UOT 004.042

Şıxəliyev R.H.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
ramiz@science.az

İNTERNET TRAFİKİNDƏ TRENDLƏRİN AŞKARLANMASININ BİR ÜSULU HAQQINDA

Məqalə İnternet trafikində trendlərin aşkarlanması məsələsinə həsr olunmuşdur. Bunun üçün ardıcıl şablonların aşkarlanması alqoritminin istifadə edilməsi təklif edilir. İnternet trafikində trendlərin aşkar edilməsi kompüter şəbəkələrinin idarə edilməsi zamanı düzgün qərarların qəbul edilməsi üçün çox vacibdir və onların monitorinqi üçün baza tələblərin və mümkün metrikalarının seçilməsinə imkan verər.

Açar sözləri: *İnternet trafiki, İnternet trafikində trendlərin aşkarlanması, ardıcıl şablonların aşkarlanması, tez-tez rast gəlinən elementlər toplusu.*

Giriş

Son on illik ərzində İnternet şəbəkəsinin topoloji mürəkkəbliyinin, istifadəçilərinin, tətbiqlərinin, xidmətlərinin sayının sürətlə artması İnternet trafikinin strukturunda böyük dəyişikliklərə və həcminin həddindən artıq böyüməsinə səbəb olmuşdur.

Müasir İnternet trafiki İnternetə qoşulmuş və onun dəstəklədiyi bütün qurğuların (sistem trafiki), müxtəlif tətbiqlərin və istifadəçilərin yaratdığı böyük həcmdə informasiya axınından ibarətdir. İnternet trafiki istənilən şəbəkənin fəaliyyətinin çox vacib faktiki göstəricilərindən biridir və istifadəçilərin, şəbəkələrin fəaliyyəti haqqında informasiya daşıyır. İnternet trafiki müxtəlif trafiklərin birləşməsindən ibarət olduğu üçün heterogen xarakterə malikdir. Digər tərəfdən, informasiya texnologiyalarının sürətlə inkişafı ilə əlaqədar olaraq yeni tətbiqlər meydana çıxır, digərləri isə istifadədən çıxır, mobil İnternet və sosial şəbəkələr sürətlə inkişaf edir və istifadəsi genişlənir. Yəni İnternet trafikinin təbiəti dinamik və tez-tez dəyişir. Bu şəraitdə şəbəkənin planlaşdırılması, idarə edilməsi, şəbəkə resurslarının bölünməsi, şəbəkə trafikinin idarə edilməsi, şəbəkənin təhlükəsizliyinin təmin edilməsi və s. kimi məsələlərin həlli çətinləşir. Bu məsələlərin həlli üçün İnternet trafikində baş verən qlobal və caritrendlərinaşkar edilməsivə analizi çox vacibdir, lakin yuxarıda göstərilən şəraitdə bu çox çətin məsələyə çevrilir.

İnternet trafikində trendlərin aşkar edilməsişəbəkə trafikinin analizinin çox vacib hissəsidir və kompüter şəbəkələrinin xarakteristikalarının müəyyən edilməsi üsullarından biridir. İnternet trafikindətrendlərin dəqiq aşkar edilməsişəbəkə trafikinin proqnozlaşdırılması vəidarə edilməsi, İnternet xidmətləri üçün şəbəkə resurslarının (buraxma zolağının) effektiv bölünməsi, şəbəkələrin planlaşdırılması, anomaliyaların, hücumların aşar edilməsivə s.kimi məsələlərin effektiv həlli üçün çox vacibdir. Həmçinin İnternet trafikində trendlərin dəqiq aşkar edilməsi və analizi şəbəkə istifadəçiləri üçün yüksək QoS-un təmin edilməsinə imkan verir [1]. Bunun üçün İnternet trafiki daimi monitorinq edilməlidir.

Adətən trendlər sadə xətti reqressiyanın köməyi ilə aşkar edilir. Lakin bu üsulla İnternet trafikindəki trendləri aşkar etmək bir qədər mürəkkəb məsələdir. Çünki İnternet çox mürəkkəb sistemdir və İnternet trafikində trendlər lokal və qlobal xüsusiyyətlərə malik olur. Buna görə də İnternet trafikində trendlərin aşkaredilməsi və analizi üçün trafikdə tez-tez rast gəlinən müəyyən tipik xarakteristikalar – şablonlarmüəyyən olunmalıdır. Başqa sözlə, İnternet trafikində trendlərin aşkarlanması trafikdə tez-tez rast gəlinən müəyyən tipik xarakteristikaları – şablonları əldə etmədən mümkün deyil. Buna görə də İnternet trafikində trendləri aşkarlamaq üçün ən uyğun yanaşma ardıcıl şablonların aşkarlanması (Mining Sequential Patterns) alqoritminin istifadəsi olardı [2, 3]. Ardıcıl şablonlar verilənlər axınında tez-tez rast gəlinən verilənlər toplusu ardıcıllığıdır. Ardıcıl şablonların aşkarlanmasının əsas məqsədizamandan asılıolandavranış prosesinin başa düşülməsidir. Tez-tez rast gəlinən şablonların aşkarlanması verilənlərdə ümumi

maraqlı trendlərin, assosiativ qaydaların, korrelyasiyanın və digər münasibətlərin aşkarlanmasına imkan verir [4], həmçinin ardıcıllıqların klassifikasiyası və klasterləşdirilməsi üçün istifadə edilə bilər.

Məqalənin əsas məqsədi İnternet trafikində trendlərin aşkar edilməsidir və bunun üçün ardıcıl şablonların aşkarlanması alqoritminin istifadəsi təklif olunur. Bu yanaşma İnternet trafikində baş verən müxtəlif trendləri aşkar etməyə imkan verər.

İnternet trafikin xarakteristikalarının analizi

İnternet trafikində ardıcıl şablonların aşkarlanması üçün əvvəlcə onun xarakteristikalarının müəyyən edilməsi çox vacibdir. Bu xarateristikalar İnternet trafikini təsvir edən xüsusiyyətlərin analizi nəticəsində əldə edilə bilər [5]. İnternet trafikin təhlili göstərir ki, o, mürəkkəb dinamik prosesdir və müxtəlif protokollar və tətbiqlər (məsələn, DNS, FTP, WINS, ARP sorğuları, NetBIOS, HTTP, P2P, SMTP, POP3, Telnet seansları və s.) tərəfindən generasiya olunan çoxlu sayda qarşılıqlıəlaqəlixarakteristikalara malik olan axınlardan ibarətdir. Axınların əsas komponentləri ünvan və protokollardır. [6]-da axın iki host arasında paketlər mübadiləsi ardıcıllığı kimi təyin olunmuşdur və beş elementdən (mənbənin IP-ünvanı, mənbənin port nömrəsi, təyinatın IP-ünvanı, təyinatınport nömrəsi və protokolun növü) ibarət olan kortej kimi təyin olunur.

İnternet trafikin analizi bir neçə mücərrəd səviyyələrdə həyata keçirilə bilər. Bu səviyyələrə misal kimi portların nömrəsi, paketin məzmunu, paketin başlığı və trafikin həcmi səviyyələrini göstərmək olar. Bu zaman hər bir şəviyyədə İnternet trafikin xarakteristikaları fərqlənir, məsələn, İnternet trafik paket səviyyəsində paketlərin ölçüsü və paketlərarası zaman intervalları ilə, trafikin həcmi səviyyəsində isə kanalın ötürmə intensivliyi və buraxma qabiliyyəti kimi kəmiyyətlərlə xarakterizə olunur.

Məlumdur ki, IP protokolu İnternet mühitində istifadə edilən istənilən növ tətbiqlərin paketlərinin göndərilməsi üçün universal protokoldur və trafikin daşınması üzrə bütün yük onun üzərinə düşür. Həmçinin məlumdur ki, IP protokolunun fəaliyyətini təmin edən iki əsas nəqliyyat protokolu – TCP və UDP protokolu mövcuddur və İnternet trafiki əsasən TCP-trafikdən ibarətdir. Bu hal IP-trafikin əsas xüsusiyyətlərindən biridir və təxminən 98,2% IP-trafik TCP protokolu haqqında informasiyaya malikdir. Bununla yanaşı, İnternet mühitindəmultimedia və interaktiv tətbiqlərin, həmçinin RTP (Real Time Protocol) və RSVP(Resource Reservation Protocol)kimi genişlənmiş protokolların geniş istifadəsi UDP-trafikinin həcminin böyüməsinə gətirib çıxarır.

İnternet trafikinə böyük çoxölçülü axın kimi baxmaq olar. Bu TCP və IP protokollarının növbəti xüsusiyyətlərinə əsaslanır: TCP və IP başlıqları müxtəlif formatlarda inkapulyasiya olunmuş müxtəlif informasiyalar daşıyır. Bu informasiyalar paketlərin təyinatından və ötürülən məlumatdan asılı olaraq dəyişə bilər. Trafikin paylanması, tətbiqlərin istifadəsi, İnternet-xidmətlərin istifadəsi, protokolların istifadəsi, istifadəçi fəaliyyəti və s. şablonlar İnternet trafikində trendlərin aşkarlanması üçün istifadə edilə bilər.

İlk dəfə İnternet trafikin xarakteristikalarının müəyyənləşdirilməsi məsələlərinə [7, 8]-də baxılmısdır və əsasən axınlar və onları yaradan tətbiqi protokollar arasındakı qarşılıqlı əlaqə müəyyən edilmişdir. Bu işlər göstərdi ki, bir neçə protokolların xüsusiyyətlərinin təsviri üçün təsadüfi dəyişənlərin analitik modelləri istifadə edilə bilər.

[9]-da İnternet trafikin xarakteristikalarının müəyyən edilməsi üçün hostların aktivliyinin statistik analizi metodu təklif edilir. Bu metod əsasında hostların aktivlik şablonları müəyyən edilir.

Ardıcıl şablonların aşkarlanması alqoritmləri

Ardıcıl şablonların aşkarlanması məsələsinin həlli üçün işlənmiş ilk alqoritmlər AprioriAll, AprioriSome və DynamicSome alqoritmlərdir və Rakeş Arval və Ramakrişman Srikant [2]

tərəfindən işlənmişdir. Bütün bu alqoritmlərdə ardıcıl şablonların aşkarlanması məsələsinə iqtisadi sahədə, bazarın analizi kontekstində baxılmışdır, yəni müştərilərin müəyyən tranzaksiyalar çərçivəsində malların alınması zamanı tez-tez rast gəlinən şablonların aşkarlanmasına baxılmışdır.Sonralar onun istifadəsi genişlənərəkdigər sahələrə: telekommunikasiya, biologiya, tibb (məsələn, DNT-nin tədqiqində) və s. sahələrinə tətbiq edilməyə başlanmışdır.Bunun üçün müxtəlif alqoritmlər işlənmişdir, məsələn, GSP(Generalized Sequential Pattern) [3], SPADE(Sequential Pattern Discovery using Equivalent Class) [10], PrefixSpan [11] və s. kimi alqoritmlər işlənmişdir və nəticədə bu sahə qısa müddətdə çox inkişaf etmişdir.

$X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n, t(X_i) \leq t(X_j), i \leq j$ şəkilindəki ardıcıllıq ardıcıl şablonlar adlanır. Burada, X hadisə və ya hadisələr çoxluğu, t isə zamandır. Beləliklə, ardıcıl şablonlar zamana görə nizamlanmış hadisələr çoxluğudur. Ardıcıllıqdakı hadisələrə o qədər tez-tez rast gəlinir ki, bu da onlar arasında əlaqələrin olmasını deməyə imkan verir. Belə əlaqələrin analizi müəyyən qaydaların aşkarlanmasına imkan verir. Yəni, bir şablondan müəyyən hadisələrin baş verməsi çox böyük ehtimalla bu şablondan olan digər hadisələrin və ya hadisənin baş verməsinə gətirir. Bu zaman birinci qrup əsas hadisələr, ikinci qrup, yəni baş verməsi gözlənilən hadisələr isə məqsəd hadisələr adlanır.

Ardıcıl şablonlar nəzəriyyəsi assosiativ qaydalar nəzəriyyəsindən yaranmışdır və müəyyən dərəcədə oxşardırlar. Çünki, hər iki halda elementlər toplusu, tranzaksiyalar anlayışları, dəstək və etibarlılıq kimi kəmiyyət xarakteristikaları istifadə olunur. Tez-tez rast gəlinən şablonların aşkarlanması üçün isə Apriori alqoritminin müxtəlif modifikasiyaları istifadə edilir. Lakin ardıcıl şablonlar və assosiativ qaydalar arasında əsaslı fərqlər mövcuddur. Məsələn, assosiativ qaydalarda sadəcə tranzaksiyalarda elementlərin birlikdə rastgəlmə faktı maraq kəsb edir və bu zaman elementlərin rastgəlmə ardıcıllığına baxılmır. Əksinə, ardıcıl şablonlarda elementlərin rastgəlmə ardıcıllığı mühüm rol oynayır, çünki belə nəzərdə tutulur ki, əvvəlki elementlər sonrakı elementlərin meydana gəlmə ehtimalına təsir göstərir.

Ümumi halda, ardıcıl şablonların aşkarlanması məsələsi elə S ardıcıllığının tapılmasından ibarətdir ki, I verilənlər çoxluğu üçün dəstəyin verilmiş $minsup$ hədd qiymətində $sup(s) \geq minsup$ olsun. Bununla yanaşı, böyük həcmdə verilənlər çoxluğundan tez-tez rast gəlinən ardıcıllıqların aşkarlanması çox mürəkkəb məsələdir, çünki bu zaman axtarış fəzası həddindən artıq böyük olur. Məsələn, m atributları üçün k uzunluqlu $O(m^k)$ sayda tez-tez rast gəlinən ardıcıllıqlar mövcuddur [12]. Qeyd etmək lazımdır ki, böyük uzunluqlu ardıcıllığa malik olan onlayn İnternet trafikində ardıcıl şablonların aşkarlanması çox mürəkkəb məsələdir və həlli üçün böyük vaxt və yaddaş tələb olunur.

Ardıcıl şablonların aşkarlanması məsələsini çox çətin edən və böyük vaxt sərfiyyatına gətirən müəyyən səbəblər mövcuddur. Birincisi, şablonlar tək bir element üçün deyil, həmçinin elementlər çoxluğu üçün formalaşdırılır. İkincisi, nə şablonda olan elementlər çoxluğunun, nə də elementlər çoxluğunda olan elementlərin sayı aprior olaraq məlum olmur. Üçüncüsü, şablonların formalaşdırılması verilənlər çoxluğuna daxil olan elementlərin istənilən kombinasiyasının yerdəyişməsi əsasında yaradıla bilər [13].

Bütün bunlarla yanaşı, ardıcıl şablonların aşkarlanmasının bəzi əsas konsepsiyaları mövcuddur və bunlar aşağıdakılardır [2]:

1. Tutaq ki, $I = \{x_1, \ldots, x_n\}$ elementlər çoxluğudur və hər bir element müəyyən atributlara– xarakteristikalara malikdir (məsələn, qiymət, məsafə, period və s.). x elementinin A atributunun qiyməti xA kimi işarə olunur. Elementlər çoxluğu boş olmayan elementlər altçoxluğundan ibarətdir və k elementdən ibarət olan elementlər çoxluğu $k-$elementlər çoxluğu adlanır.

2. $S = \langle X_1 \dots X_l \rangle$ ardıcıllığı nizamlanmış elementlər çoxluğudur. Ardıcıllıqdakı $X_i (1 \leq i \leq l)$ elementlər toplusu çoxluğu tranzaksiya adlanır. X tranzaksiyası xüsusi atributa malik ola bilər, məsələn, zaman nişanına malik ola bilər və $X_i t$ kimi işarə edilir. Zaman nişanı tranzaksiyanın başvermə vaxtını göstərir. $S = \langle X_1 \dots X_l \rangle$ ardıcıllığı üçün tranzaksiyaların zaman nişanı $X_i t < X_j t, 1 \leq i < j \leq l$ kimi işarə olunur.

3 Ardıcıllıqdakı tranzaksiyaların sayı ardıcıllığın uzunluğu adlanır. Uzunluğu l olan ardıcıllıq l-ardıcıllıq adlanır. $S = \langle X_1 \dots X_l \rangle l$-ardıcıllıq üçün $L(S) = l$. Digər tərəfdən i-ci element$S[i]$ kimi işarə olunur. Hər bir element bir elementlər toplusu daxilində yalnız bir dəfə, müxtəlif elementlər toplusundaisə bir neçə dəfə rast gəlinə bilər.

4. $S = \langle X_1 \dots X_l \rangle$ardıcıllığı $S' = \langle Y_1 \dots Y_m \rangle (n \leq m)$ ardıcılllığının altardıcıllığı adlanır.$1 \leq i_1 < .. < i_n \leq m$üçün elə $X_1 Y_{i_1} \dots X_n Y_{i_n}$tam ədədlər mövcuddur ki, S' ardıcıllığı S ardıcıllığının superardıcıllığı adlanır.Əgər S ardıcıllığı digər heç bir ardıcıllığın altardıcıllığı deyilsə, onda o, maksimal ardıcıllıq adlanır. Bəzi hallarda bütün ardıcıl hadisələrin şablonlarının tapılması deyil, yalnız onlardan maksimal olanlarının tapılması tələb olunur.

5. Ardıcıllıq verilənlər bazası (SDB – Sequence Database)(sid, S) kortejlər çoxluğundan ibarətdir. Burada sid (sequence-id)ardıcıllığın identifikatorudur, S isə ardıcıllıqdır. Əgər γ ardıcıllığı ardıcıllıq verilənlər bazasındakı (sid, S) kortejinin altardıcıllığıdırsa, onda deyillər ki, (sid, S) korteji γ ardıcıllığına malikdir.γ ardıcıllığına malik ardıcıllıq verilənlər bazasındakı(sid, S) kortejinin sayı γ ardıcıllığının dəstəyi adlanır və $sup(\gamma)$ kimi işarə olunur.Tutaq ki, dəstəyin həddqiymətini göstərən $\min\_sup$ müsbət ədəd verilmişdir və əgər $sup(\gamma) \geq \min\_sup$, onda γardıcıllığı ardıcıllıq verilənlər bazasının ardıcıl şablonudur. Ardıcıl şablonların aşkarlanması məsələsi verilmiş SDB ardıcıllıq verilənlər bazası və $\min\_sup$ həddqiyməti üçünbütün ardıcıl şablonlarıntapılmasından ibarətdir.

6. Bir kliyentin bütün tranzaksiyaları tarixə, zamana və ya müraciətin nömrəsinə görə nizamlanmış ardıcıllıq şəklində göstərilə bilər. Belə ardıcıllığı kliyent ardıcıllığı adlandırırlar. Formal şəkildə bu aşağıdakı şəkildə yazılır.

Tutaq ki, kliyent zamana görə nizamlanmış bir neçə $T_1, T_2, \dots, T_k$ tranzaksiyaları həyata keçirib. Onda T_i tranzaksiyasındakı hər bir elementlər çoxluğu $I(T_i)$ şəklində işarə edilir və verilmiş kliyent üçün hər bir kliyent ardıcıllığını isə $< I(T_1), I(T_2, \dots, I(T_k)) >$ kimi yazılır. Digər sözlə, kliyent ardıcıllığı verilmiş kliyent tərəfindən həyata keçirilmiş tranzaksiyalara müvafiq elementlər toplusu ardıcıllığıdır.

Əgər S ardıcıllığı kliyent ardıcıllığında mövcuddursa, onda bu ardıcıllıq kliyent tərəfindən dəstəklənən adlanır. Onda ardıcıllığın dəstəyi onu dəstəkləyən kliyentlərin sayı ilə müəyyən edilir.

Ardıcıl şablonların aşkarlanması prosesi aşağıdakı mərhələlərdən ibarətdir [2]:

1. *Çeşidləmə*. İlkin verilənlər bazasının tranzaksiyaları kliyentlərin identifikatoruna, hər bir kliyentin tranzaksiyası isə tarixə, zamana və yaxud ziyarətin nömrəsinə görə çeşidlənir. Beləliklə, ilkin verilənlər bazası kliyent ardıcıllıqları bazasına çevrilir.

2. *Tex-tez rast gəlinən elementlər toplusunun axtarılması.* Bu mərhələdə bütün L tez-tez rast gəlinənelementlər toplusu axtarılır. Eyni zamanda tez-tez rast gəlinən bütün 1-ardıcıllıqlar axtarılır, çünki bu sadəcə $\{< l > | \, l \in L\}$ çoxluğudur. Sonra tez-tez rast gəlinən elementlər toplusu hərflər, tam ədədlər və ya ikilik ardıcılllıq şəklində (məsələn, a, b, 2, 3, və s.) alternativ təsvirə çevrilir. Bu cür təsvirin istifadəsi məsələnin alqotitmik həllini sadələşdirməyə imkan verir.

3. *Çevrilmə.* Tez-tez rast gəlinən ardıcıllıqlardan hansılarının kliyentardıcıllığında olması müəyyən edilir. Bunun üçün kliyen tranzaksiyaları ardıcıllıqları onun tez-tez rast gəlinən elementləri toplusuçoxluğu ilə əvəz edilir. Əgər kliyent tranzaksiyasında heç bir tez-tez rast gəlinən elementlər toplusu yoxdursa, onda çevirmə nəticəsində bu tranzaksiyaya ümumiyyətlə baxılmır.Əgər kliyent ardıcıllığında heç bir tez-tez rast gəlinən elementlər toplusu yoxdursa,

onda bu ardıcıllığa da baxılmır.Çevrilmədən sonra hər bir kliyent ardıcıllığı tez-tez rast gəlinən elementlər toplusunnun$\{l_1, l_2, \dots, l_n\}$ çoxluğu şəklində təsvir edilir, burada l elementlər toplusudur.

Çevrilmiş ilkin D verilənlər bazasını D_T kimi işarə edilir. Sistemin disk yaddaşının həcmindən asılı olaraq çevrilmə bütün verilənlər üzərində və ya hər dəfə kliyent ardıcıllığı bazadan oxunan zaman həyata keçirilə bilər.

4. Tez-tez rast gəlinən ardıcıllığın axtarılması. Tez-tez rast gəlinən elementlər toplusu çoxluğunu istifadə etməklə tez-tez rast gəlinən ardıcıllıqlar axtarılır. Bu mərhələ ən çətin mərhələlərdən biridir, çünki tez-tez rast gəlinən ardıcıllıqların tapılması üçün böyük həcmdə elementlər toplusu arasından böyük sayda mümkün ardıcıllıqlara baxılmalıdır.

Tez-tez rast gəlinən ardıcıllığın axtarılması üçün istifadə edilən alqoritmlərin ümumi strukturu ondan ibarətdir ki, onlar verilənlər üzərindən bir neçə dəfə keçid edirlər. Bu zaman hər bir keçid yeni mümkün ardıcıllıqların generasiyası üçün əvvəlki tapılmış ardıcıllıqlar toplusundan başlayır və bu yeni mümkün ardıcıllıqlar namizəd-ardıcıllıqlar və ya sadəcə namizədlər adlanır. Bunun üçün onların dəstəyi hesablanır və keçid başa çatdıqdan sonra aşkarlanmş namizədlərin həqiqətən də tez-tez rast gəlinən olduğu yoxlanılır. Aşkar olunmuş tez-tez rast gəlinən namizəd-ardıcıllıqlar yeni keçid üçün başlanğıc olacaq. Birinci keçiddə 2-ci mərhələdə tapılmış tez-tez rast gəlinən elementlər toplusunun minimal dəstəyini ödəyən bütün 1-ardıcıllıqlar axtarılır. Bunun üçün iki alqoritmin – dolğun və qismən axtarış alqoritmlərinin istifadəsi mümkündür. Dolğun axtarış alqoritmi maksimal olmayan ardıcıllıqlar daxil olmaqla bütün tez-tez rast gəlinən ardıcıllıqları axtarır. Daha sonra qeyri-maksimal ardıcıllıqlar çıxarılır. Dolğun axtarışı həyata keçirən alqoritmlərdən biri AprioriAll alqoritmidir və Apriori alqoritminə əsaslanır.

Bundan başqa, qismən ardıcıllıqların axtarılması üçün AprioriSome və DynamicSome kimialqoritmlər mövcuddur. Bu alqoritmlər ancaq maksimal ardıcıllıqlarla işləyir ki, bu da qeyri-maksimal ardıcıllıqların emalından qaçmağa imkan verir və ardıcıllıqları düz və əks mərhələdə emal edir. Birinci mərhələdə müəyyən uzunluqlu bütün tez-tez rast gəlinən ardıcıllıqlar, ikinci mərhələdə isə tez-tez rast gəlinən ardıcıllıqlar tapılır. Bu ardıcıllıqlar arasındakı əsas fərq onların düz mərhələdə namizəd-ardıcıllıqların formalaşdırılması üçün istifadə etdiklərіprosedurlardadır.

Tez-tez rast gəlinən ardıcıllığın axtarılması üçün istifadə edilən bu və ya digər alqoritmlər [2, 14–16]-da geniş analiz edilmişdir.

5. Maksimal ardıcıllıqların axtarılması. Bu mərhələdə tez-tez rast gəlinən ardıcıllıqların arasından maksimal ardıcıllıqların axtarışı həyata keçirilir. Bəzi hallarda qeyri-maksimal ardıcıllıqların hesablanmasına sərf edilən vaxtın azaldılması üçün bu mərhələni əvvəlki mərhələ ilə birləşdirilir.

İnternet trafikində trendlərin aşkarlanması məsələsi

Bu gün böyük əminliklə demək olar ki, istənilən kompüter şəbəkəsində, xüsusilə də İnternet şəbəkəsində trafikin böyük hissəsi istifadəçilərin təşəbbüsü ilə yaranır. Buna görə də, İnternet trafikində bu və ya digər trendlərin baş verməsində istifadəçitrafiklərinin xarakteristikalarının bilavasitə təsiri vardır. Adətən, istifadəçi trafiki çoxlu sayda müxtəlif əlamətlərlə xarakterizə olunur və bu əlamətlər monitorinq verilənləri kimi istifadə edilir. Bu əlamətlər keyfiyyətində paketlərin nəqliyyat (TCP-protokolunun), şəbəkə (IP-protokolunun) və tətbiqi səviyyələrin xüsusiyyətləri, məsələn, mənbənin IP-ünvanı, mənbənin port nömrəsi, təyinatın IP-ünvanı, təyinatın port nömrəsi, istifadə olunan protokolların növü(məsələn, HTTP, FTP, SMTP və s.), istifadə olunan tətbiqlərin növü (məsələn, WWW, FTP, P2P, email və s.) xidmətin növü, paketlərin sayı, daxil olan və göndərilən trafikin həcmi və sürəti, zaman və s. istifadə olunur.

Bununla yanaşı, İnternet trafiki müxtəlif böyük siniflərə (məsələn, Chat, Interactive, VoIP, P2P və s.) bölünür və yaxudprotokollar (FTP, HTTP, SSH və s.) əsasında daha kiçik siniflərə bölünür.

İnternet istifadəçilərinin fəaliyyətinin müxtəlifliyinə baxmayaraq, onların trafikində ümumi trendlərin aşkar edilməsi mümkündür.Buna görədə İnternet trafikində trendlərin aşkar edilməsi məsələsinə istifadəçi trafiklərində trendlərin aşkar edilməsi məsələsi kimi baxmaq olar, yəni istifadəçilərin İnternetdən istifadəsinin analizi (Internet usage mining) məsələsi kimi baxmaq olar. Bu məsələnin həllinin ilkin mərhələsiİnternet trafikindaimimonitorinqedilməsi və istifadəçilərin İnternetdən istifadəsi haqqındakıverilənlər loq fayllarda toplanmasıdır.

Qeyd edək ki, bu loq fayllardakı verilənlər istifadəçilərin İnternet trafiklərini xarekterizə edir vəbu trafiklərdəki trendlərin aşkar edilməsinəimkan verir. Bütün deyilənləri nəzərə alaraq İnternet trafikində trendlərin aşkar edilməsi məsələsiİnternet trafikində ardıcıl şablonların aşkarlanması məsələsi kimitəyin edilir.

Tutaq ki, İnternet istifadəçilərinin trafikinin monitorinqini təsvir edən D verilənlər bazası verilmişdir. Bu bazanın hər bir yazısı istifadəçi trafikini təsvir edir.Bu zaman belə bir məhdudiyyət əlavə edilir ki, eyni bir zaman anında istifadəçi yalnız bir trafik yarada bilər. Hər bir trafik növbəti sahələrdən (atributlardan) ibarətdir: istifadəçi identifikatoru (istifadəçinin IP-ünvanından), tarix/zaman, istifadə olunan protokol(məsələn, HTTP, FTP, SMTP və s.), tətbiq (məsələn, WWW, FTP, P2P, email və s.), daxil olan və göndərilən trafikin həcmi, sürəti və s. Bu atributları$I = \{x_1, \ldots, x_n\}$ çoxluğu kimi işarə edirik. Ardıcıllığı isə $S = \langle X_1 \ldots X_l \rangle$ şəklində işarə edirik və nizamlanmış atributlar toplusundan ibarətdir. Ardıcıllıqdakı $X_i (1 \le i \le l)$ atributlar toplusunutrafikadlandırırıq.

Lakin, çox zaman şəbəkə monitorinqi verilənlərini adi standart verilənlər bazası şəklində təsvir etmək mümkün olmur. Məsələn, toplanmış paketlər ixtiyari məlumatlara malik ola bilər və paketlər ardıcıllığından TCP axınını əldə etmək üçün paket başlığının istənilən sahəsinə və ya tətbiq sahəsinin başlıqlarına baxıla bilər.

Bütün yuxarıda deyilənləri nəzərə alaraq İnternet trafikində ardıcıl şablonların aşkarlanması məsələsi,verilmişİnternet istifadəçilərinin trafikinin monitorinqini təsvir edən D verilənlər bazası və $\min\_sup$ hədd qiyməti üçün bütün ardıcıl şablonlar arasından maksimal ardıcıllıqların tapılması məsələsi kimi təyin edilir. Bu məsələnin həlli nəticəsində tapılmış maksimal ardıcıllıqlar ardıcıl şablonlar adlanır.

İnternet trafikində ardıcıl şablonların aşkarlanması alqoritmi

İnternet istifadəçilərinin trafikinin monitorinqini təsvir edən D verilənlər bazasında ardıcıl şablonların aşkarlanması üçün istifadəçilər tərəfindən müəyyən minimum dəstəyə malik olan bütün ardıcıllıqlar arasından maksimal ardıcıllıqlar tapılmalıdır. Bu məsələni həll etmək üçün yuxarıda göstərilən alqoritmiitifadə edərək alırıq:

1. İnternet istifadəçilərinin trafikinin monitorinqini təsvir edən D verilənlər bazası istifadəçi identifikatoruna (istifadəçinin IP ünvanına görə), hər bir istifadəçi trafiki isə zamana (qısamüddətli trendlərin aşkarlanması zamanı) və ya tarixə (uzunmüddətli trendlərin aşkarlanması zamanı) görə çeşidlənir. Beləliklə, İnternet istifadəçilərinin trafikinin monitorinqini təsvir edən Dverilənlər bazası istifadəçi ardıcıllıqları bazasına çevrilir.
2. Birinci mərhələdə alınmış istifadəçi ardıcıllıqları bazasındabütün L tez-tez rast gəlinən elementlər toplusu axtarılır. Eyni zamanda tez-tez rast gəlinən bütün 1-ardıcıllıqlar axtarılır, çünki bu sadəcə $\{< l > | \, l \in L\}$ çoxluğudur. Sonra tez-tez rast gəlinən elementlər toplusu hərflərin köməyi ilə (məsələn, HTTP – a,TCP – b, SMTP – c, FTP – d və s.) alternativ təsvirə çevrilir. Bu cür təsvirin istifadəsi məsələnin alqoritmik həllini sadələşdirməyə imkan verir.
3. Tez-tez rast gəlinən ardıcıllıqlardan hansılarının istifadəçi ardıcıllığında olması müəyyən edilir. Bunun üçün istifadəçi trafikləri ardıcıllıqları onun tez-tez rast gəlinən elementlər toplusu çoxluğu ilə əvəz edilir. Əgər istifadəçi trafikində heç bir tez-tez rast gəlinən elementlər toplusu

yoxdursa, onda çevirmə nəticəsində bu trafikə ümumiyyətlə baxılmır. Əgər istifadəçi ardıcıllığında heç bir tez-tez rast gəlinən elementlər toplusu yoxdursa, onda bu ardıcıllığa da baxılmır. Çevrilmədən sonra hər bir istifadəçi ardıcıllığı tez-tez rast gəlinən elementlər toplusunun $\{l_1, l_2, \dots, l_n\}$ çoxluğu şəklində təsvir edilir, burada l elementlər toplusudur.

Çevrilmiş ilkin D verilənlər bazası D_T kimi işarə edilir və sistemin disk yaddaşının həcmindən asılı olaraq çevrilmə bütün verilənlər üzərində və ya hər dəfə istifadəçi ardıcıllığı bazadan birbaşa oxunan zaman həyata keçirilə bilər.

4. Bu mərhələdə İnternet trafikində tez-tez rast gəlinən elementlər toplusu çoxluğunu istifadə etməklə tez-tez rast gəlinən ardıcıllıqlar axtarılır.

5. Nəhayət sonuncu mərhələdə İnternet trafikində tez-tez rast gəlinən ardıcıllıqların arasından maksimal ardıcıllıqların axtarışı həyata keçirilir.

4-cü mərhələdə bütün böyük S ardıcıllıqlar çoxluğu tapıldıqdan sonra bu çoxluqda maksimal ardıcıllığın tapılması üçün aşağıdakı alqoritm istifadə edilə bilər. Tutaq ki, ən uzun ardıcıllığın uzunluğu n-dir, onda:

for $(k = n; k > 1; k -)$ **do**

foreach $\boldsymbol{k}$-sequence s_k **do**

S ardıcıllığından bütün s_k ardıcıllıqları silinsin.

Verilmiş ardcıllıqda bütün alt ardıcıllıqların tez tapılması üçün verilənlərin strukturu (heş-ağac) və alqoritmi [17]-də verilmişdir.

Tez-tez rast gəlinən ardıcıllığın axtarılması üçün işlənmiş alqoritmlərin [2, 11, 14–16] analizi nəticəsində İnternet trafikində ardıcıl şablonların, dolayısı ilə trendlərin aşkarlanması üçün PrefixSpan alqoritminin istifadəsi təklif edilir.

PrefixSpan alqoritminin əsas ideyası verilənlər bazasının tez-tez rast gəlinən elementlərə uyğun olaraq bir neçə kiçik verilənlər bazasına proyeksiya olunması və şablonların uzunluğunun artırılmasından ibarətdir [11].

Bunun üçün verilmiş bazadan bütün tez-tez rast gəlinən elementlər tapılır, onlar cari şablona əlavə edilir, bununla da yeni tez-tez rast gəlinən ardıcıllıqlar alınır və bundan sonra proyeksiya olunmuş bazalar əsasında böyük uzunluqlu tez-tez rast gəlinən ardıcıllıqlar axtarılır. PrefixSpan alqoritminin psevdokodu listinq 1-də göstərilmişdir.

PrefixSpan(s, D|s):

Parametrləri: s – ardıcıl şablonlardır;
$D|_s$ – ilkin D bazasının s-proyeksiyasıdır, əgər s boş<> ardıcıllıq deyilsə, əks halda $D|_s = D$.

1. *$D|_s$-dən bütün x tez-tez rast gəlinən elə elementlər tapılsın ki:*

a) *x tez-tez rast gəlinən ardıcıllıq yaradılması üçün s-ə daxil olan sonuncu elementlər toplusuna əlavə oluna bilər;*

b) *X elementlər toplusu tez-tez rast gəlinən ardıcıllıq yaradılması üçün s-ə əlavə oluna bilər.*

2. *Hər bir tez-tez rast gəlinən x elementi üçün:*
   yeni s' şablon yaradılması üçün x elementi s-ə əlavə olunur;
   s' şablonunu nəticəyə əlavə olunur.

3. *Hər bir yeni s' şablonu üçün:*
   $D|_{s'}$ s'-proyeksiyasının qurulması;
   PrefixSpan (s',$D|_{s'}$)proseduru işə salınır.

Listinq 1. PrefixSpan alqoritmi

D bazasındakı bütün ardıcıl şablonların tapılması üçün PrefixSpan (<>, D) proseduru istifadə olunur.

Proyeksiya olunmuş bazaların yaradılması böyük həcmdə verilənlər bazası ilə işləyən zaman bu alqoritmin məhsuldarlığına təsir edə bilər. Buna görə də, fiziki proyeksiyanın əvəzinə psevdoproyeksiya istifadə edilir və nəticədə alqotimin işləmə sürəti xeyli yüksəlir, həmçinin alqoritmin işləməsi üçün daha kiçik həcmdə yaddaş tələb olunur [18].

Alqoritmin fəalliyyəti nəticəsində tapılmış bütün maksimal ardıcıllıqlar verilmiş İnternet istifadəçilərinin trafikinin monitorinqini təsvir edən D verilənlər bazasında aşkar edilmiş ardıcıl şablonları göstərir.

Nəticə

İnternet trafikində trendlərin aşkarlanması kontekstində trafiklərin analizi məsələsində ardıcıl şablonların aşkarlanması alqoritmlərinin istifadəsi çox vacibdir. İnternet trafikində aşkarlanan şablonların ardıcıllığına əsaslanaraq bu və ya digər şablonların rast gəlinməsi haqqında qabaqcadan proqnoz vermək olar və bu da kompüter şəbəkələrinin effektiv idarə edilməsi üçün düzgün qərarların qəbul edilməsinə imkan verər.

Minnətdarlıq

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişaf Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – Qrant № EİF-RİTN-MQM-2/İKT-2-2013-7(13)-29/27/1.

Ədəbiyyat

1. Zeng B. D., Zhang W. Li, Zhang M., Hong Q. An adaptive sampling methodology for internet traffic data measurement / Proceedings of the International Conference on Communication Software and Networks, 2009, Feb. 27–28, pp.215–218.
2. Agrawal R., Srikant R. Mining Sequential Patterns // Journal Intelligent Systems, 1997, vol.9, no.1, pp.33–56.
3. Agrawal R., Srikant R. Mining sequential patterns: Generalizations and performance improvements / Proceedings of the 5th International Conference on Extending Database Technology, 1996, pp.1–17.
4. Han J., Kamber M., Data mining: concepts and techniques. Morgan Kaufmann, 2006.
5. Lamparter O. and Stauffer B., A network traffic measurement tool / Proceedings of the 10th International Conference on Telecommunications, 2003, Feb. 23-Mar., vol.1, pp.1078–1083.
6. Zhanh L., Tang J. Characterization and performance study of IP traffic in WDM networks // Computer communications, 2001, no.24, pp.1702–1713.
7. Paxson V. Empirically derived analytic models of wide-area TCP connections / IEEE / ACM Trans. Netw., 1994, vol.2, no.4, pp.316–336.
8. Paxson V. and Floyd S., Wide area traffic: the failure of Poisson modeling / IEEE/ACM Trans. Netw., 1995, vol.3, no.3, pp.226–244.
9. Karagiannis T., Papagiannaki K., Faloutsos M. BLINC: multilevel traffic classification in the dark / Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Pomputer Communications, 2005, New York, pp.229–240.
10. Zaki M. J. Spade: An efficient algorithm for mining frequent sequences // Machine Learning 2001, vol.42, no.1–2, pp.31–60.
11. Pei J., Han J., Mortazavi-Asl B., etc. Mining sequential patterns by pattern-growth: The prefixspan approach / IEEE Transactions on Knowledge and Data Engineering 2004, vol.16, no.11, pp.1424–1440.
12. Zaki M.J. Scalable data mining for rules, Technical Report Ph.D. Dissertation, University of Rochester, New York, 1998.
13. Ming-Yen Lin, Suh-Yin Lee. Interactive Sequence Discovery by Incremental Mining // An International Journal of Information Sciences-Informatics and Computer Science, 2004, vol.165, no.3–4, pp.187–205.

14. Mabroukeh RN., Ezeife C. I. A taxonomy of sequential pattern mining algorithms // Journal ACM Computing Surveys, 2010, vol.43, no. 3.
15. Chandra V. Shekhar Rao, Sammula P. Survey on Sequential Pattern Mining Algorithms / International Journal of Computer Applications, 2013, vol.76, no.12, pp.24–31.
16. Parikh M., Chaudhari B. and Chand C., A Comparative Study of Sequential Pattern Mining Algorithms // International Journal of Application or Innovation in Engineering and Management, 2013, vol.2, no.2, pp.103–109.
17. Agrawal R. and Srikant R., Mining sequential patterns. Research Report RJ9910, IBM Almaden Research Center, San Jose, California, October 1994.
18. Pei J., Han J., Mortazavi-Asl B., etc. PrefixSpan: mining sequential patterns efficiently by prefix projected pattern growth / Proceedings of the 17th International Conference on Data Engineering, 2001, pp.215–226.

УДК 004.042

Шыхалиев Рамиз Х.
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

Об одном методе обнаружения трендов в интернет-трафике

Статья посвящена задаче обнаружения трендов в интернет-трафике. Для этого предлагается использовать алгоритм обнаружения последовательных шаблонов. Обнаружение трендов в интернет-трафике необходимо для принятия правильных решений при управлении компьютерных сетей и позволит выбрать базовые требования и возможные метрики для их мониторинга.

Ключевые слова: *интернет-трафик, обнаружение трендов в интернет-трафике, обнаружение последовательных шаблонов, набор часто встречаемых элементов.*

Ramiz H. Shikhaliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
ramiz@science.az

One method for Internet traffic trends detection

Article devoted to the problem of trends detection in Internet traffic. For solution of this problem we has proposed to use the algorithm of sequential patterns mining. Detecting trends in Internet traffic is very necessary to make the right decisions in the management of computer networks and to allow choosing the basic requirements and possible metrics to monitor them.

Key words: *Internet traffic, detection of Internet traffic trends, sequential patterns mining, itemsets.*