sistemos naudotojui. Taip pat, sistemos, duomenų junginiams ir duomenų svarumo reitingams paremti, gali pasinaudoti natūraliosiomis hierarchijomis ir ontologijomis, taip suteikdamos naudotojams galimybė gauti reikiamą informacijos detalumo lygi.

1.2. Iteracijų prioritizavimas

Šiuolaikinė analizė ir jos darbo eiga susideda iš daug įvairių žingsnių, todėl pats svarbiausias aukštos kokybės analitinės sistemos projektavimo ir kūrimo principas yra iteracijos. Šie žingsniai apima: funkcijų kūrimą (angl. „Feature Engineering“), modelio pasirinkimą, parametrų reguliavimą ir efektyvumo inžineriją (angl. „Performance Engineering“). Be to, minėtieji žingsniai yra

iš prigimties iteratyvūs ir varomi grįžtamuoju ryšiu (angl. „Feedback-driven“). Tačiau, didelė dalis ekspertų ekspertizių nėra paverčiamos analizėmis, nes modernios analizės darbo procesas reikalauja labai daug darbo jėgos. Yra skiriamos didelės lėšos, komandos ir daug aukštos kvalifikacijos mokslininkų, kurie dažnai atlieka vienodas, varginančias ir nuobodžias užduotis. Todėl, greitųjų duomenų sistemų užduotis turėtų būti: pasitelkusios išskirtinai iteracinį darbo procesą – pagreitinti grįžtamuoju ryšiu varomus procesus ir nuleisti reikalingų analizei atlikti žinių barjerą.

Sistemos projektavimo procesas turėtų suteikti sąlygas iteracijomis bei grįžtamuoju ryšių paremtam kūrimui. Yra geriau sistemai išlikti lanksčiai, nei tobulai tiksliai. Tai reiškia, kad yra geriau suteikti sistemai naudingus numatytuosius nustatymus bei padaryti analizės darbo procesą lengvai konfigūruojama, nei užkrauti naudotojams naštą tai daryti patiems. Greitųjų duomenų sistemos iš naudotojų turėtų beveik nereikalauti jokio darbo, suteikdamos jiems lengvai naudojamus ir reguliuojamus analitinius įrankius. Turi būti vadovaujamasi mąstymo, kad pirmas modelis retai bus paskutinis, tam geri numatytieji nustatymai gali padėti gauti greitą grįžtamąjį ryšį bei rezultatus. Tačiau galingi sistemos įrankiai skirti patyrusiems ir aukštos kvalifikacijos sistemos naudotojams galintys konfigūruoti sudėtingąją analitinio darbo proceso pusę turėtų būt taip pat suteikiami.

Dauguma dabartinių sistemų gebančių atlikti tam tikrą užduotį gerai yra labai retai pritaikomos kitų užduočių sprendimui. Todėl, greitųjų duomenų sistema turėtų būt kuriama moduli ir iteratyviai plečiama. Sistemos naudotojai neturi atlikti vienodų, klaidoms atsirasti aplinkybes suteikiančių, nuobodžių ir varginančių užduočių, šios užduotys turėtų būti automatizuojamos sistemos, taip nuimant naštą nuo naudotojų. Visas naudotojo dėmesys turėtų būt sutelktas į taikomąją sritį ir į jai specifines žinias. Šie du principai suteikia net ir geriausiems analitiniams procesams galimybę didinti darbo mastą ir duomenų kiekį.

1.3. Skaičiavimų prioritizavimas

2. MacroBase duomenų analizės sistema

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.

[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/.

[LKI11] Sophie Lovell, Klaus Kemp ir Jonathan Ive. *Dieter Rams: As Little Design as Possible*. Phaidon Press, 2011.

[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.

[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.