Şəbəkə trafikinin klasterizasiya metodu haqqında

Babək Nəbiyev

AMEA İnformasiyaTexnologiyaları İnstitutu

babek@iit.ab.az

***Xülasə*— Kompüter şəbəkələrinin təhlükəsizliyinin təmin olunması və prosesin optimallaşdırılması üçün bir çox vasitələr mövcuddur. Məlumdur ki, təhlükələrin yaranmasının əsas səbəblərindən biri şəbəkə trafikində anomal və qeyri profil trafikin generasiya olunmasıdır. Bunları nəzərə alaraq, məqalədə şəbəkə trafikində davranış profilinin müəyyən olunması üçün şəbəkə trafikinin klasterizasiya metodu təklif olunur. Davranış profilinin müəyyən olunması üçün *K*-ortalar klasterizasiya metodu tətbiq olunur.**

***Açar sözlər*— *şəbəkə trafiki; klasterizasiya; davranış profili; anomal trafik*.**

I. GİRİŞ

İnternet vasitəsilə qloballaşan dünyada hər bir resurs və ya informasiyanı sürətlə əldə etmək çox asan olumuşdur. Bu informasiya cəmiyyəti baxımından çox müsbət haldır. Amma generasiya olunan informasiyanın heç də hamısı lazımlı və məqsədəuyğun olmur. Bu isə öz növbəsində kompüter şəbəkələrində lazımsız yük yaradaraq əlaqə kanallarının əlyetənlik qabiliyyətini aşağı salır. Bu hadisə şəbəkədən istifadə prosesini davranış profilinə uyğunlaşdırmayan korporativ şəbəkələrin geci-tezi qarşılaşa biləcəyi hadisələrdəndir.

Symantec şirkətinin 2014-ci ildə verdiyi hesabata əsasən, bir gün ərzində veb-resurslarda qarşısı alınan hücumların sayı 586700-ə bərabərdir [1]. Bunları nəzərə alaraq, şəbəkə istifadəçilərinin korporativ resurslardan düzgün istifadə edərək təhdidlərlə qarşılaşmaması, məhdud olan informasiya kanalını yersiz yükləməməsi və faydalı iş qabiliyyətinin yüksəldilməsi üçün şəbəkə trafikinin klasterizasiyası metodu əsasında şəbəkə trafikində davranış profilinin (bundan sonra, davranış profili) müəyyən olunması təklif olunur. Şəbəkə trafikinin monitorinq vasitəsilə alınan verilənləri klasterizasiya dəyərləri əsasında analizi aparılaraq müyyən trafikin davranış klasterləri əldə oluna bilər və bu prosesin reallaşdırılması *K*-ortalar klasterizasiya alqoritmi vasitəsilə həyata keçirilir.

II. ƏLAQƏDAR TƏDQİQATLARIN ANALİZİ

Şəbəkə trafikinin identifikasiyası və kateqoriya-laşdırılması şəbəkənin idarə olunmasının əsas elemetlərindən biridir, buna axının prioritetləşdirilməsi, trafikin formalaşdırılması və siyasəti, diaqnostik monitorinqi misal gətirmək olar.

Bütün dünyada IP-şəbəkələr vasitəsilə böyük həcmli informasiya ötürülür və qəbul olunur. Mütəxəssislər bütün bu prosesi nəzarətdə saxlayır və bunun vasitəsilə təhdidləri müəyyən edərək aradan qaldırırlar. IP-paket başlığıda daxil olan funksiya və parametrlər şəbəkə və istifadəçilər haqqında bir çox informasiya əldə etməyə imkan verir. Bundan əlavə IP-paketlərin başlıqlarını analiz edərək şəbəkənin idarə olunması və optimimallaşdırılma, təhdidlərin aradan qaldırılmasında, yeni servislərin yaradılmasında istifadə etmək mümkündür. [2]-də IP-paketlərin başlıqlarından istifadə edərək şəbəkənin axın prosesini və istifadəçilərin davranış profilini geniş formada izah edən çoxsəviyyəli klasterizasiya metodu təklif olunur. Əlavə olaraq demək lazımdır ki, IP-paketlərin başlıqlarından istifadə edərək aparılan analiz prosesi istifadəçilərin şəxsi məlumat toxunulmazlığını təmin edir.

Şəbəkə trafiki və ya ümumiyyətlə şəbəkə haqqında toplanmış loqlar vasitəsilə anomaliyaların və təhdilərin aşkarlanması üçün istifadə etmək olar. Bu proses üçün müxtəlif metodlardan vasitələrdən istifadə olunur. Məsələn, [3]-də *K*-ortalar klasterləşmə alqoritmindən istifadə edərək trafik axınında anomaliyaları aşkarlamaq təklif olunur. Şəbəkə trafikinin işarələnməmiş verilənləri iki klasterə bölünür, bunlar normal və anomal trafiklərdir. Yeni monitorinq verilənlərində anomaliyaların aşkarlanması əsasında effektiv məsafənin seçilməsi üçün müəyyən olunmuş klasterlərdə şablon olaraq ağırlıq mərkəzi istifadə olunur.

Mərkəzi idarəetmə olmadan özü təşkilatlanan və nəzarət prosesi olmayan klasterləşmə metodu ən yeni yaxınlaşmalardan biridir. Bunu üçün [4]-də qarşılıqlı əlaqəyə əsaslanan qarışqa davranışı metodu istifadə olunur. Bu metodun üstünlüyü ondan ibarətdir ki, ilkin verilənlərə və ya klasterlərin sayının əvvəlcədən müəyyən olunmasına ehtiyac yoxdur. Virtual qarışqalar hər biri ayrı-ayrılıqda şəbəkəni təqdqiq edərək klasterləmə prosesini yerinə yetirirlər. Amma bu yeni metod olduğuna görə aparılan prosesin dəqiqlik əmsalı şübhə doğurur.

"Machine learning" yanaşması şəbəkə trafikində anomal axınların unikal statistik xarakteristikalara əsaslanaraq müəyyən olunması üçün geniş istifadə olunur. Qeyri-səlis klasterləşdirmə ənənəvi klasterləşdirməyə nəzərən daha çevikdir, müdaxilələrin aşkarlanması və verilənlərin təbii emalı üçün daha məqsədəuyğundur [5].

Bir çox klasterləşdirmə metodları müdaxilələrin aşkarlanması üçün normal və anomal trafikin ayrılmasını nəzərdə tutur. Klasterləşdirmə metodları trafik sessiyalarının fərqlərini və oxşarlıqlarını tapmaq, onların hər birini müvafiq qruplara bölərək təsnif etmək üçün tətbiq edilir [6]. Bu qruplar onlara verilmiş nişanlar ilə təmsil olunur. Daha sonra bu nişanlar daxil olan şəbəkə trafikinin növünü proqnozlaşdırmaq üçün istifadə olunur.

Şəbəkə trafikinin tez və dəqiq indentifikasiyası QoS-un idarə edilməsi, şəbəkə təhlükəsizliyinin monitorinqi və s. funksiyalar üçün ən vacib məsələlərdən biridir. Lakin son zamanlar P2P-dən istifadə edən qovşaqlar çoxalıb və onlar

müxtəlif portlardan istifadə edərək özünü hər hansı qurğu, lazımlı məlumat axını və ya sifrlənmiş məlumat axını altında gizlədərək lazımsız informasiya axınını generasiya edirlər. Bu halda klassik yanaşmalar sayılan "port mapping" və ya "payload analysis" yanaşmalarının istifadəsi effektiv deyil. Alternativ yanaşma şəbəkədə TCP trafiki ilə əlaqədar ilk bir neçə paket daxilində davranışı tədqiq edərək klassifikasiya etməkdir. Bu gələcəkdə bütün informasiyanı klasterləşdirərək identifikasiya prosesini asanlaşdırmaq üçün istifadə etməyə imkan verərdi [7].

III. K-ORTALAR ALQORİTMİ

Şəbəkə trafikinin klasterizasiyası üçün biz *K*-ortalar alqoritmindən istifadə edəcəyik. Buna səbəb *K*-ortalar alqoritmin klasterizasiya məsələsini həll etmək üçün çox sürətli və sadə olmasıdır. Tutaq ki, $X = \{x_1, \dots, x_n\}$ verilənlər toplusu n trafik sessiyalarından ibarətdir. Hər bir trafik-sessiyası d-ölçülü Evklid fəzasında "verilənlər nöqtəsi"-dir. $x_i = (f_1, \dots, f_d)$, i trafik-sessiyası üçün $f_1, \dots, f_d$ olduğu halda d əlamətlərin qiymətidir. Əsas məqsəd trafik-sessiyalarını klasterlərə bölməkdir. Bu proses zamanı *n* "verilənlər nöqtəsi" ilə müvafiq K klaster "centroid"-ləri arasındakı məsafə minimum olmalıdır. Hər bir klasterin centroid adlanan mərkəzi μ_k var və bu qrupun təmsilçisi kimi hesab edilə bilər.

Belə ki, *K*-ortalar alqoritminin girişi *nxd* verilənlər matrisi, *K* klasterlərin sayı və ilkin verilənlər isə centroidlərdir. Alqoritmin aşağıdakı mərhələləri var:

1. İlk olaraq centroid qruplarını təmsil edəcək *K* nöqtələr müəyyən olunmalıdır.
2. Hər bir verilənlər nöqtəsi ilə ən yaxın centroid arasında Evklid məsafəsini hesablamaq üçün (1) tənliyindən istifadə olunur:

$$dist(x, y) = \left(\sum_{i=1}^{n} (x_i - z_i)^2\right)^{\frac{1}{2}} \quad (1)$$

3. Bütün nöqtələr müəyyən olunduqdan sonra, *K* centroidlərin mövqeləri yenidən hesablanır, bu o deməkdir ki, müəyyən qrupda bütün nöqtələrin $\mu_k$ortası yenidən təyin olunur.
4. 2-ci və 3-cü bənd centroidlər mövqeyin dəyişməyənədək təkrarlanmalıdır.

IV. KLASTERLƏRİN SAYININ SEÇİLMƏSİ

Bu bölmədə *K*-ortalar alqoritmini tətbiq etməzdən qabaq klasterlərin sayını necə seçdiyimiz izah olunacaq. Birinci, nöqtə və cetroid arasındakı məsafəni müəyyən edən klaster daxili məsafə ölçülür. Bundan sonra bütün bu məsələrin ortalaması müəyyən olunur (2):

$$intra = \frac{1}{N} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - x_i\|^2 \quad (2)$$

buradan, *N* sessiyaların (nöqtələrin) sayı, K klasterlərin sayı, $z_i$isə C_i klasterin centroididir. Sonra isə klasterlər arası məsafə ölçülməlidir və onlar bir birlərindən nə qədər uzaq olsa o qədər yaxşıdır. Bunun üçün düstur (3) istifadə olunur:

$$inter = \min\left(\|z_i - z_j\|^2\right), i = 1,2, \dots, K-1; \quad j = i+1, \dots, K \quad (3)$$

K-ortalar üçün klasterlərin *K* sayını müəyyən etmək məqsədilə aşağıdakı (4) düsturdan istifadə etmək lazımdır:

$$validity = \frac{intra}{inter} \quad (4)$$

V. VERİLƏNLƏRİN TƏSVİRİ VƏ EMALI

Verilənlər 5000-dən çox IP ünvandan ibarət olan real şəbəkə mühitində toplanıb və bu şəbəkədə özlüyündə bir neçə xırda şəbəkəyə bölünür. İstifadəçilərin konfedensiallığının pozulmaması məqsədilə ilə AzScienceNet şəbəkəsinin istifadəçi siyasətinə əsaslanmış və əlavə olaraq bütün verilənlər adsızlaşdırılmışdır. Bu verilənlər 7 dəyişəndən ibarətdir, bunlar cədvəl 1-də göstərilmişdir. Bu dəyişənlərin nəzərə alınması ilə aparılan klasterizasiya prosesi zamanı ən dolğun nəticə klaster daxili məsafənin kiçik, klaster xarici məsafənin isə böyük olduğu halda əldə ediləcək. Əsas məqsəd isə AzScienceNET şəbəkəsinin davraniş profilinin müəyyən edilməsidir.

CƏDVƏL 1. Klasterizasiya dəyişənlərinin təsviri

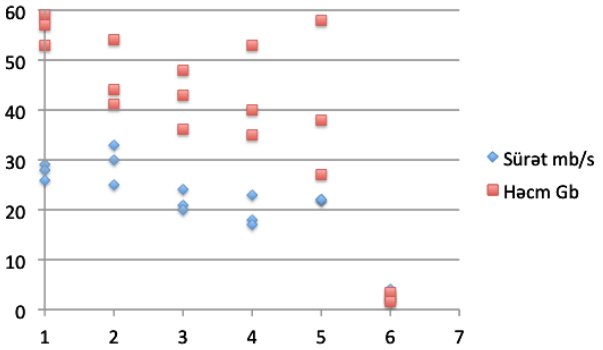
İndeks	**Dəyişənlərin təsviri**
1	Zaman nişanı
2	İP ünvan
3	Kontentin həcmi
4	Veb-resurslar
5	Port nömrəsi
6	Protokol tipi
7	Müraciətlərin sayı

Cədvəl 1-də göstərilən 7 dəyişəni aşağıdakı kimi izah etmək olar:

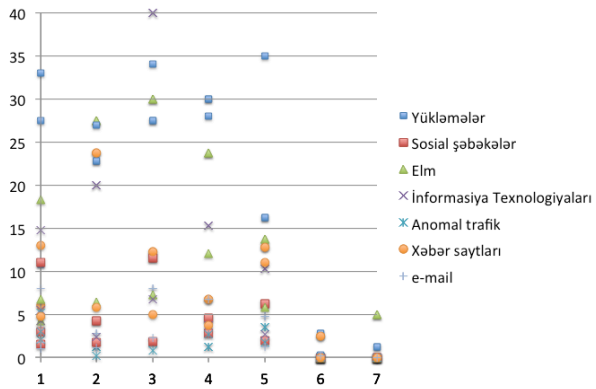
1. Zaman nişanı – bu ümümiyyətlə informasiya texnologiyaları sahəsində hər hansı bir informasiya növünün onun yaranma, ləgv olma, göndərilmə və ya qəbul olunma vaxtını qeyd etmək üçün simvollar və ya şifrələnmiş informasiya ardıcıllığıdır.
2. IP-ünvan – informasiya üçün müraciət edənin və müraciətə cavab verən resursların IP-ünvanları qeyd olunur.
3. Kontentin həcmi – Göndərilən və qəbul olunan bütün paketlərin kontent həcmi qeydə alınaraq ümumi trafikin həcmini müəyyən etmək üçün vacibdir.
4. Veb-resurslar – şəbəkə istifadəçiləri tərəfindən müraciət olunan ver resursların birinci səviyyəli domen adları, domen adları və daxil olduqları liklər qeyd olunur.
5. Port nömrəsi – kompüter şəbəkələrində portlar IP-ünvanlarla və protokollarla əlaqəlidir. Port hər bir ünvan və protokol üçün 16-bit rəqəm kimi identifikasiya olunur. Port nömrəsini bilməklə trafikin növünü müəyyən etmək olur.

6. Protokol tipi – protokollar resurslar arasında qarşılıqlı əlaqə qaydalarını müəyyən edir. Kompüter şəbəkələrində protokollar informasiyanın paketlər formasında göndərilməsi və qəbul olunması üçün kommutasiya metodu kimi istifadə olunur. Protokolların tipinin müəyyən olunması ilə trafikin xüsusiyyətlərinin də müəyyən olunması mümkün olur.
7. Müraciətlərin sayı – istifadəçilərin hansı veb-resursa və nə qədər klik etdiyi nəzərdə tutulur. Bu istifadəçinin hansı veb resursda nə qədər vaxt keçirdiyini müəyyən etmək üçün istifadə oluna bilər.

Bu dəyişənlərin nəzərə alınması ilə AzScienceNet-də bir lokal şəbəkə üzrə (şərti olaraq, A istiqaməti) şəkil 1 və şəkil 2-də göstərilən mənzərə yaranmışdır.



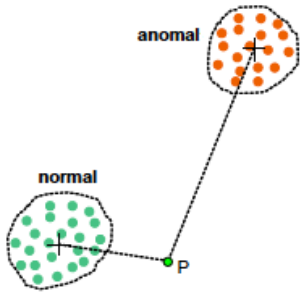
Şəkil 1. A istiqaməti üzrə trafikin sürətinin və həcminin müqayisəsi



Şəkil 2. A istiqaməti üzrə şəbəkə trafikinin profili

Toplanan loq-faylların analizindən sonra, *K*-ortalar alqoritmi əsasında anomal trafikin müyyən olunması prosesi yerinə yetirilir. Bu proses şəbəkə istiqamətləri və ya ümumi olaraq bütün şəbəkəyə tətbiq oluna bilər.

Evklid məsafəsinə əsaslanaraq, klasterlərin centroidləri arasındakı məsafələr hesablanır. Əgər trafik normal sayılırsa, o normal, əgər anomal sayılırsa, anomal klasterə yaxın olur. Bu proses şəkil 3-də ikiölçülü fəzada göstərilmişdir: *P* obyekti normal klasterə yaxın olduğu üçün normal trafik sayılır. Məsafəyə əsaslanan klassifikasiya müəyyən xarakteristikalar əsasında anomal trafiki müəyyən etməyə kömək edir.



Şəkil 3. Normal və anomal klasterlərə misal

NƏTİCƏ

Bu məqalədə, AzScienceNet şəbəkəsində trafik verilənlərinin klasterizasiyası əsasında davranış profilinin müəyyən olunması üsulu işlənib hazırlanmışdır. Bunun üçün klasterizasiya metodları arasında ən sürətli və sadə model olan *K*-ortalar alqoritmi seçilmişdir. Bu prosesin əsas məqsədi şəbəkə resurslarının məqsədəuyğun paylanması, şəbəkə trafikinin optimallaşdırılması, anomal aktivliyin mənbəyinin müəyyən olunması və təhdidlərin vaxtında aradan qaldırılmasını təmin etməkdir.

ƏDƏBİYYAT

[1] http://www.itu.int/en/ITUD/Cybersecurity/Documents/Symantec_annual_internet_threat_report_ITU2014.pdf

[2] P. Kumpulainen, K. Hätönen, O. Knuuti, T. Alapaholuoma, “Internet traffic clustering using packet header information,” Joint Int-l IMEKO TC1+ TC7+ TC13 Symposium, 2011, pp. 13-20.

[3] M.Gerhard, L. Sa, C. Georg, “Traffic anomaly detection using K-means clustering,” Proceedings of performance, reliability and dependability evaluation of communication networks and distributed systems, 4GI/ITG-Workshop MMBnet, 2007, pp. 25-33.

[4] T. Ekola, M. Laurikkala, T. Lehto, H. Koivisto, “Network traffic analysis using clustering ants,” Proceedings World Automation Congress, vol. 17, 2004, pp. 275-280.

[5] D. Liu, C-H. Lung, I. Lambadańs, N. Seddigh, “Network traffic anomaly detection using clustering techniques and performance comparison,” Proc. the 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2013, pp.1-4.

[6] R. Shokri, F. Oroumchian, N. Yazdani, “CluSID: a clustering scheme for intrusion detection improved by information theory,” Proc. of the 7th IEEE Malaysia International Conference on Communications and IEEE International Conference on Networks, 2005, pp.553-558.