MOKSLAS – LIETUVOS ATEITIS
SCIENCE – FUTURE OF LITHUANIA
ISSN 2029-2341 print / ISSN 2029-2252 online
http://www.mla.vgtu.lt

ELEKTRONIKA IR ELEKTROTECHNIKA
ELECTRONICS AND ELECTRICAL ENGINEERING

2014 6(2): 162–167

Sistemų inžinerija, kompiuterių technologija T 120
System Engineering, Computer Technologies T 120

http://dx.doi.org/10.3846/mla.2014.22

KOMPIUTERIŲ TINKLO SRAUTO ANOMALIJŲ ATPAŽINIMAS MAKSIMALIOS ENTROPIJOS METODU

Dalius Mažeika[1], Saulius Jasonis[2]

Vilniaus Gedimino technikos universitetas
El. paštas: [1]dalius.mazeika@vgtu.lt; [2]saulius.jasonis@bdc.lt

Santrauka. Straipsnyje nagrinėjama kompiuterių tinklo srauto anomalijų atpažinimo problema. Kompiuterių tinklo srautui stebėti pasirenkama NetFlow technologija, o anomalijos aptinkamos maksimalios entropijos metodu. Metodas pritaikytas NetFlow pateikiamiems duomenims apdoroti. Sukurta programinė priemonė ir atliktas eksperimentinis metodo tinkamumo tyrimas analizuojant „Cisco“ maršrutizatoriaus srauto duomenis. Metodas patobulintas siekiant pagreitinti skaičiavimus, tačiau neprarandant tikslumo. Nustatyta, kad metodas yra jautrus įvairaus tipo tinklo srauto nuokrypiams ir gali būti sėkmingai taikomas tinklo srauto anomalijoms aptikti.

Reikšminiai žodžiai: anomalijų atpažinimas, entropija, NetFlow, kompiuterių tinklai.

Įvadas

Kompiuterių tinklai tampa kritine informacinių technologijų infrastruktūros dalimi, todėl ypač aktualu, kad jis funkcionuotų patikimai, saugiai ir stabiliai. Duomenų perdavimo srautas dėl tam tikrų priežasčių gali sutrikti t. y. nukrypti nuo įprastų duomenų perdavimo srauto pasiskirstymo laike ir sukelti duomenų perdavimo anomalijų. Dažnai anomalijos pažeidžia tinklo išteklių vientisumą, konfidencialumą ir pasiekiamumą. Jos sukelia tinklo saugos grėsmių ir mažina teikiamų paslaugų kokybę, todėl svarbu anomalijas laiku identifikuoti ir pritaikyti tinkamą sprendimą joms neutralizuoti.

Anomalijos gali kilti dėl įvairių priežasčių, kurios klasifikuojamos į tris apibendrintas kategorijas, t. y. anomalijos, įvykusios dėl antplūdžio, su piktavališka veikla ir kenksmingu programiniu kodu susijusios anomalijos bei anomalijos, atsirandančios dėl fizinių ar programinių tinklo infrastruktūros problemų (Barford, Plonka 2001; Miluocheva, Muller 2003). Nagrinėjant tinklo anomalijų priežastis, buvo nustatyta, kad 20 % anomalijų atsiranda dėl suplanuotų tinklo tvarkymo darbų (Markopoulou *et al.* 2004). Apie 30 % visų anomalijų atsiranda dėl maršrutizatorių ar perdavimo terpės gedimų. Taigi anomalijų priežastys ir jų poveikio apimtis gali būti įvairios, todėl svarbu atskirti anomalijas, atsiradusias dėl planuotų tinklo priežiūros darbų, nuo anomalijų, kurioms poveikį daro neplanuoti tinklo įrangos gedimai ar piktavališka veikla, bei įvertinti anomalijų poveikio apimtį.

Apibendrinant teigtina, kad tinklo srauto anomalijų atpažinimas – svarbi tinklo stebėjimo sistemos dalis, leidžianti didinti tinklo patikimumą ir saugumą, todėl svarbu taikyti tokius anomalijų atpažinimo metodus, kurie būtų jautrūs tinklo srauto nuokrypiams, užtikrintų aukštą rezultatų patikimumą ir greitaveiką.

Duomenų surinkimo metodai

Kompiuterių tinklo anomalijoms atpažinti reikalingi duomenys, surenkami iš tinklo įrenginių. Duomenų patikimumas – vienas iš svarbesnių kriterijų, veikiančių atpažinimo rezultatų tikslumą. Egzistuoja du pagrindiniai tinklo srauto stebėjimo metodai: aktyvus ir pasyvus.

Aktyvūs metodai dažniausiai taikomi kokybinėms tinklo charakteristikoms nustatyti, pavyzdžiui, vėlinimui ar pralaidumui, tačiau gali būti naudojami ir anomalijoms reikalingiems duomenims surinkti. Aktyvaus matavimo metodų pranašumas tas, kad jiems nereikalinga specializuota techninė įranga, o programinė įranga yra nesudėtinga. Tačiau reikia atkreipti dėmesį į tai, kad taikant aktyvius metodus siunčiami ICMP, TCP ar UDP paketai trikdo tinklo darbą, todėl gauti duomenys bus netikslūs, o anomalijų atpažinimo rezultatai turės didesnę paklaidą.

Pasyvūs metodai tinklo srautams stebėti naudoja specializuotą techninę įrangą, pavyzdžiui, duomenų srautus persiunčiančius maršrutizatorius arba tinklo pasiklausymo

2014 © Straipsnio autoriai. Leidėjas VGTU leidykla „Technika“.
Šis straipsnis yra atvirosios prieigos straipsnis, turintis Kūrybinių bendrijų (*Creative Commons*) licenciją (CC BY-NC 4.0), kuri leidžia neribotą straipsnio ar jo dalių panaudą su privaloma sąlyga nurodyti autorių ir pirminį šaltinį. Straipsnis ar jo dalys negali būti naudojami komerciniams tikslams.

įrenginį, kuris duomenų srautą nukreipia į tinklo stebėjimo stotį. Pasyvūs metodai turi tokių pranašumų:

- stebint tinklo srautus nėra siunčiama papildomų duomenų paketų, kurie sutrikdytų stebimo tinklo darbą (Landfeldt *et al.* 2000);
- surenkama išsami informacija apie tinklo protokolo parametrus ir perduodamos informacijos turinį, kurią galima efektyviai panaudoti anomalijų atpažinimui.

Įvertinus aktyvų ir pasyvių metodų pranašumus ir trūkumus, buvo nuspręsta duomenims rinkti taikyti pasyvųjį metodą, pritaikant NetFlow paketų srauto ilgalaikius duomenis.

NetFlow paketą sudaro septyni reikšminiai laukai: siuntėjo IP adresas, gavėjo IP adresas, siuntėjo prievado numeris, gavėjo prievado numeris, protokolo numeris, TOS reikšmė, priimančios sąsajos SNMP indeksas (Cisco Systems 2007). NetFlow sraute visų paketų reikšminių laukų reikšmės sutampa. Jei maršrutizatorius sukuria paketą, kuriame nors vieno iš reikšminių laukų reikšmė skiriasi, tuomet sukuriamas naujas srautas. NetFlow srautų duomenys dažniausiai naudojami srauto pralaidumui nustatyti ir, srautui viršijus nustatytas ribas, kuriamas įspėjimas. Tinklo stebėjimo sistemos paprastai neanalizuoja srauto pralaidumo pasiskirstymo laike, todėl neišnaudojamos visos NetFlow naudojimo galimybės.

Anomalijų atpažinimo metodai

Kompiuterių tinklo srautas yra labai dinamiškas, todėl aptikti anomaliją, kurią charakterizuoja nedidelis tinklo srauto pokytis, yra sunku. Atlikus tinklo srauto anomalijų atpažinimo metodų analizę buvo išskirtos tokios pagrindinės metodų grupės:

- klasifikavimo metodai;
- klasterizavimo metodai;
- statistiniai metodai;
- nuokrypiu grįsti metodai.

Klasifikavimo metoduose sukuriamas klasifikavimo modelis, kuris apmokomas naudojant specialius mokymui skirtus teisingus ir anomaliją apibūdinančius duomenis. Kitu žingsniu apmokytas modelis taikomas nežinomiems duomenims ar įvykiams klasifikuoti. Klasifikavimo metodai skirstomi į prižiūrimus ir pusiau prižiūrimus. Pirmiesiems reikalingos žinios tiek apie normalias, tiek apie anomalias duomenų klases, o pusiau prižiūrimiems metodams reikia žinių tik apie normalias duomenų klases. Klasifikavimo metodų grupei priklauso taisyklėmis pagrįsti metodai, neuroniniai tinklai, Bajeso metodas, maksimalios entropijos, SVM (angl. *Support Vector Machine*) pagrindu veikiantys metodai. Klasifikavimo metodų pranašumai – tai gebėjimas aptikti dar nežinomas anomalijas, didelis tikslumas, greita klasifikavimo fazė, kai turimas apmokytas modelis.

Klasterizavimo metodų tikslas – sugrupuoti panašius duomenis į klasterius, o nepriklausantys klasteriams duomenys arba labai maži klasteriai laikomi anomalijomis. Klasterizavimo metodai skirstomi į pusiau prižiūrimus, kai iš anksto sudaromi duomenų klasteriai, apibūdinantys normalią sistemos veiklą, ir neprižiūrimus metodus, kai po klasterizavimo reikalingi papildomi žingsniai, įvertinantys klasterių dydžius ir atstumus tarp jų, padedantys surasti anomalijoms priklausančius taškus. Pagrindiniai klasterizavimo metodų trūkumai – tai imlumas skaičiavimo ištekliams, šio tipo metodų neveiksmingumas, jei normalūs taškai nesiburia į klasterius.

Statistiniai metodai remiasi prielaida, kad normalių įvykių duomenys yra stochastinio modelio didelės tikimybės zonoje. Siekiant nustatyti duomenų anomalijas, tikrinama, ar jie priklauso sudarytam modeliui. Duomenys, turintys mažą tikimybės vertę, laikomi anomaliais. Statistiniai metodai yra dviejų tipų: parametriniai ir neparametriniai. Taikant parametrinius metodus, normalūs duomenys ir galimos duomenų anomalijos sugeneruojamos iš pagrindinių parametrinių skirstinių, o parametrai randami mokymo metu. Statistiniai neparametriniai metodai daro prielaidą, kad modelio struktūra nėra žinoma iš anksto ir ji surandama pagal turimus duomenis. Tokio tipo metodai taiko mažiau prielaidų apie duomenis nei parametriniai metodai, todėl yra tikslesni. Pagrindinis statistinių modelių trūkumas tas, kad parametriniai įverčiai dažniausiai neatitinka realių duomenų pasiskirstymo, todėl šių modelių tikslumas labai priklauso nuo pradinio duomenų pasiskirstymo.

Nuokrypiu pagrįsti metodai – tai tokie metodai, kuriuose atliekami vidutinės reikšmės ir nuokrypio nuo jos skaičiavimai. Metodai tarpusavyje skiriasi vidutinės reikšmės skaičiavimo algoritmu. Vienas populiariausių nuokrypių metodų – eksponentinio svorinio slankiojo vidurkio metodas (Roughan *et al.* 2004). Jis atitinka žemojo dažnio signalų filtrą, nes panaikina aukštojo dažnio triukšmus, palikdamas tik bazinę žemojo dažnio dedamąją. Taip pat dažnai naudojamas ir eksponentinio svorinio slankiojo vidurkio skaičiavimo metodas, kuris papildomai įvertina duomenų srautų periodiškumą.

Atlikus anomalijų atpažinimo metodų analizę buvo nuspręsta tinklo srauto anomalijų atpažinimo tyrimus atlikti taikant maksimalios entropijos metodą. Šis metodas yra lankstus, t. y. galima ieškoti anomalijų grupuojant duomenis pagal skirtingus kriterijus, taip pat sugeba atpažinti anomalijas, esant įvairiems tinklo srauto nuokrypiams.

Šiam metodui reikia daug skaičiavimo išteklių, todėl šio tyrimo tikslas – ne tik pritaikyti metodą NetFlow paketų srauto analizei, bet ir modifikuoti jį, siekiant sutrumpinti skaičiavimo laiką, kartu neprarandant tikslumo.

Santykinės entropijos skaičiavimo algoritmas

Maksimalios entropijos metodas atpažįsta anomalijas kompiuterių tinklo sraute, lygindamas srauto pasiskirstymą konkrečiu laiko momentu su tipinio srauto pasiskirstymu tam tikrame laiko intervale. Apskaičiuojant santykinę srautų entropiją, atpažįstamos anomalijos. Maksimali entropija sukuria mažiausiai kintančio pasiskirstymo modelį pagal pateiktus ribojimus. Taikant maksimalios entropijos metodą, naudojami du duomenų masyvai. Pirmąjį sudaro tinklo srauto tikimybiniai duomenys, kai nėra anomalijų. Šis masyvas konstruojamas iš ilgalaikių srauto duomenų ir naudojami modeliui apmokyti. Antrąjį masyvą sudaro planuojami tirti empiriniai srauto duomenys, t. y. atitinkamos srauto tikimybės. Turint šiuos duomenis, galima apskaičiuoti entropiją kaip pasiskirstymo vienodumo matą:

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x), \qquad (1)$$

čia $p(x_i)$ ir $\tilde{p}(x_i)$ – atitinkamai įvykių tikimybės apmokymo modelyje ir empirinių duomenų modelyje; $p(y \mid x)$ – santykinė tikimybė. Entropija aprėžta iš apačios, t. y. neneigiama. Norėdami parinkti tinkamą modelį, siejantį apmokymo ir empirinį pasiskirstymus, ieškosime modelio p^* su maksimalia entropija $H(p)$, t. y. spręsime optimizavimo uždavinį:

$$p^* = \arg\max_p H(p) = \arg\max_p \left(-\sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \right). \qquad (2)$$

Optimizavimo uždavinio ribojimai yra tokie:

$$p(y \mid x) \geq 0 \quad \forall\, x, y; \qquad (3)$$

$$\sum_y p(y \mid x) = 1 \quad \forall\, x; \qquad (4)$$

$$\sum_{x,y} \tilde{p}(x) p(y \mid x) f(x,y) = \sum_{x,y} \tilde{p}(x,y) f(x,y), \qquad (5)$$

čia $f(x,y)$ – indikatoriaus funkcija, priimanti reikšmes 1 arba 0, priklausomai nuo to, įvyko įvykis ar ne.

Norint sukurti optimalų pasiskirstymo modelį, reikia atlikti tokius žingsnius: parinkti indikatorių funkcijų $f_i(x, y)$ aibę taip, kad jos atspindėtų pagrindinius lyginamųjų modelių skirtumus ir apskaičiuoti parametrus λ_i. Tai yra reikšmės, minimizuojančias Kulbako ir Lioblerio divergenciją pasiskirstymo $p(x_i)$ atžvilgiu. Šiuos du žingsnius Yu Gu *et al.* (2005) rekomenduoja atlikti vykdant iteracinį algoritmą, kurį sudaro šie žingsniai: indikatoriaus funkcijų parinkimas, parametrų skaičiavimas, tikslumo tikrinimas. Parametrų skaičiavimo ir tikrinių indikatorių funkcijų parinkimo algoritmas leidžia stebėti entropijos pokyčius pildant empirinių duomenų masyvą naujais įrašais ir koreguojant empirinių duomenų pasiskirstymą. Tačiau tokiam algoritmui reikia daug skaičiavimo išteklių, todėl pasiūlytas kitas parametrų λ_i skaičiavimo būdas. Dažniausiai empirinių duomenų masyvas nėra nuolat pildomas, todėl empirinius duomenis galima klasifikuoti ir suskirstyti juos į kelias lenteles. Indikatoriaus funkcija $f_i(x, y)$ parodys, ar *i*-tojo bandymo metu yra pastebėtas atitinkamas įvykis ir šios funkcijos reikšmė pasikeis tik tuomet, kai bus atrasta duomenų anomalija.

Pirmu žingsniu pradinėms parametrų λ_i reikšmėms priskiriamos nulinės reikšmės:

$$\lambda_i = 0, \; i \in \{1, 2, \ldots, n\}. \qquad (6)$$

Toliau ieškome $\Delta\lambda_i$ reikšmių spręsdami lygčių sistemą:

$$\Delta\lambda_i = \frac{1}{M} \log \frac{\tilde{p}(x_i, y_i)}{\sum_{x_i, y_i} \tilde{p}(x) p(y \mid x) \exp(M \Delta\lambda_i)}, \qquad (7)$$

čia $i \in \{1, 2, \ldots, n\}$, M – indikatorių funkcijų $f(x, y)$ suma, t. y. skirtingų pasiskirstymo kriterijų atžvilgiu stebėtų įvykių skaičių suma. Radus $\Delta\lambda_i$ tikrinama, ar pasiektas reikiamas tikslumas, t. y. jei sąlyga:

$$\max \left| \Delta\lambda_i \right| < \varepsilon \qquad (8)$$

netenkinama, tuomet vykdoma nauja iteracija:

$$\lambda_i = \lambda_i + \Delta\lambda_i. \qquad (9)$$

Suradus λ_i sudaromas optimalus pasiskirstymo modelis:

$$p^*(y_i \mid x_i) = \exp(\lambda_i) \Big/ \sum_i \lambda_i. \qquad (10)$$

Turint optimalų modelį, randamos srautų entropijos ir apskaičiuojama santykinė entropija.

Duomenų klasifikavimas

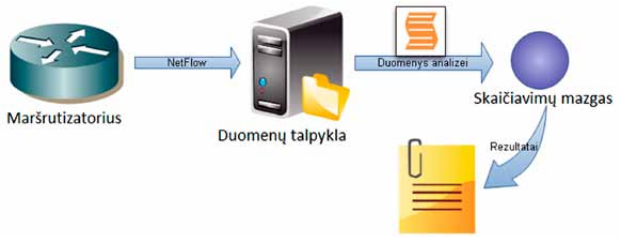
NetFlow paketų srautą sudaro trys pagrindinės paketų grupės: TCP, UDP ir ICMP. Komunikacijoms naudojami prievadai, pasiskirstę intervalu nuo 1 iki 65 535. Analizuojant šiuos duomenis pagal prievadus ir paketų tipus, susidaro didelės apimties duomenų masyvai. Norint pagreitinti apdorojimą ir sumažinti skaičiavimo trukmę, siūloma klasifikuoti paketus pagal prievadus į grupes.

Grupuojant prievadus buvo atsižvelgta į IANA apibrėžtas prievadų grupes. Klasifikuojant paketus, prievadų numeriai, patenkantys į intervalą nuo 1 iki 1023, buvo skirstomi po 10 prievadų į kiekvieną grupę. Kadangi 80-asis prievadas sudaro didžiąją dalį tinklo srauto, šiam prievadui sudaroma atskira grupė. Taigi suskirstę šį prievadų intervalą, gauname 104 grupes. Paketai, kurių prievadai patenka į intervalą nuo 1024 iki 49 151, grupuojami po 100 prievadų į vieną grupę. Iš viso gaunamos papildomos 482 prievadų grupės. Paketai, kurių prievadų numeriai priklauso intervalui nuo 49 152 iki 65 535, priskiriami vienai grupei. Susumavus visas grupes, iš viso gaunamos 587 paketų grupės.

Galima srauto paketus klasifikuoti ne tik pagal prievadus, bet ir pagal siuntėjo bei gavėjo IP adresus, IP adresų ir prievadų poras, tačiau tokie klasifikavimo principai nebuvo nagrinėti.

Srauto anomalijų tyrimo modelis

Siekiant patikrinti maksimalios entropijos metodo tinkamumą tinklo srauto anomalijoms atpažinti, buvo sudarytas eksperimentinio tyrimo stendas. Jį sudarė „Cisco“ maršrutizatorius, duomenų talpykla ir duomenų apdorojimo mazgas (1 pav.).

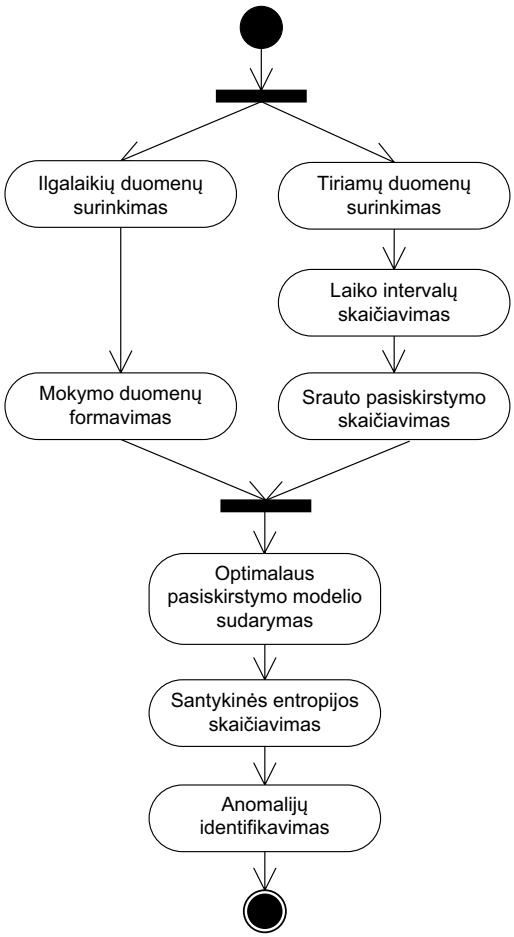


1 pav. Tinklo srauto anomalijų atpažinimo tyrimo schema
Fig. 1. Scheme of network traffic anomalies detection

Maršrutizatoriaus registruojami tinklo srauto duomenys NetFlow formatu siunčiami į nuotolinę duomenų talpyklą. Skaičiavimo mazgas, prieš paimdamas srauto duomenis, juos transformuoja į tekstinį formatą, pašalina netinkamas analizuoti eilutes ir atlieka entropijos skaičiavimus. Gauti rezultatai pateikiami vartotojui arba tinklo stebėjimo sistemai.

Skaičiavimų mazge veikiančiame anomalijos atpažinimo programiniame prototipe realizuotas algoritmas, kurį sudaro tokia veiksmų seka (2 pav.):

1. Ilgalaikių stebėjimo duomenų surinkimas.
2. Modeliui apmokyti skirtų duomenų paruošimas ir modelio apmokymas.
3. Analizuojamų duomenų paruošimas, suskirstant juos pagal laiko intervalus.



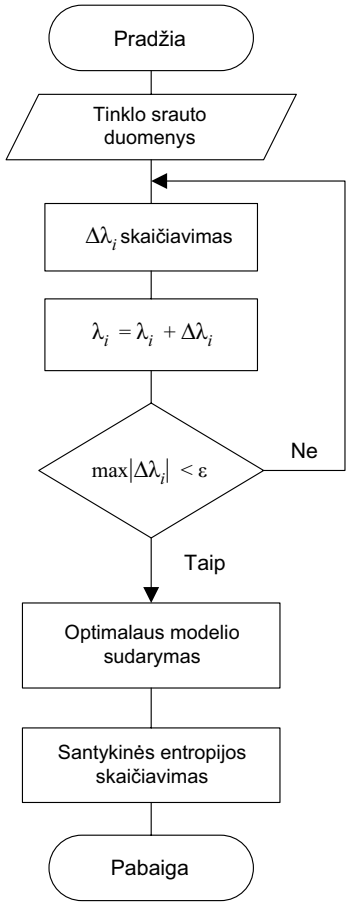
2 pav. Anomalijos atpažinimo algoritmas
Fig. 2. Anomalies detection alogrithm

4. Santykinės entropijos skaičiavimas.
5. Rezultatų analizė ir atvaizdavimas.

Identifikavus didžiausią santykinės entropijos pokytį tam tikrame laiko intervale, atliekamas laiko intervalo skaidymas į mažesnius laiko langus ir vėl kartojamas anomalijos atpažinimo algoritmas.

Santykinė entropija skaičiuojama naudojant iteracinį algoritmą (3 pav.). Nuskaičius nagrinėjamo laiko intervalo srauto duomenis, skaičiuojama λ_i reikšmė. Imamas kiekvienas laiko langas atskirai ir pradedamas iteracijos procesas. Apskaičiuojamas kiekvieno duomenų matricos elemento $\Delta\lambda_i$ ir koreguojamos λ_i reikšmės. Jei užsibrėžtas tikslumas pasiektas, tuomet sudaromas optimalus modelis.

Tikrinant modelio eksperimentą pastebėta, kad, turint žemo tikslumo λ_i reikšmes, modelis jau perteikia bendrą pasiskirstymą ir, skaičiuojant galutinę eksperimento srauto entropiją, jos reikšmės daro nedidelę įtaką.



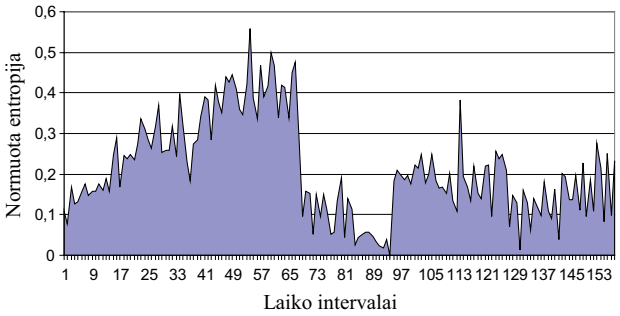
3 pav. Santykinės entropijos skaičiavimo algoritmas
Fig. 3. Relative entropy calculation algorithm

Nustatyta, kad naudojant skaičiavimų tikslumą 0,5 (iteracijų skaičius 20) ir tikslumą 0,001 (iteracijų skaičius 2000) galutinė entropijos vertė gaunama tokia pati, nors tarpinės λ_i reikšmės skiriasi. Todėl galima teigti, kad, norėdami trumpinti anomalijos atpažinimo skaičiavimų laiką, galime mažinti λ_i tikslumą ir atitinkamai iteracijų skaičių, nes entropijos reikšmės iš esmės nusistovi, kai λ_i tikslumas artimas 0,5.

Eksperimentinis tyrimas

Eksperimentiniam tyrimui naudoti „Cisco“ maršrutizatoriaus tinklo srautų įrašai. Mokymui buvo skirti visi vienos dienos duomenys. Tiriamieji duomenys buvo kitos dienos srauto duomenys, suskaidyti į laiko intervalus po 10 minučių. Gauti 174 laiko intervalai. Vidutiniškai kiekvienam laiko intervalui apdoroti prireikė 15 iteracijų, kai tikslumas 0,5. Tačiau padidinus tikslumą iki 0,1, reikiamų iteracijų skaičius išauga iki 2000, todėl pailgėja ir skaičiavimo trukmė. Kai λ_i tikslumas yra 0,5, kiekviena anomalijos atpažinimo algoritmo iteracija užtrunka 1 s, o visos iteracijos užbaigiamos per 15 s. Tiriant naudoti 174 laiko intervalai, todėl bendra srauto apdorojimo trukmė buvo 2610 s. Santykinės entropijos skaičiavimai buvo atliekami asmeniniu kompiuteriu, turinčiu dviejų branduolių Intel Pentium 4, 3 GHz dažnio procesorių ir 4 GB operatyviosios atminties.

4 pav. pateiktas normuotas santykinės entropijos kitimo grafikas. Jame matyti, kad laiko intervalu nuo 1 iki 15 ir nuo 67 iki 98 normuotos santykinės entropijos reikšmės labai kinta. Pažymėtina, kad intervaluose santykinės entropijos reikšmės kinta kitaip. Jis susijęs su staigiu srauto sumažėjimu, kuris buvo tą dieną įvykęs dėl ryšio kanalo planinio remonto darbų. Tuo remiantis galima teigti, kad pateiktas maksimalios entropijos metodas yra jautrus ir fiksuoja tiek staigius, tiek lėtus tinklo srauto pokyčius. Siekiant išvengti klaidingo anomalijos atpažinimo, reikėtų įvesti slenkstines normuotos santykinės anomalijos reikšmes, kurias viršijus būtų laikoma, kad įvyko anomalija. Stebint entropijos pasikeitimus skirtingais aspektais, taip pat galima tiksliau nustatyti anomaliją veikiančius veiksnius ir identifikuoti anomalijos tipą.



4 pav. Normuotos santykinės entropijos kitimas, apdorojant banduomosius NetFlow duomenis
Fig. 4. Variation of rational relative entropy when testing NetFlow data are used

Išvados

1. Tinklo anomalijų atpažinimas – svarus aspektas didinant tinklo patikimumą ir saugumą.
2. Apžvelgus anomalijų atpažinimo metodus galima teigti, kad didelio tikslumo metodas reikalauja daug kompiuterinių išteklių. Greiti anomalijų atpažinimo metodai sunkiai geba aptikti nežinomas anomalijas ir duoda dideles paklaidas.
3. Tyrimui atlikti buvo pritaikytas ir patobulintas maksimalios entropijos metodas. Atliktos metodo modifikacijos, atsižvelgiant į tinklo srauto duomenų ypatumus, leidžia supaprastinti λ_i koeficientų skaičiavimą ir nagrinėti kompiuterių tinklo srautą nedideliais paketais, todėl skaičiavimo trukmė labai sumažėja ir entropijos pasikeitimus galima stebėti beveik realiuoju laiku.

4. Atlikus skaičiavimų eksperimentus su maršrutizatoriaus tinklo srauto duomenimis nustatyta, kad metodas yra jautrus netgi nežymiam tinklo srauto pasikeitimui.

Literatūra

Barford, P.; Plonka, D. 2001. Characteristics of network traffic flow anomalies, in *Proc. of ACM SIGCOMM Internet Measurement Workshop*, 1–2 November, 2001, Burlingame, 69–73.

Miluocheva, I.; Muller, E. 2003. A practical approach to forecast Quality of Service parameters considering outliers, in *1st Int. Workshop on Inter-Domain Performance and Simulation,* 21–21 February, 2003, Salzburg, Austria, 163–172.

Markopoulou, A.; Iannaccone, G.; Bhattacharyya, S.; Chuah, C.; Diot, C. 2004. Characterization of Failures in an IP Backbone, in *IEEE INFOCOM*, 7–11 March, 2004, Hong Kong. http://dx.doi.org/10.1109/INFCOM.2004.1354653

Landfeldt, B.; Sookavatana, P.; Seneviratne, A. 2000. The case for a hybrid passive/active network monitoring scheme in the wirel, in *8th IEEE International Conference on Networks*, 5–8 September, 2000, 139–147. http://dx.doi.org/10.1109/ICON.2000.875781

Cisco Systems. 2007. *NetFlow Services Solutions Guide* [interaktyvus], [žiūrėta 2014 m. balandžio 2 d.]. Prieiga per internetą: http://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/netflow/nfwhite.html

Roughan, M.; Griffiny, T.; Mao, M.; Greenbergx, A.; Freeman, B. 2004. IP forwarding anomalies and improving their detection using multiple data sources, in *ACM SIGCOMM workshop on Network Troubleshooting*, 30 August – 03 September, 2004, Portland, 307–312.

Yu Gu, A.; McCallum, A.; Towsley, D. 2005. Detecting anomalies in network traffic using maximum entropy estimation, in *Proceeding of the 5th ACM SIGCOMM conference on Internet Measurement,* 19–21 October, 2005, USENIX Association, Berkeley, 345–350.

NETWORK TRAFFIC ANOMALIES DETECTING USING MAXIMUM ENTROPY METHOD

D. Mažeika, S. Jasonis

Abstract

The problem of traffic anomalies in computer networks is analyzed. NetFlow packets are used as network traffic data and maximum entropy methods is used for anomalies detection. Method detects network anomalies by comparing the current network traffic against a baseline distribution. Method is adopted according to NetFow data and performace of the method is improved. Prototype of anomalies detection system was developed and experimental investigation carried out. Results of investigation confirmed that method is sensitive to deviations of the network traffic and can be successfully used for network traffic anomalies detection.

Keywords: anomalies detection, entropy, NetFlow, computer network.