Veb İstifadəçilərin Profilinin Qurulması Metodu

Babək Nəbiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
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 şəbəkə trafikində istifadəçilərin profilinin müəyyən olunması üçün vasitə işlənib hazırlanmışdır. İstifadəçilərin profilinin müəyyən olunması üçün K-orta klasterizasiya metodu tətbiq olunmuşdur.**

Açar sözlər – ***şəbəkə trafiki, klasterizasiya, istifadəçi profili, anomal trafik***

I. GİRİŞ

İnternet vasitəsilə qloballaşan dünyada hər bir resurs və ya informasiyanı sürətlə əldə etmək çox asan olmuşdur. Bu informasiya cəmiyyəti nöqteyi nəzərindən çox müsbət haldır. Amma bildiyimiz kimi generasiya olunan informasiyaların heç də hamısı 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 gec-tez qarşılaşa biləcəyi hadisələrdəndir.

Symantec şirkətinin 2014-ci ildə verdiyi hesabata əsasən [1], bir gün ərzində veb-resurslarda qarşısı alınan hücumların sayı 586700 – ə bərabərdir. 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 klasterizasiya metodu əsasında şəbəkə trafikində istifadəçilərin davranış profilinin (bundan sonra istifadəçi profili) müəyyən olunması təklif olunur. Şəbəkə trafikinin monitorinqi vasitəsilə alınan verilənlərin klasterizasiya dəyərləri əsasında analizi aparılaraq müyyən istifadəçinin davranış klasterləri və bu prosesin reallaşdırılması K-Orta klasterizasiya alqoritmi vasitəsilə həyata keçirilir.

II. ƏLAQƏDAR TƏDQİQATLARIN ANALİZİ

Şəbəkə trafikinin identifikasiyası və kateqoriyalaşdırılması şəbəkə idarəolunmasının əsas elemetlərindən biridir, buna axının prioritetləşdirilməsi, trafikin formalaşdırılması və siyasəti, diaqnostik monitorinqi misal gətirmək olar.

Bütün dünyada IP şəbəkələr vasitəsilə böyük həcmli informasiyalar ötürülür və qəbul olunur. Mütəxəssislər bütün bu prosesi nəzarətdə saxlayır və bunun vasitəsilə təhdidləri müəyyən edərək aradan qaldırırlar. IP paketlər istifadəçilər haqqında bir çox informasiyanı əldə etməyə imkan verir. IP paketləri analizin edərək şəbəkənin idarəolunması və optimimallaşdırılması, təhdidlərin aradan qaldırılması zamanı informasiya mənbəyi kimi 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 loq-fayllar vasitəsilə anomaliyaların və təhdilərin aşkarlanmasını həyata keçirmək olar. Bu proses üçün müxtəlif metodlardan, vasitələrdən istifadə olunur. Məsələn, [3]-də K-Orta 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. Bunun 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 tədqiq 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. VEB LOQ FAYLLAR

Verilənlər 5000-dən çox IP ünvandan ibarət olan AzScienceNet şəbəkə mühitində toplanıb və bu şəbəkə də özlüyündə bir neçə xırda şəbəkəyə bölünür. İstifadəçilərin məxfiliyinin pozulmaması məqsədilə ilə AzScienceNet şəbəkəsinin istifadəçi siyasətinə əsaslanmış və əlavə olaraq istifadəçilər kimliyi haqqında verilənlər adsızlaşdırılmışdır. Bu verilənlər 10 dəyişəndən ibarətdir [7], bunlar cədvəl 1-də göstərilmişdir.

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

1. Unix zaman damğası – bu Squid tərəfindən əməliyyatın qeyd olunduğu vaxtdır və bir qayda olaraq əməliyyatın həyat dövrünün bitdiyi vaxtı yəni, müraciət olunan informasiya təmamilə əldə olunduğu vaxtı qeyd edir.
2. Əməliyyat zamanı - əməliyyatda keşdə nə qədər zaman keçirdiyini qeyd edir. Yəni, HTTP paketlərin əməliyyat başlayandan son bayt ötürülənə qədər olan zaman kəsiyi. Unix zaman damğasının qeyd etdiyi vaxtdan əməliyyat zamanını çıxdığınız halda əməliyyatın başlama anını əldə etmək mümkündür.
3. Lokal IP ünvan – informasiya üçün müraciət edən IP ünvanlar qeyd olunur.
4. Nəticə kodları – özündə müraciətlərin cavablandırılması, imtina olunması və s. haqqında məlumat toplayır.
5. Həcm (bayt) - Göndərilən və qəbul olunan bütün paketlərin kontent həcmi qeydə alınır. Bu ümumi trafikin həcmini müəyyən etmək üçün vacibdir.
6. Sorğu metodu – adətən böyük hərflərlə yazılmış kiçik ingilis sözlərindən ibarət olur GET, HEAD və s. Bu metodlar əsasında veb resurs istifadəçinin nə üçün sorğu göndərdiyini müəyyən edir və ona uyğun cavab verir.
7. URL (Uniform Resource Locator) – şəbəkə istifadəçiləri tərəfindən müraciət olunan URL liklər qeyd olunur.
8. Ierarxiya kodu – müraciətlərin emal formaları haqqında məlumat verir. Məsələn müraciət bir başa cavablandırıldı və ya partnyor serverə göndərildi və s.
9. Təyinat IP-si – müraciətə cavab verən IP ünvan.
10. Məzmun – HTTP cavabın başlığında yerləşir və obyektin məzmun növünü göstərir.

İndeks	Dəyişənlərin təsviri
1	Unix zaman damğası
2	Əməliyyat zamanı (ms)
3	Local IP
4	Nəticə kodları
5	Həcm (bayt)
6	Sorğu metodu
7	URL
8	İerarxiya kodu
9	Təyinatlı IP
10	Məzmun

Cədvəl 1. Squid loq dəyişənlərinin təsviri

Bütün bu məlumatlar Squid proxy server vasitəsilə toplanır. Squid proxy server [8] - şəbəkə trafikinin loq fayllarının toplanması və idarə olunması prosesini reallaşdırmaq üçün istifadə edilir. Squid proxy server açıq kodlu proqram təminatıdır və bir gün ərzində internetlə işləyən istifadəçilərin sayı 2000-dən çox olan böyük şəbəkələrdə istifadə olunması əlverişlidir. Squid proxy serverin əsas üstünlüyü keşlənən proxy server olmasıdır, bu halda müraciət olunan resurslar keşdə toplanır və onlara yenidən müraciət olunduğu halda emal prosesi nisbətən daha sürətlə sona çatır. Bu da öz növbəsində şəbəkənin əlyetənliyinə müsbət təsir göstərir. Squid proxy server vasitəsilə toplanan loq fayllar xüsusi verilənlər bazasında toplanaraq analiz prosesində istifadə olunur.

IV. LOQ MƏLUMATLARIN TƏMİZLƏNMƏSİ

Squid proxy server vasitəsilə toplanan loq fayllar müxtəlif interpretasiyalar üçün geniş imkanlar yaradır. Bu da öz növbəsində bir növ loq fayldan müxtəlif məqsədlər üçün istifadə edilməsinə şərait yaradır. Cədvəl 2-də squid loq serverin topladığı verilənlər göstərilmişdir. Amma konkret olaraq bu məqalə çərçivəsində squid proxy serverin təqdim etdiyi 10 dəyişənin hamısına ehtiyac yoxdur. İnformasiya təhlükəsizliyi yanaşmasında istifadəçi profilinin identifikasiyası üçün müraciətin məzmunu, təyinat ip-si, http ierarxik kodu, sorğu metodu və nəticə kodları kimi məlumatlar gərəkli deyil. Ona görədə loq-faylların analizi zamanı emal prosesinin asanlaşdırılması və sürətləndirilməsi üçün bu dəyişənlər nəzərə alınmır.

V. İSTIFADƏÇİ PROFİLİNİN İDENTİFİKASİYASI

İstifadəçinin profili dedikdə - müraciət olunan veb resurslar əsasında tərtib olunan maraqlar vektoru və tematik seçimlər nəzərdə tutulur istifadəçilərin tematik profillər toplusu matris yaradır. Bu matrisdə hər bir sətir istifadəçi, hər bir sütun isə əlamətləri göstərir.

UNIX zaman damğası	Əməliyyat zamanı (ms)	Local IP	Nəticə kodları	Həcm (bayt)	Sorğu metodu	URL	İyerarxiya kodu	Təyinat İP-si	Məzmun
1444780867.298	39	10.100.80.51	TCP_MISS/200	10946	GET	http://pagead2.googlesyndication.com	HIER_DIRECT	216.58.208.98	application/x-shockwave-flash
1444795608.042	3598	10.100.80.23	TCP_MISS/301	567	POST	http://v.icecentury.com/	HIER_DIRECT	54.169.165.185	text/html
1444795738.177	222	10.100.80.14	TCP_MISS/304	318	GET	http://code.createjs.com	HIER_DIRECT	23.77.228.124	application/x-javascript
1444799392.183	38	10.100.80.61	TCP_MISS/200	345	HEAD	http://ds.download.windowsupdate.com	HIER_DIREC	188.43.72.35	application/octet-stream

Cədvəl 2. Squid loq serverin topladığı verilənlər

İstifadəçilərin davranış kateqoriyalarına daxil olan resurslara müraciətlərin tezliyindən və daxil olan trafikin həcmindən asılı olaraq əlamətlərin qiyməti hesablanır. Modelin keyfiyyətinin yüksəldilməsi üçün xüsusiyyətlərin normallaşdırılması prosesi aparılaraq [0;1] aralığına gətirilir.

Əlamətlərin lahiyələndirmə prosesi bitdikdən sonra, modelin qurulması üçün daha informativ və dolğun əlamətlər seçilir. Bu emal olunan məlumatların həcmini azaldır, yenidən təlim prosesindən keçməməsinə şərait yaradır və ümumiyyətlə, modelin keyfiyyətini yüksəldir. Baxılan halda resurlar tematik kateqoriyalara uyğun qruplaşdırılır. Məlumdur ki, bir tematik kateqoriyaya aid ola bilən resurslar müxtəlif mənbələrdə yerləşə bilər.

Data mining məsələsinin ilkin mərhələsi əlamətlərin layihələndirilməsidir (feature engineering). Bu ən məsuliyyətli və əziyyətli mərhələ olması ilə yanaşı, prosesin nəticəsinə birbaşa təsir edir. Baxılan halda obyeklər kimi şəbəkə istifadəçiləri, əlamətlər kimi isə onların müraciət etdiyi veb-resurslar nəzərdə tutulur. Bu əlamətlərin təsviri nəticəsində istifadəçilərin tematik profili formalaşır. Nəticədə informativ əlamətlərdən ibarət olan istifadəçilər/kateqoriyalar matrisini əldə edirik. Alınmış matris böyük ölçülərə malikdir (cəd. 3),

	Kat1/ həcm	Kat2/ həcm	Kat3/ həcm	Kat1/ zaman	Kat2/ zaman	Kat3/ zaman	Kat1/ müra.	Kat2/ müra.	Kat3/ müra.
İst 1	12Gb	6Gb	800Mb	126dəq	98dəq	22dəq	5355	4742	1586
İst 2	14Gb	4,8Gb	350Mb	148dəq	71dəq	18dəq	10163	3102	1475
İst 3	3,1Gb	2,7Gb	787Mb	78dəq	38dəq	28dəq	608	1554	3217

Cədvəl 3. İnformativ əlamətlərdən ibarət olan istifadəçilər/kateqoriyalar matrisini

bununla belə seyrək matris formasındadır (sparse matrix).

VI. K-ORTA ALQORİTMİ

İstifadəçi profilinin klasterizasiyası üçün biz K-Orta alqoritmini istifadə edəcəyik. Buna səbəb K-Orta alqoritmin klasterizasiya məsələsini həll etmək üçün çox sürətli və sadə olmasıdır. Əgər, $X = \{x_1, \dots, x_n\}$ isə verilənlər toplusu n trafik sessiyalarından ibarətdir. Hər bir trafik-sessiyası d-ölçülü Evklid mühitində bir veriləndir. $x_i = (f_1, \dots, f_d)$, i trafik-sessiyası üçün $f_1, \dots, f_d$ olduğu halda d xüsusiyyətlər dəyəridir. Əsas məqsəd trafik-sessiyalarını ilə klasterlərə bölməkdir. Bu proses zamanı, *n* verilənlərlə müvafiq K klaster "sentroid"-ləri arasındakı məsafə minimum olsun. Hər bir klasterin sentroid kimi tanınan mərkəzi μ_k var və bu qrupun təmsilçisi hesab edilə bilər.

Belə ki, K-orta alqoritminin girişi "nxd" verilənlər matrisi n – dən ibarətdir, K klasterlərin sayı və ilkin verilənlər isə sentroidlərdir. Burada alqoritmin aşağıdakı mərhələlərdən ibarətdir:

1. İlk olaraq "sentroid" qruplarını təmsil edəcək K nöqtələr müəyyən olunmalıdır;
2. Hər bir verilən ilə ən yaxın "sentroid" arasında Evklid-məsafəsini hesablamaq üçün (1) tənliyindən istifadə olunur:

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

3. Bütün nöqtələr müəyyən olunduqdan sonra, K sentroidlərin mövqeləri yenidən hesablanır və bu o

deməkdir ki, bütün nöqtələrin ortası yenidən təyin olunur.

4. 2-ci və 3-cü bənd sentroidlər mövqeyini dəyişməyənədək təkrarlanmalıdır.

VII. KLASTERLƏRİN SAYININ SEÇİLMƏSİ

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

$$daxili = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}\|x - x_n\|^2 \qquad (2)$$

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

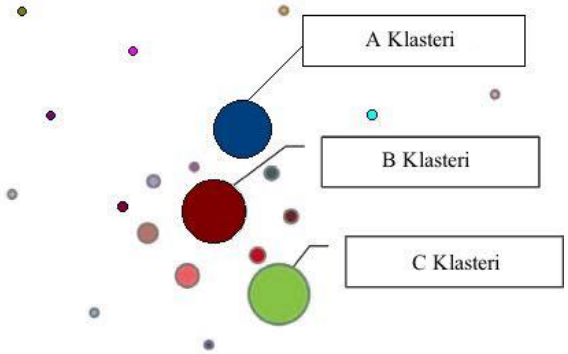
$$arası = min\left(\left\|z_i - z_j\right\|^2\right), i = 1,2,\dots,K-1; \quad j = i+1,\dots,K \qquad (3)$$

Düstur (4) ilə verilmiş etibarlılıq-ölçüsünun minimum dəyəri, K klasterlərin sayını müəyyən etmək üçün istifadə olunur:

$$\text{ölçü} = \frac{daxili}{arası} \qquad (4)$$

VIII. TƏTBİQ

Klasterizasiya modelinin tətbiqi nəticəsində, müəyyən klasterlər formalaşmışdır. Klasterləri əsasən formalaşdıran sosial şəbəkələr, video resurslar və elmi-praktiki resurslardır (şək. 1). Şəkil 1 – də göstərilən nəticə 20 klaster üçün bigml.com resursu [9] vasitəsilə əldə edilmişdir. Ən çox müaraciət olunan A klasteri elmi-praktiki resurslardan ibarətdir. 2-ci ən çox müraciət olunan B klasteri isə sosial şəbəkələrdir. C klasteridir isə, video resurslara olan müraciətlərdən ibarətdir. Digər klasterlərə isə müraciətlər daha azdır. Çünki, istifadəçilərin çoxsu lazım olan informasiyanın çox hissəsini sosial şəbəkələrdən və video resurslardan alırlar.



NƏTİCƏ

Bu məqalədə, AzScienceNet istifadəçilərinin, klasterizasiyası əsasında profilinin müəyyən olunması vasitəsi işlənib hazırlanmışdır. Bunun üçün klasterizasiya metodlarının içində ən sürətli və sadə model olan K-Orta seçilmişdir. Bu prosesin əsas məqsədi şəbəkə resurslarının məqsədə uyğun 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] Kumpulainen P., Hätönen K., Knuuti O., Alapaholuoma T., **Internet traffic clustering using packet header information** / Joint International IMEKO TC1+ TC7+ TC13 Symposium, Jena, Germany, 2011, pp. 13-20

[3] Gerhard M., Sa L., Georg C., **Traffic Anomaly Detection Using K-Means Clustering** / In Proceedings of performance, reliability and dependability evaluation of communication networks and distributed systems, 4GI/ITG-Workshop MMBnet, Hamburg, Germany, 2007, pp. 25-33

[4] Ekola T., Laurikkala M., Lehto T., Koivisto H., **Network traffic analysis using clustering ants** / Proceedings. World Automation Congress, v. 17, Seville, Spain 2004, pp. 275-280

[5] Duo Liu, Chung-Horng Lung, Lambadaṅs I., Seddigh N. **Network traffic anomaly detection using clustering techniques and performance comparison** / Proceedings the 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Canada, 2013, pp.1-4

[6] Shokri, R., Oroumchian F., Yazdani N., CluSID: a clustering scheme for intrusion detection improved by information theory / Proceedings of the 7th IEEE Malaysia International Conference on Communications and IEEE International Conference in Networks, Kuala Lumpur, Malasia, 2005, pp.553-55

[7] http://wiki.squid-cache.org/SquidFaq/SquidLogs

[8] http://www.squid-cache.org/Intro/why.html

[9] http://www.bigml.com